CollageVis: Rapid Previsualization Tool for Indie Filmmaking using Video Collages

Hye-Young Jo
Chung-Ang University
Seoul, Republic of Korea
hyeyoungjo@cau.ac.kr

Ryo Suzuki
University of Calgary
Calgary, Canada
ryo.suzuki@ucalgary.ca

Yoonji Kim
Chung-Ang University
Seoul, Republic of Korea
yoonjikim@cau.ac.kr

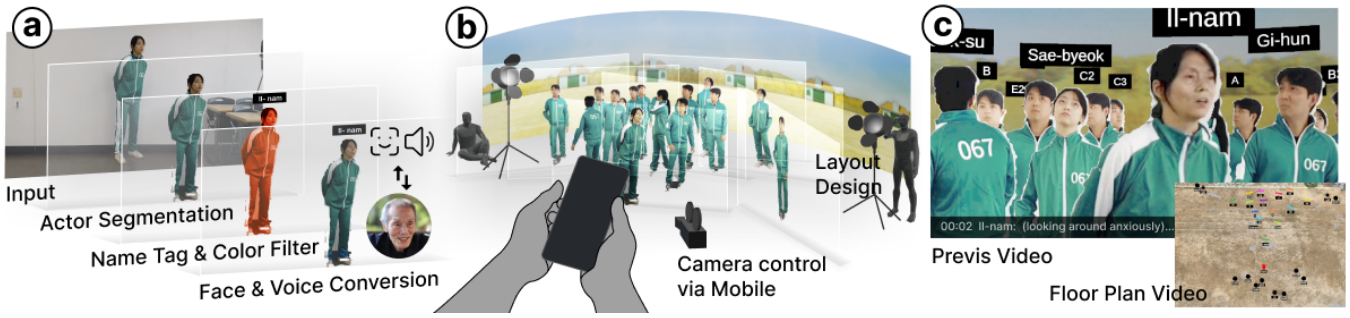


Figure 1: *CollageVis* composites a film previs through two user interfaces: ⓐ collage board and ⓑ virtual stage. The collage board segments actors from the input videos and assigns roles by tagging names, applying color filters, and changing faces and voices. The virtual stage places the video layers, staff, and lighting equipment in 3D space and allows the user to record shots using a mobile as a proxy for the virtual camera. The ⓒ output of *CollageVis* includes previs video and floor plan video.

ABSTRACT

Previsualization, previs, is essential for film production, allowing cinematographic experiments and effective collaboration. However, traditional previs methods like 2D storyboarding and 3D animation require substantial time, cost, and technical expertise, posing challenges for indie filmmakers. We introduce *CollageVis*, a rapid previsualization tool using video collages. *CollageVis* enables filmmakers to create previs through two main user interfaces. First, it automatically segments actors from videos and assigns roles using name tags, color filters, and face swaps. Second, it positions video layers on a virtual stage and allows users to record shots using mobile as a proxy for a virtual camera. These features were developed based on formative interviews by reflecting indie filmmakers' needs and working methods. We demonstrate the system's capability by replicating seven film scenes and evaluate the system's usability with six indie filmmakers. The findings indicate that *CollageVis* allows more flexible yet expressive previs creation for idea development and collaboration.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642575

CCS CONCEPTS

• **Human-centered computing** → **Graphical user interfaces**; **Visualization toolkits**.

KEYWORDS

previsualization, storyboard, indie filmmaking

ACM Reference Format:
Hye-Young Jo, Ryo Suzuki, and Yoonji Kim. 2024. *CollageVis*: Rapid Previsualization Tool for Indie Filmmaking using Video Collages. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642575

1 INTRODUCTION

Previsualization, also known as *previs* [1], is essential for film production. It helps filmmakers conceptualize film sequences beforehand, allowing them to experiment with different staging and art directions. By quickly translating a director's idea into visual images, previs facilitates effective collaboration and discussion with the production team. For example, by leveraging 2D storyboards and 3D character animations [8], filmmakers can explore diverse framing techniques, camera movements, and editing, fostering seamless collaboration among film crews to establish a unified vision [43].

However, creating previs requires substantial costs, time, and effort, making it difficult for indie filmmakers to create it, as they do not have enough budget, technical expertise, and human resources. For instance, creating a 2D storyboard requires drawing skills, including a deep understanding of human anatomy, which cannot be acquired in a short period of time. Similarly, making 3D

[1] https://en.wikipedia.org/wiki/Previsualization

animations requires diverse skill sets, such as modeling, rigging, and animation in 3D Digital Content Creation (DCC) tools like Maya. Thus, many film directors need to hire previs artists and constantly communicate with them to convey their visions, which is costly and time-consuming.

To solve this problem, researchers have developed automatic text-to-graphic generation tools and virtual reality systems to make previs without drawing or 3D DCC tool skill sets. For instance, Chen et al. [7] developed a previs tool that converts scripts into storyboards using reference images from a movie database, and Kim et al. [22] generated 3D character animations from scripts by synthesizing relevant character gestures from an animation library. Other researchers [13, 20, 36] investigated Virtual Reality (VR) for a scene simulation using ready-made 3D characters. These approaches show great potential in reducing previsualization task loads. However, automatic previs generation heavily relies on pre-existing datasets, which may limit directors' creativity and require effort in modifying the script to obtain the desired previs results. Also, while VR systems offer more immersive environments and natural body interaction than desktop tools, they demand additional investments in purchasing and adapting to VR equipment, and 3D characters in previs often lack expressiveness due to static facial expressions.

In contrast, according to our formative interview, indie filmmakers prefer rough video recording and editing rather than 2D storyboards or 3D animations. For example, we learned that they often simply visit the prospective shooting location and record *test videos*, as this approach is more intuitive for filmmakers, given their familiarity with handling cameras. These test videos not only give a closer representation of the final film but also facilitate planning the spatial positioning of actors, crew, and cameras. Still, making test videos requires film crews to meet in person, which causes scheduling issues and delays. Directors also need to put extra time into composing video clips. Hence, a tool for creating test videos without physical constraints and cumbersome video editing would benefit indie filmmakers.

In this paper, we present *CollageVis*, a system that enables indie filmmakers to create previs using video collages. With *CollageVis*, users can rapidly composite video clips, lay them out on the virtual stage as 2D planes, and explore various camera movements. Our system draws inspiration from two video editing techniques, *clone effect* [2] and *deepfake* [3], to create video collages without multiple actors and crews. The *CollageVis* comprises two user interfaces: collage board and virtual stage (Figure 1.ⓐ, ⓑ). The collage board automatically segments the actor from the video, making it a cut-out video layer like a paper doll. Users can gather layers of actors on a single board to compose them into scenes. The collage board offers actor differentiation features, such as name tagging, color filters, face swapping, and voice modulation, to differentiate a single actor with distinct roles across multiple layers. Though the collage board is enough for static shots, for more complex scenes where spatial positioning and camera movement are crucial, users can transition to the virtual stage for layout exploration. In this interface, users can place the video layers, mock-up staff, and lighting equipment in the scene for effective shot planning. Then, users can experiment with various camera transitions using always camera-facing 2D video planes and 3D environments composed of panorama images or scanned/modeled 3D data of potential shooting locations.

These features were developed based on formative interviews by reflecting indie filmmakers' needs and working methods. We also demonstrate the system's capability by replicating seven film scenes that feature different filming aspects, such as the number of actors, different shooting environments, and various camera-work sequences. We also evaluate the system's usability with six indie filmmakers. The findings indicate that *CollageVis* allows more flexible yet expressive previs creation for idea development and collaboration.

Finally, our work provides the following contributions:

- A formative study with six indie filmmakers to identify the difficulties of using existing previs tools in indie filmmaking and deduce the design goals encompassing the user's practical needs.
- A system, *CollageVis*, that supports filmmakers to quickly create video storyboards by compositing short video clips in real-time and enhancing the communicative aspect of previs through actor differentiation filters and a floor plan.
- A cinematic simulation capability test of the system by replicating seven different scenes in existing films.
- A user study with six indie filmmakers to review the system's usability and potential usage.

2 RELATED WORK

Our work builds upon the previsualization research by enabling the rapid creation of video storyboards using real-time video compositing. This section reviews the prior work supporting 2D and 3D previs creations and outlines the inspiring HCI research using video compositing techniques.

2.1 2D Previs Tools

A storyboard is a traditional previs format visualizing the position and action of characters in 2D frames. There are many commercial storyboard tools like *Storyboard Pro* [2] providing general support for sketching and animating sketches. However, acquiring relevant drawing skills and cultivating a sense of cinematic composition takes considerable time and effort [7]. So, there are two approaches to enable the rapid creation of 2D storyboards: drawing-supporting and automatic storyboarding tools.

Drawing-supporting Tools: Researchers developed tools that support drawing focusing on specific elements, such as facial expression [46], aging features [28], and animating clothes [27]. Shi et al. [46], for example, developed an automatic facial expression creation tool that generates six face variations expressing different emotions when given a user-drawn neutral face. Although these tools speed up the sketching process, they still require the user to have an anatomical understanding for initial drawing and spend a lot of time on labor-intensive drawing processes.

[2]Clone effect in video editing refers to a technique where multiple instances of an actor appear within the same frame to depict a character interacting with a doppelganger.
[3]Deepfake technology indicates hyper-realistic videos created using deep learning techniques that manipulate an individual's image and audio to resemble someone else [49]

Automatic Storyboarding Tools: On the other hand, some researchers [7, 19] proposed to leverage the existing movie images that already contain rich cinematographic language. Jo et al.[19] developed an AI agent generating sketches from the user's story description by referencing movie trailers. Similarly, Chen et al. [7] proposed *inspire-and-create* framework, which retrieves relevant images from cinematic images according to the input script. Recently, AI-infused storyboard tools like *Krock.ai* [24] that convert text into images have been commercialized. The user can erase unimportant parts or apply a consistent cartoon style for visual consistency. However, despite their convenience, automatic 2D storyboard tools come with inherent limitations. These tools heavily rely on pre-existing image data; consequently, they are unable to generate entirely new visual content, restricting the user's creative freedom. Also, they require users to put effort into modifying their script to create the *right* images.

2.2 3D Previs Tools

A computer-generated 3D animation is a predominant form of previs, especially in big-budget movies [50] because of its flexibility and technical capability to test in the virtual environment. Despite the positives, generating 3D previs necessitates more experts than 2D storyboarding as it requires broad expertise in various 3D DCC tools [36]. Filmmakers need to create virtual scenes, model characters, and animate them using complex key animation techniques using tools like Maya and Blender, which is difficult for non-technical filmmakers [13]. Thus, some of the recent industry tools like Previs Pro [29] and CineTracer [51] aim to simplify certain manual processes by providing 3D modeling and animation libraries. On the other hand, researchers tackled this problem by translating physical mockups or scripts to 3D animations or simulating pre-made 3D assets directly in virtual reality.

3D Previs with Physical Mockups: Some researchers [16, 17, 33, 37, 47] suggested using tangible objects such as dolls and legos and converting the user's physical rehearsals with these objects to the 3D environment. Shin et al.[47] developed an AR system that converts paper items into 3D animations. Also, Horiuchi et al.[17] proposed a tabletop interface that reflects the physical interaction with the dolls to the virtual scene in real time. The usage of these physical mockups enables non-technical filmmakers to intuitively create 3D previs; however, they have a limited range of expressions and require time and cost to prepare tangible objects beforehand.

Automatic 3D Animations: Other researchers [10, 21, 22, 30] developed automation tools that generate 3D previs based on a parameterized storyboard or script by logically synthesizing 3D character animations. Kim et al.[22] developed *ASAP* system that decomposes a script into different paragraphs (e.g., action, character, and dialogue) and composes 3D character actions in the virtual environment. Similarly, Marti et al.[32] suggested visualizing characters' actions in 3D in their script writing tool, *CARDINAL*, to support visual understanding of the authored story. These 3D previs automation tools effectively reduce the time spent in 3D character animation. However, they require a complete script/storyboard to begin with, support limited character animations based on the predetermined animation library, and lack expressiveness compared to videos with real actors.

3D Previs in Virtual Reality: Meanwhile, several researchers [13, 20, 36] proposed using virtual reality to directly manipulate and simulate 3D scenes in the game engine like *Unreal*. The value of the real-time game engine's interactive control over traditional animation has been recognized by earlier researchers [18, 42]. But, researchers like Muender et al.[36] and Galvane et al.[14] later incorporated VR technologies into the game engine and suggested the concept of a user standing in multiple roles (e.g., director, photographer, editor) to single-handedly make an entire 3D previs. The sequel research also explored using VR for technical previs, such as controlling the configuration of camera rigs [13] and acting practice [20]. Despite the advantage of direct manipulation in these works, they require a VR system with a head-mounted display and motion capture gears, causing a financial burden to indie filmmakers. Moreover, virtual avatars with static facial expressions often fail to deliver the actor's emotions.

2.3 Live Video Compositing

Video compositing is combining multiple source video footage into a single integrated video. It is commonly used in the feature film production's visual effects (VFX) to create a convincing visual narrative to viewers [3]. The final composite includes various elements that are shot separately and layered on top of each other in a specific order with respective alpha mattes (i.e., masks). The VFX compositors use advanced video editing tools such as Nuke to pull an animated mask from video footage in the post-production process.

Researchers have attempted to do video compositing in real-time for dynamic storytelling [9, 11, 12] and interactive Mixed Reality content authoring [25, 38, 40, 41]. *Improv Remix* [11] applied live video compositing to modern theatrical improvisation using Kinect and projection wall and *LACES* [12] system uses traditional compositing techniques (e.g., rotoscoping, and chroma keying) to live streaming videos for rapid casual video editing. Similarly, Nebeling et al.[40] and Müller et al.[38] used live video compositing techniques in AR prototyping research. These AR prototyping systems recorded a miniature clay, used a chroma key to acquire the alpha matte, and triggered the clay animation on the camera view to simulate the AR experience.

In summary, in contrast to the prior works, *CollageVis* does not require drawing proficiencies nor restrict the user's creativity to the preexisting dataset, enabling the user to generate novel videos. In addition, *CollageVis* does not entail the preparation of physical mockups, a complete script, or virtual reality setups. Furthermore, using video has an advantage over the traditional previs method (2D storyboarding/3D animation) as it already has a familiar workflow to filmmakers, and it can deliver the actor's rich expressions as it is. Lastly, the *CollageVis* supports unique features that are exclusively designed for the film previs on top of live video compositing techniques. For example, *CollageVis* has actor classification features (e.g., name tag, color filter, face swap, voice filter) to facilitate communication between film crews. It also takes the composite to the

virtual stage, enabling further experimentation of cinematography with a virtual camera.

3 FORMATIVE STUDY & DESIGN GOALS

Although each of the previsualization tools described in Section 2 holds intrinsic value, it remains uncertain whether these tools are practically employed in indie filmmaking. There may be difficulties directors encounter when implementing these tools within the specific context of indie films, where there is a heavy burden on time and budget management.

3.1 Study Design

To better understand the common practices and challenges of indie film previs and the need for a new tool, we conducted semi-structured interviews with six indie film directors aged 26-41 (M : 34.67, SD : 6.47, 1 female and 5 males) who have worked in the film industry for 2-17 years (M : 8.00, SD : 5.48). For the previs, most of them preferred 2D storyboarding (P2, P4, P5, P6) or test video (P2, P3, P5, P6), and one participant (P1) usually skipped a detailed storyboarding and just drew a floor plan per scene.

During the interviews, we asked them to describe a past previs experience, its purpose, and which tool they selected and why. Also, we showed them the images and videos of various existing previs tools—2D storyboarding tools, 3D animation tools, VR previs systems, AI previs (script-to-storyboard or script-to-3D-animation) tools—and asked their opinion on applying them to their films. The interviews lasted an hour online (Zoom), and participants were compensated 30 USD in local currency.

3.2 Interview Findings

With the participants' consent, we recorded all interviews, transcribed them, and conducted a thematic analysis. The five emerging themes we identified are as follows:

Theme 1. Previs Is Important but Often Neglected Due to Time Constraints. All participants agreed that previs is valuable even in indie films. It offers the chance to delve deeper into directing techniques (P1, P2, P5), enables the film crew to better grasp the director's intentions (P2, P3, P4), and facilitates efficient communication among the crew members (P1-P6). However, they complained about the considerable time and effort involved in previs creation. *P2: "Personally, it takes me at least a month to create a storyboard for a short film. It is very time-consuming and stressful, especially because I'm not good at drawing."* They said that many indie filmmakers cut back on or omit previs because of time constraints and extremely tight budgets. *P1: "I'm pretty good at drawing, but I just don't have enough time."*, *P3: "I'd like to hire a storyboard artist, but I don't have the money or time to go back and forth with multiple revisions."* In general, all participants expressed their willingness to engage in previs as long as it doesn't entail a lengthy production process.

Theme 2. 3D Previs Is Not Always Necessary, but a Planning Layout in 3D Space Is Essential. When asked about their experience with 3D previs, two participants (P2, P3) said they had the experience, but it was for music video and animation, not film, and all participants expressed a preference for 2D storyboards over 3D previs. They said that creating 3D previs is not only a difficult and lengthy process (P3) but also unnecessary for their films (P1-P5). *P1: "I'd consider 3D simulation if there is a complex scene with multiple actors or a dynamic scene with a disposable prop. But my work is fairly static."* Nevertheless, since storyboards alone were insufficient to explicitly represent the length of shots, they often made the storyboard into a video reel (i.e., animatics). *P3: "With animatics, the crew can see the duration of each shot and use it as a reference for on-set editing."* Also, instead of making sophisticated 3D previs, participants (P1-P5) drew simple shot planning diagrams to plan the movement of the camera and actors. *P2: "For tricky shots, I sketch a floor plan since storyboard doesn't show how the lighting should be set up or how the camera moves in the actual space."*

Theme 3. The Test Video Is a More Familiar and Clear Previs Method Than Storyboarding, But Lacks Flexibility. Besides hand-drawn storyboards, participants said indie filmmakers often record test videos. This test video was considered superior to the 2D storyboard for various reasons. First, using a camera is more "easy and familiar"(P2) than drawing for non-technical filmmakers. Moreover, as the test video closely resembles the final film, it intuitively conveys crucial information such as spatial and temporal details (P5), actors' motion (P6), and camera movement (P2) to the film crew. In addition, since it features a real actor on set, it enables a head start to the post-production process. *P3: "It is the clearest (previs). It's like a rehearsal to position actors in a frame."* However, compared to storyboarding, which can be done at any time and location without additional human resources, making test videos poses other challenges. It requires coordinating the schedules of actors and staff (P2, P5, P6) on the right day for different weather conditions (P6). *P2: "The biggest pitfall is the cost. In indie films, it's unlikely to have pre-meetings, let alone field test shoots. So, in the last film, I made a test video with a couple of staff instead of real actors (to save cost)."* Also, it takes considerable time and effort to edit the test video footage (P2).

Theme 4. VR Systems Have Issues of Accessibility and Expressiveness. All participants were skeptical about using VR systems for their previs-making for various reasons. First of all, none of the participants had a VR device, so accessibility was lacking at the moment. Second, participants expressed their hesitance towards the 3D interface, finding it rather "intimidating" (P3). *P1: "Working in 3D is not easy. For me, it takes just as much effort as drawing a storyboard." P5: "I've tried VR a couple of times, but it (controller) wasn't easy."* Third, the final result of the VR previs system, which was 3D character animation, lacked expressiveness compared to drawings or actors. *P3: "I don't like (the look and feel of) the result. It lacks something [...] a sort of aura, image, and expression."*

Theme 5. Text-Based AI Systems May Impose Constraints on the Extent of Creative Expression. The participants expressed ambivalent feelings toward the AI systems that instantly convert the script to a 2D storyboard or 3D animation. They thought the automatic generation of previs was extremely efficient, but they were concerned about becoming too reliant on the outcomes generated by AI (P1-P3, P6). *P1: "As painful as the previs-making process can be, it actually helps me refine my vision through trial and error. I'm afraid it's going to make my work unoriginal."* They also pointed out that they cannot create entirely new images because the AI

systems "rely on pre-existing data"(P2, P4). In addition, participants (P3-P6) felt that it would be challenging to guide the AI system to generate images to their liking. *P5: "I'll spend more time fixing it (image). I think this would be good for inspiration, but not for actual previs."* Nevertheless, participants were optimistic about the future of AI-based previs, saying AI could help them "develop their creative vision" (P1) and "reduce previs costs" (P3) as long as it could maintain their creative intention.

3.3 Design Goals

The results of our formative interviews suggest that existing previs tools should incorporate the following Design Goals (DG) to better cater to indie filmmakers' needs.

- DG1. Save costs and efforts.
- DG2. Facilitate the design of spatial layout.
- DG3. Leverage familiar practice and overcome its shortcomings.
- DG4. Utilize readily available devices.
- DG5. Maintain creative intention when incorporating AI.

Though there are various approaches for previs authoring, we decided to attempt to build a new tool that builds upon indie filmmakers' familiar practices of test videos, which has been less explored.

4 *COLLAGEVIS* SYSTEM

CollageVis is a rapid previs tool for indie filmmakers that allows the creation of on-demand test videos (i.e., video collages) using video compositing techniques. It takes multiple test video clips and background images as input and renders a composited outcome with a corresponding floor plan. In this section, we introduce our system *CollageVis*, explain the design rationale behind key features, and provide a detailed authoring walk-through to illustrate the benefits of each feature.

4.1 Overview

Figure 2 shows the *CollageVis* system's overall workflow. *CollageVis* primarily runs on a laptop, with a mobile device serving as a supplementary tool for camera-related functions. The system has two main user interfaces: collage board and virtual stage. In the collage board, the mobile is used as an IP camera, streaming the video to the main laptop interface that composites the video collage (Figure 2.ⓐ, ⓑ). On the other hand, in the virtual stage, the mobile is used as a proxy for the virtual camera, navigating 3D spaces and controlling camera settings (Figure 2.ⓒ). At the same time, the virtual stage laptop interface helps design the overall layout by placing characters, cameras, staff, and lighting equipment and exports the floor plan along with video collages (Figure 2.ⓓ).

Collage board: This interface collects test video clips and composes them in real time. The system shows a live composition of previous video layers and real-time video streams to provide temporal cues for acting. Each video recording is saved as an image sequence and can be moved and trimmed in the timeline pane. Besides video collage, the system generates a script draft by converting audio input to text and composing a dialogue between different characters to use as a subtitle for a video collage. The system provides various actor differentiation filters to distinguish characters; the users can set the character's name using a name tag and apply a color filter to each layer. Also, the user can set the character's gender and change the voice of each video clip. Furthermore, the user can change the face in the video clip to the desired face by uploading an actor's profile picture. All these filters can turned on and off. In addition, users can set the video background by selecting the preset image or adding a new image to the board. Each composition can be rendered into a video collage or can be sent to the next interface, the virtual stage, to explore the spatial aspects.

Virtual stage: To support further exploration in layout and camera movement design, *CollageVis* provides functionality to lay out video layers collected from the collage board in 2.5D space. On the stage, flat 2D video layers stand like paper dolls, along with 3D staff and lighting mockups in the 3D environment. Users can set the scene's time by selecting a spherical sky image from the dropdown menu. Then, the user can change the 3D environment by combining two components: a cylindrical panorama image and 3D data. After defining the scene environment, users can move 2D video layers by dragging them to the desired locations using a mouse, and can add 3D components such as virtual cameras, staff, and lighting mockups to design the overall layout.‘ Finally, using a mobile camera puppet application, the user can move the virtual camera by tilting the mobile. The mobile app supports four camera modes: steady, hand-held, hold, and track. The steady mode moves the virtual camera smoothly, while the hand-held mode reflects the raw motion of the user's hand to the virtual camera. The hold mode acts as a static camera on a tripod, and the track mode simulates the dolly cam effect moving from one position to the other. Additionally, the mobile app has four sliders to adjust the height and simulate the camera's panning, tilting, and zooming effects. The user can export the result as a floor plan and multiple video collages recorded through each virtual camera.

4.2 System Design Rationale

Our system has four key features that are designed to meet the five Design Goals (DG1-5) highlighted in Section 3.3.

Feature 1. Live Video Composition. To decrease the production time (DG1), we designed a live video compositing feature using commodity devices (DG4). This automates data management and reduces the post-editing efforts typically required in traditional test video workflows (DG3).

Feature 2. Character Differentiation Filters. To save the cost of in-person rehearsals and avoid scheduling conflicts, *CollageVis* allows directors to use the same actor for multiple characters (DG1). We designed various actor differentiation filters for each video input, including name tag, coloring, face swap, and voice modulation for clear communication among film crews.

Feature 3. 2.5D Spatial Design. To facilitate spatial exploration of the target scene across time and space (DG2), we designed a 2.5D virtual stage. The 3D environment simulating various space, time, and weather conditions was devised to prevent multiple visits to potential shooting locations. We also support additional floor plan elements, such as the layout of the background space and the

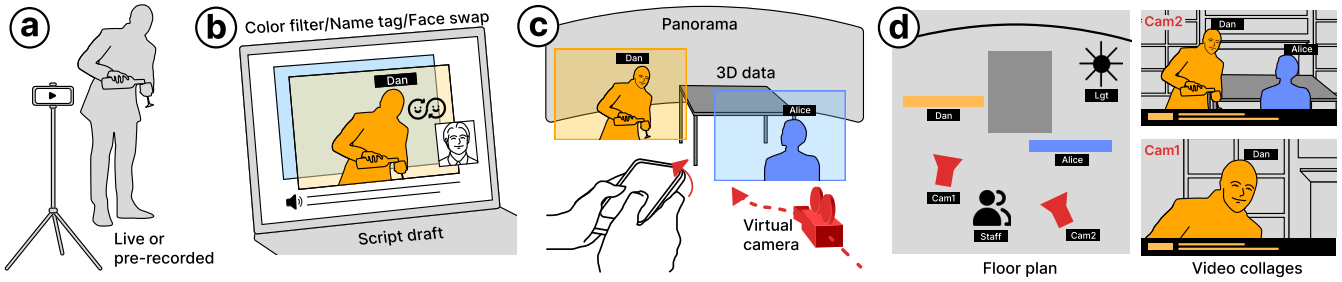


Figure 2: *CollageVis* workflow: ⓐ Record test videos, ⓑ Compose videos and apply actor differentiation filters, ⓒ Layout video clips in virtual space and record shots using mobile camera puppet, ⓓ Export the floor plan and video collage per each virtual camera.

placement of lights and staff, so that each team member can know their position in advance.

Feature 4. Virtual Camera Control with Mobile. In order to give filmmakers creative control over shot framing and camera work (DG5) while allowing easy navigation in the virtual stage, we designed a separate application for camera control (Figure 4.ⓑ), similar to previous works using physical mockups to move virtual characters [17, 33, 37].

4.3 Authoring Walk-through

The system supports both live streaming and pre-recorded video inputs for authoring. The user can upload the pre-recorded video when there is an already casted actor who is in a remote place to avoid scheduling issues and the extra cost of traveling. Here, we hypothesize a more general and familiar way of making test videos: an indie film director (Lucy) recording a staff (Martin) using live streaming mode. They are making a simple conversation scene between two characters, Jack and Alice.

Step 1. Recording Videos While Watching Real-Time Composition. After setting the mobile on a tripod, the staff acts out the character Jack. The director watches the streaming video on the laptop and presses the *Record/Stop* button to save the character layer (Figure 3.ⓐ). The director changes the backdrop to the terrace image. Then, looking at the real-time composition on the collage board interface (Figure 3.ⓑ), the director goes on to record the second layer, and the staff acts out the next character, Alice. After recording all the characters, they watch the video collage together, verify the timing of the conversation, and make adjustments (e.g., moving and trimming the clips) in the timeline view when necessary. Looking at the script draft on the right side, they edit the generated dialogue, adding the details for the production. The director jots down how the actor should express emotion, and the staff adds the necessary props and equipment to prepare for shooting.

Step 2. Differentiating Characters via Name Tags, Color, Face, and Voice Filters. As the staff acted out two characters, the video collage looks like a conversation between twins or doppelgangers. To prevent potential miscommunication among film crews, the director assigns each video layer a character name tag (Jack, Alice) using the *Layer* pane. Moreover, to make it easy to distinguish each

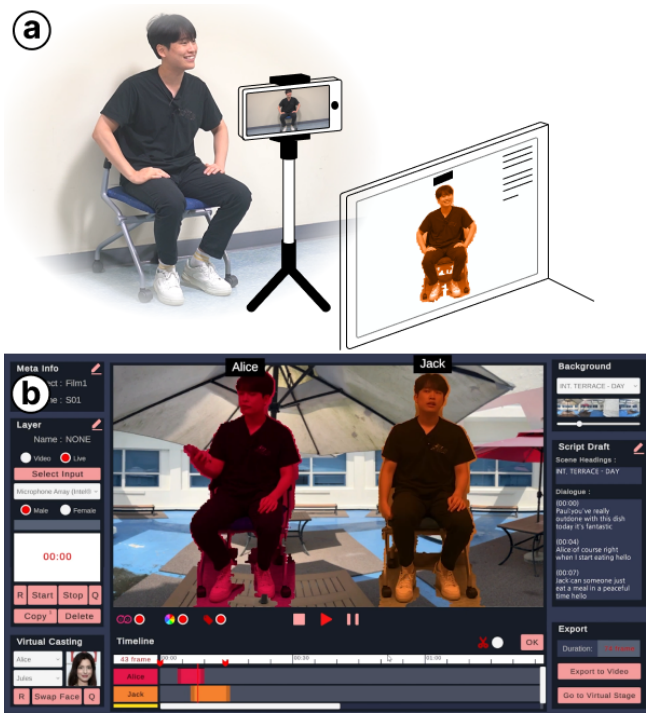


Figure 3: Collage board: ⓐ live compositing setup, ⓑ laptop user interface

role at a glance, even in a full shot, the director turns on the color filter. The system colorizes the Alice layer in red and the Jack layer in orange (Figure 3.ⓑ). Also, for the character Alice, the director set the gender to female to change the audio's pitch higher than the original, as the staff is male pretending to be the female character. Then, the director and staff upload the profile pictures of actors to the system and swap the staff's faces in the video with the actors' to see the look and feel of the different actor combinations in a frame. Since they want to test different layouts for the characters and cameras, they decide to move the composition to the virtual stage interface instead of exporting.

Step 3. Placing Characters, Cameras, Staff, and Lighting Equipment in the 3D Environment. On the virtual stage, character layers, Jack and Alice, from the collage board are loaded as a standing 2D plane. Looking at the *Virtual Stage* view pane in the upper left corner of the interface (Figure 4.ⓒ), the director and staff determine the scene environment. First, they select clear sky texture from the Sky dropdown menu to shoot at noon. Then, they upload a couple of panorama images and 3D scanned data of the prospective locations that they took a couple of days ago to the system. After testing different locations, they decide to shoot on the terrace. Since they want to film on a clear day, they leave the Rain/Snow toggle off. After setting the environment, they position each video layer from the *Layout* view pane at the bottom left (Figure 4.ⓒ). Besides video layers, they add additional components to the virtual stage, like a 3D mock-up of virtual cameras, staff, and lighting gears, to plan effective movement paths for the production team. For example, to add a Director Of Photography (DOP) on the virtual stage, they set the name as 'DOP,' click the *Create* button, and drag the mock-up to the desired position, looking from the bird-eye view.

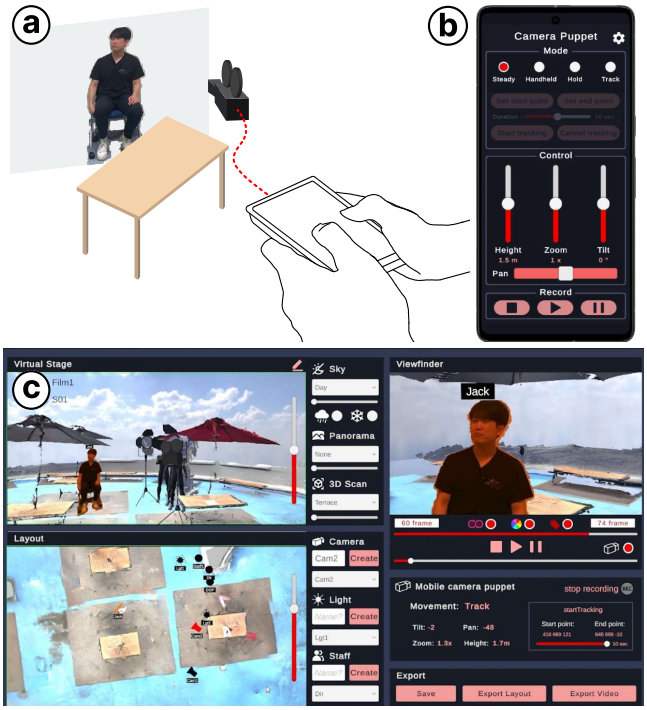


Figure 4: Virtual stage: ⓐ camera puppet control setup, ⓑ mobile application, ⓒ laptop interface

Step 4. Designing Virtual Camera Movements via Mobile. Instead of recording static shots, the director wants to film in a more dynamic manner, experimenting with various camera techniques. So, she opens up the mobile camera puppet application (Figure 4.ⓑ). She moves the virtual camera to the desired position by tilting the mobile left/right and forward/backward in the *Steady* mode. To capture the terrace from the character Jack's point of view, the director selects the *Hand-held* mode and gently shakes the mobile as if it is a character's head. The virtual camera reflects this unfiltered hand motion and captures the scene in the *Viewfinder* pane in the upper right corner (Figure 4.ⓒ). Then, to record a smooth camera transition from the character Jack to Alice, the director manipulates the *Pan* slider in a *Hold* mode, simulating a camera pan on a tripod. Lastly, the director wants to frame the character Jack and express his emotional panic using a *Dolly Zoom* [4] camera technique. So, the director selects the *Track* mode. In this mode, she first zooms in on Jack using *Zoom* slider and clicks the *Set Start Point* button. Second, she zooms out, moves the virtual camera closer to Jack by tilting the mobile forward, and clicks the *Set End Point* button. Third, she sets the tracking duration to 5 seconds and clicks the *Start Tracking* button. The character Jack remains its size in the frame while the background continuously becomes smaller in scale, creating perspective distortion. During the simulation, she realizes the staff mock-up appears when she zooms out. So, she drags them to a further position. After all the experimentation and recording of various shots using multiple virtual cameras, the director exports the top view and each camera's viewfinder as an animated floor plan and video collages.

5 IMPLEMENTATION

We built *CollageVis* system on a laptop and a mobile using Unity3D (2021.3.18f1) with C# and integrated Python (3.10.11) for backend tasks such as image segmentation and face swapping. For the prototype, we used the ProArt Studiobook laptop, equipped with an Intel Core i9 12900H CPU at 2.5GHz, 16GB of RAM, and an NVIDIA GeForce RTX 3070 Ti (8 GB) Laptop graphic card, and the Galaxy A53 mobile.

Collage board: For the real-time video composition in collage board, *CollageVis* streams mobile camera input to the laptop and composes them using MediaPipe's *SelfieSegmenter* model [5]. The collage board separates the foreground image from the background and saves it as an image sequence with an alpha mask. In order to save the storage while maintaining acceptable quality, we set the video to HD resolution (1280x720) and processed it at 5fps. To tag the character name in the video collage, we track the user's head using MediaPipe's *Holistic* solution [1]. Furthermore, we integrated the deepfake technology, SimSwap [6], to seamlessly transfer facial features from a given photo to the video layer. This model was chosen for high identity performance and superior attribute preservation. In addition, Google Cloud Speech-to-Text API [6] was used for the generation of a script draft from the video's sound input, and the WORLD vocoder API [34] was employed for voice alterations.

Virtual Stage: When users move to the virtual stage, *CollageVis* loads each character's data, an image sequence, on a 2D plane and distributes them on a virtual space. The virtual space uses a giant sphere and cylinder to compose a sky and backdrop of the stage. We used High Dynamic Range Image (HDRI) to texture the sphere and panorama image to texture the cylinder. The prototype has several preset HDRI images from PolyHaven [7] and panorama images we

[4] https://en.wikipedia.org/wiki/Dolly_zoom
[5] https://developers.google.com/mediapipe/solutions/vision/image_segmenter
[6] https://cloud.google.com/speech-to-text
[7] https://polyhaven.com/hdris

took with a mobile (e.g., terrace, office, street). For the 3D data, the system supports both scanned and modeled data (we show how we applied both types in the Section 6). For the weather simulation, we used Unity's basic particle system. To control the virtual camera's movement and settings using a mobile, we built a secondary mobile application using Unity3D. This application reads the IMU data of the mobile and sends it to the laptop via UDP communication. Finally, for seamless composition between 2D and 3D elements during the recording, we applied a *Billboard effect* to video layers so that they always rotate towards the virtual camera.

6 CINEMATIC SIMULATION EVALUATION: RECREATING REFERENCE FILM SCENES

In this section, we introduce previs examples we created using *CollageVis* for the existing film scenes to assess the *CollageVis* system's ability to perform the three types of simulation: actor, environment, and camerawork.

Figure 5 shows the processes and outputs of video collages depicting seven well-known film scenes: the dinner scene from the Mr & Mrs. Smith (2005), the red light green light scene from The Squid Game (2021), the boy with apple scene from The Grand Budapest Hotel (2014), the rain scene from the If Only (2004), the peach scene from the Parasite (2019), the jazz club scene from the LaLaLand (2016), and the subway scene from the Joker (2019).

6.1 Actor Simulation

CollageVis can change the actor's face to simulate the potential casting beforehand and duplicate the actor's body to simulate the crowd scene. Figure 5.ⓐ shows an example previs for the dinner scene of Mr & Mrs. Smith (2005), where we took a picture of the actor, Brad Pitt and applied a face swap filter. Normally, indie directors receive profile pictures from actors and imagine how the actors would fit in their visions because of the limited time and money resources. With *CollageVis*, the director can record oneself or one of the staff quickly and then change the stand-in's face with the actor's face to test the look and feel of the shot. This can also be used to communicate with the already casted actor to visually deliver the detailed action the director wants. Meanwhile, Figure 5.ⓑ illustrates a crowd scene previs for the red light green light scene from the Squid Game (2021). Here, the director records three staff standing and clicks a copy button to duplicate each test video 50 times. Then, the director lays them out on a virtual stage and records from the bird's eye view, slightly tilted downwards. This body cloning is fast and useful to simulate simple crowd scenes.

6.2 Environment Simulation

Moreover, *CollageVis* can change the background panorama image or 3D data to test different shooting locations and add the simple effect of rain/snow to simulate desired weather conditions.

To create a previs for scenes where the character's movement path is important, the *CollageVis* allows the director to upload the 3D data and place each video layer inside the environment accordingly. For example, in the boy with apple scene from The Grand Budapest Hotel (2014), the character Dmitri appears from the second floor, walks down the stairs, and passes the hallway to arrive in front of the table (Figure 5.ⓒ). We uploaded a similar environment 3D model to the system and replicated this transition. The 3D data can be fully computer-generated assets like this example scene when there is a 3D artist in the film crew, or it can be rough 3D scan data as we did in the rest of the example scenes. We scanned various environment such as home (Figure 5.ⓐ), playground (ⓑ), street (ⓓ), mension (ⓔ), piano room (ⓕ), and subway (ⓖ) using free 3D scanning app [8] with iPhone 13 Pro. In addition to the location, the director can set the time of the scene by selecting sky textures and can toggle the rain/snow button to simulate the weather condition as shown in Figure 5.ⓓ. We replicated the raining scene from If Only (2004), where Samantha and Ian run on a rainy street to catch the taxi.

6.3 Camerawork Simulation

Lastly, the *CollageVis* can simulate various camera movements such as tracking, panning, and handheld motion. To recreate the famous peach scene from Parasite (2019), characterized by the smooth track motion of the camera following Ki-jung passing by Moon-gwang sleeping on the couch, we used the track mode of the mobile camera puppet application (Figure 5.ⓔ) We first set the start point and end point of the tracking by clicking the button in the desired positions and then set the duration using the slider. Finally, we clicked the start tracking button to simulate the smooth camera transition, like the dolly cam effect in cinematography, following Ki-jung's walking movement. Meanwhile, for the jazz club scene from the LaLaLand (2016), we used the pan function on the mobile app (Figure 5.ⓕ). To imitate the whip pan camera motion, quickly switching between Mia and the people dancing and Sebastian playing the piano, we move the pan slider from left to right. Finally, to reflect the shaking motion in the subway scene from the Joker (2019) (Figure 5.ⓖ), we used a handheld mode in the mobile app. We moved a mobile phone up and down as if we were in a moving train while recording the shot.

6.4 Simulation Constraints

Despite the *CollageVis* system's ability to simulate the aforementioned various scenes, we found two prerequisites for the smooth simulation. First, users should record the full body of an actor (or stand-in) at the same distance to remove the tedious process of matching one video layer's size to the other. Second, framing characters from different perspectives requires preparing multiple source video clips from various angles. This is the limitation of the 2.5D environment; all the video layers are 2D planes in the 3D environment, always rotating towards the active virtual camera, which means the actor appears at the same angle from the input video regardless of the virtual camera's position. For instance, in the sample scene of the Joker (Figure 5.ⓖ), the director first shot the Joker's right side of the face to capture him looking at the incident between three men and the woman in the back and then showed the left side to capture him from the woman's perspective. To replicate these two shots, we had to record the actor twice from left and right angles. Unlike us, the potential user may not have clear target shots in mind. In this case, users can easily add new video clips from the desired perspective when they notice the necessary scenes later at

[8] https://3dscannerapp.com

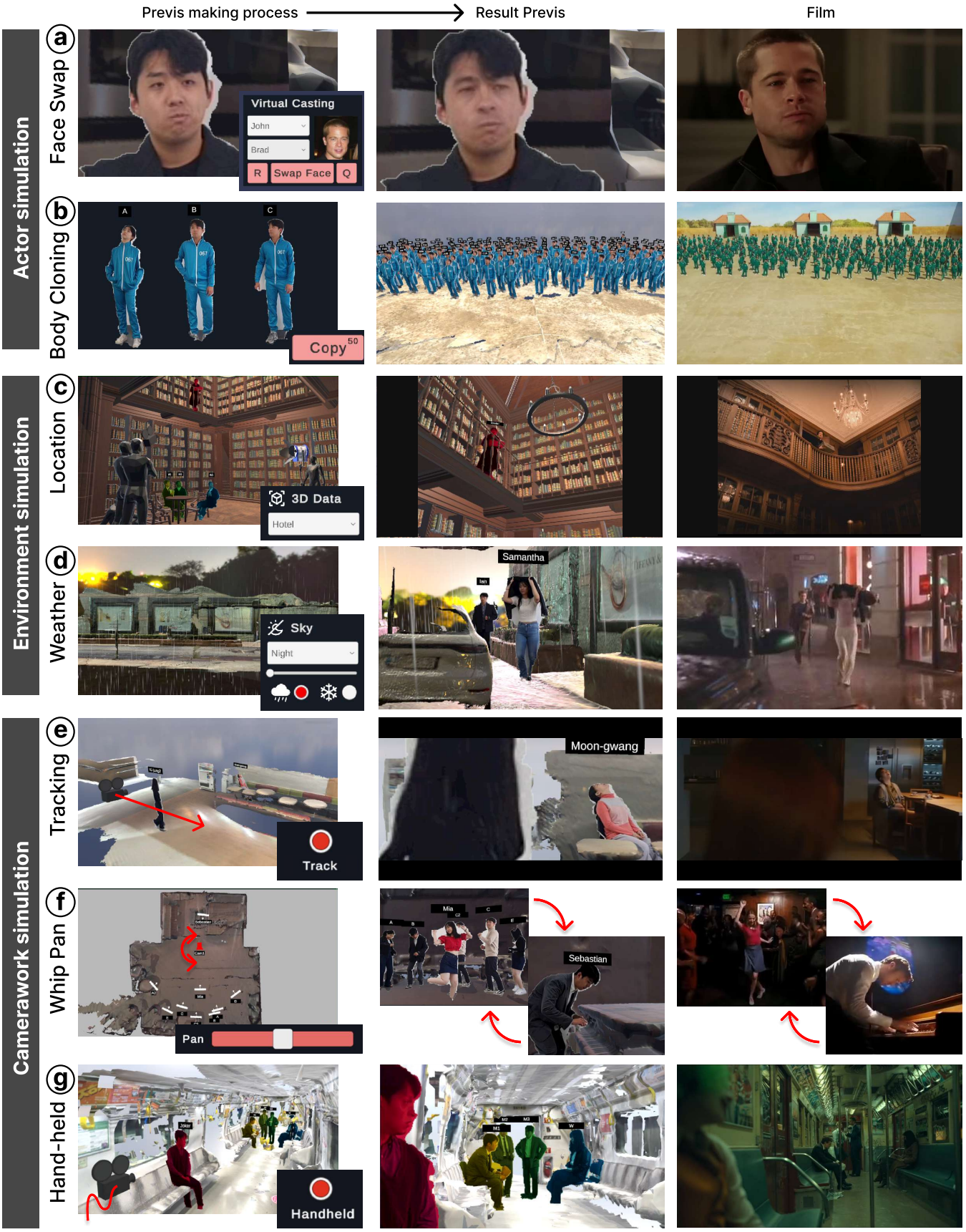


Figure 5: Seven previs making processes and examples for existing film scenes simulating actor (ⓐ face swap and ⓑ body cloning), environment (ⓒ location and ⓓ weather), and camerawork (ⓔ tracking, ⓕ whip pan, and ⓖ handheld).

any time, without casting the same actor, as they can use the face swap function.

6.5 System Performance Evaluation

The *CollageVis* system uses off-the-shelf solutions for subfeatures such as image segmentation (MediaPipe's *SelfieSegmenter* model), name tag [1], face swap [6], audio-to-script generation (Google Cloud Speech-to-Text API). Table 1 shows the performance of each subfeature in seven previs examples. The image segmentation and name tagging features worked in real-time with a latency of 33.46 ms. While the name tagging showed high accuracy in all seven cases, the quality of the segmented outline was heavily jagged in some frames, such as the actor moving fast in the Grand Budapest Hotel example and the actor wearing dark-colored clothes in poor lighting conditions in the Parasite example. Also, the model failed to extract the alpha mask when the actor turned her back from the camera in the If Only previs. As to the face-swapping feature, it required a considerable amount of time to complete processing, with an average processing time of 35 seconds for a 5-second input video. Furthermore, when the actor's face was too small in a frame, the change was marginal. In that case, we cropped the input video, applied face-swapping, and then superimposed it to the original video, which created a substantial improvement. Finally, the audio-to-script feature was applied offline with a 1-second delay after the recording was finished to focus computing power on the visual components. Unfortunately, three out of seven scenes did not include dialogue, as we chose the target film scene based on visual variety. Nevertheless, the mean accuracy of the remaining four scenes (calculated by counting correctly guessed words over all words) was high at 94.30 (SD : 7.86) and showed occasional errors in detecting foreign names and interjections.

It is worth mentioning that the previs authoring time does not necessarily correlate with the duration of the target film scene. Factors influencing the authoring time include the number of video layers, the duration of video clips with the face swap filter applied, and the complexity of the camera work.

7 INDIE FILM DIRECTOR REVIEW STUDY

To evaluate the usability and the efficacy of the *CollageVis* prototype and compare its workflow with traditional previs making, we conducted a user study with indie film directors. As there is no clear counterpart previs tool that has similar capability with the *CollageVis* system, we focus on the qualitative expert review instead of a comparison study against a baseline similar to prior works [26, 48].

7.1 Participants

We recruited six indie film directors aged 21-41 (M : 29.67, SD : 8.73, 1 female and 5 males) who have 1-10 years (M : 4.00, SD : 3.29) of experience in filmmaking and have made at least one previs for their films. Three out of the six directors (P3, P4, and P6) had previously participated in our formative interviews. Their preferred way of previs differed as follows: three participants (P3, P4, and P6) exclusively used hand-drawn 2D storyboards. In addition to a 2D storyboard, P1 also made a test video to explore camera transition, while P5 created a 3D animation using Maya and Unreal Engine to test camerawork and lighting. Meanwhile, P2 relied solely on Maya to create 3D previs. As to the level of technical skill, most participants lacked confidence in drawing and had no experience in 3D graphics. But, two (P2, P5) were competent in both drawing and 3D graphics, especially. P2 was the most skillful in previs making, being a film academy student preparing to become a professional previs artist. Participants were compensated 40 USD in local currency.

7.2 Study Design

To enable users to focus on evaluating the usability of the *CollageVis* system, we provided video clips and asked users to compose a scene depicting a family dinner, involving more than two characters in various locations. We prioritized allowing participants to explore every feature of the system and aimed to reduce the workload for participants in creating a story and video shooting within the limited time frame of the user study.

Task. Participants were asked to create a family dinner scene using provided video clips. They were allowed to record and add new video clips if needed. Using the collage board interface, participants were asked to edit the timing of each video layer, apply name tags and color filters, and change the actor's face and voice to their liking. Then, they were asked to move the scene to the virtual stage. In the virtual stage interface, participants selected the time, weather, and 3D scan data to define the scene environment and position video layers (=actors), lighting gears, and staff in the selected environment. Finally, participants used a mobile phone to move a virtual camera and record various shots.

Resources. We provided 36 video clips of an actor saying relevant lines such as "It's too spicy. water please" and "Can't someone just eat a meal in a peaceful time. (answering the phone) Hello?". Also, we prepared two panorama images and 3D scan data of the indoor room and outdoor terrace for the background (Figure 6).

Study Procedure. The study was carried out in an empty studio in our institution. Upon arrival, participants filled out a demographic survey and were introduced to the system interfaces with a demo video. After the introduction, we explained the available resources for the task. Then, participants created the scene for up to 60 minutes. After completing the task, participants provided feedback by answering three questionnaires: custom questionnaire similar to the prior work [26] comparing the *CollageVis* with their prior previs practices, System Usability Scale (SUS) [4] and NASA Task Load Index (NASA-TLX) [15] questionnaires to evaluate the system's general usability and assess the user's perceived workload. At the end, we conducted a 30-minute semi-structured interview. We asked participants to share their views on the system's value and limitations, particularly regarding its practical integration in their workflow (e.g., What is the advantage/disadvantage of output results from *CollageVis*, and how will you use them?). The study lasted 2 hours in total.

7.3 Results

Table 2 showcases the user-created videos (previs and floor plan), output video length, and authoring time. All participants completed the task within time except P1. P1 spent 44 minutes crafting dialogue on the collage board interface, leaving insufficient time for

Reference Film	Duration (second)	Image Segmentation	Name Tag	Face Swap Speed (seconds/second)	Audio to Script	Authoring Time (minute)
Mr & Mrs. Smith	26	100%	100%	7	83%	12
Squid Game	8	100%	100%	8	NA	27
The Grand Budapest Hotel	21	87%	96%	5	100%	18
If Only	20	88%	100%	6	100%	15
Parasite	16	82%	100%	8	94%	20
LaLaLand	15	100%	100%	6	NA	12
Joker	14	100%	100%	7	NA	10

Table 1: System performance in seven previs examples recreating reference film scenes.

Figure 6: Sample user study materials in two virtual scenes: ⓐ outdoor terrace, ⓑ indoor room.

recording on the virtual stage. The visual outcomes exhibited distinct variations in camera movement design and shot composition. Participants utilized different camera movements, including static zoom (P1, P6), dolly zoom (P2), and circular dolly track (P3). In addition, unlike others framing one character at a time, two participants (P4, P6) captured two characters in a single frame using over-the-shoulder shots.

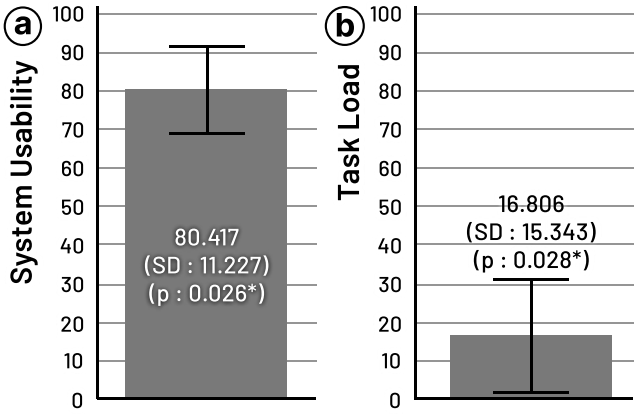


Figure 7: SUS and NASA-TLX Questionnaire results: ⓐ system usability and ⓑ task workload. Mean scores, standard deviation, and p-value of one-sample Wilcoxon signed-rank test against median (50) are displayed on the bars.

Figure 7 and 8 summarize the participants' responses to SUS, NASA-TLX, and custom questionnaires. We conducted a one-sample Wilcoxon signed-rank test to determine if there was a significant difference between the median of participants' ratings and the median of the scale (4 for the custom questionnaire, the median of a 7-point Likert scale, and 50 for SUS and NASA-TLX questionnaires, the median of scales converted to the 0-100 range).

The participants' mean score for the usability on the SUS questionnaire was high at 80.417 (M : 80.417, SD : 11.227), while the perceived task load on NASA-TLX was low at 16.806 (SD : 15.343), both out of 100 (Figure 7). The median values of SUS and NASA-TLX were both significantly different from 50, with a strong effect size (SUS: $Z = 2.226$, $p = 0.026^*$, $r = 0.909$, NASA-TLX: $Z = -2.201$, $p = 0.028^*$, $r = 0.899$). The average rating on our custom questionnaire was 5.850 (SD : 0.987) on a 7-point Likert scale, and the median was significantly different from 4, $Z = 2.201$, $p = 0.028^*$, with a strong effect size ($r = 0.899$) (Figure 8).

Overall, all participants reported that *CollageVis* was easy to learn and its' diverse features were useful. The participants responded that it was easy to use (P1-P6) and useful for idea development (P1, P4, P6), plan production (P3-P6), and collaboration (P1-P6).

> *P5: "It's just faster. Taking a video is much easier than drawing or adding a keyframe animation to 3D characters."*
>
> *P4: "I really liked that I can see the result in various ways (video collage per each virtual camera and floor plan). It will be helpful to communicate with [...(art director, director of photography, and actor)...]"*

1. Simplifying The Previs Creation Process With Some Limitations. Looking at overall preference (Q1), although there was no statistical difference in the median ($Z = 1.913$, $p = 0.056$), most participants except P2 preferred the *CollageVis* over their accustomed previs methods (M : 5.500, SD : 1.378). All participants were willing to use the *CollageVis* for their next films (M : 6.000, SD : 1.265), and the median willingness (Q2) showed a statistical difference ($Z = 2.060$, $p = 0.039^*$, $r = 0.841$). A unique aspect of *CollageVis* is that "it starts from videos to make a video (film)." The majority of participants saw this positively as it is a familiar working method of making test videos (P1), and it replaces the tedious hand-drawing process (P3-P6).

Using flat video clips was our intention to reduce complexity and a welcomed feature by participants (P3, P4, P6); however, P2, who is more experienced in 3D graphics, saw this as a major setback because it means the tool cannot frame actor in different camera angle or lighting from the input video. Meanwhile, P5, who has technical skills similar to P2's, remained positive, saying that "it is enough for most of my work."

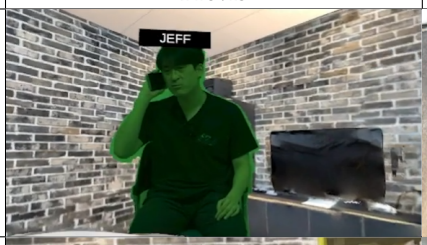
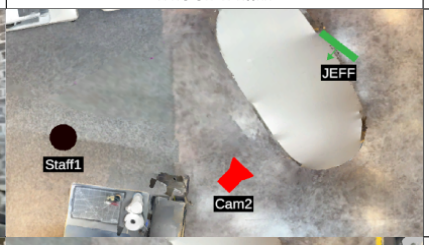
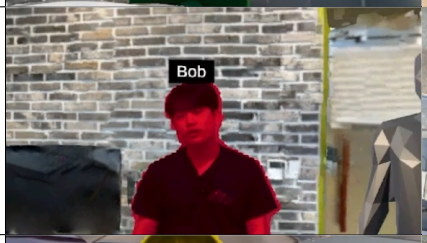
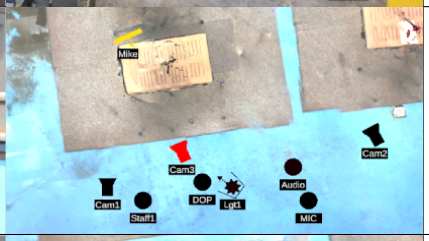
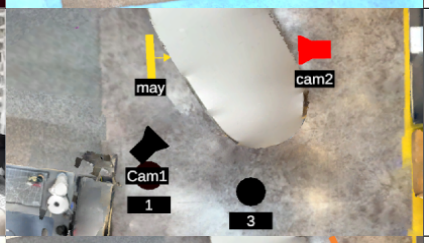
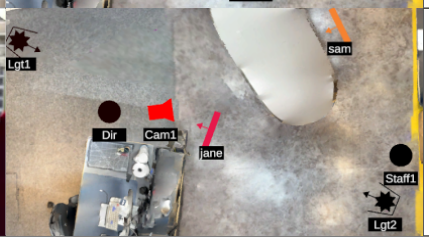
Participant	Previs	Floor Plan	Length	Authoring Time
P1			12 sec	60 min
P2			12 sec	19 min
P3			15 sec	41 min
P4			18 sec	30 min
P5			18 sec	38 min
P6			17 sec	37 min

Table 2: User study task results: user-created video-based previs and floor plan examples depicting family dinner scenes.

2. Beneficial for Creative Exploration, but Requires Video Preparation. Looking at exploration (Q6), all participants agreed that *CollageVis* encourages creative exploration (M : 6.000, SD : 0.894), and the median exploration response showed a statistical difference ($Z = 2.220, p = 0.026^{*}, r = 0.906$). Participants said that the creativity of filmmakers can be found in narrative construction and visual storytelling.

> *P2: "The same story can be directed to evoke different feelings (from spectators). [...] How the characters move, how the camera frames, and*

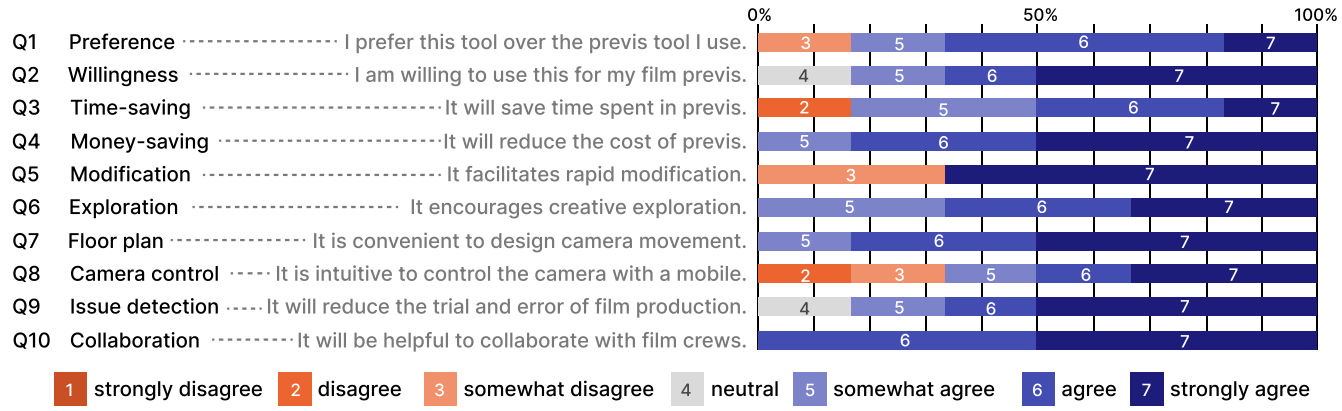


Figure 8: Custom Questionnaire results: 100% stacked bar chart (n = 6) of participants' ratings on "Do you agree or disagree with the following statements when you compare *CollageVis* with your familiar previs tool?" (Q1-Q6) and "Do you agree or disagree with the following statements about *CollageVis*?" (Q7-Q10).

> *how the light casts [...] It all comes down to the message you want to deliver."*

While they believed that *CollageVis* would not support their creativity in story building, they anticipated that it could support the refinement of their creative vision by enabling the rapid generation of visual variations. Participants believed that the video format would motivate them to think about nuanced details in character actions (P4, P6) and camera movements (P1-P6), which can be challenging with static drawings.

> *P6: "I can explore various ways to better express the character. For example, a character jaded with work could smoke a cigarette like this (hunched over) or like this (leaning back in the chair). It creates a different vibe."*

> *P2: "I usually jot down 'Zoom In' on the drawing, but with this, I can vividly imagine the specifics [...] how the camera and the actor move in the shot [...]"*

Finally, many participants appreciated the virtual stage interface simulating different shooting environments because it allowed them to see the look and feel of the shot in different weather (P3) and to test out various shooting locations (P1, P4, P6) before solidifying their vision.

Nevertheless, in modification question (Q5), the responses were split into two groups of score 3 (two participants) and 7 (four participants), resulting in an average of 5.667 (SD : 2.066). The median was not statistically significant ($Z = 1.622, p = 0.105$). This is because modifying some of the features like dialogue, character's face, environment, and camera framing was much easier than the storyboarding, as we expected, but modifying the character's action and full body appearance, like clothing and hairstyle, was not easy. It required the user to reshoot another video clip, and some felt it was cumbersome.

> *P4: "I think changing the drawing is much faster. If I want to make a male character into a female, I can just draw a ponytail."*

So, participants wished that they could use drawing (P4) or photography (P6) when they could not prepare the new video input.

3. Minimizing Trial and Error and Promoting Effective Collaboration. As shown in the questions of floor plan (Q7), the ability to design camera movement in *CollageVis* is a well-liked feature (M : 6.333, SD : 0.816), with statistical significance ($Z = 2.232, p = 0.026^{*}, r = 0.911$). Participants thought using the top view of the virtual stage and adding different components like actors, staff, cameras, and lights were comfortable and intuitive (P1). Participants also said looking at the camera movement in the top view would help them plan cinematography, especially when there are multiple cameras (P3, P6). For similar reasons favoring the floor plan, participants gave high scores to the issue detection (Q9) (M : 6.000, SD : 1.265) ($Z = 2.060, p = 0.039^{*}, r = 0.841$) and collaboration (Q10) (M : 6.500, SD : 0.548) ($Z = 2.251, p = 0.024^{*}, r = 0.919$).

Making previs while checking the floor plan made participants spot the potential problem of arrangement. For example, P2 and P5, who chose the scan data of the confined room as the background instead of the open space of the terrace, had to rearrange the lighting gears and staff to frame the actor without obstacles.

> *P2: "It made me think ahead how to plan production."*

Participants thought that exporting in two formats of video collage (camera view) and floor plan (layout view) would be useful when they are communicating with various creatives, from non-technical actors and art teams to technical staff like the director of cinematography.

However, it's worth mentioning that using a mobile application to control camera movement received different feedback depending on the user's expertise (Q8) (M : 5.000, SD : 2.098) ($Z = 1.163, p = 0.245$). Most participants said it was intuitive, and some mentioned it is similar to playing a mobile game (P4) or tripod adjustment (P6). But, P2 and P5, skilled at the 3D DCC tools, wanted to use a laptop with a mouse instead of a mobile, saying it was uncomfortable and slow. Nevertheless, all participants appreciated the hand-held

camera mode that can reflect their physical motion of moving the mobile into the virtual camera.

4. Reducing Manual Labor and Increasing Experimental Expression. When asked about how the *CollageVis* would influence their filmmaking workflow, all participants expected that it would reduce previs authoring time by replacing manual tasks of drawing and keyframe animation with video recording. They thought this saved effort could be used to increase the quality of films. Furthermore, some participants (P4, P6) believed they could easily develop high-quality visuals so that later it would be possible for them to repurpose their previs as different forms of art (e.g., publication, exhibition), similar to what great directors do with their storyboards.

Participants also expected that they would be able to attempt more experimental cinematographic techniques in their films with *CollageVis*. This was because the system allowed them to work remotely without temporal and spatial constraints.

> *P6: "(Compared to on-set rehearsal,) I have more time to experiment with everything without consuming the crew's time [...] If I had this tool, I would have tested various camera movements like extremely speedy transition and frequent cut scenes before shooting (explaining a scene in his film) [...]"*

> *P1: "With this tool, I can quickly visualize it, decide if I like it or not, and try again (for unique expression)."*

Consequently, participants said they would make a careful decision after iterative testing.

> *P3: "Before production, I will spend more time to test different framings, add symbolic objects in the frame, planting subtle clues for the story."*

Regarding its rapid prototyping capability, participants suggested that the *CollageVis* would be useful for the previsualization of other narrative media productions such as commercials (P4), animations (P1, P5), and theatrical performances (P6). Furthermore, some participants mentioned possible use cases beyond narrative media in dance choreography design (P5) and film education (P2, P6).

In summary, our user study showed that *CollageVis* is easy to use and valuable for creative exploration and effective collaboration. The study revealed four primary insights:

(1) Participants found that the *CollageVis* is more effective than 2D storyboarding. However, they noted it couldn't replace 3D previs due to its inability to simulate lighting.
(2) In terms of creative exploration, participants praised the system's capabilities to quickly capture the character's action, as well as environmental conditions. Yet, they wished for more input options (e.g., drawing, photography) other than video for easier modification.
(3) The layout design feature was well-liked by all participants as it is useful for communication among the film crew, but controlling the virtual camera using a mobile received different feedback depending on the participant's technical expertise. Non-technical participants generally liked a mobile control for simplicity, while participants familiar with 3D DCC tools wished to switch to mouse control.
(4) Compared to their current filmmaking workflow, participants expected that *CollageVis* would encourage them to experiment more with visual expressions, such as on-screen actions, camera movements, and background settings, before the production.

8 LIMITATION & FUTURE WORK

Current film previsualization practices are divided into analog methods (e.g., storyboarding, physical rehearsal using maquettes or stand-ins) and digital methods (e.g., 3D animations in game engines and virtual reality environments). Both approaches impose a substantial burden on indie filmmakers in terms of time and budget management. *CollageVis* bridges the gap by transitioning the physical rehearsal from the traditional workflow of test videos to the virtual space. User study results confirm that *CollageVis* demonstrates its potential in minimizing manual labor and enhancing communicative aspects in future filmmaking workflows while maintaining a low cost. However, *CollageVis* needs further improvement regarding the following limitations to fully leverage the benefits of both analog and digital approaches.

1. Enhancing Input Flexibility and Modification. Similar to rapid prototyping methods in user experience design, where high-fidelity development cycles are shortened using paper-based prototyping [23] or the Wizard of Oz technique [39], *CollageVis* replaces drawing in 2D storyboard and keyframing character animation in 3D previs to simple video recording.

Although this saves time and lowers the technical barrier of creating anatomically convincing human figures and actions, the input video could not be partially modified to illustrate the characters' personalities through different accessories. We consider further supporting the post-processing of video, such as augmenting costumes and hairstyles [27].

In addition, as the current prototype only segments human figures, sometimes the props (e.g., a wine glass on the actor's hand) disappear. We plan to support dynamic annotation on video like in [52, 53]. This would allow the user to draw essential props in the environment or on a character's hand without having to record another video clip. We can also support various input types (e.g., static pictures, drawings, and 3D props) for fast modification.

2. Supporting Video Perspective Change. As indicated in Section 6.4, the user needs to record the same acting multiple times for different perspectives. To relieve this, future work can apply image manipulation techniques [44], which can change the image's perspective by clicking and dragging them to desired positions.

3. Enhancing Video Composition Quality. Our current prototype uses MediaPipe's SelfieSegmenter model [31] to extract the foreground from the background. We chose this model for its real-time processing capability and easy implementation, but this model creates a jagged edge depending on the lighting quality of the input video. During the user study, most participants did not mind the low quality since it was for previs. However, one participant (P1)

commented that the low quality was distracting when envisioning the final look. For refined image segmentation, offline image segmentation models (e.g., DeepLabV3+ [5]) could be employed. Alternatively, as a more lightweight solution, we can achieve visual consistency by stylizing the output video collage, like the prior work [7] did using cartoon style.

4. Varying Camera Control Modalities. The virtual stage of *CollageVis* provided a mobile to control the virtual camera instead of a mouse to lower the technical barrier in 3D navigation in digital space and to reflect the manual motion in the hand-held camera mode. From the user study, we confirmed that directors appreciated using the mobile device to replicate motion in virtual camera work, such as hand-held and dolly-cam effects. However, some directors, familiar with the 3D interface, expressed a preference for tilting, panning, and zooming using the mouse on the laptop. To enhance the usability of manipulating the virtual camera, we plan to offer interfaces on both mobile and laptop.

5. Applying Video Pose Estimation for 3D Previs. Although 3D previs is not common in indie films, as described in the formative interviews, transitioning to 3D character animations from 2D videos could improve the system's fidelity because it can simulate different lighting conditions and camera angles. Recent advancements in video-based motion tracking technologies [35, 45] show the great potential for automatic conversion of 3D character animations from video inputs. Unlike prior 3D previs systems that relied on manual keyframe animation or full-body motion capture system [13, 20, 36], using video input for 3D animation would combine the benefits of using 2D video and 3D characters. This can enable full rendering capabilities such as 360° camera tilting and lighting while maintaining the indie film's low budget and the actor's natural performance.

6. In-Depth Assessment in Actual Film Previs and Beyond Film. Lastly, despite the positive responses from indie film directors, our user study has a relatively small sample size and was conducted in the lab with a designed task. Future work should attempt to test *CollageVis* for actual film previs in the wild.

Furthermore, since *CollageVis* does not require a complete script like the text-based previs tools [22, 32], we can further explore potential usage beyond the context of narrative media production, such as making music videos or designing group dance choreography.

9 CONCLUSION

In this paper, we presented *CollageVis*, a rapid previsualization tool for indie film production using video collages. We first interviewed six indie film directors to understand the practical needs and challenges in previs making. Then, we demonstrated how *CollageVis* addresses them by accelerating the test video editing process with real-time video composition. Also, we further improve its communicative aspect by adding actor differentiation filters and the ability to design floor plans in 2.5D space. We created seven previs samples for existing various film scenes to assess the *CollageVis* system's capability of simulating actors in various locations with camera movements and reported the technical constraints. Finally, we evaluated our prototype's usability with six film directors and found its value in developing ideas and planning production for collaboration.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1C1C1011962). Also, this research received support from the Chung-Ang University Research Grants in 2021. We would also like to express our gratitude to actor Chan Young Hong and undergraduate students Chan Hu Wie, Dong-Uk Kim, Yejin Jang, and Yurim Son for portraying various characters.

REFERENCES

[1] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204* (2020).
[2] Toon Boon. 2022. Storyboard Pro. 2D storyboard & animatics software. Retrieved February 6, 2023 from https://www.toonboom.com/products/storyboard-pro
[3] Ron Brinkmann. 2008. *The art and science of digital compositing: Techniques for visual effects, animation and motion graphics*. Morgan Kaufmann.
[4] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.
[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).
[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2003–2011.
[7] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. 2019. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2236–2244.
[8] Hye-Kyung Choi and Sae-Hong Cho. 2014. Development of effective previsualization authoring tool using conversion technologies—based on film storyboard application. *Cluster computing* 17 (2014), 585–591.
[9] Edirlei Soares De Lima, Bruno Feijó, and Antonio L Furtado. 2018. Video-based interactive storytelling using real-time video compositing techniques. *Multimedia tools and Applications* 77 (2018), 2333–2357.
[10] David Elson and Mark Riedl. 2007. A lightweight intelligent virtual cinematography system for machinima production. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 3. 8–13.
[11] Dustin Freeman and Ravin Balakrishnan. 2016. Improv Remix: Mixed-Reality Video Manipulation Using Whole-Body Interaction to Extend Improvised Theatre. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 533–542.
[12] Dustin Freeman, Stephanie Santosa, Fanny Chevalier, Ravin Balakrishnan, and Karan Singh. 2014. LACES: live authoring through compositing and editing of streaming video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1207–1216.
[13] Quentin Galvane, I-Sheng Lin, Fernando Argelaguet, Tsai-Yen Li, and Marc Christie. 2019. Vr as a content creation tool for movie previsualisation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 303–311.
[14] Quentin Galvane, I-Sheng Lin, Marc Christie, and Tsai-Yen Li. 2018. Immersive previz: Vr authoring for film previsualisation. In *ACM SIGGRAPH 2018 Studio*. 1–2.
[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[16] Robert Held, Ankit Gupta, Brian Curless, and Maneesh Agrawala. 2012. 3D puppetry: a kinect-based interface for 3D animation.. In *UIST*, Vol. 12. 423–434.
[17] Yosuke Horiuchi, Tomoo Inoue, and Ken-ichi Okada. 2012. Virtual stage linked with a physical miniature stage to support multiple users in planning theatrical productions. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 109–118.
[18] Ryosuke Ichikari, Keisuke Kawano, Asako Kimura, Fumihisa Shibata, and Hideyuki Tamura. 2006. Mixed reality pre-visualization and camera-work authoring in filmmaking. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE, 239–240.
[19] Sihyeon Jo, Zhenyuan Yuan, and Seong-Woo Kim. 2022. Interactive Storyboarding for Rapid Visual Story Generation. In *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 1–4.

[20] Ching-Yu Kang and Tsai-Yen Li. 2021. One-Man Movie: A System to Assist Actor Recording in a Virtual Studio. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 84–91.

[21] Mubbasir Kapadia, Seth Frey, Alexander Shoulson, Robert W Sumner, and Markus H Gross. 2016. CANVAS: computer-assisted narrative animation synthesis.. In *Symposium on computer animation*. 199–209.

[22] Hanseob Kim, Ghazanfar Ali, and Jae-In Hwang. 2021. ASAP: Auto-generating Storyboard And Previz with Virtual Humans. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 316–320.

[23] Han-Jong Kim, Chang Min Kim, and Tek-Jin Nam. 2018. Sketchstudio: Experience prototyping with 2.5-dimensional animated design scenarios. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 831–843.

[24] Krock.io. 2023. Storyboard AI. Automatic storyboard generation software. Retrieved June 20, 2023 from https://krock.io/storyboard-ai/

[25] Tobias Langlotz, Mathäus Zingerle, Raphael Grasset, Hannes Kaufmann, and Gerhard Reitmayr. 2012. AR record&replay: situated compositing of video content in mobile augmented reality. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 318–326.

[26] 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*. 1–13.

[27] Jingyi Li, Wilmot Li, Sean Follmer, and Maneesh Agrawala. 2021. Automated Accessory Rigs for Layered 2D Character Illustrations. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 1100–1108.

[28] Sicheng Li, Haoran Xie, Xi Yang, Chia-Ming Chang, and Kazunori Miyata. 2022. A Drawing Support System for Sketching Aging Anime Faces. In *2022 International Conference on Cyberworlds (CW)*. IEEE, 1–7.

[29] Ghostwheel LLC. 2019. Previs Pro. 3D storyboard & augmented reality simulation software. Retrieved November 9, 2023 from https://www.previspro.com/

[30] Amaury Louarn, Marc Christie, and Fabrice Lamarche. 2018. Automated staging for virtual cinematography. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–10.

[31] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[32] Marcel Marti, Jodok Vieli, Wojciech Witoń, Rushit Sanghrajka, Daniel Inversini, Diana Wotruba, Isabel Simo, Sasha Schriber, Mubbasir Kapadia, and Markus Gross. 2018. Cardinal: Computer assisted authoring of movie scripts. In *23rd International Conference on Intelligent User Interfaces*. 509–519.

[33] Ali Mazalek and Michael Nitsche. 2007. Tangible interfaces for real-time 3D virtual environments. In *Proceedings of the international conference on Advances in computer entertainment technology*. 155–162.

[34] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems* 99, 7 (2016), 1877–1884.

[35] Move.ai. 2022. Move.ai. AI video motion capture software. Retrieved February 22, 2023 from https://www.move.ai/

[36] Thomas Muender, Thomas Fröhlich, and Rainer Malaka. 2018. Empowering creative people: Virtual reality for previsualization. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[37] Thomas Muender, Anke V Reinschluessel, Sean Drewes, Dirk Wenig, Tanja Döring, and Rainer Malaka. 2019. Does it feel real? Using tangibles with different fidelities to build and explore scenes in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[38] Leon Müller, Ken Pfeuffer, Jan Gugenheimer, Bastian Pfleging, Sarah Prange, and Florian Alt. 2021. Spatialproto: Exploring real-world motion captures for rapid prototyping of interactive mixed reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[39] Michael Nebeling and Katy Madier. 2019. 360proto: Making interactive virtual reality & augmented reality prototypes from paper. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[40] Michael Nebeling, Janet Nebeling, Ao Yu, and Rob Rumble. 2018. Protoar: Rapid physical-digital prototyping of mobile augmented reality applications. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Michael Nebeling, Shwetha Rajaram, Liwei Wu, Yifei Cheng, and Jaylin Herskovitz. 2021. Xrstudio: A virtual production and live streaming system for immersive instructional experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[42] Michael Nitsche. 2008. Experiments in the use of game technology for previsualization. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*. 160–165.

[43] Jeffrey A Okun, VES Susan Zwerman, et al. 2020. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Routledge.

[44] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

[45] Rokoko. 2022. Rokoko Video. AI video motion capture software. Retrieved February 22, 2023 from https://www.rokoko.com/products/video

[46] Yang Shi, Nan Cao, Xiaojuan Ma, Siji Chen, and Pei Liu. 2020. Emog: Supporting the sketching of emotional expressions for storyboarding. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[47] Midieum Shin, Byung-soo Kim, and Jun Park. 2005. AR storyboard: an augmented reality based interactive storyboard authoring tool. In *Fourth IEEE and ACM international symposium on mixed and augmented reality (ISMAR'05)*. IEEE, 198–199.

[48] 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*. 166–181.

[49] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[50] Hock Hian Wong. 2012. Previsualization: assisting filmmakers in realizing their vision. In *SIGGRAPH Asia 2012 Courses*. 1–20.

[51] Matt Workman. 2019. Cine Tracer. Cinematography Simulator created with Unreal Engine. Retrieved August 15, 2023 from https://www.cinetracer.com/

[52] Zhijie Xia, Kyzyl Monteiro, Kevin Van, and Ryo Suzuki. 2023. RealityCanvas: Augmented Reality Sketching for Embedded and Responsive Scribble Animation Effects. *arXiv preprint arXiv:2307.16116* (2023).

[53] 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 Transactions on Graphics (TOG)* 42, 4 (2023), 1–12.