Constructing Shared Space: Designing for Asymmetry in Collaborative Mixed Reality

Nels Numan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

March 13, 2026

I, Nels Numan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Mixed-reality technologies enable embodied collaboration across distance, yet collaborators often differ in their interfaces, viewpoints, and physical contexts. These *asymmetries* shape what participants can perceive, reference, and act upon, creating challenges for mutual understanding and coordination. Environmental representations, the computational constructs that determine what each collaborator perceives about an environment and how views are related across participants, play a central role in mediating these asymmetries. This thesis investigates how environmental representations in asymmetric collaborative mixed reality can be constructed, augmented, and transformed to incorporate contextual information, examining their influence on mutual understanding, coordination, and task performance.

The research comprises three empirical investigations combining system design with user studies. The first study examines an asymmetric configuration in which a remote virtual reality user collaborates from within a static reconstruction of a local augmented reality user's workspace (n = 30, in dyads). Findings reveal that shared space is established through ongoing grounding work rather than delivered by the reconstruction alone, with technological asymmetries shaping spatial practices of exploration and attention. The second study addresses task-driven augmentation in site-specific outdoor augmented reality authoring, where pre-captured representations are often incomplete and temporally misaligned with current site conditions. Through the design and evaluation of CoCreatAR, an asymmetric collaborative authoring system (n = 32, in pairs), findings demonstrate that anchored visual augmentation supports repair and verification of proxy representations, increasing developer confidence while

introducing coordination overhead. The third study explores representational transformation through SpaceBlender, a pipeline that synthesises unified virtual environments from multiple users' physical surroundings using generative scene completion. A preliminary evaluation (n = 20, in pairs) indicates that structural coherence is a prerequisite for transformed spaces to remain usable for navigation and grounding, while contextual cues from familiar spaces show potential to become coordination resources.

This thesis contributes empirical evidence on how asymmetric configurations shape the establishment and maintenance of shared space, systems that augment and transform environmental representations to address representational limitations, and implications for the design of collaborative mixed-reality systems that must bridge asymmetries of perspective, context, and control. The findings point to future directions including broadening explorations of asymmetric configurations beyond the device pairings examined, exploring artificial intelligence as a mediator of asymmetry, and leveraging advances in environmental reconstruction and generative techniques to address the representational limitations identified in this work.

Impact Statement

As mixed-reality technologies become increasingly prevalent in remote work, education, and entertainment, enabling people to collaborate effectively across distance remains a fundamental challenge. When collaborators use different devices or occupy different physical spaces, they often struggle to establish mutual understanding because they perceive and interact with shared environments in fundamentally different ways. This thesis investigates how the computational representations of environments that underpin mixed-reality collaboration can be designed to bridge these differences, with implications for how future collaborative technologies support distributed teams.

This research has achieved academic impact through publication in leading peer-reviewed venues in human-computer interaction. The three core investigations have been published at ACM VRST 2022 [2], ACM CHI 2025 [15], and ACM UIST 2024 [12], with Chapter 4 receiving an Honourable Mention Award at CHI 2025. Chapter 3 provides empirical evidence on how technological differences shape spatial practices and coordination when one collaborator is virtually relocated into another's workspace, offering guidance for systems that embed remote users into local environments. Chapter 4 demonstrates how real-time visual context sharing between remote and on-site collaborators can address long-standing challenges in location-based content creation, informing the design of future collaborative authoring workflows. Chapter 5 introduces a novel approach to generating shared virtual environments from multiple users' physical surroundings, opening new research directions at the intersection of generative artificial intelligence and collaborative systems. Beyond these core contributions, Ubiq-Genie [5], a framework for integrating generative artificial intelligence into

collaborative mixed-reality applications, has been released as open-source software and adopted by multiple independent research projects [22, 10], enabling further scholarship and student projects. International research collaborations supported by European Union Horizon 2020 funding [4] have further extended the reach of this work and supported opportunities for future cross-institutional partnerships.

The research also demonstrates strong potential for commercial and professional practice impact through multiple patent filings. The collaborative augmented reality authoring approach described in Chapter 4, conducted during an internship at Niantic Labs, has an associated patent filing (US19/451,480) and received significant attention from the AR development community upon announcement. The generative environment blending approach described in Chapter 5 has an associated patent filing (US18/756,717) by Microsoft Research. A complementary exploration conducted alongside this work, BlendScape [13], applies similar generative techniques to video-teleconferencing environments and has received an Honourable Mention Award at ACM UIST 2024; this direction has particular relevance for widely used platforms such as Microsoft Teams. Additional internships at Google have resulted in further patent filings, published research [17, 21], and the release of XR Blocks [19], an open-source framework for building extended reality experiences on the web.

Collectively, this body of work can benefit individuals and organisations who rely on remote collaboration by informing the design of platforms that better accommodate diverse devices, physical contexts, and interaction capabilities. As distributed collaboration continues to grow across professional, educational, and social domains, the findings, systems, and design implications contributed by this thesis offer a foundation for building mixed-reality experiences that improve mutual understanding, coordination, and ultimately the quality of remote collaborative work.

Acknowledgements

This work was supported by the European Union's Horizon 2020 Research and Innovation programme as part of project RISE under grant agreement No. 739578.

First and foremost, I want to express my deepest gratitude to my primary supervisor, Anthony Steed, for his continuous support and insightful guidance throughout this research journey. His knowledge and encouragement have been instrumental to my growth as a researcher. I am truly grateful for the chance to work under his supervision and cherish our many discussions that have shaped the research direction and my broader understanding of the field.

I would also like to recognise my secondary supervisor, Simon Julier, for his incredible enthusiasm and expertise, which never failed to inspire me and helped me think outside the box. His feedback and suggestions sparked so many new ideas and directions, both inside and outside my PhD research area.

I would like to thank everyone in the Virtual Environments and Computer Graphics group at UCL for creating such a welcoming and inspiring environment. Special thanks to my colleagues and friends, including Ziwen, Daniele, David, Sebastian, Ben, Klara, Lisa, Kalila, Jingyi, and Phoenix, for the countless discussions and all the fun along the way.

I have been fortunate to collaborate with many talented researchers during my PhD through various collaborations and internships, many of whom have become dear friends. At Microsoft Research, my mentors Bala Kumaravel, Nic Marquardt, and Andy Wilson taught me so much about research, from formulating ideas that look beyond the state of the art to developing them into solid contributions. Your mentorship during my internship and beyond has meant so much to me. I would also like to thank my fellow intern, Shwetha Rajaram, for

being a wonderful collaborator whom I learned so much from. To all my fellow interns and friends at Microsoft Research, thank you for the great memories.

At Niantic Labs, I would like to thank my mentors Jessica Van Brummelen and Gabriel Brostow for their exceptional guidance and support every step of the way. Working with you both was an incredible learning experience, and I am so grateful for your mentorship. I would also like to thank Jaewook Lee, my fellow intern, for the inspiring chats and fun times during and after the internship.

At Google, I would like to extend my heartfelt thanks to Mar Gonzalez-Franco for her trust and incredible mentorship in research and beyond. Thank you for pushing my thinking and always making space for questions and learning. I would also like to thank Eric Gonzalez and Harish Kulkarni for being such wonderful hosts and mentors. Your feedback, encouragement, and genuine care made a huge difference in my journey. Thank you to Ruofei Du for so warmly welcoming me to the Bay Area, and for constantly challenging me with questions that stretched my thinking. I would also like to thank my colleagues, fellow interns, and friends at Google, including Andrea Colaco, Xiaoan Liu, Hongyu Mao, Geonsun Lee, Yuran Ding, David Li, Xun Qian, and many others, for contributing to such a meaningful experience.

This work would not have been possible without the generosity of my study participants. I am very grateful to everyone who shared their time, experiences, and thoughtful feedback.

To my family and friends—thank you for filling this journey with love and support. To my friends, thank you for always being there to celebrate the highs and lift me through the lows. To my sister, you have been there every step of the way, listening to every worry and cheering for every win. I am so lucky to have you. To my mom, thank you for teaching me to follow my passions and for believing in me even when I doubted myself. I am forever grateful. And to my wife, Suhyun, thank you for your infinite love, patience, and dedication—from standing by me through the toughest times to celebrating every achievement, big or small. I truly could not have done this without you by my side.

Research Paper Declaration

This thesis is based on work that has been previously published elsewhere. A full list of publications during the research period of this dissertation is provided in the Appendix (Chapter A).

Chapter 3 is based on the following publication [2]:

> **Numan**, **Nels** and Steed, Anthony. 2022. Exploring User Behaviour in Asymmetric Collaborative Mixed Reality. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. VRST '22: 28th ACM Symposium on Virtual Reality Software and Technology. ACM, New York, NY, USA, 11. doi:10.1145/3562939.3565630

Chapter 4 is based on the following publication [15]:

> **Numan**, **Nels**, Brostow, Gabriel, Park, Suhyun, Julier, Simon, Steed, Anthony, and Van Brummelen, Jessica. 2025. CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). ACM, New York, NY, USA, (Apr. 25, 2025). ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3714274
>
> *Received an Honourable Mention Award* | *Related Patent Filing* US19/451,480

Chapter 5 is based on the following publication [12]:

> **Numan**, **Nels**[*], Rajaram, Shwetha[*], Kumaravel, Balasaravanan Thoravi, Marquardt, Nicolai, and Wilson, Andrew D. 2024. SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (UIST '24). ACM, New York, NY, USA, (Oct. 13, 2024). doi:10.1145/3654777.3676361
>
> *Related Patent Filing* US18/756,717

All three publications have been peer-reviewed. The copyright for each publication is held by the respective authors, with publication rights licensed to the Association for Computing Machinery (ACM). The division of research contributions among co-authors is outlined as a footnote in each chapter.

Contents

1 Introduction **27**

1.1 Problem Statement 31

1.2 Research Approach 33

1.3 Dissertation Structure 36

2 Background **37**

2.1 Foundational Concepts of Mediated Collaboration 37

2.1.1 Context 39

2.1.2 Copresence, Social Presence, and Related Concepts . . . 40

2.1.3 Awareness 42

2.1.4 Common Ground 43

2.2 Early Audio-Visual Approaches to Mediated Collaboration 45

2.2.1 Capturing Users and Their Communication Cues 45

2.2.2 Capturing and Representing Action 46

2.3 Asymmetry as a Design Condition in Collaborative Systems . . . 48

2.3.1 Taxonomies of Asymmetry in Collaborative Work and Games 49

2.3.2 Asymmetry as a Source of Awareness Breakdown 50

2.3.3 Asymmetry as a Deliberate Design Strategy 52

2.4 Foundational Concepts of Immersive Experience 54

2.4.1 Definitions and Taxonomies of Mixed Realities 54

2.4.2 Immersion 57

2.4.3 Presence 58

2.4.4 Embodiment 61

2.5 Conceptual Models of Collaborative Mixed Realities 62

2.5.1 Taxonomies of Shared Space in Collaborative Systems . . 62
2.5.2 Boundary-Based Spatial Coupling 63
2.5.3 Volumetric Spatial Coupling 65
2.5.4 Task-Driven Extensions of Volumetric Coupling 66
2.6 Types of Asymmetry in Collaborative Mixed Realities 68
2.6.1 Technological Asymmetry: Interface Differences 70
2.6.2 Role Asymmetry 76
2.6.3 View and Information Asymmetry 78
2.6.4 Environmental Context Asymmetry 81
2.7 Representational Strategies for Managing Asymmetry 83
2.7.1 Augmenting Shared Space to Support Grounding and Awareness 84
2.7.2 Providing Context Through Environmental Representation 86

3 Exploring User Behaviour in Asymmetric Collaborative Mixed Reality 95
3.1 Context and Motivation 96
3.2 Hypotheses 97
3.3 System Design 99
3.3.1 Apparatus 99
3.3.2 Virtual Replica of Physical Space 99
3.3.3 Coordinate Space 99
3.3.4 Experimental Space 100
3.3.5 Recordings 101
3.3.6 Interactions 101
3.4 Method 101
3.4.1 Design 101
3.4.2 Participants 102
3.4.3 Task 102
3.4.4 Procedure 102
3.4.5 Measures 103
3.5 Results 105

3.5.1 Differences Across HMD Types . . . 106
3.5.2 Relations Between Measures . . . 109
3.6 Discussion . . . 112
3.6.1 Insights from the Statistical Analysis . . . 113
3.6.2 Grounding Work and Breakdowns in Spatial Referencing 116
3.6.3 Emergent Coordination Strategies and Shared-Space Management . . . 117
3.6.4 Static Environmental Representation . . . 118
3.6.5 Limitations . . . 119
3.7 Conclusion . . . 120

4 CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration 122
4.1 Context and Motivation . . . 123
4.2 Related Work: AR Experience Authoring . . . 125
4.2.1 In-Situ Authoring . . . 126
4.2.2 Ex-Situ Authoring . . . 127
4.2.3 Collaborative Approaches to AR Experience Authoring . . 128
4.3 Formative Study . . . 129
4.3.1 Method & Participants . . . 129
4.3.2 Findings . . . 131
4.4 CoCreatAR System . . . 137
4.4.1 System Requirements . . . 137
4.4.2 Scene View and Localisation . . . 139
4.4.3 Real-Time In-Situ View Streaming and Audio Communication . . . 140
4.4.4 Spatial References with 3D Cursors . . . 140
4.4.5 Auto-Networked Objects and Scene Updates . . . 140
4.4.6 Spatial Context Capture with Coarse Meshes and 3D Snapshots . . . 141

4.4.7 Annotation of the Real World with Surface Draw and Air Draw . . . 142
4.5 Exploratory User Study . . . 142
4.5.1 Baseline . . . 144
4.5.2 Participants . . . 145
4.5.3 Task . . . 146
4.5.4 Procedure . . . 149
4.5.5 Measures and Analysis . . . 150
4.6 Exploratory User Study Results . . . 151
4.6.1 Questionnaire Results . . . 151
4.6.2 Interview Results . . . 154
4.7 Discussion . . . 162
4.7.1 Enhancing Environmental Understanding Through Capture and Communication Tools . . . 162
4.7.2 Task Load and Multitasking Challenges . . . 164
4.7.3 Engagement and Creative Exploration . . . 166
4.7.4 Confidence in End-User Experience Through In-Situ Feedback . . . 167
4.7.5 Enabling Hybrid Approaches to AR Experience Authoring 168
4.8 Limitations . . . 169
4.9 Conclusion . . . 170

5 SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending 172

5.1 Context and Motivation . . . 173
5.2 Related Work: 3D Scene Generation . . . 175
5.3 SpaceBlender System . . . 178
5.3.1 Requirements . . . 178
5.3.2 System Overview . . . 180
5.3.3 *Stage 1*: From 2D Images to 3D Layout . . . 182

5.3.4 *Stage 2*: Iterative Blending Guided by Geometric Priors and Contextual Prompts . . . 186
5.4 Preliminary User Study . . . 190
5.4.1 Task . . . 191
5.4.2 Recruitment . . . 191
5.4.3 Implementation and Setup . . . 192
5.4.4 Procedure . . . 192
5.4.5 Data Collection & Measures . . . 193
5.5 Results . . . 195
5.5.1 Self-Reported Questionnaire Results . . . 196
5.5.2 Benefits and Limitations of Environments for the Clustering Task . . . 197
5.5.3 Future Requirements and Potential Use Cases for Blended Spaces . . . 199
5.6 Discussion and Future Work . . . 201
5.6.1 Positioning SpaceBlender Within Shared Space Construction Models . . . 201
5.6.2 Opportunities for Supporting VR Telepresence with Blended Environments . . . 202
5.6.3 Opportunities to Improve and Extend SpaceBlender . . . 206
5.6.4 Limitations . . . 207
5.7 Conclusion . . . 209

6 Discussion **211**
6.1 Addressing the Research Questions . . . 211
6.1.1 Perception and Enactment of Shared Space . . . 211
6.1.2 Limitations of Pre-Captured Environmental Representations . . . 213
6.1.3 Augmenting Environmental Representations . . . 216
6.1.4 Transforming Environmental Representations . . . 217
6.2 Future Work . . . 219

6.2.1 Expanding Empirical Scope and Ecological Validity . . . 219
6.2.2 Broadening Explorations of Asymmetric Configurations . 219
6.2.3 Toward Artificial Intelligence as a Collaborator in Mixed Reality . . . 221
6.2.4 Leveraging Advances in Environmental Reconstruction and Representation . . . 225

7 Conclusion **227**

Appendices **230**

A Publications **230**

B Appendix – Chapter 3 **234**
B.1 Post-Study Questionnaire . . . 234
B.2 Velocity Filter Analysis . . . 235
B.3 Supplementary Correlation Tables . . . 235

C Appendix – Chapter 4 **242**
C.1 Formative Study: Tools Summary . . . 242
C.2 System Walkthrough . . . 243
C.3 Task Design: AR Prototypes in Two Locations . . . 245
C.3.1 *Location* A: Welcoming Tourists to the City . . . 245
C.3.2 *Location* B: Promoting and Celebrating the Upcoming Cookie Party . . . 247
C.3.3 Task Components – Issues . . . 248
C.4 Full Results of CoCreatAR User Study . . . 249
C.4.1 Engagement . . . 250
C.4.2 Task Load . . . 251
C.4.3 Confidence in Authored Result . . . 252
C.4.4 Feature usage during the sessions . . . 254
C.5 Survey and Semi-Structured Interview Questions . . . 254
C.5.1 Formative Study Questions . . . 254

C.5.2 User Study 258

D Appendix – Chapter 5 **264**

D.1 Contextually Adaptive Prompt Inference: Full System Prompt . . 264

D.2 ControlNet-Layout: Additional Output Samples 266

D.3 Comparison of ControlNet-Layout and LooseControl 267

D.4 Influence of Submesh Shapes on Geometric Prior Shape 268

D.5 Examples of Spaces Generated with Various Numbers of Submeshes 269

D.6 Preliminary User Study: Questionnaire Items 270

D.7 Preliminary User Study: Walkthrough Scenarios 271

D.8 Preliminary User Study: Semi-Structured Interview Questions . . 272

D.9 Preliminary User Study: Full Analysis 272

D.9.1 Possible Actions 272

D.9.2 Self-Location 273

D.9.3 Co-Presence 273

D.9.4 Impact of Environmental Factors 273

List of Figures

1.1 Overview of asymmetry types in collaborative mixed reality (CMR). 30

2.1 The Reality—Virtuality (RV) continuum by Milgram et al. [129]. . 54
2.2 Subdimensions of the Reality—Virtuality (RV) continuum by Milgram et al. [129]. 55
2.3 Revised dimensions of the Reality—Virtuality (RV) continuum by Skarbez et al. [130]. 56
2.4 Benford et al.'s [46] classification of shared spaces along three dimensions: transportation, artificiality, and spatiality. Image adapted from the original. 64
2.5 Illustrative example of technological asymmetry in CMR. One collaborator uses an augmented reality (AR) head-mounted display (HMD) and hand tracking for manipulation of a cube (left), while the other uses a virtual reality (VR) HMD and controllers (right). 70
2.6 Illustrative example of role asymmetry in CMR. A remote collaborator assumes an expert role with oversight and guidance capabilities (left), while the local collaborator performs hands-on manipulation (right). 76
2.7 Illustrative example of view and information asymmetry in CMR. 79
2.8 Illustrative example of environmental context asymmetry in CMR. 81

3.1 Conceptual overview of our CMR system, in which an AR and a VR user can collaborate remotely. 95

3.2 Overview of the collaborative virtual environment (CVE) used by Steed et al. [52] 97

3.3 Overview of an AR user's (left) and a VR user's (right) perspectives. 101

3.4 Plot of head positions for each HMD type for the two periods of Trial 14. 118

4.1 Overview of CoCreatAR collaboration between ex-situ developers and in-situ users. 122

4.2 The five main site-specific AR issue types we aimed to address with CoCreatAR. 130

4.3 The number of participants who reported using various methods to document issues in their AR experiences when in situ. 132

4.4 The number of participants who reported looking for various issues in their AR experiences when in situ. 133

4.5 In-situ user interface of CoCreatAR. 136

4.6 Ex-situ user interface of CoCreatAR. 138

4.7 *In-situ* user interface of the baseline system. 143

4.8 Overview of the study procedure. 146

4.9 Screenshots of the prototypes designed for *Location* A and *Location* B. 148

4.10 Plot of engagement and task load scores. 153

4.11 Stacked bar plot illustrating response proportions for items in which participants indicated perceived performance and confidence for each task component across *collaboration mode* and *location* combinations. 154

4.12 Overview of CoCreatAR feature usage during Phase 1, shown as *ex-situ* perspective screenshots. 158

5.1 Overview of SpaceBlender. 172

5.2 A bird's-eye view of two meshes that failed to blend due to a lack of geometric guidance and insufficient context during the iterative mesh completion process. 179

5.3 Outer view of a room generated by TEXT2ROOM. 180

5.4 Overview of *Stage 1* components of SPACEBLENDER. 181

5.5 Comparison of unaligned and aligned submeshes. 183

5.6 Overview of *Stage 2* components of SPACEBLENDER. 186

5.7 Comparison of outputs generated with varying weights of ControlNet depth and layout models. 187

5.8 Overview of the environments used for the GENERIC3D and TEXT2ROOM conditions. 190

5.9 Captures of participants manipulating sticky notes while represented by Ubiq avatars in the individual (left) and collaborative (right) task phases. 192

5.10 Overview of SPACEBLENDER meshes generated from input images uploaded (with green outline) or selected by participants (no outline). 194

5.11 Plot of score distributions for Possible Actions, Self-Location, Copresence, and task impact factors. 195

5.12 Plot of participant preferences per condition. 196

5.13 Examples of errors in SPACEBLENDER pipeline components: (A) protrusion of an outdoor structure, (B) distorted table, (C) depth estimation error, (D) depth alignment error, (E) depth estimation error, (F) floor expansion error. 204

5.14 Photograph of an illustration from Do Ho Suh's *Perfect Home*. . . 206

6.1 Extending the extent of world knowledge dimension 215

6.2 Overview of *AdjustAR* [18], which corrects misaligned AR content at runtime. 222

B.1 Yaw velocity of the AR and VR HMDs with different sliding-window sizes. 235

C.1 The number of participants who reported using various AR development/design tools on the formative survey. 242
C.2 The number of participants who reported using various AR frameworks on the formative survey. 243
C.3 The virtual objects in Prototype A. 245
C.4 The virtual objects in Prototype B. 246
C.5 Box plot representing the distributions of feature counts per group. 254

D.1 Additional results of the ControlNet-Layout model. 266
D.2 Comparative image generation output of the recent LooseControl model and our ControlNet-Layout model. 267
D.3 Visual explanation of the impact of submesh count and shape on the shape of the geometric prior mesh and final blended space. . 268
D.4 Examples of spaces generated with various numbers of submeshes. 269
D.5 Overview of environments used in the walkthrough segment of the semi-structured interview. 271

List of Tables

2.1 The computer-supported cooperative work (CSCW) Space–Time matrix, adapted from Johansen [72]. 38

3.1 Results of paired t-tests for all variables between HMD conditions ($N = 30$). 107

3.2 Regression of leadership on IAS, reported separately by period, device, and sex. 108

3.3 Results of paired t-tests for all variables between trial periods ($N = 30$). 109

3.4 Spearman's rank-order correlations (r_s) and p-values between head rotation velocity and social metrics ($N = 26$). 111

B.1 Post-trial questionnaire based on prior work [68, 52, 53]. 234

B.3 Spearman correlation coefficients between head rotation velocity and leadership and talkativeness variables for the *second* trial period. 237

B.4 Spearman correlation coefficients between head rotation velocity and leadership and talkativeness variables for the *AR HMD* condition. 238

B.5 Spearman correlation coefficients between head rotation velocity and leadership and talkativeness variables for the *VR HMD* condition. 239

B.6 Results of t-tests for all variables between the two HMD type conditions (AR and VR), split by the two trial periods. 240

B.7 Results of t-tests for all variables between the two trial periods, split by the HMD type conditions (AR and VR). 241

C.1 Design patterns, referent types, alignment types, and issue types in AR element design. 249

D.1 Post-task questionnaire for measuring Self-Location and Possible Actions. 270

Acronyms

3D Gaussian splatting (3DGS) A method for representing 3D scenes using a collection of Gaussian functions, allowing for efficient rendering and manipulation of complex geometries and appearances. 92, 176, 177, 225

augmented reality (AR) A type of mixed reality that overlays digital content onto the physical world, enhancing the user's perception of their real environment. 17, 18, 22, 30, 34, 54, 57, 69–72, 74–77, 80, 85, 89, 91, 95, 96, 98–102, 106–120, 122–133, 137–148, 151, 162, 166–170, 173, 212, 216, 220, 222–224, 226, 227, 242–245, 247–252

augmented virtuality (AV) A type of mixed reality where objects, environments, or people are integrated into a virtual environment. 54, 57

Cave Automatic Virtual Environment (CAVE) A room-sized immersive virtual reality (VR) environment where users can experience and interact with 3D computer-generated environments projected onto the walls and floor of the room. 72, 80

collaborative mixed reality (CMR) A type of mixed reality that supports multiple users interacting within a shared mixed reality (MR) environment.. 17, 28–32, 34, 37, 48, 53, 61–63, 66, 68–71, 74, 76, 79–81, 83, 84, 86, 90–92, 94, 95, 98, 104, 113, 115, 137, 171, 172, 175, 201, 211, 217, 220, 221, 223, 225, 227–229

collaborative virtual environment (CVE) Virtual environments (VEs) that allow multiple users to interact and collaborate, often in virtual reality (VR). 18, 40, 41, 46, 47, 57, 60, 62, 95–97, 99, 100, 105, 116, 118, 174

computer-supported cooperative work (CSCW) An interdisciplinary field that studies how people work together using computer technology, focusing on the design and use of collaborative systems and tools. 21, 27, 28, 36–40, 42, 43, 46, 48–52, 62, 63, 68, 79

degrees of freedom (DoF) The degrees of freedom that define the movement and orientation of an object in 3D space, typically including translation along the X, Y, and Z axes (3DoF) and rotation around those axes (6DoF). 74, 76

eXtended reality (XR) A broad term that encompasses all forms of immersive technologies, including augmented reality (AR), virtual reality (VR), and mixed reality (MR). 57

field of view (FOV) The extent of the observable environment that can be seen at any given moment, typically measured in degrees. In the context of MR, it refers to the area visible through a head-mounted display (HMD) or other viewing device. 85, 116, 117, 164, 212

global positioning system (GPS) A satellite-based navigation system that provides location and time information anywhere on or near the Earth, commonly used for mapping and tracking in various applications, including augmented reality (AR). 245

handheld display (HHD) A handheld device that combines a display and input capabilities, often used in augmented reality (AR) applications to provide users with a portable interface for interacting with digital content overlaid on the physical world. 51, 71, 73, 76, 77, 97

head-mounted display (HMD) A wearable device that places a screen or screens in front of the user's eyes, often used in virtual reality (VR) and augmented reality (AR) applications to provide immersive visual experiences. 17, 18, 21, 30, 58, 70–73, 76, 77, 80, 97, 98, 100–103, 105–109, 113, 115–119, 164, 191, 192, 207

large language model (LLM) A type of artificial intelligence model that is trained on vast amounts of text data to understand and generate human-like language. Examples include OpenAI's GPT series and Google's Gemini models. 175, 177, 180, 182, 184, 185, 189

mixed reality (MR) An umbrella term encompassing both augmented reality (AR) and virtual reality (VR), referring to technologies that blend physical and digital environments to varying degrees. 27, 28, 33, 37, 54, 55, 57–59, 62–65, 69, 77, 81, 83, 86–88, 103, 104, 120, 202, 207, 209, 214, 223

multimodal large language model (MLLM) An extension of large language models (LLMs) that can process and generate content across multiple modalities, such as text, images, and audio, enabling more versatile interactions and applications. 222

neural radiance field (NeRF) A neural network-based technique for representing 3D scenes as continuous volumetric functions, enabling high-quality rendering and view synthesis from sparse input images. 92, 176, 225

optical see-through (OST) A type of augmented reality (AR) display that allows users to see the real world directly through transparent optics, with digital content overlaid onto their view. 72

Place Illusion (PI) The sensation of being in a real place, contributing to the sense of presence in virtual reality (VR). 58, 59, 61

Plausibility Illusion (Psi) The illusion that the events and interactions occurring within a virtual environment are actually happening, contributing to the sense of presence in virtual reality (VR). 58, 59, 61, 82

Reality—Virtuality (RV) The continuum that encompasses the full spectrum of experiences from the real world to a fully virtual environment, including augmented reality (AR), mixed reality (MR), and virtual reality (VR). 17, 54–57, 59

simultaneous localisation and mapping (SLAM) A computational technique used in robotics and augmented reality (AR) to simultaneously build a map of an unknown environment while tracking the device's location within that environment. 87

three-dimensional (3D) Relating to or denoting an object or space that has three dimensions (e.g., height, width, and depth). 27, 28, 35, 74, 96, 99, 174, 175, 216, 217, 228

two-dimensional (2D) Relating to or denoting an object or space that has two dimensions (e.g., height and width). 27, 28, 74, 96, 174, 175

video see-through (VST) A type of augmented reality (AR) display that uses cameras to capture the real world and then displays a composited view on a screen, blending the real and virtual elements. 117

virtual environment (VE) A computer-generated simulation of an environment, often in three-dimensional space. 61

virtual reality (VR) A type of mixed reality that immerses users in a fully digital environment, typically through the use of head-mounted displays (HMDs) that block out the physical world. 17, 18, 30, 34, 35, 55, 57–59, 69–76, 85–87, 89, 91, 95, 96, 98–102, 105–120, 172, 174, 180, 191, 192, 212, 213, 217, 220, 221, 227, 228

vision-language model (VLM) An extension of large language models (LLMs) that can directly process visual content alongside text, often used for tasks such as image captioning and visual question answering.. 177, 180, 182–185

visual positioning system (VPS) A technology that uses visual data from cameras to determine a device's position and orientation in space, often used in location-specific augmented reality (AR) applications. 123, 131, 134, 135, 137, 139

Chapter 1

Introduction

Distributed collaboration has become routine across contemporary work, education, and entertainment. Collaborators now coordinate across physical locations and time zones, often using a wide range of devices [23, 24]. A key focus in computer-supported cooperative work (CSCW) research has been how mediated settings shape the establishment and maintenance of *common ground*, the shared knowledge and interactional understandings that collaborators rely on to interpret one another and coordinate action [25]. Common ground is sustained through ongoing *grounding work*, as collaborators signal, assess, and repair understanding during interaction. This process depends on what the system renders *mutually available*, meaning information that participants can jointly access and use as shared evidence for reference and for assessing alignment [26].

Many systems, due to technical and practical limitations, organise collaboration around a digital interaction space that remains only loosely coupled to participants' physical environments. In video conferencing, for example, interaction is confined to two-dimensional (2D) screens, separating the shared digital space from the spaces in which participants act. This separation reduces the mutual availability of embodied and spatial context that often anchors reference and interpretation in co-located interaction. Spatial reference frames become less stable, task state becomes harder to perceive, and embodied cues that support coordination and awareness are less accessible than in co-located settings [27, 28, 29, 30].

Mixed-reality (MR) systems are often positioned as a response to these limitations by supporting embodied interaction in three-dimensional (3D) environments that span physical and virtual realities [31, 32, 33]. In principle, spatial

representations in MR can make surrounding environments and participants' embodied actions jointly perceivable and available for reference, reintroducing resources for awareness, coordination, and grounding that can be difficult to sustain through 2D interfaces alone. The resulting design problem is to determine which spatial and contextual cues become mutually available, and how these cues can support grounding across collaborators' situated actions and interpretations.

In practice, however, collaborative MR (CMR) rarely provides symmetric access to these cues [34, 33, 35]. Many CMR scenarios involve collaborators with different perceptual access, representational fidelity, and interaction capabilities, often shaped by differences in physical setting, device type, and role within the activity, such as a remote expert supporting a local collaborator [36, 37, 38, 39]. Differences in knowledge and perspective are inherent to collaboration [40], yet CMR mediates how they take effect and can itself introduce further *asymmetry* in what collaborators can perceive, how they can act, and which aspects of context become mutually available. As a result, asymmetry is a structural condition of many CMR systems [41, 42, 43, 44]. Consequently, differences in devices, roles, viewpoints, and physical settings become grounding problems by altering the spatial cues available as shared evidence and the stability of reference frames across collaborators.

> *"Space is so fundamental to our everyday experience that it permeates the way we think." – Dourish [45]*

A key theme across spatial approaches to CSCW is that familiar spatial layouts can act as an interactional scaffold by leveraging everyday spatial practices for coordination [46, 47, 45]. As Harrison and Dourish argued, "spatially-organised systems will support spatially-managed behaviors" [47]. However, a shared environment is not guaranteed by what a representation depicts or how it is structured. A geometric 3D model does not automatically result in agreement about what matters in the scene, which actions are appropriate, and how spatial relations should ground reference and action. Harrison and Dourish [47]

distinguished *space* from *place*, arguing that collaborative conduct depends on socially organised understandings of what a setting is and what actions it makes intelligible. Building on this framing, Dourish [48, 45] cautioned against treating context as a stable set of attributes that systems can simply encode, arguing instead that "context and activity are mutually constitutive" [48].

In CMR, this implies that capturing and rendering spatial structure does not guarantee convergence on what the representation means, what is task-relevant within it, or how it should support grounding. Representations condition the resources available for grounding, while collaborators establish their meaning through situated interaction [47, 45]. When representations are partial, unstable, or misaligned, collaborators must interpret what is shown, negotiate relevance, and repair misunderstandings [25]. For example, early media space systems provided only partial and intermittently informative views of remote activity due to camera framing, occlusion, and limited capture of task-relevant detail, which meant that collaborators often inferred off-screen action and repaired misunderstandings through interaction [27, 28]. As Luff et al. stated, even sophisticated collaborative systems can "fracture the relation between conduct and the environment in which it is produced and understood" [49, p. 73].

This dissertation focuses on *environmental representations*, the computational constructs that shape what each collaborator can perceive about an environment, how that environment is structured for interaction, and how relationships across collaborators' views are maintained. Environmental representations can take multiple forms, including geometric, image-based, symbolic, and hybrid representations that combine these elements [50]. In CMR, they organise access to contextual cues and help stabilise the reference frames that support spatial language, deictic action, and mutual awareness [29, 49]. Their role becomes especially important in asymmetric configurations, where devices, roles, viewpoints, and physical settings influence what can be rendered, perceived, and treated as shared evidence. A shared environment is established through collaborative practice, as participants develop a working sense of what the set-

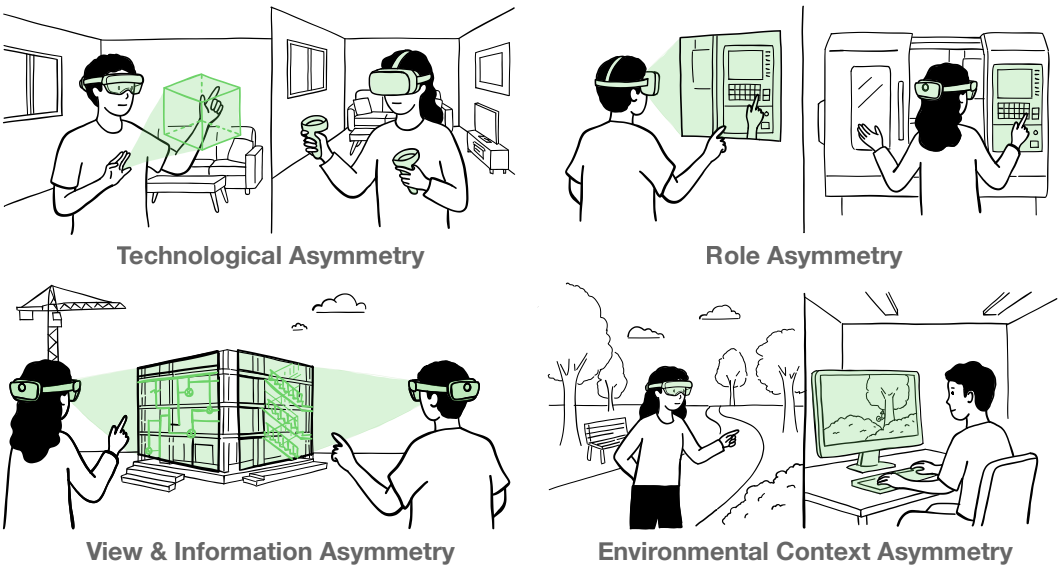


Figure 1.1 Overview of asymmetry types in CMR.

ting affords, what counts as relevant, and how representational cues should be interpreted over time [47]. Representational choices therefore condition how collaborators maintain awareness of others' focus and changes, coordinate action, and repair breakdowns in reference and awareness [51, 47].

In this dissertation, *asymmetry* (i.e., heterogeneity) refers to systematic differences in collaborators' access to views, capabilities, information, or control that arise from technical configuration, physical setting, or role structure [42]. To situate the research goal, we distinguish four recurring asymmetry types in CMR: a) *technological asymmetry*, where collaborators use different devices and interaction languages (e.g., augmented reality (AR) head-mounted displays (HMDs) versus virtual reality (VR) HMDs or desktop tools), resulting in differences in mobility, field of view, input expressivity, and output fidelity; b) *role asymmetry*, where collaborators adopt roles with different responsibilities and privileges (e.g., a remote expert versus an in-situ worker), often aligned with device capabilities and task demands; c) *view and information asymmetry*, where collaborators have different views of the workspace or differential access to task-relevant information, shaping what can be referenced and what is mutually visible; d) *environmental context asymmetry*, where collaborators differ in their access to, and interpretation of, one another's immediate surroundings, includ-

ing the spatial arrangement of artefacts and place-level constraints that shape situated action and communication. Figure 1.1 illustrates these asymmetry types. These asymmetry types frequently co-occur and can have consequential effects on collaboration in CMR [43, 44].

For example, *technological asymmetry* can reconfigure participation and authority by shifting initiative, attention, and perceived leadership toward collaborators with broader views or greater control over the shared environment [52, 53, 54, 55]. Such shifts can undermine mutual awareness or marginalise certain contributions when asymmetries are not accounted for. At the same time, *role asymmetry* and *view and information asymmetry* can be leveraged to support division of labour, coordination efficiency, and performance gains through complementary perspectives and affordances [56, 57]. Prior work therefore increasingly treats asymmetry as a structural condition that can be shaped through design. This perspective has informed the design of shared spaces that selectively expose, transform, or align collaborators' access to information and action. Examples include creating intentional *view and information asymmetry* through object visibility modifiers that emphasise or suppress elements for different collaborators [58, 59]; mitigating *environmental context asymmetry* through reconstructions of local physical context that improve remote interpretability and deictic grounding [60, 61, 62]; and accommodating *technological asymmetry* by adapting interfaces to differences in input expressivity, interactional dimensionality, and representational dimensionality [63, 64].

This perspective positions asymmetry as a representational concern in CMR, beyond a difference in device configurations alone. The following problem statement turns from this broader design context to the role of environmental representations in supporting grounding across uneven access to context.

1.1 Problem Statement

Conceptual models and surveys have established asymmetry as a recurrent condition of CMR [42, 43, 44, 65]. Controlled studies have further shown that asymmetric configurations can shape collaboration dynamics and performance [55,

56, 53, 52]. Yet less is known about how environmental representations mediate collaboration under such conditions. In particular, prior work has only partially addressed how representational choices affect the construction, maintenance, and repair of shared space across divergent devices, viewpoints, roles, and physical contexts. This gap has direct consequences for CMR design, because environmental representations condition (1) which spatial and contextual cues become available as shared evidence for grounding and reference, and (2) how collaborators sustain awareness of one another's actions and focus across uneven access to context. These issues motivate the research objective and questions outlined below.

> **Research Objective** To investigate how environmental representations in asymmetric CMR can be constructed, augmented, and transformed to incorporate contextual information, and to examine their influence on mutual understanding, coordination, and task performance.

To address this objective, we pose the following research questions.

> **Research Question 1** How is shared space perceived and enacted in asymmetric CMR scenarios?

We examine how collaborators experience and interpret shared space when their viewpoints, interfaces, and situated contexts differ, including how these differences shape coordination practices and role dynamics in unidirectional configurations, such as embedding a remote collaborator into a representation of a local collaborator's workspace.

> **Research Question 2** What limitations do pre-captured environmental representations present for supporting mutual understanding and task coordination in asymmetric CMR?

We identify where representational strategies break down, technically and

interactionally, when applied to asymmetric configurations, including breakdowns in awareness, reference, and representational consistency as environments change.

> **Research Question 3** How can environmental representations be augmented to support specific collaborative tasks across asymmetric configurations?

We investigate task-driven augmentation in settings where environmental dynamics and partial access are central, asking how hybrid representations and spatial annotations can support grounding, coordination, and shared understanding during synchronous collaboration.

> **Research Question 4** How can computational techniques transform spatial representations to support shared understanding across distributed MR contexts?

We examine how computational transformation can reshape or synthesise new forms of spatial representations for collaborators that are distributed across different physical contexts.

These research questions guide the dissertation and are operationalised in different ways across the empirical chapters. Each chapter specifies its operationalisation and measures, and Chapter 6 relates these operationalisations back to the research questions.

1.2 Research Approach

Methodologically, this dissertation draws inspiration from the *extended user-centred iterative design process* proposed by Harper et al. [66], which framed *understand*, *study*, *design*, *build*, and *evaluate* as recurring activities within iterative system development. While different stages assumed primary importance at different points across the dissertation, *understand* provided a recurring foundation, anchored in the review of prior work in Chapter 2 and revisited in each

chapter to specify task-relevant context. The work addressing RQ1 and RQ2 emphasises *understand* and *study* through the characterisation of collaboration under asymmetry and the identification of representational limitations. In contrast, RQ3 and RQ4 focus on *design*, *build*, and *evaluate*, developing and assessing augmentation and transformation techniques for environmental representations.

In Chapter 3, we operationalise RQ1 and RQ2 through an asymmetric AR–VR collaboration scenario in which a remote VR user is virtually relocated into a reconstructed model of the in-situ AR user's environment [2]. Prior CMR research examined user behaviour and experience under asymmetry, including effects related to immersion, embodiment, role assignment, and interface capability [67, 68, 52, 53, 55, 54]. However, the literature provides limited direct evidence for the specific unidirectional configuration in which one collaborator's physical environment serves as the shared workspace by relocating the remote partner into a virtual replica. We isolate this representational condition by using a pre-captured workspace model that remains static during collaboration, and by combining subjective reports with behavioural measures derived from recorded interaction. We evaluate leadership emergence as an indicator of interface bias following prior work [67, 68, 52, 53], and we analyse complementary behavioural and experiential measures to characterise coordination under technological and environmental context asymmetry. These findings motivate the subsequent chapters by clarifying which coordination resources become fragile when representations are incomplete or stale, informing our shift toward task-driven augmentation for dynamic contexts (Chapter 4) and representational transformation for collaboration across multiple distinct places (Chapter 5).

In Chapter 4, we operationalise RQ2 and RQ3 in the context of site-specific outdoor AR experience development, where environmental change and partial access are common [15]. This chapter examines a practical mismatch between two authoring contexts. Ex-situ (i.e., remote) development commonly relies on pre-captured large-scale representations that support expressive editing in

familiar tools, yet omit dynamic characteristics and situational cues needed to verify placement, usability, and feasibility in the real world. In-situ (i.e., on-site) authoring, in contrast, provides direct access to these cues, while typically limiting expressive authoring capabilities and computational resources. We first conduct a formative study of professional practice, comprising a survey ($n = 25$) and follow-up interviews ($n = 5$), to characterise how teams cope with missing or outdated context and costly testing cycles. Building on these findings, we develop CoCreatAR, an asymmetric collaborative authoring system that augments remote work with anchored real-world context, including shared visuospatial context and annotations, to support grounding, reference, and verification during synchronous collaboration. We then evaluate CoCreatAR in a comparative user study ($n = 32$, in pairs) against an asynchronous baseline reflecting current practice. The results indicate that task-driven augmentation can improve awareness of on-site conditions and support authoring decisions under asymmetry, while increasing engagement and confidence in the authored outcome.

In Chapter 5, we operationalise RQ4 by moving beyond configurations that privilege a single participant's environment to examine representational transformation that incorporates contextual cues from multiple distinct places [12]. Prior approaches for collaboration across dissimilar sites often rely on simplified or anchor-based representations, which can limit the richness of contextual cues available for collaboration across places [69, 70, 65]. This chapter explores how computational techniques can instead synthesise context-rich shared environments without requiring collaborators to operate within a single host environment. We introduce SpaceBlender, a pipeline that takes images of multiple users' physical surroundings and generates a unified 3D environment for VR telepresence by combining geometric reconstruction with generative 3D scene completion. We evaluate the suitability of the generated environments through a comparative study ($n = 20$, in pairs) using a collaborative affinity diagramming task across three environment types. Participants' reports and observed strate-

gies characterise how visually rich computationally generated environments shape comfort, navigability, and the use of environmental cues for coordination. Furthermore, the results provide early insights into when synthesised contextual detail supports collaboration and when representational artefacts and inconsistencies impose costs, providing initial guidance for systems that transform environmental representations to reduce environmental context asymmetry across multiple places.

1.3 Dissertation Structure

The remainder of this dissertation is organised as follows. **Chapter 2** reviews prior work on mediated collaboration and CSCW foundations, with a focus on spatial organisation, context, and asymmetry in collaborative systems. **Chapters 3 to 5** present the three empirical investigations that constitute the core of this dissertation, advancing from characterising how shared space is established under heterogeneous interfaces and static representations, to designing and evaluating augmentations that support grounding in dynamic partially accessible contexts, and finally to exploring computational transformation that constructs a shared environment from distributed contextual inputs. **Chapter 6** synthesises findings across these investigations and reflects on their implications for our research questions and directions for future work. **Chapter 7** concludes by summarising the dissertation's contributions.

Chapter 2

Background

In this chapter, we review prior work that frames our investigation of spatial representations for asymmetric CMR. We first introduce foundational CSCW concepts for analysing mediated coordination: context, copresence, social presence, awareness, and common ground (Section 2.1). We then trace early audio-visual approaches to synchronous distributed work and their limitations in conveying interaction and workspace cues (Section 2.2). Next, we synthesise how asymmetry has been treated in CSCW as both a source of breakdown and a deliberate design condition (Section 2.3). We then review foundations of immersive experience, covering MR definitions and taxonomy alongside immersion, presence, and embodiment (Section 2.4). Following this, we examine conceptual models of shared space construction in the context of CMR (Section 2.5). Finally, we discuss types of asymmetries in CMR and prior studies and systems that have examined or leveraged such asymmetries (Section 2.6).

2.1 Foundational Concepts of Mediated Collaboration

Much of what makes collaboration effective in face-to-face settings operates implicitly, through stable access to shared context and continuously available social cues. Introducing computational mediation disrupts this implicit grounding by filtering, re-encoding, or delaying information that collaborators ordinarily rely on. As a result, coordination practices that are routine in co-located work often require explicit support or reconfiguration. Within this context, the field of CSCW examines how cooperative work is organised and how computational media can support it. CSCW has been described as “an endeavor to understand the nature

	Same *time*	**Different *time***
Same *place*	Face-to-face interaction (e.g., meeting rooms, wall displays)	Asynchronous interaction (e.g., shift work logs, project boards)
Different *place*	Synchronous distributed interaction (e.g., video conferencing, shared editors)	Asynchronous distributed interaction (e.g., email, version control)

Table 2.1 The CSCW Space–Time matrix, adapted from Johansen [72].

and characteristics of cooperative work with the objective of designing adequate computer-based technologies" [71].

Within CSCW, collaboration is considered a socio-technical phenomenon. It involves interdependent activities situated within organisational, material, and technological arrangements, and it is continually coordinated through communication. Consequently, CSCW research attends to both system design and the situated practices through which collaborators share information, negotiate responsibilities, and maintain mutual intelligibility.

A common way to distinguish collaborative situations considers where and when participants interact. Johansen's space-time framework categorises collaborative activity along two dimensions: whether collaborators share the same or different physical locations, and whether their actions occur at the same or different times [72]. These dimensions define four classes of collaboration that differ in the degree of spatial and temporal alignment among participants. Table 2.1 summarises these categories.

In this thesis, our focus is on synchronous distributed interaction, which places especially strong demands on mediating systems. In such settings, participants often expect the medium to support practices that are taken for granted in face-to-face collaboration, including access to shared visual context, deictic reference, and rapid repair of misunderstandings. However, early and contemporary CSCW systems alike have been marked by inefficiencies and limitations in practice, and users have frequently avoided or worked around them [73]. For example, Olson and Olson [73] reported that audio-only conferencing frequently failed to support tightly coupled work, as participants struggled to identify speak-

ers, track references, and repair misunderstandings in real time. This prompted teams to adopt more formalised interaction practices to maintain coordination. In this section, we draw on several foundational constructs from CSCW and communication theory: context, social presence and copresence, awareness, and common ground. These constructs have shaped decades of CSCW research and are closely related to each other.

2.1.1 Context

Context has long been understood as foundational to communication. Bateson [74] characterised meaning as emerging from the relation between an utterance and the ecological system of interaction in which it occurs, highlighting the role of "context markers" through which participants frame the nature of their activity. These markers are continually negotiated, such that contexts are maintained as interaction unfolds.

Within ubiquitous computing, context has been conceptualised as an informational resource that computational systems can capture and exploit. Abowd et al. [75] defined context as "any information that can be used to characterise the situation of an entity" and used this definition to articulate how systems may adapt their behaviour in response to such information. Their formulation provides a pragmatic foundation for engineering context-aware applications, motivating advances in sensing, modelling, representation, and infrastructure.

Subsequent work has questioned whether such representational conceptions adequately capture how context functions in everyday interaction. Dourish [48] argued that when context is treated as predefined or externally specifiable information, technical systems risk misinterpreting the circumstances of human action. Drawing on ethnomethodological analysis, Dourish [48] proposed that contextuality is an interactional property: the relevance and boundaries of contextual features emerge through participants' ongoing efforts to make each other's actions accountable and intelligible. From this perspective, context is an interactional accomplishment that is continuously produced and re-specified through ongoing activity.

These two perspectives highlight different facets of the problem. Representational approaches provide mechanisms through which systems can make aspects of a situation available for use and enable adaptive behaviour. Interactional perspectives explain how people draw on available resources, including those produced by technical systems, as part of sustaining coordinated activity. In this thesis, we use "context" to refer to system-level representations of situational information and to the interactional work through which collaborators render that information relevant to their joint activity.

2.1.2 Copresence, Social Presence, and Related Concepts

The sense of "being together" with others in mediated environments has been examined across related research disciplines, which can lead to terminological confusion. Terms such as *copresence* and *social presence* have been frequently used interchangeably, despite having emerged from different theoretical concerns and being operationalised in distinct ways [76, 77]. This overlap reflects the intersecting influences of sociology, communication studies, CSCW, and immersive technology, which have shaped one another while retaining different assumptions about what it means to "be together" [78, 79].

Within sociology, copresence was originally articulated as a structural condition of interaction. Goffman used the term to describe situations in which individuals are physically co-located and mutually perceptible, emphasising that "copresence renders persons uniquely accessible, available, and subject to one another" [80].

As the concept was taken up in research on collaborative virtual environments (CVEs), it was applied to capturing the subjective experience of being with others despite physical separation [78]. This shift introduced ambiguity between copresence as an objective configuration supported by technology and copresence as an experiential quality of interaction. Zhao [79] addressed this by distinguishing between the *mode* of copresence, referring to the spatial and technological context, and *sense* of copresence, referring to the experiential outcome that may arise within that context.

A series of studies conducted as part of the *COVEN* project on CVEs were among the first to explicitly operationalise and measure this sense of copresence in the form of a questionnaire [67, 52, 68]. Items targeted participants' sense of being together, the extent to which interaction was experienced as engagement with other people rather than with a mediating computer system, and the degree to which the virtual collaboration resembled working together in a real-world setting. Notably, these studies identified a positive association between the construct of copresence and presence (the sense of being in the virtual environment itself) [68, 52], which we discuss further in Section 2.4 and explore in Chapter 3.

Early work on mediated communication, particularly in media spaces, typically framed togetherness through the lens of *social presence*. Building on foundational communication theories, social presence was framed as a property of media that shaped how salient and immediate others appeared, and the extent to which they were experienced as psychologically available during interaction [81, 76]. Within this framing, research focused on how communication technologies modulated interpersonal engagement, for example by supporting awareness, intimacy, or involvement, without requiring a shared spatial frame of reference. As a result, many views of social presence treated spatial co-location as secondary, directing attention instead to the relational and communicative conditions under which mediated interaction felt socially connected.

As research progressed, social presence frameworks increasingly incorporated copresence as one element within broader relational contexts. A prominent example is the *Networked Minds* framework introduced by Harms and Biocca, which remains one of the most widely used operationalisations of social presence. While the framework includes copresence as a subdimension, its scope extends beyond the minimal belief that another person is simply *present*. Instead, it models social presence through interrelated aspects of attention, perceived understanding, and interpersonal interdependence. These aspects describe how collaborators direct attention toward one another, interpret communicative intent and affect, and experience their actions and feelings as mutually

consequential [82]. Through this formulation, social presence was treated as a dynamic relational process rather than a single experiential judgement [76], and it implicitly incorporated elements commonly associated with awareness, which we discuss next.

2.1.3 Awareness

Awareness is a foundational concept in CSCW and refers to how collaborators maintain an understanding of others' activities in order to coordinate their own actions. Dourish and Bellotti characterised awareness as "an understanding of the activities of others, which provides a context for your own activity" [83].

As noted by Carroll et al. [84], the variety of awareness concepts in CSCW reflects the field's multidisciplinary origins, each with distinct informational contexts and demands. For example, *mutual awareness* concerns participants' reciprocal orientation to one another's presence, actions, and perceptual access [26]; *passive awareness* emphasises information that becomes available as a by-product of interaction with shared artefacts rather than through explicit communication [83]; *activity awareness* extends awareness beyond individual actions to include goals, plans, and task dependencies over time [85]; *situational awareness* focuses on perception, comprehension, and projection in dynamic environments [86]; and *workspace awareness* addresses how collaborators maintain an understanding of others' actions, locations, and interactions within a shared workspace [29]. Together, these perspectives illustrate how awareness has been conceptualised to represent different aspects of collaborative work.

In this thesis, we are specifically concerned with collaborative systems in which interaction is organised around shared spatial representations. Within this context, *workspace awareness* is particularly relevant [29]. This focus provides a conceptual bridge to later sections of this chapter, which examine shared spaces as key mechanisms through which collaboration is structured and negotiated.

Gutwin and Greenberg [29] described workspace awareness through four defining characteristics:

> 1. *Awareness is knowledge about the state of an environment bounded in time and space.*
> 2. *Environments change over time, so awareness is knowledge that must be maintained and kept up to date.*
> 3. *People interact with and explore the environment, and the maintenance of awareness is accomplished through this interaction.*
> 4. *Awareness is a secondary goal in the task, that is, the overall goal is not simply to maintain awareness but to complete some task in the environment.*

This perspective aligns closely with Dourish's [48] interactional conceptualisation of context, in which participants actively produce and sustain the conditions that make one another's actions intelligible. Accordingly, we use "awareness" as an umbrella term for knowledge of others' activities, with a focus on workspace awareness when analysing interaction in shared visual environments throughout this thesis.

2.1.4 Common Ground

Common ground originates in communication theory and refers to the knowledge that participants mutually believe they share [25]. It constitutes the body of joint knowledge on which collaborators rely to communicate and coordinate effectively [85]. Within CSCW, common ground is closely related to awareness, and empirical work often treated the two as intertwined. Carroll et al. [84] described common ground as one facet of broader activity awareness, and subsequent CSCW work highlighted the difficulty of sharply separating common ground from forms of *mutual awareness* on a conceptual level and in practice [87].

Relevant across these concepts is the interactional process of *grounding*. Clark and Brennan [25] characterised communication as a continual process of grounding, in which contributions are evaluated against a context-dependent

grounding criterion. Evidence of adequate understanding may be explicit, but more often appears implicitly in the next relevant action. Grounding proceeds incrementally and remains contingent on ongoing evidence of understanding, with repair initiated when such evidence is absent.

This process is often illustrated through conversational examples. Consider the following exchange adapted from Clark and Brennan [25]:

> ***A**: Where can I get a hold of you?*
> ***B**: You can reach me through phone or email.*
> ***A**: I mean, what is your office number?*
> ***B**: Oh, it is 401B.*

Here, **B**'s initial response provides negative evidence of understanding, as it does not address the specific information sought by **A**. The clarification initiates repair, and **B**'s subsequent response supplies information that is conditionally relevant to the revised contribution.

In this thesis, we treat *grounding* as the core interactional process through which participants achieve and sustain mutual orientation. *Common ground* and *mutual awareness* are used descriptively to refer to states of understanding that participants demonstrably treat as shared for the purposes of ongoing coordination.

Summary and Implications

This section established the analytic vocabulary that structures the dissertation's approach to asymmetric collaboration. Context, treated both as system-representable information [75] and as an interactional accomplishment [48], informs how we analyse environmental representations as resources that collaborators actively draw upon and maintain. Copresence, social presence, and awareness provide constructs for evaluating how representational choices shape collaborators' sense of togetherness and their capacity to interpret one another's actions [29, 82], which we operationalise across the empirical studies in Chapters 3 to 5. Grounding and common ground provide the interactional frame for

understanding how collaborators achieve and repair mutual orientation under partial or uncertain shared access [25], a dynamic that is relevant to the asymmetric configurations examined throughout this thesis. These constructs recur in the subsequent sections and remainder of the dissertation.

2.2 Early Audio-Visual Approaches to Mediated Collaboration

In 1968, Engelbart and English [88] presented a seminal demonstration of the *oN-Line* system, a software and hardware platform that integrated many of the foundational components of modern personal computing and collaboration technologies, including real-time editing, shared workspaces, and video conferencing. This demonstration, later referred to as "The Mother of All Demos", foreshadowed the subsequent emergence of research on computer-mediated collaboration systems.

From the late 1980s through the 1990s, synchronous computer-mediated collaboration systems gained momentum, particularly those that attempted to support everyday work across distance using continuous or on-demand audio and video. These systems, known as *media spaces* [89, 90], sought to approximate fundamental characteristics of co-located interaction through networked audio-visual links [27]. Early studies investigated whether such systems could effectively support distributed collaboration, informal communication, and shared problem solving.

2.2.1 Capturing Users and Their Communication Cues

A series of early works found that audio-visual media improved upon voice- or text-only channels by supporting more natural conversational behaviour [91, 92]. Studies reported that video enhanced collaborators' ability to complement speech with non-verbal information, facilitating turn-taking, disambiguation, and coordination [93, 94]. For example, Isaacs and Tang [93] observed that video allowed participants to use facial expressions and gestures to regulate conversational flow and express stance, contributing to smoother interaction.

At the same time, the limitations of video-mediated communication were substantial. Restricted views, limited resolution, and lagging frame rates often hindered participants' ability to perceive critical cues such as subtle gestures or mutual attentional focus [93, 94, 95, 92]. Moreover, several studies noted difficulties in supporting serendipitous interaction in media spaces, such as lightweight hallway-style encounters or peripheral social contact [96, 93]. These observations contributed to a growing recognition in the mid-to-late 1990s that simple point-to-point video links were insufficient for reproducing the ecological qualities of co-present interaction. In response, some systems began to explore more spatially organised designs. For example, the desktop-based system *CU-SeeMe VR* [97] embedded live audio and video into a simple CVE that offered proximity-based audio and navigable layouts.

Although these systems experimented with spatial structure, subsequent analyses suggested that many early designs still overlooked the social meaning embedded in physical spatial organisation. Buxton [98] later argued that effective mediated interaction requires respecting distinctions among person space, task space, and reference space, and showed how fixed camera placements and monolithic display configurations frequently violated the spatial conventions through which collaborators signal role, approach, and engagement. Together, these limitations and exploratory efforts set the stage for more fully spatialised and immersive CSCW systems, which we examine in detail in Section 2.6.

2.2.2 Capturing and Representing Action

A parallel thread of research investigated collaboration on physical or spatially organised tasks, where effective performance is highly dependent on sharing a view of the task environment. Systems in this category focused on transmitting the workspace itself instead of users' faces [99, 100, 101, 36].

Clark and Krych [102] demonstrated that remote helpers were more effective when they could see a worker's task environment during a construction activity. Workspace visibility increased the use of gestures, midcourse utterances, and other non-verbal resources tied closely to action in the environment.

Similarly, Fussell et al. [103] found that shared visual workspace video supported faster task completion than audio-only communication in a robot-assembly task. While they observed no significant benefit of a head-mounted camera over a fixed scene-oriented view, their analysis pointed to limitations of camera placement, resolution, and alignment with participants' actual visual focus. Importantly, Fussell et al. [104] showed that even when video improved grounding behaviour and reduced the need for explicit verbal descriptions, audio-video-mediated collaboration could still be less efficient than side-by-side work. Their study identified reduced field-of-view awareness and challenges in interpreting partial workspace cues as key contributors to inefficiency. These issues, including the problem of view coupling and the need for view independence, have continued to shape subsequent research on CVEs [105] and are discussed further in Section 2.6.

The constraints of single-view systems prompted researchers to examine whether access to multiple viewpoints could better support collaborative action. Gaver et al. [106] investigated this through a multi-camera media space applied to collaboration around physical artefacts and found that collaborators strongly preferred task-focused views, such as a "desk" or "dollhouse" view, over facial ones. At the same time, they struggled with establishing shared frames of reference, aligning perspectives, and coordinating orientation across discontinuous views.

Beyond single-view and multi-view systems, several efforts explored more complex configurations to overcome these limitations. For example, Fuchs et al. [107] introduced a "sea of cameras" approach to derive dense 3D scene information for *telepresence* [108], while Raskar et al. [109] proposed image-based modelling and spatially immersive displays as part of a broader vision for the "office of the future". These approaches extended early workspace-sharing concepts toward fully three-dimensional representations, a trajectory we discuss in Section 2.6.

Summary and Implications

This section showed how early media spaces and workspace-sharing systems exposed coordination difficulties when collaborators had partial, delayed, or misaligned access to interactional and workspace cues [93, 104, 28]. These difficulties motivated the pursuit of symmetry as a design ideal and highlighted the importance of what collaborators can perceive about one another's actions and environments. The limitations of single-view and multi-view systems in supporting shared reference and natural coordination foreshadow challenges that persist in contemporary CMR, where asymmetries of device and context can produce comparable breakdowns. The trajectory from point-to-point video toward spatially organised and three-dimensional representations provides historical context for the immersive systems reviewed in Section 2.6 and motivates the examination of asymmetry as a design condition in Section 2.3.

2.3 Asymmetry as a Design Condition in Collaborative Systems

In this thesis, we use *asymmetry* to describe systematic differences in collaborators' access to perspectives, capabilities, information, or benefits that arise from technical or situational conditions. While asymmetries also occur in face-to-face interaction and are routinely managed through communicative negotiation and signalling [40], physical copresence affords comparatively aligned access to shared artefacts and ongoing activity [110, 111].

Early CSCW research frequently treated these properties of copresence as a point of reference, motivating efforts to reproduce comparable conditions in mediated settings. One influential example of this orientation is the notion of *What You See Is What I See* (WYSIWIS), introduced in the context of shared workspaces to ensure that collaborators had identical views and equivalent opportunities to act on shared artefacts [112]. However, even in this early work, strict symmetry was quickly recognised as impractical for real collaborative tasks, leading to the proposal of *relaxed WYSIWIS*, which explicitly permits controlled

divergences in view, timing, and participation to support parallel work, focus, and privacy [113].

The continued relevance of symmetry as an aspiration, and its persistent limitations, is evident in more recent CSCW research on hybrid work. Bjørn et al. [114] argued that achieving symmetry in synchronous hybrid interaction is fundamentally impossible, as participants necessarily have only partial access to complex artefact ecologies spanning multiple sites, devices, and configurations. They proposed that asymmetry should be acknowledged as a defining condition of contemporary collaborative work. To review how asymmetry has been treated as a design condition, this section examines how asymmetry has been conceptualised in CSCW research, first as a source of breakdown, and subsequently as a condition to be managed, structured, and, in some cases, deliberately exploited through design.

2.3.1 Taxonomies of Asymmetry in Collaborative Work and Games

As CSCW researchers increasingly recognised the persistence of asymmetry, efforts emerged to systematically characterise its forms and sources [41, 42, 115, 116, 117]. In the context of media spaces, Voida et al. [41] identified six types of asymmetry: *media* (e.g., audio vs. video), *fidelity* (quality of transmission), *participation* (intentional vs. incidental), *engagement* (effort invested), *benefit* (who gains value), and *place* (physical context). Notably, these categories cut across both social and technical aspects of mediated interaction.

More recently, Bréhault et al. [42] performed a systematic review of asymmetry in collaborative systems. Their resulting framework characterised asymmetry along interaction-centric dimensions such as *input modality* (e.g., mouse vs. gesture), *output modality* (e.g., VR headset vs. tablet), *dimensionality* (e.g., 2D screen vs. 3D immersive display), *mobility* (e.g., seated desktop user vs. moving AR user), *display space* (e.g., small mobile display vs. large wall display), *content access* (e.g., viewing only vs. having control to act on content), *content layout* (e.g., horizontal tabletop vs. vertical monitor arrangement), and *shared feedback*

(e.g., representing users via avatars vs. 2D video). Based on these dimensions, the authors classified recent systems and discussed how each type of asymmetry can shape collaboration.

Games research offers a complementary perspective by focusing on types of asymmetry and on how asymmetry structures coordination dynamics [117, 118]. Harris et al. [117] examined forms of asymmetry in multiplayer games, where players have fundamentally different interfaces, abilities, or challenges, and argued that such differences can be deliberately leveraged to cultivate interdependence. Beyond categorising asymmetries of ability, information, interface, and responsibility, this work articulated patterns of interdependence, such as unidirectional or bidirectional dependence and temporal coupling. These concepts provide a basis for describing how asymmetry shapes the flow of action and coordination over time, an aspect that is less explicitly addressed in CSCW taxonomies on asymmetry.

2.3.2 Asymmetry as a Source of Awareness Breakdown

Early CSCW research frequently approached asymmetry through its negative consequences for awareness, grounding, and coordination. Studies of video-mediated communication showed that mediated settings disrupt the fine-grained coordination of gaze, gesture, and timing that underpins face-to-face interaction. Heath and Luff [28] demonstrated how participants in video links struggled to interpret embodied conduct when actions become detached from their local context of production and reception. Similarly, Gaver [119] argued that media spaces are inherently *anisotropic*, meaning that their affordances vary by direction and configuration. As a result, visibility and audibility are often non-reciprocal, undermining assumptions of mutual access that are routine in copresent interaction [119]. Such asymmetries complicate conversational grounding and increase the effort required to establish shared understanding.

These issues extend beyond interpersonal cues to include cues oriented toward the shared environment. Research on workspace awareness showed that when collaborators have different views, representations, or interaction

capabilities, their ability to interpret one another's actions in a shared workspace is diminished. Gutwin and Greenberg [29] showed that deictic references such as "look at this" become ambiguous under divergent views, requiring additional communicative work to reconstruct what others can see or do. Similar dynamics arise in heterogeneous (i.e., asymmetric) device collaboration, where differences between desktop and handheld display (HHD) interfaces produce asymmetric access to shared state and complicate coordination [120]. In such settings, asymmetry shapes what collaborators can perceive and how initiative and participation are negotiated.

The concept of *fractured ecologies* provides a unifying concept for understanding these breakdowns across both interpersonal and workspace-oriented cues [49]. Luff et al. [49] introduced the term to describe collaborative settings in which participants inhabit perceptually and materially misaligned environments, such that actions that appear visible or interpretable to one party are opaque to another. Later work extended this analysis to more flexible and asymmetric configurations, showing how incongruent locations, resources, and representations further complicate the establishment of a common focus. From this perspective, asymmetry emerges from broader misalignments between how collaboration is produced and perceived across sites, rather than from isolated interface differences.

Symmetry as a design response. Awareness breakdowns identified in early CSCW research motivated design approaches that treated the restoration of symmetry as a primary objective, particularly in hybrid collaboration that spans co-located and remote participants.

Paay et al. [121] exemplified this orientation through the concept of *blended interaction spaces*, which aimed to align spatial experience across distributed sites through tightly coordinated physical and technical design. Their *BISi* prototype relied on matched room geometry, furniture, and camera configurations to support comparable access to gaze, gesture, and deictic reference.

Inspired by the concept of WYSIWIS, Grønbæk et al. [122] introduced *MirrorBlender*, which enabled collaboration through a shared canvas that mixes participant video feeds with meeting artefacts, allowing co-located and remote participants to see the same spatial configuration and to perform deictic gestures within it. While this approach mitigated certain breakdowns, it also illustrated the degree of interface complexity and active coordination required to sustain symmetry in practice.

Other hybrid systems pursued similar goals of experiential parity through spatial means. Tang et al. [123] provided remote participants with viewpoints that more closely approximated those of co-located collaborators, supporting more natural orientation and attention to shared activity. Panda et al. [124] likewise sought to reduce asymmetries by physically representing remote participants around a shared table using large screens and enabling them to look around and be perceived as spatially present.

These works reaffirmed symmetry as a persistent design aspiration for addressing awareness breakdowns. They also highlighted its limits. Restoring comparable access to cues and context in mediated collaboration requires increasingly elaborate representational strategies, and even then, symmetry remains partial and contingent [114].

2.3.3 Asymmetry as a Deliberate Design Strategy

Building on these conceptual developments, a range of CSCW systems treated asymmetry as a deliberate design strategy.

Early examples demonstrated how controlled information asymmetry can structure collaboration. The *Lumisight* table displayed different information on different sides of a shared tabletop, shaping how participants contributed knowledge and negotiated decisions [125]. Similarly, the *Flashlight Jigsaw* system assigned players asymmetric abilities to reveal parts of a shared puzzle, requiring close coordination to solve the task [126]. In both cases, asymmetry was used to create interdependence and to structure collaborative problem solving.

More complex applications of deliberate asymmetry have appeared in task-oriented collaborative systems. For example, Convertino et al. [127] introduced a role-based emergency management planning system in which participants from different agencies worked with both private role-specific views and a shared public representation. By selectively distributing information and explicitly supporting role-linked annotations, the system reduced coordination costs while maintaining common ground and activity awareness. This work illustrates how asymmetry can be used to scaffold division of labour, responsibility, and expertise in complex cooperative tasks.

Across these examples, interface design and interface type, including differences in input and output modalities, have emerged as primary mechanisms through which intentional role or information asymmetries are introduced.

Summary and Implications

This section established asymmetry as a persistent condition of mediated collaboration that can disrupt grounding and awareness when it misaligns collaborators' access to interactional and workspace cues [28, 49], while also serving as a deliberate design resource for structuring interdependence and role differentiation [125, 127]. The taxonomies reviewed here provide dimensions for characterising asymmetric configurations, which we extend in Section 2.6 to the specific context of CMR. For this dissertation, the dual framing of asymmetry as both a source of breakdown and a design lever motivates the empirical approach taken in Chapter 3, which examines how technological asymmetry shapes collaborative behaviour, and the constructive approaches in Chapters 4 and 5, which develop representational strategies that manage asymmetry rather than eliminate it. Before examining these CMR-specific concerns, Section 2.4 reviews foundational concepts of immersive experience that underpin the interfaces explored in this thesis.

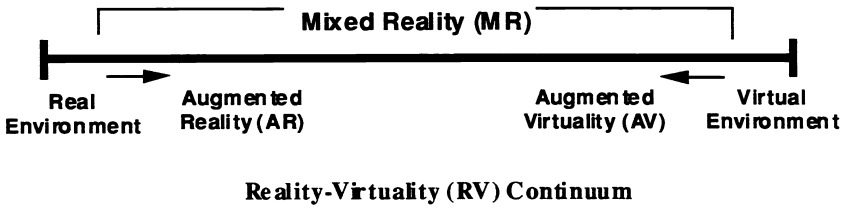


Figure 2.1 The Reality—Virtuality (RV) continuum by Milgram et al. [129]. MR encompasses both AR and augmented virtuality (AV), situated between the real and virtual extremes. Reproduced with permission.

2.4 Foundational Concepts of Immersive Experience

Since the earliest visions of immersive computing, researchers have examined how advances in display, tracking, and interaction shape the character of users' lived experience within mediated worlds. In his formulation of the "ultimate display," Sutherland [128] described a system that would "serve as a looking-glass into the mathematical wonderland constructed in computer memory" and that would "serve as many senses as possible". This vision highlighted both the technical ambition of immersive systems and the experiential consequences of blending computational and physical realities. In this section, we review definitions of immersive technology, as well as foundational concepts such as immersion, presence, and embodiment.

2.4.1 Definitions and Taxonomies of Mixed Realities

Early definitions of MR emerged from efforts to situate emerging display technologies relative to both the physical world and fully synthetic virtual environments. Milgram et al. [129] formalised this relationship through the Reality—Virtuality (RV) continuum, illustrated in Figure 2.1, which positions MR as the broad class of environments lying between purely real and purely virtual extremes. Within this continuum, AR refers to systems in which virtual content is added to a primarily real environment, whereas AV denotes environments that are largely virtual but incorporate or represent aspects of the real world.

The RV model was supplemented with a multidimensional taxonomy cap-

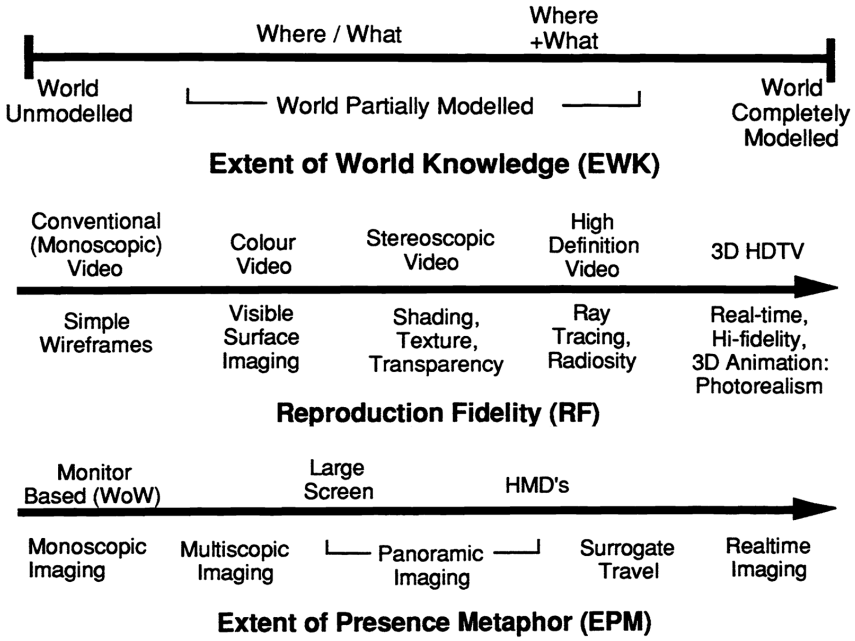


Figure 2.2 Subdimensions of the Reality—Virtuality (RV) continuum by Milgram et al. [129]. From top to bottom: *extent of world knowledge, reproduction fidelity,* and *extent of presence metaphor.* Reproduced with permission.

turing key properties of MR systems [129]. This taxonomy introduced three orthogonal dimensions: *extent of world knowledge, reproduction fidelity,* and *extent of presence metaphor.* Together, these dimensions describe how much a system knows about the physical environment, how faithfully it renders real and virtual content, and whether interaction is framed from an egocentric or exocentric viewpoint. As shown in Figure 2.2, this framework enabled finer distinctions between systems that might otherwise appear similar when placed solely along the RV axis.

Subsequent work revisited and extended this framework in light of advances in sensing, rendering, and interaction. Skarbez et al. [130] argued that the original RV continuum is effectively discontinuous, since even highly immersive virtual environments necessarily involve a mixture of real and virtual stimuli at the level of human perception. From this perspective, MR is better understood as encompassing most technology-mediated experiences, including conventional VR, because users always perceive virtual content in conjunction with aspects of the physical world and their own bodily state. This reframing shifts emphasis

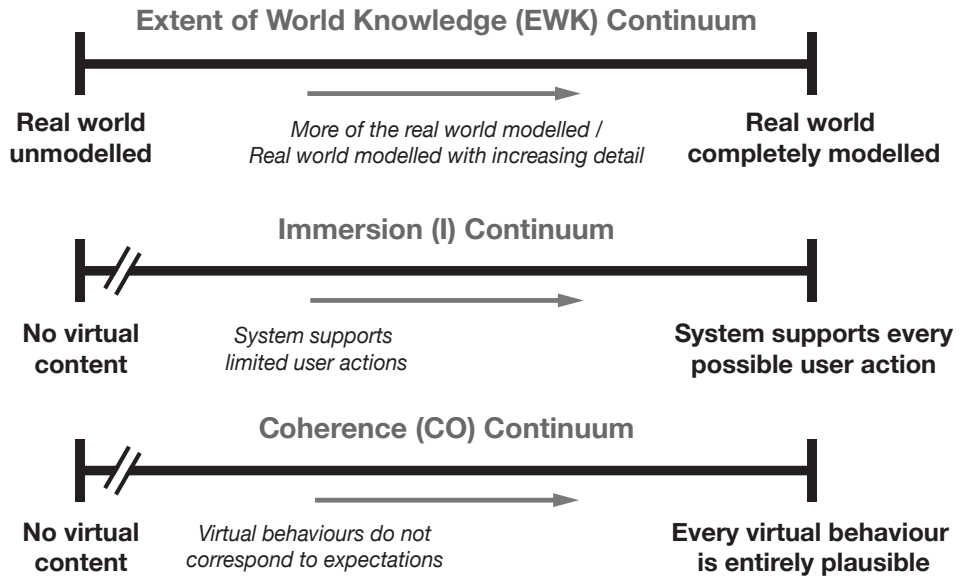


Figure 2.3 Revised dimensions of the Reality—Virtuality (RV) continuum, adapted from Skarbez et al. [130]. Relative to Milgram and Kishino's original dimensions, Skarbez et al. retained *extent of world knowledge*, combined *reproduction fidelity* and *extent of presence metaphor* into *immersion*, and introduced *coherence* as a third dimension oriented toward the consistency of the overall perceptual experience.

from display technology alone toward the coherence of the resulting perceptual experience.

Building on this critique, Skarbez et al. [130] also revisited the associated taxonomy and proposed a revised set of dimensions intended to characterise mixed reality experiences rather than display hardware alone. As illustrated in Figure 2.3, the authors retained *extent of world knowledge* as a key axis, since it captures the degree to which a system models and responds to its real-world surroundings, and they argued that Milgram and Kishino's *reproduction fidelity* and *extent of presence metaphor* are closely related in practice and can be combined into a single dimension of *immersion*, grounded in the view that immersion reflects the set of valid actions supported by a system (see Section 2.4.2). To address what they identify as a missing concern for the integrity of the resulting experience, they introduced *coherence* as a third dimension describing how consistently the system's intended cues and behaviours form a plausible percept, including internal coherence within virtual content and external coherence between virtual

and real-world dynamics (see Section 2.4.3).

Despite the well-established nature of the RV continuum, recent analyses have highlighted persistent ambiguity in how terms such as "mixed reality (MR)" and "eXtended reality (XR)" are used across research and industry. Speicher et al. [131] showed that MR has been variously defined as a superset of AR and VR, a synonym for spatial AR, or a distinct category requiring environmental understanding and interaction. Rauschnabel et al. [132] similarly found that eXtended reality (XR) is often used interchangeably with MR or as an umbrella term for all immersive technologies, including AR, VR, and AV.

In this thesis, we adopt the broad definition as proposed by Skarbez et al. [130], using MR as an umbrella term for immersive experiences that blend real and virtual elements, including both AR and VR.

2.4.2 Immersion

While *immersion* was initially defined as a psychological state of "perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences" [133], subsequent work in CVE research has emphasised a more technical definition.

In current literature, immersion is widely accepted as an *objective* description of the technology rather than a psychological state [134, 135, 136]. Slater and Wilbur [134] characterised *immersion* as the extent to which a display system can deliver an illusion of reality that is *inclusive*, *extensive*, *surrounding*, and *vivid* to the senses of a human participant. Inclusiveness refers to the degree to which physical reality is shut out, extensiveness to the range of sensory modalities supported, surrounding to the extent of panoramic coverage, and vividness to the resolution, fidelity, and information content of the simulated sensory energy. Under this definition, immersion depends on both the fidelity and the breadth of sensory modalities supported by the system, including the quality of body tracking and the provision of a virtual body that enables egocentric perception [134].

Crucially, this system-centric definition allows researchers to quantify im-

mersion objectively. For instance, a system that provides stereoscopic visual output coupled with low-latency head and body tracking, as is typical of contemporary HMDs, is objectively more immersive than a desktop monitor because it better realises sensorimotor matching, since motor actions such as turning one's head or walking result in appropriate changes to the visual and auditory displays [134].

2.4.3 Presence

Minsky [108] introduced the term *telepresence* to describe the sense of "being there" at a remote location through a mediated robotic interface. This concept was later adapted to virtual environments, where *presence* (i.e., virtual presence) came to denote the subjective experience of being located within a virtual space [137, 138, 139, 133, 135]. This sense of "being there" is also frequently referred to in the literature as *spatial presence*, emphasising the spatial component of the experience.

2.4.3.1 Dimensions and Extensions of Presence

After its initial formulation, presence and its underlying mechanisms became a major focus in the MR literature [135, 140, 141, 133, 142, 143].

Witmer and Singer [133] argued that presence depends on both *involvement*, understood as a psychological state of focused attention on a coherent set of stimuli, and a subjective form of *immersion*, which they interpreted more broadly as the perception of being enveloped by the environment.

In a now widely adopted model, Slater [141] revisited presence as a primary experiential quality of VR and proposed a two-component framework consisting of Place Illusion (PI) and Plausibility Illusion (Psi). They defined PI as the qualia of having the sensation of being in a real place, and Psi as the illusion that the scenario being depicted is actually occurring, despite the user's explicit knowledge that the environment is simulated. In recent work, Slater et al. [144] reiterated that PI and Psi jointly characterise presence in VR and positioned them, together with the illusion of ownership over the virtual body (see Section 2.4.4) and the

illusion of *copresence*, or virtual togetherness with others (see Section 2.1.2), as the key illusions of VR.

PI is constrained by the sensorimotor contingencies afforded by the immersive system. If the system responds to the user's motor actions, such as head movement, in a way that is consistent with natural perception, PI can be sustained [141, 144]. Psi, in contrast, is shaped by how well the scenario matches the user's expectations about how situations and agents should behave, particularly in terms of whether events are contingent on the user's actions and whether social and causal dynamics are credible. Hofer et al. [142] built upon the concept of Psi by arguing that plausibility arises from the consistency of the environment with the user's prior knowledge and expectations. They distinguished between *internal plausibility*, defined as consistency within the fictional world, and *external plausibility*, defined as consistency with real-world knowledge. Violations of these consistencies can break the illusion of plausibility. However, Hofer et al. [142] reported that users may sometimes accommodate anomalies without fully disrupting their sense of presence, as found in a study that introduced high levels of implausible events (e.g., furniture on the ceiling, snow in indoor settings) [142].

In contrast to VR, presence in MR received comparatively less systematic attention until recently [145]. Wienrich et al. [143] argued that MR complicates presence because users simultaneously perceive and act within tightly coupled physical and virtual spaces, which can provide competing reference frames for self-location. Building on the PI/Psi framework, Wienrich et al. [143] proposed that MR experiences must support presence across the entire RV continuum rather than in a purely virtual space, and showed empirically that inconsistencies between physical and virtual layers can reduce presence even when each layer is individually coherent [143]. Relatedly, Westermeier et al. [146] demonstrated that in MR settings, perceptual incongruities such as mismatched lighting or occlusions can affect both plausibility and presence. This reinforces the view that PI and Psi remain useful constructs in MR, but that their realisation is constrained by additional cross-reality consistency requirements.

2.4.3.2 Formation and Maintenance of Presence

As discussed above, presence and its underlying constructs are dynamic and evolve as the user interacts with the CVE. Several models have been proposed to explain how presence is formed and maintained over time.

Slater and Steed [135] described presence as a stochastic selector mechanism grounded in Gestalt psychology. They argued that, at any given moment, a user processes two competing streams of sensory data: signals originating from the physical environment (R) and signals originating from the virtual environment (V). Presence is determined by which of these streams is selected as the dominant Gestalt figure. At any specific instant, presence is therefore binary, in that the user is effectively present in either the virtual or the real environment. Across time, however, it can be described as a probability distribution reflecting fluctuations in attentional dominance.

Within this framework, the maintenance of presence is closely tied to avoiding disruptions that shift attentional selection back to the physical world [135]. Slater and Steed [135] characterised such disruptions as "Breaks in Presence" (BIPs), which occur when sensory inconsistencies such as tracking latency or external interruptions draw attention away from the virtual environment.

A different model framed presence as a constructive cognitive process. Wirth et al. [147] proposed a two-level model in which presence emerges through successive stages of mental processing [147]. The first level involves the construction of a *Spatial Situation Model*, which represents the spatial layout of the virtual environment based on available sensory cues. The second level involves the evaluation of a perceptual hypothesis, referred to as the "Medium-as-PERF" (Primary Egocentric Reference Frame) hypothesis. According to this model, presence arises when the user accepts the virtual environment as the primary reference frame for *self-location* and *action*, thereby discounting contradictory sensory information from the physical world.

2.4.4 Embodiment

Embodiment concerns how users experience acting through a body within a virtual environment. In virtual environments (VEs), the user's physical body is often represented by a virtual body or avatar. Interaction with such representations can give rise to a *sense of embodiment*, defined as the experience that the properties of a virtual body are processed as if they were properties of one's own biological body [148].

The sense of embodiment is commonly described as comprising three interrelated components [148]. The *sense of self-location* refers to the experienced location of the self within space. The *sense of agency* denotes the experience of initiating and controlling actions, arising when the movements of the virtual body correspond coherently with the user's intentions and resulting sensory feedback. Finally, the *sense of body ownership* refers to the experience that the virtual body is one's own body, in the sense that it is perceived as the source of sensations and actions.

Providing a virtual body serves as an important anchor for immersive experience. Slater et al. [149] argued that a virtual body can function as a unified reference for visual and proprioceptive information, such that the self-representation in the virtual environment supplies a coherent frame through which sensorimotor contingencies are organised. Slater et al. [150] complemented this with empirical evidence, showing that the extent and richness of whole-body movement, including head rotations and repeated bending, are positively associated with reported presence.

Summary and Implications

This section reviewed foundational concepts that inform how immersive interfaces shape user experience in CMR. Immersion, defined as the sensorimotor contingencies a system supports [134], provides an objective basis for comparing the display and tracking configurations examined in Chapter 3. Presence, particularly the constructs of PI and Psi [141], explains how coherence and plausibility shape whether users treat a mediated space as the primary frame for percep-

tion and action—a consideration that becomes critical when environmental representations diverge from physical reality, as examined in Chapters 4 and 5. Embodiment captures how agency and self-location are established through virtual bodies [148], which affects how collaborators interpret one another's actions and establish reference within shared space. Together with the constructs reviewed in Section 2.1, these concepts provide a foundation for evaluating CMR as a platform for asymmetric collaboration.

2.5 Conceptual Models of Collaborative Mixed Realities

Building on the foundations of MR as discussed earlier, this section reviews conceptual models that describe how CMR systems construct *shared space* across asymmetric physical and virtual contexts. We focus on three complementary strands: (i) taxonomic dimensions for positioning collaborative systems in a broader design space of shared space, (ii) mechanisms for constructing shared space through boundary-based and volumetric coupling, and (iii) conceptualisations of asymmetry as a structural condition of CMR, with particular emphasis on asymmetric collaboration.

2.5.1 Taxonomies of Shared Space in Collaborative Systems

Benford et al. [46] introduced one of the earliest spatial taxonomies for CSCW, aiming to compare and relate emerging spatial approaches such as media spaces, spatial video conferencing, CVEs, and telepresence systems (Figure 2.4). This taxonomy characterised shared-space systems along three orthogonal dimensions. First, *transportation* concerns where collaboration is situated, describing the extent to which users experience entering a remote environment versus having remote information introduced into their local physical setting. Second, *artificiality* concerns what is shared, capturing the extent to which the shared space is composed of synthetic representations versus content derived from the physical world. Third, *spatiality* concerns how shared space is structured, describing the degree to which a system supports fundamental spatial properties such as

containment, topology, navigation, orientation, and movement, ranging from minimal notions of *place* to fully navigable shared spatial frames of reference. This dimension provides a conceptual bridge between early CSCW systems that supported only weak spatial cues (see Section 2.2) and later spatial approaches that offer stronger cues (see Section 2.6). Together, these dimensions disentangle complementary aspects of shared space, allowing systems to be compared in terms of what content is shared, where participants are situated with respect to one another, and how strongly a common spatial structure is established.

In a review of CMR, Ens et al. [43] used these dimensions as a conceptual foundation for situating systems within physical and virtual space. The authors extended their classification framework by introducing dimensions that more fully represent the human experience of collaborative activity. These additional dimensions include the focus of coordination (categorised as an environment, workspace, object, or person), the symmetry of interaction roles and technical capabilities, and the temporal nature of the interaction (asynchronous vs. synchronous). Finally, they introduced a scenario dimension to categorise systems into five distinct use cases, such as *remote expert* or *shared workspace*.

2.5.2 Boundary-Based Spatial Coupling

While Benford et al.'s [46] taxonomy provides a descriptive framework for situating collaborative systems, it does not specify the mechanisms through which shared space is constructed across physical and virtual environments. Addressing this gap, their subsequent work introduced the concept of *MR boundaries* as a mechanism for understanding how distinct spaces are coupled through interfaces that mediate perception and action [33]. MR boundaries conceptualise shared space as a structured interface that regulates how participants gain access to remote people, objects, and activity, thereby influencing awareness, coordination, and control.

Koleva et al. [35] refined this concept with properties that characterise how MR boundaries can be configured to support different styles of cooperative activity. They organised these properties into three groups. *Permeability* describes

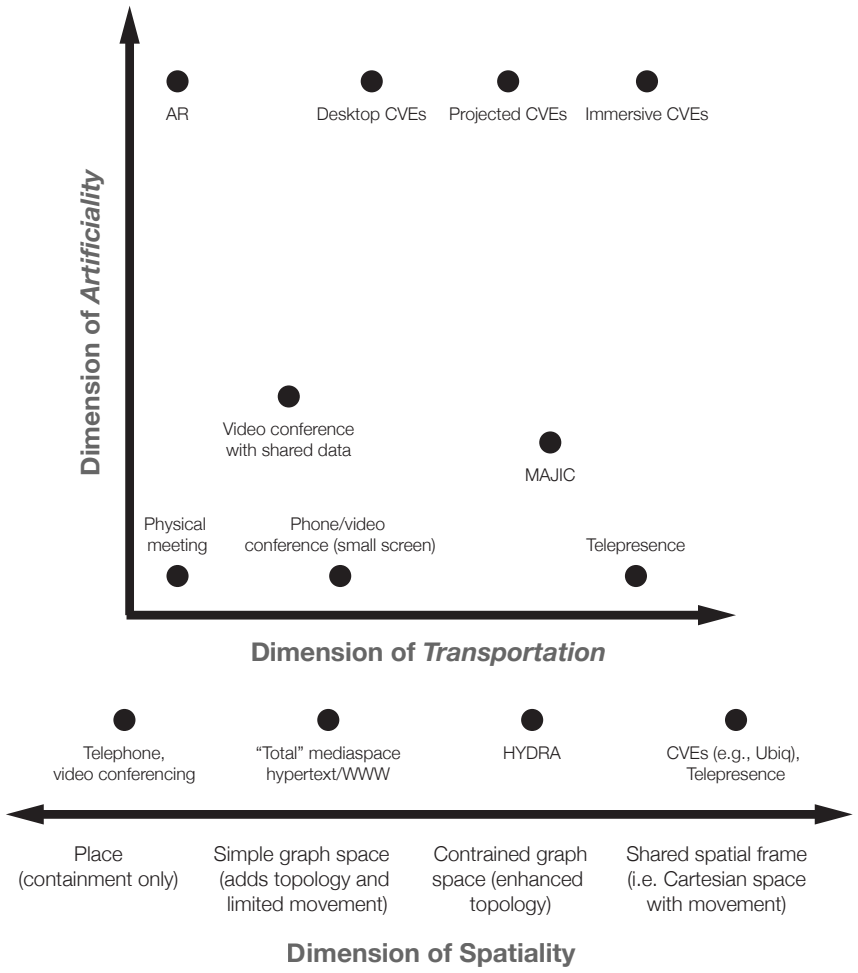


Figure 2.4 Benford et al.'s [46] classification of shared spaces along three dimensions: transportation, artificiality, and spatiality. Image adapted from the original.

how sensory information passes through a boundary, including what is visible, audible, or physically traversable, and how such information may be attenuated, transformed, or aggregated. *Situation* captures the spatial characteristics of a boundary, such as its location and orientation relative to participants, its mobility, and whether it is segmented into multiple regions with different properties. *Dynamics* concerns the temporal aspects of boundaries, including their lifetime and the extent to which their properties can be reconfigured over time. Together, these properties make explicit that MR boundaries are flexible spatial constructs whose design directly shapes participants' spatial understanding, privacy, and opportunities for interaction.

2.5.3 Volumetric Spatial Coupling

As shared space construction methods become more elaborate, boundaries between MR environments provide limited explanatory power for how environments are coupled. In response, researchers have proposed volumetric models that describe how portions of one environment are incorporated into another through geometric transformation and partial overlap [65, 69, 151]. Ohl [69] addressed this through the concept of *immersive extent*, modelling "tele-immersion" as a volumetric coupling between environments rather than as a boundary relation. Immersive extent describes how a *scene volume* is extracted from a remote *location volume*, warped, and included relative to a local location. Similar to MR boundaries, immersive extent is directional. Specifically, the immersive extent from one collaborator to another can differ from the reverse direction, allowing asymmetric spatial relationships to be modelled explicitly.

Ohl [69] decomposed immersive extent into three stages that parameterise this directed coupling:

- *Scene extraction* specifies which portions of the remote location are captured and represented. The author described extraction in terms of semantic segmentation and illustrated coarse but generalisable variants such as extracting the full location, the user, or only the head.

- *Scene warping* then applies a spatial distortion to both the extracted scene volume and the enclosing location volume, preserving their mutual relationships while altering scale or geometry. The framework distinguished identity, upscaling, downscaling, and more general non-similarity warps, highlighting how warping changes the perceived relationship between remote and local space.

- *Scene inclusion* places the warped scene volume relative to the local location volume through a Euclidean transformation. The author characterised inclusion relations such as disjoint, adjacent, partial, perfect, subset, and superset [69]. This final stage determines whether and how volumes over-

lap, constraining which forms of direct spatial interaction are possible.

Two constructs emerging from this framework are particularly relevant for this thesis [69]. First, *directly shared space* was introduced as the intersection volume in which physical and perceived remote spaces overlap, providing a precise way to reason about when direct interaction with remote elements becomes possible. Second, the framework discussed how changes in extraction, warping, or inclusion can shift where users experience remote content as located relative to their local environment, motivating the concept of *immersive location shift* as a way to reason about the perceived relocation of remote activity within a user's action space. These constructs are especially useful for analysing asymmetric CMR, where collaborators may not only have different views but also operate within differently structured shared volumes, with direct consequences for grounding and coordination (Section 2.1).

2.5.4 Task-Driven Extensions of Volumetric Coupling

A complementary, systems-driven approach was provided by Herskovitz et al. [70], who analysed how CMR systems established spatial correspondence across multiple physical environments. They distinguished *object-centric*, *perspective-driven*, and *mesh-based* approaches for anchoring and relating remote content across rooms and validated this framework through the implementation of prototypes.

Building on this task-driven perspective, Wong et al. [65] subsequently extended and formalised the approach for collaboration across physically dissimilar environments. The proposed *Spatial Heterogeneity Framework* operationalised shared space construction through four linked elements: *activity zones*, a *heterogeneity ladder*, *blended proxemics*, and an *MR solutions matrix* [65]. Relative to Ohl's [69] decomposition of volumetric coupling into scene extraction, warping, and inclusion, this framework provides a procedure for selecting the coupled region, characterising spatial mismatch, and determining which embodied cues must remain interpretable for a given activity [65].

- *Activity zones* ground coupling in the spatial organisation of the task. The framework partitions each environment into an *active* zone (task-relevant) and an *inactive* zone, and further distinguishes *public* and *private* regions within the active zone. It then defines a *blended zone* as the public region intended to correspond across sites, and an *independent zone* that can remain visible without one-to-one mapping. This formulation makes explicit, in task terms, which portions of a location are selected for coupling, and it supports asymmetric, discontiguous, and evolving coupling targets as tasks shift.

- The *heterogeneity ladder* describes the mismatch between two blended zones by the transformation required to map one onto the other. The framework orders cases from *causal* and *rigid* through *affine*, *non-linear continuous*, *non-linear discontinuous*, and *categorical* mapping [65]. This framing connects directly to volumetric warping in Ohl's model, while making discontinuities and categorical correspondences explicit design cases when physical layouts cannot be related by a single continuous transformation [69, 65].

- *Blended proxemics* specify which spatial and embodied cues the shared environment should preserve across sites. The framework enumerates six proxemic requirements: *physical manipulation* of corresponding objects, *one-to-one scale* of action, *spatial deictic cues*, *congruent body position and trajectory* between landmarks, *consistent vicinity* around selected landmarks, and *verbal referencing* to shared objects [65]. These requirements clarify which aspects of embodied coordination are threatened as heterogeneity increases.

- The *MR solutions matrix* situates representative system classes by the heterogeneity levels they assume and the blending techniques they use to satisfy the targeted proxemic requirements. The framework contrasts, for example, systems that rely on rigid similarity (e.g., Holoportation [152]) with

approaches that handle affine differences (e.g., Blended Whiteboard [153]) and strategies that compensate for higher heterogeneity through warping, redirection, or relocation [65]. As a synthesis, the matrix links zone selection, mismatch characterisation, and proxemic requirements to a concrete space of implementation strategies, making volumetric coupling decisions analysable as task-conditioned trade-offs in heterogeneous spaces.

Summary and Implications

The models reviewed in this section provide complementary perspectives on how CMR systems construct shared space, supplying vocabulary for describing the representational choices examined throughout this dissertation. Taxonomic dimensions such as transportation, artificiality, and spatiality [46] situate the systems developed in Chapters 4 and 5 within a broader design space. Boundary-based models foreground key concepts such as permeability and directionality, which we use to discuss how asymmetry is instantiated [33, 35], while volumetric coupling models provide precise vocabulary for reasoning about scene extraction, warping, and inclusion [69], operations that inform the environmental representation strategies in Section 2.7. Task-driven extensions link shared space construction to required system affordances such as spatial referencing and physical manipulation [65], connecting these models to the grounding and awareness concerns from Section 2.1. This dissertation builds on these foundations by examining how environmental context, beyond symbolic anchors alone, shapes the construction and enactment of shared space under asymmetric conditions.

2.6 Types of Asymmetry in Collaborative Mixed Realities

As discussed in Section 2.3, early CSCW research identified asymmetry as both a design challenge and a resource for structuring collaboration. This dual role persists in CMR, where differences in devices, environments, and roles frequently shape collaborative activity. In this section, we introduce four asymmetry types that recur across CMR systems and studies: technological (interface) asymmetry,

role asymmetry, view and information asymmetry, and environmental context asymmetry. The subsequent sections examine representative systems for each of these types.

Prior surveys documented the prevalence of asymmetric configurations in CMR and adjacent cross-reality settings, primarily treating symmetry as a classificatory attribute. Ens et al. [43] coded symmetry alongside other system dimensions and reported that asymmetric configurations were common, particularly in remote expert settings. Sereno et al. [44] operationalised technology symmetry and role symmetry in collaborative AR, and reported that these attributes frequently aligned across surveyed work. Complementary cross-device combinations have also been synthesised through the lens of hybrid user interfaces, which characterised asymmetric roles and recurring patterns for distributing content and control across components [154]. Broader reviews of asymmetric interaction proposed finer-grained classifications of asymmetry across interactive systems, offering a wider set of categories than we adopt here [42].

Several surveys examined asymmetry within more narrowly scoped domains or related problem spaces. Thomsen et al. [155] analysed asymmetric collaboration in immersive learning and training environments, emphasising differentiated roles and interfaces in instructional settings. In multiplayer VR games, Rogers et al. [118] proposed a taxonomy of asymmetric systems that extended earlier work on asymmetric play [117], focusing on role complementarity and experiential balance. More broadly, Brudy et al. [24] explored heterogeneous device ecologies and distributed interfaces beyond MR, highlighting interaction design and coordination challenges that also surface in CMR configurations.

In practice, interface, role, view, and environmental context asymmetries frequently co-occur [43, 44]. Any categorisation therefore reflects a pragmatic structuring choice rather than an orthogonal decomposition. We prioritise three well-established asymmetries that recur across CMR work, and we introduce environmental context asymmetry as a distinct type to foreground differences in access to, and interpretability of, collaborators' physical surroundings. To avoid

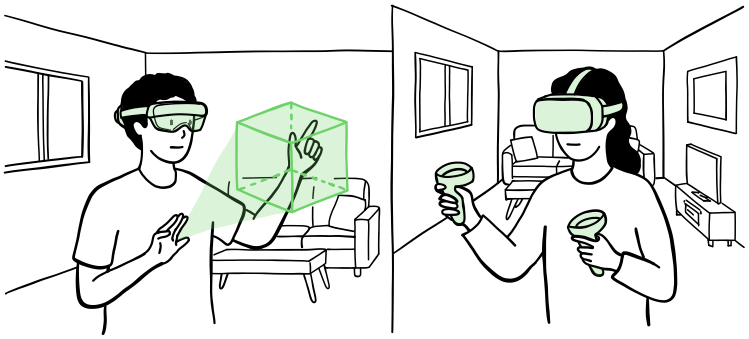


Figure 2.5 Illustrative example of technological asymmetry in CMR. One collaborator uses an AR HMD and hand tracking for manipulation of a cube (left), while the other uses a VR HMD and controllers (right).

duplicating systems across subsections, we organise this review by the asymmetry type that is most relevant to a system's design and evaluation, and we treat co-occurring asymmetries as secondary conditions that are cross-referenced where needed.

2.6.1 Technological Asymmetry: Interface Differences

Technological asymmetry refers to collaborators using different interaction and display technologies, leading to differences in input expressivity, output fidelity, and constraints such as mobility and field of view (Figure 2.5). Common reasons for interface asymmetry include device availability, task requirements, and user preferences. Similar to Bréhault et al. [42], we note that studies and systems often approached interface asymmetry from one of two perspectives: (i) as a constraint that must be managed to enable effective collaboration despite differing capabilities, or (ii) as an opportunity to leverage complementary strengths of different technologies to enhance collaborative outcomes [42]. We observe that these perspectives often overlap in practice. For instance, a system may treat a certain technological configuration as a challenge to be mitigated but ultimately identify ways to exploit complementary capabilities that arise from the asymmetry [56]. In what follows, we discuss common pairings of output and input modalities in asymmetric CMR systems through representative systems and studies.

2.6.1.1 Output Modality Asymmetry

Output modality asymmetry describes collaborative configurations in which participants access shared tasks, spaces, or artefacts through fundamentally different display technologies. In asymmetric CMR systems, these differences commonly manifest as pairings between immersive and non-immersive displays, such as AR HMDs used for situated action and desktop or VR displays used for remote inspection, guidance, or oversight. Prior work has repeatedly adopted such asymmetric output configurations to balance mobility, immersion, visual fidelity, and accessibility, often treating output differences as a deliberate design choice rather than an incidental constraint.

A prominent cluster of systems is grounded in *physical* environments, where AR supports local, situated action while remote collaborators use desktop or VR displays to access a mediated view of the workspace. Early work already explored these configurations. Ahlers et al. [156] presented a distributed AR platform for collaborative interior design in which each participant ran a replicated, desktop-based AR interface to view and manipulate spatially registered virtual furniture in a camera-mediated view of a physical room. Notably, due to technical limitations, remote users received static images of the scene rather than a real-time view [156]. Billinghurst et al. [34] explicitly examined asymmetries between wearable HMDs and desktop monitors in collaborative tasks, motivated by scenarios in which a mobile fieldworker is assisted by a deskbound expert, and emphasised how differences in display properties shaped what each role could perceive and contribute [34]. Related mobile AR work such as *MARS* [157] supported multiple display configurations, including outdoor use with HMDs and HHDs alongside indoor desktop and immersive interfaces. This reflects how output modality could allocate mobility and context-preserving overlays to roaming users while supporting authoring and broader overviews on stationary setups.

Johnson et al. [158] provided a systematic comparison of two asymmetric output modalities for remote physical collaboration, comparing AR HHDs and

AR HMDs for the local worker while a remote helper consistently used a desktop display. Their controlled study showed that head-mounted output better supported dynamic, high-mobility tasks by providing a continuous view of the workspace, enabling more proactive assistance and marginally faster completion times. In contrast, handheld output afforded a more stable and visually detailed view in static tasks, supporting subtle visual communication and shaping collaborators' perceptions of success.

Later systems increasingly paired AR and VR HMDs for remote collaboration around physical tasks. *CoVAR* placed the local collaborator in situ using an AR HMD, while the remote collaborator used a VR HMD to inspect and interact with a reconstructed representation of the local environment, using the asymmetry to combine local access to the physical world with remote viewpoint freedom and immersive spatial understanding [60]. *SharedSphere* similarly paired an optical see-through (OST) AR HMD for a local host with a VR HMD for a remote guest, enabling the guest to explore a live panoramic view with rotational independence while the host could observe the guest's view direction and gestures [61].

A second physically grounded body of work used immersive output for the remote participant while local participants perceived that remote collaborator through situated or robotic output in the destination environment. *Beaming* explicitly embraced technological asymmetry by reconstructing a physical environment for a remote visitor using an immersive display, such as an HMD or Cave Automatic Virtual Environment (CAVE), while the destination site could render the visitor through "destination interventions" such as projections, avatars, or embodied surrogates [159]. *VROOM* similarly combined immersive VR output for the remote user, based on a live 360°-video view captured from a telepresence robot, with AR output for the local user that overlaid a life-sized avatar of the remote participant onto the robot body [160]. *VRoxy* further advanced this embodied telepresence direction by enabling a remote VR user to navigate and collaborate within a large physical environment through a robotic proxy, while the co-located collaborator relied on the robot's embodied displays and

gestures rather than adopting immersive output [161]. Across these systems, output modality was designed to provide remote users with immersive access while reducing the burden of wearing an HMD for local participants.

A complementary set of systems is grounded in *virtual* contexts, where one user experiences a scene through immersive VR output while others access it through non-HMD displays. *ShareVR* enabled co-located experiences in which an HMD user was immersed in VR, while non-HMD participants viewed and engaged with the virtual world through floor projections, tracked HHDs acting as movable windows, and mirrored screens [162]. In distributed settings, *TransceiVR* paired a VR HMD user with an external collaborator using a touchscreen tablet, using asymmetric output to enable assistance, exploration, and annotation for existing VR applications without requiring all collaborators to wear headsets [63].

Pan et al. [53] examined output modality asymmetry in collaborative mixed-reality design tasks by comparing AR–AR, AR–VR, AR–VR with a full virtual body, and AR–desktop configurations. Their study showed that asymmetric output modalities, particularly AR–VR and AR–desktop pairings, led to unequal contributions during three-dimensional interaction, with the more immersive participant often assuming a leadership role. Notably, this effect was mitigated when embodiment cues were strengthened through a virtual body representation. These findings indicate that output modality asymmetry can systematically shape collaborative roles and power relations in spatial tasks, extending the implications of asymmetric displays beyond perceptual access to organisational structure within collaboration.

Finally, work on VR–desktop collaboration studied output modality asymmetry as a practical configuration driven by device availability and complementary affordances. Tong et al. [56] examined collaborative problem solving in a visualisation system where one participant used a VR HMD and the other used a 2D desktop display, reporting unanticipated role differentiation and perceived benefits aligned with the distinct representational capacities of the two output modalities. Related findings were reported in a gaming context by Jung et al.

[163], who compared asymmetric VR–desktop play under collaborative and competitive conditions, showing that immersive VR output consistently enhanced reported experience while desktop output afforded performance advantages in competitive tasks [163].

Across this body of work, output modality asymmetry appears as a recurring strategy for allocating perceptual resources and structuring collaboration. VR interfaces were often used in situations requiring spatial inspection, high-level overview, or embodiment, while AR supported situated action within the physical workspace, and desktop, tablet, projection, or robotic displays broadened accessibility. Beyond perceptual access, several studies also reported behavioural implications of output modality asymmetry, including differences in initiative, communication patterns, contribution, and leadership, indicating that display configurations can actively shape how collaborators act, coordinate, and assume roles during shared tasks. We explore these behavioural implications further in Chapter 3.

2.6.1.2 Input Modality Asymmetry

In line with Bréhault et al. [42], we treat an input modality as a coupling between a device and an interaction language, written as ⟨device, language⟩ [164]. For example, a standard mouse typically affords a 2D point-and-click language by default, while a six-degrees of freedom (DoF)-tracked controller often affords direct 3D manipulation. In asymmetric CMR, input modality differences are often coupled to output modality asymmetry, since display form factors and sensing constrain what kinds of actions can be expressed. In what follows, we discuss common pairings of input modalities in asymmetric CMR systems through representative systems and studies.

Stafford et al. [165] presented "god-like" techniques in which tabletop users used hands and physical props to create and position virtual objects and referential cues for an outdoor AR user, who interacted with these augmentations through a wearable interface and dedicated manipulation input [165]. In this configuration, output modality directly supported these interactions, as the in-

door user acted over an overview representation, while the outdoor user acted through a situated first-person view of the task environment.

Remote expert systems frequently instantiated input modality asymmetry as a combination of physical manipulation on one side and 2D deictic or sketch input on the other [166, 167]. Kim et al. [166] studied a system in which a local participant manipulated physical objects while a remote participant used a mouse on a desktop interface to provide a live pointer and draw annotations over a shared video view. Their results indicated that augmenting speech and video with such 2D input channels improved reported copresence-related measures relative to a baseline without these cues, and they discussed trade-offs between transient pointer cues and more persistent annotation cues [166].

Another common occurrence in systems centred on one user's physical space is locomotion. In this case, a local user navigates their physical environment through natural walking, while a remote user relies on device-dependent navigation techniques [39, 168, 169, 62]. For example, Tian et al. [62] paired a local AR user who navigated their physical environment through walking with a remote VR user who used a controller joystick for locomotion. Earlier work that relied on local user-driven 360°-video capture was often even more constrained, as the remote user could only navigate the captured view rotationally [170, 171]. Follow-up work such as *VirtualNexus* explored how to address this constraint by bringing parts of the environment to the user instead of moving the user through the environment [172].

TransceiVR demonstrated the value of adding an asymmetric 2D input language that could be integrated into the VR user's 3D interaction context [63]. This system enabled an external collaborator to draw annotations on a tablet and highlight controller inputs, with these cues rendered to the VR user during task performance. In their study, the authors reported improvements over a baseline that relied on conventional VR mirroring and verbal communication, including higher task success and reductions in completion time, errors, and workload [63].

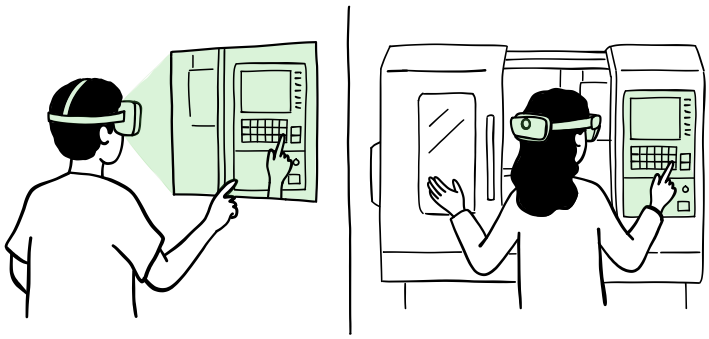


Figure 2.6 Illustrative example of role asymmetry in CMR. A remote collaborator assumes an expert role with oversight and guidance capabilities (left), while the local collaborator performs hands-on manipulation (right).

Lastly, Grandi et al. [173] compared collaborative object manipulation across co-located VR–VR, AR–AR, and VR–AR conditions, studying input asymmetry through tracked six-DoF controllers for the VR user and touch plus device-motion techniques for the tablet-based AR user [173]. They reported that the VR–VR condition was fastest, followed by VR–AR, and hypothesised that this was due to a more intuitive interaction mapping for VR controllers.

2.6.2 Role Asymmetry

Role asymmetry refers to differences in collaborators' roles that are stipulated by interface capabilities or by design choices. Designated roles can benefit the collaboration process. For example, Zhu and Zhou argued that "[b]ased on the findings of social psychologists a role-based collaborative system allows users to know and fulfill their obligations while respecting the rights and authority of other users in collaboration" [174]. Several proposed CMR systems explicitly incorporated role-specific capabilities and limitations as a means of structuring collaboration [54, 175, 176, 57].

Similar to the previously discussed work by Grandi et al. [173], Olin et al. [54] conducted a study of an asymmetric collaborative VR system that enabled an HMD user to collaborate with an HHD user in an object manipulation task. In this configuration, the HMD user was represented by a full avatar, whereas the HHD user was represented as a floating head. The HMD user could create, manipulate, and delete objects, while the HHD user could create and delete

objects and highlight them. In addition, the HHD user had access to a building plan outlining the task solution. As such, asymmetries of interface, role, and information access were deliberately enforced. In their evaluation, the authors found that HHD users were perceived as leaders and hypothesised that this effect may have stemmed from their privileged access to task-relevant overview information, a concept elaborated in the following section. As an alternative explanation, the authors also suggested that the higher relative eye level of the HHD user with respect to the HMD user could have contributed to this leadership effect [177].

Related evidence that role and information assignment shaped perceived leadership was provided by Norman et al. [55]. In their system, two co-located users collaborated using MR HMDs, while a third participant joined remotely via a desktop-based AR interface that supported visual guidance cues such as pointing and virtual markers but did not provide direct object manipulation. The study compared a coordinator condition, in which the remote participant had access to both local participants' preferences and was expected to mediate them, against a condition where this information was distributed symmetrically. While task completion time did not differ, the coordinator condition increased the remote participant's use of visual cues and reported mental effort, and local participants perceived the remote participant as more dominant. Together with Olin et al. [54], this work illustrates how role asymmetry can emerge from coupled asymmetries in agency and information access, where privileged overview affords coordination authority even in the absence of direct control.

Similar dynamics of information-mediated role differentiation were also observed in co-located settings. *ShARe* combined an AR HMD with projection to include a non-HMD participant, resulting in asymmetries in both visibility and control over shared content [178]. Because the HMD wearer could regulate what was visible to others through head orientation, visibility itself became a role-defining resource. In application examples, this led to observer-like roles with access to hidden information, while non-HMD participants could directly

manipulate the physical setup without access to that information. This work demonstrates that role asymmetry can arise from perceptual and representational control even under physical copresence.

Whereas the preceding systems relied on emergent differences in information access and interaction, *XRDirector* framed role asymmetry as an explicit coordination strategy for immersive authoring [57]. The system defined filmmaking-inspired roles with distinct responsibilities and capabilities, including a director role with full scene control alongside specialised roles such as actor, camera, and light. In workshops and reenactments, role partitioning supported coordination by distributing labour and reducing interaction complexity for non-director roles, while assigning the director as an orchestration hub. At the same time, *XRDirector* showed that structured role assignment did not eliminate coordination work, particularly when collaborators required stable shared reference points across devices, motivating additional role support such as a dedicated camera perspective (see Section 2.6.3).

Role asymmetry was also used to combine complementary viewpoints and manipulation capabilities within co-located interaction. Le Chénéchal et al. [176] introduced "Giant" and "Ant" roles for multi-scale 3D manipulation, where the Giant performed fast, coarse positioning from an external overview while the Ant operated from within the object to perform finer-grained rotations and scaling. The design further accommodated optional helper and spectator roles.

2.6.3 View and Information Asymmetry

Asymmetric information refers to situations in which one actor has access to information that at least one other actor lacks, a phenomenon widely studied in economic theory and communication [179, 180]. In collaborative work, such asymmetry is closely tied to limits of perception, attention, and coordination. Expanding communication across multiple channels can lead to information overload [180]. Consequently, collaborative systems may deliberately introduce information asymmetry to filter or prioritise information and reduce per-user burden, with risks of misunderstandings and degraded performance when col-

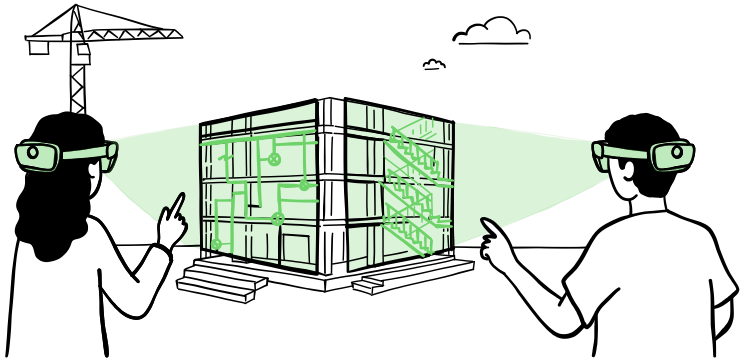


Figure 2.7 Illustrative example of view and information asymmetry in CMR. Both collaborators are viewing information overlaid on a construction site, but one can view electrical schematics (left) while the other sees building blueprints for staircases (right).

laborators cannot reconstruct what others know or see.

Within CSCW, these asymmetries can exceed those found face-to-face and may arise across sensory channels, including visual, audio, haptic, and olfactory modalities [181]. Jää-Aro and Snowdon [182] captured this design space through the concept of subjectivity, defined as the ability to add viewer-dependent features to a system. Jää-Aro and Snowdon [182] argued that fully objective representations are overly restrictive for collaboration around complex data. Building on this view, we focus on visual manifestations of asymmetric information, commonly termed *subjective views* [183, 182]. Snowdon et al. [183] modelled subjective views as transformations of shared data that structured how interaction propagated between views and the underlying representation.

Smith and Mariani [184] implemented subjective views in *DIVE* [32] by defining subjective modifiers that controlled object appearance on a per-user basis. Building on this capability, Steed et al. [58] introduced modifiers including invisible, transparent, wireframe, dim, and bright. These mechanisms illustrated a pragmatic form of view-dependent filtering, where collaborators could emphasise or suppress objects without changing the shared world state.

Subjective views have also been used to manage selective disclosure. Butz et al. [59] considered privacy in multi-user augmented environments where objects could be public or private, and introduced two view-dependent interaction

metaphors for inspecting and changing that state. *Privacy (and publicity) lamps* acted as virtual spotlights whose beams designated sets of objects to hide from or reveal to remote users. *Vampire mirrors* reflected only public objects, enabling users to review privacy state at a glance and to change it by interacting with the mirror representation of objects.

In collaborative scientific visualisation, Park et al. [185] reported how multiple perspectives in a CAVE supported collaboration around multivariate data by allowing collaborators to tailor viewpoints and representations to their immediate analytic needs. At the same time, they noted that view divergence shifted effort toward coordination, since collaborators had to establish correspondences between perspectives to maintain shared understanding.

Pettersson et al. [186] provided complementary evidence that selective visibility could improve collaboration when it aligned with task structure. They evaluated layer partitioning in a co-located 3D visualisation task by giving each collaborator access to a subset of layers rather than the full multilayer view. Dyads completed the task faster under partitioning, with similar error rates, which they attributed to reduced visual clutter and fewer occlusion-resolution demands, even though partitioning required additional communication to integrate partial views.

Salimian et al. [187] examined view divergence in a CMR setting by varying view congruence between co-located tabletop collaborators and a remote HMD user. Their less congruent condition supported inspection and manipulation without requiring strict view alignment and was often preferred, while it maintained shared reference through consistent spatial mappings.

Borowski et al. [64] addressed view asymmetry across AR and 2D displays by introducing a shared representational substrate that stored elements in 3D while rendering them in heterogeneous ways. A key contribution of this work was a configurable projection plane that projected 3D content onto a 2D infinite canvas, enabling 2D collaborators to view and manipulate spatially grounded content through a planar representation while preserving meaningful spatial rela-



Figure 2.8 Illustrative example of environmental context asymmetry in CMR. One collaborator (left) is situated in a park, while the other (right) is in an office and is viewing the other's streamed view.

tionships. The system further supported cross-dimensional awareness through representations that mapped 3D poses into 2D and 2D actions back into 3D, helping collaborators interpret each other across dimensionalities.

Overall, information asymmetry can be understood as a design trade-off between reducing local complexity and increasing the need for explicit grounding, and it is often tightly coupled to other forms of asymmetry such as technological and role asymmetry.

2.6.4 Environmental Context Asymmetry

We define *environmental context asymmetry* as systematic differences in collaborators' access to, and interpretability of, one another's immediate surroundings, including the spatial arrangement of artefacts and place-level constraints that shape situated action and communication. The construct is closely related to workspace awareness (Section 2.1.3) because it concerns the perceptual basis on which collaborators interpret others' location, actions, and references within a workspace. Within the shared-space models reviewed in Section 2.5, environmental context asymmetry can be understood as a consequence of how shared space is constructed and regulated. In boundary-based terms, it arises when MR boundaries have limited permeability or selectively filter environmental cues, such that remote collaborators cannot recover the situational contingencies that make actions accountable. In volumetric terms, it reflects directional choices

about scene extraction, warping, and inclusion that determine which parts of a location are made present, how they relate geometrically to the remote user, and whether collaborators have access to a sufficiently structured volume for deictic reference and interpretation.

Voida et al. [41] introduced a closely related notion through the concept of *asymmetry of place* (see Section 2.3), emphasising that collaborators may inhabit widely different physical settings and that these differences shape what becomes available and actionable through mediated interaction. Bréhault et al. [42] later introduced the complementary notion of *shared feedback asymmetry* through *environment sharing*, defined in terms of whether users have the same view of each other's environment. Environmental context asymmetry bridges these perspectives by foregrounding situations in which a collaborator's surroundings and local constraints are not perceptually legible to others, despite remaining consequential for interpretation and action.

Relatedly, Wong et al. [188] introduced the *Jamais Vu effect* to characterise breakdowns in interpretability that occurred when collaborators coordinated across dissimilar spaces and observed behaviour that appeared implausible because relevant differences in spatial layout and furniture were not available to them. The authors explicitly described these moments as breaking the illusion of copresence (Section 2.1.2), since collaborators could no longer treat observed actions as grounded in a shared intelligible space [188]. In this respect, Jamais Vu aligns with the concepts of "Breaks in Presence" and Psi (Section 2.4.3), where mismatches between observed behaviour and expected environmental contingencies disrupt users' sense of "being together" in a coherent situation. Although introduced in the context of heterogeneous layouts and spatial anchors, the concept extends to broader forms of environmental context asymmetry in which actions remain visible but lose their accountable relation to the environment that produces them [188, 41].

Empirical work provides direct evidence that unequal access to a collaborator's surroundings impairs spatial reference and increases coordination effort.

For example, Müller et al. [189] examined how, when collaborators lacked sufficient shared visual context, references to physical objects and locations became harder to establish, shifting interaction toward more explicit negotiation of what was visible and where actions were directed. Furthermore, Johnson et al. [190] studied remote collaboration in unfamiliar environments and showed how representational support shaped coordination when helpers had limited access to spatial task context. They varied how spatial task state was presented to the helper (map-like versus list-like) and whether MR guidance assisted the worker during object referencing, demonstrating that these supports changed how effectively collaborators could coordinate under constrained shared spatial context.

Summary and Implications

This section provided a vocabulary of asymmetry types grounded in prior CMR literature, organising recurring differences in interfaces, roles, views, and environmental context into analysable categories [43, 44, 42]. These asymmetries frequently co-occur, such that changes intended to address one mismatch can reconfigure related coordination resources, including authority, participation, and what can be treated as mutually available. We define *environmental context asymmetry* to foreground situations in which collaborators have systematically different access to the place-level contingencies that render action interpretable, connecting asymmetry taxonomies to the shared space construction models in Section 2.5 and to the grounding and awareness concerns in Section 2.1. This definition directly informs the dissertation's empirical focus. Chapter 3 examines how technological asymmetry shapes leadership and coordination when one collaborator lacks direct access to the other's physical environment. Chapters 4 and 5 develop representational strategies that address environmental context asymmetry through augmentation and transformation, which we review next.

2.7 Representational Strategies for Managing Asymmetry in Collaborative Mixed Reality

Asymmetric CMR configurations introduce persistent differences in what collaborators can perceive, reference, and act upon (Section 2.6). While earlier sections

discussed how shared space can be constructed across heterogeneous systems, this section focuses more narrowly on how spatial representations are *adapted* once asymmetry is taken as a given. In line with the scope of this dissertation, we do not aim to survey all mechanisms proposed for managing asymmetry. Instead, we synthesise two representational levers that recur across prior work and that directly motivate the empirical and system-building investigations in Chapters 3–5: (i) augmenting shared space to support grounding and awareness under partial access, and (ii) transforming spatial representations to reconcile heterogeneous physical contexts.

2.7.1 Augmenting Shared Space to Support Grounding and Awareness

Even when collaborators operate over a shared underlying representation, asymmetries in viewpoint, sensing, field of view, and physical access can produce systematic gaps in awareness and grounding. A common response is to augment the shared space with additional cues or representations that make others' actions and local context more legible. Such augmentation is typically selective and task-driven, reflecting the need to balance informational sufficiency against overload in already complex environments.

A key challenge in asymmetric CMR is enabling collaborators to establish what actions, utterances, or gestures refer to when access to space, viewpoint, and perceptual context is uneven. Chastine et al. [191] framed this challenge as *inter-referential awareness* and highlighted the importance of techniques that support referencing across divergent perspectives. In the following, we organise augmentation approaches by whether referential intent is conveyed implicitly through attention cues or explicitly through deliberate actions such as pointing and annotation.

2.7.1.1 Attention cues as implicit referential actions

One class of augmentation approaches conveys referential intent implicitly by making collaborators' *attention* legible, using gaze- and view-based cues as prox-

ies for action, focus, and communicative intent. A range of systems explored sharing eye gaze, head orientation, or derived attention regions to support target identification and joint focus. Erickson et al. [192] demonstrated that gaze rays can support visual target identification in collaborative AR, while also showing that accuracy and latency critically shape both performance and users' subjective confidence. Piumsomboon et al. [193] extended this line of work in an asymmetric AR–VR setting, showing that combinations of awareness cues, particularly field of view (FOV) frustums paired with head-gaze rays, improved task performance, usability, and copresence, while also shifting interaction away from verbal negotiation. Subsequent work refined how attention cues are represented and interpreted. Bovo et al. [194] proposed the *Cone of Vision*, a probabilistic attention cue derived from fixation behaviour rather than FOV or gaze indicators, and showed that such abstraction can reduce cognitive effort while supporting smoother transitions between independent and joint work.

Jing et al. [195] and Jing et al. [196] demonstrated that bidirectional and near-gaze visualisations, which adapt cue presentation to users' ongoing gaze behaviour and place cues close to the locus of attention, increased confidence, joint attention, and copresence while reducing perceived distraction. In a VR-based collaborative 360° video editing system, *CollaVR*, Nguyen et al. [197] showed that sharing viewports and supporting lightweight perspective alignment, together with explicit task-specific cues such as activity icons and shared annotations, enabled collaborators to rely on implicit deictic references such as "this" or "that", reducing the need for explicit verbal grounding.

2.7.1.2 Explicit referential actions through gestures and annotations

Complementing attention-based approaches, other systems externalised intent more explicitly through deliberate referential actions such as pointing, sketching, and annotation. These techniques foreground what a collaborator *means* rather than where they are looking, and they often trade immediacy for precision. Early wearable teleconferencing work already demonstrated the importance

of explicit pointing under asymmetric access. Bauer et al. [198] studied expert guidance through a head-mounted video link and showed that pointing acts dominated communication and were decisive in the vast majority of interactions, far outweighing verbal instruction alone. A study of MR navigation by Stafford et al. [199] similarly found that visual cue-based techniques substantially outperformed audio-only guidance, with gesture- and pointer-based cues enabling faster and more expressive communication across exocentric and egocentric viewpoints [199].

In a video-based CMR system, Kim et al. [166] showed that simple spatial cues such as pointers and drawing annotations significantly improved copresence and coordination compared to voice-only communication, with pointers being both preferred and efficient for bidirectional interaction. Addressing more challenging cases of viewpoint divergence, Oda and Feiner [200] introduced *GARDEN*, a technique that reconstructs and shares a local 3D volume around a referent, achieving substantially higher accuracy than laser pointers or arrows when collaborators viewed the scene from different perspectives, albeit at the cost of additional interaction steps. In live 360° panorama-based collaboration, Teo et al. [201] found that explicit visual annotations outperformed virtual hand gestures alone for conveying precise instructions in spatially complex scenes. Finally, *TransceiVR* demonstrated how action-level cues can be retrofitted onto existing single-user VR applications, enabling external collaborators to provide sketches, highlights, and controller-level cues that improved task success and reduced errors under asymmetric control and visibility [63]. Across these systems, explicit referential artefacts functioned as stabilising elements that anchor intent in shared representations, particularly when attention cues alone were insufficient.

2.7.2 Providing Context Through Environmental Representation

Environmental representations are a common response to *environmental context asymmetry* in CMR, where collaborators have systematically different access

to, and interpretability of, one another's surroundings (Section 2.6.4). These representations also interact with technological, role, and information asymmetries by shaping what each interface can render, what each role can contribute, and what contextual detail is disclosed (Section 2.6). By mediating what is visible, navigable, and referable about a remote workspace, they directly affect grounding and workspace awareness under partial access (Section 2.1). Shared space construction models provide analytic vocabulary for these choices. MR boundaries characterise selective permeability and dynamics (Section 2.5.2), and volumetric coupling frames extraction, warping, and inclusion across sites (Section 2.5.3).

2.7.2.1 Environmental Representation Types

We organise prior work by the primary representation used to convey environmental context to remote collaborators. The categories highlight trade-offs between capture friction, spatial structure, and the extent to which the representation supports navigation and deictic reference under asymmetry (Section 2.6).

360° video and panoramic capture. Panoramic capture provides a low-friction means of sharing environmental context in remote collaboration by prioritising immediate wide-angle visual coverage over explicit spatial structure. Early systems already treated viewpoint independence as central. Kasahara and Rekimoto [202] generated an out-of-body view by integrating first-person frames via simultaneous localisation and mapping (SLAM) into a wider surrounding scene and allowed the remote helper to control a virtual viewpoint. Subsequent work moved from stitched panoramas toward full spherical capture. Kasahara and Rekimoto [203] and Nagai et al. [204] equipped a mobile user with an omnidirectional camera rig and streamed a stabilised 360° view to a remote VR user, who could freely look around while remaining decoupled from the local user's head motion. These systems demonstrate how omnidirectional video can provide broad situational visibility and support basic spatial referencing, while leaving the underlying spatial structure largely implicit.

Live 360° video systems further reinforced this decoupling by capturing a full spherical view around the local user. Early explorations demonstrated the potential of sharing panoramic experiences for social and remote collaboration contexts [205, 61]. Several systems then added interaction and awareness cues to reduce the coordination costs introduced by view independence. Lee et al. [61] presented *SharedSphere*, which augmented live 360° panorama sharing with view-awareness cues and gesture visualisations so collaborators could interpret one another's focus despite independent viewing directions. Speicher et al. [206] enabled remote users to draw annotations over a live 360° stream and ground them in place by projecting those annotations back onto physical surfaces in the local environment using gnomonic projection. More recently, Zaman et al. [170] examined remote assistance that used 360° video as the primary environmental representation, highlighting its strengths for rapid situational understanding alongside persistent limitations in precise spatial reference and depth interpretation.

Overall, 360° video increases access to remote place context, behaving like a highly permeable MR boundary (Section 2.5.2) with limited support for stable volumetric inclusion beyond an image shell (Section 2.5.3). In terms of the asymmetry types in Section 2.6, it primarily reduces *environmental context asymmetry* through wide visual coverage, while often preserving or amplifying *view asymmetry* through deliberate viewpoint independence. This can increase the grounding work required for precise deictic reference (Section 2.6.3).

Geometric and volumetric capture. Geometric capture methods represent the environment as point clouds or meshes, enabling remote collaborators to navigate, maintain stable reference, and ground instructions within a shared spatial frame. Where panoramic views privilege boundary permeability, geometric reconstructions privilege spatial structure, strengthening shared reference for grounding under view and environmental context asymmetry (Section 2.6.3, Section 2.6.4).

Gao et al. [207] demonstrated an approach that used mobile tracking to integrate depth keyframes into a stitched 3D point cloud for remote guidance, combined with a real-time 2D video stream as a reference channel for dynamic scene changes. Building on panoramic approaches, subsequent systems leveraged reconstruction methods to provide more explicit spatial structure, projecting panoramas [208] or live 360° video [209] onto reconstructed meshes or planes to provide a navigable scene with richer visual context.

A broad area of work has explored live volumetric capture using depth sensors or depth inference methods to generate real-time point clouds or fused meshes [62, 210, 211, 152, 212, 213, 214]. For example, Tian et al. [62] combined live point clouds with pre-scanned mesh models and introduced virtual replicas of task objects as spatial communication cues, enabling remote experts to demonstrate complex assembly actions through manipulable stand-ins. Orts-Escolano et al. [152] presented *Holoportation*, a room-scale volumetric telepresence system that fused multi-view RGB-D streams from dense depth camera arrays into temporally consistent textured meshes, enabling life-sized embodied telepresence for remote users in AR and VR. Irlitti et al. [213] introduced a room-scale volumetric telepresence system that captured people and objects with depth camera arrays and triangulated reconstructions into meshes, supporting interaction with objects across remote and co-located configurations.

Geometric and volumetric approaches make site coupling more explicit than panoramic video, since the resulting representations support navigation and persistent spatial anchors. Under the immersive extent model (Section 2.5.3), these systems operationalise shared space construction through decisions about what volume is captured, how it is filtered or scaled, and where it is included relative to the remote collaborator's action space. These choices often reduce environmental context asymmetry and view asymmetry for remote guidance by stabilising reference frames, yet they introduce constraints tied to sensing, bandwidth, and update latency. They also enable role- and information-asymmetric allocations through selective capture and abstraction that prioritise task-relevant

entities such as hands, work surfaces, or instrumented objects (Section 2.6.2).

Symbolic and semantic environmental models. Beyond representations that aim to visually approximate an environment, symbolic and semantic models represent the environment as a structured set of discrete entities and relations [50]. Spaces can be encoded as labelled regions, surfaces, objects, affordances, and constraints, enabling systems to treat the environment as structured state rather than as appearance. Unlike panoramic or geometric approaches, these abstractions are particularly suited to deliberate *information* and *role* asymmetry, for example by giving an expert an interpretable overview while shielding irrelevant or sensitive context from a worker (Section 2.3.3, Section 2.6.2).

Across prior work, symbolic environmental models commonly took the form of 2D floorplan-like encodings, segmented maps, planar surface abstractions, or relational graphs. Lehment et al. [215] derived orthographic height maps and segmented them into categories such as free space, work surfaces, and obstacles, illustrating how symbolic partitions can express navigability and functional regions. At a higher semantic level, Guo et al. [216] used virtual stand-in objects (e.g., sofas, chairs) to virtualise real-world spaces for unidirectional CMR. Surface-centric encodings further emphasised actionable substrates, for example by representing selected physical surfaces as planes [217] or by treating local whiteboards as anchors for a reconfigurable 2D canvas [153]. This aligns with task-driven shared space construction (Section 2.5.4). Symbolic and semantic models can support correspondence at the level of activity zones, affordances, and constraints without requiring rigid spatial equivalence.

2.7.2.2 Representational operations for managing environmental context asymmetry

If representation types describe what is made perceptible about a remote workspace, representational operations describe how shared space is composed, transformed, and kept intelligible across asymmetric devices and places (Section 2.5.2, Section 2.5.3). In practice, operations both reduce environmental

opacity to support grounding and introduce role- or information-specific divergence to manage complexity (Section 2.6).

Composition and layering of representations. Hybrid systems combine multiple environmental representations to balance spatial structure, visual fidelity, and update responsiveness, often layering a navigable scaffold with higher-fidelity views and specialised update mechanisms. In terms of Section 2.6, layering is a pragmatic response to technological constraints while maintaining sufficient environmental legibility to sustain grounding and workspace awareness.

Komiyama et al. [169] introduced *JackIn Space*, which combined first-person camera views with a surrounding 3D point cloud reconstructed from external depth sensors. Remote users could switch between an egocentric view for local detail and an out-of-body view for broader situational awareness, enabling dynamic shifts between situated inspection and overview coordination. Young et al. [218] introduced a mobile variant by combining incremental point clouds with live 360° video, supporting transitions between exocentric and egocentric views.

Teo et al. [219] presented a hybrid CMR system that combined live 360° video and a precomputed 3D reconstruction to support remote collaboration between a local AR user and a remote VR user. Through a user study, the authors showed that 360° video improved task performance and social presence, while the 3D reconstruction complemented it with better spatial understanding, motivating dynamic switching between the two modes in a single system.

Thoravi Kumaravel et al. [220] explored phase-specific layering in *Loki*, where instructors and learners alternated between live volumetric views, recorded sequences, and conventional video windows across observation, coaching, and review, selecting the representation that best matched each phase of instruction.

More recent approaches have used neural scene representations to increase

photorealism while retaining real-time handling of dynamics. Sakashita et al. [221] blended a neural radiance field (NeRF) trained on frames obtained from a head-mounted camera for high-fidelity static structure with live point clouds for dynamic elements such as hands, and found this to be particularly effective for fine-grained manipulation tasks such as floral arrangement. Huang et al. [222] pursued a related layered strategy using a pre-trained 3D Gaussian splatting (3DGS) scene for room-scale navigability and fidelity, while incorporating dynamic changes through tracked object updates and projected 360° video, supported by a hidden mesh for collision and physics.

Transforming and blending representations across sites. Transforming and blending representations across sites most commonly appears in bidirectional or multi-directional CMR, where collaborators must coordinate action and reference across multiple physical contexts. These systems do not transmit one participant's environment into another. Instead, they modify which portions of each site are represented, how they are aligned, and how actions are mapped, such that a shared workspace remains usable under spatial heterogeneity. These patterns can be read as design choices over extraction, warping, and inclusion (Section 2.5.3) and over the selection of blended zones and correspondence level (Section 2.5.4). In what follows, we distinguish object- and surface-centric alignment, action-level retargeting, and mutual space synthesis as three recurring transformation patterns. For a broader review of space transformation techniques in CMR, we refer to Wong et al. [65], which emphasises symbolic and coordinate-space transformations.

A first group of systems performs transformation by aligning task-relevant objects or surfaces across sites, treating them as anchors for collaboration without attempting full-room equivalence. Grønbæk et al. [217] aligned selected physical surfaces and allowed users to switch or overlay alignments as activities moved between anchors. Grønbæk et al. [153] similarly grounded collaboration in physical whiteboards linked to a shared reconfigurable canvas, with view

transformations supporting joint and individual navigation. These systems were primarily surface-centric.

Related systems extended this logic to richer task spaces. In an instructional context, Johnson et al. [38] decomposed the environment into an active space and a peripheral space, representing the novice's immediate task area as a live 3D point cloud and embedding it within a static reconstruction of the expert's familiar workspace to balance awareness and cognitive effort. Irlitti et al. [213] aligned remote locations around shared physical features, such as a table, and rendered people and objects volumetrically, enabling interaction with physical, volumetric, and virtual objects within a hybrid workspace. Across these systems, transformation operates by selecting and aligning task-relevant elements, reducing environmental context asymmetry where it is most consequential and leaving non-critical place differences intact.

A second, more fine-grained form of transformation operates on embodied action and movement. Kang et al. [223] retargeted pointing and touching motions to corresponding remote objects, preserving deictic intent despite layout mismatch. Relatedly, Congdon et al. [224] redirected user movement by mapping locomotion trajectories between dissimilar floorplans and enabled collaborators to walk naturally while remaining aligned to corresponding regions. These approaches preserve interactional meaning by transforming action mappings, mitigating breakdowns from heterogeneous layouts by keeping references interpretable under view and environmental context asymmetry (Section 2.6.3).

A third group synthesises a shared space from multiple inputs and moves beyond the alignment of existing environments [225, 226]. For example, Keshavarzi et al. [225] extracted semantic scene structure from participants' environments and generated a mutual virtual scene that maximised shared functional areas, shifting from alignment toward generative construction. In this setting, environmental context asymmetry is re-specified into a task-oriented shared space that prioritises functional correspondence. This treatment aligns with asymmetry as a deliberate design condition in collaborative systems (Section 2.3).

Summary and Implications

This section examined how asymmetric CMR systems support awareness and reference when collaborators have unequal access to space, viewpoint, and environmental context. Prior work shows that implicit cues such as gaze and view frustums, and explicit cues such as pointing and annotation, can make referential intent visible and reduce ambiguity in deictic communication [193, 200]. Yet their effectiveness depends on anchoring to environmental representations that preserve spatial relationships and support interpretation during ongoing action [207, 219]. Work on spatial transformation further indicates that maintaining usable reference across dissimilar environments can require altering correspondence or constructing new shared frames, with trade-offs between contextual richness and interpretive effort [224, 65, 225]. These observations directly motivate the dissertation's constructive contributions. Chapter 3 characterises collaborative behaviour under static environmental reconstruction to establish baseline coordination dynamics. Chapter 4 investigates how augmenting environmental representations with live context and spatial annotations supports collaboration when the environment changes. Chapter 5 examines how synthesising shared environments from multiple physical contexts affects awareness, reference, and coordination across distributed sites.

Chapter 3

Exploring User Behaviour in Asymmetric Collaborative Mixed Reality

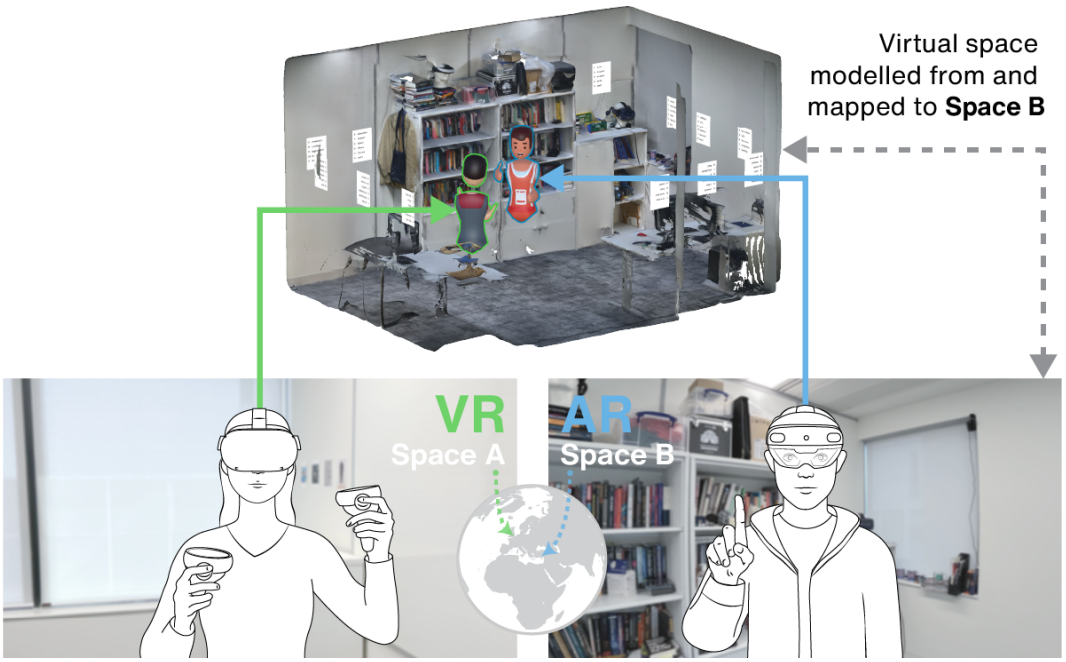


Figure 3.1 Conceptual overview of our CMR system, in which an AR and a VR user can collaborate remotely. The CVE contains a static 3D model of *Space B*. Users are represented with avatars. Line illustrations by Suhyun Park (artist).

This chapter contributes to the dissertation by addressing RQ1 and RQ2. These research questions examine how asymmetries in technology and environmental context shape interaction in asymmetric CMR, as well as the limitations that current environmental representations impose on mutual understanding and task coordination. This work was published at ACM VRST 2022 [2][1].

[1]This paper was co-authored with my primary supervisor, Anthony Steed. I led the system design and implementation, experimental design, data collection, analysis, and writing. Anthony Steed provided guidance at each stage of the work and contributed to writing and editing.

3.1 Context and Motivation

We examine an asymmetric AR–VR configuration in which a remote VR user collaborates with an in-situ AR user while being embedded in a reconstructed representation of the AR user's workspace. This configuration is increasingly common in practice for remote collaboration and expert guidance [60, 62, 38]. At the same time, the literature provides limited empirical accounts of collaboration dynamics for the specific case in which the remote user relies on a *pre-captured* representation that remains static during the session and in which interfaces to this representation differ across collaborators. This gap motivates closer examination of how shared reference is established and maintained under these constraints.

One recurring theme in prior work is that asymmetric configurations can bias participation and perceived influence, often discussed in terms of leadership and talkativeness. Early CVE studies using a shared puzzle task reported strong immersion-associated leadership and talkativeness effects during virtual sessions [67, 68]. These effects weakened or disappeared when the same groups performed comparable tasks in the real world, suggesting that the observed leadership advantage depended on properties of the mediated setting rather than stable interpersonal roles [67, 68]. Steed et al. [52] extended this line of work to distributed sites and reported that leadership ratings were closely associated with talkativeness and individual differences, with immersion effects that varied across participants (see Figure 3.2).

Subsequent work examining technological asymmetry refined this picture by showing that leadership effects depend on how interface differences intersect with task structure, representational cues, and access to spatial information. Pan et al. [53] reported leadership effects in asymmetric AR–VR collaboration that emerged specifically during 3D spatial interaction when the VR user was represented only through controller models. In this AR–VR condition, asymmetry in output modality coincided with unequal contribution and perceived leadership, yet these effects did not appear during 2D interaction and were mitigated when

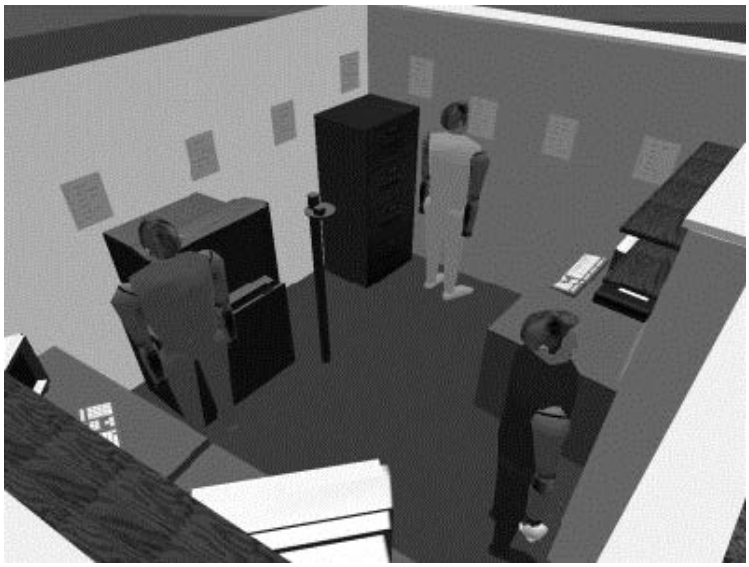


Figure 3.2 Overview of the CVE used in Steed et al. [52], where posters with words were placed on the walls. Reproduced with permission. ©1999 IEEE.

embodiment cues were strengthened through the introduction of a virtual body representation. Norman et al. [55] showed that leadership dynamics can be shaped through role assignment. Assigning a remote participant a coordinator role increased engagement and use of visual communication cues, alongside differences in perceived dominance and mental effort, indicating that participation patterns can be structured through role asymmetry in addition to interface configuration. Olin et al. [54] provided qualitative evidence that spatial configuration and perceptual cues such as eye-level position can shape leadership under technological asymmetry. In their study, HHD users were frequently perceived as leaders despite using less immersive technology, particularly when positioned above their HMD-using collaborators and when holding exclusive task-relevant information.

3.2 Hypotheses

This chapter operationalises RQ1 and RQ2 through hypotheses that capture potential interaction bias via *leadership emergence* and related behavioural markers. Following prior studies that used leadership ratings and talkativeness to characterise contribution and influence [52, 68, 67], we formulate:

Hypothesis 3.1 AR users obtain a higher participant-rated leadership score than VR users.

Hypothesis 3.2 AR users obtain a higher participant-rated talkativeness score and speak more words than VR users.

Additionally, the VR users in our system had access to an additional input modality for locomotion through joystick-based navigation, alongside physical movement. This asymmetry in input expressivity and mobility constitutes a further dimension of technological asymmetry that may influence how participants explore the shared environment and take initiative during collaboration. We therefore hypothesise:

Hypothesis 3.3 VR users travel further than AR users.

To test these hypotheses, we conducted an exploratory experiment in which dyads collaborated using a VR HMD and an AR HMD while solving a distributed word puzzle task adapted from Slater et al. [68] and Steed et al. [52]. We analysed questionnaire responses and behavioural measures derived from recorded interaction data. We found no support for H3.1, partial support for H3.2, and support for H3.3. Post hoc analyses further revealed relationships between presence and copresence, accord and copresence, leadership and talkativeness, and head rotation velocity and measures of leadership and talkativeness. These results provide an empirical account of how asymmetric interfaces and a static environmental representation shape interaction in asymmetric CMR. They also formed a basis for the subsequent chapters' exploration of how environmental representations can be augmented and transformed to better support coordination under asymmetry.

3.3 System Design

Our system enables asymmetric collaboration between an AR user and a VR user. It was implemented in Unity (version 2020.6.26f1) and built on Ubiq [227], an open-source social VR framework for research. Each user was embodied as a random avatar from the Ubiq avatar module. These avatars are cartoon-like, with simple shapes and only a head, upper torso, and floating hands. The floating hands have a mitten-like shape and include an animation that closes the hand. Avatar appearance was device-specific and remained unchanged throughout each trial. Avatar movement was networked using the networking module of Ubiq over a local Wi-Fi network to which both devices were connected. Both users were connected through a spatialised audio connection (stereo with attenuation) established through voice over IP (VoIP) using Ubiq.

3.3.1 Apparatus

For the AR user, we used a Microsoft HoloLens 2. Hand tracking was implemented to allow AR users to control their avatar's hands with their own hands. Mixed Reality Toolkit [228] (MRTK) was used for hand tracking and configuration of the system. For the VR user, we used a Meta Quest 2. Avatar hands were controlled with the default controllers. Interactions were controlled by Ubiq's interactions module. Both users had control over the hand-closing animation.

3.3.2 Virtual Replica of Physical Space

Dyads collaborated in a CVE containing a static 3D virtual replica of the physical space of the AR user. This 3D model was created with a fifth-generation 12.9-inch iPad Pro running iOS 15 equipped with a LiDAR sensor, using an application called *3D Scanner App* [229]. Post-processing of the model was performed in Blender [230].

3.3.3 Coordinate Space

Both users had an egocentric point of view. They were scaled to their normal height and could move independently about the CVE. They were free to choose

```
// Calculate the rotation on the y axis for alignment
var rotation = Quaternion.Euler(
    0f,
    Quaternion.FromToRotation(
        p2.transform.position - p1.transform.position,
        q2.transform.position - q1.transform.position).eulerAngles.y,
    0f);

// Rotate the recorded device points around origin to align orientation
var p1PositionR = p1Position.RotateAroundPivot(Vector3.zero, rotation);
var p2PositionR = p2Position.RotateAroundPivot(Vector3.zero, rotation);

// Find the translation between the centroids of Q1/Q2 and P1R/P2R
var translation = ((q1Position + q2Position) * 0.5f) - ((p1PositionR +
↪  p2PositionR) * 0.5f);

// Then apply rotation and translation to parent of tracked MR device
```

Listing 1 Pseudo code for alignment based on two known points, $Q1$ and $Q2$, adapted from McGill et al. [231].[2]

how to stand relative to their collaborator.

To establish a shared coordinate system between the AR user and the VR user, the CVE was aligned with the physical room of the AR user. We used a manual alignment procedure based on two known points $Q1$ and $Q2$, located on two corners of a large table, as described by McGill et al. [231] and shown in Listing 1. Upon the first launch of the system on the AR HMD, an initial calibration was required. This was completed by dragging two spheres to their known locations on the corners of the large table in the room. *Space pins* from the world-locking tools module from MRTK were set once the initial setup was completed to lock the virtual coordinate space to the real world. The calibration persisted across sessions.

3.3.4 Experimental Space

Dyads carried out the task in two separate spaces on the same floor of a university building. The VR user and the AR user were located in *Space A* and *Space B*, respectively, as shown in Figure 3.1. For the AR user, virtual posters were placed

[2]We correct a minor typographical error in the pseudocode from McGill et al. [231] concerning the rotation of the second device point.

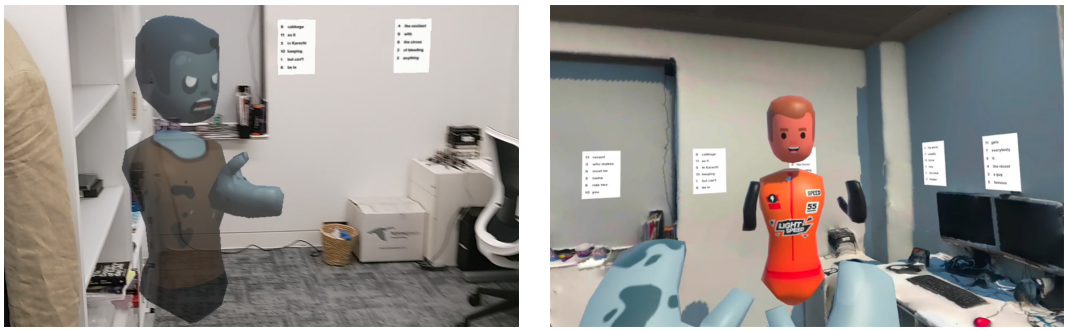


Figure 3.3 Overview of an AR user's (left) and a VR user's (right) perspectives. *Note. These images are not temporally aligned.*

along the walls. These posters were placed in corresponding locations in the 3D model for the VR user. The 3D model of the room was visible only to the VR user. Perspectives of an AR user and a VR user are shown in Figure 3.3.

3.3.5 Recordings

The record and replay module from *Ubiq-Exp* [1] was integrated with the system. This module recorded all network messages in addition to microphone audio, which enabled post hoc analysis of the trials, including extraction of additional metrics after trials were completed.

3.3.6 Interactions

Users could freely move through the space but could not interact with any virtual objects. Both users could use their avatar's hands for simple gestures. The AR user could move through the physical room by walking, whereas the VR user could move with the joystick on the left controller or could walk within the boundaries of the physical space. The VR user could not collide with any objects in the virtual room.

3.4 Method

3.4.1 Design

Our study followed a within-subjects design, where the order of the conditions was counterbalanced using a Latin square. We assessed the effect of one independent variable, the HMD type used during the task, which was a VR HMD

(Meta Quest 2) or an AR HMD (Microsoft HoloLens 2). Each participant within a dyad performed the task using each HMD by switching after the first trial period.

3.4.2 Participants

We recruited fifteen dyads, resulting in 30 participants in total. Participants were recruited through mailing lists and posters distributed across a university campus. Participants were in the range of 18–24 years old ($N = 21$) or 25–34 years old ($N = 9$), with 15 women and 15 men. Eleven of the fifteen dyads knew each other prior to the experiment. Most participants were non-native English speakers ($N = 18$). Furthermore, half of the participants played (online) video games weekly ($N = 15$), and some participants used VR ($N = 7$) or social VR ($N = 2$) weekly.

3.4.3 Task

The task design of this experiment was analogous to the work of Slater et al. [68] and Steed et al. [52] (see Figure 3.2), but with dyads instead of groups of three. The word puzzle task involved identifying and ordering same-numbered words across virtual posters hung around a (virtual) room. Each poster contained a list of five to six words, preceded by a number that indicated which of the eleven riddles it belonged to. Participants were asked to solve the riddles by finding all same-numbered words and putting them in the right order. Participants could check their solution at any time by saying it out loud, after which the experimenter told them whether it was correct. The model and posters used are available at `https://osf.io/fwru9/`.

3.4.4 Procedure

The study was led by two experimenters, the main experimenter and an assistant experimenter. Before the experiment, participants were asked to read through an information sheet and fill out a pre-study questionnaire. Each participant was assigned to one of two conditions at random for the first trial period. Participants switched conditions for the second trial period. Each period lasted 15 minutes.

Before the first trial period, each participant was led to the assigned space

and was instructed on how to use the HMD by an experimenter. Once both participants were instructed, the main experimenter was seated in front of a PC in a separate space, running a desktop client of the system. Represented as a random avatar, the main experimenter then explained the task verbally to the participants. Once the task was clear, the experimenter's avatar disappeared and the trial started.

Once the first trial period ended, both participants were assisted in removing their HMDs and were led to desks where they filled out a post-trial questionnaire on a PC. Once finished, they received new instructions and performed the task under the other condition. After the second trial period, participants filled out the same questionnaire.

After each trial, an unstructured interview was conducted to learn more about the participants' experiences and preferences. Participants received a £10 voucher as compensation.

3.4.5 Measures

3.4.5.1 Pre-trial Questionnaire

The pre-trial questionnaire consisted of two sections. The first section covered demographics, including age, gender, and experience with video games and MR. The second section contained the Interaction Anxiousness Scale (IAS) questionnaire [232]. Following the studies that we draw upon [68, 52], this questionnaire was included to obtain a measure of social anxiety, which has been theorised to influence leadership capability. We also asked participants to complete the BFI-10 questionnaire [233], which measures personality in terms of five dimensions that have been associated with leadership capability [234, 235]. However, results from the latter questionnaire were not considered in the present study, as our sample size was too small to jointly analyse the five dimensions of the scale without risking overfitting statistical models. A 5-point Likert scale was used for both questionnaires.

3.4.5.2 Post-trial Questionnaire

After each trial period, participants filled in a post-trial questionnaire. The statements of the post-trial questionnaire are shown in the Appendix (Table B.1). These statements are based on previous work that measured leadership emergence in asymmetric CMR [68, 52, 53]. We employed the following measures: talkativeness, leadership, presence, copresence, and accord (that is, *group harmony*). As most dyads had a prior relationship, we excluded two accord-related questions that were not applicable.

Given the absence of an established questionnaire for measuring presence across the spectrum of MR, we used a modified version of the presence-related questions from [52, 68]. Specifically, we prepended "virtual" to the word "office" to differentiate between the virtual and physical spaces.

To measure leadership, each participant rated their own and their partner's degree of leadership in carrying out the task on a scale from 1 to 100. On the same scale, each participant rated their own and their partner's talkativeness. Ratios were calculated based on the given scores to obtain leadership and talkativeness scores that summed to 1 within each dyad. The participant-rated leadership and talkativeness scores used in the analysis were computed as the average of each participant's self-rating and the rating received from the partner.

Presence and copresence were rated on a 7-point Likert scale, where "7" represents a high level of the respective construct. We counted the number of "6" and "7" responses to form presence and copresence scores. Accord was also rated on a 7-point Likert scale, where "7" represents a high degree of accord. Accord scores were defined as the average of responses.

3.4.5.3 Behavioural Measures

Motion and audio data were collected during the trial. For post hoc analysis, we defined behavioural measures based on the recorded data: number of words spoken, distance travelled, and median head rotation velocity. These measures provide insight into users' activity levels, which could reveal how the MR interface shaped user behaviour.

1. *Number of words spoken* was calculated by transcribing each participant's speech and counting the resulting words. This measure provided an objective indicator of talkativeness, which has been associated with leadership in prior work [68, 52].

2. *Distance travelled* was calculated based on the distance that each participant's avatar travelled within the CVE. This measure provided an objective indicator of user activity.

3. *Median head rotation velocity* was calculated based on avatar head movement driven by HMD sensors. We attempted to mitigate sensor noise by applying a moving average filter to the data (recorded at 30 Hz). Furthermore, the signal of the VR HMD contained dropped frames, which we removed by linearly interpolating between neighbouring values. An example of the filtered signals is shown in the Appendix (Figure B.1). For analysis, we took the median of the filtered signals of the pitch, roll, and yaw axes.

The optimal sliding window size to recover the true signal is unknown, and visual inspection provided limited insight. We therefore performed the statistical analysis using different window sizes to assess the sensitivity of the results to potential filter-induced bias. Our main findings are based on a window size of three frames, selected based on visual inspection of plots and recordings. We report statistics for larger window sizes (5, 10, and 15 frames) in the Appendix (Table B.2–B.5).

3.5 Results

In this section, we first report results that highlight differences across HMD types, including the results relevant to our hypotheses. Next, we present results that highlight differences across the two trial periods. Finally, we report notable relationships between measures.

As we used modified versions of existing questionnaires, we assessed reliability using Cronbach's alpha. The results showed that the copresence subscale had good reliability (0.85 for eight items), whereas the presence (0.68 for six

items) and accord (0.69 for five items) subscales had acceptable reliability for use in exploratory research [236].

Some data were excluded from the analysis. First, behavioural measures based on recorded data are missing for Trial 10 due to a system error. Second, responses to accord-related questions from one participant in Trial 6 were excluded because all responses had the value "1", and these items appeared at the end of the questionnaire.

3.5.1 Differences Across HMD Types

Paired-samples two-tailed t-tests were used to assess whether there were statistically significant differences between the means of questionnaire responses and behavioural measures when participants used an AR HMD versus a VR HMD.

Based on boxplots, outliers were detected that were more than 1.5 box lengths from the edge of the box. However, with the exception of the excluded outliers mentioned above, inspection of their values did not indicate extreme values, and they were retained in the analysis. The assumption of normality was not violated for any variables, as assessed by Shapiro–Wilk tests. The means, standard deviations, and statistical test results for each variable are shown in Table 3.1. Our key observations are as follows.

- We found *no* statistically significant differences in subjective measures between AR and VR users, including leadership and talkativeness scores.
- AR users spoke a significantly higher number of words than VR users ($p = .020$).
- VR users travelled a significantly greater distance than AR users ($p = .036$).
- AR users had a significantly higher median head rotation velocity than VR users in yaw ($p = .009$), roll ($p < .001$), and pitch ($p < .001$), indicating that AR users turned their heads more frequently or at a higher angular velocity.

Results for differences between HMD types by trial period are reported in the Appendix (Table B.7). These results indicate that the difference in median

Table 3.1 Results of paired t-tests for all variables between HMD conditions ($N = 30$). Significance levels: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

	AR		VR					
Variable	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Presence	2.47	1.85	2.37	1.81	29	0.20	.843	0.06
Copresence	3.13	2.74	3.73	2.24	29	-1.00	.327	-0.24
Accord	5.84	0.89	5.91	0.91	29	-0.50	.620	-0.08
Leadership	0.51	0.08	0.49	0.08	29	0.89	.381	0.15
Talkativeness	0.50	0.08	0.50	0.08	29	0.06	.954	0.01
Words Spoken	854.63	280.36	782.93	286.03	25	2.49	.020*	0.25
Distance Travelled	144.09	53.69	178.92	88.69	27	-2.20	.036*	-0.47
Median Yaw	0.23	0.05	0.20	0.06	27	2.83	.009**	0.55
Median Roll	0.07	0.02	0.05	0.02	27	4.77	< .001***	0.77
Median Pitch	0.08	0.02	0.06	0.02	27	5.61	< .001***	0.81

head rotation is statistically significant only in the first trial period. In addition, the mean difference in distance travelled is present in both trial periods, but is statistically significant only in the second trial period.

Notably, we did not observe a statistically significant difference between the leadership scores of AR and VR users. Following the analytic strategy used by Steed et al. [52], we examined whether social anxiety (IAS) was associated with leadership. As in their analysis, we fitted separate regression models for each trial period, device condition, and participant sex. Given that leadership scores were dyad-normalised ratios in $(0, 1)$, we used beta regression (`betareg` [237]) and report fitted relationships on the response scale.

As summarised in Table 3.2, the relationship between IAS and leadership was statistically significant only for female participants in the subgroup that used AR in period 1 and VR in period 2. In both periods, higher IAS corresponded to lower predicted leadership. Over the IAS range observed in this subgroup, the model implies an approximate 0.20 decrease in predicted leadership on the 0–1 ratio scale. We did not observe a statistically significant relationship between

IAS and leadership for male participants in any condition.

Table 3.2 Regression of leadership on IAS, reported separately by period, device, and sex. Each row reports the fitted mean-model function on the response scale, $\hat{\mu} = \sigma(\beta_0 + \beta_1 \cdot IAS)$, where $\sigma(x) = 1/(1+e^{-x})$. IAS_{min} and IAS_{max} denote the observed IAS range within each subgroup. $\Delta\hat{\mu}$ reports the model-implied change in predicted leadership across this observed range, computed as $\hat{\mu}(IAS_{\text{max}}) - \hat{\mu}(IAS_{\text{min}})$. Non-significant IAS slopes are shown in grey. Significance levels: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Period	HMD	Sex	IAS_{min}	IAS_{max}	Fitted function	$\Delta\hat{\mu}$
1	AR	Female	32	48	$\hat{\mu} = \sigma(2.125 - 0.052 \cdot IAS)$	-0.206*
		Male	27	47	$\hat{\mu} = \sigma(0.190 - 0.004 \cdot IAS)$	-0.022
	VR	Female	22	43	$\hat{\mu} = \sigma(-0.390 + 0.014 \cdot IAS)$	0.075
		Male	22	56	$\hat{\mu} = \sigma(0.130 - 0.008 \cdot IAS)$	-0.065
2	AR	Female	22	43	$\hat{\mu} = \sigma(0.034 + 0.001 \cdot IAS)$	0.007
		Male	22	56	$\hat{\mu} = \sigma(0.121 - 0.005 \cdot IAS)$	-0.039
	VR	Female	32	48	$\hat{\mu} = \sigma(1.935 - 0.051 \cdot IAS)$	-0.201***
		Male	27	47	$\hat{\mu} = \sigma(-0.031 + 0.003 \cdot IAS)$	0.014

Paired-samples two-tailed t-tests were used to assess whether there were statistically significant differences between the means of questionnaire responses and behavioural measures between the first and second trial periods.

Based on boxplots, outliers were detected that were more than 1.5 box lengths from the edge of the box. However, with the exception of the excluded outliers mentioned above, inspection of their values did not indicate extreme values, and they were retained in the analysis. The assumption of normality was not violated for any variables, as assessed by Shapiro–Wilk tests.

The means, standard deviations, and statistical test results for each variable are shown in Table 3.3. The results indicate that only the mean of median yaw head-rotation was significantly higher in the second trial period than in the first trial period ($p = .048$). Differences between trial periods by HMD type are reported in the Appendix (Table B.2 and Table B.3). These results indicate that the increase in median head rotation (for all three axes) in the second trial period was observed only for the VR condition.

Table 3.3 Results of paired t-tests for all variables between trial periods ($N = 30$). Significance levels: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

	Trial Period 1		Trial Period 2					
Variable	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	Cohen's *d*
Presence	2.03	1.47	2.80	2.06	29	-1.59	.122	-0.42
Copresence	3.87	2.52	3.00	2.45	29	1.47	.153	0.35
Accord	5.82	0.99	5.93	0.81	29	-0.78	.442	-0.13
Leadership	0.50	0.09	0.50	0.07	29	0.00	1.000	0.00
Talkativeness	0.50	0.11	0.50	0.06	29	0.00	1.000	0.00
Words Spoken	832.65	267.27	805.89	300.90	25	0.55	.585	0.09
Distance Travelled	168.61	87.30	154.41	60.44	27	0.84	.410	0.19
Median Yaw	0.20	0.06	0.22	0.05	27	-2.07	.048*	-0.42
Median Roll	0.06	0.03	0.07	0.02	27	-1.27	.213	-0.27
Median Pitch	0.07	0.03	0.08	0.02	27	-1.65	.110	-0.34

3.5.2 Relations Between Measures

In this section, we report notable relationships between variables identified in the statistical analysis. Unless otherwise stated, we used Spearman's rank-order correlations to assess these relationships. Preliminary analysis indicated monotonic relationships based on visual inspection of scatter plots. Where relevant, we report relationships by HMD type and by trial period.

Across measures, we observed significant positive relations between presence and copresence, accord and copresence, and participant-rated leadership and participant-rated talkativeness. We also observed significant relations between head rotation velocity and social metrics, including positive correlations between participant-rated leadership and median yaw (both AR and VR) and median roll (AR only), and between participant-rated talkativeness and median yaw (both AR and VR). In addition, words spoken correlated positively with head rotation velocity in the AR condition and, by trial period, in Trial 2 for median yaw and median pitch. We did not observe significant relations between dyad-level accord and task performance.

3.5.2.1 Presence and Copresence

Presence and copresence were positively correlated in both conditions, with a strong correlation in AR ($r_s(28) = .471$, $p = .009$) and VR ($r_s(28) = .685$, $p < .001$). When analysed by trial period, presence and copresence were positively correlated in the first period ($r_s(28) = .647$, $p < .001$) and the second period ($r_s(28) = .555$, $p = .001$).

3.5.2.2 Accord and Copresence

Accord and copresence were positively correlated in both conditions, with a strong correlation in AR ($r_s(27) = .603$, $p < .001$) and a moderate correlation in VR ($r_s(27) = .419$, $p = .020$). When analysed by trial period, accord and copresence were positively correlated in the first trial period ($r_s(27) = .760$, $p < .001$), but not in the second trial period ($r_s(28) = .330$, $p = .075$).

3.5.2.3 Leadership and Talkativeness

Participant-rated leadership and participant-rated talkativeness were positively correlated in both conditions (AR $r_s(28) = .764$, $p < .001$; VR $r_s(28) = .764$, $p < .001$). These correlations were equal because leadership and talkativeness scores were dyad-level ratios that were symmetric within each dyad. When analysed by trial period, leadership and talkativeness were positively correlated in the first trial period ($r_s(28) = .952$, $p < .001$) and the second trial period ($r_s(28) = .605$, $p < .001$).

Participant-rated leadership was not significantly correlated with words spoken in the VR condition ($r_s(28) = .199$, $p = .320$) or the AR condition ($r_s(28) = .374$, $p = .054$). When analysed by trial period, leadership and words spoken were positively correlated in the first trial period ($r_s(28) = .490$, $p = .011$), but not in the second trial period ($r_s(28) = .200$, $p = .308$).

3.5.2.4 Head Rotation Velocity and Leadership

As shown in Table 3.4, participant-rated leadership was positively correlated with median yaw in both conditions (AR $r_s(26)_{(3f)} = .492$, $p = .008$; VR $r_s(26)_{(3f)} = .458$, $p = .014$). In addition, leadership was positively correlated with median roll

Table 3.4 Spearman's rank-order correlations (r_s) and p-values between head rotation velocity and social metrics ($N = 26$). Significance levels: $^{*}p < .05$, $^{**}p < .01$. AR and VR labels indicate condition-specific subsets.

		Median Yaw		Median Roll		Median Pitch	
Metric	**Condition**	r_s	p	r_s	p	r_s	p
Leadership	AR	.492	.008**	.466	.012*	.305	.114
(Rated)	VR	.458	.014*	.283	.144	.291	.134
	Trial 1	.500	.007**	.496	.007**	.410	.030*
	Trial 2	.495	.007**	.300	.121	.128	.517
Talkativeness	AR	.397	.036*	.340	.077	.225	.250
(Rated)	VR	.445	.018*	.327	.090	.220	.262
	Trial 1	.503	.006**	.486	.009**	.378	.048*
	Trial 2	.284	.143	.082	.680	-.094	.634
Words Spoken	AR	.487	.011*	.398	.041*	.421	.030*
(Count)	VR	.295	.135	.168	.401	.292	.139
	Trial 1	.342	.088	.269	.183	.320	.111
	Trial 2	.515	.006**	.295	.128	.391	.041*

in the AR condition ($r_s(26)_{(3f)} = .466$, $p = .012$). When analysed by trial period, leadership was positively correlated with median yaw in Trial 1 ($r_s(26)_{(3f)} = .500$, $p = .007$) and Trial 2 ($r_s(26)_{(3f)} = .495$, $p = .007$). Trial 1 also showed positive correlations between leadership and median roll ($r_s(26)_{(3f)} = .496$, $p = .007$) and median pitch ($r_s(26)_{(3f)} = .410$, $p = .030$). The remaining correlations did not reach statistical significance (Table 3.4).

3.5.2.5 Head Rotation Velocity and Talkativeness

As shown in Table 3.4, participant-rated talkativeness was positively correlated with median yaw in both conditions (AR $r_s(26)_{(3f)} = .397$, $p = .036$; VR $r_s(26)_{(3f)} = .445$, $p = .018$). When analysed by trial period, talkativeness was positively correlated with all three axes in Trial 1 (median yaw $r_s(26)_{(3f)} = .503$, $p = .006$; median roll $r_s(26)_{(3f)} = .486$, $p = .009$; median pitch $r_s(26)_{(3f)} = .378$, $p = .048$). No correlations reached statistical significance in Trial 2 (Table 3.4).

The number of words spoken was positively correlated with all three axes in

the AR condition (median yaw $r_s(26)_{(3f)} = .487$, $p = .011$; median roll $r_s(26)_{(3f)} = .398$, $p = .041$; median pitch $r_s(26)_{(3f)} = .421$, $p = .030$). When analysed by trial period, the number of words spoken was positively correlated with median yaw ($r_s(26)_{(3f)} = .515$, $p = .006$) and median pitch ($r_s(26)_{(3f)} = .391$, $p = .041$) in Trial 2. The remaining correlations did not reach statistical significance (Table 3.4).

3.5.2.6 Number of Words Spoken and Participant-rated Talkativeness

Participant-rated talkativeness was positively correlated with words spoken in the AR condition ($r_s(26) = .385$, $p = .047$), but not in the VR condition ($r_s(26) = .179$, $p = .372$). When analysed by trial period, talkativeness and words spoken were positively correlated in the first trial period ($r_s(26) = .439$, $p = .025$), but not in the second trial period ($r_s(26) = .259$, $p = .183$).

3.5.2.7 Accord and Task Performance

Dyad-level accord was not significantly correlated with task performance (number of solved riddles) in the first trial period ($r_s(26) = -.364$, $p = .057$) or the second trial period ($r_s(28) = .236$, $p = .209$).

3.6 Discussion

This discussion returns to RQ1 and RQ2 by interpreting the results as evidence about how dyads established shared space in an asymmetric AR–VR configuration and where the combination of static environmental representation and interface differences constrained grounding and coordination. While our task context was intentionally stable and cognitively lightweight, the observed coordination practices and breakdowns expose representational boundary conditions that motivate the augmentations introduced in Chapter 4.

A useful way to frame these findings is in terms of an *asymmetry profile*. The collaboration was asymmetric in output modality (AR versus VR), input expressivity and mobility (natural walking versus joystick locomotion), and environmental context (the VR user operated within a static reconstruction of the AR user's workspace). The results below can be read as consequences of this

combination. Dyads generally succeeded under low task demands, yet they incurred recurring overhead when establishing what was mutually visible and when resolving deictic references.

3.6.1 Insights from the Statistical Analysis

We now interpret the statistical results presented in Section 3.5 in relation to our hypotheses and research questions.

3.6.1.1 Role Emergence and Participation

We did not find sufficient evidence to support H3.1, as the results in Section 3.5.1 do not show a statistically significant difference between participant-rated leadership scores for AR and VR users. This outcome may indicate that the difference in immersion between the two device types was not large enough to yield a measurable effect on leadership emergence in the present task. In particular, both collaborators wore an HMD and had comparable interaction capabilities in the context of the task. This contrasts with prior work on asymmetric collaboration in VR, where one user had an immersive VR experience while the other used a substantially less immersive desktop or handheld interface [52, 68, 238, 54].

Several factors could contribute to this outcome. First, previous work on leadership emergence under asymmetric collaboration studied groups of three users [52, 68], whereas our study examined dyads. Dyadic collaboration may require less coordination due to direct interdependence between collaborators. Future work could examine how group size moderates leadership emergence in CMR. Second, social distance within dyads may have shaped interaction [239]. Third, although social anxiety could plausibly influence leadership, the regression results in Table 3.2 are mixed and largely statistically non-significant, which limits interpretability.

We found partial support for H3.2. Participant-rated talkativeness scores did not differ significantly between AR and VR users. However, the mean number of words spoken was significantly higher for AR users than for VR users. This provides objective evidence consistent with H3.2, despite the absence of sup-

port from the subjective measure. The divergence between these measures is reflected in the fact that participant-rated talkativeness and words spoken were significantly related only in the first trial period (Section 3.5.2.6). This discrepancy raises questions about measurement choice and motivates combining subjective and behavioural measures in user studies involving complex social constructs.

Taken together, these results suggest that in this relatively lightweight task, the interface asymmetries were more visible in behavioural traces of participation (words spoken) than in participant attributions of social role. This pattern is consistent with the idea that dyads can maintain a functional division of labour without stable role differentiation, while still exhibiting subtle participation imbalances that may become more consequential as task demands increase.

3.6.1.2 Social Experience: Presence, Copresence, and Accord

As noted in Section 3.4.5.2, it remains unclear whether the questionnaire provides meaningful insight into presence and copresence for AR users. The corresponding results should therefore be interpreted with caution. Nevertheless, Table 3.1 shows no statistically significant difference in presence scores between the AR and VR conditions. We also did not observe a statistically significant difference in copresence scores between AR and VR.

Presence and copresence are conceptually distinct, yet prior work reported mixed evidence on whether they covary [240, 52]. Our results (Section 3.5.2.1) indicate a positive association, aligning with prior studies with comparable designs [68, 52, 238]. Similar to theories offered in these works, one plausible explanation in our setting is that both constructs were likely jointly supported by stable task demands, sustained mutual visibility of an embodied collaborator, and a largely coherent shared scene provided by the registered reconstruction. Under these conditions, moments of increased presence may also have increased opportunities for reciprocal attention and contingent turn-taking, which participants could have interpreted as “being together” in the shared space.

We also observed a positive association between copresence and accord in the first period (Section 3.5.2.2), consistent with prior findings [68, 52]. While we

cannot establish causal direction, one interpretation is that a stronger sense of being together coincided with a stronger sense of harmony, or vice versa. This association was not observed in the second trial period, suggesting that it may weaken as participants acclimatise to the system, the collaborator, or the task.

As participants collaborated with the same partner across both periods, we expected accord to increase over time. Table 3.3 shows a small increase in accord in the second period, but this change was not statistically significant. This pattern may be influenced by the fact that most dyads knew each other prior to the experiment.

In contrast to prior work [68, 52], we did not observe a positive association between accord and task performance (Section 3.5.2.7). Prior social relationships may again be relevant, as enjoyment of collaboration may depend less on task outcomes when collaborators already have an established relationship.

Overall, these findings tentatively suggest that in a stable low-demand environment, subjective social experience may remain relatively robust to AR–VR asymmetry, even when the interaction mechanics create measurable differences in exploration and participation. Further research is needed to clarify the roles of presence, copresence, and accord in task performance under asymmetric CMR.

3.6.1.3 Exploration Behaviour: Locomotion and Visual Search

Table 3.1 shows that VR users travelled a significantly longer distance than AR users, supporting H3.3. This result highlights a key consequence of *technological asymmetry* in input expressivity. The VR interface afforded joystick-based navigation, allowing participants to reposition more readily than AR users, who were constrained by natural walking. In addition, several participants commented that poster readability was better with the AR HMD than with the VR HMD, which may have led VR users to move more to find an appropriate reading distance. For example, a participant using AR in Trial 8 stated:

AR "I can see much better with this headset because last time I couldn't really read the posters unless I'm really close to it."

Table 3.1 also indicates that AR users tended to have higher median head

rotation velocity than VR users. This may reflect compensatory visual search. Because AR users moved less, they may have relied more on head rotation when scanning posters, with yaw expected to dominate for such behaviour. This interpretation is consistent with the larger yaw effect size in Table 3.1. Alternatively, the narrower FOV of the AR HMD may have required more head rotation to observe objects and spatial references [241].

In Section 3.5.2.4, head rotation velocity in yaw was associated with leadership score. This suggests that participants perceived as leaders may have more actively surveyed the CVE. Yaw rotation was also associated with talkativeness in the first trial period, which is consistent with the possibility that more active visual exploration coincided with greater verbal contribution. This pattern did not replicate in the second period.

Together with the locomotion difference, these results suggest that technological asymmetries shape the spatial information participants can readily acquire and relay during grounding. Joystick navigation supports rapid viewpoint changes in VR, whereas higher head rotation in AR may reflect scanning under mobility and FOV constraints.

3.6.2 Grounding Work and Breakdowns in Spatial Referencing

During the study, we observed that several dyads referred to elements of the environment to coordinate conversation and collaboration. Some dyads also began by establishing whether they saw the same virtual posters in the same locations. In doing so, dyads used elements around the office, read out words shown on the posters, or counted posters. This explicit grounding behaviour helped to establish a common frame of reference for subsequent collaboration [25]. As Müller et al. [189] noted, shared visual landmarks are critical for coordination. An example from the first period of Trial 2 is shown below.

AR “I don’t know if the posters are in the same positions, but I can see `except`.”
VR “There’s like a door, right? There’s a door and a clock right above it.”
AR “Yes, yeah, same. So it’s the same.”

On two occasions, dyads questioned whether poster positions or contents

were changing over time, yet continued after a brief pause. In other cases, verbal spatial references produced confusion, as participants were often unsure what the other participant was pointing at. In such moments, dyads required additional time to re-establish common ground. This indicates a limitation of our system [198], which supported only pointing with the avatar's hand. This limited support for *inter-referential awareness* [191] meant that participants often spent substantial time negotiating what the other was pointing at.

These breakdowns also indicate that even with a registered virtual replica, the *environmental context asymmetry* remained substantial. Because the VR user could only see a static version of the room, both collaborators faced uncertainty about what aspects of the environment were mutually available at any moment. As a result, shared space was not "given" by the reconstruction but was maintained through ongoing verification and repair.

Other obstacles to spatial referencing were the limited FOV and low brightness of the AR HMD. Future work could explore video see-through (VST) HMDs, enabling evaluation of the AR and VR conditions using the same HMD. During the second period of Trial 2, a participant commented:

VR "You're so much more visible to me this time around, whereas last time I had to focus to find you."

Prior work addressed related problems by adding gestural or gaze cues, which improved task performance and usability [193, 242, 243, 49, 244, 61, 245]. We explore such augmentations in Chapter 4.

3.6.3 Emergent Coordination Strategies and Shared-Space Management

We recorded motion data from participant avatars. Beyond objective measures such as distance travelled, we visually inspected recorded head positions.

Although objects were generally avoided, participants frequently collided with each other. At times, this caused interruptions. Two participants appeared particularly aware of their collaborator bumping into them, reflected in com-

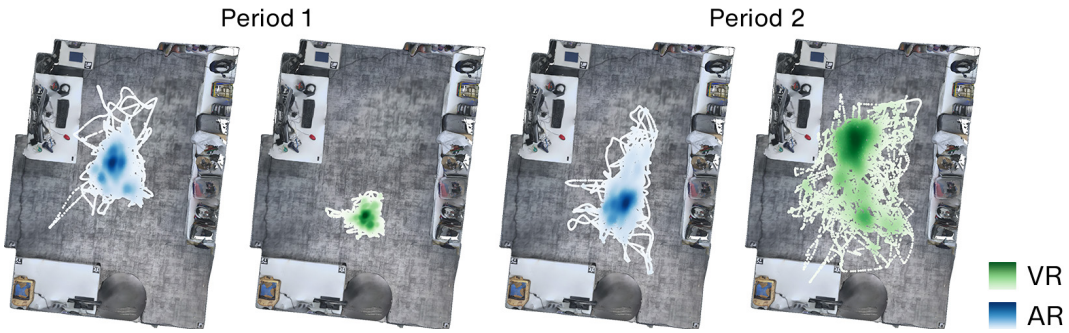


Figure 3.4 Plot of head positions for each HMD type for the two periods of Trial 14. Background is a top view of the CVE.

ments about obstruction and apologies when obstruction was noticed. This aligns with observations in prior work [68]. These comments did not appear to be specific to an HMD type.

In four trials, participants verbally agreed on a *divide and conquer* strategy, and this strategy was apparent in head position plots. Figure 3.4 shows an example, where participants primarily moved within distinct regions of the room. Overall, strategies and movement patterns did not appear to vary with trial period or HMD type.

3.6.4 Static Environmental Representation

For the VR user, the CVE contained a static 3D model of the AR user's physical space, as shown in Figure 3.1. Because participants used both HMD types in a randomised order, they observed their collaborator's perspective in the second trial period, which prompted notable reactions. For example, most participants who first performed the task in VR did not immediately recognise that they had entered the physical version of the CVE when they entered Space B during setup for the second trial period. In some cases, participants did not realise that the collaborator was located in the physical space during the first trial period.

AR "Ohhh... So this is where you were the whole time."

This observation suggests two interpretations of environmental influence during the task: 1) the representational gap between the static reconstruction and the physical space did not noticeably affect participants' sense of shared space, or 2)

participants did not actively draw on environmental references to ground their collaboration.

Participants who used the VR HMD in the second trial period also commented on the differences between the static reconstruction and the physical space. At the start of the second period of Trial 1, a participant described a sense of uncanniness after switching to VR.

VR "It's quite interesting doing it. When you are first in the real room and then in this one, which is a reproduction of that room. I feel like that changed my perspective."

VR "Cause now I'm sort of, you know, predisposed to the space already, but it's a slightly uncanny thing."

These reactions highlight that the reconstruction shapes participants' interpretation of where collaboration is taking place and what can be assumed to be shared. For the participants, the model's static nature further implies that mutual knowledge of the environment is, at best, approximate. This helps explain why participants sometimes treated poster locations and contents as potentially unstable and why they engaged in explicit verification work even when the replica was nominally registered to the room.

3.6.5 Limitations

Several limitations constrain the presented results. First, our study did not include co-located or symmetric conditions, which could enable additional comparisons [173]. Second, device type was confounded with locomotion. The VR user could translate with a joystick in addition to physical walking, whereas the AR user relied on natural walking. This asymmetry reflects a common deployment constraint, as collaborators often have mismatched navigable spaces and redirected walking remains uncommon in practice [224, 246]. However, it likely contributed to the observed distance differences and may have affected presence, given prior evidence that natural walking can yield higher presence than controller-based "flying" [247]. Future work could decouple display modality from locomotion by matching locomotion capabilities across collaborators.

Third, as discussed in Section 3.4.5.2, the reliability of the presence and copresence scales remains uncertain, given that these scales were developed for VR and VR differs conceptually from AR. Although questionnaires for presence in MR have been proposed [248, 249], measuring presence and copresence in MR remains an open problem [144]. Fourth, participants had verbal contact with the experimenter during the experiment, which could lead to breaks in presence [135]. Future work could implement alternative mechanisms for communicating solution correctness. Fifth, switching between conditions and spaces may have influenced outcomes in the second trial period. For this reason, we reported and discussed period-specific results where relevant. Finally, participants were represented by random cartoon-like avatars, which could have influenced interaction [250]. Future work could examine how avatar properties, including gender and fidelity, shape collaboration under asymmetry.

3.7 Conclusion

This chapter contributed empirical grounding for RQ1 and RQ2 by examining how dyads enacted shared space in an asymmetric AR–VR configuration in which a remote VR user collaborated from within a static reconstruction of the AR user's workspace. We reported an exploratory study and analysed questionnaire responses alongside behavioural measures derived from recorded interaction data to characterise how technological and environmental context asymmetries related to coordination and participation.

Consistent with prior work that linked immersion and interface differences to leadership emergence [52, 68, 238, 53], we tested whether AR users would obtain higher leadership and talkativeness ratings. Questionnaire responses did not indicate a statistically significant difference in participant-rated leadership or talkativeness between AR and VR. However, AR users spoke more words than VR users, suggesting that behavioural traces can diverge from participant ratings in this setting. In addition, VR users travelled further, consistent with the availability of joystick locomotion as an asymmetric input modality. AR users exhibited higher head rotation velocity overall, with period-specific analyses indicating

that this difference was concentrated in the first trial period.

Post hoc analyses suggested several associations between subjective social measures and behavioural activity. Within the limits of the questionnaire measures, presence and copresence covaried in both conditions and both trial periods, and accord and copresence covaried in the first trial period. Leadership and talkativeness were strongly related by construction of the dyad-normalised ratings, and head rotation velocity was positively associated with leadership and talkativeness in several exploratory analyses. Given the number of relationships tested, these correlations should be interpreted as hypothesis-generating rather than confirmatory. Together, the results motivate the use of behavioural measures as complementary evidence, particularly where alignment between participant ratings and behavioural counts varies by condition and trial period.

More broadly, the findings highlight limitations of static environmental representations and minimal referential support for sustaining mutual understanding under asymmetry. Dyads frequently engaged in explicit verification of shared landmarks and experienced ambiguity during deictic reference, indicating that shared space was established through ongoing grounding work rather than delivered by the reconstruction alone. These observations motivate the subsequent chapter's focus on task-driven augmentation of environmental representations in more dynamic contexts, where representational staleness and mobility constraints become more consequential for coordination.

Chapter 4

CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration

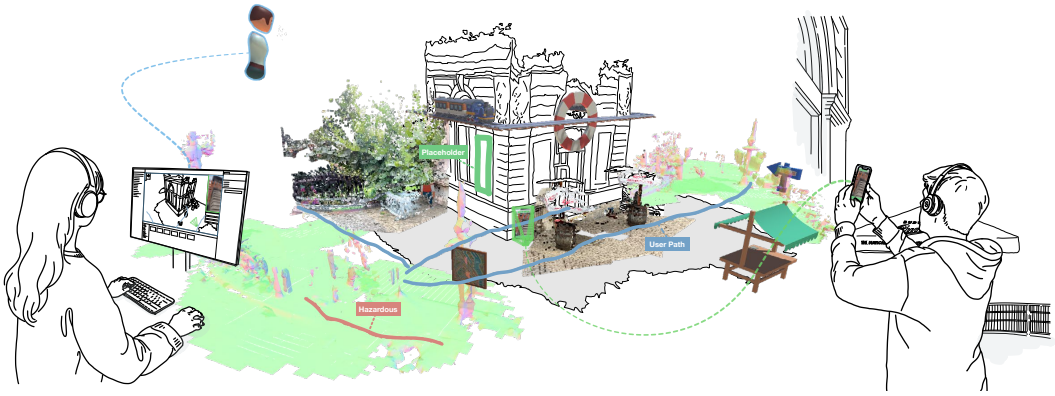


Figure 4.1 CoCreatAR supports collaboration between ex-situ developers (left) and in-situ users (right) during the development of site-specific AR experiences. During design iteration and user experience refinement, the in-situ user can synchronously share critical real-world context with a remote partner. This includes audio, video, meshes, RGB-D point clouds, annotations, and feedback on usability and performance. The additional anchored context allows the ex-situ developer to refine the prototype using familiar tools, without requiring repeated visits to the target site.

This chapter develops the dissertation by operationalising RQ2 and RQ3 in a context where environmental change and partial access are common. We examine asymmetric collaboration during site-specific outdoor AR experience authoring, where representational incompleteness and contextual variation shape what collaborators can treat as mutually available. This setting makes the limitations of static or weakly contextualised environmental representations consequential

for coordination, reference, and verification, which provides a direct investigation of RQ2. We then investigate RQ3 by introducing task-driven augmentations, including explicit spatial cues and implicit awareness cues, and by evaluating how these cues support grounding and authoring decisions during synchronous collaboration. A paper based on this work was published at ACM CHI 2025 [15][1].

4.1 Context and Motivation

Outdoor AR experiences are often designed for specific locations, spanning a range of settings from controlled environments such as theme parks [251] to dynamic public urban spaces [252]. Developing these experiences is inherently complex because the user experience is closely intertwined with the physical environment. However, developers often work remotely, with only intermittent access to the target location. This chapter therefore uses site-specific AR authoring to examine when environmental representations fail to support mutual understanding under asymmetry (RQ2), and how task-driven augmentation can restore referential and awareness resources through collaboration (RQ3).

Ex-situ (i.e., *off-site*, *remote*) development workflows typically rely on pre-captured 3D models of the environment obtained from prior scans or geospatial data to design and position AR content. Tools such as *Google Geospatial Creator* [253] and *Niantic Remote Content Authoring* [254] facilitate this process by providing site-specific environmental representations at a large scale. Additionally, these tools provide a visual positioning system (VPS) that can align virtual content with the real-world target environment. While these tools effectively support development, their reliance on static representations often results in missing critical contextual information.

As discussed in Section 2.1.1, context is relevant both as system-representable

[1]The majority of this research was performed during an internship at Niantic Labs. I was responsible for the system design and implementation, experimental design, data collection, analysis, and writing. My mentors at Niantic Labs (Jessica Van Brummelen and Gabriel Brostow) and supervisors at UCL (Anthony Steed and Simon Julier) provided guidance and feedback throughout the project, with Jessica Van Brummelen supporting the thematic analysis of the formative study. Suhyun Park assisted with data collection, thematic coding of video recordings, and the creation of illustrations.

information and as an interactional accomplishment [75, 48]. As such, context is particularly important when creating experiences for outdoor settings. For instance, dynamic elements such as lighting conditions, moving objects (e.g., vehicles, temporary structures), and pedestrian activity are typically absent from pre-captured 3D models. In practice, these dynamics can reduce workspace awareness and increase grounding work because developers must continually re-establish what is currently true at the site and which features are task-relevant for the ongoing activity (Section 2.1.3). This absence can lead to inconsistencies between virtual elements and the real environment, which can ultimately degrade the user experience [255]. Consequently, developers frequently resort to repeated on-site visits, which can be costly, time-consuming, and logistically challenging.

In-situ (i.e., *on-site*) authoring tools, in contrast, enable AR content creation and testing directly within the target environment, providing immediate environmental context awareness [256, 257, 258]. However, these tools often lack the flexibility and expressive power of remote development environments, particularly for large-scale outdoor experiences that require precise object placement, complex interactions, or considerations of user flow and safety. Moreover, mobile devices used for in-situ authoring often have limited computational resources and interaction capabilities, which makes it difficult to manage complex assets or scripting tasks [259, 260, 261]. This mismatch creates a practical asymmetry between authoring contexts, in which in-situ creators have privileged access to situated environmental cues, while ex-situ developers have privileged access to expressive authoring tools and computational resources.

To better understand the challenges of ex-situ authoring of site-specific outdoor AR experiences, we conducted a formative study comprising a survey (n = 25) and follow-up interviews (n = 5) with industry professionals. Our findings highlighted common issues, including missing or outdated environmental representations, time-consuming iterative testing cycles that necessitate site visits, and challenges in capturing and communicating contextual information among

team members.

Building on these insights, we explored collaborative approaches that integrate the advantages of both ex-situ and in-situ authoring. By enabling synchronous collaboration between ex-situ developers and in-situ collaborators, we sought to address a set of major challenges we identified. Inspired by the concept of pair programming [262], we structured collaborator roles so that the in-situ user captures and transmits real-time environmental context while the ex-situ developer remotely utilises advanced authoring tools.

To support this workflow, we developed CoCreatAR, an asymmetric collaborative AR authoring system that facilitates real-time editing of site-specific AR content. CoCreatAR enables ex-situ developers to work synchronously with in-situ users by providing enhanced spatial information, anchored annotations, and communication tools to bridge the gap between remote development and the physical target environment.

We conducted a user study (n = 32, in pairs) to compare synchronous (CoCreatAR) and asynchronous collaborative authoring approaches in terms of task load, engagement, and confidence in the authored result. In the synchronous condition, ex-situ and in-situ collaborators used CoCreatAR to work together in real time. In the asynchronous condition, development proceeded sequentially without immediate interaction, reflecting current practices identified in our formative study. Our findings suggest that authoring with CoCreatAR improves the integration of real-world context, increases developers' confidence in the accuracy and feasibility of their AR designs, and leads to greater engagement and creativity among team members.

4.2 Related Work: AR Experience Authoring

While Chapter 2 provides a comprehensive review of foundational concepts and prior work relevant to this thesis, this section adds material on related work concerning AR experience authoring tools and related collaborative approaches.

AR authoring tools span a wide spectrum of fidelity levels, supporting different development stages and levels of expertise. As categorised by Nebeling and

Speicher [263], these range from low-fidelity prototyping tools that require little to no technical expertise [257, 264, 265] to high-fidelity authoring tools, such as Unity [266] and Unreal Engine [267], which are commonly used for developing production-ready applications.

Despite their utility, existing AR authoring tools face challenges that limit their effectiveness. Several surveys and studies have identified key issues, including difficulties in spatial navigation, adapting to changing environmental contexts, and integrating context-specific considerations into AR experiences [268, 269, 270, 255]. A key limitation, as noted by Nebeling [255], is the insufficient incorporation of real-world context during authoring, which can disrupt user experience and immersion. To incorporate this context, authoring tools generally adopt one of two approaches: *in-situ* authoring, which supports real-time creation in the target environment, and *ex-situ* authoring, which relies on proxy representations or recordings of the target environment.

4.2.1 In-Situ Authoring

Tools that incorporate in-situ interfaces, such as the mobile applications of Adobe Aero [257], Reality Composer [271], and Unity Mars [258], as well as other prior work [272, 256], build on the *WYXIWYG (What You Experience Is What You Get)* paradigm introduced by Lee et al. [273]. This paradigm emphasises the benefits of allowing users to experience and verify authored content in situ, such as immediate testing and adjustment.

In a user study, Lee and Kim [259] found that their in-situ system supported efficient and precise arrangement of virtual content in real-world settings. However, they noted a key limitation: abstract tasks, such as programming behaviours, were better supported by traditional ex-situ desktop environments. Langlotz et al. [256] developed a mobile AR authoring system that focused on enabling spontaneous content creation in unprepared environments, including outdoor settings. While highlighting the low barriers to entry that in-situ authoring provides, the authors also emphasised the importance of integrating in-situ authoring with ex-situ desktop-based interfaces for refinement. Vargas González et al. [260] com-

pared in-situ AR and ex-situ desktop authoring tools in scenario-based training contexts and similarly found that, while both approaches offered comparable usability and task completion times, desktop tools were perceived as more efficient, particularly for tasks requiring global context.

4.2.2 Ex-Situ Authoring

Ex-situ authoring tools, such as *DART* [274], *ScalAR* [275], *DistanciAR* [276], and *Corsican Twin* [277], offer broader authoring capabilities than in-situ tools by providing 3D scene editors, scripting environments, flexible testing workflows, and asset integration. These tools represent environmental context using pre-captured data, including 3D models [277, 275, 278, 254, 253], video [274, 279], sensor data [274], and 360° footage [264]. For example, *DART* [274] supported recording and playback during authoring by synchronising real-time video and sensor data capture, while *DistanciAR* [276] supported in-situ 3D reconstruction for capturing environmental context and enabled ex-situ authoring and testing through different model visualisation modes. However, a key limitation of these approaches is their reliance on static environmental representations. These models capture only a fixed moment in time and do not account for dynamic changes. As a result, authoring errors and spatial misalignment can occur.

Recent ex-situ authoring tools for outdoor AR experiences, such as *Niantic Remote Content Authoring* [254] and *Google Geospatial Creator* [253], adopt similar approaches by providing 3D models of large outdoor sites. However, outdoor environments present additional challenges due to a higher degree of dynamic change. Unlike indoor spaces, outdoor settings are continuously affected by shifting lighting, weather, seasons, and moving elements such as people and vehicles. These dynamic factors impact the accuracy and relevance of pre-captured data, making static environmental representations particularly problematic for authoring outdoor AR content. While in-situ testing can mitigate some challenges, it is often constrained by logistical factors such as time, cost, and accessibility. Moreover, in-situ authoring of outdoor AR experiences is complicated by the scale and complexity of outdoor spaces [280, 4, 281], which

makes in-situ adjustments more difficult to execute.

Overall, our analysis indicates that current AR authoring tools remain limited in bridging the gap between ex-situ flexibility and in-situ contextual awareness. Ex-situ tools rely on static proxies that fail to capture the dynamic nature of outdoor spaces, while in-situ tools, despite offering real-time context, lack the expressiveness and flexibility required for comprehensive authoring, as we found in our formative study (Section 4.3). This disconnect forces authors to switch between ex-situ and in-situ development, which leads to inefficiencies and challenges in maintaining contextual accuracy. To address this gap, we propose CoCreatAR, a system that leverages the flexibility of ex-situ authoring while incorporating real-time environmental awareness from in-situ users. By enabling a synchronous collaborative workflow between ex-situ and in-situ roles, our approach supports contextually aware authoring without compromising the richness and advanced capabilities of ex-situ tools.

4.2.3 Collaborative Approaches to AR Experience Authoring

Given the complex and multidisciplinary nature of AR experience development, collaboration is integral to the authoring process, as noted by Krauß et al. [268]. AR application development typically involves multiple roles, such as interaction designers, content creators, and developers, each contributing distinct expertise. It can also involve non-technical collaborators such as clients or end-users, who can help shape the design and content [270]. Motivated by role-based collaboration, Nebeling et al. [57] introduced *XRDirector*, a system that enables users to manipulate virtual objects from different subjective viewpoints in AR and VR. Designed for indoor immersive storytelling, *XRDirector* enabled multiple contributors to author experiences from perspectives aligned with their respective roles.

Beyond role-based collaboration, prior work has examined different collaboration modes for authoring AR experiences. Guo et al. [282] explored synchronous and asynchronous collaboration in AR authoring through their system, *Blocks*. Their findings indicated that synchronous collaboration produced higher

engagement, particularly when participants collaboratively created shared AR structures in real time. Asynchronous collaboration offered greater flexibility, while it introduced challenges related to maintaining awareness of others' contributions.

While prior systems demonstrate different approaches to collaborative authoring, they do not adequately address the challenges of site-specific AR development, particularly for outdoor environments where ex-situ authors must take in-situ perspectives into account. In designing CoCreatAR, we sought to support this workflow by facilitating collaboration between ex-situ developers and in-situ collaborators without requiring specialised technical skills beyond familiarity with handheld AR. Our system leverages subjective views, similar to *XRDirector*, while it applies them to gathering in-situ context and feedback in outdoor, site-specific authoring. Additionally, we compare synchronous and asynchronous workflows, as examined in *Blocks*, including specific hypotheses on engagement and task load (Section 4.5).

4.3 Formative Study

To inform the design of CoCreatAR, we conducted formative surveys and interviews with experienced site-specific AR developers, artists, designers, and testers. Inspired by prior investigations of AR development workflows [283, 269, 268], we sought to characterise current workflows and key challenges in outdoor site-specific AR experience development.

4.3.1 Method & Participants

We recruited 25 survey respondents through internal mailing lists, AR development social media channels, and targeted emails to professional outdoor AR agencies. The survey included quantitative and qualitative questions about the projects respondents had worked on, the tools they used, and their typical workflows. One survey question prompted respondents to list issues encountered during in-situ testing. To provide an initial set of response options, we defined issue types through a brainstorming session and included an open-ended question

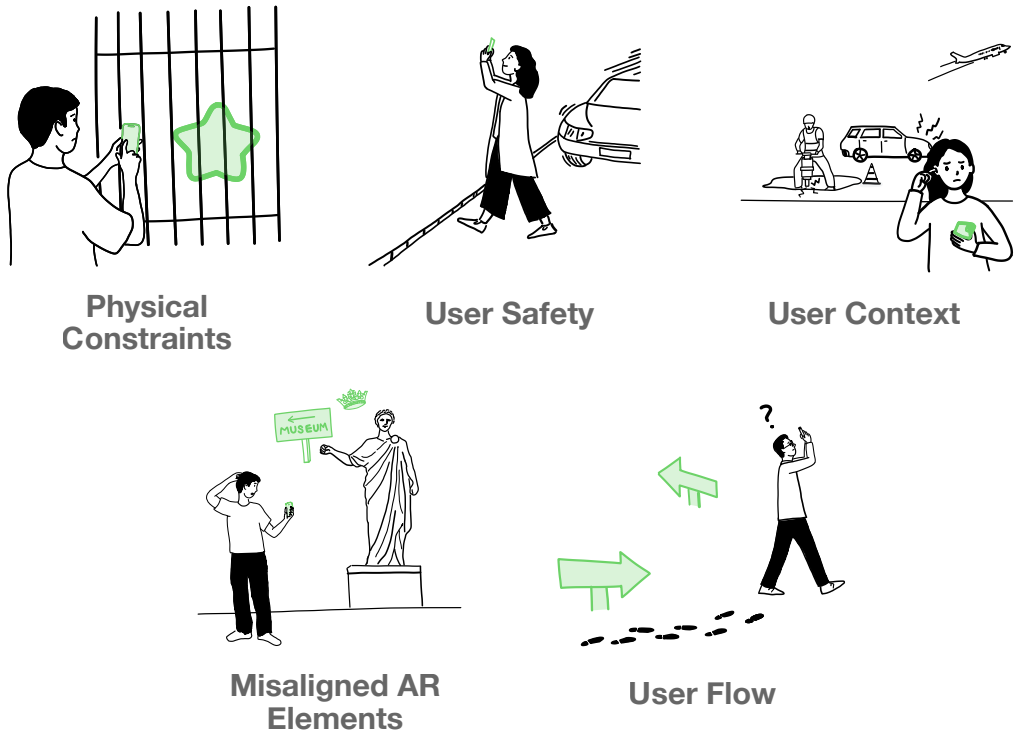


Figure 4.2 The five main site-specific AR issue types we aimed to address with CoCreatAR. We refer to Section 4.3 for further details on the full range of issues identified in our formative study.

for respondents to report additional issues they had encountered.

The resulting issue types were as follows: (A) Physical constraints (e.g., blocked paths and restricted areas), (B) User safety (e.g., hazardous environments), (C) Misaligned AR elements (e.g., misplaced anchors, occlusion, or perspective issues), (D) User context (e.g., environmental noise or lighting conditions), (E) Hardware performance (e.g., device processing and rendering power limitations), (F) Registration issues (e.g., localisation and tracking errors), (G) User flow (e.g., user position or perspective affecting AR interaction), (H) Socio-cultural appropriateness of the area or experience, and (I) Semantic interaction (e.g., whether the system correctly recognises objects in the world). See Figure 4.2 for visual illustrations of these issues.

Five of the survey respondents agreed to complete an interview, in which we asked semi-structured follow-up questions to obtain further detail. The author of this dissertation and a collaborator conducted a reflexive thematic

analysis to analyse the qualitative results. After developing initial codes, the authors iteratively refined the codebook while reviewing transcripts, following the methods described by Braun et al. [284]. Each interview participant received a gift card valued at £25, and survey respondents had a 1-in-10 chance of winning a gift card valued at £30. The survey and interview materials are available in the Appendix (Section C.5).

4.3.2 Findings

We structure our findings by first outlining typical workflows, followed by a review of major hurdles, and concluding with a discussion of future tools that participants considered useful. These findings inform our understanding of how practitioners develop site-specific AR, support us in establishing requirements for our proposed system (Section 4.4), and provide a baseline for our study (Section 4.5).

4.3.2.1 Workflow walkthrough and tools

One of the first steps participants took was to **acquire context about the target location** for their AR experience. This typically involved downloading publicly available pre-captured meshes or visiting the site to scan the location themselves. Participants primarily used *Niantic Geospatial Browser* [285] (64.0%) or *Google Geospatial Creator* [253] (28.0%) to obtain pre-captured meshes, which provided spatial information about the location to support the design and development process. We refer to these pre-captured meshes as *location meshes* throughout the rest of this dissertation. These representations helped participants place and carefully align virtual content with real-world structures and surfaces. Later, when at the site, end-users could use a VPS to localise and interact with the experience.

Following this, participants typically **worked ex situ to design and build the experience** based on the *location mesh*. For development, they most often used *Unity* (96%) and *Blender* (76%). Additionally, they employed AR frameworks to support features such as plane detection, occlusion meshes, and semantic

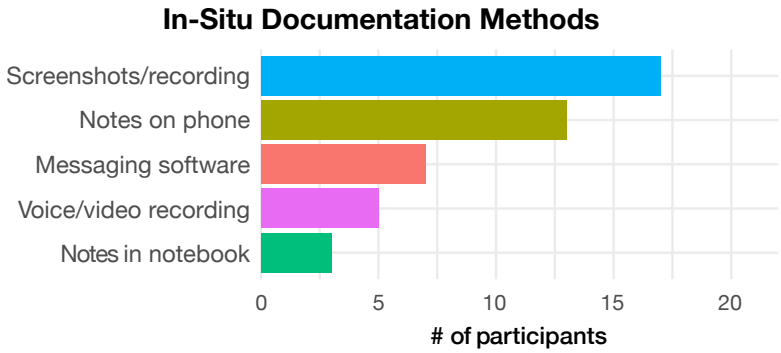


Figure 4.3 The number of participants who reported using various methods to document issues in their AR experiences when in situ.

segmentation. The most commonly used framework was *Niantic SDK for Unity* (80.0%), followed by *ARFoundation* (76.0%). A summary of the tools participants used is available in the Appendix (Section C.1).

Since ex-situ tools could not provide full contextual awareness of the location (e.g., dynamic objects, crowding, and lighting conditions), **participants or their team members often travelled back and forth** between their ex-situ development location and the target environment. When in situ, participants logged information about issues with their AR experience to address them later or relay them to their team. On average, they reported spending 36% (SD = 24.7) of their time testing the experience in situ and working with between 0 and 12 team members (M = 4.16; SD = 2.97). Notably, one interview participant (P4) considered bringing their "computer with [their] phone" to the site to address alignment issues in situ, yet noted that they "do not think [they] should" for practical reasons such as weather and crowding, which other participants also cited as constraints (P1, P2, P5).

Among the issues described in Section 4.3.1, *Physical Constraints*, *User Safety*, and *Misaligned AR Elements* were the most commonly reported[2] (see Figure 4.4). To record these issues in situ, participants typically took screenshots, screen recordings, or typed notes (see Figure 4.3). Some participants also received recordings or notes from team members who visited the site. After initial context

[2]Only one participant described an additional issue type in the survey, "Game Flow," which we categorise under "User Flow."

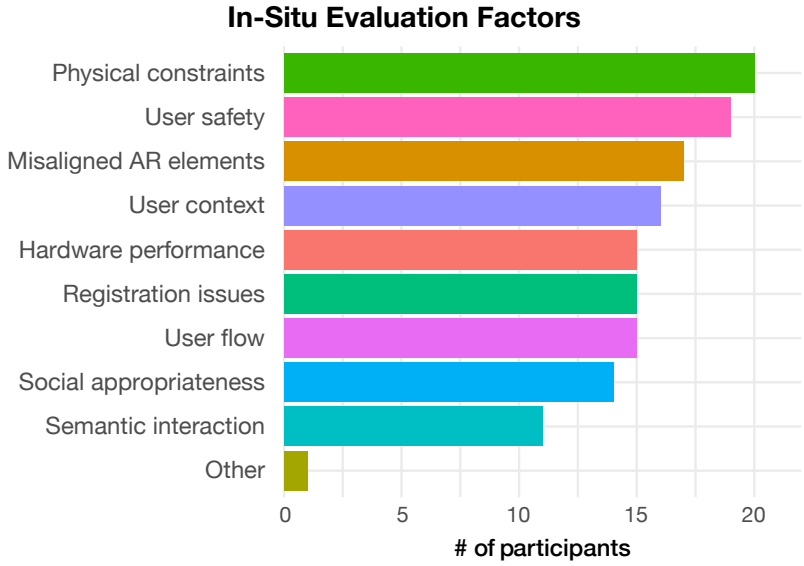


Figure 4.4 The number of participants who reported looking for various issues in their AR experiences when in situ. See Section 4.3.1 for descriptions of the issues.

gathering, participants or their team members continued iterating on their AR experiences, travelling between locations as needed until completion.

4.3.2.2 Challenges with context capture

In the interviews, all five participants mentioned **difficulties with pre-captured meshes**, whether obtained from geospatial browsers or personally scanned. For instance, P2 described pre-captured *location meshes* as "old" and not up to date with the current state of the real world. P4 noted that *location meshes* were "always a challenge because there are mesh holes everywhere [...] so the geometry falls through." Similarly, P5 stated that "the maps are very incomplete," while noting that the alternative of scanning the location manually was "quite labour intensive."

Participants also mentioned **difficulties with scanning**. P1 described how "often the test scans do not come out great" and how selecting alternative target locations was sometimes necessary due to poor scan quality. P5 noted that scan processing times could be lengthy, which reduced iteration efficiency. Two participants also commented on the difficulty of the scanning process due to social awkwardness (P3) or poor weather conditions (P5).

Poor-quality meshes contributed to **alignment issues** between virtual ob-

jects and the real world. This occurred because degraded meshes reduced localisation accuracy, and because poor mesh geometry hindered precise placement when working ex situ. For example, P5 described this limitation: "It's like you try to figure out where [the virtual object] has to be [ex-situ], and then every time you go [in-situ], your viewpoint is a bit different because you are looking at it from street level or it's slightly tilted [...] we always ended up having to reposition the thing compared to the initial [placement]." All participants mentioned alignment issues as a recurring problem.

4.3.2.3 Missing contextual information

All five participants mentioned **challenges in acquiring contextual information missing from meshes**. For example, P3 and P4 described how locations found online might not actually be accessible, or how previously accessible locations might become restricted over time. Three participants mentioned that some locations were hard to conceptualise ex situ without visiting in person, either because they were too complex or large (P1), or too cramped (P2, P3).

P1, P2, and P4 mentioned challenges related to **lighting conditions**, particularly at different times of day. This concern arose both from an end-user experience perspective (e.g., P1: "You do not want someone staring into the sun") and in terms of VPS localisation accuracy (e.g., P4: "we generally test throughout the day to try to make sure that it is going to work under any lighting"). P1 also highlighted "audio contamination" as a factor that could affect the experience, but which could not be discovered without being at the target location. Additionally, P1, P2, and P5 mentioned adverse weather conditions affecting in-situ testing.

One crucial piece of contextual information that all five participants mentioned gathering in situ was **the area's safety**. This included assessing foot and vehicle traffic, as well as the potential for theft. For example, P3 described how even though "[there] might be plenty of areas where things are walkable [... this still] means very heavy foot traffic, [so] you still have to be careful." P1 pointed out that if users are "getting jostled, or [using] a very expensive phone, then

someone could just snatch it."

All participants mentioned the need to assess **the level of crowding at a location**, as high foot traffic could impede both the user experience and VPS localisation. P2 described changing the target site due to crowding. **Moving objects** also posed a challenge for both the experience and localisation. As objects at the location changed, such as cars, signs, or even "posters" (P1), VPSs struggled to match the environment with the *location meshes*.

4.3.2.4 User experience challenges

All five participants reported **difficulties identifying end-user experience challenges**, particularly when developing ex-situ. P1 and P2, for example, emphasised the importance of understanding the "human scale" to ensure that objects were appropriately reachable and scaled, which was difficult to achieve without being in situ. Participants also noted visibility challenges. P1, P2, and P3 described how "tight FOVs," long distances, or sun direction could affect visibility. Gameplay considerations also required in-situ testing; P2 stated that users should not have to run during the experience, while P1 mentioned ensuring that gameplay events were not "too close" or "too far away."

4.3.2.5 Workflow challenges

Due to these difficulties, four of the five participants described **travelling to target locations as a significant challenge**. Participants characterised these trips as "inefficient" (P1), "a lot of effort" (P3), and "costly" (P5). These trips could also be extensive, with participants travelling to different cities (P2, P3) and even different countries (P5) for testing.

Other workflow challenges included collecting and sharing feedback from in-situ testing. P1 described difficulties in physically taking notes while testing the experience, explaining issues over audio, and annotating video recordings. P2 and P5 described communication challenges when working with team members from different backgrounds. P2 also mentioned "hassles" related to uploading, downloading, and sharing information while in situ.

4.3.2.6 Desired features for future tools

As participants described these challenges, they also suggested potential solutions. To address poor context capture, P2 and P4 proposed **methods to tag areas with semantic or other labels**. Others suggested **collaborative approaches to context gathering**. For instance, P1, P4, and P5 suggested streamlined methods for sharing and viewing *location meshes* with team members. P4 proposed a way to combine and edit meshes captured by multiple team members to improve accuracy. Additionally, P2 and P4 expressed interest in a system that allowed team members to collaborate in situ and ex situ via live streaming of location data and voice communication.

Participants also expressed interest in tools that improve the authoring process. Four of the five participants expressed interest in **editing objects in situ**. P2 and P4 suggested **procedural placement of objects based on semantic data**. P3, P4, and P5 wanted better methods for simulating the site ex situ, with P5 stating, "The more you can do without actually being on the spot, the better."



Figure 4.5 In-situ user interface of CoCreatAR. (A) Users can scan the environment to obtain a *Coarse Mesh* of the surroundings; (B) Users can tap Capture to take a *3D Snapshot*; (C) Users can tap Capture or Annotate to access feature sub-menus from the main menu, and tap and hold on the screen to spawn a *3D Cursor*; (D) Users can create drawings projected onto surfaces with *Surface Draw*; (E) Users can create a trajectory or a 3D drawing with *Air Draw* by moving their smartphone.

4.4 CoCreatAR System

Based on our formative study, we developed CoCreatAR, a collaborative authoring system for outdoor AR experiences designed to facilitate real-time interaction between ex-situ developers and in-situ collaborators. The system enables ex-situ creators, who typically design and develop the experience asynchronously in Unity [266], to collaborate synchronously with in-situ users who experience the AR content directly in the field. By integrating real-time communication, contextual reference tools, and spatial data capture, CoCreatAR aims to reduce the need for repeated on-site visits during iterative design of site-specific AR experiences. While CoCreatAR supports multiple ex-situ and in-situ users, in this section and our user study we focus on a two-user scenario in which a single ex-situ user collaborates with one in-situ partner.

CoCreatAR primarily utilises two frameworks: the *Niantic SDK for Unity* [254], identified in the formative study as the most widely used outdoor AR development kit, and *Ubiq* [227], an open-source library for CMR. The *Niantic SDK for Unity* provides VPS localisation, depth estimation, meshing, and occlusion handling, whereas *Ubiq* facilitates real-time networking, avatar representation, and peer-to-peer communication, including WebRTC-based audio and video streaming.

4.4.1 System Requirements

Based on challenges reported by participants in our formative study (Section 4.3) and previous work, we identified five requirements for the development of CoCreatAR.

R1: Reduce the need for frequent on-site visits in the development and testing iteration loop Our formative study revealed that participants required frequent on-site visits to gather real-world context and feedback. CoCreatAR aims to reduce travel burden by allowing ex-situ users to collaborate synchronously with in-situ partners, enabling remote adjustments based on immediate feed-



Figure 4.6 Ex-situ user interface of CoCreatAR. (A) All objects under `NetworkedScene` are automatically synchronised between ex-situ and in-situ users; (B) *3D Snapshots* captured by the in-situ user; (C) The *location mesh* of *Location A*; (D) *Coarse Mesh* captured by the in-situ user; (E) Live feed of the in-situ user's screen, including AR content; (F) The *3D Cursor* of the ex-situ user, projected into world space; (G) Close-up of the *3D Cursor* of the ex-situ user as seen in the scene view; (H) List of annotations and spatial captures, persistently saved in the scene for later review; (I) Sample assets that can be added to the scene at runtime.

back and live validation.

R2: Support incorporation of additional or updated real-world spatial information. Interviewees highlighted limitations of static meshes, which are often outdated or incomplete and can lead to alignment issues. CoCreatAR aims to address this limitation by enabling in-situ users to capture and relay real-time updates about the site, such as changes in the environment.

R3: Facilitate the capture of real-world context beyond spatial data. Participants emphasised the importance of capturing additional real-world context, including lighting conditions, pedestrian flow, and safety concerns. CoCreatAR incorporates tools that enable in-situ users to capture this contextual information through annotations, real-time audio, snapshots, and a live video feed.

R4: Enable live persistent scene adjustments and additions during testing and prototyping. Participants described returning to their workplace to make adjustments after in-situ testing as challenging. To mitigate this challenge, CoCreatAR supports live modifications during testing sessions, allowing developers to make immediate adjustments based on in-situ feedback.

R5: Integrate captured data into existing development workflows. Participants reported difficulties incorporating in-situ feedback into the virtual scene representations within their development tools, particularly in Unity. CoCreatAR addresses this issue by integrating captured spatial data and annotations directly into Unity and anchoring them to the *location mesh*, enabling developers to incorporate feedback without manual import procedures.

4.4.2 Scene View and Localisation

CoCreatAR enables both users to view the same AR content from their respective familiar perspectives. In-situ users can interact with the AR experience and a CoCreatAR UI layer within their real-world environment via a smartphone (Figure 4.5), whereas ex-situ users observe the scene remotely through Unity's scene editor (Figure 4.6).

At launch, in-situ users complete a step-by-step VPS-based localisation process to align the AR experience with their surroundings. The system continuously localises, prompting the in-situ user to confirm alignment by overlaying a semi-transparent pre-captured *location mesh* onto the real world. The in-situ user can adjust its transparency to verify overlap before finalising localisation. Once confirmed, the full scene, including the in-situ user's avatar, loads into the ex-situ user's scene view. The ex-situ user can similarly adjust the *location mesh* transparency to better view AR elements, spatial captures, and annotations throughout the collaboration process (Figure 4.6J). If needed, localisation can be restarted via the settings panel to correct drift that may occur during collaboration.

4.4.3 Real-Time In-Situ View Streaming and Audio Communication

CoCreatAR supports real-time peer-to-peer video and audio communication to facilitate collaboration between users. A peer-to-peer audio channel enables verbal communication. Additionally, a live feed of the in-situ user's screen is streamed to the ex-situ user, who can view this feed in a floating panel within the Unity scene view (Figure 4.6E). This feed presents both the real-world camera view and the rendered AR elements, enabling the ex-situ user to understand the interaction between AR content and the physical environment.

Ex-situ users can also capture screenshots of the in-situ user's screen, which are saved and displayed in the Unity scene and anchored to the *location mesh* (Figure 4.6K). These captures align with the in-situ user's perspective and can mark specific moments, errors, or areas of interest for later review.

4.4.4 Spatial References with 3D Cursors

To support spatial referencing during the collaborative process, CoCreatAR incorporates networked *3D Cursors* that each user can control. These cursors are projected into shared world space based on screen inputs. In-situ users can tap and hold anywhere on their screen to spawn and control a cursor in the 3D environment, while ex-situ users can hover over the live in-situ view panel within Unity to achieve the same. Cursors appear as semi-transparent spheres with axis indicators, supporting spatial referencing. Green represents in-situ cursors, while blue represents ex-situ cursors (see Figure 4.6F).

Both users can create persistent cursor markers: ex-situ users by clicking within the Unity editor and in-situ users by double-tapping on their screen. This functionality supports alignment tasks and spatial referencing during collaboration.

4.4.5 Auto-Networked Objects and Scene Updates

A core feature of CoCreatAR is the automatic synchronisation of all scene changes (Figure 4.6A). Any modifications made by the ex-situ user in Unity, such

as adjusting the position, rotation, or properties of AR objects, are immediately reflected in the in-situ user's view. This functionality extends Ubiq's networking capabilities to transmit network messages whenever changes in object transforms, materials, or user-controllable serialisable script parameters are detected. CoCreatAR persistently saves all modifications within the Unity scene file, enabling continued iteration and version control after the collaborative session concludes.

4.4.6 Spatial Context Capture with Coarse Meshes and 3D Snapshots

In-situ users have two tools for capturing additional spatial context to supplement the ex-situ user's representation of the real world: *3D Snapshots*, which are point clouds based on world-projected RGB-D frames, and *Coarse Meshes*, which are untextured 3D meshes.

The *3D Snapshot* feature uses the depth submodule of *Niantic SDK* to capture a depth image alongside the current camera frame from a LiDAR-enabled iPhone model. We note that *Niantic SDK* also supports 3D data capture on other devices and can extend beyond the LiDAR sensor's range. The depth data is combined with RGB information to generate a point cloud, which is transmitted to the ex-situ user's Unity scene. There, it is processed and visualised as a coloured world-space point cloud anchored to the *location mesh*. For in-situ users, the point cloud appears in white and is semi-transparent to provide feedback without distracting from the overall experience. This feature is particularly useful for capturing smaller objects or areas requiring finer detail, such as user interfaces or situated virtual objects that demand precise positioning.

For larger-scale spatial data capture, CoCreatAR provides a coarse 3D mesh capture feature, extending the *Niantic SDK* meshing submodule. Unlike *3D Snapshots*, the *Coarse Mesh* feature captures the environment's geometry without colour information. This method is well-suited for representing larger structures such as buildings, or open areas such as streets, where surface texture may be less relevant. Mesh updates are streamed to the ex-situ user's Unity editor in

real time, with network message coalescing applied to reduce network load. The meshing process automatically terminates after 15 seconds to prevent excessive network usage, and can be restarted at any time. The coloured mesh material represents surface normals to improve shape visualisation.

Both *3D Snapshots* and *Coarse Meshes* are persistently saved within the Unity scene for future reference, organised into separate timestamped objects for each collaborative session (Figure 4.6H).

4.4.7 Annotation of the Real World with Surface Draw and Air Draw

CoCreatAR also enables in-situ users to provide contextual feedback through annotations. The system supports two types of annotations: *Surface Draw* and *Air Draw*. *Surface Draw* projects the in-situ user's sketches onto real-world surfaces using depth estimates, while *Air Draw* allows users to draw freely in 3D space by moving their smartphone through the environment.

Annotations are colour-coded and can be tagged with predefined labels such as *hazard* or *user flow*, or with custom labels that are automatically assigned distinct colours for identification. All annotations are streamed to the ex-situ user's Unity editor in real time, where they appear as world-space objects and are saved for future iterations. Like spatial captures, annotations are persistently stored in the Unity scene, which allows contextual information to be revisited and used to inform subsequent development steps (Figure 4.6H).

4.5 Exploratory User Study

In this study, we aimed to explore how paired users co-author site-specific AR experiences using CoCreatAR, compared to a typical authoring workflow. In addition to assessing effectiveness and user experience, we sought to identify key directions for future research.

We employed a within-subjects counterbalanced design, comparing *two collaboration modes*: Async (baseline) and Sync (CoCreatAR, ours). For each session, we assigned one participant to the ex situ role and the other to the in

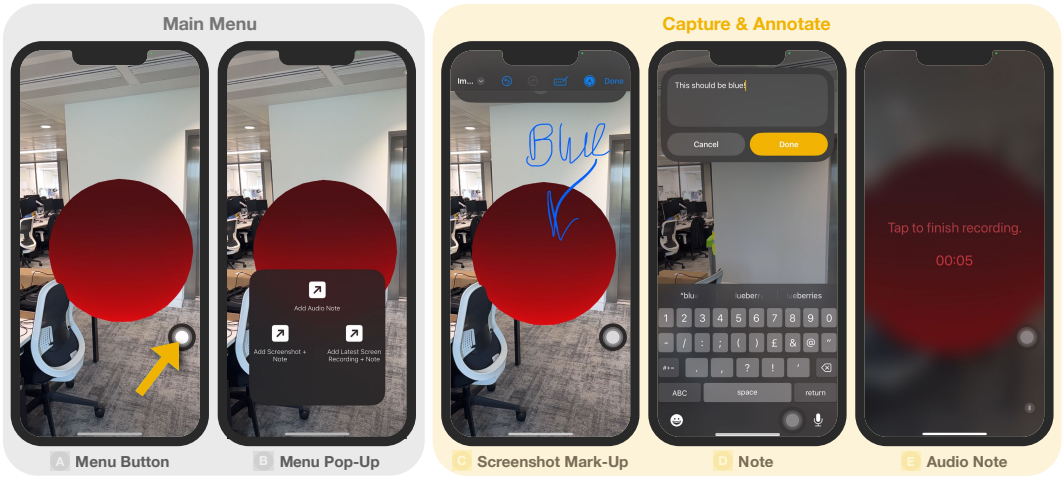


Figure 4.7 *In-situ* user interface of the baseline system. (A) The menu button (highlighted by a yellow arrow), which is the iOS *AssistiveTouch* button; (B) The menu where users can easily access screenshot, audio note, and screen recording features; (C) Interface for annotating screenshots; (D) Interface for attaching typed notes to screenshots or recordings; (E) Interface for recording audio notes. *Note: the interface for screen recordings is not shown, as the UI is simply a red dot in the upper right corner.*

situ role, based on their background experience. Participants were compensated with gift cards valued at £40 for study participation, which lasted around two hours. The study was approved by the UCL Computer Science Research Ethics Committee (Study ID UCL/CSREC/R/17).

To control for order and environmental effects, we counterbalanced both *collaboration mode* (Sync or Async) and *location* (*Location* A or *Location* B). Each pair began by completing the task using either the Sync or Async *collaboration mode*, with the in situ user physically at either *Location* A or *Location* B, while the ex situ user participated remotely. Participants then completed the task a second time using the alternate *collaboration mode*, with the in situ user at the alternate *location*.

We formulated the following hypotheses to examine how synchronous asymmetric collaboration changes workload, engagement, and confidence relative to an asynchronous workflow baseline, grounded in our formative study and prior work on collaborative AR authoring and social flow [282, 268, 286, 287]. We expected awareness cues in CoCreatAR to reduce ex situ users' need for explicit

grounding [83], while increasing coordination demands for in situ users, who must manage attention and action across the physical site, the AR experience, and the remote authoring context in parallel. We further hypothesised higher engagement under synchronous collaboration, consistent with prior work on collaborative AR authoring [282], and higher ex situ confidence, given that our formative study indicated delayed or incomplete site feedback as a primary driver of uncertainty during ex situ iteration.

> **Hypothesis 4.1** Ex-situ users will report a **lower** level of task load for Sync compared to Async.

> **Hypothesis 4.2** In-situ users will report a **higher** level of task load for Sync compared to Async.

> **Hypothesis 4.3** Ex-situ and in-situ users will report a **higher** level of engagement for Sync compared to Async.

> **Hypothesis 4.4** Ex-situ users will report a **higher** level of confidence in authored experiences for Sync compared to Async.

4.5.1 Baseline

Our baseline condition, Async, was designed to reflect the typical workflow identified in the formative study (Section 4.3). In this workflow, developers commonly receive or create notes, recordings, and screenshots with insights on improving the AR experience. To align with this workflow, we included these elements in the baseline condition. In our study, the in situ participant gathered this feedback using a series of custom iPhone *shortcuts* [288], activated via the iOS *AssistiveTouch* button. These shortcuts supported the creation of voice recordings, annotated screenshots, and screen recordings of the AR application, each automatically saved to the iOS *Notes* app. The user interface of this baseline

system is shown in Figure 4.7.

Once the in situ participant finished gathering information about the experience, the notes were sent to the ex situ participant. To recreate a typical developer environment, the baseline system for the ex situ user was based on Unity [266] and *Niantic SDK for Unity* [254]. The Unity editor contained the *location mesh*, an initial prototype of a site-specific AR experience, and a set of sample assets, including objects and materials from the public domain[3]. More information on the study procedure is available in Section 4.5.4.

4.5.2 Participants

We recruited 32 participants (16 pairs) through internal mailing lists and social media platforms. Participants self-reported their gender, with 19 identifying as men and 13 as women. The mean age of participants was 30.16 years (SD = 9.78). Participants were primarily students and researchers from fields related to computer science, virtual and augmented reality, and interactive media, with a few professionals from sectors such as government, consulting, and game development. For in situ participants, no prior experience was necessary. For ex situ participants, basic Unity experience was required. To increase our participant pool and obtain a broad range of experience in Unity, similar to the demographics identified in our formative study, we enabled ex situ participants to participate remotely from anywhere in the world using a high-speed internet connection. All participants collaborated in English.

Given the exploratory nature of our study, we selected a sample of 32 participants, consistent with related research on collaborative and authoring systems [63, 220, 57]. Moreover, the requirement for participants with Unity experience imposed recruitment constraints, limiting the feasibility of a larger sample [289]. Given the study's scale and exploratory nature, the statistical results should be interpreted with due consideration of potential biases and limitations, as outlined in Section 4.8.

[3] `https://kenney.nl/assets/`

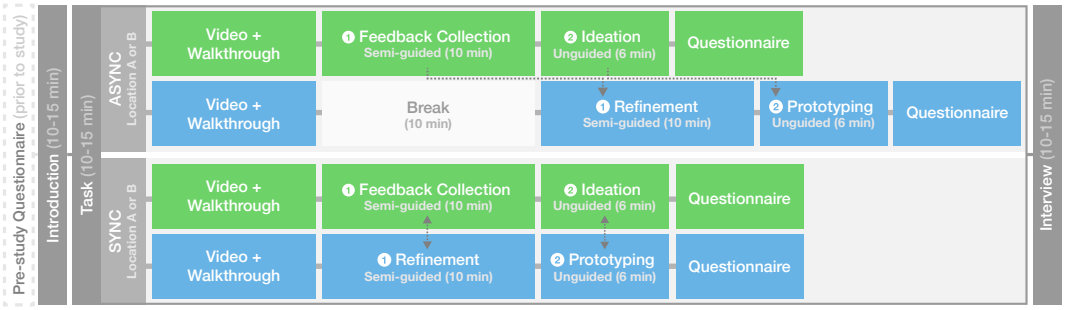


Figure 4.8 Overview of the study procedure. The process begins with a general introduction (left), followed by the task phase, performed under two conditions (SYNC (i.e., COCREATAR) and ASYNC), and concludes with a post-study interview (right). Green represents the procedure components of the in situ participants, blue represents the procedure components of the ex situ participants, and grey blocks represent components completed by both. ❶ represents Phase 1 (Feedback Collection & Refinement) and ❷ represents Phase 2 (Ideation & Prototyping) of the task. The arrows represent the exchange of task-related information across pairs. Note: both *collaboration modes* (SYNC and ASYNC) and *locations* were counterbalanced across pairs.

4.5.3 Task

The task involved a two-part collaborative process aimed at improving and expanding predefined site-specific outdoor AR prototype experiences, which are described in Section 4.5.3.1.

In *Phase 1:* ***Feedback Collection & Refinement***, participants focused on identifying and addressing issues within the prototype. The in situ participant primarily gathered feedback and contextual insights through real-world interaction with the prototype, while the ex situ participant refined the prototype by implementing necessary corrections based on the feedback provided. In the ASYNC condition, the feedback gathered by the in situ participant was available to the ex situ participant *after* the feedback collection phase. To accommodate this order of actions in the ASYNC condition, the ex situ participant had a short break prior to starting this phase while the in situ participant gathered feedback. In the SYNC condition, this process occurred synchronously.

In *Phase 2:* ***Ideation & Prototyping***, which lasted six minutes, participants transitioned to brainstorming and implementing extensions or enhancements to the prototype. During the ASYNC condition, the in situ participant ideated

independently, noting ideas and gathering relevant contextual artefacts, which were then shared with the ex situ participant for prototyping. In contrast, in the SYNC condition, ideation and prototyping occurred synchronously.

4.5.3.1 Prototype AR experiences

The prototype experiences used in the study were designed based on the environmental characteristics of the two selected outdoor locations (*Location* A and *Location* B), informed by both the *location mesh* and supplementary information from *Google Maps* and *Google Street View*. To maintain consistency across locations and ensure generalisability to other site-specific AR applications, the prototypes adhered to a set of predefined design dimensions. These dimensions, grounded in prior work and insights from our formative study, were applied to create similar AR prototypes for each location, ensuring participants could address comparable issues across both locations.

1. **Design Patterns** *used as design guidelines for the representation of each AR element, based on [290]:* Glyphs (navigation elements, indicative of actions with spatial alignment to referents), Decals (information related to referent context), Trajectories (e.g., paths, outlines, arrows), Labels (contextual annotations), Ghosts (overlays linked to referents), and Audio (non-visual cues providing additional context).

2. **Referent Types** *used as design guidelines for the physical referent of each AR element*: Small objects (e.g., mobile elements such as signs or markers), Large objects (e.g., statues, lampposts, or trees), Large planes (e.g., open spaces or flat surfaces), and Building structures (static physical structures with specific entry points or facades).

3. **Alignment Types** *used as design guidelines for the placement of each AR element*: Overlap (directly aligned with referents), Proximity (placed relative to nearby referents), and Surface (mapped directly onto physical surfaces).

4. **Issue Types** *used as a design guideline to incorporate common design issues, based on the five most prominent issues identified in Section 4.3:* Physical

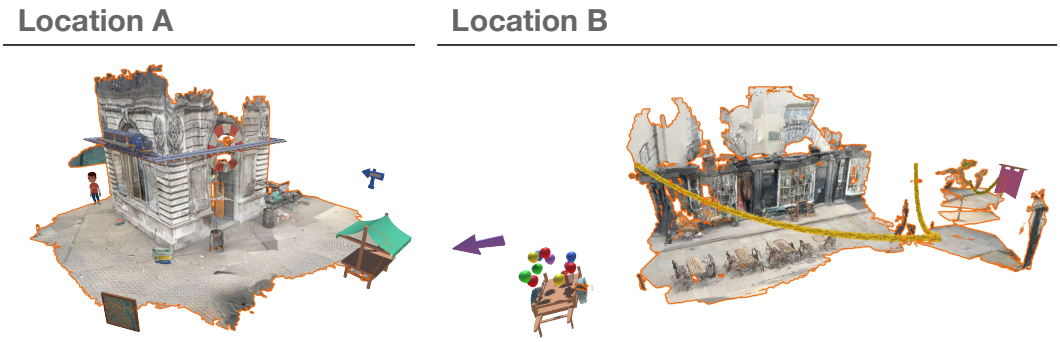


Figure 4.9 Screenshots of the prototypes designed for *Location* A and *Location* B. The *location mesh*, to which the visible virtual low-poly AR elements are anchored, is outlined in orange.

constraints, User safety, Misaligned AR elements, User context, and User flow, as visualised in Figure 4.2.

Both selected locations were within walking distance of University College London (*Location* A and *Location* B). *Location* A, located near a major transit station, featured high foot traffic, a nearby road, and elevated levels of environmental noise. The theme of the AR experience at *Location* A was "*Welcoming Tourists to the City*," reflecting its proximity to a transit hub. *Location* B was situated along a narrow street with a local cafe, characterised by significant foot and bicycle traffic and lower environmental noise. The theme for this location was "*Promoting and Celebrating the Upcoming Cookie Party*," aligning with the atmosphere of nearby restaurants and cafes. For each location, seven issues covering the range of issue types, as previously defined, were identified and verified through on-site testing prior to the user study. An overview of the prototypes is shown in Figure 4.9, with detailed descriptions available in the Appendix (Section C.3).

To provide both participants with hints regarding issues in the prototype, both in situ and ex situ participants received a list of seven *task clues* in Phase 1. This list contained fictional observations made by "early testers" of the prototype experience, offering participants a semi-guided path through the experience while still requiring further investigation to ensure consistency across pairs. Both task clue lists are available in the Appendix (Section C.3). Participants were

allowed to start at any point on the list and could skip or revisit clues as desired.

4.5.4 Procedure

An overview of the study procedure is shown in Figure 4.8. Prior to the study session, participants completed a questionnaire covering demographics and their background in smartphone-based AR, 3D game engines, and 3D editors.

For each session, the in situ participant took part in person using an iPhone 13 Pro, while the ex situ participant participated remotely by connecting to a PC with all necessary tools pre-installed via the low-latency screen-sharing software *Parsec*[4].

Upon arrival, the in situ participant was welcomed by an experimenter, while another experimenter connected with the ex situ participant via video call. The in situ participant was then brought to a room where they joined the same video call. The experimenter provided a general introduction to the study context and procedure. Before the first condition, the ex situ participant was guided through a setup document containing log-in information and instructions for forwarding their microphone to the PC. Once the ex situ participant connected, latency statistics were recorded, ranging between ~35 and 120ms (including network, decoding, and encoding latency) across participants.

Each pair completed the task twice: once under the Sync and once under the Async *collaboration mode* conditions, with one condition occurring in *Location* A and the other in *Location* B. At the start of each condition, participants watched a tutorial video tailored to their role and condition. The in situ participant then proceeded to either *Location* A or *Location* B, assisted by an experimenter, where they received a brief walkthrough of the Sync or Async system. Meanwhile, the ex situ participant received a similar walkthrough remotely, along with an introduction to the prototype experience that provided background information on its components and initial design decisions.

In the Async condition, during Phase 1 (Feedback Collection & Refinement), the in situ participant had ten minutes to gather feedback on the prototype

[4] `https://parsec.app`

experience and any relevant real-world context necessary to address the issues listed on the task clue list. The in situ participant then proceeded to Phase 2 (Ideation & Prototyping), where they collected artefacts related to new ideas for improving or extending the prototype. Concurrently, the ex situ participant began Phase 1 by refining and correcting the prototype based on the other participant's feedback artefacts. After six minutes, the in situ participant's Phase 2 concluded, and their artefacts were made available to the ex situ participant for Phase 2, which commenced after Phase 1 (refinement) was completed. The in situ participant then returned indoors, guided by the experimenter, and both participants completed post-task questionnaires separately.

The same process was followed for the SYNC condition, except that the in situ and ex situ participants progressed through both phases simultaneously. Upon completing both conditions, participants were interviewed separately to discuss their experiences across conditions.

4.5.5 Measures and Analysis

Each collaborative session was recorded from the perspective of the ex situ participant. In the SYNC condition, this recording included the streamed view of the in situ participant along with the audio captured from both participants. The recordings were coded to track each participant's feature usage throughout the session. An overview of feature usage is provided in the Appendix (Section C.4.4) and is discussed in Section 4.7.1.

The post-task questionnaire included the NASA Task Load Index (TLX) [291] to assess workload across conditions using a 7-point Likert scale. All TLX items were included for in situ participants, whereas the item measuring physical demand ("How physically demanding was the task?") was excluded for ex situ participants due to their seated, computer-based role in both conditions. An overall task load score was calculated by averaging the responses to the included items.

To measure engagement, we employed the short-form version of the User Engagement Scale (UES-SF) [292], focusing on the subscales of *focused attention*,

perceived usability, and *reward*. The *aesthetic appeal* subscale was omitted as it was deemed irrelevant. An overall engagement score was obtained by averaging the 5-point Likert scale responses across subscales, adjusting for reverse-coded items.

Additionally, we included several custom items in the post-task questionnaire, which are provided in the Appendix (Section C.5.2). For ex situ participants, custom items assessed their *overall confidence* in the final AR experience as end-users would encounter it in situ, confidence in successful execution of task components that influence end-user experience, and specific feedback on the goals of the different task phases. Both participant groups also rated perceived overall task performance using a custom item.

Before concluding the study, both participants were interviewed separately. The post-study interviews focused on evaluating their experiences and preferences regarding the systems and workflows in the Sync and Async conditions. Key topics included overall preference, confidence, communication quality, advantages and limitations of each method, and suggestions for improving system features. The full interview scripts are available in the Appendix (Section C.5.2). The author of this dissertation and a collaborator used an affinity diagramming approach to analyse and synthesise themes from the interview transcripts.

4.6 Exploratory User Study Results

Building on the formative study and resulting requirements, we present the results of our evaluation of CoCreatAR following the user study design detailed in Section 4.5. We first report the self-reported questionnaire results, followed by four themes derived from a thematic analysis of interviews with ex-situ and in-situ participants. Plots that show an overview of the usage of specific CoCreatAR features are provided in the Appendix (Section C.4.4).

4.6.1 Questionnaire Results

We conducted statistical analyses using linear mixed-effects models, accounting for participant variability as a random effect and controlling for confounding

factors such as task location and prior experience. All models converged, and model assumptions were verified (linearity, normality, and homoscedasticity). Where applicable, we report the estimated coefficients (β), standard errors (SE), p-values, and 95% confidence intervals (CI). Unless stated otherwise, each analysis includes 32 observations ($n = 32$). In this section, we provide full details for statistically significant and marginally statistically significant results, whereas an overview of all statistical results is available in the Appendix (Section C.4).

Cronbach's alpha values indicated acceptable internal consistency for engagement and task load measures. Engagement items yielded alpha values of 0.740 for in-situ users and 0.764 for ex-situ users. Task load items had alpha values of 0.812 and 0.772 for in-situ and ex-situ users, respectively.

4.6.1.1 Engagement

The analysis revealed a significant effect of *collaboration mode* on self-reported engagement for both ex-situ and in-situ participants. Ex-situ participants reported significantly higher engagement in the SYNC condition ($\beta = 0.500$, SE = 0.140, $p < 0.001$, 95% CI [0.226, 0.774]). A similar effect was observed for in-situ participants, who reported significantly higher engagement in the SYNC condition ($\beta = 0.340$, SE = 0.130, $p = 0.009$, 95% CI [0.085, 0.595]). A plot of engagement scores is shown in Figure 4.10A.

While no significant effect of *location* was found for ex-situ participants ($p = 0.196$), in-situ participants reported significantly higher engagement at *Location* B ($\beta = 0.299$, SE = 0.130, $p = 0.022$, 95% CI [0.044, 0.554]). *Prior experience* with AR or 3D editing tools did not significantly influence self-reported engagement for either group (ex-situ: $p = 0.093$; in-situ: $p = 0.231$).

4.6.1.2 Task load

Self-reported task load showed no significant difference between SYNC and ASYNC *collaboration modes* for either ex-situ participants ($\beta = 0.313$, SE = 0.309, $p = 0.312$) or in-situ participants ($\beta = 0.213$, SE = 0.345, $p = 0.538$). A plot of task load scores is shown in Figure 4.10B.

Location did not significantly impact self-reported task load for ex-situ

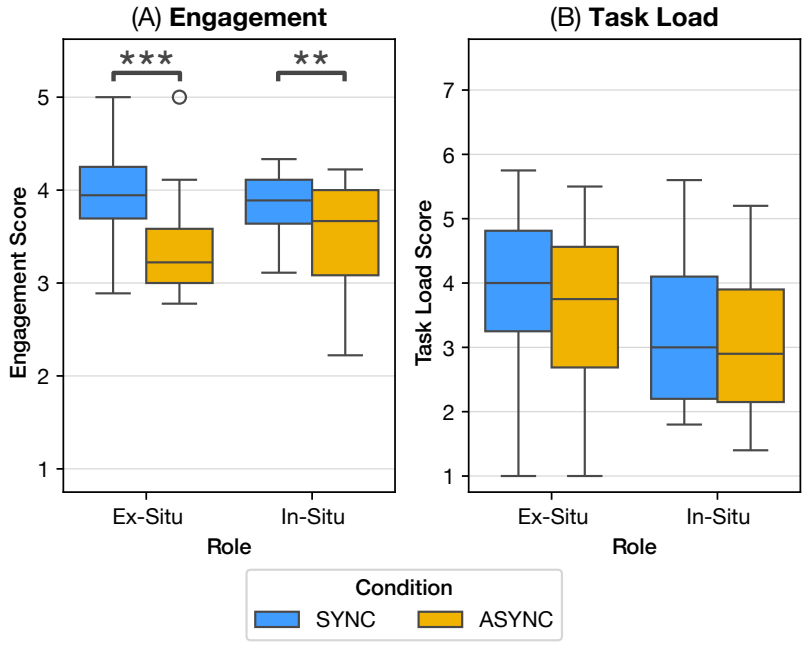


Figure 4.10 (A) Engagement scores for each role per condition, with significantly higher engagement in the Sync condition for both in-situ and ex-situ participants. (B) Task load scores for each role (ex-situ and in-situ) per condition. No significant differences were found between conditions. Levels of statistical significance: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

participants ($p = 0.419$). However, in-situ participants reported a marginally significant decrease in task load at *Location* B compared to *Location* A ($\beta = -0.637$, SE = 0.345, $p = 0.065$, 95% CI [−1.314, 0.039]). *Prior experience* did not significantly affect self-reported task load for ex-situ participants ($p = 0.162$). However, in-situ participants showed a marginally significant association between lower task load and greater AR experience ($\beta = -0.320$, SE = 0.176, $p = 0.070$, 95% CI [−0.666, 0.026]).

4.6.1.3 Confidence in authored results

Confidence in the number of issues fixed and overall self-reported confidence were both significantly influenced by *collaboration mode*. Participants in the Sync condition reported fixing more issues with confidence than those in the Async condition ($\beta = 3.000$, SE = 0.373, $p < 0.001$, 95% CI [2.270, 3.730], $n = 24$). Additionally, overall confidence scores were significantly higher in the Sync condition ($\beta = 1.125$, SE = 0.387, $p = 0.004$, 95% CI [0.367, 1.883]). An overview

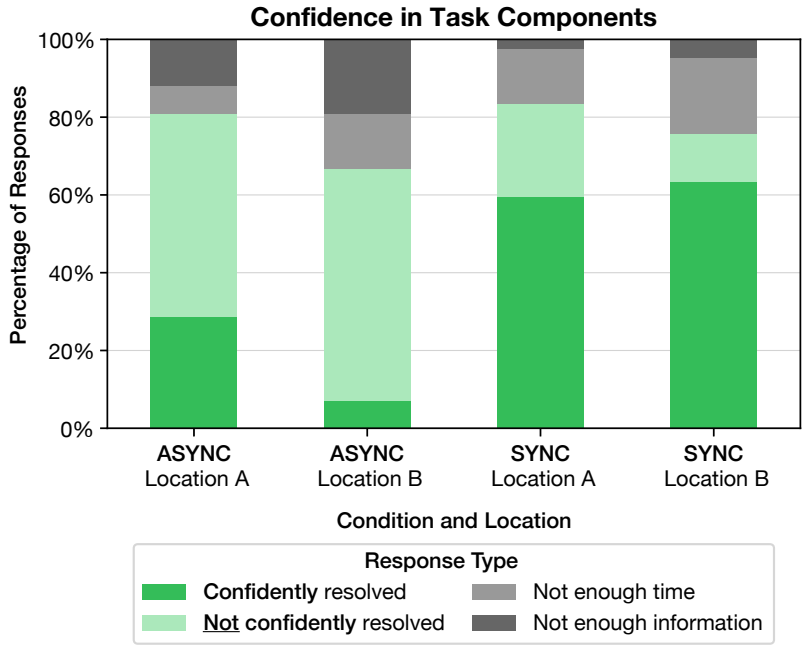


Figure 4.11 Stacked bar plot illustrating response proportions for items in which participants indicated perceived performance and confidence for each task component across *collaboration mode* and *location* combinations.

of the proportion of participant responses regarding perceived performance and confidence for each task component is shown in Figure 4.11.

A marginal effect of *location* was observed for ex-situ participants, with fewer issues confidently fixed at *Location* B ($\beta = -0.667$, SE = 0.373, $p = 0.074$, 95% CI [−1.397, 0.064], $n = 24$). However, *location* did not significantly affect overall confidence scores ($p = 0.747$). *Prior experience* did not significantly impact the number of confidently fixed issues ($p = 0.096$) or overall confidence scores ($p = 0.466$).

4.6.2 Interview Results

At the start of the interview, each participant was asked which system they would prefer to use if they were to perform a similar task again. Among in-situ participants, the majority (12 out of 16) expressed a preference for the Sync condition, citing advantages such as real-time feedback, improved communication, and higher engagement. Two participants indicated that their preference depended on the scenario, while two preferred the Async condition, highlighting the ability to focus without the distractions of real-time interaction.

Similarly, ex-situ participants largely preferred the SYNC condition (12 out of 16), with many emphasising the value of real-time collaboration and immediate feedback. Three participants indicated that their preference depended on the context, and one preferred the ASYNC condition due to the slower pace, which allowed more time to refine contributions.

Our thematic analysis of interview transcripts resulted in four main themes comparing the synchronous COCREATAR and asynchronous baseline experiences. We denote in-situ participant quotes with an 'I' (e.g., I2) and ex-situ participants with an 'E' (e.g., E8).

4.6.2.1 Collaboration with in-situ users facilitated the integration of real-world context into ex-situ users' design process.

In the ASYNC condition, all ex-situ participants struggled to integrate real-world context into their design decisions due to missing information. Ex-situ users were uncertain about object placement and its interaction with the environment, as they could not verify details without live input. One participant noted, "I'm not 100% sure about the real way, if it's low enough or is this blocking other ways because I need to see it from the real end" (E7). Lacking live visual and spatial information, ex-situ participants relied on limited artefacts provided by in-situ participants, restricting their ability to account for dynamic environmental elements. One participant explained, "I tried to put the map onto that light post but I'm not sure whether that's on the post or not" (E7).

In contrast, the SYNC condition enabled participants to gain a clearer understanding of real-world context, leading to more informed decision-making in the design process (E12, E10, E9, I5, E17). One ex-situ participant emphasised the importance of real-time observation, particularly for user flow, explaining that although an object appeared well-positioned, real-world feedback revealed that it interfered with pedestrian flow: "We had a situation where people were walking through one of the objects. It looked like an ideal location for the object on the map [*location mesh*], but in real time and in real life, it wasn't going to work because it was in a walkway" (E10).

The integration of real-world context through synchronous communication also extended to addressing safety and navigation concerns. One ex-situ participant described uncertainty about whether a virtual cart placed in the scene would interfere with foot and cycling traffic, as there were two real obstacles on either side of it. They noted that this uncertainty was mitigated by being able to check the scene in real time with the in-situ participant (E12). Another participant highlighted the value of hearing environmental noise in real time, which influenced decisions about setting audio levels in the AR experience: "I need to know the population density there, like what's the environment like, should I put the speaker very loud or not at all" (E9).

Participants also reported that changes in the environment were easier to account for in the SYNC condition. One ex-situ user noted that objects had shifted between the time the location mesh was captured and their session, which they would not have noticed without real-time feedback. As they explained, "The barrel, [...] had actually moved... I wouldn't have been able to tell that from the picture or the scan [location mesh]" (E17). Following this realisation, they asked the in-situ participant to place cursor markers indicating the correct positions of the objects on the barrel.

Ex-situ participants also emphasised that synchronous collaboration allowed them to indirectly experience contextual elements of the real-world environment. One participant underscored the importance of experiencing real-world audio and visual elements: "You can actually hear what they're hearing as well, which is quite important" (E7). This sense of presence contributed to increased understanding of spatial relationships within the environment. One participant mentioned, "I got a much better idea of the space when the other person was there," as the in-situ user provided real-time feedback and visual information that could not have been fully captured asynchronously (E16). The combination of real-time video, real-time 3D captures of the environment, and verbal feedback gave ex-situ users a more comprehensive view of the environment, improving their understanding of how the design fit within real-world

context (E7, E9, E10, E12, E16, E17).

4.6.2.2 Immediate feedback and changes supported confidence in design decisions, mutual understanding, and perceived accuracy of the outcome.

Participants noted that the SYNC condition facilitated more informative feedback and iterative adjustments, with many additionally highlighting that this strengthened their confidence in the authored outcome (I2, E3, E5, E7, I10, E12–17). The ability to communicate in real time supported quicker, informed decisions, with several participants noting that the immediate feedback loop reduced guesswork when refining the AR prototype. As I16 stated, "I think that, very strongly, I'd be more confident that the positions that we put those objects, they actually line up with the real world a lot better." This confidence was echoed by other ex-situ participants who mentioned that making changes in real time felt "much, much quicker" (E3), "was a lot more dynamic" (E12), and that "it felt like, once we were done, it was very complete" (E10).

The ability to communicate synchronously and see changes immediately increased confidence and improved mutual understanding among pairs. Participants emphasised that the visual nature of real-time interaction removed ambiguity, leading to more accurate placement of virtual objects. Participant I14 noted that discrepancies were minimised: "I don't think we had a lot of discrepancies in what we discussed about [in the SYNC condition], which means probably we had a good sync about the scene."

Beyond the real-time feed, COCREATAR features such as the *3D Cursor*, annotation tools, and the ability to capture *Coarse Meshes* and *3D Snapshots* improved mutual understanding between ex-situ and in-situ users. Figure 4.12 presents examples of feature usage throughout the user study. Participant I16 noted that using the annotation tool to mark exact locations allowed them to "draw exactly where" changes were needed, acting as "the bridge between being there in reality and the scene that [ex-situ user] is seeing." Similarly, I5 emphasised that "drawing in real time is much easier than just explaining it," particularly when

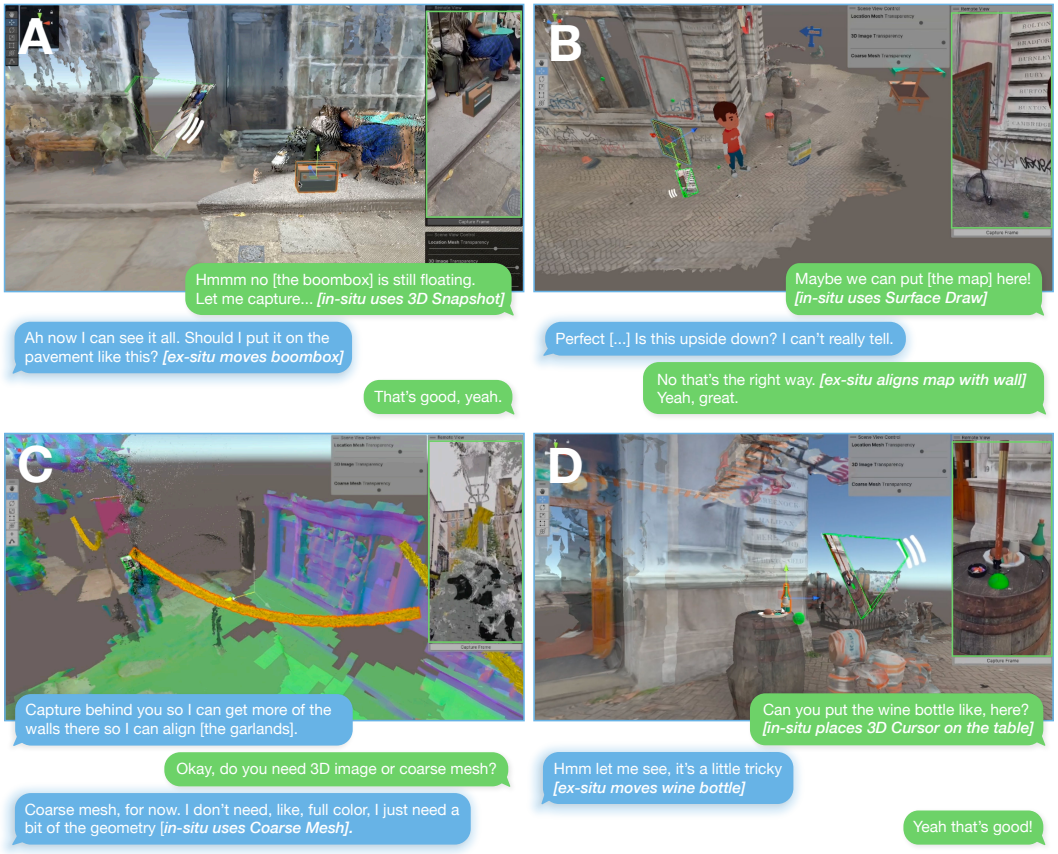


Figure 4.12 Overview of CoCreatAR feature usage during Phase 1, shown as *ex-situ* perspective screenshots. Participant conversations appear in colour-coded speech bubbles: in-situ (green) and ex-situ (blue). Speech bubbles with a *glow* indicate utterances concurrent with the screenshot. (A) Alignment of a boombox using spatial context captured with *3D Snapshot*. (B) The *ex-situ* participant repositioning the map on the wall as specified by the *in-situ* participant via *Surface Drawing*. (C) Alignment of a garland to a previously unmapped street region using *Coarse 3D Mesh*. (D) Alignment of misplaced food items based on *in-situ* input using *3D Cursor*.

identifying hazardous areas. The 3D cursor helped reduce ambiguity by allowing participants to visually reference key elements and create placeholders for alignment (I2, I5, I6, I8–10, I12, I14, I16–17, E6, E8, E10–12, E14, E16–17). Additionally, E3 highlighted how the ability to capture *Coarse Meshes* and *3D Snapshots* enabled the in-situ user "to add more detail to the scene," helping ex-situ users make more informed design decisions.

In contrast, the Async condition often resulted in lower confidence due to the lack of immediate feedback and reliance on static information (E9, E12). As E12 described, "With the asynchronous one, I don't have any information in terms

of how much should I move... there's no feedback whatsoever." The absence of real-time interaction forced participants to work with limited information, leading to hesitant decisions. E5 emphasised this challenge: "There was some ineffective information for me in the notes," indicating that static instructions without live verification did not provide the necessary context for accurate design decisions. This lack of real-time verification made it difficult to gauge whether changes were correctly applied. E4 highlighted this issue, stating, "You can't see the change you made, like, in the real world."

In-situ participants also expressed uncertainty about whether they had captured the right content or communicated their observations clearly (I6, I12, I14, I17). One in-situ participant stated, "I wasn't sure it was clear enough" (I14), while another noted that the lack of immediate dialogue required them to provide excessive detail, which still might not ensure accurate interpretation (I17).

4.6.2.3 Synchronous authoring increased engagement and encouraged creative exploration through collaborative interaction.

In the Sync condition, participants frequently reported higher levels of engagement, often attributing this to the sense of real-time collaboration and mutual decision-making (I3, I8, I12, I15, E3, E5, E12, E14, E17). For instance, I10 described the synchronous condition as "a lot more fun and engaging" due to the opportunity to work together with another person. Similarly, I14 noted that the synchronous session was "more enjoyable" and led to "higher engagement" because of the collaborative nature of the task. This sentiment was echoed by E17, who highlighted the satisfaction of "working together" and described the experience as akin to "live game testing." This framing suggests that seeing immediate reactions and feedback from the in-situ user created a more interactive and stimulating process. The ability to see ideas realised in real time contributed to the creative experience, as noted by multiple participants (E9, I17).

Synchronous collaboration also supported a sense of teamwork and shared ownership of the final result, contributing to a more engaging authoring process

(E7, I10, I12, E16). One participant explained that "it was fun to chat to another person while I was doing it," which made the task feel more like a collaborative endeavour rather than an individual effort (I6). This increased engagement was reflected in the enthusiasm with which participants approached the synchronous conditions. One participant stated that they felt "buzzing" with excitement after completing the task together with their collaborator (I12).

In contrast, the ASYNC condition was perceived as less engaging by most participants (I3, I4, E4, I6, E7, I8–I10, I12, E12, I14–I17, E14, E16). Participant I12 compared the two conditions, explaining that the asynchronous condition felt "more like a task as opposed to fun," a sentiment echoed by other participants who described the asynchronous process as more isolated and procedural (I4, I7, E14, I17). E12, for example, mentioned that the asynchronous condition felt "quicker but less satisfying," as there was no immediate feedback or dynamic interaction.

In-situ participants frequently expressed dissatisfaction with working towards a goal without seeing the final result (I3, I4, I5, I6, I7, I12, I16). For example, I16 stated, "It was a weird feeling leaving and then not seeing it change." The lack of real-time collaboration in the ASYNC condition also constrained opportunities for creative exploration. One participant explained that without the ability to brainstorm with another person, the task became "more about checking a box" rather than experimenting with new ideas (I10).

Collaborative brainstorming emerged as a key driver of engagement in the synchronous sessions. Participants reported that having someone to exchange ideas with in real time supported the creative process and led to more diverse and spontaneous solutions (I9, I10, I17, E14). Participant I9 emphasised the ease of "drafting concepts" together in the synchronous condition, stating that "we could brainstorm easily" and quickly iterate on suggestions. These interactions facilitated creative exploration and strengthened the sense of shared authorship, which several participants valued (I17, E10).

4.6.2.4 Multitasking overwhelmed some participants in synchronous collaboration, while asynchronous workflows were seen as more suitable for certain scenarios.

Although the synchronous condition generally increased engagement and collaboration, some participants reported that the complexity of multitasking within the synchronous workflow was overwhelming (I8, E16, I17). Some participants also noted that the SYNC condition increased pressure, as they felt their collaborators had to wait for them to complete tasks (I4, E12, E14).

Participant E16 pointed out that managing multiple tasks simultaneously, such as communicating with the in-situ user, observing the scene, and making design adjustments, led to cognitive overload: "I felt like I was doing too many things at once. Talking to [the in-situ user] and trying to watch the environment was sometimes just too much" (E16). Similarly, I10 expressed that for users unfamiliar with the system, synchronous interactions might feel overwhelming, particularly due to the need to navigate while processing real-time feedback: "For someone who's not experienced, it was just a lot. You're trying to follow directions, but there's so much going on" (I10). This aligns with the feedback of several other participants less familiar with AR, who wished they had more time to practise ahead of the task (I3, I6, I10, I13, E10, E11, E16).

Several participants indicated that the synchronous condition was well-suited to scenarios where immediate feedback was critical, while the asynchronous condition was more suitable when working in overwhelming environments. Participant I10 described how synchronous interaction could become overwhelming in crowded environments with high noise levels, explaining that "if you're dealing with a busy place and someone is talking in your ear, it gets really overwhelming" (I10).

Participants also recognised that both synchronous and asynchronous workflows had their strengths depending on the scenario. While some participants felt overwhelmed by multitasking in the synchronous condition, they still ac-

knowledged its value for tasks requiring rapid decision-making or creative brainstorming. E12 noted, "It's definitely harder when you have to do everything at once, but I still think the synchronous one is better when you're trying to bounce ideas off someone" (E12). Participant I16 described the SYNC condition as most applicable for complex experiences, whereas the ASYNC condition might be more suitable "if it was, like, a smaller experience with one object." E9 also described scenarios where asynchronous workflows could be beneficial, such as when in-situ participants already have "enough information to already produce what I want" (E9).

Moreover, participants reflected on how asynchronous and synchronous workflows or features could be complementary and applied at different stages of the development process (I5, E15, E16). For example, E15 highlighted a hybrid approach: "I kind of see it as you start with asynchronous, then you do synchronous to refine it, to collect feedback on your experience." Other participants described potential for a hybrid system that flexibly integrates asynchronous and synchronous workflows while incorporating CoCreatAR capturing and annotation features (I5, E15). E16 also proposed enabling lightweight in-situ editing in addition.

4.7 Discussion

In this section, we reflect on the results of the user study, discuss our findings in the context of our formative study and related work, and offer recommendations for future authoring systems for site-specific outdoor AR experiences.

4.7.1 Enhancing Environmental Understanding Through Capture and Communication Tools

Our findings indicate that synchronous collaboration using CoCreatAR significantly enhanced the ability of ex-situ participants to integrate real-world context into their design process. In particular, ex-situ participants in the synchronous condition demonstrated a more comprehensive understanding of environmental dynamics such as user flow and spatial relationships. In this subsection, we

contextualise how each feature group contributed to this outcome, drawing on insights from post-study interviews and observations made when coding feature usage counts (see the Appendix, Section C.4.4).

Participant feedback emphasised that real-time video and audio communication was the most critical feature of CoCreatAR. It provided ex-situ participants with immediate visual context from the in-situ user's perspective and supported stable communication. This was especially valuable when in-situ participants felt overwhelmed, enabling them to carry out the task nonetheless. Notably, background noise captured by the in-situ user's microphone proved useful for decisions regarding audio components (Section 4.6.2.1), an issue that, according to P1 of our formative study, could only be resolved when in situ (Section 4.3.2.3). This observation aligns with prior arguments that context is not exhaustively specifiable at capture time and becomes relevant through activity as collaborators negotiate what matters for the task at hand [48]. Additionally, some ex-situ participants mentioned that background audio increased their sense of presence in the in-situ environment.

Surface Draw was commonly used due to its task relevance, as objects were typically tied to physical referents. It was often used as a referencing tool to indicate precise placement locations on surfaces (e.g., the map on the wall in Figure 4.12B), in line with early work on referential annotation [293]. *Air Draw* was primarily used to indicate user paths and, occasionally, to create 3D wireframe placeholder objects during brainstorming. Labelled annotations were used less often than colour annotations, likely because ex-situ participants were not expected to revisit annotations asynchronously at a later stage during the study, unlike in real-world scenarios where annotations may be revisited over time.

Coarse Mesh was often used to capture large areas (e.g., buildings in Figure 4.12C), while the *3D Snapshot* feature was favoured for specific points of interest (e.g., a curb and table setting in Figure 4.12A). Although the number of *3D Snapshot* captures was higher overall, the longer capture time required for *Coarse Meshes* balanced the frequency of their use. The lack of colour in *Coarse*

Meshes was not described as a major issue, though one participant captured over ten *3D Snapshots* to obtain a large coloured point cloud. *3D Cursors* were mainly used by in-situ participants for object alignment (Figure 4.12D) and to support deictic references, while ex-situ participants used *3D Cursors* to guide in-situ users' movements or capture actions. In line with prior work on workspace awareness, these cursor-based cues provided lightweight action visibility that reduced ambiguity about what a collaborator was attending to or acting upon [198, 293].

Recommendations for future work Enhancing spatial representations and integrating richer metadata layers could expand CoCreatAR's potential for complex real-world locations. Future work could leverage 3D Gaussian splatting techniques [294] and their temporal extensions [295, 296] to support more intricate design tasks and more precise alignment. Automatically capturing metadata such as pedestrian flow [297] or hazards [298] during on-site visits could streamline context handling for synchronous and asynchronous ex-situ authoring.

Future work could also leverage insights from Fussell et al. [104] to enhance collaboration in CoCreatAR. For instance, visualising collaborators' actions, such as object selection or transformation, could improve task awareness. Alternative in-situ interfaces, including HMDs with a larger FOV, might offer a more comprehensive perspective and support natural interactions. Furthermore, hand tracking could enable intuitive gestures, as observed in our study when participants occasionally employed deictic gestures in front of the device camera [299, 293].

4.7.2 Task Load and Multitasking Challenges

While our results regarding task load are inconclusive, they offer early insights into the impact of task context, multitasking demands, and individual differences. A slight trend visible in Figure 4.10B suggests higher task load in the Sync condition for both roles, although no statistically significant differences were observed across conditions. Notably, qualitative feedback revealed that a subset

of participants, particularly those with less prior experience, expressed feeling overwhelmed during the SYNC condition. This feedback provides preliminary evidence that challenges **H4.1**, suggesting that ex-situ participants may experience higher, rather than lower, task load in the SYNC condition. At the same time, qualitative observations align with **H4.2**, as in-situ participants reported multitasking challenges consistent with higher task load in the SYNC condition. Based on the interviews, we conclude that an increase in task load during the SYNC condition may arise from a substantial volume of information flow combined with the demands of concurrent coordination and communication. This is consistent with prior research on awareness cues and grounding processes, which highlights a trade-off between rich information and processing costs [25, 300]. For in-situ users, this challenge appeared compounded by the need to safely navigate the real world, which has been recognised as inherently demanding [301].

Some participants reported that limited familiarity with the system contributed to their sense of being overwhelmed. We hypothesise that these experiences may be shaped by participants' background knowledge and individual personality traits [302]. As we detail in Section 4.8, follow-up research could examine these hypotheses further in larger-scale studies.

Participants' qualitative feedback also highlighted potential *process losses*, defined as productivity reductions that occur when individuals collaborate synchronously rather than working independently [303]. Specifically, some participants reported that the multitasking demands of the SYNC condition, such as communicating with their collaborator while performing tasks, led to perceived inefficiencies. Related challenges included waiting for collaborators to complete tasks (e.g., annotation by an in-situ user or scene editing by an ex-situ user) and feeling pressured to avoid blocking their partner's progress (Section 4.6.2.4). Guo et al. [282] similarly observed potential process losses in synchronous collaboration during an authoring task, noting that skill mismatches within pairs can limit creativity. While we did not observe direct evidence of this particular form of process loss, we theorise that clearly defined roles and responsibilities, together

with the semi-guided nature of the task, may have mitigated these effects.

Recommendations for future work Future work could benefit from more extensive onboarding processes to familiarise users with system features and clarify collaborators' roles. Additionally, future iterations of CoCreatAR could address challenges related to task load by streamlining workflows. For instance, simplified scene editing for ex-situ users could be supported through modular templates [304] or AI-driven authoring tools [305, 275, 306]. In parallel, in-situ users could benefit from AI-assisted annotation methods, such as transcribed audio notes [307, 308].

To improve orientation and reduce cognitive load for in-situ users in complex environments, future systems might integrate visual aids such as mini-maps [309] or directional cues [310], in addition to enhanced visual indicators for collaborator actions as noted in Section 4.7.1. Lastly, for ex-situ users, features such as user-perspective scene rendering [311] and selective visualisation of information layers [312, 313] could help reduce visual overload.

4.7.3 Engagement and Creative Exploration

Both ex-situ and in-situ participants reported higher engagement in the Sync condition, supporting **H4.3**. Real-time interaction supported effective teamwork and creative exploration, which the vast majority of participants described as more enjoyable than the Async workflow. This finding aligns with prior research on *social flow*, which found that interdependent team settings amplify engagement compared to solitary workflows [286].

Prior work on in-situ AR game level editing [287] found that users were highly engaged in both creating AR content and observing others interact with it. Similarly, in our study, in-situ users were engaged by experiencing and interacting with the AR content as it materialised in their environment, while ex-situ users likened the process to live game testing, emphasising how immediate in-situ feedback served as a source of inspiration (Section 4.6.2.3).

The advantages of synchronous collaboration observed in our study are

supported by findings from Guo et al. [282], who identified synchronous authoring as a driver of engagement and creativity compared to asynchronous authoring. Notably, their work also reported a user preference for *co-located* over *remote* synchronous collaborative authoring, which relied on standard video conferencing. This preference underscores the importance of designing remote collaboration tools that support person-oriented cues alongside task-oriented views, a recurring concern in CSCW work [98]. For instance, I3 expressed a desire to see their collaborator, particularly when working together for the first time.

Recommendations for future work While CoCreatAR currently employs avatars to simulate the embodied perspective of co-located collaboration, future research could investigate enhancing interpersonal connection through spatial video-based avatars, as demonstrated in recent work [314, 315]. Another avenue for future research is to explore how creativity can be supported in larger groups, which offer potential for richer creative outcomes [316] and greater productivity. Although larger groups may face greater coordination overhead, strategies such as role specialisation, as employed by XRDirector [57], could streamline workflows in larger teams.

4.7.4 Confidence in End-User Experience Through In-Situ Feedback

Our findings demonstrate that participants in the Sync condition reported significantly higher confidence in the authored AR experiences than those in the Async condition, supporting **H4.4**. Participant feedback indicated that this confidence stemmed from the ability to iteratively verify the designed experience in real time, ensuring alignment with the spatial and contextual dynamics of the real-world environment. In contrast, participants in the Async condition struggled with uncertainty due to reliance on static data and delayed or incomplete feedback, which often led to hesitation in their design decisions. In terms of RQ2, this uncertainty reflects a representational limitation of asynchronous context-sharing approaches for site-specific outdoor authoring, since partial captures

and delayed updates shift effort towards interpretation and repair during design decision-making.

This positive influence of representative in-situ feedback aligns with prior research in ubiquitous computing that emphasised the critical role of situated evaluation in achieving a reliable understanding of how applications behave in real-world contexts [317, 318], an approach that our work both embraces and extends.

Notably, much of this research has focused on asynchronous techniques such as event logging and recordings to collect in-situ feedback [317, 319]. Our findings contribute a complementary perspective by highlighting the opportunity to evaluate and adapt an application within the same session, a workflow that our user study suggested is effective for iterative refinement of site-specific outdoor AR experiences.

Recommendations for future work Future work could explore how author confidence and perceived quality of AR end-user experiences evolve over multiple revisits of a site, especially under dynamic environmental conditions. Additionally, research could focus on developing quantitative metrics that represent the quality of AR experiences, as explored by ARCHIE [320], to provide authors with additional objective insights. While quantitative metrics typically require controlled environments or ground truth sources, metrics derived from user behaviours (e.g., frequency of manual relocalisation triggers) could be a practical alternative. Finally, author confidence could be supported by incorporating constraints for contextual adaptation, such as semantic referents and rules [275, 258], which have been shown to improve confidence when used alongside virtual simulation environments for indoor AR tutorial authoring [275].

4.7.5 Enabling Hybrid Approaches to AR Experience Authoring

Despite the advantages of synchronous collaboration, participants identified some scenarios where asynchronous workflows were preferable. In contexts

with overwhelming environmental conditions, or when tasks did not require immediate feedback, asynchronous methods offered more flexibility and reduced pressure. This insight underscores the importance of providing flexible authoring workflows that can adapt to the diverse needs of authoring teams [268], including collaboration across *place* and *time* [72].

Recommendations for future work Studying the impact of switching between synchronous and asynchronous modes on collaborative processes and authoring outcomes could reveal opportunities for optimisation. Furthermore, to support flexible workflows, future work could explore designing hybrid systems that integrate session management and version control for iterative collaborative authoring [321, 322]. Additionally, techniques that visualise synchronous collaboration sessions as artefacts for asynchronous follow-up work could prolong the value of collaborative sessions [323, 324, 325, 326].

4.8 Limitations

While our study provides valuable early insights into synchronous and asynchronous collaborative authoring for site-specific outdoor AR experiences, several limitations should be noted.

First, there was variability between the two locations used in the study, which differed in environmental characteristics such as foot traffic, noise levels, and spatial layout. Although we counterbalanced the assignment of locations to conditions, maintained a consistent task across locations, and did not identify a major impact of location in our analyses, these differences may still have influenced participants' experiences and performance.

Second, the sample size of our user study was small, with 32 participants forming 16 pairs. This limited the power of our statistical analyses to detect certain effects. For instance, while qualitative data indicated that some novice participants reported increased task load, this observation may not have been fully reflected in the quantitative data due to the small sample and variability in participants' experience levels. These findings highlight the need for future

research with larger sample sizes and specific target groups to examine the role of prior experience and explore strategies for lowering barriers to AR authoring, in line with recommendations by Ashtari et al. [269].

Third, the novelty of the system compared to current methods may have led to participant response bias [327], potentially influencing measures such as engagement. Longitudinal studies with extended and repeated exposure could help isolate these effects.

Finally, our study was limited to a specific type of site-specific AR experience and a particular set of tasks. Consequently, the findings may not generalise to other types and scales of AR application development, such as those involving larger teams or multiple co-located users collaborating on extensive projects. Future research could examine a broader range of AR authoring scenarios that vary in task type and group size, including those with multiple in-situ and ex-situ users engaged in synchronous distributed authoring (e.g., guided tours over larger areas) and environmental exploration.

4.9 Conclusion

This chapter examined site-specific outdoor AR authoring as a setting where a persistent gap exists between where experiences are authored and where they are ultimately encountered. In current workflows, authors often work ex situ using proxy representations of a target site, yet many real-world constraints that shape the end-user experience emerge only in situ and change over time. Based on the findings of our formative study, we developed CoCreatAR to support asymmetric collaboration around these representational gaps, and we operationalised RQ2 and RQ3 by analysing breakdowns in existing strategies and evaluating task-driven augmentations for incorporating meaningful context during authoring.

In response to RQ2, our findings show that environmental representations commonly used for ex-situ authoring remain insufficient for outdoor tasks that depend on temporal and situational constraints. While geospatial tools can provide high-fidelity geometric *location meshes*, these models are typically static and can present a fractured or incomplete version of reality that omits access

restrictions, pedestrian and vehicle flow, lighting variation across times of day, environmental noise, and physical changes at the site. To explore RQ3, we investigated a collaborative approach that treats this authoring gap as a coordination problem between an ex-situ author and an in-situ partner who can observe and probe the target setting. CoCreatAR enabled collaborators to selectively repair and enrich the *location mesh* during synchronous sessions through environmental context imagery and spatial interaction cues, including coarse meshes that extend missing geometry, 3D snapshots that capture local detail around small referents, and anchored drawings and shared 3D cursors that make constraints and targets explicit for joint interpretation. In our evaluation, participants used these features to resolve ambiguity at moments when static geometry did not support placement, feasibility judgements, or verification, indicating how task-driven augmentation can support grounding and coordination across asymmetries of access and control.

Beyond these dissertation-level research questions, the user study tested four workflow-specific hypotheses about synchronous collaboration with CoCreatAR relative to an asynchronous baseline. We found support for H4.3 and H4.4, with significantly higher self-reported engagement for both roles in the Sync condition and significantly higher ex-situ confidence in the authored outcome. In contrast, task load results were inconclusive, providing no support for H4.1 and H4.2 in the quantitative analyses. Qualitative findings nevertheless indicated that synchronous work introduced multitasking demands and coordination overheads, particularly for in-situ users navigating busy outdoor settings and for participants with limited familiarity with the tools. These results motivate future CMR authoring systems that retain rapid mechanisms for contextual repair and verification, while providing stronger support for information layering, pacing, and hybrid workflows that combine synchronous and asynchronous work.

Chapter 5

SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending

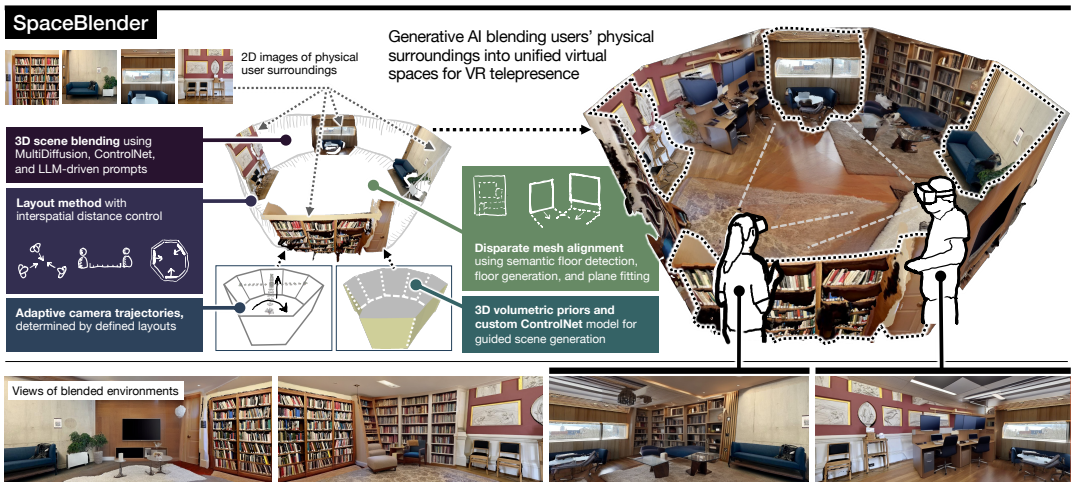


Figure 5.1 Overview of SpaceBlender, a pipeline that extends state-of-the-art generative AI models to blend users' physical surroundings into unified virtual environments for VR telepresence.

A fundamental challenge in distributed CMR is the inherent *environmental context asymmetry* that arises when collaborators operate from distinct physical settings. As discussed in Chapter 1, distributed collaboration requires coordination of activity as well as coordination of the spaces in which that activity unfolds. Many CMR systems address this issue by privileging one collaborator's physical context over others. A common strategy, examined in Chapter 3, uses a unidirectional configuration in which a remote VR collaborator is virtually

relocated into a reconstruction of a local AR user's physical environment [60, 62, 2, 4]. While this approach supports task-centric assistance and grounding within the host's domain, it provides limited support for collaboration that depends on resources or contextual cues distributed across multiple environments. In such cases, environmental context asymmetry becomes problematic, since only one participant's context can be represented, interacted with, and used for grounding. In this chapter, we investigate an alternative approach that aims to reduce environmental context asymmetry by synthesising shared virtual environments that blend multiple collaborators' physical surroundings. This work was published at *ACM UIST 2024* [12][1]

5.1 Context and Motivation

As discussed in Section 2.5, prior work proposed many approaches for constructing shared space across dissimilar sites, as well as taxonomies for classifying these approaches [69, 70, 65]. Boundary-based models treat coupling as a configurable interface that controls what perceptual and interactional information passes between spaces, without requiring the environments themselves to be visually reconciled [33, 35, 328] (Section 2.5.2). Volumetric coupling approaches instead establish correspondence by selecting task-relevant regions and anchors, then applying geometric and coordinate-space transformations such as extraction, warping, inclusion, alignment, and action retargeting [69, 70, 65, 217, 153, 223, 224, 213, 225, 226] (Section 2.5.3). This class of approaches typically reduces environments to symbolic, segmented, or anchor-based representations, which lowers the fidelity of the physical context conveyed. This leaves a gap for transformation approaches that preserve visually rich physical context while supporting collaboration across dissimilar environments, without requiring ex-

[1]The majority of this research was conducted during an internship at Microsoft Research as part of the *Transcendence* project. The work was carried out in collaboration with a fellow intern, Shwetha Rajaram, and resulted in two co-first-authored publications at ACM UIST 2024 [12, 13]. Each of us led a complementary project. Shwetha Rajaram focused on BlendScape [13] and I focused on SpaceBlender [12]. We both contributed to research framing, discussion, writing, and editing, and led the technical development and evaluation of our respective projects. Our mentors at Microsoft Research (Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andy Wilson) provided guidance throughout and contributed to writing and editing.

plicit boundary traversal. We address this gap by exploring *space blending*, a representational transformation that synthesises context-rich shared environments across multiple physical places while accounting for the rendering and consistency requirements of comfortable VR experiences.

Our design goal is to support mutual *place* awareness across sites, where each collaborator encounters their own surroundings alongside cues from other collaborators' settings within a unified environment. This framing follows a place-centric view in which collaborative action draws on locally meaningful situational cues that cannot be fully specified in advance, since participants appropriate what is available to establish reference, interpret activity, and coordinate conduct [47]. We are further motivated by prior work suggesting that bringing familiar environmental context into remote collaboration may strengthen information recall [329, 330, 331, 332] as well as feelings of comfort [38].

Recent progress in generative AI has introduced new computational possibilities for creating 2D and 3D content. Diffusion-based approaches have enabled low-effort creation of 3D objects [333, 334], panoramic images [335, 336, 337, 338], and complete 3D scenes [339, 340, 341, 342, 343, 344], often from minimal input such as text prompts and images. These developments are being integrated into commercial VR development workflows to accelerate content creation [345, 346] and enable end-user interaction and customisation [347, 348, 13], offering an alternative to manual content creation. In this work, we extend a recent generative 3D scene generation framework to automatically create CVEs for VR telepresence that blend multiple users' physical surroundings into a unified 3D space.

Building on existing generative models to achieve this goal revealed two primary challenges. First, many models aim to produce fully synthetic output that is not grounded in real-world spaces [341, 349]. Models that attempt grounding typically require input beyond text and images, such as camera trajectories, and they generally only support grounding in a single space [339, 342, 338], which is insufficient for blending distinct physical contexts. Second, generated 3D meshes are often not optimised for use as VR environments. During development, we ob-

served usability issues including non-navigable pathways, distracting visual and geometric artefacts, and perceptually incoherent spaces that reduced comfort during collaboration (see Figure 5.3).

To address these challenges, we created SpaceBlender, a pipeline that leverages and extends state-of-the-art generative AI techniques to blend users' physical surroundings into unified virtual spaces suitable for VR telepresence. SpaceBlender transforms user-provided 2D images of distinct spaces into context-rich 3D environments through a multi-stage process. First, it transforms 2D images into 3D meshes using depth estimation, depth alignment, and backprojection. It then applies a RANSAC-based alignment technique to align disparate meshes while maintaining a uniform floor level. Finally, it performs diffusion-based space completion guided by geometric priors and text prompts produced by an large language model (LLM) acting as an interior architect.

As a preliminary assessment of the suitability of SpaceBlender environments for telepresence, we conducted a comparative study with 20 participants. In pairs, participants performed a VR-based affinity diagramming task in three virtual environments: *(1)* Generic3D, a generic low-poly room; *(2)* an environment generated with Text2Room [339]; and *(3)* a SpaceBlender environment that incorporated images of familiar physical locations provided by participants. Participants reported higher physical comfort and navigability in the Generic3D and SpaceBlender compared with Text2Room. Participants attributed these differences to greater consistency in room geometry. Some participants used recognisable environmental features in the SpaceBlender as reference points during the clustering task. Participants also described potential use cases for familiar or personal contextual details in remote collaboration, and they requested improvements to visual quality and realism to better support these use cases.

5.2 Related Work: 3D Scene Generation

While Section 2.7.2 provides a comprehensive overview of prior work on environmental representations in CMR systems, this section provides additional context on recent advancements in computational 3D scene generation that underpin

the technical contributions of SpaceBlender.

The domain of computational 3D scene generation has been substantially advanced by recent developments in generative methods. Unlike traditional procedural generation techniques that depend on predefined rules and asset libraries to assemble 3D environments [350, 351, 352, 353], recent approaches can generate novel spaces via learned generative models. These approaches use a range of scene representations, including well-known explicit formats such as 2D panoramic views [335, 336, 337] and meshes [339, 354, 341, 342], as well as recent implicit representations such as NeRF [355, 343], 3DGS [356], and signed distance functions (SDFs) [357].

At the time of this research, most recent scene generation methods operate primarily in the 2D image domain due to the wide availability of image-caption datasets and state-of-the-art generation models trained on these [358]. By leveraging depth estimation models [359, 360], 2D images can be transformed into 3D representations by predicting a depth value for each pixel and backprojecting these values into 3D space. Using models that support panoramic images, combined with image upscaling techniques that mitigate the limited output resolution of image generation models (e.g., [361, 362]), systems such as Skybox AI [363] and LDM3D [336, 337] can generate high-quality 360° skyboxes from user-provided text prompts. However, such techniques yield environments that do not remain spatially consistent during navigation due to their single-view generation approach, which makes them less suitable for VR telepresence.

Other image-based scene generation methods address this limitation by generating multi-view image sequences [339, 338, 349]. One such method is Text2Room, which, given a camera trajectory with matching text prompts and an optional starting image, iteratively expands a 3D mesh through text-conditioned inpainting of 2D rendered views of the mesh [339]. In each step, this method estimates a depth image for a rendered view, aligns it with known depth values, and backprojects the resulting depth values to integrate them into the mesh. This process is repeated for each viewpoint in the camera trajectory. Finally, an

adaptive trajectory is defined at runtime based on a set of viewpoints with the highest number of missing pixels, which is used to complete remaining holes in the mesh. While this approach produces high-quality scenes with improved multi-view consistency during navigation, the generated scenes still exhibit structural and visual irregularities, as well as contextual repetitions (e.g., multiple bathtubs in a single bathroom), which limit the navigability and realism required for telepresence (see Figure 5.3).

To improve geometric consistency and controllability, *MVDiffusion* [338] as well as recent work conducted in parallel to our research such as *Ctrl-Room* [341] and *ControlRoom3D* [342] take additional inputs that act as spatial constraints, such as untextured 3D meshes [338, 343] or semantic scene maps [342, 341]. However, in typical camera-based telepresence systems, such priors are commonly unavailable. Furthermore, like Text2Room, these systems cannot process multiple disparate images. SpaceBlender advances these concepts by accepting multi-image inputs from distinct spaces associated with collaborating users, while leveraging contextual inferences from a vision-language model (VLM) [364], an LLM [365], and a semantic segmentation model [366] to autonomously generate a suitable layout, geometric prior, and text prompts for environment generation directly from user-provided image frames.

Since conducting the research presented in this chapter, several works have further advanced 3D scene generation along dimensions that are directly relevant to our problem setting. First, video diffusion models have been increasingly adopted for synthesising long-range camera-controllable videos that remain consistent across views [367, 368]. Second, new methods have improved controllability and layout specification through the use of scene graphs and higher-level semantic constraints, enabling text-driven and graph-driven manipulation of object placement and room layout [369, 370, 371, 372]. Finally, industry has begun to operationalise such generative capabilities at scale, with closed-source foundational models such as World Labs' *Marble* [373] targeting generative 3D world construction using 3DGS representations for applications in simulation, enter-

tainment, and autonomy. We discuss how these developments can be integrated into future iterations of SpaceBlender in Section 5.6.

5.3 SpaceBlender System

This section details the SpaceBlender pipeline, which integrates images of the physical context of multiple users into cohesive virtual environments. SpaceBlender builds on the Text2Room pipeline due to its extensibility through the use of off-the-shelf 2D image models, its transparent iterative generation process, and its ability to initialise generation from single 2D input images, which are commonly available in telepresence scenarios.

We begin by outlining the system requirements in Section 5.3.1, followed by an overview of the proposed SpaceBlender pipeline in Section 5.3.2. We then provide a detailed description of its two stages in Sections 5.3.3 and 5.3.4.

5.3.1 Requirements

The assumptions, implementation details, and limitations of the Text2Room framework pose substantial barriers to generating unified spaces from disparate input images for use in VR. Below, we articulate the requirements of SpaceBlender, outline the associated challenges, and describe our approach to addressing them.

Requirement 1: Enabling multiple disparate spatial inputs from diverse perspectives and locations.

Challenge: Text2Room accepts at most one input image and does not register the resulting mesh in a global coordinate space, while SpaceBlender must process and align multiple images captured from different viewpoints to support scene blending.

Approach: We introduce a floor plane alignment process that identifies the floor of each mesh by backprojecting semantic labels into 3D space, which are then used for global alignment (Section 5.3.3.2). When no floor is visible, we propose a technique to synthesise floor sections prior to alignment. The aligned meshes

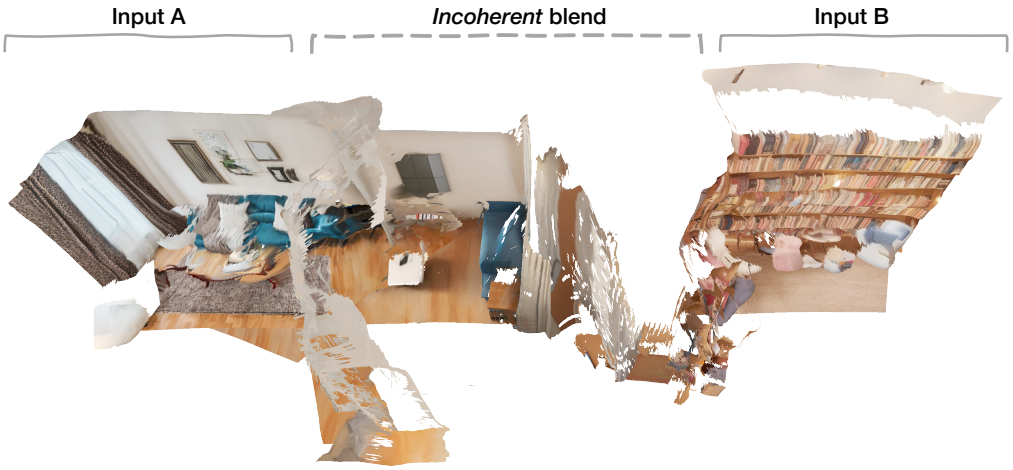


Figure 5.2 A bird's-eye view of two meshes that failed to blend due to a lack of geometric guidance and insufficient context during the iterative mesh completion process.

are then arranged using a parameter-based layout technique (Section 5.3.3.3).

Requirement 2: Enabling coherent scene blending for realistic and context-rich spaces.

Challenge: Text2Room uses low-resolution square images in its iterative view inpainting process. This limits the reference frame of the inpainting model, leading to mismatched mesh segments between disparate meshes and resulting in harsh geometric and visual boundaries and artefacts, as shown in Figure 5.2.

Approach: We incorporate MultiDiffusion-based [361] image inpainting to complete wider images, expanding the model's contextual window and enabling smoother blends.

Challenge: Text2Room's lack of control over room shape causes issues when blending disparate spaces, as shown in Figure 5.2.

Approach: We introduce a geometric prior defined as the convex hull of the disparate meshes (Section 5.3.3.4) and a custom ControlNet model for guided scene generation (Section 5.3.4.1).

Requirement 3: Enabling users to create blended environments without extensive manual configuration.

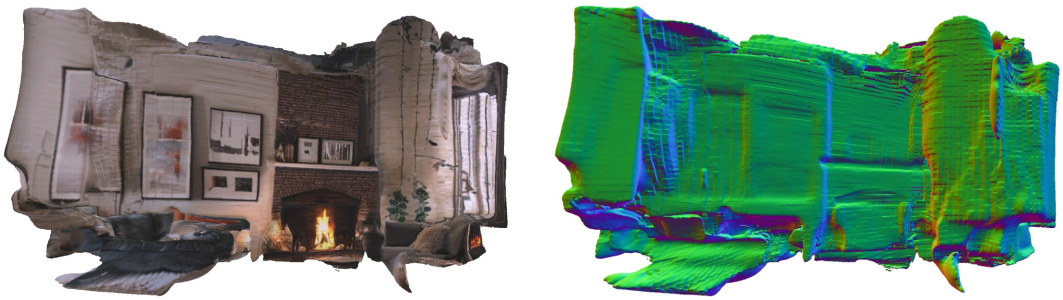


Figure 5.3 An outer view of a room generated by Text2Room (left: colour texture, right: normal map). The room has an irregular shape with non-navigable pathways and visual artefacts that can detract from user comfort in VR.

Challenge: The manual configuration required by Text2Room for trajectory and prompt adjustments is infeasible for SpaceBlender's intended application context of VR telepresence for end-users. Defining trajectories and prompts is time-consuming and requires specialised expertise.

Approach: We introduce contextually adaptive prompt inference based on a VLM and an LLM (Section 5.3.3.5), as well as adaptive trajectories (Section 5.3.4.3), enabling cohesive and automated space blending.

Requirement 4: Supporting core VR usability requirements for end-users, including comfortable navigation and viewing.

Challenge: The depth estimator used by Text2Room commonly produces slanted and discontinuous floors and walls (see Figure 5.3), which is problematic for VR navigation and spatial orientation.

Approach: To achieve a consistent room structure, SpaceBlender performs semantic segmentation on inpainted images and copies depth values for wall, floor, and ceiling pixels from the rendered depth image of the geometric prior. These values also inform depth completion for the remaining pixels.

5.3.2 System Overview

The SpaceBlender pipeline processes n input images to produce a 3D mesh that integrates the context of each image into a cohesive blended environment. The pipeline has two main stages. The first runs once per generation, while the

second is iterative, similar to TEXT2ROOM. Below, we provide a brief overview of these stages, with detailed descriptions in the following subsections.

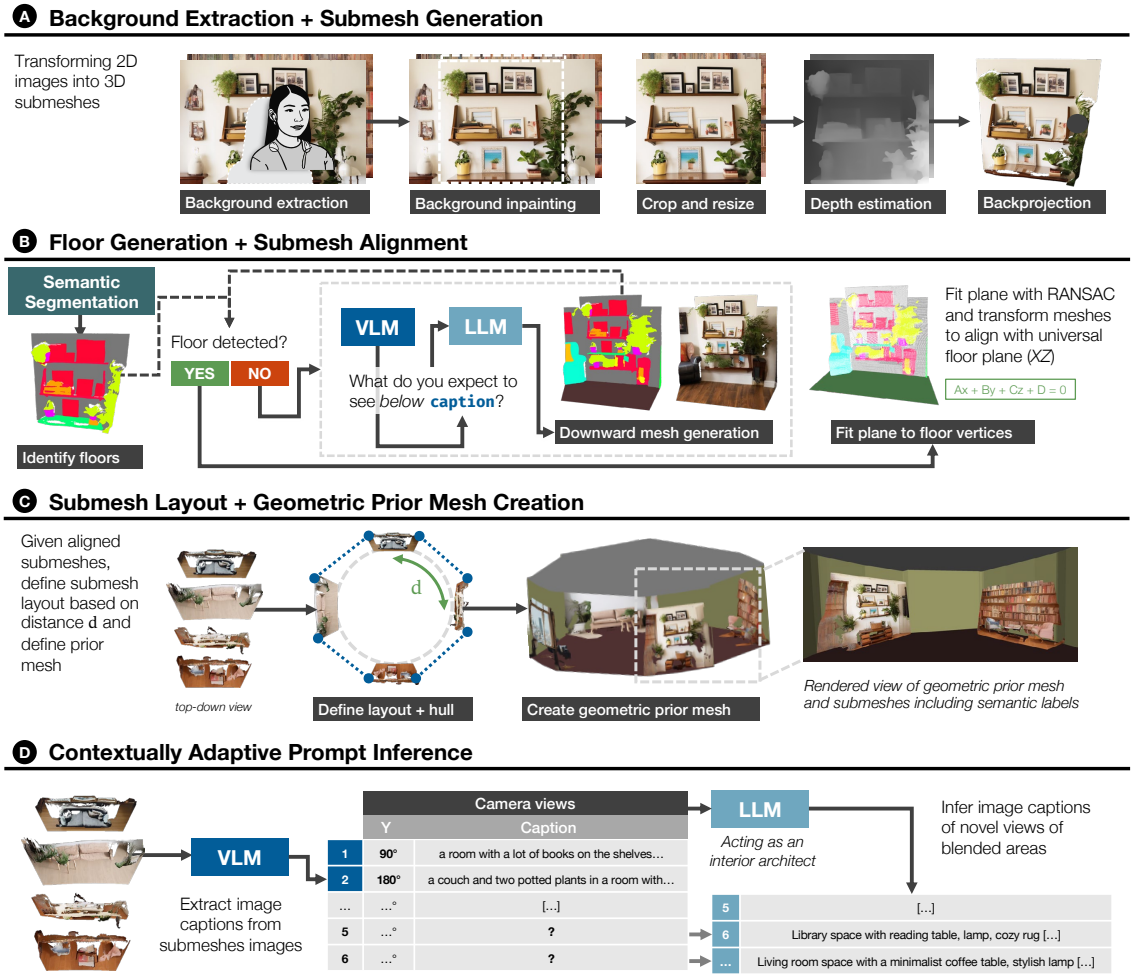


Figure 5.4 Overview of *Stage 1* components as described in Section 5.3.3.

In *Stage 1*, each input image is first preprocessed, after which depth values are estimated for each pixel and backprojected to create a 3D mesh (Section 5.3.3.1). We refer to the resulting n meshes as *submeshes* throughout the remainder of this dissertation. Next, the submeshes are aligned to a common floor plane using a RANSAC-based method applied to floor vertices identified by a semantic segmentation model (Section 5.3.3.2), optionally including a floor generation step when no floor is visible in the image. The aligned submeshes are then positioned using a parameter-based layout technique (Section 5.3.3.3), after which a geometric prior mesh is created to define the shape of the blended space

(Section 5.3.3.4). Finally, text prompts describing the blended regions (i.e., the empty space between submeshes) are generated with an LLM based on captions inferred by a VLM (Section 5.3.3.5).

In *Stage 2*, the submeshes are blended through a process that repeatedly inpaints and integrates 2D rendered views of the mesh. For each iteration, geometric image priors are rendered based on the submesh layout defined in *Stage 1* to guide the shape of the space (Section 5.3.4.1). These are combined with the text prompts generated in *Stage 1* to guide the content and appearance of the space (Section 5.3.4.2). Once blending completes, an adaptive mesh completion trajectory is followed to fill remaining gaps in the environment (Section 5.3.4.3).

Implementation Like the original Text2Room implementation, we use Stable Diffusion 1.5 [358] for image generation and inpainting and IronDepth [360] for depth estimation and inpainting. SpaceBlender uses the BLIP-2 [364] VLM, the GPT-4 [365] LLM, and the OneFormer [366] semantic segmentation model. We decoupled the image inpainting process by using a local API endpoint of the A1111 WebUI server[2] to provide sufficient GPU memory for the various models in our pipeline. The server and pipeline run on separate machines, each equipped with an NVIDIA RTX 4090 GPU in our local setup. Generating a SpaceBlender environment with this configuration takes approximately 55–60 minutes.

5.3.3 *Stage 1*: From 2D Images to 3D Layout

The first stage establishes the spatial structure of the blended environment. It begins with image preprocessing and depth estimation, converting 2D images into 3D submeshes. These submeshes are aligned to a common floor plane and arranged in a circle, after which a geometric prior is defined. Finally, prompts for the blended regions are generated by an LLM.

5.3.3.1 From 2D Images to 3D Submeshes

In this step, shown in Figure 5.4A, the n input images are projected into 3D space. First, a semantic segmentation model is used to detect people in each input

[2] https://github.com/AUTOMATIC1111/stable-diffusion-webui

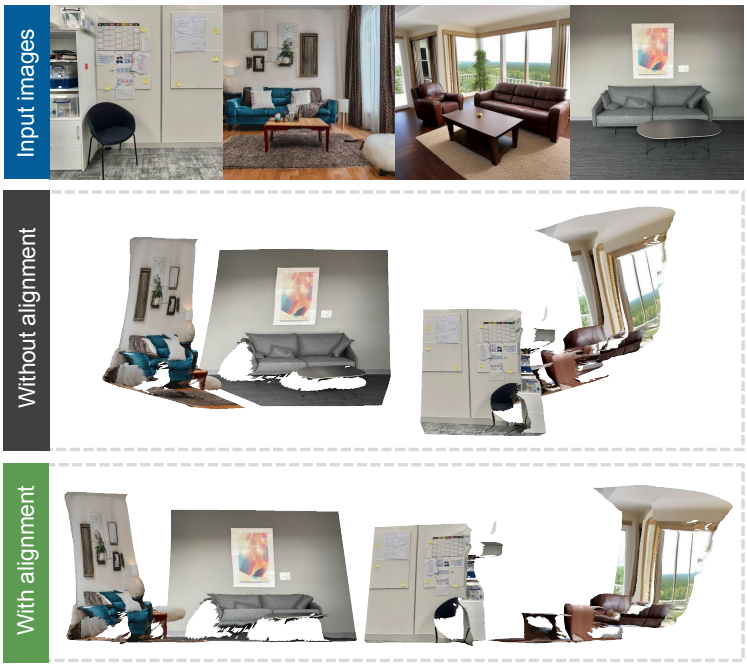


Figure 5.5 Comparison between unaligned submeshes and submeshes aligned with our semantic floor alignment technique. The unaligned spaces have floors at different levels and inclines that can be jarring to navigate.

image. If a person is detected, that region is removed and inpainted using a prompt inferred by the VLM. The resulting image is then cropped to 512×512 pixels to ensure compatibility with subsequent models in the pipeline. Next, using TEXT2ROOM's existing functionality, a 3D mesh is created. This involves using the depth estimation model to estimate depth values, aligning them with known depth values, and backprojecting them into 3D space along with the colours of the input image.

5.3.3.2 Submesh Alignment

We introduce a floor plane alignment technique to reconcile differing perspectives in input images, ensuring the spatial consistency required for scene blending (Figure 5.4B). First, labels for floor-like objects (e.g., floor, carpet) within each submesh are obtained using a semantic segmentation map derived from the input image. These semantic labels are then backprojected into 3D space, replacing the submesh's RGB colours with semantic label values. To handle discrepancies between the depth estimation and semantic segmentation outputs,

floor vertices more than 0.3 m above or below the median Y-coordinate of the floor-like vertices are excluded.

Next, RANSAC is used to identify a plane corresponding to these floor-like vertices. To ensure that a candidate plane corresponds to a floor, we apply three heuristics: (1) the plane's orientation must be within 45° of the target plane normal; (2) the normal vector must have a positive Y-component; and (3) the extent of the inlier points in the X and Z axes must be at least 0.5 m. After selecting the best floor plane candidate, we rotate the mesh to align the plane's normal with the Y-axis. Next, we translate the floor to $Y = 0$ and set the minimum Z-coordinate to 0. Figure 5.5 shows an example of submeshes aligned to a common floor plane.

Floor Generation If input images lack a visible floor (e.g., only a wall is shown), which prevents floor plane fitting, we use a generative technique to extend the submesh downwards to create a floor for alignment. This method follows a five-step trajectory that interpolates between the following movements: gradually looking downwards (from −5° to −30°), moving backwards (from 1 m to 1.5 m), and moving upwards (from 0.3 m to 1 m) relative to the initial camera view. For each step, we use a floor description generated by the LLM based on the image caption produced by the VLM. If floor generation fails after ten attempts, the submesh remains unaligned for the remainder of the generation process.

5.3.3.3 Submesh Layout

With submeshes aligned to a common floor plane, each submesh is transformed to form a layout resembling an open space (**Figure 5.4C**). This is achieved by positioning the front side (i.e., the side facing the camera used to capture the input image) of each submesh on the perimeter of a circle as seen from a top-down view. The diameter of this circle is determined by a configurable parameter d, which controls the distance between the submeshes. Each submesh faces the centre of the unified space, ensuring a clear line of sight between all submeshes. This design choice was made in consideration of the importance of mutual

awareness in collaborative scenarios [51, 83].

5.3.3.4 Geometric Prior Mesh

Given the aligned submeshes, a geometric prior mesh is generated to define the shape of the blended space (**Figure 5.4C**). This involves defining a mesh based on the convex hull of the submesh layout, with faces assigned to represent the floor, walls, and ceiling. The height of this mesh is set to the height of the tallest submesh, or 2.5 m if none is taller. The floor, ceiling, and wall faces are coloured based on their respective semantic label colours from the ADE20K dataset [374]. This mesh is used to render prior images for the iterative image inpainting process, as described in Section 5.3.4.

As the convex hull forms straight walls between submeshes, the number of submeshes and their shapes directly affect the overall shape of the geometric prior, and consequently the blended space. For example, four submeshes with straight walls create an octagon-like shape, while four submeshes with pronounced corners result in a square-like shape. A visual explanation and examples of spaces with different numbers of input images and submesh shapes are available in the Appendix (Figures D.3 and D.4).

5.3.3.5 Contextually Adaptive Prompt Inference

In preparation for the iterative mesh generation process in *Stage 2*, text prompts are generated to describe the intended contents of the blended regions **Figure 5.4D**. This begins by obtaining image descriptions for each submesh using the VLM, along with a rotation value indicating each submesh's relative orientation from a top-down view. This information is then passed to an LLM instructed to act as a creative interior architect and photographer who designs spaces with diverse contexts and appearances while avoiding repetitive objects. Based on the rotation values and image descriptions, the LLM returns a set of new image descriptions paired with rotation values corresponding to the regions to be blended (i.e., the empty regions between submeshes). The full system prompt for this LLM is provided in the Appendix (Section D.1).

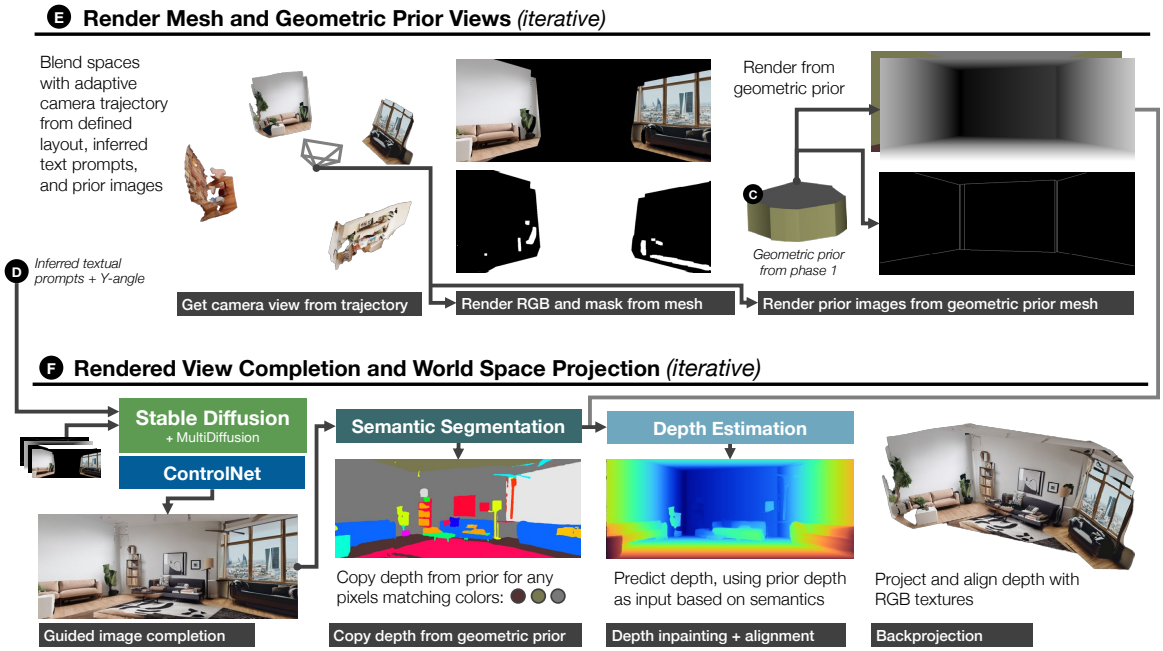


Figure 5.6 Overview of *Stage 2* components as described in Section 5.3.4.

5.3.4 *Stage 2*: Iterative Blending Guided by Geometric Priors and Contextual Prompts

Using the submesh layout, geometric prior mesh, and text prompts from Stage 1, this stage iteratively blends the disparate submeshes into a unified environment.

5.3.4.1 Room Shape Guidance with Geometric Prior Images

To enable geometrically coherent space blending, SpaceBlender uses a set of prior images to guide the iterative text-conditioned image inpainting step of the mesh blending and completion process via ControlNet [375] (**Figure 5.6E**). Each time the process renders a view of the mesh for inpainting, it also renders a set of prior images from the same camera viewpoint based on the geometric prior mesh aligned with the submesh layout. SpaceBlender generates three types of prior images, each with a distinct effect on the inpainting output:

- **Depth Prior:** This prior can act as a strong constraint that encourages spaces with shapes and contents consistent with the geometric prior. It is defined by rendering the relative depth for a specified view of the geometric prior mesh.

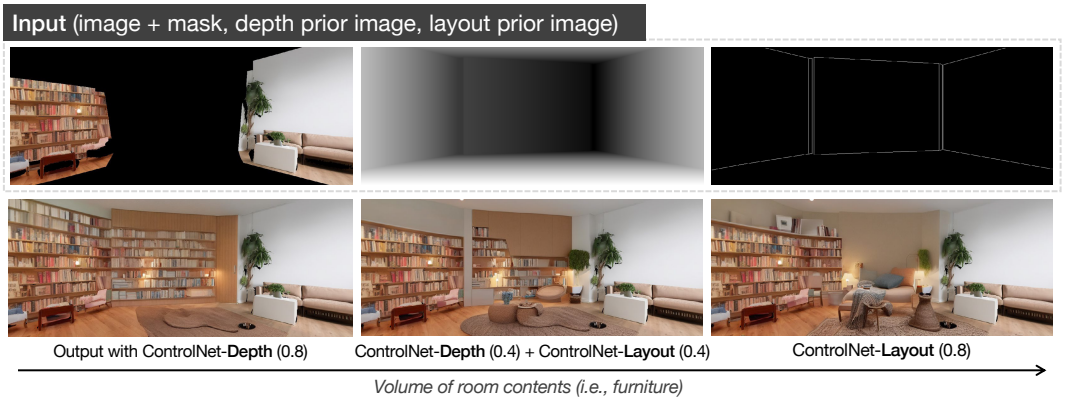


Figure 5.7 Comparison of outputs generated with varying weights of ControlNet depth and layout models, showing how the prior influences the output (generated with a fixed seed). Top: inputs including the input image, depth prior image, and layout prior image rendered from the geometric prior. Bottom: results for different weights are indicated in parentheses.

- **Layout Prior:** This prior constrains the shape of the space while allowing the content (e.g., furniture) to be generated more freely. It is defined by computing depth gradients using the Sobel operator to form surface normals based on the depth prior. The magnitude of these normals is then computed and processed with Canny edge detection to produce an image that outlines the geometric prior mesh, with white edges corresponding to walls, floor, and ceiling on a black background.

- **Semantic Prior:** This prior can act as an additional strong constraint to guide the semantic content of inpainted views. The current geometric prior mesh in SpaceBlender defines semantic labels for walls, floor, and ceiling, which makes it suitable as room-layout composition guidance when an empty open space is desired.

These prior images can be stacked and combined using Multi-ControlNet[3], which allows the influence of each prior on the output to be adjusted. For example, using only the layout prior guides the model to generate a space with a specific room structure while allowing room contents (e.g., furniture) to be generated freely. A depth prior can be added to guide the inpainting model

[3] https://github.com/Mikubill/sd-webui-controlnet#multi-controlnet

towards generating objects that are commonly positioned near walls (e.g., sofas and bookshelves). Figure 5.7 illustrates the influence of depth and layout priors on the output, with additional examples shown in the Appendix (Figure D.1).

While the depth and semantic priors are used with pre-trained ControlNet models[4], the layout prior is used with a custom ControlNet model, `ControlNet-Layout`. We describe the training process for this model below.

Training ControlNet-Layout We trained `ControlNet-Layout` on a dataset of 13,182 images. Instead of using an existing dataset, we constructed our own by using the pre-trained ControlNet segmentation model[4] to generate images from semantic maps inferred from SUN-RGBD [376] and LSUN [377, 378], resized to 512×512 pixels. This process was repeated several times with unique seeds to improve image quality and diversify the dataset by generating multiple images per semantic map. Training was initialised with the weights of the pre-trained M-LSD model of ControlNet[4] and used a learning rate of 1×10^{-5} and a batch size of 4. Training was halted after one epoch due to satisfactory performance on the validation set. Additional examples of `ControlNet-Layout` output are provided in the Appendix (Figure D.2), including a comparison to a related model from parallel work [379].

5.3.4.2 Iterative Space Blending

Shown in **Figure 5.6F**, this step iteratively unifies disparate submeshes according to the geometric prior images and prompts defined in Section 5.3.3.5. To support SpaceBlender's blending capabilities, we broaden the context window of the inpainting model by increasing image resolution from 512×512 (as used by Text2Room) to 512×1280 while maintaining a rendering camera field of view of 55°. This is enabled by an A1111 WebUI plugin implementation of MultiDiffusion[5] [361].

By increasing the width of images generated during blending, the inpainting model receives a broader environmental reference frame, which allows it to

[4] `https://huggingface.co/lllyasviel/ControlNet`
[5] `https://github.com/pkuliyi2015/multidiffusion-upscaler-for-automatic1111`

blend the space between neighbouring submeshes in a single inpainting step (see Figure 5.7). This yields higher-fidelity results compared to stepwise blending, which can produce harsh boundaries and artefacts such as those shown in Figure 5.2.

However, due to the circular layout method, when SpaceBlender is given three or fewer input images, the distances between submeshes can prevent blending in a single step. In these cases, SpaceBlender uses LLM-generated prompts to create intermediate submeshes with an image generation model and integrates them into the layout to bridge gaps between submeshes and support blending. The rightmost example in the Appendix (Figure D.3) demonstrates this approach.

5.3.4.3 Mesh Completion Trajectory

The initial iterative blending process produces a mesh that integrates the submeshes horizontally, defining the blended space from a central perspective. At this stage, the floor and ceiling are largely incomplete, and the mesh contains gaps that require completion. To address this, we employ an additional set of camera trajectories.

First, interpolation-based trajectories are generated to cover missing sections of the floor and ceiling. Next, trajectories are defined for each submesh. These paths interpolate the camera position and rotation, starting from a central position within the unified space and initially directed towards a specific submesh. Each interpolation trajectory ends at the submesh centre, facing the left or right adjacent submesh. Throughout this process, the text prompt passed to the inpainting model is set to the prompt for the blended region that is most closely aligned with the camera view.

Finally, an additional trajectory simulates a user looking around the unified space from the centre point of that user's submesh, ensuring that the mesh accounts for gaps visible from typical user perspectives. To approximate natural gaze variation, we introduce a controlled degree of randomness into these viewpoints. Once all trajectories are complete, the blended space is ready for use in a

VR telepresence system.

5.4 Preliminary User Study

Our preliminary user study explored the effects of *space blending* in the context of a collaborative affinity diagramming task within a VR telepresence environment. The study used a within-subjects design with three conditions that emphasised different visual and geometric qualities of virtual environments. Through the selection of conditions below, we sought to explore variation along the dimensions of environmental visual and geometric complexity, fidelity, and familiarity, and to examine their impacts on participant behaviour and strategies.

- **GENERIC3D** served as a baseline representing low-poly environments commonly used in current social VR platforms such as Meta Horizon and Rec Room. This space was designed with 3D models from the public domain[6].
- **TEXT2ROOM** served as a baseline representing an environment generated with a state-of-the-art 3D scene generation method, containing salient landmarks that contrast with the simpler landmarks of the GENERIC3D condition. This environment was produced using the TEXT2ROOM framework [339], which we extended to develop SPACEBLENDER.
- **SPACEBLENDER** represented our pipeline with participant-provided input images of spaces familiar to them.



Figure 5.8 Overview of the environments used for the GENERIC3D and TEXT2ROOM conditions.

The order of conditions was counterbalanced. The GENERIC3D and TEXT2ROOM environments were consistent across pairs and are shown in Figure 5.8. For the TEXT2ROOM condition, we used the pipeline's public source

[6] https://kenney.nl/assets/furniture-kit

code and trajectory files to generate twelve environments and then selected the best one based on a subjective visual assessment of geometric and visual quality. For the SpaceBlender condition, a new environment was generated for each pair to incorporate a familiar physical context for each participant, as shown in Figure 5.10. This involved collecting images via an online form sent to participants before the study, through which they could upload a photo of a familiar space (e.g., a library, cafe, living room, or desk). Participants who did not upload a photo could choose from seven photographs of spaces within our institution. Participants who reported that none of these spaces were familiar were excluded from the study. Uploaded images were cropped to a 1:1 aspect ratio, excluding personally identifiable content (e.g., portraits). The study was approved by the University College London Research Ethics Committee (Study ID UCL/CSREC/R/16).

5.4.1 Task

The task involved clustering virtual sticky notes with predefined text. Initially, each participant independently clustered twelve sticky notes from one of two datasets (*fruits* or *vegetables*) by colour (e.g., placing "banana" near "lemon"). After two minutes of individual work, participants were given three minutes to collaboratively reorganise their clusters into new groups.

5.4.2 Recruitment

We recruited 20 participants (10 pairs) through internal mailing lists. Participants (7 female, 13 male) had a mean age of 26 years (SD = 6.9) and were mainly students and professionals. Seven participants had backgrounds in computer graphics, and four in UX/HCI. All participants had used a VR HMD at least once, with eight using one monthly, and three using VR telepresence platforms monthly. Five pairs knew each other before the study. Participants received £15 gift cards as compensation.



Figure 5.9 Captures of participants manipulating sticky notes while represented by Ubiq avatars in the individual (left) and collaborative (right) task phases.

5.4.3 Implementation and Setup

The VR telepresence system for our study was built with Unity and the Ubiq framework [227], providing voice chat, networked objects, and low-poly floating-body avatars. Each participant used a Meta Quest 3 VR HMD tethered to a desktop computer with an NVIDIA RTX 4090 GPU in separate physical spaces. Controllers were used for navigation and interaction.

Two stacks of sticky notes were placed in the virtual environments. Participants could grab the top note by pressing and holding the grip button when their virtual hand was near the stack. Given the high complexity of meshes produced by generative models, sticky notes could be placed anywhere without physics constraints. The stacks were manually placed (e.g., on a table) within arm's reach of participant spawn points before the study.

5.4.4 Procedure

Upon arrival, the experimenter introduced the study structure and goals, including an explanation of how to use the VR equipment. Participants then read and signed an informed consent form and completed a pre-study questionnaire covering demographics and VR experience. Participants were guided to randomly assigned spaces and equipped with HMDs. The experimenter then joined the environment from a desktop computer in a separate space, explained the task, and guided participants through a test environment to familiarise them with the navigation controls and sticky note manipulation.

Participants were then teleported to opposite sides of an environment

matching the current condition. In the SpaceBlender condition, teleportation points were defined at the centre of the submesh generated from the image provided by the respective participant. Participants first clustered the sticky notes individually. After two minutes, the experimenter instructed them to continue the task collaboratively, combining their individual work for three additional minutes (Figure 5.9). After each condition, participants completed a post-task questionnaire.

After the final condition, participants completed a post-study questionnaire. The study concluded with the experimenter guiding participants to a shared physical space for a semi-structured interview, including a brief scenario walkthrough. This walkthrough included two scenarios: a collaborative study session and a cooking class with friends, as shown in the Appendix (Figure D.5). The interview questions are also shown in the Appendix (Section D.8).

5.4.5 Data Collection & Measures

The post-task questionnaire included questions measuring perceived spatial presence from an existing questionnaire [380], focusing on two dimensions: Self-Location, which assesses the sensation of being physically present within the virtual environment, and Possible Actions, which measures the perceived capability for interaction within that space. We also incorporated questions measuring perceived Copresence to assess participants' perceptions of sharing the virtual environment with others [82]. For each question, "Do not agree at all" was treated as 1, and "Fully agree" as 5.

Additionally, we included four custom questions to assess how various factors influenced task execution: (1) *layout*, (2) *visual quality*, (3) *level of familiarity*, and (4) *navigation controls*. Each question asked, "To what extent did [X] help or hinder you in conducting the task?", with X representing one of the factors. Participants responded using a 5-point Likert scale: *significantly hindered* (1), *slightly hindered*, *neither helped nor hindered*, *slightly helped*, and *significantly helped* (5). We averaged the items within each measure to obtain a single score per measure.

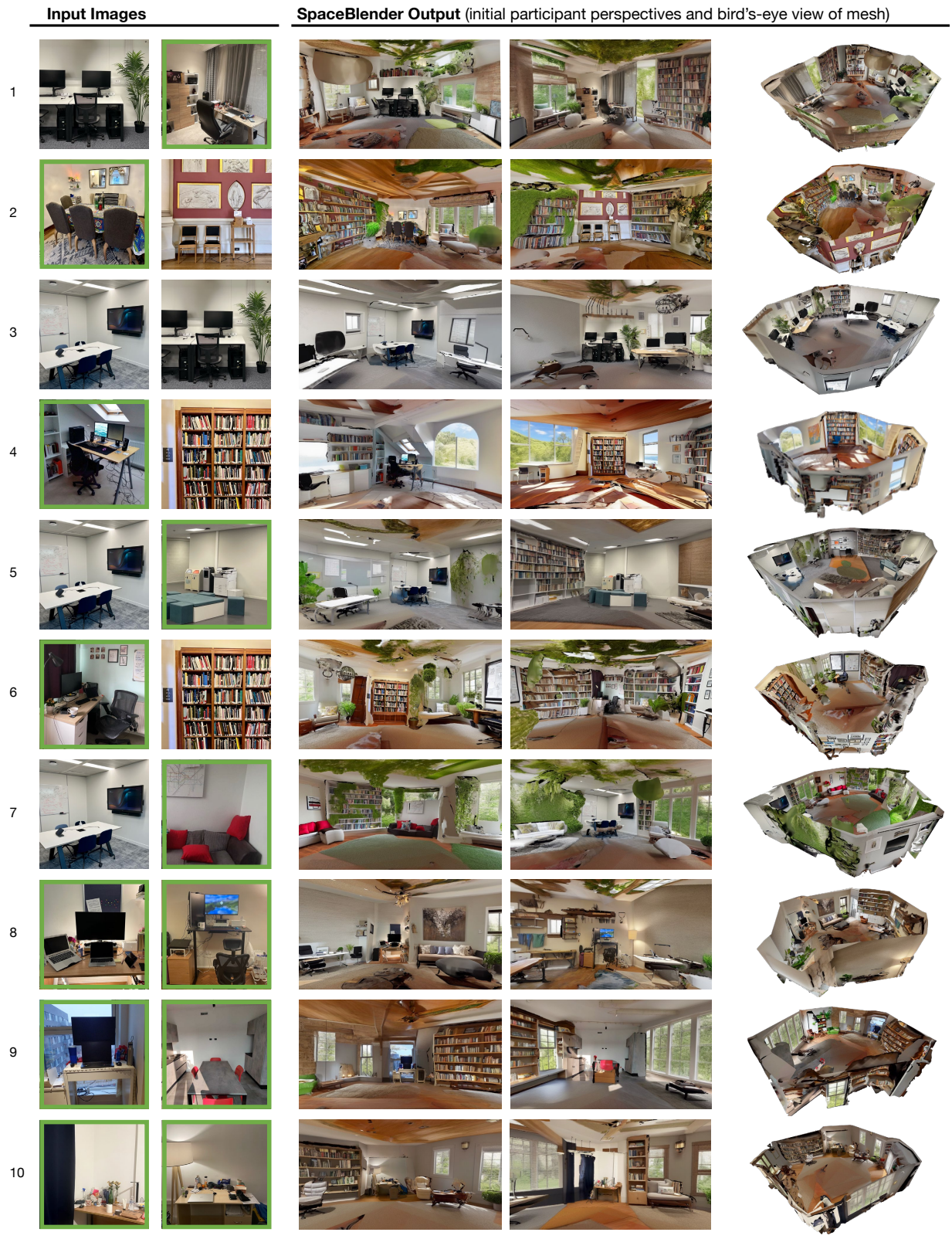


Figure 5.10 Overview of SpaceBlender meshes generated from input images uploaded (with green outline) or selected by participants (no outline).

5.5 Results

In this section, we present *(1)* a quantitative analysis of participants' self-reported measures; *(2)* qualitative themes describing the benefits and limitations of the three environments for the clustering task; and *(3)* participants' suggestions for future requirements and potential use cases of SpaceBlender environments.

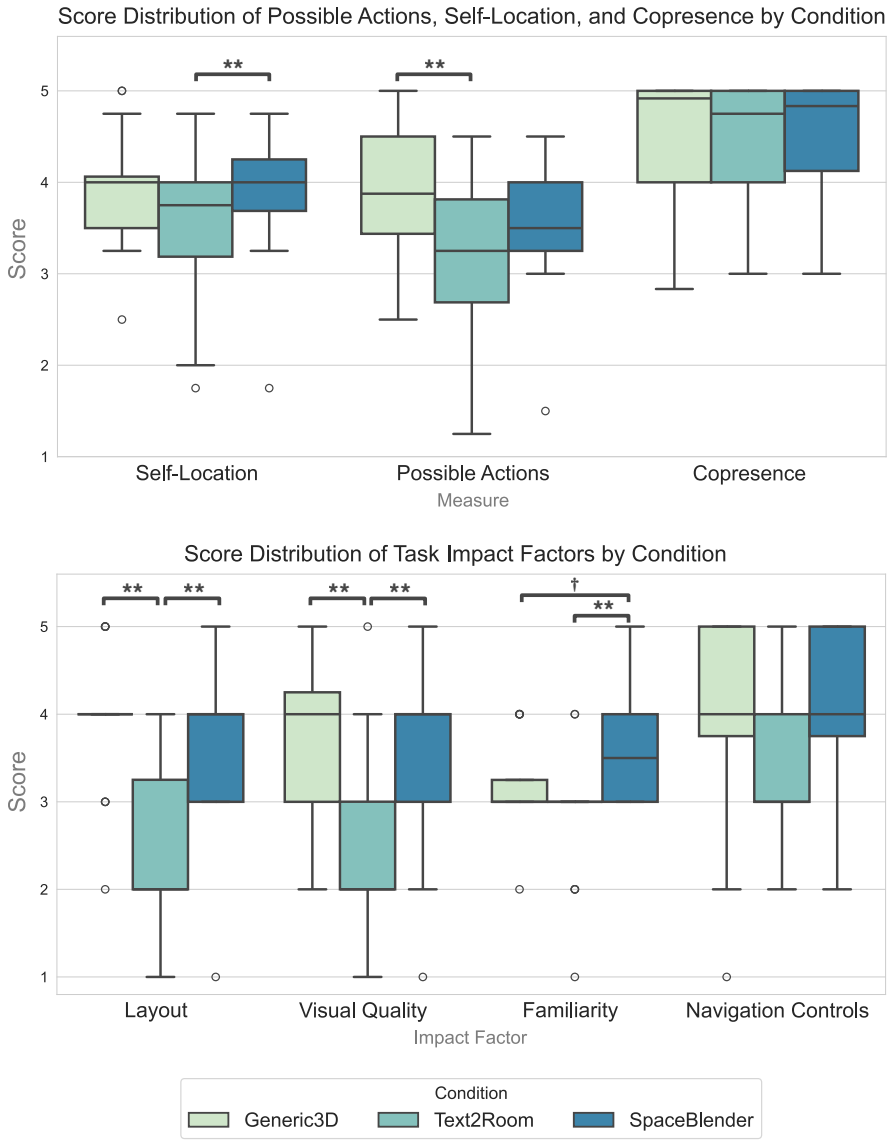


Figure 5.11 Plot of score distributions for Possible Actions, Self-Location, Copresence, and task impact factors. Levels of statistical significance: † for $p < 0.05$ (before Bonferroni correction), * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$ (after Bonferroni correction).

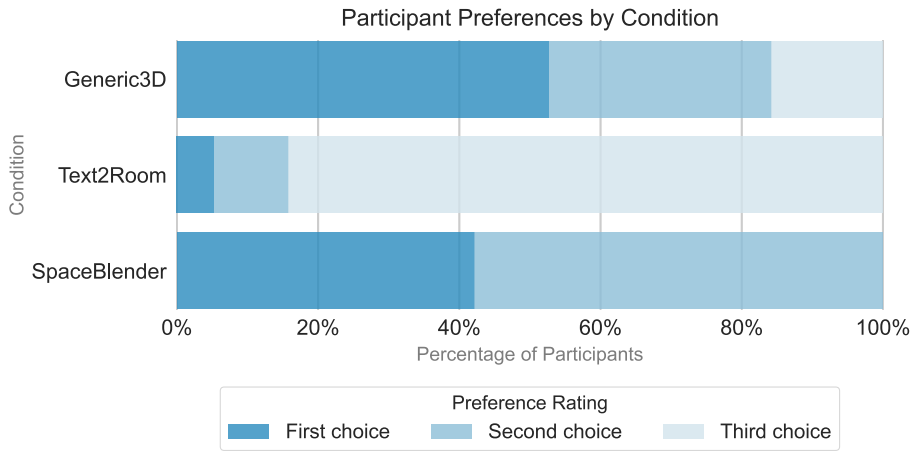


Figure 5.12 Plot of participant preferences per condition.

5.5.1 Self-Reported Questionnaire Results

To analyse the self-reported questionnaire data, we first applied the Friedman test to identify overall differences across conditions. We then conducted Wilcoxon signed-rank tests for post-hoc comparisons. Key findings are presented below, with additional results provided in the Appendix (Section D.9). Figure 5.11 shows the distributions of participants' scores for Possible Actions, Self-Location, Copresence, and task impact factors. Figure 5.12 shows participants' ranked preferences for using the Generic3D, Text2Room, and SpaceBlender environments. Most participants ranked the Generic3D environment as their first choice for completing the clustering task, followed closely by SpaceBlender, which received only first- or second-choice ratings.

Possible Actions A statistically significant difference in Possible Actions scores was found between Generic3D (M = 3.89, SD = 0.71) and Text2Room (M = 3.13, SD = 0.95) ($W = 5.0$, $p = 0.0021$), indicating a reduced perception of possible actions in the Text2Room environment.

Self-Location A statistically significant difference in Self-Location scores was found between Text2Room (M = 3.46, SD = 0.78) and SpaceBlender (M = 3.91, SD = 0.68) ($W = 0.0$, $p = 0.0039$), indicating a reduced sense of being physically present in the Text2Room environment.

Impact of Environmental Factors Among the environmental factors, we found statistically significant differences for the impact of *Layout*, *Visual Quality*, and *Familiarity*. For *Layout*, we found statistically significant differences between Generic3D (M = 3.92, SD = 0.68) and Text2Room (M = 3.20, SD = 0.75) ($W = 10.0$, $p = 0.0043$), as well as between Text2Room and SpaceBlender (M = 3.85, SD = 0.70) ($W = 4.0$, $p = 0.0036$). For *Visual Quality*, we found statistically significant differences between Generic3D (M = 4.00, SD = 0.00) and Text2Room (M = 3.17, SD = 0.75) ($W = 0.0$, $p = 0.0010$), as well as between Text2Room and SpaceBlender (M = 3.90, SD = 0.30) ($W = 3.5$, $p = 0.0049$). Finally, for *Familiarity*, we found a statistically significant difference between Text2Room (M = 3.10, SD = 0.83) and SpaceBlender (M = 3.95, SD = 0.22) ($W = 0.0$, $p = 0.0039$).

5.5.2 Benefits and Limitations of Environments for the Clustering Task

Next, we discuss themes from participants' post-task reflections on the clustering task and the extent to which the three environment styles supported their work. We refer to paired participants as P#A and P#B, where # represents their pair ID.

Environments typically played a passive role in supporting spatial organisation, but some participants adapted their clustering strategies to distinct or familiar features of SpaceBlender environments. When asked what strategies they adopted to complete the clustering task across all three conditions, most participants described using the centre of the environment as a staging area for finalised clusters. Preferences varied regarding whether ungrouped notes were stored and compared in the middle of the space or within individually assigned regions.

Although the task did not require explicit use of environmental features, some participants in SpaceBlender reported drawing on such features. First, P5A and P6A stated that the more detailed SpaceBlender environments, with distinct segments, helped them form mental models of where to organise notes: "I just put the green objects in the green area of the environment" (P6A). This

strategy contrasted with P6A's experience in GENERIC3D, where a lack of environmental cues prompted a more deliberate strategy of labelling areas of the environment for specific clusters: "We had to actually allocate areas because the wall was just gray." Additionally, several participants reported that familiar details in SPACEBLENDER environments (i.e., spatial landmarks preserved from their input images) informed their clustering strategies (P4A, P7A, P10A). P7A noted: "Familiarity helped because this is pretty much like where I work a lot" and "it just feels a bit more like comfortable, like, thinking in that area." This sentiment was echoed by P10A, who described aligning sticky notes with their own table in the virtual workspace.

Participants had mixed preferences for minimalistic and realistic environments in relation to maintaining focus on the clustering task. Most participants favoured the simplistic design of GENERIC3D, reporting that it supported sustained attention on the task. P8A described it as "the cleanest environment" and stated that it supported task efficiency. Although P3A described the GENERIC3D environment as "cartoonish" and "not real," P3A considered it the "highest quality" environment. P9A identified a benefit of combining low-fidelity environments with virtual avatars: "I know my figure in the space was artificial, so I'm more aware I'm in the game."

Some participants expressed the opposite preference: "[GENERIC3D] is my least favorite, because it didn't feel realistic" (P4B). Several participants reported that the more detailed TEXT2ROOM and SPACEBLENDER environments supported immersion: "because it feels more realistic, it feels more immersive," while also noting that visual detail could distract from task engagement.

Participants reported higher physical comfort and navigability in the SPACEBLENDER and GENERIC3D conditions compared with the TEXT2ROOM condition. All participants identified TEXT2ROOM's inconsistent floor geometry as a major source of navigational difficulty and discomfort, with P8A noting

that "stuff sticking out of the floor" significantly hindered navigation. Similarly, P3B pointed to "noise and also the distortion in the floor" in TEXT2ROOM, while noting that this issue was not present in SPACEBLENDER.

Many participants noted that both generative environments exhibited low texture resolution, noisy artefacts, and incoherent geometry, which detracted from realism and usability. Both the TEXT2ROOM and SPACEBLENDER environments were criticised for geometric inaccuracies that mapped physical objects onto the floor, such as "the table and a couple of the chair legs" (P5B).

At times, reported physical discomfort extended beyond visual annoyance. For example, P10A stated that they "did not want to move much in [TEXT2ROOM]," attributing simulator sickness to poor structural coherence and excessive clutter. Some participants reported closing their eyes while navigating in TEXT2ROOM environments (P5B, P10A, P11A, P11B).

In contrast, participants universally praised the visual and geometric consistency of the GENERIC3D environment as "navigable" and "clean." Participants described the environment's stylised consistency as visually appealing and supportive of task performance. P6A stated that it was "more spacious and a lot more easy to navigate." Similarly, P8A described it as "clean and easy to perform the task in."

5.5.3 Future Requirements and Potential Use Cases for Blended Spaces

Finally, we discuss participants' comments on scenarios in which SPACEBLENDER environments could support VR telepresence, as well as suggestions for improving blended environments.

Participants envisioned future scenarios in which blending familiar context into virtual environments could provide explicit or implicit value. After walking through additional SPACEBLENDER environments, participants discussed how SPACEBLENDER could be applied in VR telepresence settings. For professional use cases, participants proposed meeting spaces that incorporate inspiring

physical locations (e.g., "areas where you can have a coffee... and discussions") to be "more conducive for ideation" (P6B), as well as learning environments that use different parts of the blended space to structure educational activities (P3A, P5A, P6B). Participants also discussed social interactions (e.g., gaming with friends, family gatherings) and personal well-being (e.g., providing familiar environments for therapy and recording personal memories) (P3A, P3B, P4A, P8A). Some participants expressed concerns regarding the exposure of personal spaces (P3A, P3B, P7A, P8A). For instance, P3A stated that, depending on the scenario, sharing "a part of [their] life" could feel uncomfortable.

Across these scenarios, participants distinguished between SpaceBlender environments providing *explicit* and *implicit* value for collaboration in different contexts. For example, virtual training applications may require realistic depictions of users' surroundings, while social use cases may benefit from familiar details that "invoke a sense of being in a cozier space" (P8A). This distinction aligned with participants' feedback on the clustering task. Some participants reported that the SpaceBlender environment did not explicitly support their clustering strategies, while others described implicitly using environmental features (e.g., colours, familiar furniture) to establish organisation patterns.

Supporting future use cases for blended environments in VR telepresence requires quality improvements and stronger alignment with real-world spaces. While participants identified potential value in SpaceBlender environments for VR telepresence, they consistently emphasised the need to improve visual quality and realism to realise these benefits. We observed instances in which deviations from participants' memories of familiar locations caused confusion or disorientation. P9A expressed concerns about the fidelity of reconstructed input images in the blended space: "it made me feel a bit weird that I could not recognise some things on my desk." P1A described adjusting expectations due to partial reconstructions of familiar environments: "it takes a lot of getting used to if there's... somewhere you're familiar with, but then you turn around and it's not

what you're familiar with." Some participants also stated that the shape of the blended spaces did not match their expectations (P1A, P1B): "[...] closed spaces usually don't look like that" (P1B). By contrast, P4A suggested a use case that aligned with the circular shape of the blended space, imagining "[...] students in a circle, and the teacher standing in the middle."

To improve realism, some participants wanted the blended virtual environment to "connect with [their] real environment" by matching the geometry of their physical surroundings (P9B). Participants suggested that such alignment could support mixed reality collaboration for physical tasks (P1B, P3A, P10A, P10B, P11B), such as in the cooking class scenario in the Appendix (Figure D.5).

5.6 Discussion and Future Work

We structure our discussion around *(1)* avenues for further studies examining the impact of blended virtual environments on VR collaborative tasks; *(2)* themes for improving and extending SpaceBlender; and *(3)* the limitations of our work.

5.6.1 Positioning SpaceBlender Within Shared Space Construction Models

This chapter discussed boundary-based and volumetric shared space construction approaches in Section 5.1 to contextualise our approach of *space blending* as a representational transformation that preserves visually rich environmental context without privileging a single host environment. In this section, we further situate SpaceBlender within existing theoretical frameworks for shared space construction in CMR.

Within Ohl's Tele-Immersion framework (Section 2.5.3), the *scene blending* operation enabled by SpaceBlender can be interpreted as extending the framework's *scene warping* stage towards the broader *morphing* and *mixing* operations that Ohl outlined as future research directions [69]. SpaceBlender performs multi-site scene extraction from distinct locations and then applies a transformation that uses diffusion-based completion to jointly shape geometry and appearance while reconciling extracted fragments. The resulting mixed

location volume functions as the shared collaboration space, making cues from multiple sites mutually available within a single representation. This interpretation complements prior volumetric transformations that primarily relied on geometric mappings applied to symbolic, segmented, or anchor-based representations [65, 70].

Wong et al.'s task-driven Spatial Heterogeneity Framework (Section 2.5.4) offers an additional perspective on how SpaceBlender could evolve [65]. Our current pipeline generates a blended environment from environmental imagery and generic layout constraints, rather than from explicit task requirements. A promising direction for future work is to incorporate task-specific areas and activity requirements as additional inputs to generation.

Finally, participant concerns about exposing personal spaces, as described in Section 5.5.3, connect to the property of permeability in the MR boundaries framework by Koleva et al. [35] (Section 2.5.2). Permeability highlights privacy and selective disclosure as design constraints, motivating future extensions to SpaceBlender that provide users with control over what environmental detail of their personal spaces is exposed at what moment.

5.6.2 Opportunities for Supporting VR Telepresence with Blended Environments

We identify three avenues for future studies that examine how blended virtual environments shape collaborative processes in VR telepresence.

Leveraging familiar context in VR telepresence scenarios Some participants reported that recognisable elements in the SpaceBlender environment informed their clustering strategies (Section 5.5.2), including aligning notes with remembered furniture and using distinct or familiar features as reference points. In the questionnaire results, *Familiarity* was rated as significantly more impactful in SpaceBlender than in Text2Room (Section 5.5). Our study was preliminary, and the clustering task did not require explicit use of environmental features. These observations suggest that familiarity is a promising direction for future re-

search on collaborative spaces, as explored by Johnson et al. [38] in instructional settings.

Participant feedback further indicated that the desired functionality of blended spaces may vary across professional and social contexts (Section 5.5.3). In professional settings, participants envisioned meeting spaces that establish workspace context or support ideation, as well as learning environments that use different regions of a blended space to structure activities. With continued use, familiar spaces may support preparation and memory for collaborative work. The uniquely blended regions of SpaceBlender spaces may also provide a persistent artefact of collaborative activity [13, 329]. Social scenarios placed greater emphasis on comfort and meaningfulness, suggesting that place cues can support social presence through situated familiarity. These distinctions motivate evaluation beyond a single task type, using measures that capture grounding and awareness, social presence, and task performance across different collaboration genres.

At the same time, participant comments on misrecognition and disorientation in familiar scenes indicate that familiarity changes what users expect the representation to preserve. When the representation deviates from memory, the scene can become harder to interpret and can demand additional repair work to reconcile what is shown with what is known [48, 49]. This tension aligns with the concept of plausibility in presence research [141], where suspension of disbelief is challenged when virtual scenarios diverge from expected norms (see Section 2.4). The effect may be more pronounced in blended environments, where users anticipate recognisability yet encounter distortions or omissions. Future studies should therefore treat familiarity as a double-edged representational resource. It can support grounding through recognisable cues, but can also increase sensitivity to representational errors.

Structural coherence is a prerequisite for comfort and navigation in generative 3D spaces Participants expressed mixed preferences for minimalistic



Figure 5.13 Examples of errors in SpaceBlender pipeline components: (A) protrusion of an outdoor structure, (B) distorted table, (C) depth estimation error, (D) depth alignment error, (E) depth estimation error, (F) floor expansion error.

versus more realistic environments in relation to maintaining focus on the clustering task (Section 5.5.2). Many participants preferred Generic3D for its visual consistency and perceived support for task efficiency, while others preferred the more detailed generative environments for immersion or realism.

Participants' strongest negative feedback across conditions concerned navigational difficulty and discomfort in Text2Room due to inconsistent floor geometry and structural incoherence (Section 5.5). By contrast, participants reported higher comfort and navigability in Generic3D and SpaceBlender, which they attributed to more consistent geometry. These findings motivate treating geometric coherence and navigability as primary requirements for generative telepresence spaces, particularly when users are expected to move through the environment. Future work should examine how different levels of visual fidelity and environmental detail interact with task focus and physical comfort across collaboration tasks beyond affinity diagramming.

Extending SpaceBlender environments to support explicit interactions for collaborative scenarios The SpaceBlender environments primarily supported the clustering task in our study implicitly, and only some participants reported adapting their strategies to environmental features. We outline several directions for extending SpaceBlender to support collaboration more explicitly. One direction could enable users to manipulate virtual objects within the scene by semantically segmenting scene components and applying corresponding functionality, such as *drawing* on a *whiteboard* [381, 10, 153]. In line with this direction, while generating 3D scenes is currently time-consuming, future models may enable real-time user-driven changes to shared spaces [13]. These changes could include the dynamic incorporation of activity-related objects such as tables, chairs, and task-specific tools, reflecting early visions articulated by Dourish [45].

Future work may also extend the layout, geometric prior, and trajectory definition techniques to support more ecologically valid submesh arrangements, including multi-room or multi-storey structures. This could enable features such as breakout rooms and meeting context transitions [382]. To further incorporate spatial familiarity, these layouts could be modelled after existing building floor plans [371], with submeshes aligned to these layouts rather than the current parameter-based approach (Section 5.3.3.3). An artistic perspective on this direction can be found in Do Ho Suh's *Perfect Home* [383], shown in Figure 5.14, an illustration that brings multiple lived places into a single domestic proposition. This perspective suggests that blended spaces may also have value beyond multi-user telepresence, for example as single-user environments that integrate places connected to memory, work, travel, or personal identity. Such spaces could later be opened into collaborative settings, allowing users to bring personally meaningful spatial contexts into shared activity. Future work may study how spatially familiar layouts affect navigation, collaboration efficiency, spatial awareness, and the transition from personal to shared virtual environments.

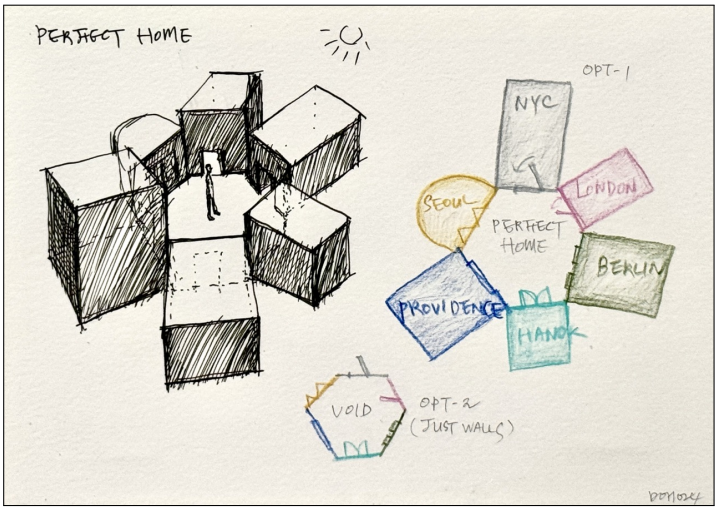


Figure 5.14 Photograph of an illustration from Do Ho Suh's *Perfect Home* [383], included in *The Genesis Exhibition: Do Ho Suh: Walk the House*, Tate Modern, London, UK. Photograph by the author. Reproduced with permission from the artist.

5.6.3 Opportunities to Improve and Extend SpaceBlender

In this section, we discuss potential ways to improve and extend SpaceBlender based on feedback shared by our study participants.

Improving the quality and physical comfort of SpaceBlender environments

Participants highlighted the need for improved visuals and geometry to support their envisioned future uses of SpaceBlender. A prevalent issue was inaccurate depth estimation and alignment, which led to objects merging with walls, floors, or ceilings, or displaying implausible depth values. These issues reduced realism and coherence (Figure 5.13A–E). Due to the iterative nature of the pipeline and the dual use of the mesh representation for rendering and completion, these inaccuracies tended to amplify during generation. One way to address this may be to replace *IronDepth* [360] with a more recent monocular depth estimation model [359, 384]. The replacement model should also support depth image *inpainting*, a task that is less commonly supported.

Additionally, semantic priors could be extended to objects to improve geometric consistency, as explored in the recent *ControlRoom3D* framework [342]. While this framework requires manual definition of priors, it could be combined

with our geometric prior definition technique and existing furniture layout synthesis methods to support a more automated approach [385, 386]. Alternatively, SDF-based methods may be used to define the mesh directly and could be extended to condition generation on the geometry of input submeshes to support blending.

Several participants also noted the low resolution of SpaceBlender environments. While the current MultiDiffusion-based inpainting process would support higher-resolution images if paired with a depth estimation and inpainting model capable of processing them, adopting more recent image generation models that produce higher-resolution images could further improve visual quality [387]. Some recent models [388, 389] also offer improved controllability, which could help prevent failure modes such as the one in Figure 5.13F, where our floor generation method produced a kitchen island rather than a floor even though the prompt specified otherwise. Lastly, recent video generation models [390, 391, 367, 368] might replace iterative image inpainting as a route to improving view coherence, reducing reliance on a mesh as the sole intermediate representation.

Aligning SpaceBlender environments with the real world Our participants commonly expressed a desire to align the blended space with their physical environment (Section 5.5.3). This may be achieved by extending SpaceBlender to accept meshes or point clouds as input that are pre-captured by users or, alternatively, captured by RGB-D cameras mounted in the user's local space in real time [214]. These representations could be registered to the user's local physical space, as explored in Chapter 3, enabling the user to observe the blended space with an MR HMD while retaining awareness and physical interaction capabilities with their surroundings.

5.6.4 Limitations

Our user study included VR novices and periodic VR users, whose perceptions may not generalise to experienced VR users. However, novice VR users helped surface critical challenges with SpaceBlender environments, particularly those

involving core VR interaction requirements (e.g., physical comfort and navigability) that also pertain to experienced users.

Some elements of our study design limited our ability to isolate the impacts of familiar context and fidelity on participants' task performance and collaboration patterns. First, we did not control the familiarity of participants' input images used to generate SpaceBlender environments, as some participants used pre-selected images rather than uploading their own. Second, our choice of baseline conditions may limit the generalisability of some findings. Several participants preferred the higher realism of the generative environments over the Generic3D condition. However, since the baseline and generative conditions had substantially different texture quality, this preference might differ under a higher-fidelity version of the Generic3D baseline. Future work could compare SpaceBlender with other environment types, including higher-fidelity virtual environments [392, 393], photogrammetry-based models, or manually designed environments. Third, the clustering task did not require participants to interact with the virtual environment explicitly, allowing participants to decide whether to use environmental features for the spatial organisation of sticky notes. We chose affinity diagramming as a comparable task across all three conditions because the Generic3D and Text2Room environments would be insufficient for tasks requiring access to users' physical surroundings. Future work may study blended spaces alongside VR tasks that require explicit environmental interaction. Lastly, given the novelty of generative 3D environments that incorporate familiar spaces, participant responses may have been subject to response bias [327].

Additionally, our sample size was insufficient to calculate order effects reliably. Participants who began with the Generic3D might have been primed to organise notes in mid-air rather than aligning them with the environment in subsequent conditions due to the environment's lack of distinctive visual features.

5.7 Conclusion

This chapter addressed RQ4 by examining how computational techniques can transform spatial representations to support shared understanding across distributed MR contexts. Rather than relocating one collaborator into a reconstruction of another's environment, we examined whether a shared environment can be synthesised that incorporates contextual cues from multiple physical places. To this end, we contributed SpaceBlender, a pipeline that uses generative AI to incorporate and extend users' physical surroundings into blended virtual environments for VR telepresence. SpaceBlender projects multiple user-provided images into 3D segments, aligns these segments to a uniform floor level to support navigability, and blends them via diffusion-based space completion guided by geometric priors and dynamically generated text prompts.

Through a preliminary within-subject study with 20 participants, we examined how a virtual baseline (Generic3D), a generative baseline (Text2Room), and SpaceBlender shaped collaboration during a clustering task. Participants reported higher physical comfort and navigability in the virtual baseline and SpaceBlender than in the generative baseline, attributing difficulties in the generative baseline to inconsistent floor geometry and structural incoherence. Many participants preferred the visual simplicity of the virtual baseline for maintaining task focus, while some used distinct segments and familiar landmarks in SpaceBlender as reference points for organising notes. Participants also described professional, social, and personal scenarios in which familiar context could add value, while emphasising the need for higher visual and geometric quality and stronger correspondence to real spaces to reduce misrecognition and disorientation.

In conclusion, these results show that multi-site imagery can be extracted, aligned, and completed into a single navigable VR workspace that preserves cues from multiple environments. They motivate future work that evaluates blended environments in tasks where place and spatial reference are consequential, and that extends the pipeline towards activity-aware and user-steerable generation,

including dynamic incorporation of task-relevant structures as collaboration unfolds (RQ3) and end-user customisation to shape shared spaces over time [13].

Chapter 6

Discussion

This dissertation investigated how environmental representations in asymmetric CMR can be constructed, augmented, and transformed to incorporate contextual information. We examined how these environmental representations influence mutual understanding, coordination, and task performance. Across three chapters, we characterised how collaborators establish and maintain shared space under technological and environmental context asymmetry, identified where representational strategies break down under partial or misaligned views, and explored how augmentation and transformation can provide coordination resources. In this chapter, we synthesise findings across these investigations, relate them to the research questions posed in Chapter 1, and outline implications for the design of asymmetric CMR systems. We then discuss limitations of the work at the dissertation level and identify directions for future research, including the emerging role of AI as a collaborator in addressing asymmetries.

6.1 Addressing the Research Questions

The four research questions introduced in Chapter 1 guided the empirical investigations reported in Chapters 3 to 5. In this section, we return to each research question and synthesise evidence from across the dissertation, relating findings to the foundational concepts and prior work reviewed in Chapter 2.

6.1.1 RQ1: How is shared space perceived and enacted in asymmetric collaborative mixed reality scenarios?

RQ1 asked how collaborators experience and interpret shared space when their viewpoints, interfaces, and situated contexts differ. We operationalised this ques-

tion in Chapter 3 through an asymmetric AR–VR configuration in which a remote VR user collaborated from within a static reconstruction of the AR user's physical workspace. Overall, the findings suggest that *shared space was established through ongoing grounding work rather than delivered by the reconstruction alone*. The replica produced a strong *candidate* shared space by offering a consistent set of spatial landmarks and object layouts, which some dyads referenced explicitly. For example, several dyads used these landmarks as anchors when agreeing on a room division for "divide and conquer" strategies. At the same time, environmental context asymmetry remained substantial once participants understood that the VR user acted in a static model while the AR user remained situated in the live physical room. Dyads responded by beginning collaboration with explicit verification of correspondences between landmarks (e.g., doors, a clock, and poster locations), by questioning whether poster contents might change over time, and by engaging in conversational repair when deictic reference produced confusion [25]. This is notable because the task was only weakly dependent on environmental features, suggesting that more demanding spatial tasks could amplify these dynamics. Future work could explore how such grounding work unfolds when the representation updates in real time, raising questions about latency, consistency, and trust in what is shown [152, 213, 62, 394, 395].

A second observation is that technological asymmetries shaped how participants enacted spatial practices of exploration and attention within the shared space. VR users travelled further than AR users, consistent with joystick locomotion supporting rapid viewpoint changes and broader exploration. AR users exhibited higher head rotation velocity, consistent with compensatory visual search under the constraints of natural walking and limited device FOV. In contrast to prior reports for more dissimilar pairings such as VR versus desktop setups [52, 68, 54], we did not observe a statistically significant difference in participant-rated leadership between AR and VR users. An open question is whether these results generalise to pairings with larger technical capability gaps, such as VR–desktop or AR–handheld configurations, where prior work reported

stronger role differentiation [54]. Furthermore, our study confounded device type with locomotion (joystick vs. natural walking), which likely influenced distance travelled and presence [247] and warrants further investigation to disentangle how each factor contributes to spatial practices.

Participant comments further suggested that prior exposure to the physical room can shape the interpretation of the virtual replica. One participant, after switching from the physical room to VR, reported feeling "predisposed to the space" and described the reconstruction as "slightly uncanny". Such reactions hint at how memory of a physical setting and perception of its representation interact, although the implications of these observations remain uncertain. We see this as motivating future work that systematically varies exposure order and duration to examine how familiarity develops over time and how it interacts with presence-related cognitive processes [396] and cross-reality memory retention effects such as memory source confusion [397].

An overarching implication for system design is that registered geometric replicas should not be treated as sufficient for shared space construction in the presence of environmental change. Even high-fidelity representations require complementary cues that signal which of their aspects can currently be treated as mutually available, even in tasks with limited dependence on environmental features.

6.1.2 RQ2: What limitations do pre-captured environmental representations present for supporting mutual understanding and task coordination in asymmetric collaborative mixed reality?

RQ2 asked where representational strategies break down when applied to asymmetric configurations. We examined this question in both Chapter 3 and Chapter 4. In Chapter 3, a static reconstruction mediated collaboration in a controlled indoor task. In Chapter 4, practitioners described challenges with pre-captured environmental representations during site-specific outdoor authoring of AR experiences, and we observed their influence during a user study in which such

representations formed the basis for remote authoring across conditions. Two limitations emerged consistently across these settings.

The first limitation concerns mutual availability. Static representations created persistent uncertainty about what could be treated as currently shared, even when they offered a coherent spatial scaffold. In Chapter 3, dyads periodically questioned whether representational content was changing, and they invested effort in negotiating what a partner could currently see or reference. In Chapter 4, ex-situ developers described uncertainty about authoring decisions because they lacked reliable information about on-site conditions. These breakdowns are consistent with the role of workspace awareness under partial views [29] and with the idea that additional grounding work becomes necessary when shared evidence is uncertain [25]. In terms of the MR boundary properties defined by Koleva et al. [35], this common failure mode reflected insufficient *permeability*, meaning that the representation did not transmit adequate information about the remote environment's current state, resulting in *environmental context asymmetry*.

The second limitation concerns temporal validity. Pre-captured meshes were often outdated, incomplete, and unable to represent dynamic context. In the formative study in Chapter 4, all five interview participants described difficulties with pre-captured *location meshes*, characterising them as "old," containing "mesh holes everywhere," or being "very incomplete." These limitations impeded authoring decisions because spatial constraints, occlusions, and environmental dynamics could not be assessed reliably. The same concern appeared in the user study, where ex-situ developers working from a static *location mesh* reported lower confidence in their authoring decisions than those who could synchronously collaborate with an in-situ partner through CoCreatAR, often attributing uncertainty to missing or outdated information. Future work could examine how these limitations manifest in other domains that depend on up-to-date environmental state, including remote collaboration in industrial [398], construction [399], and disaster response [400] settings.

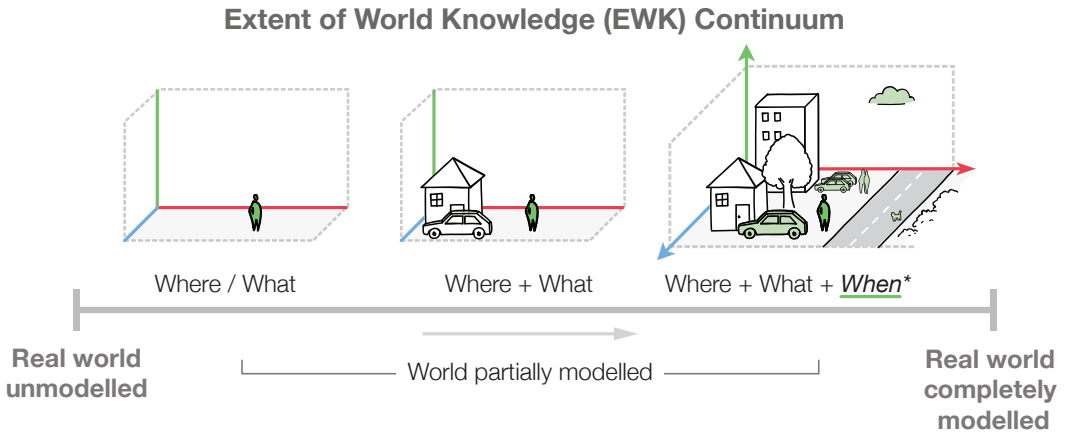


Figure 6.1 Extending the extent of world knowledge dimension from Milgram et al. [129] to explicitly include *temporal validity* as part of *world knowledge*. This extension recognises that accurate representation of "what" and "where" may also require knowledge of "when" the represented state held, particularly in dynamic environments. The green elements represent the user (and their sensed state), while the light green elements represent captured dynamic elements of the environment.

Considered together, these two limitations point to a broader observation. Pre-captured representations are useful as long as collaborators can treat them as a stable approximation of the site's current state. When that assumption is violated, either because the representation lacks coverage or because the represented state has aged, the representational gap is absorbed by additional interactional work, which becomes increasingly costly as task demands or environmental dynamics increase [401]. This observation motivated the constructive responses pursued in RQ3 and RQ4.

Based on these findings, and inspired by the *dynamics* property as proposed by Koleva et al. [35], we revisit Milgram and Kishino's *extent of world knowledge* dimension [129] that we discussed in Section 2.4. The original dimension in the taxonomy characterises knowledge in terms of "what" and "where" and omits temporal validity, that is, "when" the represented state held. We illustrate a conceptual extension that incorporates temporal validity into *world knowledge* in Figure 6.1. Future work could explore how staleness cues can be designed and integrated into representations in ways that are interpretable and actionable.

6.1.3 RQ3: How can environmental representations be augmented to support specific collaborative tasks across asymmetries of access, context, or control?

RQ3 asked how representational augmentation can support collaboration when participants have different access to a target environment, different tools for acting on it, and different control over how the shared workspace is updated. Chapter 4 operationalised this question in site-specific outdoor AR authoring, where ex-situ (i.e., remote) work typically proceeds on a pre-captured *location mesh* that is often incomplete and temporally misaligned with the current site state. Three observations emerged from this work.

First, anchored visual augmentation supported task-driven repair and verification of proxy representations. CoCreatAR layered anchored updates onto the *location mesh*, including coarse meshes and RGB-D snapshots to repair missing geometry, and 3D cursor markers and annotations to represent constraints and intended placements. In the user study, participants relied on these anchored layers when the static *location mesh* did not support verification, supported by the system's live video feed and verbal communication. With these augmentations and real-time feedback, ex-situ participants reported higher confidence in the authored outcome. This repair-oriented layering aligns with prior hybrid approaches that combined stable spatial scaffolds with channels better suited to dynamics [207, 219, 220, 221, 222]. Open questions concern how repair layers should persist across sessions, how they should be applied in more complex settings, and how they perform under constraints that demand higher placement precision.

Second, the study confirmed that implicit cues support lightweight awareness, while explicit referential actions stabilise intent when higher precision is required. The live in-situ screen stream and audio provided continuous access to situational context, supporting timely grounding work [25, 48]. When precise reference was needed, participants relied on explicit anchored cues. Surface-anchored annotations indicated placement locations, and cursors functioned as

lightweight deictic references during alignment, consistent with prior findings on explicit pointing and annotation [293, 166]. The two cue types operated in complementary roles, with implicit cues sustaining background awareness and explicit cues invoked at moments of focused coordination.

Third, the benefits of richer context came with reports of coordination overhead [127, 402, 25, 300]. For in-situ users, this overhead interacted with the demands of safely navigating the physical environment while capturing context and responding to remote requests [301]. This trade-off has implications for the design of asymmetric CMR systems beyond authoring. Augmentation strategies should be evaluated not only in terms of what is made available, but also in terms of who bears the cost of making it so. These observations motivate future interfaces that reduce orientation and interaction costs for in-situ work (e.g., mini-maps [309] and directional cues [310]), as well as reduce representational clutter for ex-situ work (e.g., user-perspective rendering [311] and selective visualisation of information layers [312, 313]).

6.1.4 RQ4: How can computational techniques transform spatial representations to support shared understanding across distributed MR contexts?

RQ4 asked how computational transformation can reshape or synthesise spatial representations for collaborators distributed across different physical contexts. We operationalised this question in Chapter 5 through SpaceBlender, a pipeline that transforms images of multiple users' physical surroundings into unified 3D environments for VR telepresence by combining geometric reconstruction with generative scene completion. Input images from distinct locations were projected into 3D segments, aligned to a uniform floor level, and completed via iterative inpainting. Within the *Tele-Immersion* framework by Ohl [69], discussed in Section 2.5.3, this operation extends the *scene warping* stage toward the "morphing" and "mixing" operations described as future directions, complementing prior transformations that relied on geometric mappings applied to symbolic or anchor-based representations [65, 70, 225].

Our preliminary study showed that structural coherence is a precondition for transformed spaces to remain usable for navigation and grounding. Participants reported higher physical comfort and navigability in SpaceBlender and the low-poly baseline environment than in the generative baseline environment [339]. Participants often attributed difficulties performing our study's simple affinity diagramming task within generative environments to inconsistent geometry and structural incoherence. This pattern is consistent with presence research emphasising the importance of sensorimotor contingencies [134] and the plausibility of virtual environments [141], and it indicates that fidelity gains from generative completion are conditional on geometric integrity at the surfaces collaborators traverse and reference. Future research could further explore how structural coherence interacts with these experiential properties in generative environments, including the impact of locomotion techniques [247].

Furthermore, we found that transformation reshapes the interpretive demands placed on collaborators and the reliability of grounding resources. Contextual cues sometimes supported coordination, for example when participants used recognisable elements from familiar spaces as reference points for organising their work [38]. However, synthesis artefacts introduced ambiguity in regions participants expected to be stable and recognisable, with several participants reporting discomfort when familiar scenes deviated from memory. This sensitivity relates to *external plausibility* [142], which refers to the consistency of a virtual environment with expectations formed by users' real-world knowledge and experiences. Familiar contextual content thus operates as a double-edged grounding resource. It can stabilise reference when preserved faithfully, and it can introduce coordination friction when distortions appear at the points of recognition that collaborators rely on most. An important direction for future work on shared space construction techniques that transform environments is how to preserve the reliability of grounding resources under transformation across sessions and realities.

6.2 Future Work

This section identifies directions for future research that extend beyond the scope of this work, organised around five themes: broadening empirical scope, exploring additional asymmetric configurations, developing design guidance for shared space construction, investigating AI as a collaborator and a mediator of asymmetry, and leveraging advances in environmental reconstruction.

6.2.1 Expanding Empirical Scope and Ecological Validity

The empirical studies in this dissertation employed controlled task designs that enabled systematic comparison, while constraining ecological validity. In Chapter 3, dyads performed a word puzzle task that placed limited demands on environmental features, which may have attenuated the consequences of representational limitations. In Chapter 4, the authoring task was grounded in professional practice, but with predefined content and constraints. In Chapter 5, the affinity diagramming task did not require explicit interaction with environmental features, leaving open questions about how blended environments support tasks with stronger implicit or explicit spatial demands.

Future work should examine how our findings generalise to tasks and configurations with greater environmental dependence, longer duration, and less structured coordination demands. This includes collaboration that unfolds over longer periods or multiple sessions, where collaborators may develop routines, repair strategies, and representational conventions that do not emerge in short sessions. Furthermore, investigations of asymmetry across larger groups [403], including scenarios with co-located and remote participants as is common in hybrid work and conferences, could reveal new dynamics in shared space construction and coordination.

6.2.2 Broadening Explorations of Asymmetric Configurations

Asymmetry is a broad concept that spans technological and social dimensions of collaborative work, as discussed in Section 2.3.1. Prior taxonomies have characterised asymmetry along numerous dimensions, including media, fidelity, par-

ticipation, engagement, and benefit [41]; input and output modality, dimensionality, mobility, display space, and content layout [42]; and ability, information, interface, and responsibility in games research [117]. This dissertation focused primarily on technological asymmetry (input and output modality), environmental context asymmetry, and secondarily on role and information asymmetry. Many other dimensions from these taxonomies remain underexplored in the context of environmental representations for CMR, raising numerous questions for future work. For example, how do differences in transmission fidelity or latency shape the reliability of environmental cues for grounding [404, 394, 395]? How do asymmetries of participation or engagement interact with representational design?

In addition to the AR–VR and AR–desktop pairings common in the CMR literature, future research should explore more extreme asymmetric configurations that reflect contemporary hybrid work practices. Many collaborative scenarios involve participants who are mobile, multitasking, or unable to attend to high-bandwidth visual interfaces. Kleinau et al. [405] explored this design space through co-design of multimodal devices for radically mobile hybrid meetings, producing largely audio-driven prototypes that do not rely on continuous visual attention. Such configurations raise questions about how spatial reference and environmental context can be conveyed when one or more collaborators lack visual access to shared representations. What minimal cues are sufficient to maintain grounding when visual channels are unavailable? How can audio-based interfaces support awareness of others' actions and environmental context? How can we design for asymmetric engagement, where some collaborators attend intermittently or in the background [406]?

Accessibility represents a particularly important direction for future work on asymmetric CMR. Collaborators may have permanent or temporary sensory, motor, or cognitive differences that require environmental representations to be rendered, navigated, or interacted with in fundamentally different ways. Future work on asymmetric CMR should examine how shared space construction can

accommodate diverse abilities, ensuring that coordination resources remain available regardless of how participants access the environment. This aligns with prior work on mixed-ability virtual workspaces [407], the application of *role asymmetry* in mixed-ability gaming contexts [408, 409], as well as explorations of usage patterns of VR among older populations [410].

6.2.3 Toward Artificial Intelligence as a Collaborator in Mixed Reality

This dissertation examined how environmental representations support collaboration between humans across asymmetric CMR configurations. As AI agents become capable of perceiving, reasoning about, and acting within shared environments, the same representational and grounding concerns that shaped our investigations also apply when AI takes part in collaboration. The interpersonal grounding work we observed across our studies, including the negotiation of mutual availability, the use of deictic reference, and conversational repair, is precisely the kind of work an AI participant must engage in if it is to support coordination rather than disrupt it. Several pieces of related work that we contributed to begin probing this space outside the main chapters of this thesis [10, 18, 17, 21]. Drawing on these efforts, we discuss two related directions. The first considers AI as a mediator that adapts representations and reduces asymmetries between human collaborators. The second considers the human-AI collaboration itself, where the AI participates in coordination and must establish common ground with the people it works alongside. These directions are intertwined, since an AI system that mediates between collaborators is also a participant whose situational understanding shapes the quality of its mediation.

6.2.3.1 Artificial Intelligence as a Mediator of Asymmetry

AI may serve as a mediator of asymmetry in CMR, operating along a spectrum from implicit background support to explicit human-AI collaboration. At the implicit end, AI systems can reduce asymmetries by adapting content or representations based on inferred context, without requiring users to issue commands.

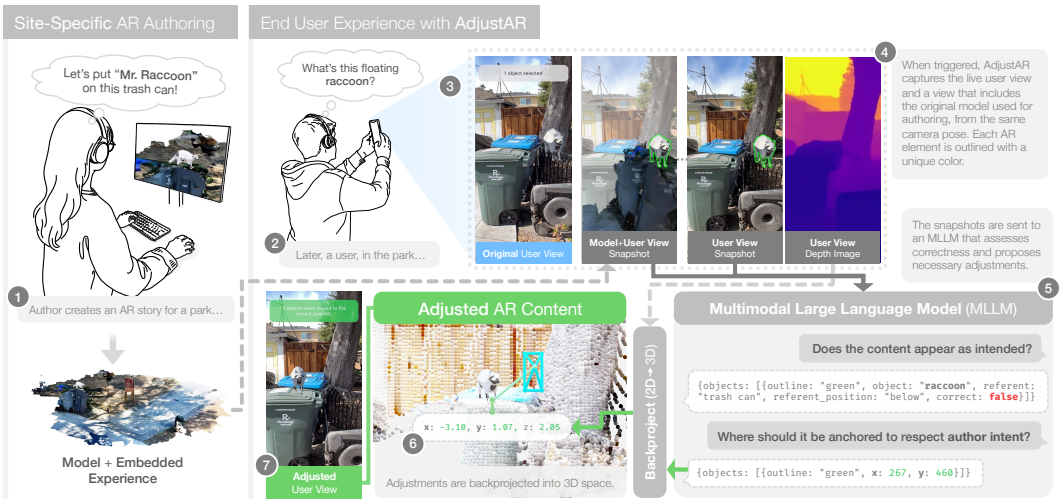


Figure 6.2 Overview of *AdjustAR* [18], which corrects misaligned AR content at runtime: (1) authors place content relative to a georeferenced 3D model of the target site; (2–3) users localise and view the scene *in-situ*, where misalignments may occur due to environmental changes; (4) the system composites live and authored views; (5) an MLLM detects misalignments and infers corrected 2D anchors; (6) corrections are backprojected into 3D and updated in the scene; (7) the user's AR view is updated.

Our follow-up work to CoCreatAR, *AdjustAR*, illustrates this position by using a multimodal large language model (MLLM) to automatically adjust site-specific AR content based on real-time environmental assessment, addressing environmental changes without explicit user intervention [18] (see Figure 6.2). A natural next step is to extend such mechanisms beyond environmental change detection toward broader situational differences, including shifts in activity, crowding, or access constraints, and to investigate whether similar mechanisms could retarget collaborative experiences across dissimilar environments while preserving the grounding resources that collaborators depend on. This direction connects to a wider body of work on context-aware adaptation, where systems infer relevant aspects of the user's situation and adjust accordingly [411, 412, 413]. In Chapter 4, we also identified more lightweight forms of implicit support, where AI could transcribe and summarise in-situ audio notes, extract task-relevant semantics, and help structure annotations in the shared representation so that remote collaborators can treat these updates as mutually available during grounding. Preparation work for collaboration could thereby be partially offloaded to AI,

reducing asymmetries of information access once collaboration begins.

At the explicit end of this spectrum, AI can function as an interactive collaborator that users directly address to overcome capability barriers. For example, our work on *DreamCodeVR* [10] as well as recent explorations [414] demonstrated how speech-driven programming can lower the individual threshold for authoring object behaviours in MR. When techniques of this kind are embedded within collaborative settings, as explored by Kobenova et al. [415], they can be appropriated as resources for addressing skill or capability asymmetries among collaborators. Speech-driven input to AI agents could extend to enabling users with disadvantaged interfaces to request transformations or augmentations of shared environmental representations that would otherwise be infeasible from their device. A handheld AR user with limited input capabilities, for example, could verbally instruct an agent to modify a shared scene on their behalf, while remote collaborators retain direct manipulation through richer interfaces. The space between these implicit and explicit positions is itself an open design territory. Mixed-initiative interaction [416, 417] offers a framing for systems that combine system-initiated adaptations with user-initiated requests, and it raises questions about when handover between modes should occur, how it should be signalled, and how authored intent persists across these transitions. These questions become more pressing in asymmetric CMR, where the appropriate balance may itself differ between collaborators based on their device capabilities, situational demands, and roles.

6.2.3.2 Common Ground and Mutual Awareness in Human-AI Collaboration

A premise underlying both the implicit and explicit positions discussed above is that the AI possesses, or can acquire, sufficient understanding of the situation to act appropriately. Across the studies in this dissertation, we observed that human collaborators expended substantial effort to establish and maintain common ground under environmental context asymmetry, drawing on conversational repair, deictic reference, and ongoing verification of mutual availability. When

an AI participates in such settings, whether as a mediator or as an addressable collaborator, it inherits these grounding demands without inheriting the conventions, shared history, or interactional competencies that human collaborators bring. The challenge of establishing common ground with AI is therefore at least as central as, and arguably more demanding than, its interpersonal counterpart.

The grounding problem in human-AI collaboration is bidirectional. The AI must form a workable model of what each human collaborator currently perceives, intends, and treats as salient, while each human must in turn form a workable model of what the AI can perceive and act on. An AI participant that lacks awareness of what is currently mutually available, and to whom, risks producing contributions that disrupt rather than support coordination. A complementary risk arises when the AI's perceptual and inferential capabilities are opaque to users, leading collaborators to over- or under-trust its contributions.

Recent work has begun to explore what kinds of context AI systems should attend to and how they should solicit it without imposing interaction overhead [17, 418, 419]. Our work on *Sensible Agent* [17] contributes to this direction by examining how proactive AR agents can infer surroundings, context, and user intent through unobtrusive multimodal sensing, allowing the agent to participate in interaction without continually requiring explicit input. Establishing such awareness depends not only on what the agent senses, but also on how that information is structured for reasoning. Our work on *Semantic Reality* [21] addresses this through a persistent model of objects and the relationships among them, giving the agent a basis for tracking which entities and dependencies are currently in play during a task. Together, these efforts illustrate two complementary requirements for AI participation in collaboration, namely the unobtrusive acquisition of contextual information and the structured representation of that information in a form that can support shared reference.

Several open questions follow from this framing. How should AI agents represent their own uncertainty about what collaborators currently see, know, or intend, and how should that uncertainty be made legible at the interface?

When an AI's situational model diverges from a collaborator's, what mechanisms allow for repair analogous to the conversational repair observed among humans? How can mutual awareness between humans and AI scale to settings with multiple human collaborators distributed across heterogeneous devices, where the agent must reason about several partial views simultaneously? These questions connect to broader research on human-AI interaction [420, 417]. Just as our studies showed that shared space among humans is established through ongoing grounding work, AI agents that take part in CMR will need to engage in equivalent work to be treated as genuine collaborators, and this represents an important direction for future research.

6.2.4 Leveraging Advances in Environmental Reconstruction and Representation

Rapid progress in environmental reconstruction and generative techniques offers new possibilities for addressing the representational limitations identified in this dissertation. Neural scene representations such as NeRF and 3DGS are improving fidelity, temporal consistency, and support for dynamic content [295, 296]. Recent work has begun incorporating object-level interactivity into reconstructed environments [421, 422], which could enable remote collaborators to act within replica spaces in ways that more closely approximate physical interaction. Open questions remain about what level of reconstruction fidelity is necessary to support effective grounding under environmental context asymmetry. How do mismatches and breakdowns in inferred affordances or partially captured dynamics affect mutual understanding and coordination? Another major challenge is how such reconstructions can be efficiently captured, particularly in highly dynamic or large-scale environments, as complicated capture setups remain a barrier to practical deployment (Section 2.7.2.2).

Complementing scene-level approaches, targeted augmentation using computer vision techniques can help address dynamic aspects of environmental context asymmetry. For example, Zhang et al. [22] built on *Ubiq-Genie* [5] to represent dynamic environmental factors, specifically people, within static vir-

tual replicas of outdoor spaces using skeletal tracking to improve the liveness of these stale environmental representations. Future work could explore how other types of dynamic context could be incorporated into shared representations to enhance grounding and task performance in domains that depend on such factors, such as site-specific AR authoring, remote assistance, and collaborative inspection.

Chapter 7

Conclusion

This dissertation investigated the construction, augmentation, and transformation of environmental representations in asymmetric CMR to incorporate contextual information. It examined how these representations influence mutual understanding, coordination, and task performance. Across three empirical chapters, we characterised how collaborators establish and maintain shared space under technological and environmental context asymmetry, identified where representational strategies break down under partial or misaligned views, and explored how augmentation and transformation can provide coordination resources that support grounding and awareness.

The first empirical investigation (Chapter 3) examined an asymmetric AR–VR configuration in which a remote VR user collaborated from within a static reconstruction of an AR user's physical workspace. Through a user study ($n = 30$, in dyads), we found that shared space was established through ongoing grounding work rather than delivered by the reconstruction alone. Dyads explicitly verified landmarks and repaired deictic references, showing that even geometrically aligned representations leave residual environmental context asymmetry that collaborators must manage through interaction. The study also showed that technological asymmetries shape spatial practices, influencing how collaborators move, search, and participate within the shared environment.

The second investigation (Chapter 4) addressed task-driven augmentation in the context of site-specific outdoor AR authoring, where environmental change and partial access are persistent constraints. A formative study with industry professionals showed that pre-captured environmental representations were often outdated, incomplete, and unable to represent dynamic context, cre-

ating uncertainty during ex-situ development. We developed CoCreatAR, an asymmetric collaborative system that augments remote work with anchored real-world context through spatial captures and annotations. A comparative user study ($n = 32$, in pairs) indicated that CoCreatAR increased ex-situ developers' confidence in authored outcomes and increased engagement for both roles. At the same time, synchronous collaboration introduced coordination overhead, particularly for in-situ users navigating busy outdoor environments. These findings show that augmentation can make task-relevant context more available, while also changing the distribution of collaborative labour.

The third investigation (Chapter 5) explored representational transformation across multiple physical contexts. We developed SpaceBlender, a pipeline that transforms images of multiple users' physical surroundings into unified 3D environments for VR telepresence by combining geometric reconstruction with generative scene completion. A preliminary user study ($n = 20$, in pairs) compared SpaceBlender environments against a generic low-poly baseline and a generative baseline during a collaborative affinity diagramming task. Participants reported higher physical comfort and navigability in SpaceBlender and the generic baseline than in the generative baseline, and some used recognisable environmental features from familiar spaces as reference points. These findings suggest that blended place cues can become coordination resources when the resulting environment remains structurally coherent enough to support navigation and grounding.

Together, these investigations show how environmental representations shape collaboration in asymmetric CMR. Static reconstructions can provide spatial scaffolds, yet leave uncertainty about what is mutually available. Augmentation strategies can layer dynamic context onto stable representations to support verification and repair, yet they also redistribute coordination work across roles. Transformation approaches can synthesise shared environments from multiple sites, yet their usefulness depends on structural coherence, recognisable contextual cues, and sufficient support for action and reference. These

findings provide a basis for future work on design guidance that relates representational choices to asymmetric participation, on reconstruction techniques that capture dynamic and task-relevant context, and on AI-mediated systems that adapt shared environments without obscuring the cues collaborators use to ground their actions. As CMR systems become more viable for the future of communication and work, the design of shared space under asymmetric conditions becomes a central problem for CMR. This dissertation contributes to that problem by showing how environmental representations can make context more mutually intelligible across asymmetric configurations, and by clarifying the representational conditions through which shared space becomes usable for joint activity.

Appendix A

Publications

This appendix lists publications co-authored by the candidate during the doctoral research period. Publications that form the basis of individual chapters are indicated accordingly.

[1] Steed, Anthony, Izzouzi, Lisa, Brandstätter, Klara, Friston, Sebastian, Congdon, Ben, Olkkonen, Otto, Giunchi, Daniele, **Numan**, **Nels**, and Swapp, David. 2022. Ubiq-Exp: a Toolkit to Build and Run Remote and Distributed Mixed Reality Experiments. *Frontiers in Virtual Reality*, 3. doi:10.3389/frvir.2022.912078.

[2] **Numan**, **Nels** and Steed, Anthony. 2022. Exploring User Behaviour in Asymmetric Collaborative Mixed Reality. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. VRST '22: 28th ACM Symposium on Virtual Reality Software and Technology. ACM, New York, NY, USA, 11. doi:10.1145/3562939.3565630.
Subject of Chapter 3

[3] Steed, Anthony, Archer, Dan, Izzouzi, Lisa, **Numan**, **Nels**, Shapiro, Kalila, Swapp, David, Lammiman, Dinah, and Lindeman, Robert W. 2023. Immersive competence and immersive literacy: Exploring how users learn about immersive experiences. *Frontiers in Virtual Reality*, 4, (Mar. 23, 2023). Retrieved Mar. 27, 2023 from https://www.frontiersin.org/articles/10.3389/frvir.2023.1129242.

[4] **Numan**, **Nels**, Lu, Ziwen, Congdon, Benjamin, Giunchi, Daniele, Rotsidis, Alexandros, Lernis, Andreas, Larmos, Kyriakos, Kourra, Tereza, Charalambous, Panayiotis, Chrysanthou, Yiorgos, Julier, Simon, and Steed, Anthony. 2023. Towards Outdoor Collaborative Mixed Reality: Lessons Learnt from a Prototype System. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). Shanghai, China, 113–118. doi:10.1109/VRW58643.2023.00029.

[5] **Numan**, **Nels**, Giunchi, Daniele, Congdon, Benjamin, and Steed, Anthony. 2023. Ubiq-Genie: Leveraging External Frameworks for Enhanced Social VR Experiences. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*.

2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE, Shanghai, China, 497–501. doi:10.1109/VRW58643.2023.00108.

[6] **Numan**, **Nels**. 2023. Towards Understanding, Alleviating, and Exploiting the Effects of Asymmetry in Collaborative Mixed Reality. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE, 991–992. doi:10.1109/VRW58643.2023.00338.

[7] Friston, Sebastian, Congdon, Ben, **Numan**, **Nels**, Brandstätter, Klara, Izzouzi, Lisa, Thiel, Felix, Zhang, Jingyi, Giunchi, Daniele, Swapp, David, and Steed, Anthony. 2023. Extending the Open Source Social Virtual Reality Ecosystem to the Browser in Ubiq. In *Proceedings of the 28th International ACM Conference on 3D Web Technology* (Web3D '23). ACM, New York, NY, USA, (Oct. 9, 2023), 1–9. ISBN: 979-8-4007-0324-9. doi:10.1145/3611314.3615903.

[8] Lu, Ziwen, Zhang, Jingyi, Shapiro, Kalila, **Numan**, **Nels**, Julier, Simon, and Steed, Anthony. 2023. Reviving the Euston Arch: A Mixed Reality Approach to Cultural Heritage Tours. In 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE Computer Society, (Oct. 16, 2023), 821–826. doi:10.1109/ISMAR-Adjunct60411.2023.00181.

Received an Honourable Mention for Best Design at ISMAR '23 Student Competition

[9] Gatti, Elia, Giunchi, Daniele, **Numan**, **Nels**, and Steed, Anthony. 2024. AIsop: Exploring Immersive VR Storytelling Leveraging Generative AI. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). (Mar. 2024), 865–866. doi:10.1109/VRW62533.2024.00229.

[10] Giunchi, Daniele, **Numan**, **Nels**, Gatti, Elia, and Steed, Anthony. 2024. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). (Mar. 2024), 579–589. doi:10.1109/VR58804.2024.00078.

[11] Giunchi, Daniele, Bovo, Riccardo, **Numan**, **Nels**, and Steed, Anthony. 2024. StreamSpace: A Framework for Window Streaming in Collaborative Mixed Reality Environments. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). (Mar. 2024), 859–860. doi:10.1109/VRW62533.2024.00226.

[12] **Numan**, **Nels**[*], Rajaram, Shwetha[*], Kumaravel, Balasaravanan Thoravi, Marquardt, Nicolai, and Wilson, Andrew D. 2024. SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (UIST '24). ACM, New York, NY, USA, (Oct. 13, 2024). doi:10.1145/3654777.3676361.

Related Patent Filing US18/756,717

Subject of Chapter 5

[13] Rajaram, Shwetha[*], **Numan**, **Nels**[*], Kumaravel, Balasaravanan Thoravi, Marquardt, Nicolai, and Wilson, Andrew D. 2024. BlendScape: Enabling End-User Customization of Video-Conferencing Environments through Generative AI. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (UIST '24). ACM, New York, NY, USA, (Oct. 13, 2024). doi:10.1145/3654777.3676326.

Received an Honourable Mention Award | *Related Patent Filing* US18/542,501

[14] Gatti, Elia, Giunchi, Daniele, **Numan**, **Nels**, and Steed, Anthony. 2025. Around the Virtual Campfire: Early UX Insights into AI-Generated Stories in VR. In *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. 2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR). (Jan. 2025), 136–141. doi:10.1109/AIxVR63409.2025.00027.

[15] **Numan**, **Nels**, Brostow, Gabriel, Park, Suhyun, Julier, Simon, Steed, Anthony, and Van Brummelen, Jessica. 2025. CoCreatAR: Enhancing Authoring of Outdoor Augmented Reality Experiences Through Asymmetric Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). ACM, New York, NY, USA, (Apr. 25, 2025). ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3714274.

Received an Honourable Mention Award | *Related U.S. Patent Filing* 19/451,480

Subject of Chapter 4

[16] Gottsacker, Matt, **Numan**, **Nels**, Steed, Anthony, Bruder, Gerd, Welch, Gregory F., and Feiner, Steve. 2025. Decoupled Hands: An Approach for Aligning Perspectives in Collaborative Mixed Reality. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI EA '25). ACM, New York, NY, USA, (Apr. 28, 2025). doi:10.1145/3706599.3720219.

[17] Lee, Geonsun, Xia, Min, **Numan**, **Nels**, Qian, Xun, Li, David, Chen, Yanhe, Kulshrestha, Achin, Chatterjee, Ishan, Zhang, Yinda, Manocha, Dinesh, Kim, David, and Du, Ruofei. 2025. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Tech-*

nology (UIST '25). ACM, New York, NY, USA, (Sept. 27, 2025), 1–22. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747748.

[18] **Numan**, **Nels**, Van Brummelen, Jessica, Lu, Ziwen, and Steed, Anthony. 2025. AdjustAR: AI-Driven In-Situ Adjustment of Site-Specific Augmented Reality Content. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (UIST Adjunct '25). ACM, New York, NY, USA, (Sept. 27, 2025), 1–4. ISBN: 979-8-4007-2036-9. doi:10.1145/3746058.3758362.

[19] Li, David, **Numan**, **Nels**, Qian, Xun, Chen, Yanhe, Zhou, Zhongyi, Alekseev, Evgenii, Lee, Geonsun, Cooper, Alex, Xia, Min, Chung, Scott, Nelson, Jeremy, Yuan, Xiuxiu, Dias, Jolica, Bettridge, Tim, Hersh, Benjamin, Huynh, Michelle, Piascik, Konrad, Cabello, Ricardo, Kim, David, and Du, Ruofei. 2025. XR Blocks: Accelerating Human-centered AI + XR Innovation. (Sept. 29, 2025). arXiv: 2509.25504 [cs]. Retrieved Oct. 5, 2025 from http://arxiv.org/abs/2509.25504. Pre-published.

[20] Zarraonandia, Telmo, Montero, Alvaro, Cuervo-Rosillo, Roberto, Díaz, Paloma, **Numan**, **Nels**, and Steed, Anthony. 2026. A Comparative Study of Using Augmented vs Virtual Reality in Collaborative Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. doi:10.1109/TVCG.2026.3679895.

[21] Liu, Xiaoan, Gonzalez, Eric J., **Numan**, **Nels**, Colaço, Andrea, Abramyan, Lucy, Zhu-Tian, Chen, Suzuki, Ryo, and Gonzalez-Franco, Mar. 2026. Semantic Reality: Interactive Context-Aware Visualization of Inter-Object Relationships in Augmented Reality. (Apr. 2026). arXiv: 2604.05265 [cs]. doi:10.48550/arXiv.2604.05265.

Appendix B

Appendix – Chapter 3

B.1 Post-Study Questionnaire

Statement	Measure
Who talked the most, you or your partner? Please rate YOUR degree of verbal communication on a scale from 1 to 100.	Talkativeness
Who talked the most, you or your partner? Please rate YOUR PARTNER's degree of verbal communication on a scale from 1 to 100.	Talkativeness
Who took the lead, you or your partner? Please rate YOUR degree of leadership while solving the riddles on a scale from 1 to 100.	Leadership
Who took the lead, you or your partner? Please rate YOUR PARTNER's degree of leadership while solving the riddles on a scale from 1 to 100.	Leadership
Please rate your sense of being in the office with posters with words on them, on the following scale from 1 to 7, where 7 represents your normal experience of being in a place. Scale: 1 (not at all) - 7 (very much)	Presence
To what extent were there times during the experience when the office filled with posters became the "reality" for you, and you almost forgot about the "real world" in which the whole experience was really taking place? There were times during the experience when the office filled with posters became more real for me compared to the "real world"... Scale: 1 (at no time) - 7 (almost all the time)	Presence
When you think back about your experience, do you think of the virtual office more as images that you saw, or more as somewhere that you visited? The virtual office seems to me to be more like... 1 (images that you saw) - 7 (somewhere that you visited)	Presence
During the time of the experience, which was strongest on the whole, your sense of being in the virtual office filled with posters, or of being in the "real world"? I had a stronger sense of being in... Scale:1 (the real world) - 7 (the virtual reality of the poster-filled office)	Presence
Consider your memory of being in the virtual office filled with posters. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By 'structure of the memory' consider things like the extent to which you have a visual memory of the field, whether that memory is in colour, the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such structural elements. I think of the virtual office as a place in a way similar to other places that I've been today... Scale: 1 (not at all) - 7 (very much so)	Presence
During the time of the experience, did you often think to yourself that you were actually just standing in a room wearing a head-mounted display or did the office filled with posters overwhelm you? During the experience I often thought that I was really standing in a room wearing a head-mounted display.... Scale: 1 (most of the time I realised I was in a room wearing a head-mounted display) - 7 (never because the poster-filled office overwhelmed me)	Presence
During the time of the experience, I felt there was a sense of being with the other person.	Co-presence
The computer interface seemed to vanish and I felt like I was directly working with the other person.	Co-presence
The experience was more like working with another person rather than interacting with a computer.	Co-presence
During the time of the experience, I forgot about the other person and concentrated on the task as if I was the only one.	Co-presence
The experience was like some other real experience of working together with another person.	Co-presence
During the time of the experience, I felt a sense of being with another person rather than just experiencing computer images.	Co-presence
The experience resembled being together with another person in a real-world setting.	Co-presence
During the time of the experience, I felt a sense of another human being interacting with me.	Co-presence
I enjoyed the experience in a similar manner to a previous real meeting that was enjoyable.	Accord
I have a desire to meet the other person.*	Accord
I was in harmony with the other person during the task.	Accord
I felt comfortable with the other person.	Accord
I would like to meet the other person again.*	Accord
I would like to form a team with the other person again.	Accord
I felt no embarrassment.	Accord

*Excluded from analysis.

Table B.1 Post-trial questionnaire based on prior work [68, 52, 53].

B.2 Velocity Filter Analysis



(a) AR HMD



(b) VR HMD

Figure B.1 Yaw velocity of the AR and VR HMDs with different sliding-window sizes for 150 frames from the first trial period. For the VR HMD, filtered frames without data (shown with zero) are linearly interpolated based on known data.

B.3 Supplementary Correlation Tables

Please find the correlation tables on the following pages.

Measure	Axis	Size	ρ	p-value
Leadership	Yaw	3	0.500	0.007**
Leadership	Yaw	5	0.503	0.006**
Leadership	Yaw	10	0.502	0.007**
Leadership	Yaw	15	0.502	0.006**
Leadership	Roll	3	0.496	0.007**
Leadership	Roll	5	0.472	0.011*
Leadership	Roll	10	0.453	0.015*
Leadership	Roll	15	0.423	0.025*
Leadership	Pitch	3	0.410	0.030*
Leadership	Pitch	5	0.380	0.046*
Leadership	Pitch	10	0.310	0.108
Leadership	Pitch	15	0.239	0.221
Talkativeness	Yaw	3	0.503	0.006**
Talkativeness	Yaw	5	0.502	0.006**
Talkativeness	Yaw	10	0.486	0.009**
Talkativeness	Yaw	15	0.478	0.010*
Talkativeness	Roll	3	0.486	0.009**
Talkativeness	Roll	5	0.458	0.014*
Talkativeness	Roll	10	0.436	0.020*
Talkativeness	Roll	15	0.421	0.025*
Talkativeness	Pitch	3	0.378	0.048*
Talkativeness	Pitch	5	0.351	0.067
Talkativeness	Pitch	10	0.283	0.144
Talkativeness	Pitch	15	0.209	0.285
Words spoken	Yaw	3	0.342	0.088
Words spoken	Yaw	5	0.344	0.086
Words spoken	Yaw	10	0.312	0.121
Words spoken	Yaw	15	0.296	0.141
Words spoken	Roll	3	0.269	0.184
Words spoken	Roll	5	0.266	0.190
Words spoken	Roll	10	0.237	0.244
Words spoken	Roll	15	0.225	0.268
Words spoken	Pitch	3	0.32	0.111
Words spoken	Pitch	5	0.309	0.125
Words spoken	Pitch	10	0.221	0.279
Words spoken	Pitch	15	0.148	0.470

Table B.2 Spearman correlation coefficients ρ between head rotation velocity and leadership and talkativeness variables calculated over different window sizes of a moving average filter for the *first* trial period. Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

Measure	Axis	Size	ρ	p-value
Leadership	Yaw	3	0.242	0.215
Leadership	Yaw	5	0.265	0.174
Leadership	Yaw	10	0.268	0.169
Leadership	Yaw	15	0.269	0.167
Leadership	Roll	3	0.358	0.061
Leadership	Roll	5	0.371	0.052
Leadership	Roll	10	0.401	0.034*
Leadership	Roll	15	0.401	0.034*
Leadership	Pitch	3	0.244	0.211
Leadership	Pitch	5	0.248	0.203
Leadership	Pitch	10	0.237	0.225
Leadership	Pitch	15	0.226	0.248
Talkativeness	Yaw	3	0.207	0.291
Talkativeness	Yaw	5	0.210	0.283
Talkativeness	Yaw	10	0.190	0.333
Talkativeness	Yaw	15	0.165	0.403
Talkativeness	Roll	3	0.197	0.315
Talkativeness	Roll	5	0.179	0.361
Talkativeness	Roll	10	0.145	0.462
Talkativeness	Roll	15	0.141	0.475
Talkativeness	Pitch	3	0.032	0.873
Talkativeness	Pitch	5	0.011	0.956
Talkativeness	Pitch	10	−0.039	0.845
Talkativeness	Pitch	15	−0.098	0.621
Words spoken	Yaw	3	0.189	0.334
Words spoken	Yaw	5	0.199	0.309
Words spoken	Yaw	10	0.180	0.361
Words spoken	Yaw	15	0.163	0.409
Words spoken	Roll	3	0.041	0.838
Words spoken	Roll	5	0.061	0.757
Words spoken	Roll	10	0.090	0.648
Words spoken	Roll	15	0.083	0.674
Words spoken	Pitch	3	0.102	0.606
Words spoken	Pitch	5	0.115	0.560
Words spoken	Pitch	10	0.067	0.734
Words spoken	Pitch	15	0.002	0.991

Table B.3 Spearman correlation coefficients ρ between head rotation velocity and leadership and talkativeness variables calculated over different window sizes of a moving average filter for the *second* trial period. Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

Measure	Axis	Size	ρ	p-value
Leadership	Yaw	3	0.284	0.143
Leadership	Yaw	5	0.298	0.124
Leadership	Yaw	10	0.274	0.158
Leadership	Yaw	15	0.272	0.161
Leadership	Roll	3	0.462	0.013*
Leadership	Roll	5	0.459	0.014*
Leadership	Roll	10	0.458	0.014*
Leadership	Roll	15	0.428	0.023*
Leadership	Pitch	3	0.287	0.138
Leadership	Pitch	5	0.270	0.164
Leadership	Pitch	10	0.225	0.249
Leadership	Pitch	15	0.184	0.349
Talkativeness	Yaw	3	0.257	0.187
Talkativeness	Yaw	5	0.264	0.175
Talkativeness	Yaw	10	0.239	0.221
Talkativeness	Yaw	15	0.222	0.256
Talkativeness	Roll	3	0.294	0.128
Talkativeness	Roll	5	0.273	0.160
Talkativeness	Roll	10	0.249	0.202
Talkativeness	Roll	15	0.227	0.245
Talkativeness	Pitch	3	0.131	0.506
Talkativeness	Pitch	5	0.099	0.615
Talkativeness	Pitch	10	0.039	0.842
Talkativeness	Pitch	15	−0.035	0.859
Words spoken	Yaw	3	0.239	0.230
Words spoken	Yaw	5	0.240	0.228
Words spoken	Yaw	10	0.217	0.278
Words spoken	Yaw	15	0.200	0.318
Words spoken	Roll	3	0.090	0.656
Words spoken	Roll	5	0.101	0.615
Words spoken	Roll	10	0.124	0.538
Words spoken	Roll	15	0.129	0.522
Words spoken	Pitch	3	0.148	0.462
Words spoken	Pitch	5	0.147	0.464
Words spoken	Pitch	10	0.078	0.698
Words spoken	Pitch	15	−0.008	0.969

Table B.4 Spearman correlation coefficients ρ between head rotation velocity and leadership and talkativeness variables calculated over different window sizes of a moving average filter for the *AR HMD* condition. Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

Measure	Axis	Size	ρ	p-value
Leadership	Yaw	3	0.460	0.014*
Leadership	Yaw	5	0.474	0.011*
Leadership	Yaw	10	0.494	0.007**
Leadership	Yaw	15	0.497	0.007**
Leadership	Roll	3	0.401	0.034*
Leadership	Roll	5	0.393	0.038*
Leadership	Roll	10	0.404	0.033*
Leadership	Roll	15	0.400	0.035*
Leadership	Pitch	3	0.372	0.051
Leadership	Pitch	5	0.358	0.061
Leadership	Pitch	10	0.323	0.093
Leadership	Pitch	15	0.279	0.150
Talkativeness	Yaw	3	0.453	0.016*
Talkativeness	Yaw	5	0.445	0.018*
Talkativeness	Yaw	10	0.443	0.018*
Talkativeness	Yaw	15	0.435	0.021*
Talkativeness	Roll	3	0.358	0.062
Talkativeness	Roll	5	0.335	0.082
Talkativeness	Roll	10	0.310	0.109
Talkativeness	Roll	15	0.315	0.102
Talkativeness	Pitch	3	0.280	0.149
Talkativeness	Pitch	5	0.269	0.167
Talkativeness	Pitch	10	0.213	0.277
Talkativeness	Pitch	15	0.163	0.408
Words spoken	Yaw	3	0.247	0.214
Words spoken	Yaw	5	0.261	0.189
Words spoken	Yaw	10	0.233	0.243
Words spoken	Yaw	15	0.215	0.282
Words spoken	Roll	3	0.179	0.370
Words spoken	Roll	5	0.196	0.327
Words spoken	Roll	10	0.187	0.349
Words spoken	Roll	15	0.170	0.397
Words spoken	Pitch	3	0.241	0.226
Words spoken	Pitch	5	0.244	0.221
Words spoken	Pitch	10	0.192	0.337
Words spoken	Pitch	15	0.150	0.455

Table B.5 Spearman correlation coefficients ρ between head rotation velocity and leadership and talkativeness variables calculated over different window sizes of a moving average filter for the *VR HMD* condition. Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

Table B.6 Results of t-tests for all variables between the two HMD type conditions (AR and VR), split by the two trial periods. Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

	HMD	AR		VR					
Period	Variable	M	SD	M	SD	df	t	p	Cohen's d
1	Presence	2.07	1.67	2	1.31	26.51	0.12	0.904	0.05
	Co-presence	3.80	2.93	3.93	2.12	25.50	-0.14	0.888	-0.05
	Accord	6.06	0.73	5.61	1.25	22.67	1.2	0.243	0.43
	Leadership	0.51	0.1	0.49	0.1	28	0.56	0.581	0.21
	Talkativeness	0.51	0.11	0.49	0.11	28	0.62	0.540	0.23
	Words spoken	884.23	322.84	781.08	197.06	19.85	0.98	0.337	0.39
	Distance travelled	157.74	51.48	179.48	113.68	18.12	-0.65	0.523	-0.25
	Median yaw	0.23	0.06	0.17	0.05	25.06	2.68	0.013*	0.91
	Median roll	0.07	0.03	0.04	0.01	18.89	3.5	0.002**	1.11
	Median pitch	0.08	0.03	0.05	0.02	20.27	3.17	0.005**	1.04
2	Presence	2.87	2	2.73	2.19	27.77	0.17	0.863	0.06
	Co-presence	2.47	2.45	3.53	2.42	28	-1.20	0.240	-0.44
	Accord	5.74	0.91	6.24	0.45	20.41	-1.89	0.073	-0.66
	Leadership	0.5	0.07	0.5	0.07	28	0.22	0.828	0.08
	Talkativeness	0.49	0.06	0.51	0.06	28	-1.07	0.295	-0.39
	Words spoken	827.14	243.52	784.64	357.46	22.93	0.37	0.717	0.14
	Distance travelled	130.45	54.19	178.37	58.43	25.85	-2.25	0.033*	-0.79
	Median yaw	0.22	0.04	0.22	0.05	24.71	0.22	0.831	0.08
	Median roll	0.07	0.02	0.06	0.02	25.35	1	0.329	0.38
	Median pitch	0.07	0.02	0.06	0.02	25.35	1	0.329	0.38

Table B.7 Results of t-tests for all variables between the two trial periods, split by the HMD type conditions (AR and VR). Significance indicated with $^{*}p < .05$, $^{**}p < .01$ $^{***}p < .001$.

	Period	1		2					
HMD	Variable	M	SD	M	SD	df	t	*p*	Cohen's d
AR	Presence	2.07	1.67	2.87	2	27.15	-1.19	0.244	-0.43
	Co-presence	3.8	2.93	2.47	2.45	27.12	1.35	0.187	0.49
	Accord	6.06	0.73	5.74	0.91	26.81	1.04	0.307	0.38
	Leadership	0.51	0.1	0.5	0.07	25.75	0.23	0.823	0.08
	Talkativeness	0.51	0.11	0.49	0.06	21.45	0.74	0.465	0.27
	Words spoken	884.23	322.84	827.14	243.52	22.29	0.52	0.611	0.2
	Distance travelled	157.74	51.48	130.45	54.19	25.93	1.37	0.184	0.51
	Median yaw	0.23	0.06	0.22	0.04	22.67	0.16	0.877	0.06
	Median roll	0.08	0.03	0.08	0.02	20.11	0	0.997	0
	Median pitch	0.08	0.03	0.08	0.02	20.11	0	0.998	0
VR	Presence	2	1.31	2.73	2.19	22.9	-1.11	0.277	-0.41
	Co-presence	3.93	2.12	3.53	2.42	27.53	0.48	0.634	0.18
	Accord	5.61	1.25	6.24	0.45	17.55	-1.84	0.083	-0.65
	Leadership	0.49	0.10	0.50	0.07	25.75	-0.23	0.823	-0.08
	Talkativeness	0.49	0.11	0.51	0.06	21.45	-0.74	0.465	-0.27
	Words spoken	781.08	197.06	784.64	357.46	20.52	-0.03	0.974	-0.01
	Distance travelled	179.48	113.68	178.37	58.43	19.42	0.03	0.974	0.01
	Median yaw	0.17	0.05	0.22	0.05	25.98	-2.59	0.016*	-0.89
	Median roll	0.04	0.01	0.06	0.02	20.90	-2.34	0.029*	-0.82
	Median pitch	0.05	0.02	0.07	0.02	22.32	-2.09	0.049*	-0.74

Appendix C

Appendix – Chapter 4

C.1 Formative Study: Tools Summary

In this section, we present two plots summarising the tools participants reported using in the formative survey to develop site-specific AR experiences. Figure C.1 illustrates that participants most frequently used Unity, Blender, and Figma in their design and development processes. Figure C.2 shows that Niantic SDK, ARFoundation, and ARCore were the most commonly used AR frameworks.

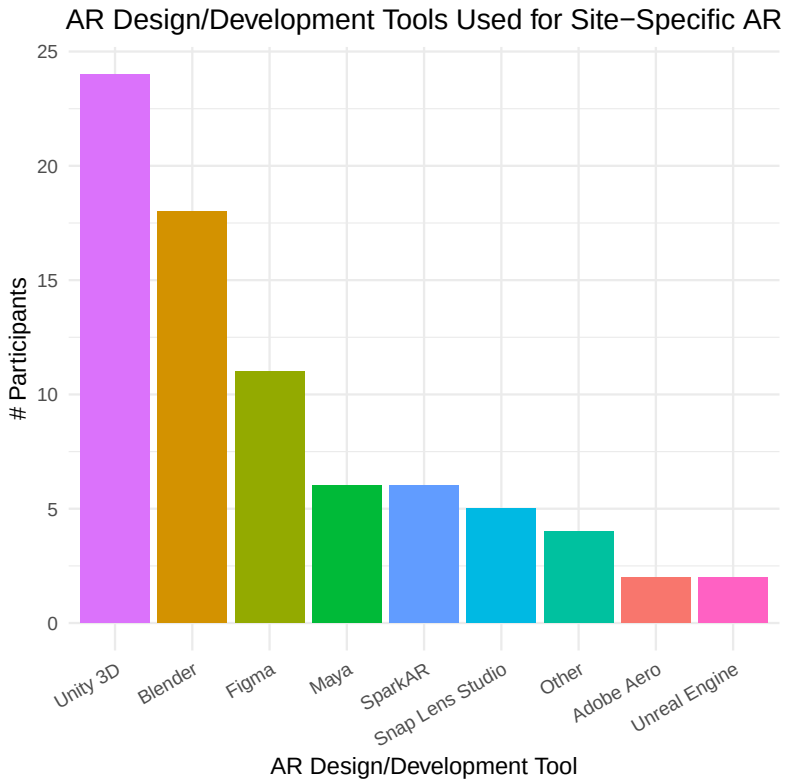


Figure C.1 The number of participants who reported using various AR development/design tools on the formative survey. Unity was the most-reported tool.

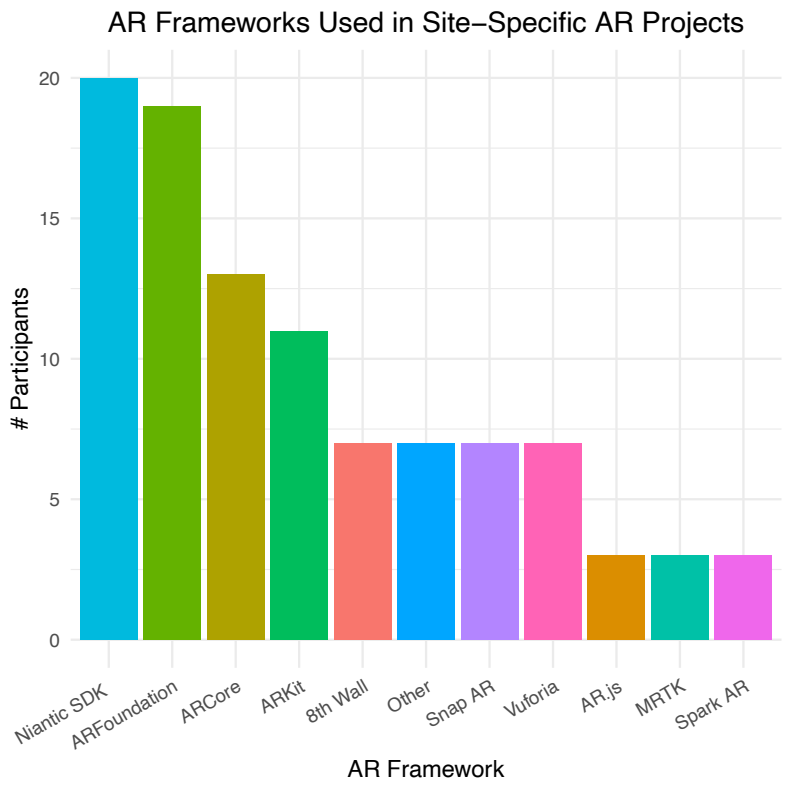


Figure C.2 The number of participants who reported using various AR frameworks on the formative survey. Niantic SDK was the most-reported AR framework.

C.2 System Walkthrough

In this scenario, a small AR agency is tasked with creating an AR tour for a university's open day. The lead developer initiates the project remotely using pre-captured spatial data and then collaborates in real time with an on-site client to ensure the AR experience accurately reflects the physical environment.

The developer begins by importing a pre-captured *location mesh* of the university campus from Niantic's Geospatial Browser into Unity. They place AR assets, such as interactive signs, narration points, and game elements, based on this mesh. Next, they initiate a collaborative session with the in-situ client. The client explores the AR tour on-site while the ex-situ developer monitors their progress remotely via a live stream of the client's screen within Unity. Early in the session, the developer notices a misalignment issue: a poster that was intended to be placed on a building wall appears embedded within the structure.

Using the **3D Cursor**, the ex-situ developer directs the in-situ client to the problematic poster. With **real-time synchronisation**, the ex-situ developer adjusts the position and colour of the poster, with the changes immediately reflected on the client's device, resolving the misalignment.

As the session continues, the developer observes—both from the *location mesh* and the client's streamed view—that certain elements are missing from the pre-captured mesh. Specifically, the entrance steps to a prominent building and some newly installed information stands are absent. To address these gaps, the developer asks the client to capture additional spatial data. The client uses **3D Coarse Mesh** capture to scan both the steps and the information stands, streaming the updated mesh in real time to the developer's Unity editor. The developer then integrates this new information, ensuring the AR content aligns with the physical space.

In another instance, the developer notices that the mesh lacks fine details of a nearby statue, including its face, hands, and the objects it holds. The ex-situ developer requests more detailed information, prompting the client to use **3D Snapshot** to capture a point cloud of the statue. This detailed point cloud is transmitted to the ex-situ developer, enabling them to refine the AR assets and ensure precise placement and detail.

Annotations are extensively used throughout the session to facilitate collaboration. At one point, the ex-situ developer asks the in-situ client to illustrate the expected flow of visitors during the open day. The client uses **Air Draw** to sketch visitor paths in the air, helping the developer visualise movement patterns and adjust the AR content accordingly.

Additionally, the client employs **Surface Draw** to propose specific locations for additional AR elements, such as interactive signs or objects. As they brainstorm together, the client sketches ideas on the ground and walls, while the developer provides input on which locations may be most engaging or logistically feasible. All annotations are synchronised and saved within the Unity scene, preserving spatial data and collaborative decisions for future iterations.

By the end of the session, the ex-situ developer has made significant real-time adjustments to the AR experience, incorporating both new spatial data and client feedback. Since the captures and annotations are saved for future use, they not only resolve immediate issues but also streamline later stages of the design process and reduce the need for further site visits.

C.3 Task Design: AR Prototypes in Two Locations

The user study involved two AR prototypes designed for distinct outdoor environments, each with a unique theme. The prototypes were developed based on the spatial characteristics of their respective sites by two authors of this work, primarily using information from the ***location mesh***, supplemented with data from Google Maps, Google Street View, and global positioning system (GPS) to inform the placement of AR elements relative to the *location mesh*. The prototypes aimed to integrate interactive and decorative AR elements with the physical surroundings while emphasising the local context of each location. Each prototype was tested on-site, and seven specific issues were identified and verified before the user study. These issues are shown in Fig. C.3 and Fig. C.4.

C.3.1 *Location* A: Welcoming Tourists to the City

Location A, situated near a major transit station, experienced high foot traffic, a nearby road, and elevated environmental noise. The AR experience at *Location*



Figure C.3 The virtual objects in Prototype A, which are described in Section C.3.1. The labels A1, A2, etc. refer to the issue types in Section C.3.3.

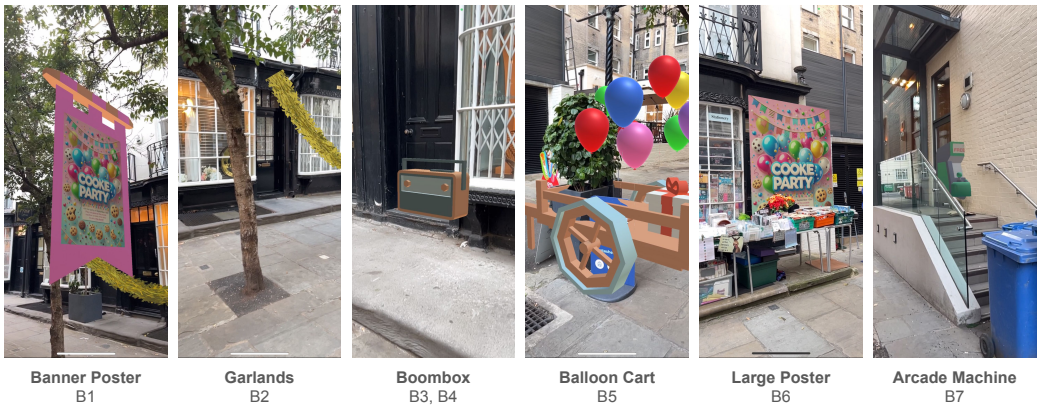


Figure C.4 The virtual objects in Prototype B, which are described in Section C.3.2. The labels B1, B2, etc. refer to the issue types in Section C.3.3.

A followed the theme *"Welcoming Tourists to the City"*, reflecting its proximity to the transit hub and its intended function as a guide for visitors. Below, we describe the virtual objects in the prototype, referencing the identified issues (e.g., A1) detailed in Section C.3.3.

- **Guide Avatar** (A1, A2): Positioned using the *location mesh*, the guide avatar was designed to interact with tourists upon arrival, providing location-specific audio information as they approached.
- **Public Transport Map** (A3): Positioned near a street lantern, the transport map was placed using the *location mesh* as a reference, supplemented by Google Street View data. It aimed to orient tourists and provide an overview of local transit options.
- **Food and Drink Items** (A4): These virtual elements were placed on a barrel table using the *location mesh* to complement the bar's physical setting.
- **Train Track Decoration** (A5): Positioned around the building's exterior using the *location mesh*, the train track emphasised the transportation-related theme of the location.
- **Information Stand** (A6): Located in an open space across from the bar, the information stand was placed using the *location mesh* and provided additional details for tourists.

- **Directional Sign** (A7): An arrow pointing toward the train station was placed using the *location mesh*, GPS, and Google Street View data. It was intended to help tourists navigate back to the station after exploring the AR experience.

C.3.2 *Location* B: Promoting and Celebrating the Upcoming Cookie Party

Location B was situated along a narrow street with a local cafe, characterised by significant foot and bicycle traffic and lower environmental noise compared to *Location* A. The AR experience at *Location* B followed the theme *"Promoting and Celebrating the Upcoming Cookie Party"*, aligning with the celebratory atmosphere of nearby restaurants and cafes. Below, we describe the virtual objects in the prototype, referencing the identified issues (e.g., B1) detailed in Section C.3.3.

- **Banner Poster** (B1): Positioned using the *location mesh*, the banner poster was designed to hang from a tree and promote the upcoming cookie party to pedestrians below.

- **Garlands** (B2): Decorative garlands were placed between tree trunks and across streetlights, with their positioning determined by the *location mesh*.

- **Boombox** (B3, B4): Placed on a stool near the cafe using the *location mesh*, the boombox provided festive background music.

- **Balloon Cart** (B5): Positioned at the centre of the street using the *location mesh*, this cart featured balloons, gifts, and cookies.

- **Large Poster with Event Information** (B6): Affixed to a nearby wall using the *location mesh*, this large poster provided details about the cookie party, including how to win a free ticket. It was designed to be clearly visible to pedestrians.

- **Arcade Machine** (B7): Placed further down the street near the event venue, the arcade machine was positioned using the *location mesh* as a reference,

supplemented by GPS and Google Street View data. A purple arrow guided users toward the machine.

C.3.3 Task Components – Issues

After developing the AR prototypes for each location remotely, we identified at least seven issues per prototype during on-site testing before the user study. We intentionally left a specific set of seven issues at each location for participants to address, selecting them to ensure coverage of the issue types identified in our formative study. These issues are listed below and classified by type in Table C.1.

To inform the design of the prototypes, we also drew inspiration from existing work on design patterns in AR [290]. As shown in Table C.1, we classified each component based on the dimensions described in Section 4.5. These issues were used in the *refinement* segment of the user study task and are detailed below.

***Location* A: Welcoming Tourists to the City**

- A1. The guide avatar is difficult to find.
- A2. The guide avatar is hard to hear over the traffic noise.
- A3. The map is inconveniently placed and near (foot) traffic.
- A4. The drinks and food on the table appear to float in the air.
- A5. The train track decoration and striped ring do not look good from all perspectives and block the bar's logo.
- A6. The information stand is in an odd place.
- A7. The sign with the arrow pointing to the train station is hard to see from the information stand.

***Location* B: Promoting and Celebrating the Upcoming Cookie Party**

- B1. The banner poster is not attached to anything.
- B2. The garlands between the trees do not appear correctly placed.

- B3. The volume of the boombox with background music is too loud.
- B4. The boombox is floating.
- B5. The balloon cart is touching/inside another object.
- B6. The large poster is hard to see, and reading it requires caution due to bike traffic.
- B7. The Free Ticket Arcade machine is out of reach.

	Design Patterns						Referent Types					Alignment Type			Issue Types				
	Glyphs	Decal	Trajectories	Label	Ghost	Audio	Small object	Small plane/area	Large object	Large plane/area	Building structure	Overlap	Proximity	Surface	Physical constraints	User safety	Misalignment	User context	User flow
A1					✓					✓			✓	✓	✓		✓		✓
A2						✓				✓			✓					✓	
A3	✓	✓							✓					✓	✓	✓		✓	✓
A4					✓			✓						✓			✓		
A5			✓		✓						✓	✓					✓		✓
A6				✓	✓					✓				✓			✓		✓
A7	✓		✓	✓									✓	✓					✓
B1	✓	✓		✓					✓			✓					✓		
B2			✓						✓		✓						✓		
B3						✓							✓					✓	
B4					✓			✓											
B5										✓				✓	✓		✓		
B6		✓									✓			✓	✓	✓	✓		✓
B7	✓		✓	✓	✓			✓						✓	✓				

Table C.1 Design patterns, referent types, alignment types, and issue types in AR element design.

C.4 Full Results of CoCreatAR User Study

This section presents the full findings from our user study. We employed linear mixed-effects models to analyse the data. Although we followed a Latin Square counterbalanced study design, we analysed the data with consideration of a

potential confounding factor, *location*. We also evaluated the data for any effect of participants' prior experience. Participant variability was modelled as a random effect, and all models converged successfully. Unless otherwise noted, the assumptions of the linear mixed-effects models were satisfied (including linearity, normality of residuals, homoscedasticity, and independence).

We separately examined the performance of ex-situ and in-situ participants, focusing on key metrics such as engagement and task load relevant to our hypotheses. As described in Section 4.6, Cronbach's alpha values indicated acceptable internal consistency for both engagement (0.740 for in-situ, and 0.764 for ex-situ) and task load items (0.812 for in-situ, and 0.772 for ex-situ).

C.4.1 Engagement

Engagement scores were analysed using mixed-effects linear models, incorporating within-subject variability (participant as a random effect), location, and prior experience with smartphone-based AR (for in-situ users) or 3D editing tools (for ex-situ users).

C.4.1.1 Engagement by Condition

For ex-situ participants, we observed a significant effect of condition on engagement scores. The estimated coefficient for the Sync condition was 0.500 (SE = 0.140, $p < 0.001$, 95% CI [0.226, 0.774]), indicating that participants in the Sync condition reported significantly higher engagement compared to those in the Async condition. The intercept of 3.042 (SE = 0.306, $p < 0.001$, 95% CI [2.443, 3.641]) represents the baseline engagement score in the Async condition. Participant variability accounted for a variance of 0.145 in engagement scores.

Similarly, for in-situ participants, engagement was significantly higher in the Sync condition, with a coefficient of 0.340 (SE = 0.130, $p = 0.009$, 95% CI [0.085, 0.595]). The intercept of 3.118 (SE = 0.232, $p < 0.001$, 95% CI [2.663, 3.574]) reflects the average engagement score for the Async condition. Participant variability accounted for a variance of 0.092 in engagement scores for in-situ participants.

C.4.1.2 Effects of Location

We included location as a potential confounding variable. For ex-situ participants, no significant effect of location was detected, with a coefficient of −0.181 (SE = 0.140, $p = 0.196$, 95% CI [−0.454, 0.093]), indicating that location did not substantially influence engagement scores.

However, for in-situ participants, location had a significant effect on engagement. The coefficient for *Location* B was 0.299 (SE = 0.130, $p = 0.022$, 95% CI [0.044, 0.554]), suggesting that participants at *Location* B exhibited higher engagement compared to those at *Location* A.

C.4.1.3 Effects of Prior Experience

Prior experience with AR or 3D editing tools did not significantly affect engagement scores for either group. For ex-situ participants, experience with 3D editing tools had a coefficient of 0.091 (SE = 0.054, $p = 0.093$, 95% CI [−0.015, 0.197]), while for in-situ participants, experience with AR had a coefficient of 0.090 (SE = 0.075, $p = 0.231$, 95% CI [−0.058, 0.238]). Neither effect reached statistical significance, suggesting that prior experience did not substantially impact engagement.

C.4.2 Task Load

Task load scores were analysed using mixed-effects linear models, with participant variability, location, and prior experience as factors.

C.4.2.1 Task Load by Condition

For ex-situ participants, condition did not significantly affect task load. The estimated coefficient for the Sync condition was 0.313 (SE = 0.309, $p = 0.312$, 95% CI [−0.294, 0.919]), indicating no significant difference in task load between the Sync and Async conditions. The intercept of 4.631 (SE = 0.724, $p < 0.001$, 95% CI [3.213, 6.049]) represents the baseline task load score in the Async condition. Participant variability accounted for a variance of 0.881 in task load.

Similarly, for in-situ participants, no significant effect of condition on task load was observed. The coefficient for the Sync condition was 0.213 (SE = 0.345,

$p = 0.538$, 95% CI [−0.464, 0.889]), with an intercept of 4.205 (SE = 0.556, $p < 0.001$, 95% CI [3.116, 5.294]). Participant variability accounted for a variance of 0.396 in task load.

C.4.2.2 Effects of Location

Location did not significantly affect task load for ex-situ participants, with a coefficient of −0.250 (SE = 0.309, $p = 0.419$, 95% CI [−0.856, 0.356]). However, for in-situ participants, location showed a marginal effect on task load, with a coefficient of −0.637 (SE = 0.345, $p = 0.065$, 95% CI [−1.314, 0.039]), indicating a tendency for lower task load at *Location* B compared to *Location* A, though this effect did not reach statistical significance.

C.4.2.3 Effects of Prior Experience

Background experience did not significantly influence task load for either group. For ex-situ participants, experience with 3D editing tools had a coefficient of −0.181 (SE = 0.129, $p = 0.162$, 95% CI [−0.434, 0.073]), while for in-situ participants, experience with AR had a coefficient of −0.320 (SE = 0.176, $p = 0.070$, 95% CI [−0.666, 0.026]). Neither effect was statistically significant, suggesting that prior experience did not substantially affect task load.

C.4.3 Confidence in Authored Result

To assess the impact of condition on participants' confidence in their authored results, we analysed the number of issues participants reported to have confidently fixed. In this case, the dependent variable represents the number of issues participants indicated they "fixed and feel confident about." In addition, we analysed the self-rated overall confidence score. This analysis included participant variability as a random effect, with location and prior experience with 3D editing tools as additional factors.

C.4.3.1 Confidence by Condition

The analysis revealed a significant effect of condition on the number of confidently fixed issues. The estimated coefficient for the Sync condition was 3.000 (SE = 0.373, $p < 0.001$, 95% CI [2.270, 3.730]), indicating that participants in the

Sync condition fixed significantly more issues with confidence compared to those in the Async condition. The intercept of 0.578 (SE = 0.710, $p = 0.416$, 95% CI [−0.813, 1.969]) represents the average number of confidently fixed issues in the Async condition. Participant variability accounted for a variance of 0.425.

In addition, participants in the Sync condition reported significantly higher overall confidence scores than those in the Async condition (SE = 0.387, $p = 0.004$, 95% CI [0.367, 1.883]), with the coefficient for Sync being 1.125. Participant variability accounted for a variance of 0.258.

C.4.3.2 Effects of Location

We included location as a potential confounding variable. For ex-situ participants, there was a marginal effect of location on the number of confidently fixed issues. The coefficient for *Location* B was −0.667 (SE = 0.373, $p = 0.074$, 95% CI [−1.397, 0.064]), suggesting that participants at *Location* B tended to fix fewer issues with confidence compared to those at Location A, though this effect did not reach conventional statistical significance.

Similarly, location did not have a significant impact on overall confidence scores, with the coefficient for *Location* B being 0.125 (SE = 0.387, $p = 0.747$, 95% CI [−0.633, 0.883]).

C.4.3.3 Effects of Prior Experience

Prior experience with 3D editing tools did not significantly affect the number of confidently fixed issues, with a coefficient of 0.195 (SE = 0.117, $p = 0.096$, 95% CI [−0.034, 0.424]). While this effect was not statistically significant, it suggests a potential positive association between prior experience and confidence in fixing issues.

The analysis also revealed no significant effect of prior experience on overall confidence scores. The coefficient for prior experience was 0.078 (SE = 0.106, $p = 0.466$, 95% CI [−0.131, 0.286]).

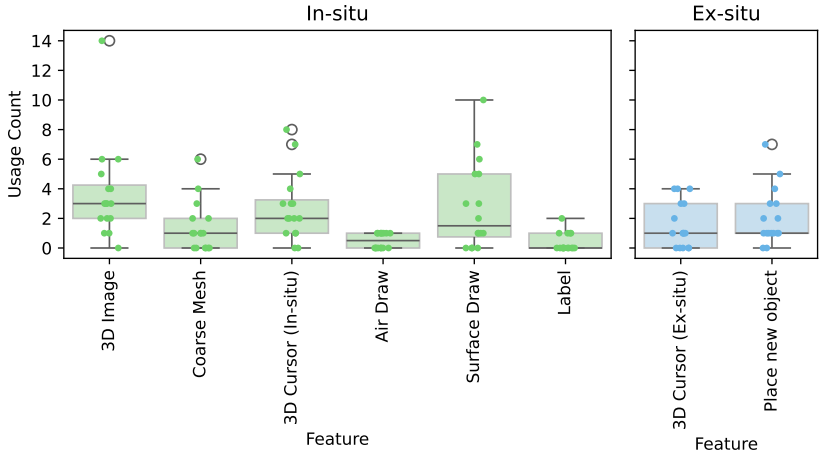


Figure C.5 Box plot representing the distributions of feature counts per group, with an additional visualisation of individual data points. Green bars and points represent in-situ features, whereas blue represents ex-situ features.

C.4.4 Feature usage during the sessions

The task recordings were systematically coded to document the usage frequency of each feature, along with qualitative observations on their application (e.g., using *Coarse 3D Mesh* to capture building geometry). The resulting distribution of feature usage counts is visualised in Fig. C.5.

C.5 Survey and Semi-Structured Interview Questions

We outline the validated questionnaires, custom questions, and semi-structured interview questions in this section.

C.5.1 Formative Study Questions

The formative study's survey and semi-structured interview questions are in Listing C.1 and C.2, respectively.

Listing C.1 The formative study's survey questions.

1. Which of the following best describes your current occupation? Please select all that apply.

[Select multiple options, including Software Engineer, Technical Artist, UX Designer, ..., Other – Write In]

2. What best describes the type of organisation you work for?

[Select one option, including Small Business, Agency, Enterprise, ..., Other - Write In]

3. How long have you been developing AR experiences?

[Select one option, including Less than a year, 1-2 years, 3-5 years, 6-10 years, Over 10 ↪ years]

4. How long have you been developing location-based AR experiences?

[Select one option, including Less than a year, 1-2 years, 3-5 years, 6-10 years, Over 10 ↪ years]

5. How would you describe the content you create? Please select all that apply.

[Select multiple options, including Long form (games, movies, other non-marketing ↪ experiences, etc), Short form (ads or marketing experiences, AR filters, product ↪ demos, etc)]

6. With brief bullet points, please describe 1-3 of your favourite location-based AR ↪ applications you have worked on. Feel free to add a link to your portfolio or ↪ projects.

[Long answer]

7. For what target group do you typically build location-based AR experiences? For ↪ example, 'adult grocery shoppers', 'gamers aged 20-25' etc.

[Long answer]

8. How many other team members typically work with you on a location-based AR ↪ project?

[Number ___ team members, on average]

9. What are the roles of the other team members working on location-based AR projects ↪ with you? Please select all that apply.

[Select multiple options, including Software Engineer, Technical Artist, UX Designer, ...,
↪ Other - Write In]

10. Which AR frameworks have you used in location-based AR projects?
[Select multiple options, including ARFoundation, ARCore, ARKit, Snap AR]

11. What development/design suites or tools have you used in your location-based AR
↪ development process?
[Select multiple options, including Unity 3D, Unreal Engine, Adobe Aero, ..., Other - Write
↪ In]

12. Which of the following tools do you use to leverage real-world information in the
↪ development process of location-based AR experiences? Please select all that
↪ apply.
[Select multiple options, including Niantic Lightship VPS, Remote Authoring, Google
↪ Geospatial Creator (combined with Unity or Adobe Aero), Google Earth (or
↪ alternatives by Apple, Microsoft), ..., Other - Write In]

13. Approximately how much time do you spend testing your location-based AR
↪ experiences on-site (i.e. physically being at the location) vs off-site (e.g. virtually
↪ testing with a mesh or playback simulator)?
[____ % of time spent testing on-site (i.e., visiting the location in the real world)]
[____ % of time spent testing off-site in a simulation (e.g., in Unity/Adobe Aero using a
↪ mesh of the environment)]

14. Which of the following factors do you look to identify when testing your AR experience
↪ on-site (i.e. at the real-world location)?
[User context (e.g. lighting, weather, noise, crowdedness, distractions), Physical constraints
↪ (e.g. real-world obstacles that were not accounted for such as parked cars,
↪ temporary structures, closed gates), ..., Other - Write In]

15. How do you document or report bugs or improvements when testing a location-based

↪ AR experience on-site (i.e. physically being at the location)? For example, by
↪ writing a note on my phone in the moment, recording my screen, or remembering
↪ and writing it down later.
[Long answer]

Listing C.2 The formative study's semi-structured interview questions.

1. When starting a new location-based AR project, what are the key elements of the
↪ physical location that you consider?
(How do these elements influence your design decisions?)

2. How do you initially conceptualize the interaction between the user, the AR content, and
↪ the physical location?
(Could you share any specific tools or techniques you find essential during this phase?)

3. You mentioned that you use tools such as [SURVEY RESPONSE]. Could you give an
↪ example of your workflow in developing a location-based AR experience?
(At what point did you test the experience?)

--- Challenges and Solutions ---
4. You mentioned [SURVEY RESPONSE] in your survey response. What have been the
↪ most significant challenges in aligning the AR content with the physical world?
(How have you addressed these challenges?)

5. Can you share an example where you had to significantly modify or even discard a
↪ feature due to unforeseen real-world elements or constraints?
(How did you navigate this setback?)

--- Testing and Refinement ---
6. Thinking about on-site vs off-site testing (e.g. Playback), [SURVEY RESPONSE]. What
↪ are the most essential factors that require you to test on-site?

7. How do you replicate physical conditions when testing off-site?

8. You mentioned you [SURVEY RESPONSE] to collect/integrate feedback from testing
↪ the experience. Could you elaborate on that a bit?

--- Collaboration and Workflow ---

9. How does your team collaborate on location-based AR projects, especially when dealing
↪ with complex, location-specific challenges?
(Are there tools or practices that you use to facilitate effective teamwork?)
(Do you ever engage in real-time collaboration through video calls or other tools?)

--- Vision for the Future and Wrap-up ---

10. What features or capabilities would your ideal toolset for developing location-based AR
↪ experiences include?
(How would these tools address the current limitations you face?)

11. Is there anything not covered in this interview that you think is crucial for understanding
↪ the development and design of location-based AR experiences that you'd like to
↪ share?

C.5.2 User Study

The user study's surveys (Listings C.3, C.4, and C.5) and semi-structured interview questions for ex-situ (Listing C.6) and in-situ (Listing C.7) participants are presented below. As described in the listings, the NASA-TLX [291] and UES-SF [292] questionnaires were used to measure task load and engagement, respectively.

Listing C.3 The pre-task survey questions.

--- Demographics ---

1. What is your age?

[Number greater than 17]

2. What is your gender?

[Select one option, including Woman, Man, Non-binary, Prefer not to say, Other - Write in]

3. What is your current occupation? If you are a student, please include your major.
[Short answer]

4. How often do you use the following? Please select the most appropriate option:
[Select one option for the following, including Several times a day, Once a day, ..., Never]
a. Augmented Reality (smartphone)
b. PC or Console Video Games

5. Have you ever used a 3D editor or game development platform, like Unity, Unreal
↪ Engine, Maya, Blender, etc.?
[Select either Yes or No]

[If yes...]
6. How often do you use 3D editors or game development platforms?
[Select one option, including Several times a day, Once a day, ..., Never]

7. Have you ever contributed to the design or development of a 3D game, movie or
↪ experience? (e.g. as a programmer, project manager, designer)
[Select either Yes or No]

[If yes...]
8. Have you ever contributed to the design or development of an augmented reality
↪ experience? (e.g., as a programmer, project manager, designer)
[Select either Yes or No]

[If yes...]
9. Have you ever contributed to the design or development of a location-based
↪ augmented reality experience? (e.g. as a programmer, project manager, designer)
[Small note:] Location-based augmented reality experiences are designed for specific
↪ physical locations, requiring users to be at that location to interact with the
↪ experience. These experiences often use a visual positioning system (VPS), such

↪ as the Niantic Lightship VPS or the Google ARCore Geospatial API, to function.

[Select either Yes or No]

10. Any other notes?

[Long answer]

Listing C.4 The post-task survey questions for ex-situ participants.

1. For each statement, please use the following scale to indicate what is most true for you.

[Select one option from Strongly disagree to Strongly agree on a 5-point Likert scale for ↪ each of the following]

a. I feel confident that the authored AR experience would function as I expect it to in the ↪ real world.

b. I felt like I was able to accurately fix the issues with the initial AR experience.

c. I felt like I was able to effectively extend the AR experience with the information given.

2. Why did you answer the way you did in the previous questions? (Please write at least 1-2 ↪ sentences.)

[Long answer]

--- Task Load ---

[To measure Task Load, we used the Mental Demand, Temporal Demand, Performance, ↪ Effort, and Frustration questions from the NASA-TLX.]

--- Engagement ---

[To measure engagement, we used the UES-SF.]

--- Task-Specific Questions ---

1. Which location did you work on?

[Select one option, including Location A or Location B]

[If Location A:]

2. Please rate how you feel you performed on the different task items. [Select one option, including Fixed it and feel confident about the result, Fixed it but do not feel confident about the result, Did not fix it because I did not have enough time, Did not fix it because I did not have enough information, Other (write below)]

a. The Guide avatar is difficult to find.

b. The Guide avatar is hard to hear over the traffic noise.

c. The map is inconveniently placed and near (foot) traffic.

d. The drinks and food on the table appear to float in the air.

e. The train track decoration and striped ring do not look good from all perspectives and block the bar's logo.

f. The information stand is in an odd place.

g. The sign with the arrow pointing to the train station is hard to see from the information stand.

[If Location B:]

3. Please rate how you feel you performed on the different task items. [Select one option, including Fixed it and feel confident about the result, Fixed it but do not feel confident about the result, Did not fix it because I did not have enough time, Did not fix it because I did not have enough information, Other (write below)]

a. The banner poster is not attached to anything.

b. The garlands between the trees do not appear correctly placed.

c. The volume of the boom box with background music is too loud.

d. The boom box is floating.

e. The balloon cart is touching/inside another object.

f. The large poster is hard to see, and reading it requires caution due to bike traffic.

g. The Free Ticket Arcade machine is out of reach.

Listing C.5 The post-task survey questions for in-situ participants.

--- Task Load ---

[To measure Task Load, we used the Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration questions from the NASA-TLX.]

--- Engagement ---

[To measure engagement, we used the UES-SF.]

Listing C.6 The semi-structured interview questions for the ex-situ participant.

--- Location ---

1. Please draw as much of the location as you can remember in a simple top-down map ↪ without looking at Unity or any other maps/tools.

We know this might be tricky! No worries if you're not accurate, just be as detailed as you ↪ can in 5 minutes. Feel free to use the drawing tool or shapes.

2. Have you ever been to this location before?

--- System ---

3. Given that you had to develop a location-based AR experience, which method (async or ↪ sync) would you use? Why?

4. Which method did you enjoy using more? Why?

5. Which method made you feel more confident about the end result? Why?

6. What were some of the pros and cons of each method (async and sync)?

   a. Are there specific cases where you would use one method vs the other?

--- Future ---

7. Are there any features you can think of that you would want to add to the sync and/or ↪ async method, e.g. to address the cons?

8. How did you feel about the effectiveness of your communication in each method (sync ↪ vs. async)? Do you feel like the capture and annotation features helped you in any ↪ way?

9. Anything else?

Listing C.7 The semi-structured interview questions for the in-situ participant.

1. Given that you had to help test a location-based AR experience, which method (async or sync) would you use? Why?

2. Which method did you enjoy using more? Why?

3. What were some of the pros and cons of each method (async and sync)?
   a. Are there specific cases where you would use one method vs the other?

4. How did you feel about the effectiveness of your communication in each method (sync vs. async)?

5. How was your experience using the features for the synchronous condition (e.g. drawing on surfaces or in the air, meshing, and 3D colour images)? Did you use them? Why or why not?
   a. Are there any additional features you can think of that you would want to add to the sync and/or async method?

6. Anything else?

Appendix D

Appendix – Chapter 5

D.1 Contextually Adaptive Prompt Inference: Full System Prompt

System Prompt You are a helpful assistant that acts like highly creative interior architect and photographer. You are given descriptions of images captured by a camera on a tripod placed exactly in the middle of an open indoor space at a height of 1.5 meters. This camera has a field-of-view of 55 degrees and only takes square images. You will be given a set of image descriptions with Y rotation values of the camera, as well as Y rotation values without a description. From your perspective as a creative interior architect, your task is to describe what you expect to see for the image that will be taken at the Y rotation value with unknown contents. You receive the rotation value and descriptions in JSON format. Do not use words such as 'blend, 'transition', 'fusion', 'mix', 'transformation' (or synonyms), but concretely describe the novel contents and salient objects you expect to see in the area without repeating objects. Use a similar format to the other given descriptions. It is not always obvious what the camera will capture as the contents of the space can be highly diverse in style and content, so please be creative and focus on coming up with new objects and artifacts that fit in. You can assume that the objects appearing in the known image description do not show up in the other image (all objects are fully contained within the image frame).

Do not include mentions of the shape of the room (e.g., corner). Use comma-separated descriptions (e.g., instead of 'On the sticky note wall, a whiteboard marker tray holds colorful pens, a spark of color in an otherwise monochrome environment.', write 'Sticky note wall with whiteboard marker holding colorful pens, monochrome environment'). Always start descriptions with '... space with' where '...' is the type of the room (e.g., living room, kitchen, etc.). Only return the descriptions with the set_description function, without any explanation. Keep your descriptions short please, without adding too many different items/objects, with 20 words or less for each description.

User Prompt The size of the room is space_size_str (WxHxL) meters and the camera is positioned in the middle. These are the Y rotation values and descriptions of the images that were already taken: y_rotations_and_descriptions. What do you expect for the following Y rotation values: y_rotations_without_descriptions? Consider the theme of "" when coming up with the descriptions

D.2 ControlNet-Layout: Additional Output Samples



Figure D.1 Additional results of the ControlNet-Layout model. Each of these output images was generated by combining Control-Layout and ControlNet-Depth with weights 0.6 and 0.3, respectively.

D.3 Comparison of ControlNet-Layout and LooseControl



Figure D.2 Comparative image generation output of the recent LooseControl model and our ControlNet-Layout model, including corresponding prior images and output images. The LooseControl prior and output images in this figure were reproduced from the paper's web page (`https://shariqfarooq123.github.io/loose-control`) with permission from the authors. The ControlNet-Layout prior images were manually created to match the room structure depicted in the LooseControl prior images.

D.4 Influence of Submesh Shapes on Geometric Prior Shape

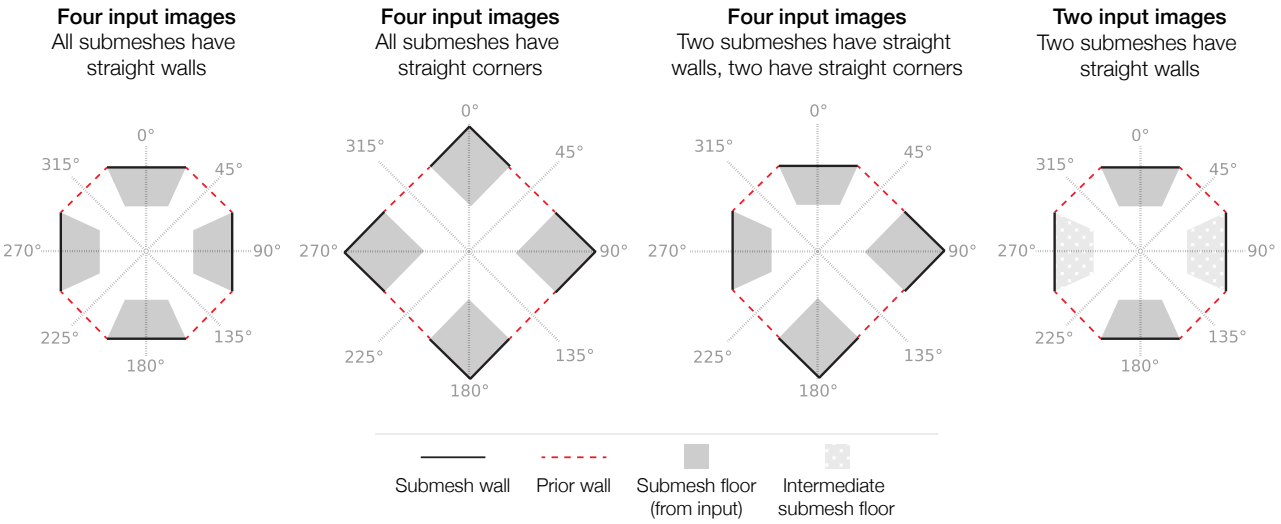


Figure D.3 Visual explanation of the impact of submesh count and shape on the shape of the geometric prior mesh and final blended space.

D.5 Examples of Spaces Generated with Various Numbers of Submeshes



Figure D.4 Examples of spaces generated with various numbers of submeshes. Left: a SPACEBLENDER mesh based on four input images, all featuring corners. Right: a SPACEBLENDER mesh based on five input images featuring a mixture of room shapes.

D.6 Preliminary User Study: Questionnaire Items

Please indicate your level of agreement with the following statements Do not agree at all; Disagree; Neutral; Agree; Fully Agree	**Category**
I felt like I was actually there in the environment of the presentation.	Self-Location
It seemed as though I actually took part in the action of the presentation.	Self-Location
It was as though my true location had shifted into the environment in the presentation.	Self-Location
I felt as though I was physically present in the environment of the presentation.	Self-Location
The objects in the presentation gave me the feeling that I could do things with them.	Possible Actions
I had the impression that I could be active in the environment of the presentation.	Possible Actions
I felt like I could move around among the objects in the presentation.	Possible Actions
It seemed to me that I could do whatever I wanted in the environment of the presentation.	Possible Actions
I noticed (my partner).	Co-presence
(My partner) noticed me.	Co-presence
(My partner's) presence was obvious to me.	Co-presence
My presence was obvious to (my partner).	Co-presence
(My partner) caught my attention.	Co-presence
I caught (my partner's) attention.	Co-presence

Table D.1 Post-task questionnaire for measuring Self-Location and Possible Actions from the Spatial Presence Experience Scale questionnaire by Hartmann et al. [380] and Co-presence items from the Networked Minds Measure of Social Presence questionnaire by Harms and Biocca [82].

D.7 Preliminary User Study: Walkthrough Scenarios



Figure D.5 Overview of environments used in the walkthrough segment of the semi-structured interview, including a *collaborative study session* scenario (left) and a *cooking class with friends* scenario (right).

D.8 Preliminary User Study: Semi-Structured Interview Questions

Questions involving clustering task:

1. You mentioned you preferred *condition*, over the others. Could you elaborate on your preference?

2. Describe the strategies you used to perform the affinity diagramming task. Did you adopt a consistent strategy in all three environments, or did you use different strategies?

Questions involving walkthrough of additional SpaceBlender environments:

1. If any, in what scenarios do you think integrating environmental context to which you have a personal connection into a virtual environment could be valuable or interesting?

2. Is there anything you wished was different about the design, composition, or function of the blended spaces you have seen that could make them more engaging, useful, or comfortable?

D.9 Preliminary User Study: Full Self-Reported Questionnaire Response Analysis

We used the Friedman test for overall comparisons, with subsequent post hoc analyses conducted via Wilcoxon signed-rank tests. All reported p-values for post hoc analyses have been adjusted using the Bonferroni correction.

D.9.1 Possible Actions

The Friedman test revealed a statistically significant difference in the scores of Possible Actions across the three conditions ($\chi^2(2) = 14.81, p < 0.0001$), suggesting that participants' perceptions of possible actions varied significantly depending on the condition. Post hoc analyses were conducted to explore these

differences further. Results showed a significant decrease (W = 5.0, $p = 0.0021$) in Possible Actions scores from GENERIC3D (M = 3.89, SD = 0.71) to TEXT2ROOM (M = 3.13, SD = 0.95), indicating a diminished perception of possible actions within the environment under TEXT2ROOM. However, the difference between GENERIC3D and SPACEBLENDER (W = 23.0, $p = 0.104$) and between TEXT2ROOM and SPACEBLENDER (W = 27.0, $p = 0.055$) did not reach statistical significance.

D.9.2 Self-Location

A similar analysis was conducted for Self-Location scores, with the Friedman test indicating a significant difference across conditions ($\chi^2(2) = 10.53, p = 0.0052$). This finding highlights that the different environments significantly affected the sense of being situated within the environment. After post hoc analysis, the comparison between GENERIC3D and TEXT2ROOM and between GENERIC3D and SPACEBLENDER did not show significant differences (W = 49.0, $p > 0.999$). However, a significant difference (W = 0.0, $p = 0.0039$) was found between TEXT2ROOM (M = 3.46, SD = 0.78) and SPACEBLENDER (M = 3.91, SD = 0.68), indicating a change in the sense of self-location between these two conditions.

D.9.3 Co-Presence

The analysis of co-presence scores using the Friedman test did not reveal a statistically significant difference across conditions ($\chi^2(2) = 1.59, p = 0.452$).

D.9.4 Impact of Environmental Factors

The assessment of environmental factors on task performance revealed several statistically significant differences across conditions.

The analysis for *Layout* showed $\chi^2(2) = 20.22, p < 0.0001$, indicating significant variations in task performance related to environmental layout. Post-hoc analysis indicated significant differences between GENERIC3D (M = 3.92, SD = 0.68) and TEXT2ROOM (M = 3.20, SD = 0.75) (W = 10.0, $p = 0.0043$). A significant difference was also observed between TEXT2ROOM and SPACEBLENDER (M = 3.85, SD = 0.70) (W = 4.0 and $p = 0.0036$). However, the difference between GENERIC3D and SPACEBLENDER did not reach statistical significance (W = 22.0, $p = 0.4842$).

For *Visual Quality*, a significant difference was found ($\chi^2(2) = 21.73, p < 0.0001$). There was a difference in perceived support for task execution between GENERIC3D (M = 4.00, SD = 0.00) and TEXT2ROOM (M = 3.17, SD = 0.75) (W = 0.0 and $p = 0.0010$), and between TEXT2ROOM and SPACEBLENDER (M = 3.90, SD = 0.30) (W = 3.5 and $p = 0.0049$). The comparison between GENERIC3D and SPACEBLENDER did not show a significant difference (W = 18.0, $p = 0.2121$).

Familiarity also demonstrated significant differences among conditions ($\chi^2(2) = 14.56, p = 0.00069$). The difference between TEXT2ROOM (M = 3.10, SD = 0.83) and SPACEBLENDER (M = 3.95, SD = 0.22) was significant (W = 0.0 and $p = 0.0039$). Comparisons between GENERIC3D and TEXT2ROOM, and between GENERIC3D and SPACEBLENDER did not achieve significance (W = 10.0, $p = 0.1566$; W = 12.0, $p = 0.0604$).

Bibliography

[22] Zhang, Jingyi, Lu, Ziwen, Zhu, Changrui, Julier, Simon, and Steed, Anthony. 2025. Do You See What I See? Bring Live Pedestrians into an Outdoor Collaborative Mixed Reality Experience. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (UIST '25). Association for Computing Machinery, New York, NY, USA, (Sept. 27, 2025), 1–14. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747667.

[23] Saatçi, Banu, Akyüz, Kaya, Rintel, Sean, and Klokmose, Clemens Nylandsted. 2020. (Re)Configuring Hybrid Meetings: Moving from User-Centered Design to Meeting-Centered Design. *Computer Supported Cooperative Work (CSCW)*, 29, 6, (Dec. 1, 2020), 769–794. doi:10.1007/s10606-020-09385-x.

[24] Brudy, Frederik, Holz, Christian, Rädle, Roman, Wu, Chi-Jui, Houben, Steven, Klokmose, Clemens Nylandsted, and Marquardt, Nicolai. 2019. Cross-Device Taxonomy: Survey, Opportunities and Challenges of Interactions Spanning Across Multiple Devices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, (May 2, 2019), 1–28. ISBN: 978-1-4503-5970-2. Retrieved Aug. 2, 2021 from https://doi.org/10.1145/3290605.3300792.

[25] Clark, Herbert H. and Brennan, Susan E. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*. Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, (Eds.) American Psychological Association, Washington, 127–149. ISBN: 978-1-55798-121-9. doi:10.1037/10096-006.

[26] Benford, Steve, Bowers, John, Fahlén, Lennart E., and Greenhalgh, Chris. 1994. Managing mutual awareness in collaborative virtual environments. In *Proceedings of the Conference on Virtual Reality Software and Technology* (VRST '94). World Scientific Publishing Co., Inc., USA, (Aug. 26, 1994), 223–236. ISBN: 978-981-02-1867-6.

[27] Bly, Sara A., Harrison, Steve R., and Irwin, Susan. 1993. Media spaces: bringing people together in a video, audio, and computing environment. *Communications of the ACM*, 36, 1, (Jan. 1, 1993), 28–46. doi:10.1145/151233.151235.

[28] Heath, Christian and Luff, Paul. 1991. Disembodied conduct: communication through video in a multi-media office environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology - CHI '91*. The SIGCHI Conference. ACM Press, New Orleans, Louisiana, United States, 99–103. ISBN: 978-0-89791-383-6. doi:10.1145/108844.108859.

[29] Gutwin, Carl and Greenberg, Saul. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)*, 11, 3, (Sept. 1, 2002), 411–446. doi:10/c43wsg.

[30] Grønb, Jens Emil, Mackay, Wendy E, Borowski, Marcel, Beaudouin-Lafon, Michel, Hoggan, Eve, and Klokmose, Clemens N. 2023. Mirrorverse: Live Tailoring of Video Conferencing Interfaces. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology. ACM, San Francisco , CA , USA. doi:10.1145/3586183.3606767.

[31] Milgram, Paul and Kishino, Fumio. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77, 12, 1321–1329.

[32] Carlsson, C. and Hagsand, O. 1993. DIVE A multi-user virtual reality system. In *Proceedings of IEEE Virtual Reality Annual International Symposium*. Proceedings of IEEE Virtual Reality Annual International Symposium. (Sept. 1993), 394–400. doi:10.1109/VRAIS.1993.380753.

[33] Benford, Steve, Greenhalgh, Chris, Reynard, Gail, Brown, Chris, and Koleva, Boriana. 1998. Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Transactions on Computer-Human Interaction*, 5, 3, (Sept. 1998), 185–223. doi:10/bgsfq7.

[34] Billinghurst, M., Bee, S., Bowskill, J., and Kato, H. 1999. Asymmetries in collaborative wearable interfaces. In *Digest of Papers. Third International Symposium on Wearable Computers*. Digest of Papers. Third International Symposium on Wearable Computers. (Oct. 1999), 133–140. doi:10/dqx596.

[35] Koleva, Boriana, Benford, Steve, and Greenhalgh, Chris. 2002. The Properties of Mixed Reality Boundaries. In *ECSCW '99: Proceedings of the Sixth European Conference on Computer Supported Cooperative Work 12–16 September 1999, Copenhagen, Denmark*. Suanne Bødker, Morten Kyng, and Kjeld Schmidt, (Eds.) Springer Netherlands, Dordrecht, 119–137. ISBN: 978-0-306-47316-6. doi:10.1007/0-306-47316-X_7.

[36] Kraut, Robert E., Miller, Mark D., and Siegel, Jane. 1996. Collaboration in performance of physical tasks: effects on outcomes and communication. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (CSCW '96). Association for

Computing Machinery, New York, NY, USA, (Nov. 16, 1996), 57–66. ISBN: 978-0-89791-765-0. doi:10.1145/240080.240190.

[37] Oda, Ohan, Elvezio, Carmine, Sukan, Mengu, Feiner, Steven, and Tversky, Barbara. 2015. Virtual Replicas for Remote Assistance in Virtual and Augmented Reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (UIST '15). Association for Computing Machinery, New York, NY, USA, (Nov. 5, 2015), 405–415. ISBN: 978-1-4503-3779-3. doi:10/ghp2qv.

[38] Johnson, Janet G, Sharkey, Tommy, Butarbutar, Iramuali Cynthia, Xiong, Danica, Huang, Ruijie, Sy, Lauren, and Weibel, Nadir. 2023. UnMapped: Leveraging Experts' Situated Experiences to Ease Remote Guidance in Collaborative Mixed Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, (Apr. 19, 2023), 1–20. ISBN: 978-1-4503-9421-5. doi:10.1145/3544548.3581444.

[39] Chenechal, Morgan Le, Duval, Thierry, Gouranton, Valerie, Royan, Jerome, and Arnaldi, Bruno. 2016. Vishnu: virtual immersive support for HelpiNg users an interaction paradigm for collaborative remote guiding in mixed reality. In *2016 IEEE Third VR International Workshop on Collaborative Virtual Environments (3DCVE)*. 2016 IEEE Third VR International Workshop on Collaborative Virtual Environments (3DCVE). (Mar. 2016), 9–12. doi:10/ggfv6n.

[40] Linell, Per. 1998. *Approaching Dialogue: Talk, Interaction and Contexts in Dialogical Perspectives*. John Benjamins Publishing. 350 pp. ISBN: 978-90-272-1833-9. Google Books: MwVFp0tVoxMC.

[41] Voida, Amy, Voida, Stephen, Greenberg, Saul, and He, Helen Ai. 2008. Asymmetry in media spaces. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (CSCW '08). Association for Computing Machinery, New York, NY, USA, (Nov. 8, 2008), 313–322. ISBN: 978-1-60558-007-4. doi:10/frmg69.

[42] Bréhault, Victor, Dubois, Emmanuel, Prouzeau, Arnaud, and Serrano, Marcos. 2025. A Systematic Literature Review to Characterize Asymmetric Interaction in Collaborative Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, (Apr. 25, 2025), 1–19. ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3713129.

[43] Ens, Barrett, Lanir, Joel, Tang, Anthony, Bateman, Scott, Lee, Gun, Piumsomboon, Thammathip, and Billinghurst, Mark. 2019. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies*. 50 Years of

the International Journal of Human-Computer Studies. Reflections on the Past, Present and Future of Human-Centred Technologies 131, (Nov. 1, 2019), 81–98. doi:10/ggrfdf.

[44] Sereno, Mickael, Wang, Xiyao, Besancon, Lonni, Mcguffin, Michael J, and Isenberg, Tobias. 2020. Collaborative Work in Augmented Reality: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. doi:10/gjkq7w.

[45] Dourish, Paul. 2001. *Where the Action Is: The Foundations of Embodied Interaction*. (Sept. 14, 2001). doi:10.7551/mitpress/7221.001.0001.

[46] Benford, Steve, Brown, Chris, Reynard, Gail, and Greenhalgh, Chris. 1996. Shared spaces: transportation, artificiality, and spatiality. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (CSCW '96). Association for Computing Machinery, New York, NY, USA, (Nov. 16, 1996), 77–86. ISBN: 978-0-89791-765-0. doi:10/bxfd98.

[47] Harrison, Steve and Dourish, Paul. 1996. Re-Place-ing Space: The Roles of Place and Space in Collaborative Systems. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (CSCW '96). Association for Computing Machinery, New York, NY, USA, (Nov. 16, 1996), 67–76. ISBN: 978-0-89791-765-0. doi:10.1145/240080.240193.

[48] Dourish, Paul. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8, 1, (Feb. 1, 2004), 19–30. doi:10.1007/s00779-003-0253-8.

[49] Luff, Paul, Heath, Christian, Kuzuoka, Hideaki, Hindmarsh, Jon, Yamazaki, Keiichi, and Oyama, Shinya. 2003. Fractured Ecologies: Creating Environments for Collaboration. *Human–Computer Interaction*, 18, 1–2, (June 1, 2003), 51–84. doi:10.1207/S15327051HCI1812_3.

[50] Steed, A., MacColl, I., Randell, C., Brown, B., Chalmers, M., and Greenhalgh, C. 2004. Models of space in a mixed-reality system. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.* Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004. IEEE, London, England, 768–777. ISBN: 978-0-7695-2177-0. doi:10.1109/IV.2004.1320228.

[51] Gutwin, Carl and Greenberg, Saul. 1996. Workspace awareness for groupware. In *Conference Companion on Human Factors in Computing Systems* (CHI '96). Association for Computing Machinery, New York, NY, USA, (Apr. 18, 1996), 208–209. ISBN: 978-0-89791-832-9. doi:10.1145/257089.257284.

[52] Steed, Anthony, Slater, Mel, Sadagic, Amela, Bullock, Adrian, and Tromp, Jolanda. 1999. Leadership and collaboration in shared virtual environments. In *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*. IEEE, 112–115. doi:10.1109/VR.1999.756941.

[53] Pan, Ye, Sinclair, David, and Mitchell, Kenny. 2018. Empowerment and embodiment for collaborative mixed reality systems. *Computer Animation and Virtual Worlds*, 29, 3–4, e1838. doi:10/gm8b8n.

[54] Olin, Patrick Aggergaard, Issa, Ahmad Mohammad, Feuchtner, Tiare, and Grønbæk, Kaj. 2020. Designing for Heterogeneous Cross-Device Collaboration and Social Interaction in Virtual Reality. In *32nd Australian Conference on Human-Computer Interaction* (OzCHI '20). Association for Computing Machinery, New York, NY, USA, (Dec. 2, 2020), 112–127. ISBN: 978-1-4503-8975-4. doi:10/gmc96w.

[55] Norman, Mitchell, Lee, Gun A., Smith, Ross T., and Billingurst, Mark. 2019. The Impact of Remote User's Role in a Mixed Reality Mixed Presence System. In *The 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry* (VRCAI '19). Association for Computing Machinery, New York, NY, USA, (Nov. 14, 2019), 1–9. ISBN: 978-1-4503-7002-8. doi:10.1145/3359997.3365691.

[56] Tong, Wai, Xia, Meng, Wong, Kam Kwai, Bowman, Doug A., Pong, Ting-Chuen, Qu, Huamin, and Yang, Yalong. 2023. Towards an Understanding of Distributed Asymmetric Collaborative Visualization on Problem-solving. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, Shanghai, China, (Mar. 2023), 387–397. ISBN: 979-8-3503-4815-6. doi:10.1109/VR55154.2023.00054.

[57] Nebeling, Michael, Lewis, Katy, Chang, Yu-Cheng, Zhu, Lihan, Chung, Michelle, Wang, Piaoyang, and Nebeling, Janet. 2020. XRDirector: A Role-Based Collaborative Immersive Authoring System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (Apr. 21, 2020), 1–12. ISBN: 978-1-4503-6708-0. doi:10.1145/3313831.3376637.

[58] Steed, Anthony, Frécon, Emmanuel, Avatare, Anneli, Pemberton, Duncan, and Smith, Gareth. 1999. The London Travel Demonstrator. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (VRST '99). Association for Computing Machinery, New York, NY, USA, (Dec. 20, 1999), 50–57. ISBN: 978-1-58113-141-3. doi:10/fnmzsw.

[59] Butz, Andreas, Beshers, Clifford, and Feiner, Steven. 1998. Of Vampire mirrors and privacy lamps: privacy management in multi-user augmented environments. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology* (UIST '98). Association for Computing Machinery, New York, NY, USA, (Nov. 1, 1998), 171–172. ISBN: 978-1-58113-034-8. doi:10.1145/288392.288598.

[60] Piumsomboon, Thammathip, Lee, Youngho, Lee, Gun, and Billinghurst, Mark. 2017. CoVAR: a collaborative virtual and augmented reality system for remote collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies* (SA '17). Association for Computing Machinery, New York, NY, USA, (Nov. 27, 2017), 1–2. ISBN: 978-1-4503-5404-2. doi:10/gf2bqk.

[61] Lee, Gun A., Teo, Theophilus, Kim, Seungwon, and Billinghurst, Mark. 2017. Mixed reality collaboration through sharing a live panorama. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications* (SA '17). Association for Computing Machinery, New York, NY, USA, (Nov. 27, 2017), 1–4. ISBN: 978-1-4503-5410-3. doi:10/ghk9ws.

[62] Tian, Huayuan, Lee, Gun A., Bai, Huidong, and Billinghurst, Mark. 2023. Using Virtual Replicas to Improve Mixed Reality Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. doi:10.1109/TVCG.2023.3247113.

[63] Thoravi Kumaravel, Balasaravanan, Nguyen, Cuong, DiVerdi, Stephen, and Hartmann, Bjoern. 2020. TransceiVR: Bridging Asymmetrical Communication Between VR Users and External Collaborators. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20: The 33rd Annual ACM Symposium on User Interface Software and Technology. ACM, Virtual Event USA, (Oct. 20, 2020), 182–195. ISBN: 978-1-4503-7514-6. doi:10.1145/3379337.3415827.

[64] Borowski, Marcel, Grønbæk, Jens Emil Sloth, Butcher, Peter W. S., Ritsos, Panagiotis D., Klokmose, Clemens Nylandsted, and Elmqvist, Niklas. 2025. Spatialstrates: Cross-Reality Collaboration through Spatial Hypermedia. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (UIST '25). Association for Computing Machinery, New York, NY, USA, (Sept. 27, 2025), 1–14. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747708.

[65] Wong, Emily, Genay, Adélaïde, Grønbæk, Jens Emil Sloth, and Velloso, Eduardo. 2025. Spatial Heterogeneity in Distributed Mixed Reality Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, (Apr. 25, 2025), 1–19. ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3714033.

[66] Harper, Richard, Rodden, Tom, Rogers, Yvonne, and Sellen, Abigail. 2008. *Being Human: Human-computer Interaction in the Year 2020*. Microsoft Research. 98 pp. ISBN: 978-0-9554761-1-2. Google Books: zNKALwAACAAJ.

[67] Tromp, J., Bullock, A., Steed, A., Sadagic, A., Slater, M., and Frecon, E. 1998. Small group behaviour experiments in the Coven project. *IEEE Computer Graphics and Applications*, 18, 6, (Nov. 1998), 53–63. doi:10.1109/38.734980.

[68] Slater, Mel, Sadagic, Amela, Usoh, Martin, and Schroeder, Ralph. 2000. Small-Group Behavior in a Virtual and Real Environment: A Comparative Study. *Presence: Teleoperators and Virtual Environments*, 9, 1, (Feb. 1, 2000), 37–51. doi:10.1162/105474600566600.

[69] Ohl, Stephan. 2018. Tele-Immersion Concepts. *IEEE Transactions on Visualization and Computer Graphics*, 24, 10, (Oct. 2018), 2827–2842. doi:10.1109/TVCG.2017.2767590.

[70] Herskovitz, Jaylin, Cheng, Yi Fei, Guo, Anhong, Sample, Alanson P., and Nebeling, Michael. 2022. XSpace: an augmented reality toolkit for enabling spatially-aware distributed collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 6, (Nov. 14, 2022), 568:277–568:302, ISS, (Nov. 14, 2022). doi:10.1145/3567721.

[71] Bannon, Liam J. and Schmidt, Kjeld. 1990. CSCW: Four Characters in Search of a Context. *Studies in Computer Supported Cooperative Work: Theory, Practice and Design*, 3–16. doi:10/gn4vff.

[72] Johansen, Robert. 1988. *Groupware: Computer Support for Business Teams*. The Free Press. ISBN: 0-02-916491-5.

[73] Olson, Gary M. and Olson, Judith S. 2000. Distance Matters. *Human–Computer Interaction*, 15, 2–3, (Sept. 2000), 139–178. doi:10.1207/S15327051HCI1523_4.

[74] Bateson, Gregory. 1972. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* Jason Aronson, Lanham, MD, US, xxvi, 545. xxvi, 545. ISBN: 978-0-87668-950-9.

[75] Abowd, Gregory D., Dey, Anind K., Brown, Peter J., Davies, Nigel, Smith, Mark, and Steggles, Pete. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Handheld and Ubiquitous Computing*. Hans-W. Gellersen, (Ed.) Springer, Berlin, Heidelberg, 304–307. ISBN: 978-3-540-48157-7. doi:10.1007/3-540-48157-5_29.

[76] Biocca, Frank, Harms, Chad, and Burgoon, Judee K. 2003. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, 12, 5, (Oct. 2003), 456–480. doi:10.1162/105474603322761270.

[77] Bailenson, Jeremy N., Swinth, Kim, Hoyt, Crystal, Persky, Susan, Dimov, Alex, and Blascovich, Jim. 2005. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoper. Virtual Environ.*, 14, 4, (Aug. 1, 2005), 379–393. doi:10.1162/105474605774785235.

[78] Schroeder, Ralph. 2002. Social Interaction in Virtual Environments: Key Issues, Common Themes, and a Framework for Research. In *The Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*. Ralph Schroeder, (Ed.) Springer, London, 1–18. ISBN: 978-1-4471-0277-9. doi:10.1007/978-1-4471-0277-9_1.

[79] Zhao, Shanyang. 2003. Toward a Taxonomy of Copresence. *Presence: Teleoperators and Virtual Environments*, 12, 5, (Oct. 2003), 445–455. doi:10.1162/105474603322761261.

[80] Goffman, Erving. 1969. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, New York. 248 pp. ISBN: 978-0-02-911940-2.

[81] Short, John, Williams, Ederyn, and Christie, B. 1976. The social psychology of telecommunications. In ISBN: 978-0-471-01581-9.

[82] Harms, Chad and Biocca, Frank. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh Annual International Workshop: Presence*. Vol. 2004. Universidad Politecnica de Valencia Valencia, Spain.

[83] Dourish, Paul and Bellotti, Victoria. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work* (CSCW '92). Association for Computing Machinery, New York, NY, USA, (Dec. 1, 1992), 107–114. ISBN: 978-0-89791-542-7. doi:10.1145/143457.143468.

[84] Carroll, John M., Rosson, Mary Beth, Convertino, Gregorio, and Ganoe, Craig H. 2006. Awareness and teamwork in computer-supported collaborations. *Interacting with Computers*, 18, 1, (Jan. 2006), 21–46. doi:10.1016/j.intcom.2005.05.005.

[85] Neale, Dennis C., Carroll, John M, and Rosson, Mary Beth. 2004. Evaluating computer-supported cooperative work: models and frameworks. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (CSCW '04). Association for Computing Machinery, New York, NY, USA, (Nov. 6, 2004), 112–121. ISBN: 978-1-58113-810-8. doi:10.1145/1031607.1031626.

[86] Nofi, Albert A. 2000. Defining and Measuring Shared Situational Awareness, (Oct. 2000), 76.

[87] Randall, Dave. 2016. What Is Common in Accounts of Common Ground? *Computer Supported Cooperative Work (CSCW)*, 25, 4, (Oct. 1, 2016), 409–423. doi:10.1007/s10606-016-9256-7.

[88] Engelbart, Douglas C. and English, William K. 1968. A research center for augmenting human intellect. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I* (AFIPS '68 (Fall, Part I)). Association for Computing Machinery, New York, NY, USA, (Dec. 9, 1968), 395–410. ISBN: 978-1-4503-7899-4. doi:10/fngqnt.

[89] Mantei, Marilyn M., Baecker, Ronald M., Sellen, Abigail J., Buxton, William A. S., Milligan, Thomas, and Wellman, Barry. 1991. Experiences in the use of a media space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology - CHI '91*. The SIGCHI Conference. ACM Press, New Orleans, Louisiana, United States, 203–208. ISBN: 978-0-89791-383-6. doi:10.1145/108844.108888.

[90] Gaver, William, Moran, Thomas, MacLean, Allan, Lövstrand, Lennart, Dourish, Paul, Carter, Kathleen, and Buxton, William. 1992. Realizing a video environment: EuroPARC's RAVE system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*. The SIGCHI Conference. ACM Press, Monterey, California, United States, 27–35. ISBN: 978-0-89791-513-7. doi:10.1145/142750.142754.

[91] Clark, Herbert H. and Schaefer, Edward F. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2, 1, (Jan. 1, 1987), 19–41. doi:10.1080/01690968708406350.

[92] Chapanis, Alphonse. 1975. Interactive Human Communication. *Scientific American*, 232, 3, 36–46. Retrieved Apr. 2, 2022 from https://www.jstor.org/stable/24949750 JSTOR: 24949750.

[93] Isaacs, Ellen A. and Tang, John C. 1994. What video can and cannot do for collaboration: A case study. *Multimedia Systems*, 2, 2, (Aug. 1, 1994), 63–73. doi:10.1007/BF01274181.

[94] Daly-Jones, Owen, Monk, Andrew, and Watts, Leon. 1998. Some advantages of video conferencing over high-quality audio conferencing. *International Journal of Human-Computer Studies*, 49, 1, (July 1, 1998), 21–58. doi:10.1006/ijhc.1998.0195.

[95] Gale, Stephell. 1989. Adding Audio and Video to an Office Environment, 10.

[96] Hudson, Scott E. and Smith, Ian. 1996. Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work - CSCW '96*. The 1996 ACM Conference. ACM Press, Boston, Massachusetts, United States, 248–257. ISBN: 978-0-89791-765-0. doi:10.1145/240080.240295.

[97] Han, Jefferson and Smith, Brian. 1997. CU-SeeMe VR immersive desktop teleconferencing. In *Proceedings of the Fourth ACM International Conference on Multimedia* (MULTIMEDIA '96). Association for Computing Machinery, New York, NY, USA, (Feb. 1, 1997), 199–207. ISBN: 978-0-89791-871-8. doi:10.1145/244130.244199.

[98] Buxton, Bill. 2009. Mediaspace – Meaningspace – Meetingspace. In *Media Space 20 + Years of Mediated Life*. Steve Harrison, (Ed.) Red. by Richard Harper. Springer London, London, 217–231. ISBN: 978-1-84882-483-6. doi:10.1007/978-1-84882-483-6_13.

[99] Tang, John C. and Minneman, Scott L. 1990. VideoDraw: a video interface for collaborative drawing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90). Association for Computing Machinery, New York, NY, USA, (Mar. 1, 1990), 313–320. ISBN: 978-0-201-50932-8. doi:10/b3xmzf.

[100] Whittaker, Steve, Geelhoed, Erik, and Robinson, Elizabeth. 1993. Shared workspaces: how do they work and when are they useful? *International Journal of Man-Machine Studies*, 39, 5, (Nov. 1, 1993), 813–842. doi:10.1006/imms.1993.1085.

[101] Kuzuoka, Hideaki. 1992. Spatial workspace collaboration: a SharedView video support system for remote collaboration capability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '92). Association for Computing Machinery, New York, NY, USA, (June 1, 1992), 533–540. ISBN: 978-0-89791-513-7. doi:10/fh5r9q.

[102] Clark, Herbert H. and Krych, Meredyth A. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50, 1, (Jan. 1, 2004), 62–81. doi:10.1016/j.jml.2003.08.004.

[103] Fussell, Susan R., Setlock, Leslie D., and Kraut, Robert E. 2003. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '03). Association for Computing Machinery, New York, NY, USA, (Apr. 5, 2003), 513–520. ISBN: 978-1-58113-630-2. doi:10.1145/642611.642701.

[104] Fussell, Susan R., Kraut, Robert E., and Siegel, Jane. 2000. Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (CSCW '00). Association for Computing Machinery, New York, NY, USA, (Dec. 1, 2000), 21–30. ISBN: 978-1-58113-222-9. doi:10.1145/358916.358947.

[105] Tait, Matthew and Billinghurst, Mark. 2015. The Effect of View Independence in a Collaborative AR System. *Computer Supported Cooperative Work (CSCW)*, 24, 6, (Dec. 1, 2015), 563–589. doi:10/ghk94p.

[106] Gaver, William W., Sellen, Abigail, Heath, Christian, and Luff, Paul. 1993. One is not enough: multiple views in a media space. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (CHI '93). Association for Computing Machinery, New York, NY, USA, (May 1, 1993), 335–341. ISBN: 978-0-89791-575-5. doi:10.1145/169059.169268.

[107] Fuchs, Henry, Bishop, Gary, Arthur, Kevin, McMillan, Leonard, Bajcsy, Ruzena, Lee, Sang, Farid, Hany, and Kanade, Takeo. 1994. Virtual Space Teleconferencing Using a Sea of Cameras. Technical Report. University of North Carolina at Chapel Hill, USA, (Apr. 1994).

[108] Minsky, Marvin. 1980. Telepresence. *OMNI Magazine*. Retrieved July 23, 2024 from https://cir.nii.ac.jp/crid/1570854175773549312.

[109] Raskar, Ramesh, Welch, Greg, Cutts, Matt, Lake, Adam, Stesin, Lev, and Fuchs, Henry. 1998. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '98*. The 25th Annual Conference. ACM Press, Not Known, 179–188. ISBN: 978-0-89791-999-9. doi:10.1145/280814.280861.

[110] Kraut, Robert E, Fussell, Susan R, Brennan, Susan E, and Siegel, Jane. 2002. Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work. In *Distributed Work*, 137–163.

[111] Holtgraves, Thomas M., Fussell, Susan R., and Setlock, Leslie D. 2014. Computer-Mediated Communication. In *The Oxford Handbook of Language and Social Psychology*. Thomas M. Holtgraves, (Ed.) Oxford University Press, (Sept. 1, 2014). ISBN: 978-0-19-983863-9. doi:10.1093/oxfordhb/9780199838639.013.020.

[112] Stefik, Mark, Foster, Gregg, Bobrow, Daniel G., Kahn, Kenneth, Lanning, Stan, and Suchman, Lucy. 1987. Beyond the chalkboard: computer support for collaboration and problem solving in meetings. *Communications of the ACM*, 30, 1, (Jan. 1, 1987), 32–47. doi:10.1145/7885.7887.

[113] Stefik, M., Bobrow, D. G., Foster, G., Lanning, S., and Tatar, D. 1987. WYSIWIS revised: early experiences with multiuser interfaces. *ACM Transactions on Information Systems*, 5, 2, (Apr. 1987), 147–167. doi:10.1145/27636.28056.

[114] Bjørn, Pernille et al. 2024. Achieving Symmetry in Synchronous Interaction in Hybrid Work is Impossible. *ACM Transactions on Computer-Human Interaction*, 31, 4, (Aug. 31, 2024), 1–34. doi:10.1145/3648617.

[115] Ouverson, Kaitlyn M. and Gilbert, Stephen B. 2021. A Composite Framework of Co-located Asymmetric Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, 5, (Apr. 22, 2021), 5:1–5:20, CSCW1, (Apr. 22, 2021). doi:10/gmddcj.

[116] Harris, John and Hancock, Mark. 2019. To Asymmetry and Beyond! Improving Social Connectedness by Increasing Designed Interdependence in Cooperative Play. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19). Association for Computing Machinery, New York, NY, USA, (May 2, 2019), 1–12. ISBN: 978-1-4503-5970-2. doi:10.1145/3290605.3300239.

[117] Harris, John, Hancock, Mark, and Scott, Stacey D. 2016. Leveraging Asymmetries in Multiplayer Games: Investigating Design Elements of Interdependent Play. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '16). Association for Computing Machinery, New York, NY, USA, (Oct. 15, 2016), 350–361. ISBN: 978-1-4503-4456-2. doi:10/gmb6r4.

[118] Rogers, Katja, Karaosmanoglu, Sukran, Wolf, Dennis, Steinicke, Frank, and Nacke, Lennart E. 2021. A Best-Fit Framework and Systematic Review of Asymmetric Gameplay in Multiplayer Virtual Reality Games. *Frontiers in Virtual Reality*, 0. doi:10/gmbvfj.

[119] Gaver, William W. 1992. The affordances of media spaces for collaboration. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work* (CSCW '92). Association for Computing Machinery, New York, NY, USA, (Dec. 1, 1992), 17–24. ISBN: 978-0-89791-542-7. doi:10.1145/143457.371596.

[120] Velez, Maria, Tremaine, Marilyn Mantei, Sarcevic, Aleksandra, Dorohonceanu, Bogdan, Krebs, Allan, and Marsic, Ivan. 2004. "Who's in charge here?" communicating across unequal computer platforms. *ACM Transactions on Computer-Human Interaction*, 11, 4, (Dec. 2004), 407–444. doi:10/cbt28q.

[121] Paay, Jeni, Kjeldskov, Jesper, and O'Hara, Kenton. 2011. BISi: a blended interaction space. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '11). Association for Computing Machinery, New York, NY, USA, (May 7, 2011), 185–200. ISBN: 978-1-4503-0268-5. doi:10.1145/1979742.1979644.

[122] Grønbæk, Jens Emil, Saatçi, Banu, Griggio, Carla F., and Klokmose, Clemens Nylandsted. 2021. MirrorBlender: Supporting Hybrid Meetings with a Malleable Video-Conferencing System. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, (May 7, 2021), 1–13. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445698.

[123] Tang, John et al. 2023. Perspectives: Creating Inclusive and Equitable Hybrid Meeting Experiences. *Proceedings of the ACM on Human-Computer Interaction*, 7, (Oct. 2023), CSCW2, (Oct. 2023).

[124] Panda, Payod et al. 2024. Hybridge: Bridging Spatiality for Inclusive and Equitable Hybrid Meetings. *Proceedings of the ACM on Human-Computer Interaction*, 8, (Nov. 7, 2024), 1–39, CSCW2, (Nov. 7, 2024). doi:10.1145/3687040.

[125] Matsuda, M., Matsushita, M., Yamada, T., and Namemura, T. 2006. Behavioral analysis of asymmetric information sharing on Lumisight table. In *First IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP '06)*. First IEEE Interna-

tional Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP '06). (Jan. 2006), 113–122. doi:10.1109/TABLETOP.2006.6.

[126] Cao, Xiang, Massimi, Michael, and Balakrishnan, Ravin. 2008. *Flashlight Jigsaw*: an exploratory study of an ad-hoc multi-player game on public displays. In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work - CSCW '08*. The ACM 2008 Conference. ACM Press, San Diego, CA, USA, 77. ISBN: 978-1-60558-007-4. doi:10.1145/1460563.1460577.

[127] Convertino, Gregorio, Mentis, Helena M., Slavkovic, Aleksandra, Rosson, Mary Beth, and Carroll, John M. 2011. Supporting common ground and awareness in emergency management planning: A design research project. *ACM Transactions on Computer-Human Interaction*, 18, 4, (Dec. 30, 2011), 1–34. doi:10.1145/2063231.2063236.

[128] Sutherland, I. 1965. The Ultimate Display. In *Proceedings of the IFIPS Congress 65(2):506-508. New York: IFIP*. CUMINCAD. Retrieved Oct. 26, 2021 from http://papers.cumincad.org/cgi-bin/works/BrowseTreefield=seriesorder=AZ/Show?c58e.

[129] Milgram, Paul, Takemura, Haruo, Utsumi, Akira, and Kishino, Fumio. 1995. Augmented reality: a class of displays on the reality-virtuality continuum. In *Telemanipulator and Telepresence Technologies*. Telemanipulator and Telepresence Technologies. Vol. 2351. International Society for Optics and Photonics, (Dec. 21, 1995), 282–292. doi:10/dh8jnv.

[130] Skarbez, Richard, Smith, Missie, and Whitton, Mary C. 2021. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality*, 2, 27. doi:10/gjrd9w.

[131] Speicher, Maximilian, Hall, Brian D., and Nebeling, Michael. 2019. What is Mixed Reality? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, (May 2, 2019), 1–15. ISBN: 978-1-4503-5970-2. Retrieved Aug. 23, 2021 from https://doi.org/10.1145/3290605.3300767.

[132] Rauschnabel, Philipp A., Felix, Reto, Hinsch, Chris, Shahab, Hamza, and Alt, Florian. 2022. What is XR? Towards a Framework for Augmented and Virtual Reality. *Computers in Human Behavior*, 133, (Aug. 2022), 107289. doi:10.1016/j.chb.2022.107289.

[133] Witmer, Bob G. and Singer, Michael J. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7, 3, (June 1998), 225–240. doi:10.1162/105474698565686.

[134] Slater, Mel and Wilbur, Sylvia. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6, 6, (Dec. 1, 1997), 603–616. doi:10.1162/pres.1997.6.6.603.

[135] Slater, Mel and Steed, Anthony. 2000. A Virtual Presence Counter. *Presence*, 9, 5, (Oct. 2000), 413–434. doi:10.1162/105474600566925.

[136] Bowman, Doug A., Gabbard, Joseph L., and Hix, Deborah. 2002. A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Presence: Teleoperators and Virtual Environments*, 11, 4, (Aug. 2002), 404–424. doi:10.1162/105474602760204309.

[137] Heeter, Carrie. 1992. Being There: The Subjective Experience of Presence. *Presence: Teleoperators and Virtual Environments*, 1, 2, (May 1, 1992), 262–271. doi:10.1162/pres.1992.1.2.262.

[138] Steuer, Jonathan. 1992. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42, 4, 73–93. doi:10/bsc5p5.

[139] Sheridan, Thomas B. 1992. Musings on Telepresence and Virtual Presence. *Presence: Teleoperators and Virtual Environments*, 1, 1, (Feb. 1, 1992), 120–126. doi:10.1162/pres.1992.1.1.120.

[140] Durlach, Nat and Slater, Mel. 2000. Presence in Shared Virtual Environments and Virtual Togetherness. *Presence*, 9, 2, (Apr. 2000), 214–217. doi:10.1162/105474600566736.

[141] Slater, Mel. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1535, (Dec. 12, 2009), 3549–3557. PMID: 19884149. doi:10/df44xc.

[142] Hofer, Matthias, Hartmann, Tilo, Eden, Allison, Ratan, Rabindra, and Hahn, Lindsay. 2020. The Role of Plausibility in the Experience of Spatial Presence in Virtual Environments. *Frontiers in Virtual Reality*, 1. doi:10.3389/frvir.2020.00002.

[143] Wienrich, Carolin, Komma, Philipp, Vogt, Stephanie, and Latoschik, Marc E. 2021. Spatial Presence in Mixed Realities–Considerations About the Concept, Measures, Design, and Experiments. *Frontiers in Virtual Reality*, 2, (Oct. 25, 2021), 694315. doi:10/gm8b8h.

[144] Slater, Mel, Banakou, Domna, Beacco, Alejandro, Gallego, Jaime, Macia-Varela, Francisco, and Oliva, Ramon. 2022. A Separate Reality: An Update on Place Illusion and Plausibility in Virtual Reality. *Frontiers in Virtual Reality*, 3. Retrieved June 27, 2022 from https://www.frontiersin.org/article/10.3389/frvir.2022.914392.

[145] Tran, Tanh Quang, Langlotz, Tobias, and Regenbrecht, Holger. 2024. A Survey On Measuring Presence in Mixed Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–38. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642383.

[146] Westermeier, Franziska, Brübach, Larissa, Latoschik, Marc Erich, and Wienrich, Carolin. 2023. Exploring Plausibility and Presence in Mixed Reality Experiences. *IEEE Transactions on Visualization and Computer Graphics*, 1–10. doi:10.1109/TVCG.2023.3247046.

[147] Wirth, Werner et al. 2007. A Process Model of the Formation of Spatial Presence Experiences. *Media Psychology*, 9, 3, (May 15, 2007), 493–525. doi:10/cqkm23.

[148] Kilteni, Konstantina, Groten, Raphaela, and Slater, Mel. 2012. The Sense of Embodiment in Virtual Reality. *Presence: Teleoperators and Virtual Environments*, 21, 4, (Nov. 1, 2012), 373–387. doi:10.1162/PRES_a_00124.

[149] Slater, Mel, Usoh, Martin, and Steed, Anthony. 1994. Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 3, 2, (May 1, 1994), 130–144. doi:10/gf8bfp.

[150] Slater, Mel, McCarthy, John, and Maringelli, Francesco. 1998. The Influence of Body Movement on Subjective Presence in Virtual Environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40, 3, (Sept. 1998), 469–477. doi:10/dbgrnz.

[151] Cools, Robbe, Han, Jihae, Esteves, Augusto, and Simeone, Adalberto L. 2025. From Display to Interaction: Design Patterns for Cross-Reality Systems. *IEEE Transactions on Visualization and Computer Graphics*, 31, 5, (May 2025), 3129–3139. doi:10.1109/TVCG.2025.3549893.

[152] Orts-Escolano, Sergio et al. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (UIST '16). Association for Computing Machinery, New York, NY, USA, (Oct. 16, 2016), 741–754. ISBN: 978-1-4503-4189-9. doi:10/gf2fpj.

[153] Grønbæk, Jens Emil Sloth, Sánchez Esquivel, Juan, Leiva, Germán, Velloso, Eduardo, Gellersen, Hans, and Pfeuffer, Ken. 2024. Blended Whiteboard: Physicality and Reconfigurability in Remote Mixed Reality Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (May 11, 2024), 1–16. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642293.

[154] Hubenschmid, Sebastian et al. 2025. Hybrid User Interfaces: Past, Present, and Future of Complementary Cross-Device Interaction in Mixed Reality. (Sept. 5, 2025). arXiv: 2509.05491 [cs]. Pre-published.

[155] Thomsen, Lui Albæk, Nilsson, Niels Christian, Nordahl, Rolf, and Lohmann, Boris. 2019. Asymmetric collaboration in virtual reality: A taxonomy of asymmetric interfaces for collaborative immersive learning. *Tidsskriftet Læring og Medier (LOM)*, 12, 20, (Mar. 31, 2019), 20, (Mar. 31, 2019). doi:10.7146/lom.v12i20.109391.

[156] Ahlers, Klaus H., Kramer, André, Breen, David E., Chevalier, Pierre-Yves, Crampton, Chris, Rose, Eric, Tuceryan, Mihran, Whitaker, Ross T., and Greer, Douglas. 1995. Distributed Augmented Reality for Collaborative Design Applications. *Computer Graphics Forum*, 14, 3, 3–14. doi:10/dj5vs7.

[157] Höllerer, Tobias, Feiner, Steven, Terauchi, Tachio, Rashid, Gus, and Hallaway, Drexel. 1999. Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system. *Computers & Graphics*, 23, 6, 779–785. doi:10.1016/S0097-8493(99)00103-X.

[158] Johnson, Steven, Gibson, Madeleine, and Mutlu, Bilge. 2015. Handheld or Handsfree?: Remote Collaboration via Lightweight Head-Mounted Displays and Handheld Devices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15: Computer Supported Cooperative Work and Social Computing. ACM, Vancouver BC Canada, (Feb. 28, 2015), 1825–1836. ISBN: 978-1-4503-2922-4. doi:10.1145/2675133.2675176.

[159] Steed, Anthony et al. 2012. Beaming: An Asymmetric Telepresence System. *IEEE Computer Graphics and Applications*, 32, 6, (Nov. 2012), 10–17. doi:10.1109/MCG.2012.110.

[160] Jones, Brennan, Zhang, Yaying, Wong, Priscilla N. Y., and Rintel, Sean. 2021. Belonging There: VROOM-ing into the Uncanny Valley of XR Telepresence. *Proceedings of the ACM on Human-Computer Interaction*, 5, (Apr. 13, 2021), 1–31, CSCW1, (Apr. 13, 2021). doi:10.1145/3449133.

[161] Sakashita, Mose, Kim, Hyunju, Woodard, Brandon, Zhang, Ruidong, and Guimbretière, François. 2023. VRoxy: Wide-Area Collaboration From an Office Using a VR-Driven Robotic Proxy. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23). Association for Computing Machinery, New York, NY, USA, (Oct. 29, 2023), 1–13. ISBN: 979-8-4007-0132-0. doi:10.1145/3586183.3606743.

[162] Gugenheimer, Jan, Stemasov, Evgeny, Frommel, Julian, and Rukzio, Enrico. 2017. ShareVR: Enabling Co-Located Experiences for Virtual Reality between HMD and Non-HMD Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17: CHI Conference on Human Factors in Computing Systems. ACM, Denver Colorado USA, (May 2, 2017), 4021–4033. ISBN: 978-1-4503-4655-9. doi:10/ggchbc.

[163] Jung, Sungchul, Wu, Yuanjie, Lukosch, Stephan, Lukosch, Heide, Mckee, Ryan Douglas, and Lindeman, Robert W. 2023. Cross-Reality Gaming: Comparing Competition and Collaboration in an Asymmetric Gaming Experience. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (VRST '23). Association for

Computing Machinery, New York, NY, USA, (Oct. 9, 2023), 1–10. ISBN: 979-8-4007-0328-7. doi:10.1145/3611659.3615698.

[164] Nigay, Laurence and Coutaz, Joëlle. 1995. A generic platform for addressing the multimodal challenge. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '95). ACM Press/Addison-Wesley Publishing Co., USA, (May 1, 1995), 98–105. ISBN: 978-0-201-84705-5. doi:10.1145/223904.223917.

[165] Stafford, Aaron, Piekarski, Wayne, and Thomas, Bruce H. 2006. Implementation of god-like interaction techniques for supporting collaboration between outdoor AR and indoor tabletop users. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE, 165–172. doi:10.1109/ISMAR.2006.297809.

[166] Kim, Seungwon, Lee, Gun, Sakata, Nobuchika, and Billinghurst, Mark. 2014. Improving co-presence with augmented visual communication cues for sharing experience through video conference. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Sept. 2014), 83–92. doi:10/gm8b9s.

[167] Norman, Mitchell, Lee, Gun, Smith, Ross T., and Billinqhurs, Mark. 2019. A Mixed Presence Collaborative Mixed Reality System. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). (Mar. 2019), 1106–1107. doi:10.1109/VR.2019.8797966.

[168] Takeuchi, Seiichiro, Hashiguchi, Kyoko, Homma, Yuki, Kajitani, Kent, and Meguro, Shingo. 2021. GIBSON: AR/VR synchronized city walking system. In *SIGGRAPH Asia 2021 XR*. SA '21: SIGGRAPH Asia 2021. ACM, Tokyo Japan, (Dec. 14, 2021), 1–2. ISBN: 978-1-4503-9075-0. doi:10/gnthcj.

[169] Komiyama, Ryohei, Miyaki, Takashi, and Rekimoto, Jun. 2017. JackIn space: designing a seamless transition between first and third person view for effective telepresence collaborations. In *Proceedings of the 8th Augmented Human International Conference* (AH '17). Association for Computing Machinery, New York, NY, USA, (Mar. 16, 2017), 1–9. ISBN: 978-1-4503-4835-5. doi:10.1145/3041164.3041183.

[170] Zaman, Faisal, Anslow, Craig, Chalmers, Andrew, and Rhee, Taehyun. 2023. MRMAC: Mixed Reality Multi-user Asymmetric Collaboration. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Oct. 2023), 591–600. doi:10.1109/ISMAR59233.2023.00074.

[171] Lee, Gun A., Teo, Theophilus, Kim, Seungwon, and Billinghurst, Mark. 2018. A User Study on MR Remote Collaboration Using Live 360 Video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Oct. 2018), 153–164. doi:10/gm8b8w.

[172] Huang, Xincheng, Yin, Michael, Xia, Ziyi, and Xiao, Robert. 2024. VirtualNexus: Enhancing 360-Degree Video AR/VR Collaboration with Environment Cutouts and Virtual Replicas. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (UIST '24). Association for Computing Machinery, New York, NY, USA, (Oct. 11, 2024), 1–12. ISBN: 979-8-4007-0628-8. doi:10.1145/3654777.3676377.

[173] Grandi, Jerônimo Gustavo, Debarba, Henrique Galvan, and Maciel, Anderson. 2019. Characterizing Asymmetric Collaborative Interactions in Virtual and Augmented Realities. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). (Mar. 2019), 127–135. doi:10/gj27xd.

[174] Zhu, Haibin and Zhou, MengChu. 2006. Role-based collaboration and its kernel mechanisms. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36, 4, (July 2006), 578–589. doi:10.1109/TSMCC.2006.875726.

[175] Ballagas, Rafael, Dugan, Thérèse E., Revelle, Glenda, Mori, Koichi, Sandberg, Maria, Go, Janet, Reardon, Emily, and Spasojevic, Mirjana. 2013. Electric agents: fostering sibling joint media engagement through interactive television and augmented reality. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13*. The 2013 Conference. ACM Press, San Antonio, Texas, USA, 225. ISBN: 978-1-4503-1331-5. doi:10.1145/2441776.2441803.

[176] Le Chénéchal, Morgan, Lacoche, Jérémy, Royan, Jérôme, Duval, Thierry, Gouranton, Valérie, and Arnaldi, Bruno. 2016. When the giant meets the ant an asymmetric approach for collaborative and concurrent object manipulation in a multi-scale environment. In *2016 IEEE Third VR International Workshop on Collaborative Virtual Environments (3DCVE)*. IEEE, 18–22. doi:10.1109/3DCVE.2016.7563562.

[177] Schubert, Thomas W. 2005. Your Highness: Vertical Positions as Perceptual Symbols of Power. *Journal of Personality and Social Psychology*, 89, 1, 1–21. doi:10.1037/0022-3514.89.1.1.

[178] Jansen, Pascal, Fischbach, Fabian, Gugenheimer, Jan, Stemasov, Evgeny, Frommel, Julian, and Rukzio, Enrico. 2020. ShARe: Enabling Co-Located Asymmetric Multi-User Interaction for Augmented Reality Head-Mounted Displays. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20: The 33rd Annual

ACM Symposium on User Interface Software and Technology. ACM, Virtual Event USA, (Oct. 20, 2020), 459–471. ISBN: 978-1-4503-7514-6. doi:10/gkd2tx.

[179] Hillier, Brian. 1997. *The Economics of Asymmetric Information*. Macmillan International Higher Education, (Apr. 7, 1997). 199 pp. ISBN: 978-1-349-25485-9. Google Books: pExdDwAAQBAJ.

[180] Allen, David G. and Griffeth, Rodger W. 1997. Vertical and Lateral Information Processing: The Effects of Gender, Employee Classification Level, and Media Richness on Communication and Work Outcomes. *Human Relations*, 50, 10, (Oct. 1, 1997), 1239–1260. doi:10.1177/001872679705001003.

[181] Jung, Sungchul, Karki, Nawam, Slutter, Max, and Lindeman, Robert W. 2021. On the Use of Multi-sensory Cues in Symmetric and Asymmetric Shared Collaborative Virtual Spaces. *Proceedings of the ACM on Human-Computer Interaction*, 5, (Apr. 22, 2021), 72:1–72:25, CSCW1, (Apr. 22, 2021). doi:10.1145/3449146.

[182] Jää-Aro, Kai-Mikael and Snowdon, Dave. 2001. How Not To Be Objective. In *Collaborative Virtual Environments: Digital Places and Spaces for Interaction*. Computer Supported Cooperative Work. Elizabeth F. Churchill, David N. Snowdon, and Alan J. Munro, (Eds.) Springer, London, 143–159. ISBN: 978-1-4471-0685-2. doi:10.1007/978-1-4471-0685-2_8.

[183] Snowdon, David, Greenhalgh, C, and Benford, Stephanie. 1995. What you see is not what i see: subjectivity in virtual environments.

[184] Smith, Gareth and Mariani, John. 1997. Using subjective views to enhance 3D applications. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST '97*. The ACM Symposium. ACM Press, Lausanne, Switzerland, 139–146. ISBN: 978-0-89791-953-1. doi:10.1145/261135.261161.

[185] Park, Kyoung S., Kapoor, Abhinav, and Leigh, Jason. 2000. Lessons learned from employing multiple perspectives in a collaborative virtual environment for visualizing scientific data. In *Proceedings of the Third International Conference on Collaborative Virtual Environments* (CVE '00). Association for Computing Machinery, New York, NY, USA, (Sept. 1, 2000), 73–82. ISBN: 978-1-58113-303-5. doi:10/cx5wp9.

[186] Pettersson, Lars Winkler, Kjellin, Andreas, Lind, Mats, and Seipel, Stefan. 2006. Effects of Layer Partitioning in Collaborative 3D Visualizations. In *Advances in Visual Computing*. George Bebis et al., (Eds.) Springer, Berlin, Heidelberg, 180–190. ISBN: 978-3-540-48631-2. doi:10.1007/11919476_19.

[187] Salimian, Hossein, Brooks, Stephen, and Reilly, Derek. 2019. MP Remix: Relaxed WYSIWIS Immersive Interfaces for Mixed Presence Collaboration With 3D Content. *Proc. ACM Hum.-Comput. Interact.*, 3, (Nov. 7, 2019), 105:1–105:22, CSCW, (Nov. 7, 2019). doi:10.1145/3359207.

[188] Wong, Emily, Grønbæk, Jens Emil Sloth, and Velloso, Eduardo. 2024. The Jamais Vu Effect: Understanding the Fragile Illusion of Co-Presence in Mixed Reality. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (DIS '24). Association for Computing Machinery, New York, NY, USA, (July 1, 2024), 2227–2246. ISBN: 979-8-4007-0583-0. doi:10.1145/3643834.3661574.

[189] Müller, Jens, Rädle, Roman, and Reiterer, Harald. 2017. Remote Collaboration With Mixed Reality Displays: How Shared Virtual Landmarks Facilitate Spatial Referencing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17: CHI Conference on Human Factors in Computing Systems. ACM, Denver Colorado USA, (May 2, 2017), 6481–6486. ISBN: 978-1-4503-4655-9. doi:10.1145/3025453.3025717.

[190] Johnson, Janet G, Gasques, Danilo, Sharkey, Tommy, Schmitz, Evan, and Weibel, Nadir. 2021. Do You Really Need to Know Where "That" Is? Enhancing Support for Referencing in Collaborative Mixed Reality Environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21: CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, (May 6, 2021), 1–14. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445246.

[191] Chastine, Jeffrey W., Nagel, Kristine, Zhu, Ying, and Yearsovich, Luca. 2007. Understanding the design space of referencing in collaborative augmented reality environments. In *Proceedings of Graphics Interface 2007* (GI '07). Association for Computing Machinery, New York, NY, USA, (May 28, 2007), 207–214. ISBN: 978-1-56881-337-0. doi:10/d49qrx.

[192] Erickson, Austin, Norouzi, Nahal, Kim, Kangsoo, Schubert, Ryan, Jules, Jonathan, LaViola, Joseph J., Bruder, Gerd, and Welch, Gregory F. 2020. Sharing gaze rays for visual target identification tasks in collaborative augmented reality. *Journal on Multimodal User Interfaces*, 14, 4, (Dec. 1, 2020), 353–371. doi:10.1007/s12193-020-00330-2.

[193] Piumsomboon, Thammathip, Dey, Arindam, Ens, Barrett, Lee, Gun, and Billinghurst, Mark. 2019. The Effects of Sharing Awareness Cues in Collaborative Mixed Reality. *Frontiers in Robotics and AI*, 6. Retrieved Apr. 1, 2022 from https://www.frontiersin.org/article/10.3389/frobt.2019.00005.

[194] Bovo, Riccardo, Giunchi, Daniele, Alebri, Muna, Steed, Anthony, Costanza, Enrico, and Heinis, Thomas. 2022. Cone of Vision as a Behavioural Cue for VR Collaboration. *Proceed-*

ings of the ACM on Human-Computer Interaction, 6, (Nov. 7, 2022), 1–27, CSCW2, (Nov. 7, 2022). doi:10.1145/3555615.

[195] Jing, Allison, May, Kieran William, Naeem, Mahnoor, Lee, Gun, and Billinghurst, Mark. 2021. eyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI EA '21). Association for Computing Machinery, New York, NY, USA, (May 8, 2021), 1–7. ISBN: 978-1-4503-8095-9. doi:10.1145/3411763.3451844.

[196] Jing, Allison, Gupta, Kunal, McDade, Jeremy, Lee, Gun A., and Billinghurst, Mark. 2022. Comparing Gaze-Supported Modalities with Empathic Mixed Reality Interfaces in Remote Collaboration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, Singapore, Singapore, (Oct. 2022), 837–846. ISBN: 978-1-6654-5325-7. doi:10.1109/ISMAR55827.2022.00102.

[197] Nguyen, Cuong, DiVerdi, Stephen, Hertzmann, Aaron, and Liu, Feng. 2017. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (UIST '17). Association for Computing Machinery, New York, NY, USA, (Oct. 20, 2017), 267–277. ISBN: 978-1-4503-4981-9. doi:10.1145/3126594.3126659.

[198] Bauer, Martin, Kortuem, Gerd, and Segall, Zary. 1999. " Where are you pointing at?" A study of remote collaboration in a wearable videoconference system. In *Digest of Papers. Third International Symposium on Wearable Computers*. IEEE, 151–158. doi:10.1109/ISWC.1999.806696.

[199] Stafford, Aaron, Thomas, Bruce H., and Piekarski, Wayne. 2008. Efficiency of techniques for mixed-space collaborative navigation. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*. 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. (Sept. 2008), 181–182. doi:10/dv6bvc.

[200] Oda, Ohan and Feiner, Steven. 2012. 3D referencing techniques for physical objects in shared augmented reality. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Nov. 2012), 207–215. doi:10.1109/ISMAR.2012.6402558.

[201] Teo, Theophilus, Lee, Gun A., Billinghurst, Mark, and Adcock, Matt. 2018. Hand gestures and visual annotation in live 360 panorama-based mixed reality remote collaboration. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (OzCHI '18). Association for Computing Machinery, New York, NY, USA, (Dec. 4, 2018), 406–410. ISBN: 978-1-4503-6188-0. doi:10.1145/3292147.3292200.

[202] Kasahara, Shunichi and Rekimoto, Jun. 2014. JackIn: integrating first-person view with out-of-body vision generation for human-human augmentation. In *Proceedings of the 5th Augmented Human International Conference* (AH '14). Association for Computing Machinery, New York, NY, USA, (Mar. 7, 2014), 1–8. ISBN: 978-1-4503-2761-9. doi:10/ghms42.

[203] Kasahara, Shunichi and Rekimoto, Jun. 2015. JackIn head: immersive visual telepresence system with omnidirectional wearable camera for remote collaboration. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*. VRST '15: 21th ACM Symposium on Virtual Reality Software and Technology. ACM, Beijing China, (Nov. 13, 2015), 217–225. ISBN: 978-1-4503-3990-2. doi:10.1145/2821592.2821608.

[204] Nagai, Shohei, Kasahara, Shunichi, and Rekimoto, Jun. 2015. LiveSphere: Sharing the Surrounding Visual Environment for Immersive Experience in Remote Collaboration. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction* (TEI '15). Association for Computing Machinery, New York, NY, USA, (Jan. 15, 2015), 113–116. ISBN: 978-1-4503-3305-4. doi:10/gm8b9j.

[205] Billinghurst, Mark, Nassani, Alaeddin, and Reichherzer, Carolin. 2014. Social panoramas: using wearable computers to share experiences. In *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications on - SA '14*. SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications. ACM Press, Shenzhen, China, 1–1. ISBN: 978-1-4503-1891-4. doi:10.1145/2669062.2669084.

[206] Speicher, Maximilian, Cao, Jingchen, Yu, Ao, Zhang, Haihua, and Nebeling, Michael. 2018. 360Anywhere: Mobile Ad-hoc Collaboration in Any Environment using 360 Video and Augmented Reality. *Proceedings of the ACM on Human-Computer Interaction*, 2, (June 19, 2018), 9:1–9:20, EICS, (June 19, 2018). doi:10.1145/3229091.

[207] Gao, Lei, Bai, Huidong, He, Weiping, Billinghurst, Mark, and Lindeman, Robert W. 2018. Real-time visual representations for mobile mixed reality remote collaboration. In *SIGGRAPH Asia 2018 Virtual & Augmented Reality on - SA '18*. SIGGRAPH Asia 2018 Virtual & Augmented Reality. ACM Press, Tokyo, Japan, 1–2. ISBN: 978-1-4503-6028-9. doi:10/gm8b9h.

[208] Teo, Theophilus, A. Lee, Gun, Billinghurst, Mark, and Adcock, Matt. 2019. 360Drops: Mixed Reality Remote Collaboration using 360 Panoramas within the 3D Scene*. In *SIGGRAPH Asia 2019 Emerging Technologies*. SA '19: SIGGRAPH Asia 2019. ACM, Brisbane QLD Australia, (Nov. 17, 2019), 1–2. ISBN: 978-1-4503-6942-8. doi:10.1145/3355049.3360517.

[209] Rhee, Taehyun, Thompson, Stephen, Medeiros, Daniel, dos Anjos, Rafael, and Chalmers, Andrew. 2020. Augmented Virtual Teleportation for High-Fidelity Telecollaboration. *IEEE*

Transactions on Visualization and Computer Graphics, 26, 5, (May 2020), 1923–1933. doi:10.1109/TVCG.2020.2973065.

[210] Viola, Irene, Jansen, Jack, Subramanyam, Shishir, Reimat, Ignacio, and Cesar, Pablo. 2023. VR2Gather: A Collaborative, Social Virtual Reality System for Adaptive, Multiparty Real-Time Communication. *IEEE MultiMedia*, 30, 2, (Apr. 2023), 48–59. doi:10.1109/MMUL.2023.3263943.

[211] Pejsa, Tomislav, Kantor, Julian, Benko, Hrvoje, Ofek, Eyal, and Wilson, Andrew. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16: Computer Supported Cooperative Work and Social Computing. ACM, San Francisco California USA, (Feb. 27, 2016), 1716–1725. ISBN: 978-1-4503-3592-8. doi:10.1145/2818048.2819965.

[212] Wang, Shengze, Wang, Ziheng, Schmelzle, Ryan, Zheng, Liujie, Kwon, YoungJoong, Sengupta, Soumyadip, and Fuchs, Henry. 2023. Bringing Telepresence to Every Desk. (Apr. 3, 2023). arXiv: 2304.01197 [cs]. Retrieved Sept. 18, 2023 from http://arxiv.org/abs/2304.01197. Pre-published.

[213] Irlitti, Andrew, Latifoglu, Mesut, Hoang, Thuong, Syiem, Brandon Victor, and Vetere, Frank. 2024. Volumetric Hybrid Workspaces: Interactions with Objects in Remote and Co-located Telepresence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–16. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642814.

[214] Lindlbauer, David and Wilson, Andy D. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18: CHI Conference on Human Factors in Computing Systems. ACM, Montreal QC Canada, (Apr. 19, 2018), 1–13. ISBN: 978-1-4503-5620-6. doi:10.1145/3173574.3173703.

[215] Lehment, Nicolas H., Merget, Daniel, and Rigoll, Gerhard. 2014. Creating automatically aligned consensus realities for AR videoconferencing. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, Munich, Germany, (Sept. 2014), 201–206. ISBN: 978-1-4799-6184-9. doi:10.1109/ISMAR.2014.6948428.

[216] Guo, Hung-Jui, Kumar, Hiranya Garbha, Kamal, Minhas, and Prabhakaran, Balakrishnan. 2024. Room2XR: Virtual Interactive Collaboration in Real-world Scenes. In *Proceedings of the 32nd ACM International Conference on Multimedia*. MM '24: The 32nd ACM In-

ternational Conference on Multimedia. ACM, Melbourne VIC Australia, (Oct. 28, 2024), 11206–11209. ISBN: 979-8-4007-0686-8. doi:10.1145/3664647.3685521.

[217] Grønbæk, Jens Emil Sloth, Pfeuffer, Ken, Velloso, Eduardo, Astrup, Morten, Pedersen, Melanie Isabel Sønderkær, Kjær, Martin, Leiva, Germán, and Gellersen, Hans. 2023. Partially Blended Realities: Aligning Dissimilar Spaces for Distributed Mixed Reality Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, (Apr. 19, 2023), 1–16. ISBN: 978-1-4503-9421-5. doi:10.1145/3544548.3581515.

[218] Young, Jacob, Langlotz, Tobias, Mills, Steven, and Regenbrecht, Holger. 2020. Mobileportation: Nomadic Telepresence for Mobile Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4, 2, (June 15, 2020), 1–16. doi:10.1145/3397331.

[219] Teo, Theophilus, Lawrence, Louise, Lee, Gun A., Billinghurst, Mark, and Adcock, Matt. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19: CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, (May 2, 2019), 1–14. ISBN: 978-1-4503-5970-2. doi:10.1145/3290605.3300431.

[220] Thoravi Kumaravel, Balasaravanan, Anderson, Fraser, Fitzmaurice, George, Hartmann, Bjoern, and Grossman, Tovi. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST '19: The 32nd Annual ACM Symposium on User Interface Software and Technology. ACM, New Orleans LA USA, (Oct. 17, 2019), 161–174. ISBN: 978-1-4503-6816-2. doi:10.1145/3332165.3347872.

[221] Sakashita, Mose, Thoravi Kumaravel, Balasaravanan, Marquardt, Nicolai, and Wilson, Andrew David. 2024. SharedNeRF: Leveraging Photorealistic and View-dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (May 11, 2024), 1–14. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642945.

[222] Huang, Xincheng, Frehlich, Dieter, Xia, Ziyi, Gholami, Peyman, and Xiao, Robert. 2025. GaussianNexus: Room-Scale Real-Time AR/VR Telepresence with Gaussian Splatting. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. UIST '25: The 38th Annual ACM Symposium on User Interface Software and Technology. ACM, Busan Republic of Korea, (Sept. 28, 2025), 1–18. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747693.

[223] Kang, Jiho, Yang, Dongseok, Kim, Taehei, Lee, Yewon, and Lee, Sung-Hee. 2023. Real-time Retargeting of Deictic Motion to Virtual Avatars for Augmented Reality Telepresence. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, Sydney, Australia, (Oct. 16, 2023), 885–893. ISBN: 979-8-3503-2838-7. doi:10.1109/ISMAR59233.2023.00104.

[224] Congdon, Ben J., Wang, Tuanfeng, and Steed, Anthony. 2018. Merging environments for shared spaces in mixed reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (VRST '18). Association for Computing Machinery, New York, NY, USA, (Nov. 28, 2018), 1–8. ISBN: 978-1-4503-6086-9. doi:10/gj6qrn.

[225] Keshavarzi, Mohammad, Zollhoefer, Michael, Yang, Allen Y., Peluse, Patrick, and Caldas, Luisa. 2022. Synthesizing Novel Spaces for Remote Telepresence Experiences. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE, Singapore, Singapore, (Oct. 2022), 524–529. ISBN: 978-1-6654-5365-3. doi:10.1109/ISMAR-Adjunct57072.2022.00111.

[226] Kim, Dooyoung, Kim, Seonji, Choi, Selin, and Woo, Woontack. 2024. Spatial Affordance-aware Interactable Subspace Allocation for Mixed Reality Telepresence. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Oct. 2024), 1256–1265. doi:10.1109/ISMAR62088.2024.00142.

[227] Friston, Sebastian J, Congdon, Ben J, Swapp, David, Izzouzi, Lisa, Brandstätter, Klara, Archer, Daniel, Olkkonen, Otto, Thiel, Felix Johannes, and Steed, Anthony. 2021. Ubiq: A System to Build Flexible Social Virtual Reality Experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (VRST '21). Association for Computing Machinery, New York, NY, USA, (Dec. 8, 2021), 1–11. ISBN: 978-1-4503-9092-7. doi:10.1145/3489849.3489871.

[228] [SW] Microsoft, Mixed Reality Toolkit (MRTK) Apr. 11, 2022. Microsoft. URL: https://github.com/microsoft/MixedRealityToolkitRetrieved Apr. 16, 2022 from.

[229] [SW] Labs, Laan, 3D Scanner App version 1.1.2, 2021. Laan Labs. URL: https://apps.apple.com/us/app/3d-scanner-app/id1419913995.

[230] [SW] Community, Blender Online, Blender - a 3D Modelling and Rendering Package 2021. Blender Foundation. URL: http://www.blender.org.

[231] McGill, Mark, Gugenheimer, Jan, and Freeman, Euan. 2020. A Quest for Co-Located Mixed Reality: Aligning and Assessing SLAM Tracking for Same-Space Multi-User Experiences. In *26th ACM Symposium on Virtual Reality Software and Technology*. VRST '20: 26th ACM Symposium on Virtual Reality Software and Technology. ACM, Virtual Event Canada, (Nov. 2020), 1–10. ISBN: 978-1-4503-7619-8. doi:10/ghwfq2.

[232] Leary, Mark R. and Kowalski, Robin M. 1993. The Interaction Anxiousness Scale: Construct and Criterion-Related Validity. *Journal of Personality Assessment*, 61, 1, (Aug. 1, 1993), 136–146. PMID: 8377098. doi:10/b6kgfh.

[233] Rammstedt, Beatrice and John, Oliver P. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 1, (Feb. 1, 2007), 203–212. doi:10/djzd32.

[234] Shahzad, Khurram, Raja, Usman, and Hashmi, Syed Danial. 2020. Impact of Big Five personality traits on authentic leadership. *Leadership & Organization Development Journal*, 42, 2, (Jan. 1, 2020), 208–218. doi:10/hfmf.

[235] Judge, Timothy A., Bono, Joyce E., Ilies, Remus, and Gerhardt, Megan W. 2002. Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, 87, 4, 765–780. doi:10/bhfk7d.

[236] Hair, Joseph F., Black, William C., Babin, Barry J., and Anderson, Rolph E. 2009. *Multivariate Data Analysis*. (7th edition ed.). Pearson, Upper Saddle River, NJ, (Feb. 23, 2009). 816 pp. ISBN: 978-0-13-813263-7.

[237] [SW] Zeileis, Achim, Cribari-Neto, Francisco, Gruen, Bettina, Kosmidis, Ioannis, Simas, Alexandre, and Rocha, Andrea, Betareg: Beta Regression version 3.1-4, Feb. 9, 2021. URL: https://CRAN.R-project.org/package=betaregRetrieved July 6, 2022 from.

[238] Axelsson, Ann-Sofie, Abelin, Åsa, Heldal, Ilona, Schroeder, Ralph, and Wideström, Josef. 2001. Cubes in the Cube: A Comparison of a Puzzle-Solving Task in a Virtual and a Real Environment. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 4, (May 1, 2001), 279–86. doi:10.1089/109493101300117956.

[239] Napier, Barbara J. and Ferris, Gerald R. 1993. Distance in organizations. *Human Resource Management Review*, 3, 4, (Dec. 1, 1993), 321–357. doi:10.1016/1053-4822(93)90004-N.

[240] Bulu, Saniye Tugba. 2012. Place presence, social presence, co-presence, and satisfaction in virtual worlds. *Computers & Education*, 58, 1, (Jan. 1, 2012), 154–161. doi:10/dkkxwr.

[241] Hindmarsh, Jon, Fraser, Mike, Heath, Christian, Benford, Steve, and Greenhalgh, Chris. 2000. Object-focused interaction in collaborative virtual environments. *ACM Transactions on Computer-Human Interaction*, 7, 4, (Dec. 2000), 477–509. doi:10/bbgzqx.

[242] Piumsomboon, Thammathip, Lee, Gun A., Hart, Jonathon D., Ens, Barrett, Lindeman, Robert W., Thomas, Bruce H., and Billinghurst, Mark. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, (Apr. 19, 2018), 1–13. ISBN: 978-1-4503-5620-6. Retrieved Sept. 16, 2021 from https://doi.org/10.1145/3173574.3173620.

[243] Li, Yuan, Lu, Feiyu, Lages, Wallace S, and Bowman, Doug. 2019. Gaze Direction Visualization Techniques for Collaborative Wide-Area Model-Free Augmented Reality. In *Symposium on Spatial User Interaction* (SUI '19). Association for Computing Machinery, New York, NY, USA, (Oct. 19, 2019), 1–11. ISBN: 978-1-4503-6975-6. doi:10.1145/3357251.3357583.

[244] Yang, Huahai and Olson, Gary M. 2002. Exploring collaborative navigation: the effect of perspectives on group performance. In *Proceedings of the 4th International Conference on Collaborative Virtual Environments* (CVE '02). Association for Computing Machinery, New York, NY, USA, (Sept. 30, 2002), 135–142. ISBN: 978-1-58113-489-6. doi:10.1145/571878.571899.

[245] Lee, Gun A., Kim, Seungwon, Lee, Youngho, Dey, Arindam, Piumsomboon, Thammathip, Norman, Mitchell, and Billinghurst, Mark. 2017. Improving Collaboration in Augmented Video Conference using Mutually Shared Gaze. *ICAT-EGVE 2017 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 8 pages. ISBN: 9783038680383. doi:10/gm8b9d.

[246] Kim, Dooyoung, Shin, Jae-eun, Lee, Jeongmi, and Woo, Woontack. 2021. Adjusting Relative Translation Gains According to Space Size in Redirected Walking for Mixed Reality Mutual Space Generation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 2021 IEEE Virtual Reality and 3D User Interfaces (VR). IEEE, Lisboa, Portugal, (Mar. 2021), 653–660. ISBN: 978-1-6654-1838-6. doi:10.1109/VR50410.2021.00091.

[247] Usoh, Martin, Arthur, Kevin, Whitton, Mary C., Bastos, Rui, Steed, Anthony, Slater, Mel, and Brooks, Frederick P. 1999. Walking > walking-in-place > flying, in virtual environments. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '99). ACM Press/Addison-Wesley Publishing Co., USA, (July 1, 1999), 359–364. ISBN: 978-0-201-48560-8. doi:10/fgsg9x.

[248] Toet, Alexander, Mioch, Tina, Gunkel, Simon N. B., Niamut, Omar, and Erp, Jan B. F. van. 2021. Assessment of Presence in Augmented and Mixed Reality: Presence in Augmented and Mixed Reality. In *CHI 2021: Workshop Evaluating User Experiences in Mixed Reality*. Conference on Human Factors in Computing Systems, CHI 2021. ACM SigCHI, (Sept. 2021). doi:10.31234/osf.io/gvkwx.

[249] Regenbrecht, Holger and Schubert, Thomas. 2021. Measuring Presence in Augmented Reality Environments: Design and a First Test of a Questionnaire. (Mar. 3, 2021). arXiv: 2103.02831 [cs]. Pre-published.

[250] Lugrin, Jean-Luc, Landeck, Maximilian, and Latoschik, Marc Erich. 2015. Avatar embodiment realism and virtual fitness training. In *2015 IEEE Virtual Reality (VR)*. 2015 IEEE Virtual Reality (VR). (Mar. 2015), 225–226. doi:10.1109/VR.2015.7223377.

[251] Legoland. 2023. Augmented reality experience. Legoland UK, (2023). https://www.legoland.co.uk/explore/theme-park/whats-new/augmented-reality-experience/.

[252] BroadcastTech. 2023. Disney & Snapchat create Kingdom Of The Planet Of The Apes AR lenses. Broadcast, (2023). https://www.broadcastnow.co.uk/tech/disney-and-snapchat-create-kingdom-of-the-planet-of-the-apes-ar-lenses/5193320.article.

[253] Google. 2024. Geospatial Creator: Make the world your canvas. (2024). https://developers.google.com/ar/geospatialcreator.

[254] Niantic. 2024. Niantic SDK for Unity. (2024). https://www.nianticspatial.com/augment/sdk-for-unity.

[255] Nebeling, Michael. 2022. XR tools and where they are taking us: characterizing the evolving research on augmented, virtual, and mixed reality prototyping and development tools. *XRDS: Crossroads, The ACM Magazine for Students*, 29, 1, (Oct. 6, 2022), 32–38. doi:10.1145/3558192.

[256] Langlotz, Tobias, Mooslechner, Stefan, Zollmann, Stefanie, Degendorfer, Claus, Reitmayr, Gerhard, and Schmalstieg, Dieter. 2012. Sketching up the world: in situ authoring for mobile Augmented Reality. *Personal and Ubiquitous Computing*, 16, 6, (Aug. 1, 2012), 623–630. doi:10.1007/s00779-011-0430-0.

[257] Adobe. 2024. Adobe Aero: Augmented Reality. (2024). https://www.adobe.com/uk/products/aero.html.

[258] Unity. 2024. Unity MARS. (2024). https://unity.com/products/unity-mars.

[259] Lee, Gun A. and Kim, Gerard J. 2009. Immersive authoring of Tangible Augmented Reality content: A user study. *Journal of Visual Languages & Computing*, 20, 2, 61–79. doi:10.1016/j.jvlc.2008.07.001.

[260] Vargas González, Andrés, Koh, Senglee, Kapalo, Katelynn, Sottilare, Robert, Garrity, Patrick, Billinghurst, Mark, and LaViola, Joseph. 2019. A Comparison of Desktop and Augmented Reality Scenario Based Training Authoring Tools. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Oct. 2019), 339–350. doi:10.1109/ISMAR.2019.00032.

[261] Lo, Wei Hong, Zollmann, Stefanie, and Regenbrecht, Holger. 2021. From off-site to on-site: A Flexible Framework for XR Prototyping in Sports Spectating. In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. 2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, Tauranga, New Zealand, (Dec. 9, 2021), 1–6. ISBN: 978-1-6654-0645-1. doi:10.1109/IVCNZ54163.2021.9653277.

[262] Hannay, Jo E., Dybå, Tore, Arisholm, Erik, and Sjøberg, Dag I. K. 2009. The effectiveness of pair programming: A meta-analysis. *Information and Software Technology*. Special Section: Software Engineering for Secure Systems 51, 7, (July 1, 2009), 1110–1122. doi:10.1016/j.infsof.2009.02.001.

[263] Nebeling, Michael and Speicher, Maximilian. 2018. The Trouble with Augmented Reality/Virtual Reality Authoring Tools. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). (Oct. 2018), 333–337. doi:10.1109/ISMAR-Adjunct.2018.00098.

[264] Nebeling, Michael, Nebeling, Janet, Yu, Ao, and Rumble, Rob. 2018. ProtoAR: Rapid Physical-Digital Prototyping of Mobile Augmented Reality Applications. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18: CHI Conference on Human Factors in Computing Systems. ACM, Montreal QC Canada, (Apr. 21, 2018), 1–12. ISBN: 978-1-4503-5620-6. doi:10.1145/3173574.3173927.

[265] Rajaram, Shwetha and Nebeling, Michael. 2022. Paper Trail: An Immersive Authoring System for Augmented Reality Instructional Experiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22). Association for Computing Machinery, New York, NY, USA, (Apr. 29, 2022), 1–16. ISBN: 978-1-4503-9157-3. doi:10.1145/3491102.3517486.

[266] [SW], Unity 2024. Unity. URL: https://unity.com/.

[267] Engine, Unreal. 2024. Unreal Engine: We make the engine. You make it unreal. (2024). https://www.unrealengine.com/.

[268] Krauß, Veronika, Boden, Alexander, Oppermann, Leif, and Reiners, René. 2021. Current Practices, Challenges, and Design Implications for Collaborative AR/VR Application Development. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, (May 7, 2021), 1–15. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445335.

[269] Ashtari, Narges, Bunt, Andrea, McGrenere, Joanna, Nebeling, Michael, and Chilana, Parmit K. 2020. Creating Augmented and Virtual Reality Applications: Current Practices, Challenges, and Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (Apr. 21, 2020), 1–13. ISBN: 978-1-4503-6708-0. doi:10.1145/3313831.3376722.

[270] Krauß, Veronika, Nebeling, Michael, Jasche, Florian, and Boden, Alexander. 2022. Elements of XR Prototyping: Characterizing the Role and Use of Prototypes in Augmented and Virtual Reality Design. In *CHI Conference on Human Factors in Computing Systems*. CHI '22: CHI Conference on Human Factors in Computing Systems. ACM, New Orleans LA USA, (Apr. 29, 2022), 1–18. ISBN: 978-1-4503-9157-3. doi:10.1145/3491102.3517714.

[271] Apple. 2024. Reality Composer: Build Augmented Reality. (2024). https://apps.apple.com/us/app/reality-composer/id1462358802.

[272] Zollmann, Stefanie, Schall, Gerhard, Junghanns, Sebastian, and Reitmayr, Gerhard. 2012. Comprehensible and Interactive Visualizations of GIS Data in Augmented Reality. In *Advances in Visual Computing* (Lecture Notes in Computer Science). George Bebis et al., (Eds.) Springer, Berlin, Heidelberg, 675–685. ISBN: 978-3-642-33179-4. doi:10.1007/978-3-642-33179-4_64.

[273] Lee, Gun A., Kim, Gerard J., and Billinghurst, Mark. 2005. Immersive authoring: What You eXperience Is What You Get (WYXIWYG). *Communications of the ACM*, 48, 7, (July 1, 2005), 76–81. doi:10.1145/1070838.1070840.

[274] MacIntyre, Blair, Gandy, Maribeth, Dow, Steven, and Bolter, Jay David. 2004. DART: a toolkit for rapid design exploration of augmented reality experiences. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology* (UIST '04). Association for Computing Machinery, New York, NY, USA, (Oct. 24, 2004), 197–206. ISBN: 978-1-58113-957-0. doi:10.1145/1029632.1029669.

[275] Qian, Xun, He, Fengming, Hu, Xiyun, Wang, Tianyi, Ipsita, Ananya, and Ramani, Karthik. 2022. ScalAR: Authoring Semantically Adaptive Augmented Reality Experiences in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing*

Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, (Apr. 28, 2022), 1–18. ISBN: 978-1-4503-9157-3. doi:10.1145/3491102.3517665.

[276] Wang, Zeyu, Nguyen, Cuong, Asente, Paul, and Dorsey, Julie. 2021. DistanciAR: Authoring Site-Specific Augmented Reality Experiences for Remote Environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, (May 7, 2021), 1–12. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445552.

[277] Prouzeau, Arnaud, Wang, Yuchen, Ens, Barrett, Willett, Wesley, and Dwyer, Tim. 2020. Corsican Twin: Authoring In Situ Augmented Reality Visualisations in Virtual Reality. In *Proceedings of the International Conference on Advanced Visual Interfaces* (AVI '20). Association for Computing Machinery, New York, NY, USA, (Oct. 2, 2020), 1–9. ISBN: 978-1-4503-7535-1. doi:10.1145/3399715.3399743.

[278] Cavallo, Marco and Forbes, Angus G. 2019. CAVE-AR: A VR Authoring System to Interactively Design, Simulate, and Debug Multi-user AR Experiences. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, Osaka, Japan, (Mar. 2019), 872–873. ISBN: 978-1-7281-1377-7. doi:10.1109/VR.2019.8798148.

[279] Leiva, Germán, Grønbæk, Jens Emil, Klokmose, Clemens Nylandsted, Nguyen, Cuong, Kazi, Rubaiat Habib, and Asente, Paul. 2021. Rapido: Prototyping Interactive AR Experiences through Programming by Demonstration. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. UIST '21: The 34th Annual ACM Symposium on User Interface Software and Technology. ACM, Virtual Event USA, (Oct. 10, 2021), 626–637. ISBN: 978-1-4503-8635-7. doi:10.1145/3472749.3474774.

[280] Imottesjo, Hyekyung, Thuvander, Liane, Billger, Monica, Wallberg, Peter, Bodell, Gustav, Kain, Jaan-Henrik, and Nielsen, Stig Anton. 2020. Iterative Prototyping of Urban CoBuilder: Tracking Methods and User Interface of an Outdoor Mobile Augmented Reality Tool for Co-Designing. *Multimodal Technologies and Interaction*, 4, 2, (June 2020), 26, 2, (June 2020). doi:10.3390/mti4020026.

[281] Cao, Jacky, Lam, Kit-Yung, Lee, Lik-Hang, Liu, Xiaoli, Hui, Pan, and Su, Xiang. 2023. Mobile Augmented Reality: User Interfaces, Frameworks, and Intelligence. *ACM Computing Surveys*, 55, 9, (Sept. 30, 2023), 1–36. doi:10.1145/3557999.

[282] Guo, Anhong, Canberk, Ilter, Murphy, Hannah, Monroy-Hernández, Andrés, and Vaish, Rajan. 2019. Blocks: Collaborative and Persistent Augmented Reality Experiences. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3, 3, (Sept. 9, 2019), 1–24. doi:10.1145/3351241.

[283] Speicher, Maximilian, Hall, Brian D., Yu, Ao, Zhang, Bowen, Zhang, Haihua, Nebeling, Janet, and Nebeling, Michael. 2018. XD-AR: Challenges and Opportunities in Cross-Device Augmented Reality Application Development. *Proceedings of the ACM on Human-Computer Interaction*, 2, (June 19, 2018), 1–24, EICS, (June 19, 2018). doi:10.1145/3229089.

[284] Braun, Virginia, Clarke, Victoria, Hayfield, Nikki, and Terry, Gareth. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*. Pranee Liamputtong, (Ed.) Springer, Singapore, 843–860. ISBN: 978-981-10-5251-4. doi:10.1007/978-981-10-5251-4_103.

[285] Niantic. 2024. How to Use the Geospatial Browser. (2024). https://lightship.dev/docs/ardk/how-to/vps/tooling/use_geospatial_browser/.

[286] Walker, Charles J. 2010. Experiencing flow: Is doing it together better than doing it alone? *The Journal of Positive Psychology*, 5, 1, (Jan. 2010), 3–11. doi:10.1080/17439760903271116.

[287] Ng, Gary, Shin, Joon Gi, Plopski, Alexander, Sandor, Christian, and Saakes, Daniel. 2018. Situated Game Level Editing in Augmented Reality. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*. TEI '18: Twelfth International Conference on Tangible, Embedded, and Embodied Interaction. ACM, Stockholm Sweden, (Mar. 18, 2018), 409–418. ISBN: 978-1-4503-5568-1. doi:10.1145/3173225.3173230.

[288] Apple. 2024. Shortcuts User Guide. Retrieved Aug. 12, 2024 from https://support.apple.com/en-gb/guide/shortcuts/welcome/ios.

[289] Lakens, Daniël. 2022. Sample Size Justification. *Collabra: Psychology*, 8, 1, (Mar. 22, 2022), 33267. Don van Ravenzwaaij, (Ed.) doi:10.1525/collabra.33267.

[290] Lee, Benjamin, Sedlmair, Michael, and Schmalstieg, Dieter. 2023. Design Patterns for Situated Visualization in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 1–12. doi:10.1109/TVCG.2023.3327398.

[291] Hart, Sandra G. and Staveland, Lowell E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Human Mental Workload. Vol. 52. Peter A. Hancock and Najmedin Meshkati, (Eds.) North-Holland, (Jan. 1, 1988), 139–183. doi:10.1016/S0166-4115(08)62386-9.

[292] O'Brien, Heather L., Cairns, Paul, and Hall, Mark. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112, (Apr. 1, 2018), 28–39. doi:10.1016/j.ijhcs.2018.01.004.

[293] Fussell, Susan R., Setlock, Leslie D., Yang, Jie, Ou, Jiazhi, Mauer, Elizabeth, and Kramer, Adam D. I. 2004. Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Human–Computer Interaction*, 19, 3, (Sept. 1, 2004), 273–309. doi:10/cgz7rd.

[294] Kerbl, Bernhard, Kopanas, Georgios, Leimkuehler, Thomas, and Drettakis, George. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42, 4, (July 26, 2023), 139:1–139:14. doi:10.1145/3592433.

[295] Wu, Guanjun, Yi, Taoran, Fang, Jiemin, Xie, Lingxi, Zhang, Xiaopeng, Wei, Wei, Liu, Wenyu, Tian, Qi, and Wang, Xinggang. 2024. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (June 2024), 20310–20320. doi:10.1109/CVPR52733.2024.01920.

[296] Mihajlovic, Marko, Prokudin, Sergey, Tang, Siyu, Maier, Robert, Bogo, Federica, Tung, Tony, and Boyer, Edmond. 2024. SplatFields: Neural Gaussian Splats for Sparse 3D and 4D Reconstruction. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part II*. Springer-Verlag, Berlin, Heidelberg, (Oct. 20, 2024), 313–332. ISBN: 978-3-031-72626-2. doi:10.1007/978-3-031-72627-9_18.

[297] Sindagi, Vishwanath A. and Patel, Vishal M. 2017. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). (Oct. 2017), 1879–1888. doi:10.1109/ICCV.2017.206.

[298] Su, Xia, Zhang, Han, Cheng, Kaiming, Lee, Jaewook, Liu, Qiaochu, Olson, Wyatt, and Froehlich, Jon E. 2024. RASSAR: Room Accessibility and Safety Scanning in Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–17. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642140.

[299] Kim, Daehwa, Mollyn, Vimal, and Harrison, Chris. 2023. WorldPoint: Finger Pointing as a Rapid and Natural Trigger for In-the-Wild Mobile Interactions. *Proceedings of the ACM on Human-Computer Interaction*, 7, (Oct. 31, 2023), 357–375, ISS, (Oct. 31, 2023). doi:10.1145/3626478.

[300] Straub, Vincent J, Tsvetkova, Milena, and Yasseri, Taha. 2023. The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence*, 2, 2, (Apr. 1, 2023), 26339137231156912. doi:10.1177/26339137231156912.

[301] Makhmutov, Munir, Asapov, Timur, and Brown, Joseph Alexander. 2021. Safety Risks in Location-Based Augmented Reality Games. In *Entertainment Computing – ICEC 2021*. Jannicke Baalsrud Hauge, Jorge C. S. Cardoso, Licínio Roque, and Pedro A. Gonzalez-Calero, (Eds.) Springer International Publishing, Cham, 457–464. ISBN: 978-3-030-89394-1. doi:10.1007/978-3-030-89394-1_39.

[302] John, Oliver P. and Srivastava, Sanjay. 1999. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research, 2nd Ed.* Guilford Press, New York, NY, US, 102–138. ISBN: 978-1-57230-483-3.

[303] Steiner, Ivan D. 1972. *Group Process and Productivity. Social Psychology.* Academic Press, New York, NY. 204 pp. ISBN: 978-0-12-665350-2.

[304] Jurgelionis, Audrius, Oppermann, Leif, Blum, Lisa, and Wetzel, Richard. 2012. Shapes, marbles and pebbles: Template-based content creation for location-based games, (July 18, 2012). Retrieved Dec. 10, 2024 from https://repository.lincoln.ac.uk/articles/conference_contribution/Shapes_marbles_and_pebbles_template-based_content_creation_for_location-based_games/25169999/2.

[305] Seeliger, Arne, Weibel, Raphael P., and Feuerriegel, Stefan. 2024. Context-Adaptive Visual Cues for Safe Navigation in Augmented Reality Using Machine Learning. *International Journal of Human–Computer Interaction*, 40, 3, (Feb. 1, 2024), 761–781. doi:10.1080/10447318.2022.2122114.

[306] Evangelista Belo, João Marcelo, Lystbæk, Mathias N., Feit, Anna Maria, Pfeuffer, Ken, Kán, Peter, Oulasvirta, Antti, and Grønbæk, Kaj. 2022. AUIT – the Adaptive User Interfaces Toolkit for Designing XR Applications. In *The 35th Annual ACM Symposium on User Interface Software and Technology*. UIST '22: The 35th Annual ACM Symposium on User Interface Software and Technology. ACM, Bend OR USA, (Oct. 29, 2022), 1–16. ISBN: 978-1-4503-9320-1. doi:10.1145/3526113.3545651.

[307] Langlotz, Tobias, Regenbrecht, Holger, Zollmann, Stefanie, and Schmalstieg, Dieter. 2013. Audio stickies: visually-guided spatial audio annotations on a mobile augmented reality platform. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (OzCHI '13). Association for Computing Machinery, New York, NY, USA, (Nov. 25, 2013), 545–554. ISBN: 978-1-4503-2525-7. doi:10.1145/2541016.2541022.

[308] Kim, Tae Soo, Kim, Seungsu, Choi, Yoonseo, and Kim, Juho. 2021. Winder: Linking Speech and Visual Objects to Support Communication in Asynchronous Collaboration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21: CHI

Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, (May 6, 2021), 1–17. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445686.

[309] Stoakley, Richard, Conway, Matthew J., and Pausch, Randy. 1995. Virtual reality on a WIM: interactive worlds in miniature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '95*. The SIGCHI Conference. ACM Press, Denver, Colorado, United States, 265–272. ISBN: 978-0-201-84705-5. doi:10.1145/223904.223938.

[310] Lee, Jaewook, Jin, Fanjie, Kim, Younsoo, and Lindlbauer, David. 2022. User Preference for Navigation Instructions in Mixed Reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). (Mar. 2022), 802–811. doi:10.1109/VR51125.2022.00102.

[311] Baričević, Domagoj, Lee, Cha, Turk, Matthew, Höllerer, Tobias, and Bowman, Doug A. 2012. A hand-held AR magic lens with user-perspective rendering. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). (Nov. 2012), 197–206. doi:10.1109/ISMAR.2012.6402557.

[312] Kersten, Mik and Murphy, Gail C. 2006. Using task context to improve programmer productivity. In *Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (SIGSOFT '06/FSE-14). Association for Computing Machinery, New York, NY, USA, (Nov. 5, 2006), 1–11. ISBN: 978-1-59593-468-0. doi:10.1145/1181775.1181777.

[313] Veas, Eduardo, Grasset, Raphael, Kruijff, Ernst, and Schmalstieg, Dieter. 2012. Extended Overview Techniques for Outdoor Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 18, 4, (Apr. 2012), 565–572. doi:10/fxnspd.

[314] Vanukuru, Rishi, Weng, Suibi Che-Chuan, Ranjan, Krithik, Hopkins, Torin, Banic, Amy, Gross, Mark D., and Do, Ellen Yi-Luen. 2023. DualStream: Spatially Sharing Selves and Surroundings using Mobile Devices and Augmented Reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, Sydney, Australia, (Oct. 16, 2023), 138–147. ISBN: 979-8-3503-2838-7. doi:10.1109/ISMAR59233.2023.00028.

[315] Qian, Xun, Tan, Feitong, Zhang, Yinda, Collins, Brian Moreno, Kim, David, Olwal, Alex, Ramani, Karthik, and Du, Ruofei. 2024. ChatDirector: Enhancing Video Conferencing with Space-Aware Scene Rendering and Speech-Driven Layout Transition. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–16. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642110.

[316] Paul B. Paulus and Bernard A. Nijstad, red. 2003. *Group Creativity: Innovation through Collaboration. Group Creativity: Innovation through Collaboration.* Oxford University Press, New York, NY, US, xiii, 346. xiii, 346. ISBN: 978-0-19-514730-8. doi:10.1093/acprof:oso/9780195147308.001.0001.

[317] Rogers, Yvonne, Connelly, Kay, Tedesco, Lenore, Hazlewood, William, Kurtz, Andrew, Hall, Robert E., Hursey, Josh, and Toscos, Tammy. 2007. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *UbiComp 2007: Ubiquitous Computing*. John Krumm, Gregory D. Abowd, Aruna Seneviratne, and Thomas Strang, (Eds.) Springer, Berlin, Heidelberg, 336–353. ISBN: 978-3-540-74853-3. doi:10.1007/978-3-540-74853-3_20.

[318] Crabtree, Andy, Chamberlain, Alan, Grinter, Rebecca E., Jones, Matt, Rodden, Tom, and Rogers, Yvonne. 2013. Introduction to the Special Issue of "The Turn to The Wild". *ACM Trans. Comput.-Hum. Interact.*, 20, 3, (July 1, 2013), 13:1–13:4. doi:10.1145/2491500.2491501.

[319] Nogueira, Júlia Carneiro de Almeida, Gomes, Alex Sandro, Filho, Amadeu Sá de Campos, and Moreira, Fernando. 2023. Effectiveness of embodied evaluation of mobile applications: A qualitative study. *Heliyon*, 9, 6, (June 1, 2023), e17043. doi:10.1016/j.heliyon.2023.e17043.

[320] Lehman, Sarah M., Ling, Haibin, and Tan, Chiu C. 2020. ARCHIE: A User-Focused Framework for Testing Augmented Reality Applications in the Wild. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, Atlanta, GA, USA, (Mar. 2020), 903–912. ISBN: 978-1-7281-5608-8. doi:10.1109/VR46266.2020.00013.

[321] Zhang, Lei, Agrawal, Ashutosh, Oney, Steve, and Guo, Anhong. 2023. VRGit: A Version Control System for Collaborative Content Creation in Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, (Apr. 19, 2023), 1–14. ISBN: 978-1-4503-9421-5. doi:10.1145/3544548.3581136.

[322] Xia, Haijun, Herscher, Sebastian, Perlin, Ken, and Wigdor, Daniel. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST '18: The 31st Annual ACM Symposium on User Interface Software and Technology. ACM, Berlin Germany, (Oct. 11, 2018), 853–866. ISBN: 978-1-4503-5948-1. doi:10/gjbvfr.

[323] Irlitti, Andrew, Smith, Ross T., Von Itzstein, Stewart, Billinghurst, Mark, and Thomas, Bruce H. 2016. Challenges for Asynchronous Collaboration in Augmented Reality. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). (Sept. 2016), 31–35. doi:10/ggfv62.

[324] Wang, Cheng Yao, Sakashita, Mose, Ehsan, Upol, Li, Jingjin, and Won, Andrea Stevenson. 2020. Again, Together: Socially Reliving Virtual Reality Experiences When Separated. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, New York, NY, USA, (Apr. 23, 2020), 1–12. ISBN: 978-1-4503-6708-0. doi:10.1145/3313831.3376642.

[325] Cho, Hyunsung, Komar, Matthew L., and Lindlbauer, David. 2023. RealityReplay: Detecting and Replaying Temporal Changes In Situ Using Mixed Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7, 3, (Sept. 27, 2023), 90:1–90:25. doi:10.1145/3610888.

[326] Wang, Ruotong, Qiu, Lin, Cranshaw, Justin, and Zhang, Amy X. 2024. Meeting Bridges: Designing Information Artifacts that Bridge from Synchronous Meetings to Asynchronous Collaboration. *Proc. ACM Hum.-Comput. Interact.*, 8, (Apr. 26, 2024), 35:1–35:29, CSCW1, (Apr. 26, 2024). doi:10.1145/3637312.

[327] Dell, Nicola, Vaidyanathan, Vidya, Medhi, Indrani, Cutrell, Edward, and Thies, William. 2012. "Yours is better!": participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). Association for Computing Machinery, New York, NY, USA, (May 5, 2012), 1321–1330. ISBN: 978-1-4503-1015-4. doi:10.1145/2207676.2208589.

[328] Ablett, Daniel. 2022. Portal Rendering and Creation Interactions in Virtual Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. ISMAR. (Sept. 2022), 9.

[329] Das, Sauvik, Lu, David, Lee, Taehoon, Lo, Joanne, and Hong, Jason I. 2019. The Memory Palace: Exploring Visual-Spatial Paths for Strong, Memorable, Infrequent Authentication. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (UIST '19). Association for Computing Machinery, New York, NY, USA, (Oct. 17, 2019), 1109–1121. ISBN: 978-1-4503-6816-2. doi:10.1145/3332165.3347917.

[330] Krokos, Eric, Plaisant, Catherine, and Varshney, Amitabh. 2019. Virtual memory palaces: immersion aids recall. *Virtual Reality*, 23, 1, (Mar. 1, 2019), 1–15. doi:10.1007/s10055-018-0346-3.

[331] Essoe, Joey Ka-Yee, Reggente, Nicco, Ohno, Ai Aileen, Baek, Younji Hera, Dell'Italia, John, and Rissman, Jesse. 2022. Enhancing learning and retention with distinctive virtual reality environments and mental context reinstatement. *npj Science of Learning*, 7, 1, (Dec. 8, 2022), 1–14. doi:10.1038/s41539-022-00147-6.

[332] Koch, Griffin E. and Coutanche, Marc N. 2024. Context reinstatement requires a schema relevant virtual environment to benefit object recall. *Psychonomic Bulletin & Review*, (Mar. 1, 2024). doi:10.3758/s13423-024-02472-w.

[333] Poole, Ben, Jain, Ajay, Barron, Jonathan T., and Mildenhall, Ben. 2022. DreamFusion: Text-to-3D using 2D Diffusion. (Sept. 29, 2022). arXiv: 2209.14988 [cs]. Pre-published.

[334] Lin, Chen-Hsuan et al. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Vancouver, BC, Canada, (June 2023), 300–309. ISBN: 979-8-3503-0129-8. doi:10.1109/CVPR52729.2023.00037.

[335] Chen, Zhaoxi, Wang, Guangcong, and Liu, Ziwei. 2022. Text2Light: Zero-Shot Text-Driven HDR Panorama Generation. *ACM Trans. Graph.*, 41, 6, (Nov. 30, 2022), 195:1–195:16. doi:10.1145/3550454.3555447.

[336] Stan, Gabriela Ben Melech et al. 2023. LDM3D: Latent Diffusion Model for 3D. (May 21, 2023). arXiv: 2305.10853 [cs]. Pre-published.

[337] Stan, Gabriela Ben Melech, Wofk, Diana, Aflalo, Estelle, Tseng, Shao-Yen, Cai, Zhipeng, Paulitsch, Michael, and Lal, Vasudev. 2023. LDM3D-VR: Latent Diffusion Model for 3D VR. (Nov. 6, 2023). arXiv: 2311.03226 [cs]. Pre-published.

[338] Tang, Shitao, Zhang, Fuyang, Chen, Jiacheng, Wang, Peng, and Furukawa, Yasutaka. 2024. MVDiffusion: enabling holistic multi-view image generation with correspondence-aware diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, (May 30, 2024), 51202–51233.

[339] Höllein, Lukas, Cao, Ang, Owens, Andrew, Johnson, Justin, and Nießner, Matthias. 2023. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 7909–7920. Retrieved July 13, 2024 from https://openaccess.thecvf.com/content/ICCV2023/html/Hollein_Text2Room_Extracting_Textured_3D_Meshes_from_2D_Text-to-Image_Models_ICCV_2023_paper.html.

[340] Zhou, Xiaoyu, Ran, Xingjian, Xiong, Yajiao, He, Jinlin, Lin, Zhiwei, Wang, Yongtao, Sun, Deqing, and Yang, Ming-Hsuan. 2024. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. (Feb. 11, 2024). arXiv: 2402.07207 [cs]. Retrieved Feb. 13, 2024 from http://arxiv.org/abs/2402.07207. Pre-published.

[341] Fang, Chuan, Hu, Xiaotao, Luo, Kunming, and Tan, Ping. 2023. Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints. (Oct. 8, 2023). arXiv: 2310.03602 [cs]. Pre-published.

[342] Schult, Jonas et al. 2024. ControlRoom3D: Room Generation using Semantic Proxy Rooms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6201–6210. Retrieved July 14, 2024 from https://openaccess.thecvf.com/content/CVPR2024/html/Schult_ControlRoom3D_Room_Generation_using_Semantic_Proxy_Rooms_CVPR_2024_paper.html.

[343] Bautista, Miguel Angel et al. 2022. GAUDI: a neural architect for immersive 3D scene generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, (Nov. 2022), 25102–25116. ISBN: 978-1-7138-7108-8.

[344] Wang, Xinpeng, Yeshwanth, Chandan, and Nießner, Matthias. 2021. SceneFormer: Indoor Scene Generation with Transformers. In *2021 International Conference on 3D Vision (3DV)*. 2021 International Conference on 3D Vision (3DV). (Dec. 2021), 106–115. doi:10.1109/3DV53792.2021.00021.

[345] [n. d.] Bezi | AI Assistance for Unity Developers & Studios. Bezi. Retrieved Dec. 12, 2025 from https://www.bezi.com/.

[346] [n. d.] Unity AI: AI Game Development Tools & RT3D Software. Unity. Retrieved Dec. 12, 2025 from https://unity.com/features/ai.

[347] Hu, Erzhen, Li, Mingyi, Hong, Jungtaek, Qian, Xun, Olwal, Alex, Kim, David, Heo, Seongkook, and Du, Ruofei. 2025. Thing2Reality: Enabling Spontaneous Creation of 3D Objects from 2D Content using Generative AI in XR Meetings. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (UIST '25). Association for Computing Machinery, New York, NY, USA, (Sept. 27, 2025), 1–16. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747621.

[348] Dang, Hai, Brudy, Frederik, Fitzmaurice, George, and Anderson, Fraser. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST

'23). Association for Computing Machinery, New York, NY, USA, (Oct. 29, 2023), 1–17. ISBN: 979-8-4007-0132-0. doi:10.1145/3586183.3606772.

[349] Fridman, Rafail, Abecasis, Amit, Kasten, Yoni, and Dekel, Tali. 2023. SceneScape: Text-Driven Consistent Scene Generation. *Advances in Neural Information Processing Systems*, 36, (Dec. 15, 2023), 39897–39914. Retrieved July 14, 2024 from https://proceedings.neurips.cc/paper_files/paper/2023/hash/7d62a85ebfed2f680eb5544beae93191-Abstract-Conference.html.

[350] Coyne, Bob and Sproat, Richard. 2001. WordsEye: an automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '01). Association for Computing Machinery, New York, NY, USA, (Aug. 1, 2001), 487–496. ISBN: 978-1-58113-374-5. doi:10.1145/383259.383316.

[351] Van Der Linden, Roland, Lopes, Ricardo, and Bidarra, Rafael. 2014. Procedural Generation of Dungeons. *IEEE Transactions on Computational Intelligence and AI in Games*, 6, 1, (Mar. 2014), 78–89. doi:10.1109/TCIAIG.2013.2290371.

[352] Sra, Misha, Garrido-Jurado, Sergio, and Maes, Pattie. 2018. Oasis: Procedurally Generated Social Virtual Spaces from 3D Scanned Real Spaces. *IEEE Transactions on Visualization and Computer Graphics*, 24, 12, (Dec. 1, 2018), 3174–3187. doi:10.1109/TVCG.2017.2762691.

[353] Cheng, Lung-Pan, Ofek, Eyal, Holz, Christian, and Wilson, Andrew D. 2019. VRoamer: Generating On-The-Fly VR Experiences While Walking inside Large, Unknown Real-World Building Environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, Osaka, Japan, (Mar. 2019), 359–366. ISBN: 978-1-7281-1377-7. doi:10.1109/VR.2019.8798074.

[354] Song, Liangchen, Cao, Liangliang, Xu, Hongyu, Kang, Kai, Tang, Feng, Yuan, Junsong, and Yang, Zhao. 2023. RoomDreamer: Text-Driven 3D Indoor Scene Synthesis with Coherent Geometry and Texture. In *Proceedings of the 31st ACM International Conference on Multimedia* (MM '23). Association for Computing Machinery, New York, NY, USA, (Oct. 27, 2023), 6898–6906. ISBN: 979-8-4007-0108-5. doi:10.1145/3581783.3611800.

[355] Zhang, Jingbo, Li, Xiaoyu, Wan, Ziyu, Wang, Can, and Liao, Jing. 2023. Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields. (May 19, 2023). arXiv: 2305.11588 [cs]. Retrieved Oct. 21, 2023 from http://arxiv.org/abs/2305.11588. Pre-published.

[356] Chung, Jaeyoung, Lee, Suyoung, Nam, Hyeongjin, Lee, Jaerin, and Lee, Kyoung Mu. 2023. LucidDreamer: Domain-free Generation of 3D Gaussian Splatting Scenes. (Nov. 23, 2023). arXiv: 2311.13384 [cs]. Pre-published.

[357] Ju, Xiaoliang, Huang, Zhaoyang, Li, Yijin, Zhang, Guofeng, Qiao, Yu, and Li, Hongsheng. 2024. DiffInDScene: Diffusion-based High-Quality 3D Indoor Scene Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4526–4535. Retrieved July 13, 2024 from https://openaccess.thecvf.com/content/CVPR2024/html/Ju_DiffInDScene_Diffusion-based_High-Quality_3D_Indoor_Scene_Generation_CVPR_2024_paper.html.

[358] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Bjorn. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, (June 2022), 10674–10685. ISBN: 978-1-6654-6946-3. doi:10.1109/CVPR52688.2022.01042.

[359] Bhat, Shariq Farooq, Birkl, Reiner, Wofk, Diana, Wonka, Peter, and Müller, Matthias. 2023. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. (Feb. 23, 2023). arXiv: 2302.12288 [cs]. Pre-published.

[360] Bae, Gwangbin, Budvytis, Ignas, and Cipolla, Roberto. 2022. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In *British Machine Vision Conference (BMVC)*. British Machine Vision Conference (BMVC).

[361] Bar-Tal, Omer, Yariv, Lior, Lipman, Yaron, and Dekel, Tali. 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning* (ICML'23). Vol. 202. JMLR.org, Honolulu, Hawaii, USA, (July 23, 2023), 1737–1752.

[362] Kang, Minguk, Zhu, Jun-Yan, Zhang, Richard, Park, Jaesik, Shechtman, Eli, Paris, Sylvain, and Park, Taesung. 2023. Scaling up GANs for Text-to-Image Synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 10124–10134.

[363] [n. d.] Skybox AI. Skybox AI. Retrieved Feb. 2, 2025 from https://skybox.blockadelabs.com/.

[364] Li, Junnan, Li, Dongxu, Savarese, Silvio, and Hoi, Steven. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (ICML'23). Vol. 202. JMLR.org, Honolulu, Hawaii, USA, (July 23, 2023), 19730–19742.

[365] OpenAI. 2023. GPT-4 Technical Report. (Mar. 27, 2023). arXiv: 2303.08774 [cs]. Prepublished.

[366] Jain, Jitesh, Li, Jiachen, Chiu, MangTik, Hassani, Ali, Orlov, Nikita, and Shi, Humphrey. 2023. OneFormer: One Transformer to Rule Universal Image Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (June 2023), 2989–2998. doi:10.1109/CVPR52729.2023.00292.

[367] Huang, Tianyu et al. 2025. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. *ACM Trans. Graph.*, 44, 6, (Dec. 4, 2025), 245:1–245:15. doi:10.1145/3763330.

[368] Ren, Xuanchi et al. 2025. Gen3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (June 2025), 6121–6132. doi:10.1109/CVPR52734.2025.00574.

[369] Yang, Xiuyu, Man, Yunze, Chen, Jun-Kun, and Wang, Yu-Xiong. 2024. SceneCraft: Layout-Guided 3D Scene Generation. *Advances in Neural Information Processing Systems*, 37, (Dec. 16, 2024), 82060–82084. doi:10.52202/079017-2608.

[370] Liu, Libin, Chen, Shen, Jia, Sen, Shi, Jingzhe, Wu, Zongkai, Hwang, Jenq-Neng, and Li, Lei. 2025. Graph Canvas for Controllable 3D Scene Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia* (MM '25). Association for Computing Machinery, New York, NY, USA, (Oct. 27, 2025), 2536–2545. ISBN: 979-8-4007-2035-2. doi:10.1145/3746027.3754554.

[371] Chen, Yiwen, Nguyen, Hieu T, Voleti, Vikram, Jampani, Varun, and Jiang, Huaizu. 2025. HouseCrafter: Lifting Floorplans to 3D Scenes with 2D Diffusion Models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, (Oct. 2025), 28440–28450.

[372] Liu, Yuheng, Li, Xinke, Zhang, Yuning, Qi, Lu, Li, Xin, Wang, Wenping, Li, Chongshou, Li, Xueting, and Yang, Ming-Hsuan. 2025. Controllable 3D Outdoor Scene Generation via Scene Graphs. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, (Oct. 2025).

[373] World Labs. 2025. Marble: A Multimodal World Model. (Oct. 2025). Retrieved Nov. 23, 2025 from https://www.worldlabs.ai/blog/marble-world-model.

[374] Zhou, Bolei, Zhao, Hang, Puig, Xavier, Fidler, Sanja, Barriuso, Adela, and Torralba, Antonio. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, (July 2017), 5122–5130. ISBN: 978-1-5386-0457-1. doi:10.1109/CVPR.2017.544.

[375] Zhang, Lvmin, Rao, Anyi, and Agrawala, Maneesh. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Paris, France, (Oct. 1, 2023), 3813–3824. ISBN: 979-8-3503-0718-4. doi:10.1109/ICCV51070.2023.00355.

[376] Song, Shuran, Lichtenberg, Samuel P., and Xiao, Jianxiong. 2015. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 567–576. Retrieved Oct. 30, 2023 from https://openaccess.thecvf.com/content_cvpr_2015/html/Song_SUN_RGB-D_A_2015_CVPR_paper.html.

[377] Zhang, Y., Yu, F., Song, S., Xu, P., Seff, A., and Xiao, J. 2016. Large-scale scene understanding challenge: Room layout estimation. http://lsun.cs.princeton.edu/2016/.

[378] Yu, Fisher, Seff, Ari, Zhang, Yinda, Song, Shuran, Funkhouser, Thomas, and Xiao, Jianxiong. 2016. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. (June 4, 2016). arXiv: 1506.03365 [cs]. Pre-published.

[379] Bhat, Shariq Farooq, Mitra, Niloy, and Wonka, Peter. 2024. LOOSECONTROL: Lifting ControlNet for Generalized Depth Conditioning. In *ACM SIGGRAPH 2024 Conference Papers* (SIGGRAPH '24). Association for Computing Machinery, New York, NY, USA, (July 13, 2024), 1–11. ISBN: 979-8-4007-0525-0. doi:10.1145/3641519.3657525.

[380] Hartmann, Tilo et al. 2016. The Spatial Presence Experience Scale (SPES): A short self-report measure for diverse media settings. *Journal of Media Psychology: Theories, Methods, and Applications*, 28, 1, 1–15. doi:10.1027/1864-1105/a000137.

[381] De La Torre, Fernanda, Fang, Cathy Mengying, Huang, Han, Banburski-Fahey, Andrzej, Amores Fernandez, Judith, and Lanier, Jaron. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (May 11, 2024), 1–22. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3642579.

[382] Gonzalez Diaz, Carlos, Tang, John, Sarkar, Advait, and Rintel, Sean. 2022. Making Space for Social Time: Supporting Conversational Transitions Before, During, and After Video Meetings. In *2022 Symposium on Human-Computer Interaction for Work*. CHIWORK 2022: 2022 Symposium on Human-Computer Interaction for Work. ACM, Durham NH USA, (June 8, 2022), 1–11. ISBN: 978-1-4503-9655-4. doi:10.1145/3533406.3533417.

[383] Suh, Do Ho. 2024. Perfect Home: London, Horsham, New York, Berlin, Providence, Seoul. (2024).

[384] Yang, Lihe, Kang, Bingyi, Huang, Zilong, Xu, Xiaogang, Feng, Jiashi, and Zhao, Hengshuang. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[385] Paschalidou, Despoina, Kar, Amlan, Shugrina, Maria, Kreis, Karsten, Geiger, Andreas, and Fidler, Sanja. 2021. ATISS: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[386] Feng, Weixi, Zhu, Wanrong, Fu, Tsu-Jui, Jampani, Varun, Akula, Arjun, He, Xuehai, Basu, S., Wang, Xin Eric, and Wang, William Yang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. *Advances in Neural Information Processing Systems*, 36, (Dec. 15, 2023), 18225–18250. Retrieved July 24, 2024 from https://proceedings.neurips.cc/paper_files/paper/2023/hash/3a7f9e485845dac27423375c934cb4db-Abstract-Conference.html.

[387] Sauer, Axel, Boesel, Frederic, Dockhorn, Tim, Blattmann, Andreas, Esser, Patrick, and Rombach, Robin. 2024. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. (Mar. 18, 2024). arXiv: 2403.12015 [cs]. Pre-published.

[388] Li, Yuheng, Liu, Haotian, Wu, Qingyang, Mu, Fangzhou, Yang, Jianwei, Gao, Jianfeng, Li, Chunyuan, and Lee, Yong Jae. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

[389] Phung, Quynh, Ge, Songwei, and Huang, Jia-Bin. 2024. Grounded Text-to-Image Synthesis with Attention Refocusing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[390] Wang, Zhouxia, Yuan, Ziyang, Wang, Xintao, Li, Yaowei, Chen, Tianshui, Xia, Menghan, Luo, Ping, and Shan, Ying. 2024. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *ACM SIGGRAPH 2024 Conference Papers* (SIGGRAPH '24). Associa-

tion for Computing Machinery, New York, NY, USA, (July 13, 2024), 1–11. ISBN: 979-8-4007-0525-0. doi:10.1145/3641519.3657518.

[391] Deng, Boyang, Tucker, Richard, Li, Zhengqi, Guibas, Leonidas, Snavely, Noah, and Wetzstein, Gordon. 2024. Streetscapes: Large-scale Consistent Street View Generation Using Autoregressive Video Diffusion. In *ACM SIGGRAPH 2024 Conference Papers* (SIGGRAPH '24). Association for Computing Machinery, New York, NY, USA, (July 13, 2024), 1–11. ISBN: 979-8-4007-0525-0. doi:10.1145/3641519.3657513.

[392] [n. d.] Spatial - Create Immersive UGC, Virtual Classrooms, Experiential Marketing. Spatial. Retrieved Feb. 2, 2025 from https://www.spatial.io/.

[393] [n. d.] Microsoft Mesh. Microsoft. Retrieved Feb. 2, 2025 from https://adoption.microsoft.com/en-us/mesh/.

[394] Delaney, Declan, Ward, Tomás, and McLoone, Seamus. 2006. On Consistency and Network Latency in Distributed Interactive Applications: A Survey—Part I. *Presence*, 15, 2, (Apr. 2006), 218–234. doi:10.1162/pres.2006.15.2.218.

[395] Friston, Sebastian and Steed, Anthony. 2014. Measuring latency in virtual environments. *IEEE transactions on visualization and computer graphics*, 20, 4, 616–625. doi:10.1109/TVCG.2014.30.

[396] Hartmann, Tilo and Hofer, Matthias. 2022. I Know It Is Not Real (And That Matters) Media Awareness vs. Presence in a Parallel Processing Account of the VR Experience. *Frontiers in Virtual Reality*, 3. doi:10.3389/frvir.2022.694048.

[397] Bonnail, Elise, Frommel, Julian, Lecolinet, Eric, Huron, Samuel, and Gugenheimer, Jan. 2024. Was it Real or Virtual? Confirming the Occurrence and Explaining Causes of Memory Source Confusion between Reality and Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, (May 11, 2024), 1–17. ISBN: 979-8-4007-0330-0. doi:10.1145/3613904.3641992.

[398] Wu, Like, Xu, Shihui, Liao, Wenjie, and Fujimura, Shigeru. 2025. Augmented Remote Assistance for Quality Inspection: A Cross-Reality Collaborative System With Virtual Replicas. *International Journal of Human–Computer Interaction*, (July 27, 2025), 1–37. doi:10.1080/10447318.2025.2534059.

[399] [n. d.] XYZ Reality | Construction delivery platform for data centres projects. XYZ Reality. Retrieved Jan. 28, 2026 from https://www.xyzreality.com/.

[400] Khanal, Shishir, Medasetti, Uma Shankar, Mashal, Mustafa, Savage, Bruce, and Khadka, Rajiv. 2022. Virtual and Augmented Reality in the Disaster Management Technology: A Literature Review of the Past 11 years. *Frontiers in Virtual Reality*, 3, (Apr. 11, 2022). doi:10.3389/frvir.2022.843195.

[401] Stern, Michael, Vergari, Maurizio, Schorlemmer, Julia, Vona, Francesco, Grieshammer, David, and Voigt-Antons, Jan-Niklas. 2025. The Impact of Spatial Misalignment and Time Delay on Collaborative Presence in Augmented Reality. In *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*. 2025 17th International Conference on Quality of Multimedia Experience (QoMEX). (Sept. 2025), 1–7. doi:10.1109/QoMEX65720.2025.11219982.

[402] Cidota, Marina, Lukosch, Stephan, Datcu, Dragos, and Lukosch, Heide. 2016. Comparing the Effect of Audio and Visual Notifications on Workspace Awareness Using Head-Mounted Displays for Remote Collaboration in Augmented Reality. *Augmented Human Research*, 1, 1, (Oct. 10, 2016), 1. doi:10.1007/s41133-016-0003-x.

[403] Benford, Steve, Greenhalgh, Chris, and Lloyd, David. 1997. Crowded collaborative virtual environments. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI '97). Association for Computing Machinery, New York, NY, USA, (Mar. 27, 1997), 59–66. ISBN: 978-0-89791-802-2. doi:10/b9cwmc.

[404] Vaghi, Ivan, Greenhalgh, Chris, and Benford, Steve. 1999. Coping with inconsistency due to network delays in collaborative virtual environments. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST '99*. The ACM Symposium. ACM Press, London, United Kingdom, 42–49. ISBN: 978-1-58113-141-3. doi:10/bszk5m.

[405] Kleinau, Julia, Grønbæk, Jens Emil Sloth, and Hoggan, Eve. 2025. Co-Designing Multimodal Tools for Radically Mobile Hybrid Meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, (Apr. 25, 2025), 1–17. ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3713993.

[406] Lee, Geonsun, Yang, Yue, Healey, Jennifer, and Manocha, Dinesh. 2025. Since U Been Gone: Augmenting Context-Aware Transcriptions for Re-Engaging in Immersive VR Meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI 2025: CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, (Apr. 26, 2025), 1–20. ISBN: 979-8-4007-1394-1. doi:10.1145/3706598.3714078.

[407] Yildiz, Zeynep and Subasi, Ozge. 2023. Virtual Collaboration Tools for Mixed-Ability Workspaces: A Cross Disability Solidarity Case from Turkey. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, (Apr. 19, 2023), 1–11. ISBN: 978-1-4503-9421-5. doi:10.1145/3544548.3580937.

[408] Gonçalves, David, Rodrigues, André, Richardson, Mike L., de Sousa, Alexandra A., Proulx, Michael J., and Guerreiro, Tiago. 2021. Exploring Asymmetric Roles in Mixed-Ability Gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). Association for Computing Machinery, New York, NY, USA, (May 7, 2021), 1–14. ISBN: 978-1-4503-8096-6. doi:10.1145/3411764.3445494.

[409] Trindade, Pedro, Gonçalves, David, Pais, Pedro, Guerreiro, Jão, Guerreiro, Tiago, and Rodrigues, André. 2025. Exploring Asymmetric Competitive Gaming For Mixed-Visual-Ability Pairs. In *Videogame Sciences and Arts*. Anabela Marto, Rui Prada, Patrícia Gouveia, Ruth Contreras- Espinosa, Alexandrino Gonçalves, Eduarda Abrantes, and Roberto Ribeiro, (Eds.) Springer Nature Switzerland, Cham, 21–39. ISBN: 978-3-031-81713-7. doi:10.1007/978-3-031-81713-7_2.

[410] Brown, Julie A. 2019. An Exploration of Virtual Reality Use and Application Among Older Adult Populations. *Gerontology and Geriatric Medicine*, 5, (Jan. 1, 2019), 2333721419885287. doi:10.1177/2333721419885287.

[411] Lindlbauer, David, Feit, Anna Maria, and Hilliges, Otmar. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST '19: The 32nd Annual ACM Symposium on User Interface Software and Technology. ACM, New Orleans LA USA, (Oct. 17, 2019), 147–160. ISBN: 978-1-4503-6816-2. doi:10.1145/3332165.3347945.

[412] Cheng, Yifei, Yan, Yukang, Yi, Xin, Shi, Yuanchun, and Lindlbauer, David. 2021. SemanticAdapt: Optimization-based Adaptation of Mixed Reality Layouts Leveraging Virtual-Physical Semantic Connections. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. UIST '21: The 34th Annual ACM Symposium on User Interface Software and Technology. ACM, Virtual Event USA, (Oct. 10, 2021), 282–297. ISBN: 978-1-4503-8635-7. doi:10/gm5z7h.

[413] Li, Zhipeng, Gebhardt, Christoph, Inglin, Yves, Steck, Nicolas, Streli, Paul, and Holz, Christian. 2024. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (UIST '24). Association for Computing Machinery,

New York, NY, USA, (Oct. 11, 2024), 1–13. ISBN: 979-8-4007-0628-8. doi:10.1145/3654777.3676470.

[414] Lee, Jaewook et al. 2025. ImaginateAR: AI-Assisted In-Situ Authoring in Augmented Reality. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (UIST '25). Association for Computing Machinery, New York, NY, USA, (Sept. 27, 2025), 1–21. ISBN: 979-8-4007-2037-6. doi:10.1145/3746059.3747635.

[415] Kobenova, Amina, DeVeaux, Cyan, Parajuli, Samyak, Banburski-Fahey, Andrzej, Amores, Judith, and Lanier, Jaron. 2025. Social Conjuring: Multi-User Runtime Collaboration with GenAI in Building Virtual Reality Worlds. In *Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology* (VRST '25). Association for Computing Machinery, New York, NY, USA, (Dec. 4, 2025), 1–3. ISBN: 979-8-4007-2118-2. doi:10.1145/3756884.3768382.

[416] Horvitz, Eric. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '99). Association for Computing Machinery, New York, NY, USA, (May 1, 1999), 159–166. ISBN: 978-0-201-48559-2. doi:10.1145/302979.303030.

[417] Allen, J.E., Guinn, C.I., and Horvtz, E. 1999. Mixed-Initiative Interaction. *IEEE Intelligent Systems and their Applications*, 14, 5, (Sept. 1999), 14–23. doi:10.1109/5254.796083.

[418] Fan, Haoxiang, Yuan, Yiwei, Hsia, Chi, Liang, Chen, and Peng, Zhenhui. 2026. Toward Personalized Proactive Agents on Smart Glasses: From Explicit Context Policies to Implicit User Differences. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems* (CHI EA '26). Association for Computing Machinery, New York, NY, USA, (Apr. 13, 2026), 1–6. ISBN: 979-8-4007-2281-3. doi:10.1145/3772363.3798509.

[419] Kim, Yoonsang, Jadeja, Devshree, Pradhan, Divyansh, Yang, Yalong, and Kaufman, Arie E. 2026. SpeechLess: Micro-utterance with Personalized Spatial Memory-aware Assistant in Everyday Augmented Reality. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). (Mar. 2026), 217–227. doi:10.1109/VR67842.2026.00044.

[420] Shneiderman, Ben and Maes, Pattie. 1997. Direct Manipulation vs. Interface Agents. *Interactions*, 4, 6, (Nov. 1, 1997), 42–61. doi:10.1145/267505.267514.

[421] Huang, Zhening, Wu, Xiaoyang, Zhong, Fangcheng, Zhao, Hengshuang, Nießner, Matthias, and Lasenby, Joan. 2025. LiteReality: Graphics-ready 3D scene reconstruction from RGB-d scans. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[422] Xia, Hongchi et al. 2025. DRAWER: Digital Reconstruction and Articulation With Environment Realism. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (June 2025), 21771–21782. doi:10.1109/CVPR52734.2025.02028.