BlendScape: Enabling End-User Customization of Video-Conferencing Environments through Generative AI

Shwetha Rajaram*
Microsoft Research
United States
shwethar@umich.edu

Nels Numan*
Microsoft Research
United States
nels.numan@ucl.ac.uk

Balasaravanan Thoravi Kumaravel
Microsoft Research
United States
bala.kumaravel@microsoft.com

Nicolai Marquardt
Microsoft Research
United States
nicmarquardt@microsoft.com

Andrew D. Wilson
Microsoft Research
United States
awilson@microsoft.com

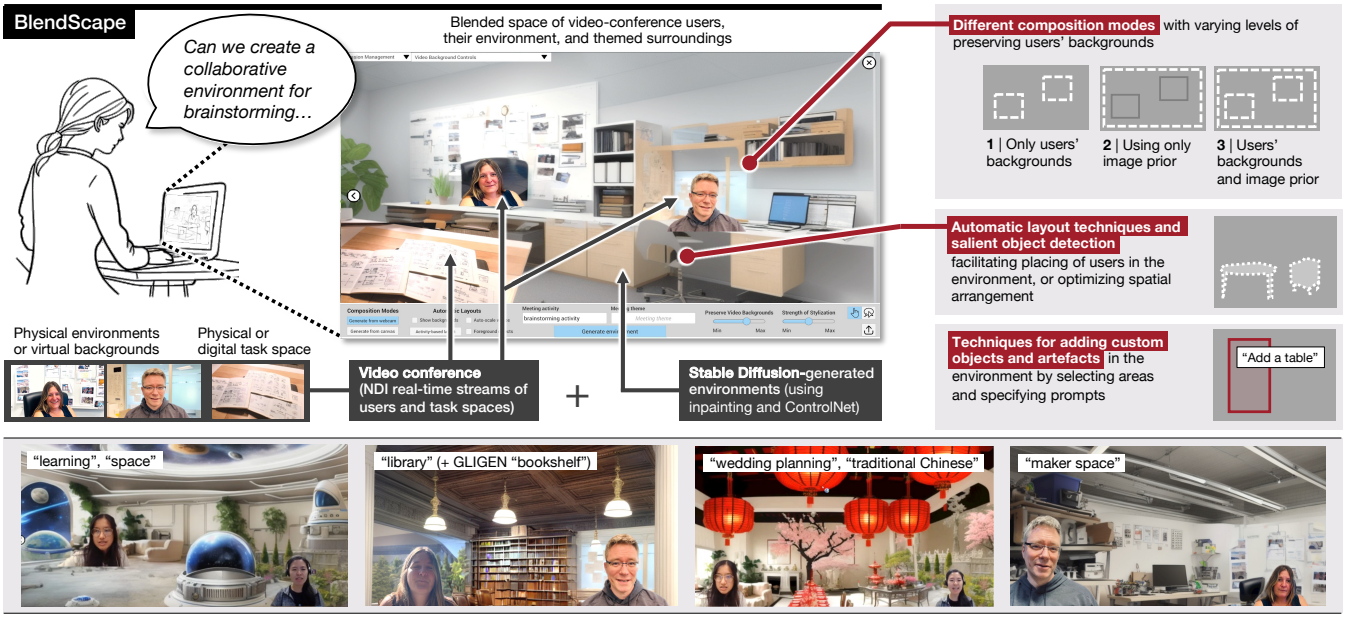

Figure 1: Overview of BlendScape, a rendering and composition system for end-users to customize video-conference environments by leveraging AI image generation techniques.

ABSTRACT

Today's video-conferencing tools support a rich range of professional and social activities, but their generic meeting environments cannot be dynamically adapted to align with distributed collaborators' needs. To enable end-user customization, we developed BlendScape, a rendering and composition system for video-conferencing participants to tailor environments to their meeting context by leveraging AI image generation techniques. BlendScape supports flexible representations of task spaces by blending users' physical or digital backgrounds into unified environments and implements multimodal interaction techniques to steer the generation. Through an exploratory study with 15 end-users, we investigated whether and how they would find value in using generative AI to customize video-conferencing environments. Participants envisioned using a system like BlendScape to facilitate collaborative activities in the future, but required further controls to mitigate distracting or unrealistic visual elements. We implemented scenarios to demonstrate BlendScape's expressiveness for supporting environment design strategies from prior work and propose composition techniques to improve the quality of environments.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Collaborative and social computing systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

*This work was done while the first two authors were interns at Microsoft Research. Both authors contributed equally to the paper.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

KEYWORDS

video-conferencing, generative AI, end-user customization

ACM Reference Format:
Shwetha Rajaram, Nels Numan, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D. Wilson. 2024. BlendScape: Enabling End-User Customization of Video-Conferencing Environments through Generative AI. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3654777.3676326

1 INTRODUCTION

Advances in video-conferencing technologies and their increasing availability over the past few decades have enabled distributed users to collaborate on activities that previously required face-to-face interaction. Since the COVID-19 pandemic, video-conferencing has gained popularity not just for facilitating professional tasks (e.g., remote work and distance learning [3, 39]), but also for health appointments [26], social gatherings [21], and hobbies [8, 62]. However, today's video-conferencing tools do not reflect the rich range of activities that they are used for, due to how they compose *meeting environments* (i.e., the "stage" or background rendered around users' videos). Users are typically placed in separate regions of a video grid within generic meeting rooms, which can lead to meeting fatigue [19], reduce user engagement [10], and disrupt interpersonal cues for mediating conversations [28, 56].

To support more expressive video-conferencing environments that are aligned with distributed collaborators' needs, we envision leveraging generative AI to enable end-users to create custom meeting environments. To understand the existing design space, we reviewed video-mediated communication research that redesigned meeting spaces to mitigate challenges with distributed collaboration (e.g., communication barriers, decreased sense of co-presence). We identify three main design strategies: *(1)* Establishing the meeting context through the environment (e.g., by rendering shared task spaces [24, 29] or thematic visuals [20, 30]); *(2)* Leveraging spatial metaphors to enhance communication (e.g., facilitating turn-taking via proxemic interactions between users [28]); *(3)* Using the environment to record a meeting history, to aid future collaboration [59]. Despite the HCI community's knowledge of effective meeting environment designs and empirical studies demonstrating their benefits for distributed collaboration, there is a lack of tool support for end-users to implement these designs in real-time. Commercial customization tools[1] require significant manual effort, making it infeasible to adapt environments as meetings progress [23].

As a step towards this vision, we developed **BlendScape, a rendering and composition system for video-conferencing participants to create environments tailored to their meeting context** (Fig. 1). We introduce two key innovations: *(1)* We ground the generation of meeting environments in real spaces that are meaningful to users by blending their physical or virtual backgrounds into a unified environment. This can serve as a mechanism for personalization [58] or to incorporate physical objects to collaborate around [29, 34]. *(2)* Capitalizing on recent advances in generative AI, we leverage image generation models to enable expressive and rapid techniques for composing environment designs. While such techniques are the subject of ongoing research, several dominant modes have emerged:

- *text-to-image*: generating an image from a given text prompt.
- *image-to-image*: generating an image from a text prompt and an *image prior* (i.e., input image), retaining features of the image prior while introducing new elements or styles consistent with the prompt.
- *inpainting*: similar to *image-to-image*, but using a mask to determine which parts of the image prior should be unchanged. The rest of the image is generated in a way that it is consistent with the fixed parts of the image prior (i.e., blended).

BlendScape uses *inpainting* to merge users' video backgrounds into blended environments and *image-to-image* techniques to transform existing images of environments to reflect the meeting purpose. To lower the barrier for end-users to generate good quality scenes, we developed multimodal interaction techniques to steer the generation of relevant visuals and composition techniques to naturally integrate users' videos within the scene.

We assessed the benefits and limitations of BlendScape's customization techniques in two steps. First, to demonstrate the expressiveness, we implemented three scenarios using BlendScape, exploring a range of professional and social collaborative activities. These scenarios incorporate a majority of environment design strategies for supporting distributed collaboration from our review of prior video-conferencing systems.

Second, we conducted an exploratory study with 15 end-users to investigate whether and how they would find value in using generative AI to customize video-conferencing environments. Through guiding participants to prototype meeting spaces for three scenarios using BlendScape, we elicited their customization preferences and explored to what extent BlendScape enabled them to achieve their design intentions. All participants could envision using generative AI techniques to facilitate a range of collaborative activities in the future (e.g., to spark creativity in professional settings or set a theme for social gatherings). However, to feel comfortable adapting environments during live meetings, they would require further controls to mitigate distracting or unrealistic visual elements. We propose improvements to BlendScape's implementation to address these limitations in future work.

Our key contributions are: *(1)* the BlendScape composition system, which enables real-time end-user customization of video-conferencing environments through generative AI-driven composition techniques; *(2)* an evaluation of BlendScape's expressiveness and considerations for empowering new design participants [32], through a study demonstrating how 15 video-conferencing users envision leveraging generative AI to personalize meeting environments and our implementation of three target use cases.

2 RELATED WORK

Our work extends prior research on *(1)* designing video-conferencing environments to support distributed collaboration; *(2)* generative AI techniques for constructing 2D and 3D environments.

[1] **Microsoft Teams TogetherMode:** https://www.microsoft.com/en-us/microsoft-teams/teams-together-mode; **Ohyay:** https://ohyay.co/

2.1 Design Strategies for Video-Conferencing Environments

Our goal with BlendScape was to provide a unified set of customization techniques for dynamically tailoring meeting environments to a wide range of collaborative activities. Here, we define the *environment* as the "stage" or background rendered around users' videos, excluding functional tools such as the chat or host controls.

To understand the design space of meeting environments, we reviewed commercial video-conferencing tools and systems from video-mediated communication research. We classified the environment design strategies these systems adopted to mitigate challenges in distributed collaboration, as demonstrated through empirical studies. Our review surfaced three main roles that meeting environments can play in supporting distributed collaboration (Fig. 2): *(1)* establishing a shared context, *(2)* enabling spatial metaphors for communication, *(3)* serving as a record or artifact of collaboration.

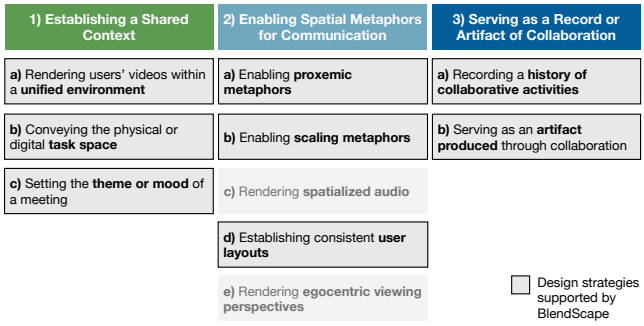

Figure 2: Classification of Environment Design Strategies: We analyzed how existing video-conferencing tools compose meeting spaces to support distributed collaboration by *(A)* depicting a shared context, *(B)* enhancing communication behaviors through spatial metaphors, *(C)* capturing a record of collaboration within the space. In Sec. 5, we use scenarios to demonstrate how BlendScape supports implementing eight of these ten design strategies (shown in bold).

We focus our review on screen-based meeting systems with 2D or 2.5D designs. Design strategies for other modalities (e.g., tabletop [57, 63] or mixed reality interfaces [25, 27, 51]) would likely introduce new dimensions to our categorization.

Establishing a Shared Context. A primary role of meeting environments is creating a shared frame of reference for distributed users. We observed three design strategies (Fig. 2A): displaying users within a unified meeting space, incorporating elements of physical or digital task spaces, and establishing a meeting theme.

Simulating co-located meetings by **rendering users' videos within a unified environment** is a popular strategy to increase distributed users' sense of co-presence [30, 56], as demonstrated both by commercial tools (e.g., Teams Together Mode, Ohyay[1]) and research systems (e.g., Waazam [30], HyperMirror [44], BISi [48]). **Conveying the task space**, i.e., the physical or digital space where artifacts are collaboratively produced [7], is critical for distributed users who would otherwise lack awareness of their collaborators' actions. Beyond traditional screen-sharing capabilities that display *digital task spaces*, MirrorBlender's [24] layerable "mirrors" (translucent representations of users' videos and shared screens) enable more natural interactions with digital artifacts, e.g., using hands to gesture around shared content. Capturing *physical task spaces* (e.g., for learning hands-on skills [37]) often requires custom hardware setups [34, 38]. To mitigate this, ThingShare [29] enables users to scan physical objects via webcams and manipulate their digital representations. To **set a common theme or mood for meeting participants**, Wazaam [30] and VideoPlay [20] introduce capture and rendering techniques for playful interactions, e.g., compositing children within storybook illustrations.

Enabling Spatial Metaphors for Communication. To enable more natural and seamless communication, recent tools leverage spatial affordances to mimic collaboration strategies from face-to-face interactions. Our review surfaced five spatial composition techniques (Fig. 2B): enabling conversational transitions through manipulating proximity or scale of users' videos, rendering spatial audio, structuring users' layouts to support turn-taking, and rendering egocentric viewing perspectives.

First, meeting rooms that resemble physical spaces afford using **proxemic interactions** to facilitate conversations: in Gather[2], users can initiate one-on-one video calls by "walking up" to each other; in OpenMic [28], users position themselves near a "Virtual Floor" to express their desire to speak. **Scaling metaphors** are commonly used to highlight specific users: Teams' and Zoom's Speaker views render active speakers at a larger size than other meeting participants; OpenMic [28] allows users to negotiate conversational transitions by increasing the size of their videos. **Rendering spatial audio** can help users follow conversation flows and support inclusion of remote participants in hybrid meetings [31, 45] (e.g., MirrorVerse's [23] "doorway" function gives users an auditory preview of breakout room discussions before they join). **Maintaining consistent user layouts**, (e.g., by seating users around a table in TogetherMode[1]) can help establish turn-taking patterns. Perspectives [56] further supports turn-taking by **rendering egocentric viewpoints** in the environment for each user, which simulates face-to-face social cues, e.g., making eye contact with speakers.

Serving as a Record or Artifact of Collaboration. Finally, we observed two examples where meeting environments document collaborative activities, providing a basis for later ideation (Fig. 2C). First, meeting tools can **record a history of user interactions, movements, and changes to the meeting environments**, which can be replayed at a later time to understand collaboration patterns (e.g., MirrorVerse's workspace record-and-replay tools [23]). Finally, the meeting environment could **be an artifact produced through users' collaboration**, e.g., for interior design or world-building scenarios. In our review of prior work, we did not find examples of video-conferencing tools that explicitly claim this functionality; however, this use case is common in VR immersive authoring tools [59, 64] which allow users to collaboratively create virtual content while situated in the virtual world itself.

While the HCI community has significant empirical knowledge on how to configure meeting environments to enhance distributed collaboration, it is still a challenge for end-users to attain these

[2]**Gather:** https://www.gather.town/

designs, due to a lack of real-time customization support in video-conferencing tools [23]. Commercial authoring tools (e.g., from Together Mode[1], Ohyay[1], and Gather[2]) are geared towards pre-meeting use, requiring users to manually craft relevant visuals and establish user layouts. Inspired by recent customization suites like MirrorVerse [23], our work aims to lower the barrier for users to dynamically create video-conferencing environments. We leverage generative AI techniques as a new approach for rapidly expressing and aligning environment designs with meeting contexts (e.g., to visualize shared task spaces or create themed visuals).

2.2 AI-Assisted Environment Generation

Our work also builds on prior approaches for enabling non-technical users to interact with generative AI models, which has been the focus of recent HCI research [1, 15, 16, 18, 42]. Based on user input, such as visuals or text-based prompts, these models can generate a wide variety of content including images [2, 43, 53, 61, 66], text [6, 50, 52], and even 3D objects [14, 22, 33, 35, 36].

Recent advances in image generation techniques allow the generation to be conditioned in a variety of useful ways, in addition to guidance from text-based prompts. For example, ControlNet uses an auxiliary model that incorporates additional data such as depth, semantic, and human pose representations [41, 65]. Prior work also explored intuitive techniques for end-users to control generative AI models, e.g., through sketching [12, 13, 65], speech [46], sliders [11], and iterative design mechanisms [15, 18, 55].

There is increased interest in harnessing image generation models to dynamically generate virtual environments. Before the rise of AI-driven approaches, environment generation relied on predefined components and procedural logic. For example, *WordsEye* converts text descriptions into 3D scenes using a database of 3D models and predefined rules [14]. *DreamWalker* substitutes real-world elements with virtual content via procedural generation, enabling VR users to safely navigate their physical surroundings [60].

These procedural approaches paved the way for the utilization of generative techniques that can be observed in recent AI-assisted creativity support tools. A recent example is *WorldSmith*, a system that investigates how multimodal image generation models can be harnessed to aid users in authoring and iteratively refining elements of fictional worlds [15]. *Opal* [42] guides users through a structured search for visual concepts to generate images for news illustrations, utilizing LLMs to tune users' prompts based on an article's content.

BlendScape builds on similar techniques, but with a distinct focus on generating creative environments for video-conferencing systems that can be aligned with meeting participants' goals. In particular, we employ *inpainting* methods to blend users' video backgrounds into unified scenes, provide multimodal input techniques for users to steer the generation, and leverage LLMs to dynamically tailor users' prompts to meeting themes and activities.

3 SCENARIO WALKTHROUGH

To illustrate how BlendScape can enable expressive meeting environments by blending physical and digital spaces, we present a scenario implemented with our system: a brainstorming session between two designers who are prototyping a mixed reality interface.

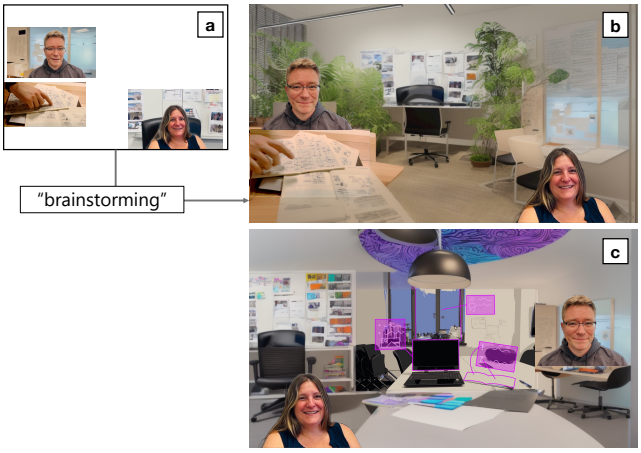

Figure 3: Scenario 1: *Design Brainstorming*. To create a unified setting for brainstorming, two designers use BlendScape to blend their webcam backgrounds with a camera feed of a physical desk *(a, b)*, enabling them to ideate around hand-drawn sketches. They later blend in elements of their digital task space, such as mock-ups of a mixed reality interface *(c)*.

The designers join a video call and add their webcam feeds to BlendScape's composition interface. One designer also adds a camera feed of their physical notebook, so they can sketch ideas during the brainstorming session (Fig. 3a). At the start of the meeting, they use BlendScape to generate a creative environment for their ideation activity. They type "`brainstorming`" in the prompt field, and a few seconds later, they see their *blended environment*: their physical surroundings are still visible, but now seamlessly extend into a unified design studio (Fig. 3b). This blended meeting space allows them to gesture and refer to physical sketches in the notebook, simulating how they might collaborate face-to-face. Later in the meeting, the designers blend a digital sketch into their physical backgrounds to preview mock-ups of their MR interface (Fig. 3c).

Many other meeting experiences are possible, from work scenarios to therapy spaces, birthday parties, or vacation planning (Appendix. A.2). As we describe next, BlendScape provides flexible techniques for creating rich meeting spaces. Later, we present additional scenarios that demonstrate BlendScape's expressiveness to enable distributed collaboration techniques from prior work (Sec. 5).

4 BLENDSCAPE SYSTEM

This section presents BlendScape, a rendering and composition system that enables meeting participants to customize video-conferencing environments. Key to our approach is blending elements of users' physical or virtual backgrounds into unified environments to allow for flexible representations of task spaces. We first outline three system requirements that guided the design of our system. Then, we describe our implementation of BlendScape's generative AI techniques for customizing meeting environments and composition techniques for enhancing visual cohesion.

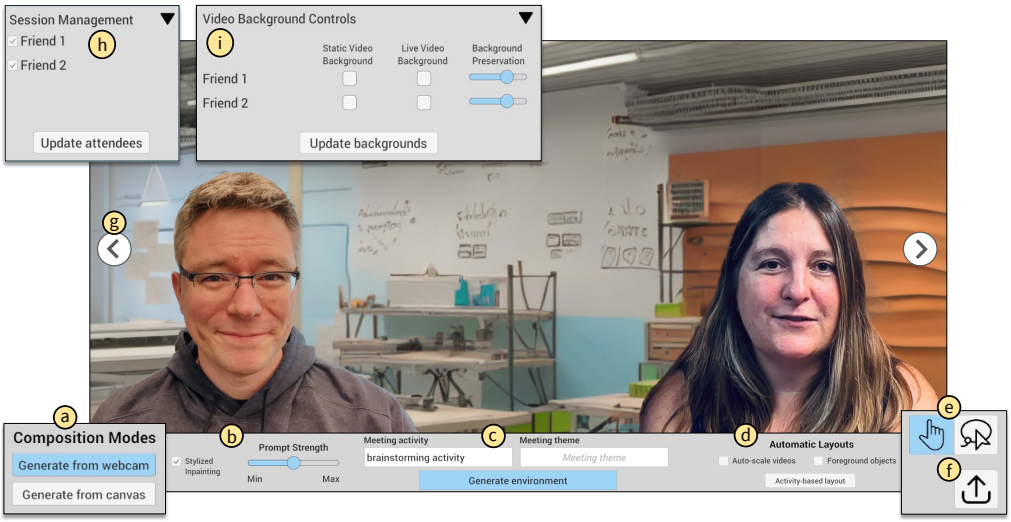

Figure 4: Overview of BlendScape interface: BlendScape offers two composition modes for creating meeting spaces *(a)*: blending webcam feeds together via *inpainting* and transforming the image on the canvas via *image-to-image*. To steer the environment generation, end-users can specify text-based prompts for the *Meeting Activity* and *Meeting Theme (c)*, control the strength of stylistic prompts *(b)*, upload custom image priors *(f)*, and modify specific regions of the scene via selection tools *(e)*. Users can return to and iterate on previous environment designs via the history tools *(g)*. The automatic layout techniques facilitate positioning users behind foreground objects in the scene *(d)*. BlendScape also provides session management tools *(h)* and per-user controls for adjusting the proportion of their video backgrounds preserved during the environment generation and toggling between displaying live webcam feeds or static frames *(i)*.

4.1 Requirements

We defined three requirements for customizing meeting environments with BlendScape, informed by our review of environment composition techniques for distributed collaboration (Sec. 2.1) and our motivating scenarios (Sec. 3, 5):

R1: Enabling users to express the meeting context through the environment. Embedding relevant visual and structural elements within video-conferencing environments can strengthen distributed users' collaborative processes (e.g., conveying shared task spaces in our *Design Brainstorming* scenario, using spatial landmarks to facilitate conversational transitions [28, 56]). BlendScape allows users to align environments to their meeting purpose by specifying text-based prompts and providing *image priors* (i.e., images of their physical surroundings or other spaces that represent their collaboration needs).

R2: Supporting convincing illusions of meeting in a shared space. To simulate face-to-face communication cues (e.g., deictic gestures [24], making eye-contact [56]), prior systems composite users within virtual environments that resemble physical spaces. To enable these designs while allowing users to incorporate their physical context, BlendScape implements two rendering techniques: *(1)* preserving and blending users' video backgrounds into a unified environment, which maintains realistic lighting, color temperature, and shadows around users; *(2) hidden surface removal* to obscure the boundaries of users' webcams among objects in the scene.

R3: Enabling coarse- and fine-grained customization of environments. Adapting meeting spaces to dynamic collaboration needs may require users to update the entire scene (e.g., to reconfigure the space for small vs. large group discussions [23]), as well as make minor adjustments (e.g., incorporating shared content [29]). In addition to *inpainting* and *image-to-image* techniques for creating new environments, BlendScape enables users to make granular changes by selecting portions of the scene to re-generate. To enable iteration on past results, we maintain a history of environments.

4.2 BlendScape Interface

Next, we provide an overview of BlendScape's composition tools (Fig. 4). BlendScape's user interface consists of a canvas in the center that displays blended environments and composites the users' videos within them. All meeting participants share the same view of this canvas; changes to the environment and position of users' video feeds are synchronized across all users.

Meeting participants can generate environments using **two composition modes** (Fig. 4a): blending the video feeds from only their webcams (using *inpainting*), or refining the image that is already present in the canvas (using *image-to-image* techniques). This image on the canvas may be a result of a previous image generation step or an image uploaded by the user (Fig. 4f). For *inpainting*, users can specify how much of their video backgrounds should be preserved through a slider (Fig. 4i).

To generate visuals relevant to the purpose of their meeting, users can enter two **prompts for a *Meeting Activity* and *Meeting Theme*** (Fig. 4c). The *Prompt Strength* slider controls the prompt weight, i.e., how much to prioritize the prompts in the environment

Figure 5: Environment Generation Techniques. BlendScape supports composing meeting spaces through *(a)* blending video feeds together via *inpainting* techniques and *(b)* transforming an input image (i.e., image prior) via *image-to-image* techniques. These composition approaches can be chained, e.g., to restyle a blended environment in the theme of a library *(c)*.

generation (Fig. 4b). BlendScape offers direct manipulation techniques (clicking, dragging, and pinching) as intuitive ways to position and scale videos, both for steering the environment generation and placing users' videos in the scene.

To make fine-grained adjustments to the environment, users can add or remove objects by selecting a region of the scene (Fig. 4e) and specifying a prompt. BlendScape also offers **automatic layout techniques** (Fig. 4d) that composite users behind objects in the scene and automatically scale their videos to match the size of the objects. BlendScape saves a **history of past environment generations and users' positions within the scene**, to enable iterating on past designs (Fig. 4g).

4.3 Generative AI Techniques for Blended Environments

To enable real-time end-user customization of video-conferencing environments, BlendScape implements two classes of AI image generation techniques: *(1)* an **inpainting technique**, which blends users' physical or virtual backgrounds into a unified environment, and *(2)* an **image-to-image technique**, which incorporates users into an overarching image that represents the meeting setting.

Generating blended environments from users' video backgrounds *(R1, R2).* BlendScape allows meeting participants to incorporate their real-world surroundings or virtual backgrounds into the shared environment as a mechanism for personalization [58] or capturing the task space where artifacts are collaboratively produced [7]. First, users can specify the proportion of their video backgrounds to retain; we mask these regions to preserve them in the blended environment (Fig. 6). Then, BlendScape performs *inpainting* to generate plausible visual details between the fixed regions, based on user-specified prompts for the *Meeting Activity* and *Meeting Theme*, e.g., "`brainstorming`" in a "`treehouse`"-themed environment (Fig. 5a). By preserving physical backgrounds, we aimed to more naturally integrate users into the blended space by matching their real-world conditions (e.g., lighting, shadows, color temperature, and webcam resolution).

Direct manipulation of video feeds (i.e., re-positioning and re-scaling) can be used to steer the generation of different styles of *inpainted* environments. For example, positioning smaller videos at the top of the screen creates a sense of environmental depth, with clear separation of foreground and background elements.

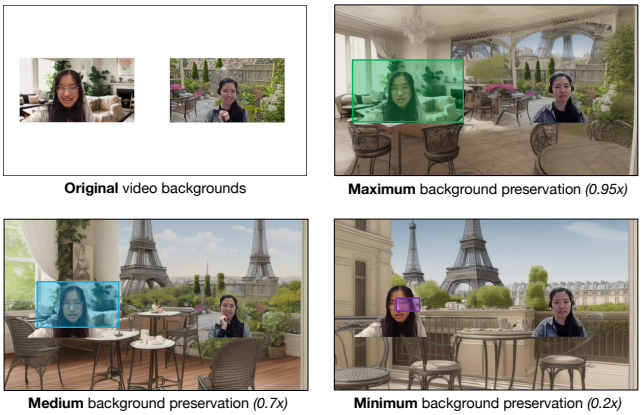

Figure 6: Masking Video Backgrounds: Users can adjust the proportion of their physical or virtual surroundings to retain in the resulting blended environments.

Driving environment generation through *image priors* *(R1).* As a second composition approach, BlendScape uses *image-to-image* techniques to transform an *image prior* (i.e., input image). This involves restyling the environment to incorporate visuals related to the *Meeting Activity* and *Meeting Theme* prompts, while preserving key structural elements in the prior, e.g., walls or background objects that convey the environment's geometry (Fig. 5b).

Meeting participants can upload image priors to set the theme of the meeting, (e.g., using an image of a library for a study group), establish a layout of users (e.g., row-based seating for academic lectures), or situate themselves in real-world meeting spaces.

Combining *inpainting* and *image-to-image* techniques for iterative environment design. *(R1, R3).* To simplify novice users' composition processes into two clear pathways for creating environments, we separated the *inpainting* and *image-to-image* techniques into two modes within BlendScape: webcam-based and canvas-based generation. However, BlendScape supports flexibly combining these techniques to iterate on prior environments. For example, performing an *image-to-image* transformation after an *inpainting* generation is an effective strategy for refining roughly-blended areas and subtly restyling an existing environment.

4.4 Improving the Composition of Blended Environments

To make the base environments generated via *inpainting* and *image-to-image* techniques more suitable for video-conferencing, we implemented three composition techniques: *(1)* **prompt enhancement** to further align environments to the meeting context and enhance the visual quality; *(2)* **granular scene-editing controls** to add visuals or correct distortions; *(3)* **hidden surface removal** techniques to integrate users' videos with foreground objects.

Generating context-informed and thematic environments through LLM-driven prompt padding *(R1, R2).* Crafting expressive prompts remains a challenge for novice users of image generation models, requiring significant trial and error [4, 12]. To lower the barrier for meeting participants to create detailed environments, BlendScape enhances user-specified prompts with keywords suggested by LLMs, as demonstrated by prior systems [4, 42].

First, we elicit text-based prompts from users to define the *Meeting Activity*, which drives the layout of the environment, and the *Meeting Theme*, which steers the generation of aesthetic elements. Then, to dynamically tailor the image generation prompts to the meeting context, we query GPT-3.5 to augment the *Meeting Activity* and *Meeting Theme* prompts with keywords for five relevant objects and five stylistic qualities that represent the meeting atmosphere. For example, for a *Meeting Activity* of "`Brainstorming Session`" and *Meeting Theme* of "`Hologram`," GPT-3.5 suggested "`Interactive Touchscreens`" and "`Holographic Whiteboards`" as objects and "`Dynamic Lighting`" and "`Seamless Integration of Virtual & Physical Elements`" as stylistic qualities.

Inspired by prompt expansion tools for artists (e.g., PromptGen[3]), BlendScape also adds a fixed set of terms to encourage high-quality visuals: "`highly detailed, intricate, sharp focus, smooth`." Appendix A.2 includes examples of GPT queries and outputs.

Compositing users in an immersive manner through *hidden surface removal* *(R2).* A limitation of capturing meeting participants via webcams is that they appear as "floating heads" with a harsh cut-off at their shoulders, due to the limited field-of-view. To more naturally integrate users within environments, BlendScape takes a similar approach as prior tools (e.g., Ohyay, TogetherMode[1]): placing users behind objects in the scene and enabling *hidden surface removal*, a rendering technique to remove portions of 3D objects that should not be visible from a particular camera perspective. We implemented an object segmentation pipeline in BlendScape to compute the salient objects in the generated environment (e.g., chairs, tables) and extract them to a foreground layer positioned in front of users' videos (Fig. 9). This creates the illusion of users sitting behind objects.

Making granular edits to the environment *(R3).* In addition to BlendScape's *inpainting* and *image-to-image* for updating the entire video-conferencing environment, we implemented a technique for making fine-grained revisions to the scene (Fig. 7). Using the selection tool, users can circle areas of the scene and specify a text-based prompt to remove content (e.g., to fix distorted areas) or add visuals relevant to the meeting scenario (e.g., chairs to accommodate new meeting participants). To regenerate the area, BlendScape uses the GLIGEN [41] inpainting model, which is trained to generate cohesive results by considering the prompt, position, and scale of the specified region in relation to its surroundings.

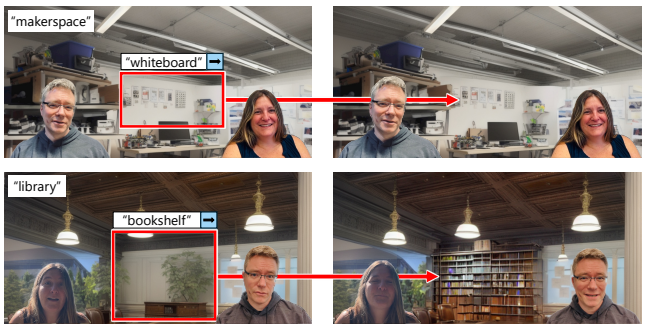

Figure 7: Granular Editing Tools: To add or remove content from the scene, users can outline a region and specify a text-based prompt.

4.5 Implementation

Figure 8 shows BlendScape's key components and system architecture. We implemented BlendScape's user interface as a Unity application integrated with Microsoft Teams. BlendScape receives video streams from individual meeting participants via Microsoft Teams' NDI streaming capabilities.

Image generation models: To enable the environment generation techniques, we used open-source Stable Diffusion [53] and ControlNet [65] models through the WebUI API[4], hosted on a PC with an AMD EPYC 7742 64-Core Processor and NVIDIA RTX A6000 GPU. For *inpainting*, we used the Realistic Vision 2.0[5] Stable Diffusion checkpoint which is fine-tuned for inpainting, further guided by a ControlNet inpainting model. For *image-to-image* generations, we used Realistic Vision's base checkpoint, guided by ControlNet Depth and Canny models to preserve the spatial layout and salient features of the image priors. Furthermore, we incorporated GLIGEN [41] to allow users to regenerate specific areas of the scene. Similar to ControlNet [65] in its goal of guiding diffusion models, GLIGEN makes region-specific edits based on textual prompts and bounding box coordinates.

[3]**PromptGen:** https://promptgen.vercel.app/

[4]**Stable Diffusion WebUI:** https://github.com/AUTOMATIC1111/stable-diffusion-webui, **ControlNet Extension:** https://github.com/Mikubill/sd-webui-controlnet

[5]**Realistic Vision 2.0**: https://huggingface.co/SG161222/Realistic_Vision_V2.0

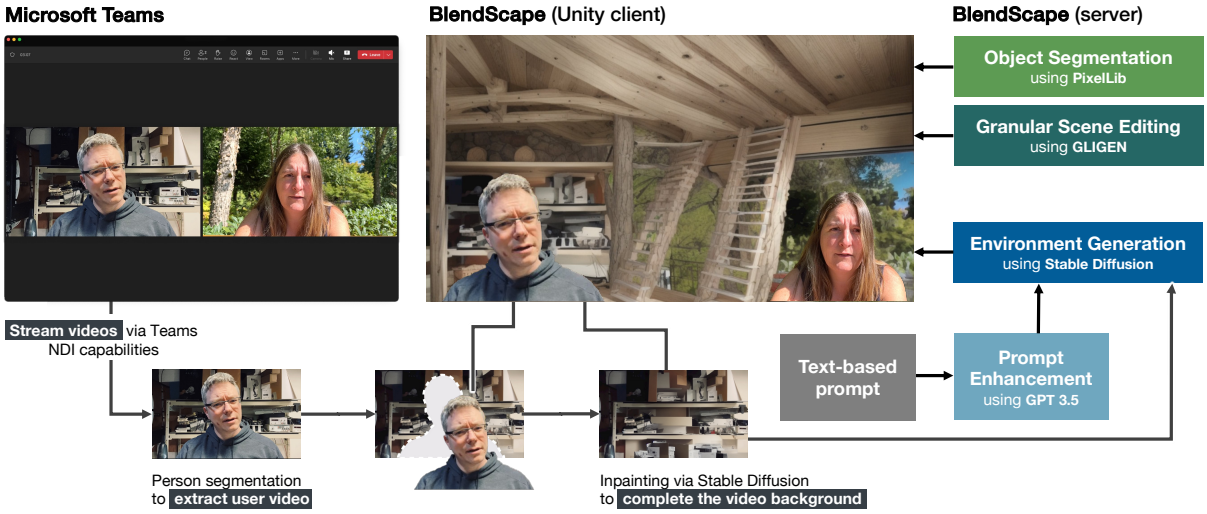

Figure 8: BlendScape Components and Architecture. The BlendScape system consists of a Unity client that serves as the main user interface. Via Microsoft Teams NDI capabilities, the Unity client receives and processes users' videos by performing person segmentation to separate the users from their video backgrounds and using inpainting techniques to fill the missing areas in the backgrounds. The Unity client connects to two servers that run *(1)* Stable Diffusion image generation processes to enable the environment generation techniques; *(2)* PixelLib for object segmentation to generate foreground objects, GPT-3.5 to enhance users' prompts with contextually-relevant keywords, and GLIGEN (grounded text-to-image generation model) for re-generating small portions of the environment.

Using these models, the average generation times were 10s for *inpainted* environments, 25s for *image-to-image* environments, and 20s for GLIGEN-enabled edits. This includes the time to enhance users' prompts with additional keywords (via GPT-3.5).

2.5D, layered scenes: The *image-to-image* generation mode transforms the entire existing environment, while the *inpainting* mode only takes users' webcam backgrounds as input. To isolate the correct scene elements for the image generation models, we implemented BlendScape as a 2.5D scene in Unity with five 2D layers staggered at different depths (Fig. 9): *(1)* foreground objects, *(2)* users' videos (separated from their video backgrounds), *(3)* users' video backgrounds, *(4)* background masks (to preserve regions of the video backgrounds), and *(5)* the generated environment. We instrumented the Unity scene with multiple cameras that render specific layers, enabling us to capture each layer separately. We use orthographic cameras (i.e., cameras that do not perform perspective rendering), so scene elements are rendered at a consistent scale even when placed at different depths.

Environment segmentation: BlendScape uses the PixelLib [49] semantic segmentation model to partition scene elements for hidden surface removal. We segment users from their video backgrounds via conventional computer vision techniques. While BlendScape displays live video feeds of the segmented users, we generate static environments using the first frame of users' videos, due to limitations with running Stable Diffusion at video rates.

Performing person segmentation leaves empty spaces in static video backgrounds, as webcams fail to capture areas of the background that are occluded by the user. We used Stable Diffusion to *inpaint* this area, simulating a continuous background (Fig. 8).

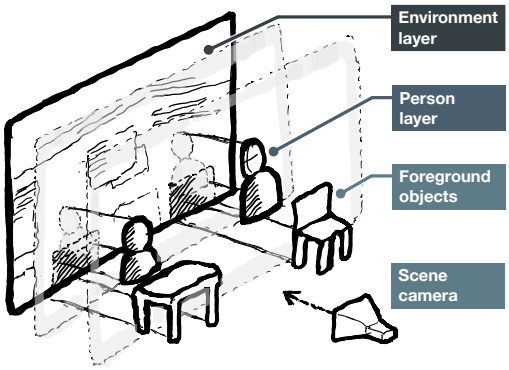

Figure 9: 2.5D, Layered Scenes. BlendScape separates the blended environment, users' videos, and foreground objects into layers and renders each layer with an orthographic camera. This enables flexibly combining the layers as input to image generation models to perform *inpainting* and *image-to-image* transformations.

5 DEMONSTRATION OF ADDITIONAL SCENARIOS

In this section, we illustrate the expressiveness of BlendScape's generative AI-enabled customization tools through prototyping meeting environments for three distributed collaboration scenarios: *Design Brainstorming*, *Remote Education*, and *Storytime with Family*.

Earlier, we reviewed the range of video-conferencing environments implemented by prior systems to enhance distributed collaborators' communication and sense of co-presence (Sec. 2.1). This review surfaced ten environment design strategies, such as incorporating representations of task spaces, enabling proxemic metaphors, or recording a history of collaborative activities. We demonstrate how BlendScape's composition techniques can be used to implement eight out of the ten design strategies (shown in bold in Fig. 2).

Our focus for BlendScape was enabling end-users to compose custom environment visuals; as such, rendering spatial sound was out of scope (Fig. 2B.iii). Currently, BlendScape's 2D image generation techniques do not support rendering egocentric viewing perspectives to simulate face-to-face communication cues (Fig. 2B.v), e.g., turning to face other users seated around a table [56]. This could be achieved via 360° panoramic or 3D environment generation approaches [40, 47].

5.1 Design Brainstorming

Recall our initial scenario from Sec. 3, where two interaction designers are remotely collaborating to prototype mixed reality interfaces (Fig. 3). They use BlendScape to **render their videos in a unified workspace** that **blends their physical and digital task spaces** into a single design studio. This allows them preview their mixed reality designs and simulate face-to-face design critiques, by verbally referring to and gesturing around hand-drawn sketches.

Implementation with BlendScape: We prototyped this scenario using four live video feeds: two webcam feeds of the designers, an external camera feed capturing a physical desk with notebooks, and a screen-capture of a tablet-based digital sketching application (Fig. 3a). We used BlendScape's *inpainting* mode to blend the users' videos and the task space feeds into seamless meeting environments (Fig. 3b, c). Finding the ideal placement of the task space feeds required some trial-and-error; for example, positioning the physical sketches at the bottom of the screen occasionally rendered them on the floor of the environment rather than on a desk.

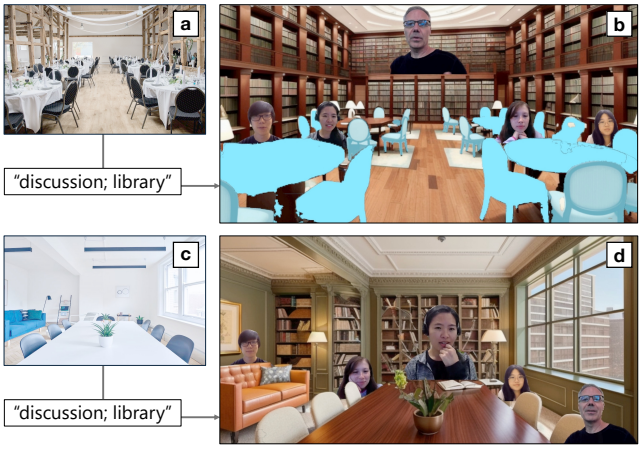

Figure 10: Scenario 2: *Remote Education*. To establish room layouts for seminar discussions, a professor restyles images with small and large tables *(a, b)* to resemble a library setting *(c, d)*. This enables students to organize and place themselves behind spatial landmarks in the scene (shown in blue).

5.2 Remote Education

Next, to demonstrate how BlendScape environments could support spatial metaphors to facilitate conversations [28], we prototyped a scenario involving a remote university course.

Scenario Description: A professor uses BlendScape to facilitate activities in their literature seminar. Before the seminar starts, they **establish the room layout** for small group discussions: they upload an image of a room with several tables and restyle it as a "`library`" (Fig. 10a, c). Students assign themselves to groups by organizing around the tables, leveraging **proxemic metaphors** (Fig. 10c). To structure a large group discussion later in the seminar, the professor restyles an image of a conference room (Fig. 10b, d). To spotlight speakers, they use **scaling metaphors** by rendering the students' videos larger and at the head of the table.

Implementation with BlendScape: We selected two images priors of event spaces that contained enough seating for 5 users and were captured from similar forward-facing perspectives as the users' webcam videos, allowing us to realistically place users behind furniture. We then used BlendScape's *image-to-image* mode with a *Meeting Theme* prompt of "`library`" to restyle the image priors.

5.3 Storytime with Family

Our final scenario is inspired by systems like Waazam [30] and VideoPlay [20], which mediate playful social interactions between children and family members.

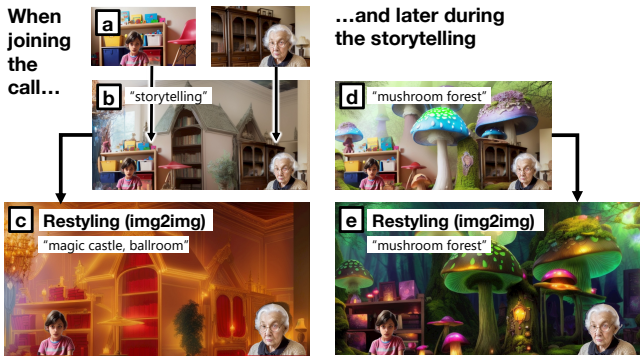

Figure 11: Scenario 3: *Storytime with Family*. A grandmother and her granddaughter use BlendScape to immerse themselves in a fairytale, using *image-to-image* techniques to restyle their blended video backgrounds into a ballroom and mushroom forest.

Scenario Description: A grandmother uses BlendScape to craft a memorable storytelling experience with her granddaughter. At the start of the call, a castle appears in the space between their video backgrounds with a prompt of "`storytelling`" (Fig. 11b). As the grandmother reads a fairy tale, she restyles the scenes to **set the theme of the meeting**; the environment transforms into magic ballroom and then a mushroom forest (Fig. 11c, e). These environments illustrating the story's progression serve as an **artifact produced through their collaboration**. A few weeks later, the granddaughter wants to write a sequel to the fairy tale. They use

BlendScape's saved environments and video positions to **replay a history of their collaboration** and extend their previous scenes.

Implementation with BlendScape: We used the *inpainting* mode to blend the users' physical environments together. We used *image-to-image* techniques with prompts of "`magic castle, ballroom`" and "`mushroom forest`" to restyle the blended environments to reflect the events of the story.

6 EXPLORATORY STUDY WITH END-USERS

To investigate whether and how end-users would find value in using generative AI to personalize video-conferencing environments, we conducted an exploratory study with 15 frequent users of traditional video-conferencing tools (e.g., Microsoft Teams, Zoom). Our goals were to *(1)* surface participants' preferences for customizing meeting environments, including visual or layout characteristics they aimed to achieve; *(2)* investigate to what extent BlendScape allowed participants to express their design intentions.

The study was approved under the Institutional Review Board (Ethics / IRB ID: #10764; Release and Compliance ID: #6755). We conducted 1-hour study sessions remotely via Microsoft Teams. Participants were compensated with $50 gift cards.

6.1 Participants & Recruitment

Advertising through internal mailing lists, we recruited individuals from our organization who use video conferencing tools at least three times per week. 15 individuals participated in our study (7 women, 8 men, majority in an age range of 25-34 years) and held a variety of job roles: UX or Product Designer (5), UX Researcher (5), Product Manager (3), Software Engineer (1), and Human Factors Engineer (1). 7 out of 15 participants used image generation models once a week or more frequently; the remaining 8 used them once a month or less frequently. 14 of 15 participants were located in North America and one was located in Asia.

6.2 Method

The study consisted of three environment composition tasks using BlendScape and a post-task discussion. In the first two tasks, we introduced participants to BlendScape's *inpainting* and *image-to-image* techniques, teaching them to generate environments based on pre-defined prompts and assess the quality in a semi-structured interview portion. This served as training for Task 3, where participants combined BlendScape's composition tools to create a series of environments for a research meeting scenario.

Set-up: We hosted BlendScape on a local machine and gave participants remote control via Microsoft Teams screen-sharing. For all tasks, we used pre-recorded videos to represent different users (simulated as NDI streams via the OBS Studio NDI Integration Tools[6]), in order to achieve relatively consistent environment generations across participants. We used a combination of real physical locations and virtual backgrounds to represent a diverse range of image priors. For Task 3, we integrated participants' webcam feeds into BlendScape so they could experience being immersed in different environments.

[6]**OBS Studio NDI Integration Tools:** https://obsproject.com/forum/resources/obs-ndi-newtek-ndi™-integration-into-obs-studio.528/

Task 1: Walkthrough & Comparison of Inpainting Techniques *(10 min)*. First, we explored blended environment designs for a *Vacation Planning* scenario, where two friends are planning a trip to Paris and generate meeting spaces to reflect landmarks they want to visit. The facilitator illustrated how even basic input (i.e., providing prompts, changing the position and scale of the image priors) can be used to steer the environment generation via BlendScape's *inpainting* techniques. Afterwards, we asked participants to experiment with different prompts and layouts.

We then showed participants three environments that preserved varying degrees of users' video backgrounds (Fig. 15). We asked participants to comment on which examples, if any, they could envision using for the *Vacation Planning* scenario with minimal changes. For the environments they could not envision using, we asked them to explain 1-2 key issues.

Task 2: Walkthrough & Comparison of Image-to-Image Techniques *(10 min)*. Next, we introduced BlendScape's *image-to-image* generation techniques using a *Game Stream* scenario, where a Minecraft player uses BlendScape to engage viewers their viewers. We started from an image of an arcade and used prompts to restyle the image prior in a Minecraft theme. Then, we combined *inpainting* and *image-to-image* techniques to blend the users' webcam backgrounds into the arcade image. Similarly to the first task, we asked participants to compare four example environments (Fig. 16), comment on which options they could envision using for the scenario, and explain key issues they observed. We intentionally included examples of roughly blended and cluttered environments to provoke discussions around the limitations of BlendScape's composition techniques.

Task 3: Designing Environments for a Progressive Meeting Scenario *(15 min)*. To explore how participants could envision adapting environment designs during a live meeting, we instructed them to compose a series of scenes for a scenario involving writing a research paper. First, we introduced a student character (using a pre-recorded video) and brought the participants' webcam feeds into BlendScape, to play the role of another student.

We facilitated the task via the following prompts: *(1)* Design an initial environment for two students to discuss the introduction of the paper; *(2)* The professor joins the call to provide feedback on the students' ideas. How would you redesign the environment to include them? *(3)* All three users are starting to feel stressed with the paper deadline approaching. How would you redesign the environment to support them?

Throughout the task, we asked the participants to think aloud to describe their design goals and assess the quality of the generated environments. We prompted them to use features that we had not yet explored (e.g., adding or removing content).

Discussion *(15 min)*. We ended with a semi-structured interview around the potential value that generative AI-enabled custom environments could provide to distributed collaboration scenarios. We asked participants to comment on specific scenarios where they could or could not envision using a system like BlendScape to personalize meeting environments. Participants also reflected on unexpected or surprising elements of the environments they generated in the previous tasks and how these aspects might support or detract from collaborative processes.

Figure 12: Participants' environment designs for the Research Paper Scenario. To support the students' and professor's writing process in our Task 3 scenario, participants adopted a variety of design strategies: blending their backgrounds into spaces that support creativity (e.g., P6's coffee shop and P8's office space), creating themed environments according to the research topic (e.g., P2's chemistry lab), and removing content to create distraction-free spaces (P13). The participants envisioned both playful and calming environments to "de-stress" the users as they approached their paper deadline, e.g., P4's rainforest with hammocks, P12's stuffed animal-themed room, and P2 & P15's indoor spaces with plants and natural lighting.

6.3 Data Collection & Analysis

For all study sessions, we collected audio transcripts & recordings, screen recordings, and images of the environments generated with BlendScape for later analysis. To analyze the environment characteristics that participants aimed to achieve, we used a thematic analysis approach [5]. One author reviewed all transcripts to create an initial codebook, specifically paying attention to participants' rationale for selecting an environment over another in the Task 1 & 2 comparison exercises and their revision strategies in Task 3. Two authors then reviewed the codes in the context of specific examples provided by the participants (including both quotes and images of the environments) to extract higher-level themes. We used an affinity diagramming approach [54] to analyze and aggregate themes from the post-task discussions around the benefits and limitations of BlendScape and future usage scenarios.

7 RESULTS

Figure 12 shows a sample of the environments generated by the 15 participants in our study. In this section, we first present four themes around participants' environment customization preferences (i.e., the visual and layout characteristics they aimed to achieve). We then discuss the benefits of BlendScape's composition techniques for expressive leverage and limitations with the time and effort required to achieve optimal designs.

7.1 Environment Customization Preferences

When assessing the quality of BlendScape's environments, participants expressed preferences for *(1)* authentic over artificial spaces; *(2)* both strong and subtle thematic elements, depending on the meeting context; *(3)* structuring collaboration through spatial layouts of users; *(4)* balancing the spatial-richness of environments.

Authentic environments were preferred over artificial, but came with higher expectations for realism. A majority of participants expressed that blending users' physical surroundings into a unified meeting space could promote co-presence while maintaining familiar aspects of their individual environments (P2, P5-6, P9, P11-14). P13 commented that "taking something personal to [them] and tweaking it" would help remote collaborators to "trick [their] minds to believe that [they're] more in the same place."

P6 cited another benefit of authentic backgrounds in creating a higher degree of realism, as users' videos appeared to be naturally framed with appropriate furniture and lighting. In environments constructed from virtual backgrounds, users sometimes appeared to be "floating in space in front of an image... like a bad Photoshop job" (P6). However, we observed a drawback to this increased realism: participants more easily identified and were more critical of flaws in how their own physical surroundings were blended (P1, P3-5, P7-8, P10), e.g., warped areas or inconsistent room geometry (Fig. 13). With fully artificial environments, some participants argued that "there's no pressure to do it well" (P8); "even if it was not as polished" they would have a "higher tolerance" for mistakes (P3).

Varied preferences for strong vs. subtle visual ties to the meeting context. A few participants appreciated having stylistic elements that closely reflect the *Meeting Activity* and *Theme* prompts they provided, e.g., a chemistry lab for the *Research Paper* scenario in Task 3. These participants embraced the at-times unrealistic visuals, arguing "that we don't need the space to look like traditional meeting spaces" (P7) which are "fixed and static... not really imaginative" (P15). However, most participants preferred simple and subtle theming to avoid distractions (P1, P7, P9, P13-15). As P1 stated, environments should "add texture to the call without pulling away from it."

Establishing user layouts through the environment was a popular design strategy, but required manual staging. Many participants placed users around spatial landmarks in the scene to reflect specific collaborative activities (P1, P3-7, P9, P13-15). For example, in the *Research Paper* scenario, P14 positioned the professor in the lower left corner "to help structure the collaboration," similar to a picture-in-picture style for online lectures. P9 placed users in chairs that were close together to "give [them] the most privacy" while having a one-on-one discussion.

However, some participants observed that BlendScape did not generate spaces to accommodate the existing layout of users in the scene. P1 and P6 had to manually reposition users to seat them in chairs, but they expected BlendScape to automatically generate chairs and tables to frame the users. This points to a limitation of how BlendScape blends environments after segmenting users from their video backgrounds, which we further discuss in Sec. 8.2.

Need to balance the spatial richness of environments with 2D representations of users. Participants found that spatial properties of BlendScape's environments (e.g., furniture layouts and lighting aligned with the geometry of the space) increased realism and lent themselves to structuring user layouts for collaboration (as discussed in the previous theme). However, many participants observed a juxtaposition between these spatially-rich environments and the strictly 2D representation of users (P5-6, P8-10, P11, P13-14). A few participants tried to tilt the users' videos to "echo the [3D] perspective in the room" (P14) and make it look like users were facing each other (P9, P13). To make our hidden surface removal technique (i.e., hiding floating heads behind foreground objects) more convincing, P14 suggested to "accept the flatness" and use environments where all furniture and users' videos are facing forwards, in the style of Teams TogetherMode[1].

7.2 Benefits and Limitations of Composition Techniques

We discuss two additional themes around participants' perceptions of the benefits and challenges with using BlendScape to compose meeting environments.

BlendScape provides expressive controls for creating environments that reflect the meeting context. Participants appreciated the ability to rapidly generate and iterate on environment designs with BlendScape, creating over 300 scenes across all three study tasks. As P13 expressed, they exerted "such minimal effort" to specify their intentions to BlendScape, and the system "does so much" to translate their prompts into rich environments.

Participants' design strategies utilized the full range of our multimodal techniques to steer image generation. To prototype environments for the *Research Paper* scenario, they added objects that represent the users' research topic (P2-3), removed objects to promote distraction-free collaboration (P10, P13, P14), and prompted for environments they associate with creativity, e.g., parks and coffee shops (P1, P5-9, P11-12, P15). To help "de-stress" the students as their deadline approached, a popular strategy was transporting the users to relaxing or playful locations, e.g., a beach, a rainforest, and a stuffed animal-themed space (Fig. 12). Participants also used the *image-to-image* mode to subtly restyle the existing scene (e.g., using prompts of "`warm`" and "`relaxing`").

Challenges with distracting scene elements, time & effort required to achieve the ideal design. To enable using a system like BlendScape for professional contexts, participants expressed a need to prevent unexpected elements "that could prove to be more distracting than helpful" (P6). These ranged from minor mistakes that "look good when you first glance at it" (P5) but become more apparent upon closer inspection (e.g., two Eiffel Towers in Fig. 15), to more significant cases where the generative AI models misunderstood participants' intent (e.g., GLIGEN inserting a playground slide rather than presentation slides for P9).

While each environment generation takes minimal effort, fixing distracting elements required several iterations (P1, P3-5, P7, P10, P12, P15). This impacted when and for which scenarios participants envisioned using generative AI to personalize meeting environments. For professional scenarios, some participants would prefer to customize meetings spaces before, rather than during, live calls (P1-3, P5), as the "the uncertainty of the visualizations might detract" from work discussions (P11). Scenarios with time pressure may also call for pre-meeting customizations: P12 expressed that their "therapist charges by the minute, so I don't want to spend time doing this during a session."

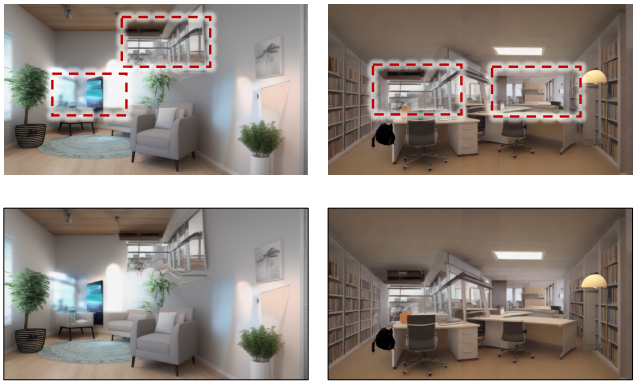

Figure 13: Imperfections in Blended Environments: BlendScape's inpainting techniques can sometimes produce warped or unrealistic room geometries when users' video backgrounds are very disparate (P7) or are captured from different perspectives (P5).

8 DISCUSSION

The scenarios we prototyped with BlendScape (Sec. 5) and our exploratory user study (Sec. 6) demonstrated our system's ability to enable a wide range of meeting environment designs. Overall, all participants expressed they could envision using AI image generation techniques to customize video-conferencing environments in the future, given further controls to prevent unrealistic visuals.

However, further research is required to operationalize similar customization techniques in live video-conferencing systems and understand how to effectively utilize them to achieve different

collaboration goals. We discuss two avenues extending our work with BlendScape: *(1)* studying the new affordances for distributed collaboration that generative AI-enabled environment composition techniques could enable; *(2)* technical improvements to address limitations of current-day image generation models.

8.1 New Opportunities to Facilitate Collaboration through Generative AI-enabled Environment Customization

While using our classification of prior work's environment design strategies to guide the implementation of BlendScape (Sec. 2.1), we observed new ways for these design strategies to manifest when using generative AI. For example, BlendScape creates literal representations of spatial landmarks (e.g., chairs to establish user layouts, walls to designate boundaries) as opposed to abstract proxemic metaphors in other tools (e.g., the Virtual Floor in OpenMic [28]). Additionally, BlendScape enables ambient representations of task spaces (e.g,. rendering physical sketches in the corner of the screen in our *Design Brainstorming* scenario), as opposed to traditional screen-sharing capabilities which fully direct users' attention towards shared content. In future work, we would find it interesting to deploy BlendScape in live meeting scenarios to study the impact of these different environment designs on users' collaboration processes, as compared to manually-crafted spaces from prior systems.

We also envision extensions to BlendScape's implementation to enable asymmetric environments and system-driven adaptations. Per-user, rather than global, customization controls could allow users to tailor meeting spaces to their personal needs (e.g., some users may prefer distraction-free spaces, while others may focus better with visual stimuli in the environment). However, these asymmetric environments should be carefully designed to maintain consistency across users when required for the collaboration scenario (e.g., preserving user layouts for turn-taking). Incorporating real-time summarization of meeting transcripts [9] could enable making the environment a more active partner in collaboration, e.g., automatically adjusting user layouts to transition to new activities or foreshadowing upcoming topics with related visuals.

8.2 Improving Upon BlendScape's Composition Techniques

We developed BlendScape to achieve a balance between usability requirements for video-conferencing and technical constraints of current image generation models. At the time of research, it was infeasible to run image generation models at video rates; as such, we generate environments from a single frame of users' video backgrounds. However, our approach still enables users to rapidly update environments (via text-based prompts and direct manipulation of video feeds) and render dynamic content by blending in external camera feeds (as demonstrated in our *Design Brainstorming* scenario, which incorporates a live feed of a physical workspace).

While we expect model performance and output quality to improve in the future, we propose extensions to BlendScape to address two key limitations that were surfaced in our study.

***(1)* BlendScape can produce environments that ignore the number and existing placement of users in the scene.** Currently, BlendScape segments and removes users from their video backgrounds before sending the backgrounds as input to Stable Diffusion (Fig. 8). This was an intentional choice to avoid generating unrealistic depictions of humans, which is a known issue with image generation models. However, this can result in environments that do not reflect the existing spatial layout of users in the scene (e.g., not containing enough furniture to seat all users).

To achieve more "people-informed" compositions, we envision using OpenPose models[7], which perform human pose estimation, to generate furniture layouts aligned with users in the scene. This approach would likely require multiple passes: first, an *inpainting* step to generate the environment, using input images with users present to support the OpenPose model. A second *inpainting* pass could detect and correct any distorted representations of people.

***(2)* The environment generation techniques can be overwhelmed when users' backgrounds and text-based prompts are in contrast.** Our study participants noticed that when users' "backgrounds are so disparate, BlendScape really had trouble integrating them into a believable environment" (P13). We observed this in particular with Task 2 (Fig. 16), where the four video streams had semantically different backgrounds (e.g., medieval castle vs. office space), with little relation to the prompt of *Game Streaming*.

To achieve more cohesive blends, one solution is to mask a larger proportion of users' video backgrounds that are relevant to the meeting prompts, thereby prioritizing these backgrounds in the resulting environment. This relevance metric could be computed by comparing the similarity between the text-based prompts and semantic descriptions of the video backgrounds.

9 LIMITATIONS

Lack of baseline comparison: Given that BlendScape introduces a new way of composing video-conferencing environments through real-time generative AI techniques, we first sought to understand the potential value that the system could provide to users through an exploratory study. The meeting scenarios we prototyped (Sec. 5) provide an initial comparison of BlendScape to existing customization suites (e.g., Together Mode, Ohyay[1]), as we demonstrated that BlendScape can be used to implement a majority of design strategies offered by these tools. While conducting a controlled baseline comparison was out of scope for our work, it would be a promising avenue for future research to understand requirements for manual vs. automated authoring workflows.

Generalizability of results to live meeting scenarios: Some aspects of our study design intended to ensure consistency across participants and support their agency during the study may limit the generalizability of our results to real meeting scenarios. In Tasks 1 & 2, we used pre-recorded videos with static backgrounds to generate similar environments for all participants, enabling us to compare their feedback and extract themes. However, dynamic video backgrounds could have elicited different design strategies from our participants (e.g., incorporating physical context changes to give collaborators awareness of each others' background activities). In Task 3, most participants chose to use virtual rather than real-world backgrounds, due to their physical surroundings being uninteresting or to preserve privacy. We find it encouraging that these participants still brainstormed a variety of scenarios where

[7] **ControlNet OpenPose:** https://huggingface.co/lllyasviel/sd-controlnet-openpose

incorporating authentic environments could provide value, e.g., to connect family members or encourage co-workers to "share their lives in different ways" (P9). Finally, we simulated multi-user scenarios via pre-recorded videos and prompted participants to envision how generative AI techniques could be used to support collaboration. Future studies with multiple participants should be conducted to surface design strategies and challenges specific to collaborative environment composition.

Study sample and novelty effects: Our participants' insights may not generalize to all future users of systems like BlendScape, as we primarily studied with individuals from a UX design and product management background. However, our participants' diverse experience using image generation models suggests that BlendScape's composition techniques are still accessible to novice users.

Considering the novelty of generative AI techniques, our studies around BlendScape are subject to participant response bias [17]. We reduced potential bias by discussing both the benefits and limitations of BlendScape with participants and probing them with examples of poorly designed environments (e.g., in Fig. 16) to elicit critiques and improvements to our system.

10 CONCLUSION

We presented BlendScape, a rendering and composition system for video-conferencing participants to tailor meeting environments to their collaboration contexts, by leveraging AI image generation techniques. Through implementing scenarios and conducting evaluations with 15 end-users, we demonstrated a rich set of meeting spaces that BlendScape can enable. Participants' feedback was encouraging in that they could rapidly express their design intentions with BlendScape and envisioned using similar techniques to facilitate meetings in the future. Based on their feedback, we proposed improvements to BlendScape to mitigate challenges with incohesive or distracting scene elements. Future studies around BlendScape could explore the impact of its customization techniques on users' collaborative processes and system-driven approaches to adapting environments (e.g., facilitating activity transitions with visuals representing future topics).

ACKNOWLEDGMENTS

We thank Amit Gulati & Henrik Turbell for their valuable feedback, Sasa Junuzovic & Pat Sweeney for their advice on working with NDI streams, and Michael Nebeling & Janet Johnson for their support throughout the project. We thank our Microsoft Research colleagues and fellow interns who participated in our video and figures. Finally, we express our gratitude to our study participants for their time.

REFERENCES

[1] Aseem Agarwala, Mira Dontcheva, Maneesh Agrawala, Steven Drucker, Alex Colburn, Brian Curless, David Salesin, and Michael Cohen. 2004. Interactive digital photomontage. *ACM Transactions on Graphics* 23, 3 (Aug. 2004), 294–302. https://doi.org/10.1145/1015706.1015718

[2] Shai Avidan and Ariel Shamir. 2007. Seam Carving for Content-Aware Image Resizing. In *ACM SIGGRAPH 2007 Papers (SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 10–es. https://doi.org/10.1145/1275808.1276390

[3] Gabrielle Benabdallah, Sam Bourgault, Nadya Peek, and Jennifer Jacobs. 2021. Remote Learners, Home Makers: How Digital Fabrication Was Taught Online During a Pandemic. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 350:1–350:14. https://doi.org/10.1145/3411764.3445450

[4] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco , CA , USA. https://doi.org/10.48550/arXiv.2304.09337 arXiv:2304.09337 [cs]

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901.

[7] Bill Buxton. 2009. *Mediaspace – Meaningspace – Meetingspace*. Springer London, London, 217–231. https://doi.org/10.1007/978-1-84882-483-6_13

[8] Carrie J. Cai, Michelle Carney, Nida Zada, and Michael Terry. 2021. Breakdowns and Breakthroughs: Observing Musicians' Responses to the COVID-19 Pandemic. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 571:1–571:13. https://doi.org/10.1145/3411764.3445192

[9] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. MeetScript: Designing Transcript-based Interactions to Support Active Participation in Group Video Meetings. *Proceedings of the ACM on Human-Computer Interaction* abs/2309.12115 (2023). https://doi.org/10.48550/ARXIV.2309.12115 arXiv:2309.12115

[10] Jaz Hee-jeong Choi and Cade Diehm. 2021. Aesthetic flattening. *Interactions* 28, 4 (2021), 21–23. https://doi.org/10.1145/3468080

[11] John Joon Young Chung and Eytan Adar. 2023. Artinter: AI-powered Boundary Objects for Commissioning Visual Arts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1997–2018. https://doi.org/10.1145/3563657.3595961

[12] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. https://doi.org/10.1145/3586183.3606777 arXiv:2308.05184 [cs]

[13] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3501819

[14] Bob Coyne and Richard Sproat. 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, 487–496. https://doi.org/10.1145/383259.383316

[15] Hai Dang, Frederik Brudy, George Fitzmaurice, and Fraser Anderson. 2023. WorldSmith: Iterative and Expressive Prompting for World Building with a Generative AI. arXiv:2308.13355 [cs]

[16] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images Using Multiple Sliders with and without Feedforward Information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3502141

[17] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is better!": participant response bias in HCI. In *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, Joseph A. Konstan, Ed H. Chi, and Kristina Höök (Eds.). ACM, 1321–1330. https://doi.org/10.1145/2207676.2208589

[18] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, R Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham Taylor. 2019. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 10303–10311. https://doi.org/10.1109/ICCV.2019.01040

[19] G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. 2021. Zoom Exhaustion & Fatigue Scale. *Computers in Human Behavior Reports* 4 (2021), 100119. https://doi.org/10.1016/j.chbr.2021.100119

[20] Sean Follmer, Hayes Raffle, Janet Go, Rafael Ballagas, and Hiroshi Ishii. 2010. Video Play: Playful Interactions in Video Conferencing for Long-Distance Families with Young Children. In *Proceedings of the 9th International Conference on Interaction Design and Children*. ACM, Barcelona Spain, 49–58. https://doi.org/10.1145/1810543.1810550

[21] Verena Fuchsberger, Janne Mascha Beuthel, Philippe Bentegeac, and Manfred Tscheligi. 2021. Grandparents and Grandchildren Meeting Online: The Role

of Material Things in Remote Settings. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 478:1–478:14. https://doi.org/10.1145/3411764.3445191

[22] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. 2020. Mesh R-CNN. arXiv:1906.02739 [cs]

[23] Jens Emil Grønbæk, Wendy E Mackay, Marcel Borowski, Michel Beaudouin-Lafon, Eve Hoggan, and Clemens N Klokmose. 2023. Mirrorverse: Live Tailoring of Video Conferencing Interfaces. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco , CA , USA. https://doi.org/10.1145/3586183.3606767

[24] Jens Emil Grønbæk, Banu Saatçi, Carla F. Griggio, and Clemens Nylandsted Klokmose. 2021. MirrorBlender: Supporting Hybrid Meetings with a Malleable Video-Conferencing System. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 451:1–451:13. https://doi.org/10.1145/3411764.3445698

[25] Jens Emil Sloth Grønbæk, Ken Pfeuffer, Eduardo Velloso, Morten Astrup, Melanie Isabel Sønderkær Pedersen, Martin Kjær, Germán Leiva, and Hans Gellersen. 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 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 456:1–456:16. https://doi.org/10.1145/3544548.3581515

[26] Dongqi Han, Denise Y. Geiskkovitch, Ye Yuan, Chelsea Mills, Ce Zhong, Amy Yo Sue Chen, Wolfgang Stuerzlinger, and Carman Neustaedter. 2023. Dr.'s Eye: The Design and Evaluation of a Video Conferencing System to Support Doctor Appointments in Home Settings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 343:1–343:18. https://doi.org/10.1145/3544548.3581350

[27] Jaylin Herskovitz, Yifei Cheng, Anhong Guo, Alanson P. Sample, and Michael Nebeling. 2022. XSpace: An Augmented Reality Toolkit for Enabling Spatially-Aware Distributed Collaboration. *Proc. ACM Hum. Comput. Interact.* 6, ISS (2022), 277–302. https://doi.org/10.1145/3567721

[28] Erzhen Hu, Jens Emil Sloth Grønbæk, Austin Houck, and Seongkook Heo. 2023. OpenMic: Utilizing Proxemic Metaphors for Conversational Floor Transitions in Multiparty Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 793:1–793:17. https://doi.org/10.1145/3544548.3581013

[29] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. 2023. ThingShare: Ad-Hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 365:1–365:22. https://doi.org/10.1145/3544548.3581148

[30] Seth E. Hunter, Pattie Maes, Anthony Tang, Kori M. Inkpen, and Susan M. Hessey. 2014. WaaZam!: supporting creative play at a distance in customized video environments. In *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014*, Matt Jones, Philippe A. Palanque, Albrecht Schmidt, and Tovi Grossman (Eds.). ACM, 1197–1206. https://doi.org/10.1145/2556288.2557382

[31] Jeremy Hyrkas, Andrew D. Wilson, John Tang, Hannes Gamper, Hong Sodoma, Lev Tankelevitch, Kori Inkpen, Shreya Chappidi, and Brennan Jones. 2023. Spatialized Audio and Hybrid Video Conferencing: Where Should Voices be Positioned for People in the Room and Remote Headset Users?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 794:1–794:14. https://doi.org/10.1145/3544548.3581085

[32] Dan R. Olsen Jr. 2007. Evaluating user interface systems research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, Newport, Rhode Island, USA, October 7-10, 2007*. ACM, 251–258. https://doi.org/10.1145/1294211.1294256

[33] Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. https://doi.org/10.48550/arXiv.2305.02463 arXiv:2305.02463 [cs]

[34] Sasa Junuzovic, Kori Inkpen, Tom Blank, and Anoop Gupta. 2012. IllumiShare: sharing any surface. In *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, Joseph A. Konstan, Ed H. Chi, and Kristina Höök (Eds.). ACM, 1919–1928. https://doi.org/10.1145/2207676.2208333

[35] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J. Mitra. 2023. HOLODIFFUSION: Training a 3D Diffusion Model Using 2D Images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 18423–18433. https://doi.org/10.1109/CVPR52729.2023.01767

[36] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2022. Simple and Effective Synthesis of Indoor 3D Scenes. https://doi.org/10.48550/arXiv.2204.02960 arXiv:2204.02960 [cs]

[37] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George W. Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 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 2019, New Orleans, LA, USA, October 20-23, 2019*, François Guimbretière, Michael S. Bernstein, and Katharina Reinecke (Eds.). ACM, 161–174. https://doi.org/10.1145/3332165.3347872

[38] Audrey Labrie, Terrance Tin Hoi Mok, Anthony Tang, Michelle Lui, Lora Oehlberg, and Lev Poretski. 2022. Toward Video-Conferencing Tools for Hands-On Activities in Online Teaching. *Proc. ACM Hum. Comput. Interact.* 6, GROUP (2022), 10:1–10:22. https://doi.org/10.1145/3492829

[39] Minha Lee, Wonyoung Park, Sunok Lee, and Sangsu Lee. 2022. Distracting Moments in Videoconferencing: A Look Back at the Pandemic Period. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 141:1–141:21. https://doi.org/10.1145/3491102.3517545

[40] Jialu Li and Mohit Bansal. 2023. PanoGen: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation. *arxiv* (2023).

[41] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[42] Vivian Liu, Han Qiao, and Lydia B. Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*. ACM, 73:1–73:17. https://doi.org/10.1145/3526113.3545621

[43] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. https://doi.org/10.48550/arXiv.1411.1784 arXiv:1411.1784 [cs, stat]

[44] Osamu Morikawa and Takanori Maesako. 1998. HyperMirror: Toward Pleasant-to-Use Video Mediated Communication System. In *CSCW '98, Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work, Seattle, WA, USA, November 14-18, 1998*, Steven E. Poltrock and Jonathan Grudin (Eds.). ACM, 149–158. https://doi.org/10.1145/289444.289489

[45] Qianqian Mu, Marcel Borowski, Jens Emil Sloth Grønbæk, Susanne Bødker, and Eve E. Hoggan. 2024. Whispering Through Walls: Towards Inclusive Backchannel Communication in Hybrid Meetings. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*. ACM, 1032:1–1032:16. https://doi.org/10.1145/3613904.3642419

[46] Nels Numan, Daniele Giunchi, Benjamin Congdon, and Anthony Steed. 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)*. IEEE, Shanghai, China, 497–501. https://doi.org/10.1109/VRW58643.2023.00108

[47] Nels Numan, Shwetha Rajaram, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D. Wilson. 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 2024, Pittsburgh, PA, USA, October 13-16, 2024*. ACM. https://doi.org/10.1145/3654777.3676361

[48] Kenton O'Hara, Jesper Kjeldskov, and Jeni Paay. 2011. Blended interaction spaces for distributed team collaboration. *ACM Trans. Comput. Hum. Interact.* 18, 1 (2011), 3:1–3:28. https://doi.org/10.1145/1959022.1959025

[49] Ayoola Olafenwa. 2021. Simplifying Object Segmentation with PixelLib Library. (Jan. 2021).

[50] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs]

[51] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew D. Wilson. 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 2016, San Francisco, CA, USA, February 27 - March 2, 2016*, Darren Gergle, Meredith Ringel Morris, Pernille Bjørn, and Joseph A. Konstan (Eds.). ACM, 1714–1723. https://doi.org/10.1145/2818048.2819965

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

IEEE, New Orleans, LA, USA, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[54] Raymond Scupin. 1997. The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. *Human Organization* 56, 2 (1997), 233–237.

[55] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. ChatPainter: Improving Text to Image Generation Using Dialogue. https://doi.org/10.48550/arXiv.1802.08216 arXiv:1802.08216 [cs]

[56] John Tang, Kori Inkpen, Sasa Junuzovic, Keri Mallari, Sean Rintel, Andrew Wilson, Shiraz Cupala, Tony Carbary, Abigail Sellen, and William Buxton. 2023. Perspectives: Creating Inclusive and Equitable Hybrid Meeting Experiences. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023).

[57] Philip Tuddenham and Peter Robinson. 2009. Territorial coordination and workspace awareness in remote tabletop collaboration. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*. ACM, 2139–2148. https://doi.org/10.1145/1518701.1519026

[58] Gina Venolia, John C. Tang, Kori Inkpen, and Baris Unver. 2018. Wish you were here: being together through composite video and digital keepsakes. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2018, Barcelona, Spain, September 03-06, 2018*, Lynne Baillie and Nuria Oliver (Eds.). ACM, 17:1–17:11. https://doi.org/10.1145/3229434.3229476

[59] Haijun Xia, Sebastian Herscher, Ken Perlin, and Daniel Wigdor. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality. In *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018, Berlin, Germany, October 14-17, 2018*. ACM, 853–866. https://doi.org/10.1145/3242587.3242597

[60] Jackie (Junrui) Yang, Christian Holz, Eyal Ofek, and Andrew D. Wilson. 2019. DreamWalker: Substituting Real-World Walking Experiences with a Virtual Reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1093–1107. https://doi.org/10.1145/3332165.3347875

[61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 4470–4479. https://doi.org/10.1109/ICCV.2019.00457

[62] Ye Yuan, Jan Cao, Ruotong Wang, and Svetlana Yarosh. 2021. Tabletop Games in the Age of Remote Collaboration: Design Opportunities for a Socially Connected Game Experience. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 436:1–436:14. https://doi.org/10.1145/3411764.3445512

[63] Johannes Zagermann, Ulrike Pfeil, Roman Rädle, Hans-Christian Jetter, Clemens Nylandsted Klokmose, and Harald Reiterer. 2016. When Tablets meet Tabletops: The Effect of Tabletop Size on Around-the-Table Collaboration with Personal Tablets. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. ACM, 5470–5481. https://doi.org/10.1145/2858036.2858224

[64] Lei Zhang, Ashutosh Agrawal, Steve Oney, and Anhong Guo. 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 2023, Hamburg, Germany, April 23-28, 2023*. ACM, 36:1–36:14. https://doi.org/10.1145/3544548.3581136

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. https://doi.org/10.48550/arXiv.2302.05543 arXiv:2302.05543 [cs]

[66] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics* 36, 4 (July 2017), 119:1–119:11. https://doi.org/10.1145/3072959.3073703

A APPENDIX

A.1 Additional Examples of Blended Environments

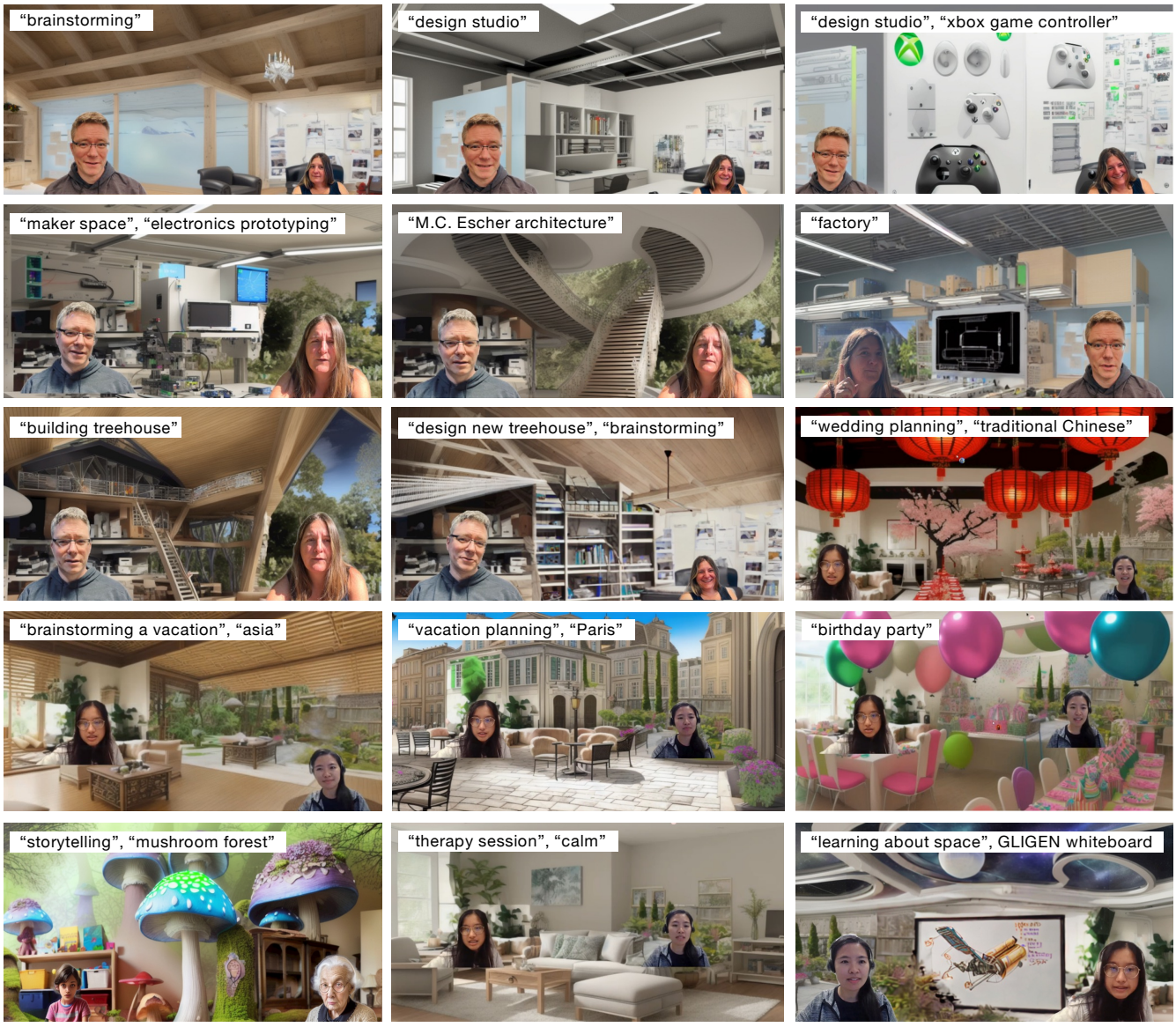

Figure 14: Examples of other generated meeting environments.

A.2 GPT-aided Prompt Crafting for Image Generation

Crafting textual prompts that result in desired output with image generation models (e.g., Stable Diffusion) requires significant expertise, and trial and error [4]. To ease the process of prompt creation for stable diffusion, BlendScape enhances users' prompts by leveraging an LLM (in this case, GPT-3.5) to augment users' prompts with additional keywords related to the *Meeting Activity* and *Meeting Theme*. The advantage of this approach is that it allows for the dynamic generation of keywords for image generation that suit the varying needs of users in the meeting.

For example, if the meeting activity is "brainstorming", and the theme of the meeting is "mushroom forest", then the system crafts a base prompt of Mushroom forest-themed environment for a brainstorming session. We then leverage single-shot prompting with GPT-3.5 to enhance the base prompt with contextually-relevant details:

System Prompt to GPT: Your task is to help the user create a Stable Diffusion prompt to generate an environment design. The user will specify an activity to occur in the environment and/or a theme for the space. You will provide a list of 4-5 types of objects to put in the environment and 4-5 distinct characteristics that describe the environment. The characteristics must be detailed and designed to generate visually appealing and cohesive results. Here is an example for a brainstorming activity:

{

Objects: "whiteboards, plants, chairs, small tables".

Environment Characteristics: "bright, open space, natural light, refreshing atmosphere, varied textures"

}

Subsequently, depending on the base prompt derived from the user's input prompt, BlendScape crafts an input to GPT-3.5:

Input Prompt (constructed by BlendScape) to GPT-3.5: Provide a list of 4-5 types of objects to put in this environment and 4-5 characteristics that describe this environment: Mushroom forest-themed environment for a brainstorming session. Return the output as comma-separated strings in JSON format: {Objects: string, Environment Characteristics: string}.

The results of such a prompting strategy to GPT results. in the following output.

{

Objects: "Mushrooms, Trees, Fairies, Moss-covered rocks, Lanterns".

Environment Characteristics: "Enchanting, Magical, Misty, Whimsical, Serene"

}

Additionally, BlendScape adds a fixed set of terms - "highly detailed, intricate, sharp focus, smooth" that we found to improve the results. In this case, the final prompt provided to Stable Diffusion is:

Mushroom forest-themed environment for a brainstorming session; Giant mushrooms, Fairy houses, Moss-covered rocks, Glowing mushrooms, Enchanted flowers; Enchanting, Magical, Misty, Whimsical, Serene; highly detailed, intricate, sharp focus, smooth.

Fig. 11 from our *Storytelling with Family* scenario shows the result of an image generation with such a prompt.

A.3 User Study Scenarios

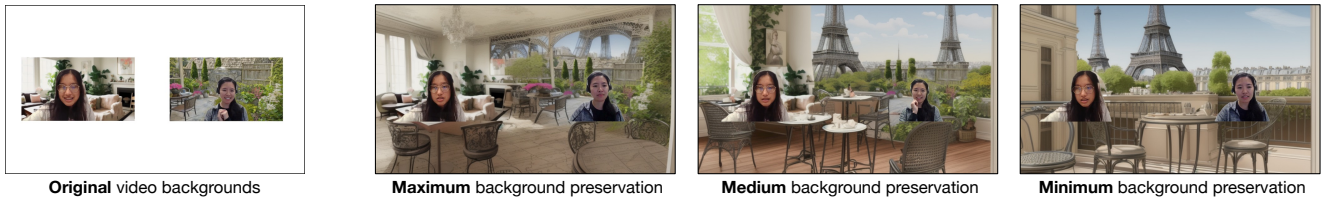

Figure 15: Vacation Planning Scenario: Two friends use BlendScape to get into the spirit of planning a trip to Paris. In Task 1 of our user study, we asked participants to compare three pre-canned environments that preserve their video backgrounds to different extents.

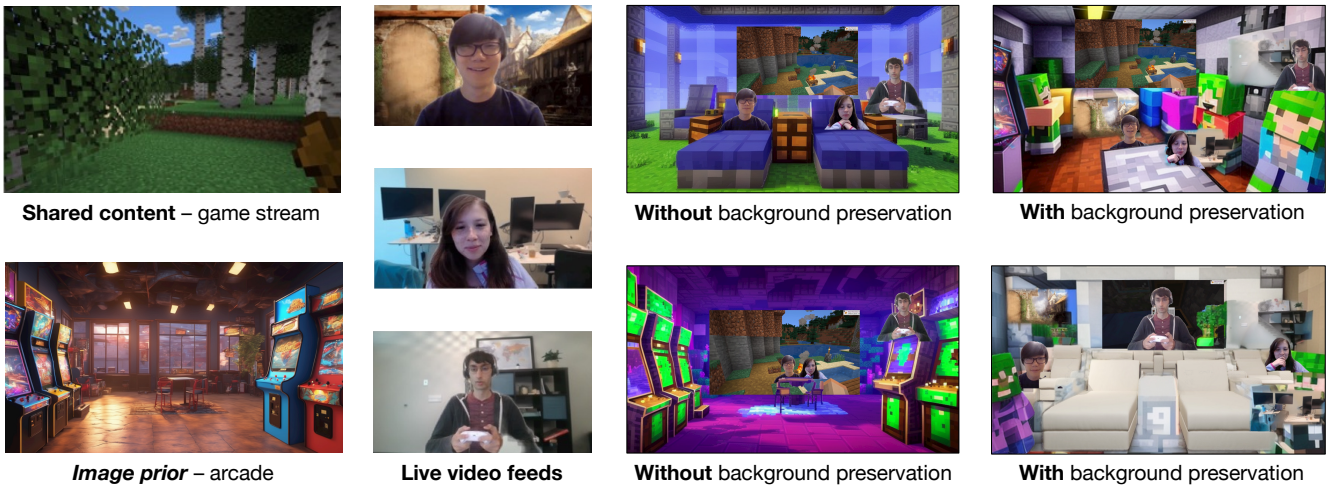

Figure 16: Game Stream Scenario: A gamer incorporates his Minecraft stream into a blended environment to make his viewers feel more connected to the gameplay. Participants compared four versions of spaces generated via image-to-image techniques, using two different priors and varying levels of background preservation.