VideoCraft: A Mixed Reality-empowered Video Generation Workflow with Spatial Layer Editing for Concept Video Creation

Boyu Li
Computational Media and Arts
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
blibr@connect.hkust-gz.edu.cn

Lin-Ping Yuan
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong SAR, China
yuanlp@cse.ust.hk

Zeyu Wang*
Computational Media and Arts
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong, China
zeyuwang@ust.hk

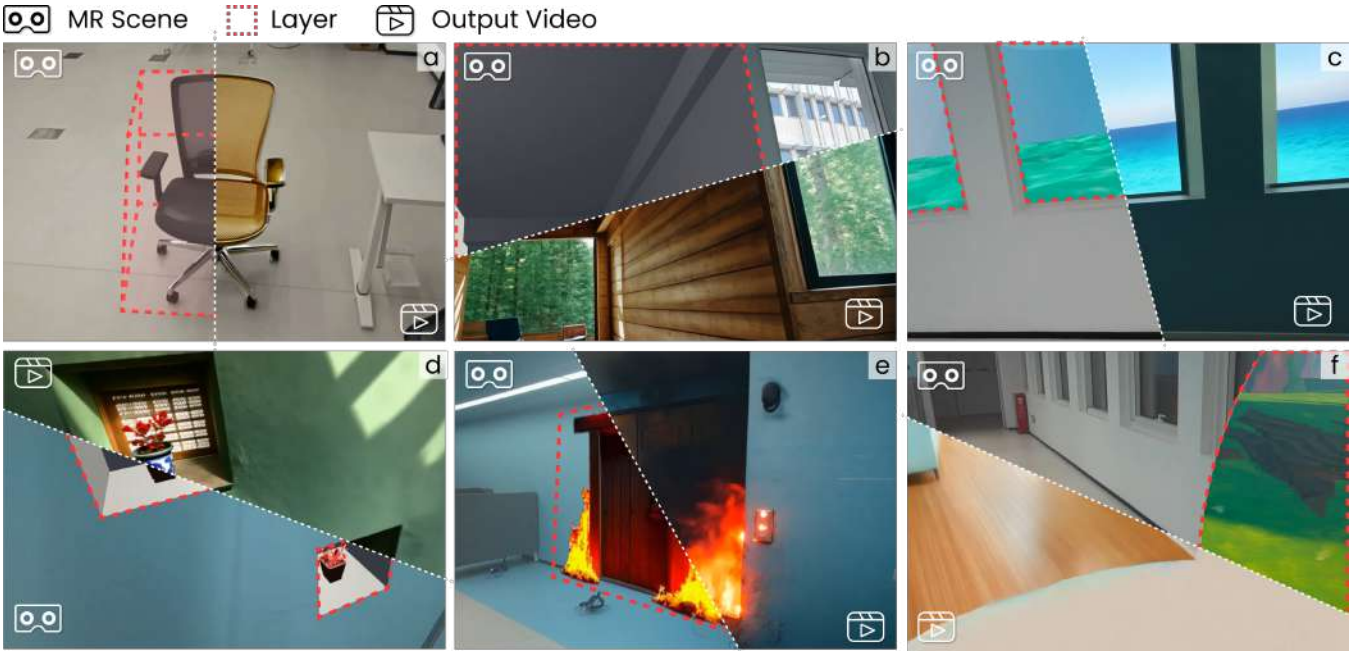


Figure 1: Overview of VideoCraft, an MR-powered video generation workflow that uses spatial layer editing to create concept videos in physical environments. Users can produce a variety of spatially grounded videos, such as: (a) transforming a chair into a golden style, (b) adding a wooden-style room, (c) changing a window view to an ocean, (d) carving a niche in the wall for a flower, (e) attaching a burning wooden door, and (f) transitioning smoothly from indoor to outdoor scenes.

Abstract

Concept videos for physical spaces are powerful tools for creators to explore and present spatial design ideas by integrating digital elements into real-world footage. While current video-to-video (V2V) generation models have eased the traditionally labor-intensive creation process, they lack support for seamlessly inserting new objects into original spaces and enabling precise spatial adjustments. To address these challenges, we propose VideoCraft, a novel mixed reality (MR)-empowered video generation workflow for concept video creation. Through a formative study, we identify key limitations in simply integrating MR and V2V models, particularly around localized editing for style and geometry. Therefore, we introduce

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
UIST '25, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747606

a spatial layer editing mechanism into the workflow, enabling intuitive spatial manipulation through layer shaping, features, and states. We evaluate VideoCraft through a controlled user study and expert interviews, demonstrating its effectiveness in enhancing spatial precision and creative control.

CCS Concepts

• **Human-centered computing** → *Interactive systems and tools*; *Mixed / augmented reality*.

Keywords

Concept Video, Video-to-Video Generation, Mixed Reality, Spatial Layer

ACM Reference Format:

Boyu Li, Lin-Ping Yuan, and Zeyu Wang. 2025. VideoCraft: A Mixed Reality-empowered Video Generation Workflow with Spatial Layer Editing for Concept Video Creation. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3746059.3747606

1 Introduction

Concept videos for physical spaces are visually engaging narratives that combine footage of real-world environments with digital overlays and visual effects to present conceptual transformations of how spaces can appear and function. They are widely used across various domains among different stakeholders. For instance, they help interior designers communicate proposed spatial layouts and design styles to clients [33], enable brands to create engaging advertising videos that captivate customers [56], and serve as entertaining short videos by allowing creators to visually reimagine everyday spaces [38]. A well-established approach to creating concept videos involves two key steps [49]: capturing real-world footage first and then editing it with tools like After Effects [8]. The editing step demands creators to manually overlay digital elements frame by frame, match colors and lighting, and adjust perspectives to ensure visual coherence and realism. Thus, the manual editing process is tedious and time-consuming, significantly slowing iteration cycles and hindering creative experimentation.

Recent advancements in Generative AI, especially video-to-video (V2V) generation models [16, 27, 31], can significantly reduce the burden of manually editing real-world footage (Figure 2-top). By taking the original footage of physical space and a text prompt as input, V2V models can automatically generate multiple concept videos in different visual styles. The generated videos effectively retain semantic information (e.g., object identities like furniture and walls) and structural information (e.g., spatial layout and object shapes) of the physical space. However, these models fall short when it comes to scene editing, particularly when creators want to introduce new objects that do not exist in the real environment. For example, while these models can transform a video of a bare-shell apartment into one with a modern renovated style (Figure 2-top), they provide limited support for adding furniture. Some recent V2V models [21, 27] alleviate this limitation by allowing creators to insert elements using text prompts and 2D masks. However, prompts and masks lack the precision and flexibility required for adjusting the position, orientation, and scale of objects in 3D spaces.

A promising solution to these limitations is leveraging Mixed Reality (MR), which allows creators to intuitively insert and manipulate digital objects within real-world environments, enhancing spatial awareness and contextual understanding, as demonstrated by many MR authoring tools such as PointShopAR [44] and SceneCtrl [54]. Additionally, recent research called VisTellAR [38] showcases the benefits of incorporating an MR pre-stage into the manual creation process of data videos, such as real-time data binding with physical objects and simplified perspective control.

Motivated by the advantages of MR and V2V models, we envision a workflow that integrates both technologies for concept video creation (Figure 2-bottom). In this workflow, MR could be used to edit the spatial configuration of physical spaces prior to video recording and then record footage that blends physical environments with digital elements. This footage would then serve as conditioning input for V2V models, which can further refine and enhance the footage through automatic editing.

To the best of our knowledge, we are the first to propose and explore such a workflow, which combines V2V generation with both virtual and physical content created in MR. Prior work [15] only focuses on real-world video input via headset or phone capture. The integration of MR with V2V remains uncommon due to the early stage of generative V2V, and their effective combination is non-trivial and still underexplored. Therefore, to understand the design considerations in building such a novel workflow, we first conducted a formative study with a simple tech probe by connecting a built-in MR furniture authoring tool in Quest 3 and a state-of-the-art V2V model [31]. Our findings indicate that the simple integration of MR and video generation is limited in enabling localized modifications of physical spaces in the generated video. Therefore, we introduce VideoCraft, which integrates a novel mechanism based on *spatial layers* to improve the workflow of MR authoring and video-to-video generation models. The spatial layer, created within the MR environment, serves as a guide for localized editing in the generated concept video. This mechanism is designed from three key aspects: (1) layer shaping, allowing users to flexibly define 3D regions for precise spatial selection in mixed-reality space; (2) layer features, enabling region-specific style editing (e.g., Figure 1-a) and geometry modification (e.g., Figure 1-b, c, d); (3) layer states, supporting dynamic and interactive behaviors to fulfill creative video requirements (e.g., Figure 1-e, f).

We evaluate VideoCraft through two studies: a user study with 12 participants and expert interviews with six professionals. The user study results show that VideoCraft facilitates the iteration design process in both spatial design and video creation, lowers the barrier for novice users, and enables more precise editing in V2V generation. The expert interviews further evaluate our workflow from a professional perspective. We identify key benefits and limitations, and offer suggestions for future enhancements to improve its practical applicability.

In summary, we make the following contributions:

- A formative study that investigates how designers use a novel workflow combining MR authoring and video-to-video generation to create concept videos for physical spaces, with a focus

Figure 2: MR and V2V for concept video of physical space. (Top) The user records a video of the real-world scene and transforms it to *"French Modern Style"* with V2V generation. (Bottom) The user places virtual elements in MR before recording, then generates a video with the same prompt. This is the initial motivation for our formative study.

on understanding user intentions, creative strategies, and the critical need for localized editing support.

- The design and implementation of VideoCraft, integrating a spatial layering mechanism in MR that enables users to define editable regions in physical space, apply localized style, and perform geometry modification.
- An evaluation of VideoCraft through a user study and expert interviews, providing insights into the role of individual components in supporting controllable and flexible editing, and expert reflections on the workflow's overall utility and design impact.

2 Related Work

2.1 Concept Video Based on Physical Space

Concept videos based on physical spaces are short visual narratives that integrate digital elements into real-world recordings to convey spatial ideas, design intentions, or conceptual transformations. Unlike video prototypes [17, 18], which are intentionally rough and sketch-like [45] to leave room for viewer imagination, concept videos are visionary and designed to impress and inspire. They often prioritize visual appeal over design detail, aiming to present a complete and compelling vision of a future or transformed space.

This emphasis on vision and storytelling makes concept videos widely applicable across domains. For example, in interior design [33], designers often begin with an empty room provided by the client and digitally add furniture, decorations, and stylistic elements. The resulting video helps clients visualize the outcome of a proposed renovation. Another use case is in creative video production for advertising or entertainment [45, 49], where virtual elements are inserted into real-world footage or the stylistic appearance of physical spaces is transformed to create imaginative scenes.

To produce such videos, different applications typically adopt different workflows, mainly falling into two categories: 3D-based and 2D-based. 3D workflows are commonly used in fields such as interior and architectural design, where designers reconstruct the physical environment using CAD drawings [1], create detailed 3D models in software like Rhino [30], and render high-quality video scenes using tools such as Blender [3]. In contrast, 2D workflows are more prevalent in post-production and visual effects tasks. Creators use tools like Adobe After Effects [8] to analyze and track camera motion, align virtual content with the spatial layout of the footage, and ensure consistency in perspective, lighting, and occlusion. These workflows are often complemented by additional techniques such as masking, recoloring [7], and texture overlays to transform the appearance of real-world environments.

While these workflows yield visually impressive results, they are time-consuming, technically demanding, and difficult to iterate, making them largely inaccessible to non-expert users. Our approach builds on these traditions by introducing AI-powered video generation into the concept video creation pipeline. This enables everyday users to rapidly prototype and visualize spatial concepts without requiring professional 3D modeling or video compositing skills.

2.2 Controllable Video Generation

Recent progress in generative video models has greatly lowered the technical barriers for concept video creation, with open-source algorithms like Stable Video Diffusion [2] and SparseCtrl [10], as well as commercial solutions such as KLING [16], Sora [25], and Runway Gen-3 [31], making it possible for novice users to synthesize high-fidelity video from simple text or image prompts. However, despite their impressive results, these systems still face substantial limitations in controllability, particularly when it comes to complex spatial design tasks.

A key challenge in controllable video generation lies in the limitations of existing input conditions used to guide the process. While text or image prompts are easy to use, they lack precise control over spatial layout and temporal events. To address this, some models introduce geometric conditions such as draggable keypoints [16], segmentation masks [21], or depth maps [5] to provide better control over structure and motion. However, these inputs remain insufficient for capturing complex, scene-level semantics. For example, keypoints may only control local object placement, while masks typically operate at the 2D frame level without awareness of scene depth or geometry. As a result, users still struggle to achieve fine-grained control over object positioning, camera motion, and global layout, especially in a 3D scene video with dynamic viewpoints or rich spatial structures.

To enhance the controllability of video generation, recent research has explored using video as an input condition [46]. This approach, commonly known as video-to-video generation, includes tasks such as video editing [9] and style transfer [4, 13]. As an input condition, videos support camera motion and scene layout, and provide stronger temporal and spatial constraints for generation. Leveraging these advantages, a typical use case involves a user recording a real-world video and inputting it into a V2V model to generate a stylized output that reflects the original scene. However, because the generated video inherits the fixed spatial structure of the input video, making localized modifications without altering the real-world environment remains challenging. To address these limitations, local video editing methods such as GenProp [21, 27] enable modifications by specifying 2D regions within individual frames. However, several constraints remain: 1) limited precision, as detailed attributes (e.g., exact shape, position, orientation, and scale) cannot be explicitly defined; 2) lack of real-time previews, requiring lengthy generation times before visual feedback; 3) inability to apply multiple edits simultaneously, forcing users to iteratively

Table 1: Comparison of representative related systems across key features, including scene-level editing, rapid prototyping, support for video generation, and output visual coherence. Unlike most previous works that simply attach rough virtual content to physical space, our system enables AI-driven reinterpretation of the MR scene to generate visually coherent video content.

	Scene-level editing	Rapid prototyping	Support video creation	Visual coherence result
After Effect [8]	✓	✗	✓	✓
PointShopAR [44]	✓	✓	✗	✗
RealityCanvas [45]	✗	✓	✓	✗
Pronto [18]	limited	✓	✗	✗
ProtoDreamer [55]	✓	✓	✗	✓
Ours	✓	✓	✓	✓

specify and generate each modification individually. Our novel workflow addresses these limitations by leveraging rich, interactive spatial inputs available in mixed reality (MR), enabling precise, real-time, and simultaneous modifications within a spatially coherent 3D environment. Combined with the expressive power of video-to-video (V2V) generation, our approach significantly enhances users' ability to efficiently create compelling and spatially consistent concept videos.

2.3 Video Authoring in XR

Dynamic content creation in XR (AR, MR, VR) environments has been extensively explored in both commercial tools (e.g., Quill [28], Vuforia [41]) and research prototype systems (e.g., RealitySketch [37], Sketched Reality [14], TimeTunnel [58], MotionMontage [11], AniCraft [19], 3D Puppetry [12], ARAnimator [48]). These systems leverage the spatial interaction capabilities of XR to support creative tasks, such as physical scene design [43, 44, 54], 3D character motion [19, 48], and interactive experience [51, 52, 57].

While XR focuses on immersive experiences, using it to create video offers a more accessible and easily shared medium. Several systems have been proposed that support video creation within XR environments [18, 23, 32, 38, 45]. For instance, Vremiere [24] supports timeline-based spherical video editing within a VR headset, allowing users to manipulate immersive video footage directly in the 3D environment. Beyond editing existing videos, other systems have begun to treat XR interactions as part of the video creation process itself. RealityCanvas [45] enables users to generate animated videos from hand-drawn sketches in AR, while Pronto [18] blends virtual and physical elements for situated video authoring. VisTellAR [38] presents a two-phase workflow that first enables users to edit data visualizations in AR, then merges these configurations with recorded footage to generate content-aware videos.

However, as Table 1 shows, most existing systems treat XR merely as a means to overlay virtual elements, such as sketches [45] or 3D models, onto the real world. This approach leads to two key limitations. First, the added virtual elements often appear visually disconnected from the physical environment due to inconsistencies in style, lighting, and material appearance [18]. Second, these systems offer little to no control over the appearance or geometry of the physical scene itself, preventing users from meaningfully transforming the environment. As a result, the generated videos resemble rough prototypes lacking visual consistency and narrative clarity, falling short of the goal of producing cohesive and aesthetically refined concept videos.

To address these limitations, we integrate video-to-video generation into the XR video creation workflow with a spatial layer-based mechanism, enabling consistent styling of virtual and physical elements and flexible edits to the appearance and geometry of the scene, thereby supporting the creation of cohesive and visually compelling concept videos.

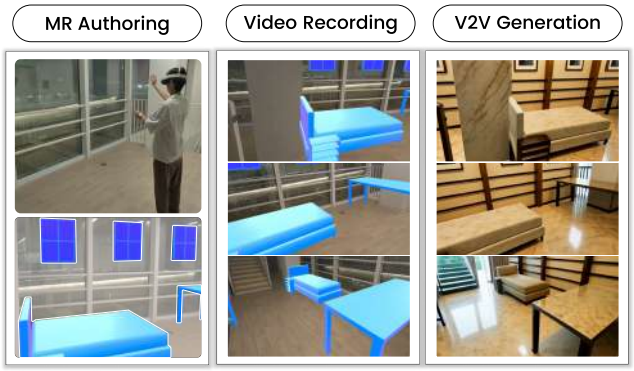


Figure 3: Tech-probe procedure. The user first places furniture in MR and records the scene. The captured video is then used as input for V2V generation, guided by a text prompt *"luxury living space."*

3 Formative Study

In the formative study, we proposed a straightforward workflow that combines MR authoring and V2V generation to support the rapid creation of concept videos based on physical space. To explore this workflow, we conducted a formative study with five creative designers. The formative study includes: 1) a semi-structured interview to understand the current practice of creating concept videos based real-world environment, 2) a tech-probe session to collect user feedback on the initial workflow to author concept videos.

Similar to recent work that begins with usable workflows to uncover design gaps [55], our tech probe directly connects MR interaction with emerging generative V2V tools to assess their practical integration into design workflows. Despite its simplicity, the

tech probe enabled meaningful creative tasks (Figure 3), revealing challenges rooted not in tool immaturity, but in fundamental design needs specific to MR with V2V authoring. Our goal was to evaluate the feasibility of this workflow, like whether AI-generated videos could achieve the intended visual effects, the unique advantages of the workflow, and the practical challenges that may arise during the process. Additionally, we aimed to analyze the limitations of existing workflows and derive design considerations to enhance user control and creative expressiveness in video prototyping.

3.1 Study Setup

Participants. Following the participant recruitment practice in previous work [19, 34, 44], we conducted a formative study with a diverse group of five creators (P1–P5; 3 males, 2 females), each participating for approximately 45 minutes. Specifically, P1 specialized in renovation concept videos for interior design and had extensive experience with AI video generation tools and VR-based design workflows. P2 focused on AI-driven video generation, particularly exploring how generative models can enhance creative storytelling. P3 had experience rendering concept videos in Blender and had previously worked with both real-time and pre-rendered 3D environments. P4 worked in film VFX and post-production, with expertise in integrating virtual elements into live-action footage. P5 was a game environment concept designer with a strong background in world-building, procedural content generation, and interactive media. All participants had at least two years of professional experience in their respective fields, with an average of 4.6 years (SD = 1.52). Three out of the five participants (P1, P3, and P5) had interdisciplinary expertise spanning areas such as interior design and VR workflows, 3D animation and real-time environments, or interactive media and procedural generation.

Procedure. This study was conducted in a controlled offline setting with two phases: a 25-minute interview and a 45-minute tech-probe session.

Phase 1: Semi-structured Interview. In Phase 1, we conducted semi-structured interviews to explore participants' typical workflows for creating concept videos based on real-world scenes. The interview covered topics such as their current processes, tools, and techniques for integrating physical spaces with digital elements, as well as the challenges they face when merging virtual content with real-world footage. We also discussed their creative flexibility, balancing technical constraints with visual storytelling, and their opinions on emerging technologies like AI-driven video generation and MR content creation. The goal was to identify key factors influencing their workflow and gather insights.

Phase 2: Tech-probe Session. In this session, we invited participants to our lab and provided several nearby indoor locations for them to choose from. Before starting, participants were given a brief tutorial to familiarize themselves with the MR device and authoring interface. After selecting a physical space, they followed our proposed workflow to create a concept video. Once the video was generated and reviewed, we conducted an exit interview to collect feedback on the workflow. This included comparisons with their existing practices, evaluations of both the creation process and the generated videos, and discussions of any challenges they encountered or suggestions for improvement.

Tech-probe: a straightforward workflow. Prior research [38] has demonstrated the potential of augmenting real-world contexts with virtual content to enhance communication and visualization. In particular, incorporating an MR pre-stage allows users to create and manipulate virtual content directly within physical environments before generating final outputs. Building on this idea, we propose a workflow that integrates MR and video-to-video generation to facilitate rapid design iteration.

As shown in Figure 3, the tech-probe workflow begins with MR content creation using the Quest 3 headset and the built-in Layout [22] app, which provides a set of preset furniture items for decorating physical spaces in mixed reality. After the MR scene is composed, a 5–10 second video of the decorated space is recorded directly through the headset. This recording is then uploaded to the website of Runway [31] on a laptop, where video-to-video generation is performed using Runway Gen-3 Turbo guided by a user-provided text prompt, typically within 60 seconds. The output is a stylized concept video that reflects both the original spatial layout and the intended visual transformation.

3.2 Interview Findings

We summarize the current practices and challenges in creating concept videos for physical spaces as follows.

Current Practices. Participants described two common approaches for creating concept videos based on physical spaces: 3D modeling and 2D post-editing. Three participants (P1, P3, P5) reported using 3D software to build virtual environments and render stylized concept videos, which offer greater creative freedom and precise control over spatial layout. P4 described a 2D-based workflow, where real-world footage was captured first and then edited using tools like After Effects for camera tracking and content modification. Additionally, there are AI-based commercial tools [27, 31] that enable object insertion and style transfer for video post-processing.

Current challenges. Despite these options, participants reported difficulties in quickly generating visually impactful concept videos. Several participants emphasized the importance of producing videos that are *"visually appealing"* (P1) and capable of *"leaving a strong impression"* (P5), especially during early-stage client communication. As P5 explained that *"clients often don't know what they want until they see something,"* making rapid concept video creation essential for aligning expectations. However, traditional methods often require extensive effort in *"tweaking the rendering effect"* (P3), which slows down the iteration process and makes it difficult to explore different spatial design ideas quickly.

Such a challenge in the laborious creation process is partially addressed by existing XR applications such as IKEA Place [26]. With simple tap-and-place functionality, these tools allow even non-professional users to visualize virtual furniture in physical spaces effortlessly. However, such applications are often limited in their editing capabilities, as they typically only support placing simplistic, often unrealistic-looking objects on flat surfaces. In contrast, our participants wish to produce high-quality concept videos as they created through traditional workflows, while enjoying the simplicity and immediacy of MR-style interactions.

3.3 Tech-Probe Findings

Drawing from participants' feedback after experiencing the tech-probe workflow, we summarize the key findings below.

Positive Feedback on the Tech-Probe. Participants generally responded positively to the proposed workflow, highlighting its potential to streamline the concept video creation process. All participants appreciated the ability to first perform rough spatial design in MR, which helped them quickly form a spatial understanding before generating higher-quality visual outputs through V2V generation. For example, P1 noted, *"It's helpful to get a basic feel of the space first, and then refine it into a polished video."* The workflow was also seen as effective in lowering the entry barrier for novice users. P4 commented, *"Even someone without a design background could easily prototype ideas in space, because the workflow is easy to understand compared with traditional modeling or video editing software."* Moreover, participants praised the intelligence of the video generation model. Instead of merely applying a uniform style transformation, the model was able to semantically recognize virtual objects and convert them into realistic representations. For example, P3 noted, *"It's impressive that the virtual table I placed was turned into a realistic wooden table, it's not just style transfer, it understands the content."*

Limitations in the Tech-Probe. While participants appreciated the overall efficiency and creativity enabled by the workflow, they also identified several key limitations that hindered creative feasibility and expressiveness during creation. We summarize these limitations as follows:

Insufficient 3D Region Selection for Targeted Modifications (C1). The current workflow lacks intuitive tools for precisely selecting specific regions within the 3D physical environment, making localized editing difficult. Although the generated video maintains overall spatial coherence, users have limited means to isolate and manipulate targeted areas. For instance, P5 expressed a desire to *"select a region on the floor for editing,"* yet the current tech-probe does not support directly defining editable areas on real-world surfaces. Similarly, P2 highlighted the need to *"manually draw a 3D region"* to specify a spatial volume for targeted modification. These examples underscore a critical gap in region selection capabilities.

Limited Local Semantic Control in Video Generation (C2). The current video generation process treats the input scene as a holistic semantic entity, making it difficult to apply independent modifications to specific regions. While effective for overall transformations, it lacks local control over individual objects or areas. For example, participants expressed the desire to make *"only the table wooden"* (P3) or to apply *"different styles to different walls"* (P5), but the system applies changes uniformly across the scene. As a result, scene outputs often deviate from users' intended compositions, especially when distinct design elements are required.

Physical Space Geometry Constraints in Video Generation (C3). The video-to-video generation model preserves the geometric structure of the input footage, which, while ensuring spatial consistency between input and output, limits the ability to reshape the physical environment. Participants reported challenges in *"widening the room"* (P4) or adding *"structural elements like a floor-to-ceiling window"* (P3), highlighting the constraints imposed by fixed input geometry. As a result, creative modifications are confined to the existing space, limiting opportunities for spatial exploration during prototyping.

Lack of Dynamic Elements in Video Generation (C4). The current video generation approach centers on static scene elements, overlooking temporal dynamics that bring scenes to life. While spatial consistency and stylistic edits are well preserved, the system lacks support for movement, interactivity, and animation. Participants noted the absence of *"vibrancy"* (P1) and expressed a desire to prototype *"not just spaces but events"* (P5). This limitation reduces the expressiveness of the resulting concept videos.

3.4 Design Requirements

We distill the following design requirements to address the limitations identified in the tech probe:

Provide flexible metaphors to define specific 3D regions for modification (R1). This could include options for creating custom shapes, selecting areas on real-world surfaces, and defining regions with varying levels of detail. It is essential to support various modification types, such as individual objects (e.g., furniture) or specific areas (e.g., windows, floor).

Support localized semantic control for targeted style (R2). It should enable users to apply distinct styles to specific regions within the concept video, such as individual objects or areas, allowing for precise semantic differentiation. This is especially important for complex regions, where fine-grained control is necessary to achieve the desired outcome.

Allow modifications of physical spatial boundaries (R3). It should enable users to alter the physical space geometry by introducing new virtual spatial elements, such as expanding rooms or adding features like windows, without being limited by the existing physical structure. Empowering designers to explore more creative and diverse spatial configurations and enhancing the prototyping process for more imaginative design possibilities.

Enable dynamic and interactable elements (R4). It should allow for the integration of dynamic elements, such as target regions and interactive behaviors, into the scene to provide a more vibrant and immersive experience, enhancing the visual appeal and creating more engaging concept videos.

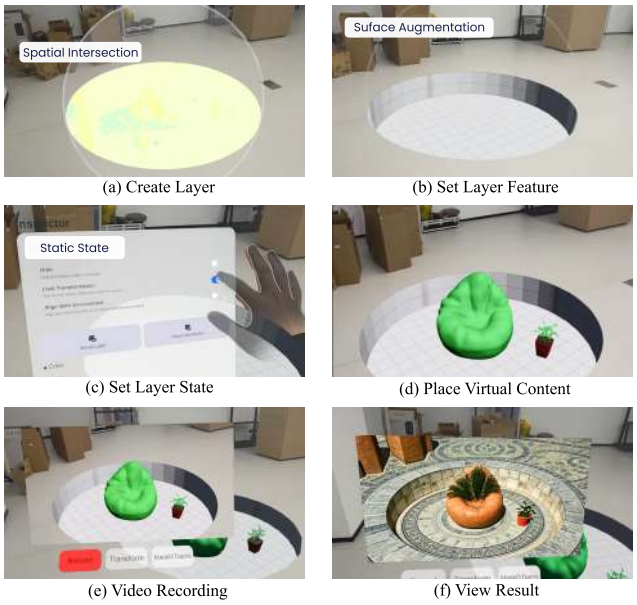
4 VideoCraft

4.1 Overview

From our tech-probe in the formative study, we identified that users need a way to perform localized editing within MR environments. Therefore, we introduce a layer-based mechanism to support spatially constrained editing and concept video authoring in MR.

This mechanism functions similarly to adjustment layers in tools like After Effects [8]. It does not alter the original physical environment directly, but acts on specific regions, enabling transformations in style, geometry, or even temporal behavior. Our layer mechanism consists of three key components:

- **Layer Shaping (R1)** is a flexible set of techniques for defining spatial layer regions in 3D space, supporting a wide range of selection needs.
- **Layer Features (R2, R3)** is a set of operations that apply style modifications or geometric changes to selected regions, enabling both visual and structural transformations.



(a) Create Layer (b) Set Layer Feature (c) Set Layer State (d) Place Virtual Content (e) Video Recording (f) View Result

Figure 4: Authoring workflow of VideoCraft for creating a concept video of a sunken seating area. (a) The user creates a circular layer by intersecting a virtual sphere with the physical floor. (b) The layer is set to *geometry modification* and assigned a sunken floor asset to simulate a recessed area. (c) The layer state is set to static. (d) The user decorates the area further with additional elements. (e) A video of the mixed space is recorded. (f) The user views the generated concept video.

- **Layer States (R4)** is an extension that enables layers to carry dynamic or interactive properties, adding temporal depth and user-driven control to the authored content.

Building on these three components, we integrate prior MR techniques into a unified spatial layer mechanism tailored for MR and V2V workflows, enabling more expressive and spatially grounded video authoring. Unlike existing video editing approaches that rely on coarse, frame-based 2D masks [21, 59], our method introduces embodied spatial interactions that occur in MR before video capture, allowing users to intuitively define, manipulate, and animate localized style and geometry edits in correspondence with the physical 3D environment.

4.2 Usage Scenario

Figure 4 showcases a usage scenario demonstrating how creators can leverage VideoCraft to efficiently create a concept video for a physical space. Imagine Eric, an interior designer, who wants a concept video to visualize his design idea for a sunken seating area reflecting ancient Greek aesthetics. To achieve this, Eric selects a relatively empty room in his apartment as the starting point. Equipped with an MR headset, he launches VideoCraft, immerses himself in the MR environment, and begins the creation process as described below.

Step 1: Defining the Editing Layer. To specify the area for the sunken platform on the ground, Eric decides to create a layer shape by spatial intersection. He selects a virtual sphere from the layer tool and scales it to an appropriate size. Switching to intersection mode, he positions the sphere so that it partially intersects with the floor. This creates a circular intersection region on the ground, which will serve as the boundary of the sunken area (Figure 4-a).

Step 2: Applying Layer Feature. After defining the circular region, Eric assigns the layer the *Geometry Modification* feature. From the asset library, he selects a recessed circular floor model and aligns it precisely with the intersection layer (Figure 4-b). In this example, *Geometry Modification* is applied for surface augmentation, allowing Eric to simulate a sunken area as part of his design concept.

Step 3: Configuring Layer State. To ensure the layer remains fixed, Eric locks its transform settings, configuring it as a static state (Figure 4-c). While the application offers a variety of preset layer animations and interactive behaviors, he opts to keep the layer static to preserve the structural realism of his sunken area design. At this point, he has successfully overlaid a virtual recessed platform onto his physical environment.

Step 4: Decorating the Scene. Eric then decorates the space by placing additional assets. He selects a sofa and a potted plant from the assets library and positions them inside the sunken area (Figure 4-d). If needed, he can repeat Steps 2, 3, and 4 to define additional layers and populate the space with more MR content, allowing for continuous iteration during the design process.

Step 5: Recording and Video Transformation. With the initial scene laid out in MR, Eric switches to the video authoring interface. He presses the [Record/Stop] button to record a 5-second first-person walkthrough of the setup (Figure 4-e) and input prompts *"ancient Greek style"* via voice. He then taps the V2V transform button to begin the generation process. While waiting for the approximately one-minute transformation, Eric walks around to experience the spatial composition from different perspectives. Once the generation is complete, he presses the [Play] button to view the resulting video with MR (Figure 4-f). The scene has been transformed into a richly styled sunken seating area, complete with Greek-inspired textures and architectural elements. If desired, he can continue editing the MR scene or refine the prompt to explore alternative stylistic variations.

4.3 Layer Shaping

The shape of the layer represents a specific region in the MR space that the user wishes to select for subsequent operations. As illustrated in Figure 5-a, to ensure flexibility and intuitive creation of these regions (**R1**), we provide several methods to address different editing scenarios:

Basic Shape. Following common practices in 3D modeling software [3, 40], we support primitive shapes such as spheres, cubes, cylinders, and cones for general, coarse regions.

Custom Shape. Users can create custom meshes by pinching to add vertices to the layer, a technique commonly used in Blender [3], allowing for finer control over the selected region.

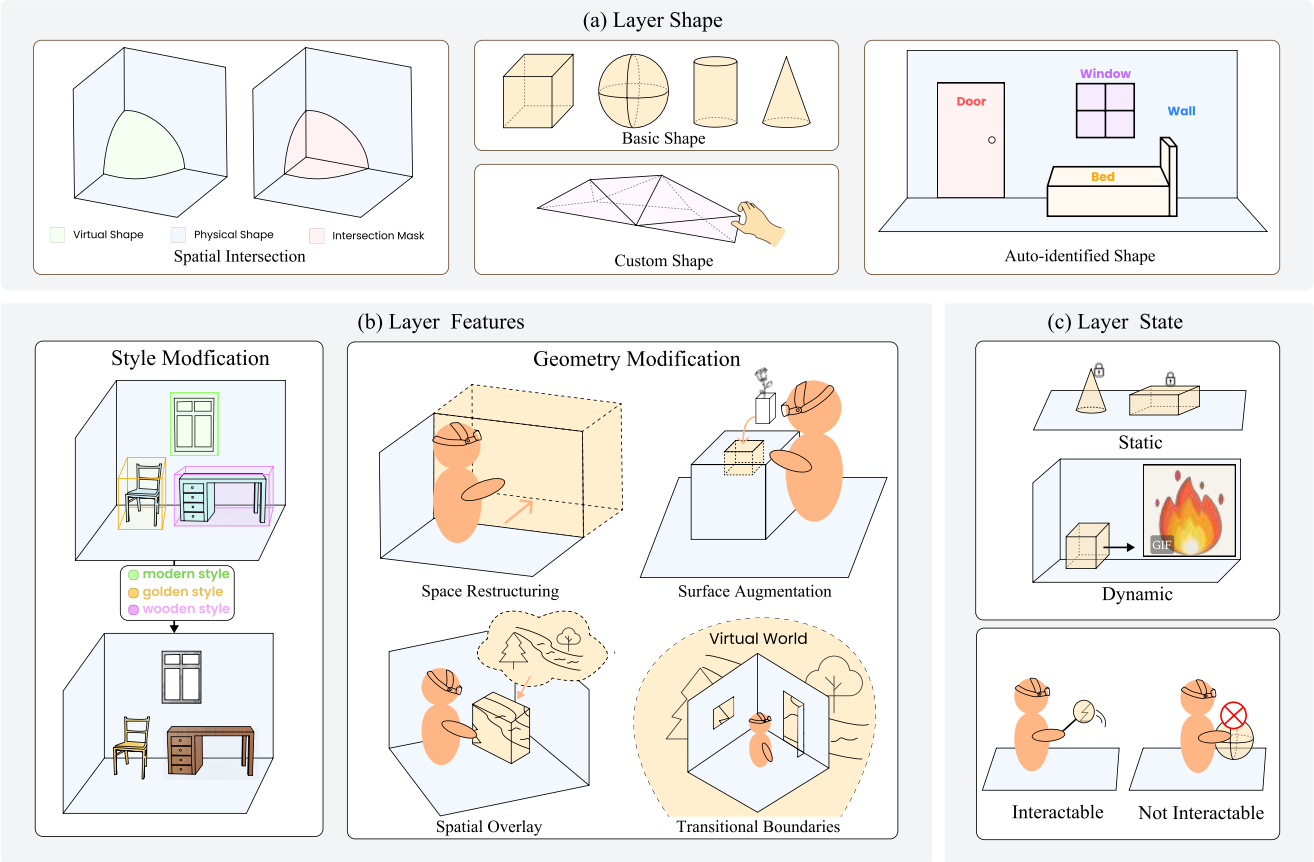


Figure 5: Design space of spatial layer. It consists of three components: layer shape, layer feature, and layer state, which together define how the layer behaves and interacts within the mixed-reality space.

Spatial Intersection. Boolean operations are a fundamental technique in computer graphics [29], allowing intersections between meshes to define new shapes based on existing surfaces [50]. We extend this concept to mixed reality by enabling intersections between virtual shapes and physical surfaces. For example, a sphere intersecting with the ground can define a circular region (e.g., Figure 4-a), providing an intuitive method for selecting surface areas on a physical surface.

Auto-Identified Shape. Leveraging Meta's automatic object recognition feature [22], which detects and outlines objects in a room, our system allows users to directly use these auto-identified elements, such as walls, doors, windows, and beds, as layer shapes.

4.4 Layer Features

4.4.1 *Style Modification.*

To achieve fine-grained semantic control over localized regions in the generated video (**R2**), we introduce the *style modification* technique, enabling users to specify distinct 3D regions in physical spaces, allowing targeted style transfer and content modification. While the workflow identified in the formative study is capable of applying overall scene transformations, it lacks the ability to make localized semantic adjustments. For the virtual object we added, we could extract masks simply by setting the material color for semantic control, similar to ProtoDreamer [55]. However, it

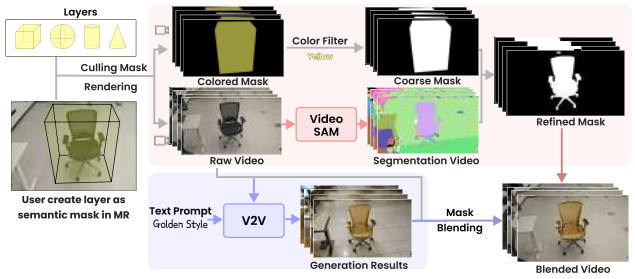


Figure 6: Pipeline of style modification. The user modifies the chair in the scene to a golden style by first defining a 3D semantic mask in MR. The mask is refined via the Video SAM algorithm and guided by a text prompt during video generation. The final result applies the refined mask to ensure localized style modification.

is challenging for objects or regions in real-world environments where object boundaries may be complex and automated segmentation methods are limited to recognizing only simple objects with moderate precision.

Inspired by PointShopAR [44], which performs point cloud selection and refinement through 3D segmentation, we adopt a similar concept of region-based refining but approach it from a video segmentation perspective. Instead of reconstructing the 3D scene, we introduce colored 3D layers as semantic masks in the MR environment to define editable regions in the output video. To ensure accurate isolation of complex shapes, we further integrate VideoSAM [6] to refine these masks.

Figure 6 illustrates the technical pipeline for localized style modification of the physical content. **Layer creation and recording:** The user defines a 3D layer to cover the target region as a semantic mask to modify specific objects. Two cameras record simultaneously: one captures the MR scene without semantic layers, while the other records only the colored masks. Culling masks keep semantic layers hidden from the user and the final video. **Generating Refined Masks:** Color filtering extracts rough masks from the mask video, while Video SAM [6] segments the raw video. A region is marked as a target if over 90% of it overlaps with a layer mask. The final refined mask is the union of identified target regions across frames. **Generation and Compositing:** Another parallel task involves applying V2V generation to the recorded raw video with user-provided prompts. Since the model regenerates the entire video globally, we leverage the refined masks to isolate specific regions in each frame, enabling localized transformations while maintaining overall scene consistency.

4.4.2 *Geometry Modification.*

Geometry modification allows users to overlay virtual layers onto physical space, enabling structural alterations that are otherwise infeasible in MR environments. When utilizing video-to-video generative models, users are often constrained by the inherent geometry of the real-world environment, which cannot be easily altered by simply adding virtual objects. Without complex scene scanning and reconstruction [54], modifying spatial structures, such as expanding a wall or adding a new room, remains challenging. As described in **R3**, our approach aims to address this limitation.

Previous research has demonstrated virtual content in MR has the potential to change the appearance of real-world content [20]. Inspired by the recently popular Portal Effect [39] in mixed reality, we adopt this mechanism to facilitate geometry modification, as its extensive application has already demonstrated its potential for transforming real-world environments. Technically, this is achieved by configuring the render queue in the order of the virtual scene, layers, and pass-through RGB stream with layers participating in depth testing without rendering visible textures. This setup selectively occludes parts of the real-world scene, allowing users to perceive the virtual elements beyond. Despite its technical feasibility, there has been little structured exploration of this approach in design. In this section, we investigate its potential as a spatial transformation framework, examining how modifying geometric structures can support video-to-video generation. The built-in scene scanning feature of the Quest 3, though rough but fast, is sufficient for our needs. As shown in Figure 5, through this conceptual exploration, we identify and categorize four distinct strategies for reshaping spatial geometry.

Figure 7: Example of space reconstruction. (Left) Expanding the original space by adding a virtual room. (Right) Top: Recorded video in MR; Bottom: Generated video based on the prompt *"A wooden cabin in the forest."* The red region marks where the layer is placed.

Spatial Reconstruction. When the existing spatial configuration does not meet functional requirements, users can manually adjust the spatial layout. For example, as shown in Figure 7, to add a new room behind an existing wall, users can define a cubic spatial intersection layer on the wall and insert a virtual room. While the inserted virtual space does not inherently possess the realism of the physical world in mixed reality, the video generation process seamlessly integrates the virtual and real elements, transforming them into a cohesive, stylistically consistent scene.

Figure 8: Example of surface augmentation. (Left) Creating recesses on the wall to place flowers. (Right) Top: Recorded video in MR; Bottom: Generated video based on the prompt *"Chinese traditional style."*

Surface Augmentation. In contrast to large-scale spatial reconstruction, which focuses on modifying overall spatial configurations, surface augmentation allows for localized geometric transformations on object surfaces. As shown in Figure 8, users can alter surface topology, such as carving a recess into an object and placing a flower into it. This enables finer-grained geometric modifications, which are particularly useful for subtle structural adjustments that enhance the creative flexibility of concept videos.

Spatial Overlay. Layers are not limited to attaching to the surfaces of physical objects; they can also exist as freestanding entities within space. Such configurations could achieve intersecting effects, like spatial portals or magical spheres that reveal glimpses of another world. For instance, as illustrated in Figure 9, a layer is placed on a virtual magnifier, enabling an experience where users can see a virtual ghost through the magnifier, an entity that would

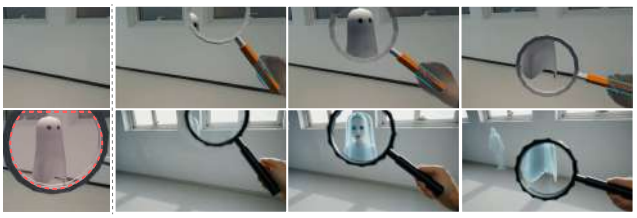


Figure 9: Example of spatial overlay. (Left) Putting a layer of spatial overlay on a magnifier to observe the invisible ghost. (Right) Top: Recorded video in MR, the user could see the virtual ghost through the magnifier in real time; Bottom: Generated video based on the prompt *"floating ghost."*

Figure 10: Example of transitional boundaries. (Left) Applying layers to detected windows. (Right) Top: Recorded mixed reality video, with the scene outside replaced by a sea view; Bottom: Generated video based on the prompt *"blue sea."*

otherwise remain invisible. This effect is further enhanced through video generation, creating a more interesting visual experience.

Transitional Boundaries. When modifying the view outside a window, transitional boundaries provide an effective method for seamlessly transitioning from the physical environment to a virtual scene. For instance, by utilizing the auto-identified window shape, a layer can be applied to the window surface, replacing the exterior view with a virtual world, as illustrated in Figure 10.

4.5 Layer State

To support diverse authoring requirements (**R4**), layers can be assigned different temporal and interactive properties through three distinct state configurations:

4.5.1 Static State. Static layers maintain fixed spatial properties (position, rotation, scale) and visual appearance throughout the video recording session. These layers serve as stable reference points for persistent environmental modifications.

4.5.2 Dynamic State. Dynamic layers exhibit time-varying transformations through programmed animations or procedural transformations, enabling rich temporal effects in the MR environment.

Geometric Transformations. The position, rotation, and scale of layers dynamically evolve along a predefined setting, enabling them to move, expand, or contract over time. For example, as shown in Figure 11, the user initially sets the layer by spatial intersection of a sphere with the physical environment. Subsequently, the layer serves as a transitional boundary for geometric modifications,

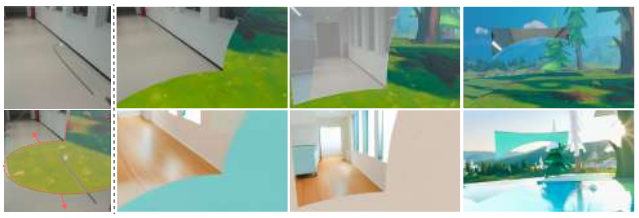


Figure 11: Example of geometric transformation. (Left) Creating a layer of transition boundary with spatial intersection that can continuously expand over time. (Right) Top: Recorded mixed reality video, with the scene transition from indoor to forest; Bottom: Generated video based on the prompt *"smooth transition style."*

Figure 12: Example of content sequences. (Left) Attaching a looping video with a transparent background to a wall layer. (Right) Top: Recorded mixed reality video of a burning wooden door; Bottom: Generated video based on the prompt *"first person view".*

smoothly expanding to facilitate a transition from an indoor scene to an outdoor scene.

Content Sequences. Frame-based animations or video streams can be mapped onto layer surfaces, enabling dynamic visual effects (e.g., an animated GIF overlay simulating flowing water). For example, we utilize a video generation model [42] to generate a short video with a transparent background and attach it to a layer on the wall, as shown in Figure 12.

Together, these techniques allow dynamic layers to create engaging visual effects, such as moving style modification, fading textures, or progressively expanding geometry modification regions, enhancing the flexibility and creativity of concept video authoring.

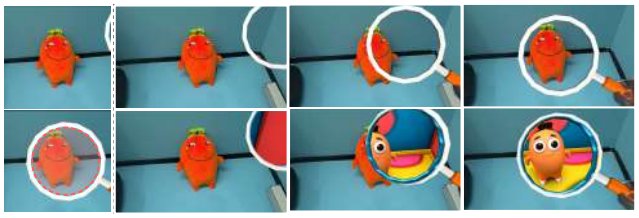


Figure 13: Example of interactive state. (Left) Attaching a layer of style modification to a magnifier. (Right) Top: Recorded mixed reality video, using a magnifier to observe the toy in first-person view; Bottom: Generated video based on the prompt *"3D cartoon style."*

4.5.3 Interactive State. In our system, the layer could also be interactable. For instance, as shown in Figure 13, a style modification layer is attached to a magnifier, allowing users to grab and manipulate the layer during video recording, and the user could interactively select which region undergoes transfer when generating. Additionally, other potential layer implementations could incorporate gesture recognition or user positioning as triggers, further enriching the creative possibilities for first-person view video. A similar example is shown in Figure 9, where the layer is interactively used as a geometry modification for spatial overlay.

5 Implementation and Interface

We implemented a prototype system with MR interface (Figure 14) that provides a unified workflow for layer-based MR authoring and video-to-video generation, enabling seamless concept video creation and transformation within a mixed reality environment.

Figure 14: Interface of our implementation, including the inspector panel, asset library, and video interface.

5.1 MR Interface

Space Authoring. Our system provides an intuitive space authoring interface that enables users to design and manipulate layers within a mixed-reality environment. The interface includes an *asset selection panel*, allowing users to browse and import various 3D models and prefabricated elements. Additionally, an *inspector panel* enables fine-grained control over object or layer attributes such as scale, rotation, and positioning. Users can also configure layer-specific attributes, including whether an object functions as a style modification for video processing or as a geometry modification layer to influence the physical scene's appearance.

Video Authoring. Beyond MR content creation, our system integrates essential video authoring capabilities, including real-time recording of the mixed reality scene, in-situ video preview in MR, video-to-video generation, and the related interface to view the generated video directly in MR. These functionalities ensure a seamless workflow from MR content design to video generation, allowing users to efficiently record, refine, and review their videos within the same immersive environment.

5.2 System Implementation

Preset Elements. To facilitate rapid prototyping, we provide a collection of preset assets, including fundamental geometric shapes and various models. We also support various virtual scenes to use in the geometry modification feature (Section 4.4.2). Additionally, we include dynamic elements with pre-configured animations, such as animated characters, water surfaces with realistic wave motion, and objects that autonomously adjust their scale and movement. Since animation creation is beyond our scope, our implementation does not include an animation authoring system. Furthermore, to facilitate seamless integration of animated overlays, we include pre-generated transparent-background videos from TransPixar [42] (Section 4.5.2), as their generation is time-intensive.

Hardware and Software Setup. Our system is developed for the Meta Quest 3, leveraging Unity as the primary development platform. The video-to-video transformation is powered by Runway Gen-3 Turbo/Alpha APIs, while computationally intensive tasks such as *Video SAM* [6] are executed on a server equipped with an NVIDIA RTX 4090.

Video Recording and Processing. For privacy reasons, Meta Quest 3 does not allow direct access to the built-in camera. As a workaround, we simply cast to stream the MR view to an external display and use Unity to record the Casting window to capture and record the video. In terms of processing time, a standard video generation without semantic masks takes approximately 30 seconds, while incorporating semantic masks increases the processing time to around one minute for the additional computation required for segmentation and video processing. To preserve the geometric structure of the physical environment as much as possible, we set the structure transformation hyperparameter in V2V generation to 0, ensuring minimal distortion to the original geometry.

6 User Study

We conducted a comparative study to evaluate VideoCraft. Since VideoCraft is the first workflow that integrates MR and V2V generation specifically for rapid concept video creation, there is no established baseline for direct comparison. Therefore, our study was designed to provide a deeper understanding of the individual contributions of key components in our workflow: MR, V2V, and layer-based editing, by isolating each and observing their respective effects on the creation process. This study design not only allows us to validate the effectiveness of the integrated workflow and understand the role of each component in shaping the overall user experience, but also enables us to investigate creativity support, as we placed no restrictions on design content, standardizing only the physical space across all conditions.

6.1 Study Design

6.1.1 Participants. We recruited 12 participants (P1-P12; 7 males and 5 females), with an average age of 25. Among them, six participants had over two years of experience in video production, while the remaining participants were novices in this domain. Unlike the formative study, which required professional designers to derive design considerations, we included novice users in this study to assess whether our tool could assist beginners in creating and generating concept videos for physical spaces. Additionally, five participants had prior experience using XR devices.

6.1.2 Study Setup. The study was conducted in a controlled lab environment, where participants used our system to create concept videos. Participants interacted with the unified application through

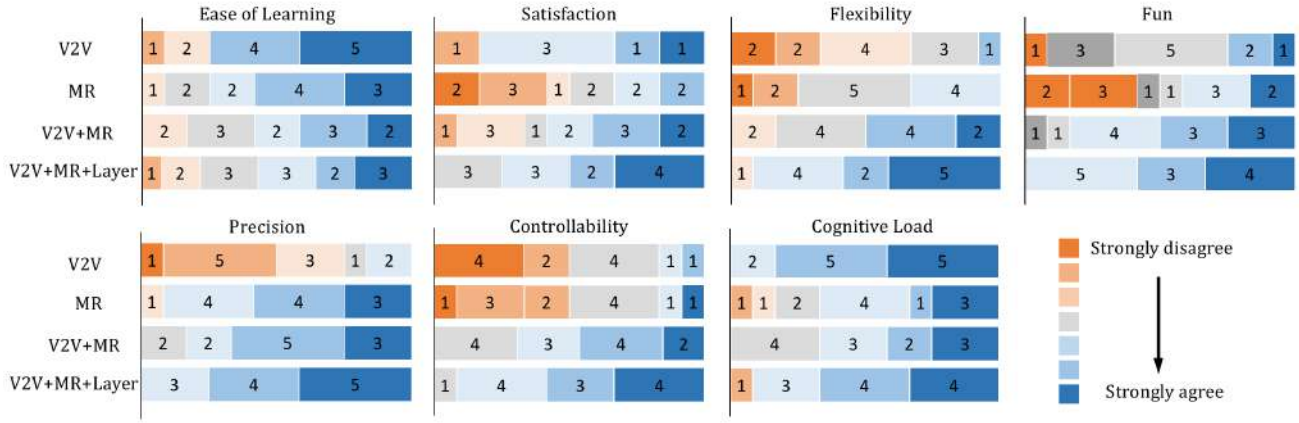


Figure 15: User ratings for four system conditions across seven evaluation criteria. Participants evaluated V2V, MR, V2V+MR, and V2V+MR+Layer using a 7-point Likert scale across seven criteria: Ease of Learning, Satisfaction, Flexibility, Fun, Precision, Controllability, and Cognitive Load. Color shading indicates the level of agreement, from orange (disagree) to blue (agree).

a Quest 3 headset, completing all steps within the MR environment, including viewing the final generated videos. We provided four physical spaces for participants to choose from, ranging in size from 8 to 20 square meters. The preset elements available for MR authoring are detailed in Section 5.2. A laptop was used to communicate with the headset, handling tasks such as video storage, invoking the V2V generation API, and performing basic video processing. It was also connected to a server responsible for running AI algorithms, including video segmentation.

6.1.3 Procedures. The user study session lasted approximately 70 minutes and was conducted individually. The study consisted of the following stages:

Tutorial Session (15 mins). Participants first completed a consent form and received a brief introduction covering the project background, the initial workflow used in the formative study, and the layer-based mechanism. They first explored the integrated V2V generation feature without MR authoring, allowing them to understand the baseline capabilities of video transformation. Subsequently, they engaged with the space authoring interface (Figure 14), learning how to browse the asset library, place virtual objects, and modify their properties using the inspector panel. Next, participants practiced spatial layer editing and completed a guided walkthrough, which included: 1) using the style modification feature to change the style of a chair (Figure 6), 2) performing a surface augmentation task (Figure 8), and 3) exploring layer states through the provided example case (Fig. 11).

Comparative Task (45 mins). Each participant compared four conditions: 1) V2V generation only, 2) MR authoring only, 3) MR authoring combined with V2V generation, and 4) MR authoring with layered editing followed by V2V generation (i.e., VideoCraft). All participants followed the above sequence to mitigate bias from early exposure to advanced features. The participants were required to use the same physical space, but the specific content and creative direction were entirely up to them. For the VideoCraft condition, participants were encouraged to make full use of the layer functionality to explore its potential. Each condition resulted in a complete video output. To reduce fatigue, participants were given the option to take a short break of up to 5 minutes between conditions.

Exit Questionnaires and Interviews (10 mins). At the end of the study, participants completed a questionnaire assessing their experience from several aspects, including ease of learning, satisfaction, flexibility, fun, precision, controllability, and cognitive load. We conducted semi-structured interviews to collect qualitative feedback on their preferences over different workflows, perceived advantages and limitations, and potential applications for our design.

6.2 Results and Findings: User Ratings

Enhanced Control through Layered Editing. Participants rated VideoCraft highest in terms of precision and controllability (Figure 15). Compared to MR-only and MR+V2V conditions, layers provided explicit spatial control. In MR-only, users could place objects but not influence how the physical scene was stylized. MR+V2V allowed global transformation, but lacked localized editing. With layers, users defined exact target regions using 3D shapes and then specified editing functions. P6 noted, *"I could change just one thing, a chair, not the whole room. That's not possible in other modes."* P8 added, *"It's clear and predictable, I know which part will be edited and how."* This fine-grained control aligned the MR setup with the intended output. Unlike earlier conditions where users had to adapt to whatever the physical space was, layers allowed users to shape the input explicitly. P2 described it as *"not just placing objects, but reshaping the space"*

Reality-Virtual Fusion Brings Desirable Effect. Participants also reported that layer-based editing significantly improved the integration between virtual and real-world elements. In the MR-only condition, virtual content appeared disconnected. P3 commented, *"Everything looked fake, like floating models."* MR+V2V improved realism but was still constrained by the original geometry. P4 observed, *"I could add things, but the space still felt the same."* Using

geometry modification layers, participants modified spatial structures, such as adding windows or expanding walls, before applying V2V generation. The resulting videos felt more cohesive. P9 noted, *"Once I edited the wall and ran generation, it looked like a real architectural change."* The ability to change not just appearance but also structure created a clearer break from the limitations of the physical space. P5 summarized, *"Adding objects is one thing, but changing the space itself, that's when it starts to feel real."*

6.3 Results and Findings: Creativity Support

Enhanced Creative Engagement. Our system supports creativity by providing an enjoyable experience, with all participants agreeing that our workflow is more fun. The interplay between mixed reality authoring, layered editing, and AI-driven stylization ignited participants' curiosity, leading them to experiment repeatedly. P7 remarked, *"I enjoyed testing new prompts just to see how far I could push the scene; it was like watching ideas come to life."* Observing immediate transformations in the generated video motivated users to think beyond conventional room designs. Some participants even introduced fantastical elements, such as floating lanterns or hidden portals, purely to see how the model would interpret them. This process of playful iteration, supported by immediate feedback from the AI, kept users engaged and often led them to explore novel design directions.

Efficiency in Iterative Design. Participants highlighted the system's capacity for rapid prototyping. Once users had set up and recorded an initial scene, they could make small adjustments, such as repositioning a virtual element or tweaking a text prompt, and quickly regenerate a transformed video. This minimal overhead between iterations facilitated a broad exploration of alternative styles or configurations. P7 noted, *"I was surprised how easy it was to go from a cozy living room to a futuristic lounge just by updating a single object and changing one line of text."* Such incremental edits, enabling major aesthetic shifts without reconstructing the entire scene, allowed both novices and professionals to refine concepts in real time, reducing the need for complex post-production and encouraging a flexible, experimental design mindset.

Contextual Semantics from Input Videos. Participants noted that the recorded video itself often carries strong semantic cues that guide the V2V model's interpretation beyond just providing a geometric shape. When participants try the example in Figure 12, a seemingly *"fake"* wooden door with animated flames attached to a wall was transformed into a convincingly burning door without any specialized text prompt. This example illustrates how, by visually suggesting a desired effect (e.g., a door on fire), the V2V can generate realistic outcomes that align with user intent. Consequently, users found that careful scene setup, rather than detailed or highly specific prompts, often produced more coherent results. As P4 noted, *"I realized that to create a spooky vibe, I just need to stage the space in a creepy way, then the AI gets it naturally."* This suggests that arranging physical and virtual props can be as effective as text prompts, offering a more intuitive way to guide video generation.

7 Expert Interview

To assess how our approach transforms concept video prototyping across industries, we conducted structured evaluations with three groups of experts: spatial designers (2 interior/architectural designers, E1-E2), MR developers (2 AR/VR engineers, E3-E5), and video creators (2 professional video producers, E6-E7). All participants had at least two years of experience in their respective fields. Each participant completed a 15-minute open-ended task after a 10-minute tutorial, followed by a 30-minute semi-structured interview. The tasks and findings are organized below.

The evaluation followed a structured three-phase approach with the three expert groups. In Phase 1, we introduced participants to VideoCraft through a 10-minute guided tutorial. Phase 2 tasked each group with prototyping a concept video for a real-world scenario, which is similar to the user study. Phase 3 involved 30-minute semi-structured interviews probing workflow efficiency, creative limitations, and cross-industry applicability. This ensured hands-on exploration of the VideoCraft's technical and creative capacities while capturing domain-specific insights.

Facilitating Real-time, On-site Collaborative Design. Traditionally, interactions between spatial designers and clients rely heavily on verbal descriptions, with designers producing concepts that clients only view after a delay. Experts (E1, E2) have pointed out the inefficiency of this workflow, noting frequent issues with *"delayed feedback," "misinterpretation,"* and *"unclear communication."* In contrast, VideoCraft allows designers to initiate design activities directly during site visits, immediately visualizing design ideas within actual physical environments. More importantly, VideoCraft also allows clients to wear the MR headset themselves and engage in hands-on creation, offering a *"first-person understanding of spatial flow"* and the ability to experiment with *"furniture placement"* and *"space planning"* from their own perspective. Experts noted that this form of active participation fosters *"mutual understanding"* and supports more *"precise dialogue"* between stakeholders, effectively bridging the gap between professional ideation and client expectation.

Expanding the Preset Library to Unlock Creative Potential. While the system offers a range of preset assets for prototyping, experts (E2–E6) expressed a strong desire for a more extensive and diverse database of virtual content. Several (E3, E6) noted that the current asset library feels *"limited"* and constrains the creative exploration of spatial and narrative ideas. Experts emphasized the importance of access to a wider variety of *"common furniture"* (E2), *"outdoor scenery elements for windows"* (E5), animated *"characters"* and *"motion assets"* (E6), and dynamic effects such as *"flames"* and *"water"* (E3, E5). Moreover, E4 highlighted the need for *"more interactive objects,"* while E3 and E4 even envisioned integrating *"game-like elements"* into the design process. Experts believe that a richer asset pool combined with semantic and geometry-aware layers would greatly broaden the design space, enabling users to construct more elaborate, expressive, and contextually rich concept videos.

Dealing with the Instability of AI Generation. Experts (E1, E5) highlighted that the primary challenge in live MR workflows is not the generation time itself, but the *"unpredictability"* and *"instability"* of AI outputs. Often, the first result is unsatisfactory, requiring *"multiple attempts"* (E1) to reach a usable outcome. This trial-and-error process can interrupt the flow of real-time interaction and

lead to *"awkward pauses"* in collaborative scenarios (E5). Such unpredictability remains a limitation for systems aiming to support seamless, interactive authoring. To mitigate this, E4 proposed incorporating mechanisms like *"fixed random seeds"* once a satisfactory result is achieved, ensuring consistency in subsequent generations and reducing redundancy in the creative process.

8 Discussion

8.1 Limitations

Misinterpretation of MR Scene. Based on our previous experiments, V2V performs well when the modifications made in MR are structurally and stylistically plausible, particularly in relatively simple and static environments. However, our current workflow occasionally produces results that are misaligned with user expectations, leading to two key failure cases. The first issue involves over-uniform stylization in heterogeneous scenes. When a static scene contains distinct target styles in different areas, the model may incorrectly apply a uniform style across all elements. Although semantic masks can partially mitigate this by isolating regions, manual refinement remains necessary. This problem becomes even more pronounced in dynamic scenes where the MR scene style evolves throughout a video. Currently, the V2V generation algorithm cannot handle such transformations effectively, resulting in abrupt style shifts. For instance, as the previous example shown in Figure 11, when transitioning from an indoor scene to an outdoor environment, the model fails to maintain a consistent outdoor style throughout the sequence. This issue is further exacerbated when complex dynamic physical elements (e.g., moving people) are present, often leading to morphing artifacts. The second issue relates to spatial misperception in modified geometries. After applying layer-based geometry modifications, such as adding virtual structures, viewpoint changes may cause the V2V model to misinterpret the 3D space as a 2D plane (Figure 16). This disrupts spatial coherence, leading to artifacts like virtual walls appearing flattened when the camera moves.

Figure 16: Example of MR scene misinterpretation. Left: Virtual objects are placed on the table, and a layer is created as a transitional boundary for geometry modification on the white wall to show a virtual wild scene. Right: The generated result using the text prompt *"wooden and magical style"* misinterprets the virtual scene as a magic circle, deviating from the user's intended design.

Spatial Precision of Layer-Based Editing. Although the layer mechanism is designed to support localized and fine-grained editing, its spatial precision remains limited in practice, particularly in visually complex scenes. When using layers as style modification, densely packed objects or irregular geometries may cause the layer shape to inadvertently include unintended elements, especially small decorative items or background clutter. This limitation arises from the fact that segmentation is performed on 2D video frames rather than based on a full 3D reconstruction of the physical environment. Without volumetric understanding, the system cannot ensure accurate object-level isolation, particularly near depth discontinuities or occluded regions.

A similar constraint applies to geometry modification tasks. Since our implementation relies on the built-in scene scanning capabilities of the Quest 3 headset, the resulting spatial mesh is typically low in resolution and geometric fidelity. As a result, geometry layers applied to intricate or irregular surfaces may lead to visual artifacts, imprecise boundaries, or failure to preserve critical structural features.

8.2 Future Work

Support for More Platforms. As a research prototype, our system is built on a mixed reality platform using head-mounted displays (HMDs), which provide intuitive 3D interaction and immersive visual experiences, closely matching how people naturally interact with physical space. However, such hardware remains expensive and less accessible to general users. In future work, we plan to extend our system to more widely available platforms, such as smartphone or tablet-based AR.

Support for 3D Representations of Physical Space. We currently use the MR environment as input without performing full 3D reconstruction, and users create virtual content directly in mixed reality. While this approach ensures responsiveness and visual quality, it limits editing flexibility. In future work, we plan to incorporate 3D reconstruction to support more precise spatial editing and viewpoint manipulation [43, 44, 54] and to experiment with 3D style transfer algorithms [47]. On the output side, the generated 2D videos effectively convey design intent but lack spatial interactivity. We aim to explore converting generated videos into 3D representations, using formats such as meshes [36] or 3D Gaussian Splatting [35] for improved visual quality and editability.

Support for More Content Creation Features. In this work, we do not focus on comprehensive content creation within MR, but rather on the overall workflow and modifications to the physical environment. To simplify virtual content placement, we provide a set of predefined presets. However, integrating richer VR content creation tools, such as custom shape modeling [28, 53], animation authoring [19, 28], virtual scene construction, or event-driven interactions [57], could greatly enhance creative freedom. These types of content would provide more diverse inputs for video-to-video (V2V) transformation. Future work could explore how videos generated from such authored content differ in style, structure, and narrative richness.

Improving V2V Models for Local Control with MR Pre-Staging. While our work primarily focuses on MR-based interaction and spatial editing, we do not deeply explore improvements to the V2V generation model itself. Existing methods typically rely on 2D conditions for localized control [21]. By introducing an MR pre-stage, our system allows users to more intuitively define 3D spatial conditions. This opens opportunities for future work to enhance V2V

algorithms, such as incorporating 3D shape embeddings directly into the generation process. Additionally, more sophisticated local prompt design could be explored to further improve controllability.

9 Conclusion

In this paper, we presented VideoCraft, a mixed reality-empowered video generation workflow with spatial layer editing for concept video creation. Our approach bridges the gap between embodied spatial interaction and high-fidelity generative content. While MR facilitates intuitive, in-situ spatial reasoning during early design exploration, the V2V models generate visually rich outputs to communicate design intent. Our spatial layer enhances creative control by supporting localized editing through intuitive 3D interaction. By embedding editable spatial layers directly into the authoring process, VideoCraft supports a seamless, spatially grounded, and iterative prototyping workflow, advancing the potential of MR-enhanced video creation for creative professionals and amateurs.

We began with a formative study using a tech-probe, placing virtual objects in MR and generating concept videos via V2V, to understand current limitations and inform the improvement of the workflow. Based on the resulting insights, we introduced a spatial layer-based mechanism that allows users to define and manipulate editable regions in physical environments. These layers enable localized style and geometric modifications, offering greater creative control during video generation. The results of our user study and expert interview confirm that VideoCraft is a controllable and flexible workflow supporting efficient concept video creation in physical space.

Acknowledgments

This work was partially supported by Guangzhou Basic Research Scheme #2024A04J4229 and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things #2023B1212010007.

References

[1] AutoDest. 2025. CAD Software | 2D and 3D Computer-Aided Design. https://www.autodesk.com/.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127 [cs.CV] https://arxiv.org/abs/2311.15127

[3] Blender. 2024. Blender: A 3D Modeling and Rendering Package. http://www.blender.org.

[4] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2024. Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning. arXiv:2305.13840 [cs.CV] https://arxiv.org/abs/2305.13840

[5] Yutao Chen, Xingning Dong, Tian Gan, Chunluan Zhou, Ming Yang, and Qingpei Guo. 2023. Eve: Efficient Zero-Shot Text-Based Video Editing With Depth Map Guidance and Temporal Consistency Constraints. *arXiv preprint arXiv:2308.10648* (2023).

[6] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and Track Anything. arXiv:2305.06558 [cs.CV] https://arxiv.org/abs/2305.06558

[7] Zheng-Jun Du, Kai-Xiang Lei, Kun Xu, Jianchao Tan, and Yotam Gingold. 2021. Video Recoloring via Spatial-Temporal Geometric Palettes. *ACM Trans. Graph.* 40, 4, Article 150 (July 2021), 16 pages. doi:10.1145/3450626.3459675

[8] Adobe After Effect. 2025. Adobe After Effects - Motion graphics software. https://www.adobe.com/products/aftereffects.html.

[9] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. 2024. Ccedit: Creative and Controllable Video Editing via Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6712–6722.

[10] Yuwei g Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2025. SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 330–348.

[11] Ankit Gupta, Maneesh Agrawala, Brian Curless, and Michael Cohen. 2014. Motionmontage: A System to Annotate and Combine Motion Takes for 3D Animations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2017–2026. doi:10.1145/2556288.2557218

[12] Robert Held, Ankit Gupta, Brian Curless, and Maneesh Agrawala. 2012. 3D Puppetry: A Kinect-based Interface for 3D Animation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 423–434. doi:10.1145/2380116.2380170

[13] Nisha Huang, Yuxin Zhang, and Weiming Dong. 2024. Style-A-Video: Agile Diffusion for Arbitrary Text-Based Video Style Transfer. *IEEE Signal Processing Letters* 31 (2024), 1494–1498. doi:10.1109/LSP.2024.3398538

[14] Hiroki Kaimoto, Kyzyl Monteiro, Mehrad Faridan, Jiatong Li, Samin Farajian, Yasuaki Kakehi, Ken Nakagaki, and Ryo Suzuki. 2022. Sketched Reality: Sketching Bi-Directional Interactions Between Virtual and Physical Worlds with AR and Actuated Tangible UI. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 12 pages. doi:10.1145/3526113.3545626

[15] Mohamed Kari, Tobias Grosse-Puppendahl, Luis Falconeri Coelho, Andreas Rene Fender, David Bethge, Reinhard Schütte, and Christian Holz. 2021. TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 69–79. doi:10.1109/ISMAR52148.2021.00021

[16] KLING. 2025. KLING AI: Next-Generation AI Creative Studio. https://klingai.com/.

[17] Germán Leiva and Michel Beaudouin-Lafon. 2018. Montage: A Video Prototyping System to Reduce Re-Shooting and Increase Re-Usability. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 675–682. doi:10.1145/3242587.3242613

[18] Germán Leiva, Cuong Nguyen, Rubaiat Habib Kazi, and Paul Asente. 2020. Pronto: Rapid Augmented Reality Video Prototyping Using Sketches and Enaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376160

[19] Boyu Li, Linping Yuan, Zhe Yan, Qianxi Liu, Yulin Shen, and Zeyu Wang. 2024. AniCraft: Crafting Everyday Objects as Physical Proxies for Prototyping 3D Character Animation in Mixed Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 99, 14 pages. doi:10.1145/3654777.3676325

[20] David Lindlbauer, Jörg Mueller, and Marc Alexa. 2017. Changing the Appearance of Real-World Objects By Modifying Their Surroundings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3954–3965. doi:10.1145/3025453.3025795

[21] Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, and Jiaya Jia. 2024. Generative Video Propagation. arXiv:2412.19761 [cs.CV] https://arxiv.org/abs/2412.19761

[22] Meta. 2025. Meta Quest 3: Mixed Reality VR Headset. https://www.meta.com/quest/quest-3/.

[23] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 267–277. doi:10.1145/3126594.3126659

[24] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. Vremiere: In-Headset Virtual Reality Video Editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5428–5438. doi:10.1145/3025453.3025675

[25] OpenAI. 2025. Sora. https://openai.com/sora/.

[26] Selcen Ozturkcan. 2021. Service Innovation: Using Augmented Reality in the IKEA Place App. *Journal of Information Technology Teaching Cases* 11, 1 (2021), 8–13. doi:10.1177/2043886920947110

[27] Pika. 2025. The idea-to-video platform that sets your creativity in motion. https://pika.art.

[28] Quill. 2016. Quill on Oculus Rift | Rift VR Games. https://quill.art/.

[29] A.A.G. Requicha and H.B. Voelcker. 1985. Boolean Operations in Solid Modeling: Boundary Evaluation and Merging Algorithms. *Proc. IEEE* 73, 1 (1985), 30–44. doi:10.1109/PROC.1985.13108

[30] Rhino. 2025. Rhino - Rhinoceros 3D. https://www.rhino3d.com/.

[31] Runaway. 2025. Runway | Tools for human imagination. https://runwayml.com/.

[32] Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. 2017. Movie Editing and Cognitive Event Segmentation in Virtual Reality Video. *ACM Trans. Graph.* 36, 4, Article 47 (July 2017), 12 pages. doi:10.1145/3072959.3073668

[33] Zichun Shao, Junming Chen, Hui Zeng, Wenjie Hu, Qiuyi Xu, and Yu Zhang. 2024. A New Approach to Interior Design: Generating Creative Interior Design Videos of Various Design Styles From Indoor Texture-Free 3D Models. *Buildings* 14, 6 (2024), 1528.

[34] Yulin Shen, Yifei Shen, Jiawen Cheng, Chutian Jiang, Mingming Fan, and Zeyu Wang. 2024. Neural Canvas: Supporting Scenic Design Prototyping by Integrating 3D Sketching and Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1056, 18 pages. doi:10.1145/3613904.3642096

[35] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. 2024. Realmdreamer: Text-Driven 3D Scene Generation With Inpainting and Depth Diffusion. *arXiv preprint arXiv:2404.07199* (2024).

[36] Adalberto L. Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality: Using the Physical Environment to Design Virtual Reality Experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. doi:10.1145/2702123.2702389

[37] Ryo Suzuki, Rubaiat Habib Kazi, Li-yi Wei, Stephen DiVerdi, Wilmot Li, and Daniel Leithinger. 2020. RealitySketch: Embedding Responsive Graphics and Visualizations in AR through Dynamic Sketching. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 166–181. doi:10.1145/3379337.3415892

[38] Wai Tong, Kento Shigyo, Lin-Ping Yuan, Mingming Fan, Ting-Chuen Pong, Huamin Qu, and Meng Xia. 2025. VisTellAR: Embedding Data Visualization to Short-Form Videos Using Mobile Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 31, 3 (2025), 1862–1874. doi:10.1109/TVCG.2024.3372104

[39] Valem Tutorials. 2024. *How to make a Portal Effect in Mixed Reality - Unity Tutorial*. Youtube. https://www.youtube.com/watch?v=BXLRprBFfNo

[40] Unity. 2024. Unity Real-Time Development Platform | 3D, 2D, VR & AR Engine. https://unity.com/.

[41] Vuforia Engine. 2025. Vuforia Engine | Create AR Apps and AR Experiences. https://vuforia.com.

[42] Luozhou Wang, Yijun Li, Zhifei Chen, Jui-Hsien Wang, Zhifei Zhang, He Zhang, Zhe Lin, and Yingcong Chen. 2025. TransPixeler: Advancing Text-to-Video Generation with Transparency. arXiv:2501.03006 [cs.CV] https://arxiv.org/abs/2501.03006

[43] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2021. DistanciAR: Authoring Site-Specific Augmented Reality Experiences for Remote Environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 411, 12 pages. doi:10.1145/3411764.3445552

[44] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2023. PointShopAR: Supporting Environmental Design Prototyping Using Point Cloud in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. doi:10.1145/3544548.3580776

[45] Zhijie Xia, Kyzyl Monteiro, Kevin Van, and Ryo Suzuki. 2023. RealityCanvas: Augmented Reality Sketching for Embedded and Responsive Scribble Animation Effects. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 115, 14 pages. doi:10.1145/3586183.3606716

[46] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2025. Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance. *IEEE Transactions on Visualization and Computer Graphics* 31, 2 (2025), 1526–1541. doi:10.1109/TVCG.2024.3365804

[47] Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. 2024. DreamSpace: Dreaming Your Room Space with Text-Driven Panoramic Texture Propagation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 650–660. doi:10.1109/VR58804.2024.00085

[48] Hui Ye, Kin Chung Kwan, Wanchao Su, and Hongbo Fu. 2020. ARAnimator: In-situ Character Animation in Mobile AR with User-defined Motion Gestures. 39, 4, Article 83 (aug 2020), 12 pages. doi:10.1145/3386569.3392404

[49] Emilie Yu, Kevin Blackburn-Matzen, Cuong Nguyen, Oliver Wang, Rubaiat Habib Kazi, and Adrien Bousseau. 2023. VideoDoodles: Hand-Drawn Animations on Videos with Scene-Aware Canvases. *ACM Trans. Graph.* 42, 4, Article 54 (July 2023), 12 pages. doi:10.1145/3592413

[50] Emilie Yu, Fanny Chevalier, Karan Singh, and Adrien Bousseau. 2024. 3D-Layers: Bringing Layer-Based Color Editing to VR Painting. *ACM Trans. Graph.* 43, 4, Article 101 (July 2024), 15 pages. doi:10.1145/3658183

[51] Zhongyuan Yu, Daniel Zeidler, Victor Victor, and Matthew Mcginity. 2023. Dynascape: Immersive Authoring of Real-World Dynamic Scenes with Spatially Tracked RGB-D Videos. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. doi:10.1145/3611659.3615718

[52] Lin-Ping Yuan, Feilin Han, Liwenhan Xie, Junjie Zhang, Jian Zhao, and Huamin Qu. 2025. "You'll Be Alice Adventuring in Wonderland!" Processes, Challenges, and Opportunities of Creating Animated Virtual Reality Stories. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 193, 21 pages. doi:10.1145/3706598.3714257

[53] Lin-Ping Yuan, Boyu Li, Jindong Wang, Huamin Qu, and Wei Zeng. 2024. Generating virtual reality stroke gesture data from out-of-distribution desktop stroke gesture data. In *IEEE Conference Virtual Reality and 3D User Interfaces*. IEEE, 732–742.

[54] Ya-Ting Yue, Yong-Liang Yang, Gang Ren, and Wenping Wang. 2017. SceneCtrl: Mixed Reality Enhancement via Efficient Scene Editing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 427–436. doi:10.1145/3126594.3126601

[55] Hongbo Zhang, Pei Chen, Xuelong Xie, Chaoyi Lin, Lianyan Liu, Zhuoshu Li, Weitao You, and Lingyun Sun. 2024. ProtoDreamer: A Mixed-prototype Tool Combining Physical Model and Generative AI to Support Conceptual Design. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 97, 18 pages. doi:10.1145/3654777.3676399

[56] Haijun Zhang, Xiangyu Mu, Han Yan, Lang Ren, and Jianghong Ma. 2023. A Survey of Online Video Advertising. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13, 2 (2023), e1489.

[57] Lei Zhang and Steve Oney. 2020. FlowMatic: An Immersive Authoring Tool for Creating Interactive Scenes in Virtual Reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 342–353. doi:10.1145/3379337.3415824

[58] Qian Zhou, David Ledo, George Fitzmaurice, and Fraser Anderson. 2024. TimeTunnel: Integrating Spatial and Temporal Motion Editing for Character Animation in Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 101, 17 pages. doi:10.1145/3613904.3641927

[59] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. 2023. Propainter: Improving Propagation and Transformer for Video Inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10477–10486.