Multimodal Direct Manipulation in Video Conferencing: Challenges and Opportunities

Josh Urban Davis
Dartmouth College
Hanover, NH, USA
josh.u.davis.gr@dartmouth.edu

Paul Asente
Adobe Research
San Jose, CA, USA
research@asente.com

Xing-Dong Yang
Simon Fraser University
Vancouver, BC, Canada
xingdong_yang@sfu.ca

Figure 1: An overview of Clio's presentation capabilities for real-time performances, presentations, and storytelling. (a) The presenter creating behavior mappings between input interaction methods (right) and output behaviors (left). (b) A menu containing images and other virtual objects. (c) The presenter selects an image using a mid-air gesture and drags it across the screen. (d) To make an image larger, the presenter uses a two-handed pan/zoom gesture. (e) The presenter uses voice commands and mid-air gestures to display text labels on the screen and draw notes on the image.

ABSTRACT

Tools supporting immersive live video conferencing (VC) have gained popularity recently across diverse application domains. A core component of the experience is augmenting video communication with multimodal interactive media. While many direct-manipulation techniques for VC communication have been proposed in existing literature, the usability and preferences for these techniques have never been formally studied. In this paper, we examine how embodied interaction democratizes content authoring, and propose a rehearsal-to-performance (RtP) framework along with a VC system, Clio, that enables performers to directly interact with their media using voice, gesture, and external devices such as tablets. We evaluate existing operation-to-modality mappings for VC communication, as well as describe novel mappings not present in the literature. A series of studies demonstrate modality preferences and potentials for incorporating real-time direct-manipulation tools to create expressive augmented VC performances.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing devices**; **Mixed / augmented reality**.

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07…$15.00
https://doi.org/10.1145/3563657.3596099

KEYWORDS

augmented reality, presentation tools, improvisation, machine vision, multi-modal interfaces, HCI

ACM Reference Format:
Josh Urban Davis, Paul Asente, and Xing-Dong Yang. 2023. Multimodal Direct Manipulation in Video Conferencing: Challenges and Opportunities. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3563657.3596099

1 INTRODUCTION

Real-time online video communication is becoming more popular in a wide range of domains, including education [14], coding [19], creativity and art making [21], video games [42], and economics [21]. This is especially prevalent following the outbreak of COVID-19, when many daily social, professional, and educational activities moved onto VC platforms and have not yet returned to the physical world [14]. Given that a wide variety of social activities that require visual communication now take place in real-time remote VCs, visual effects for augmenting live presentations have become increasingly important to support effective communication. However, there are few tools that let people easily create and interact with visual effects in VCs, and little is known about how they are used by target end users in real-world scenarios. Unlike previous research efforts, this work focuses on formally studying the design of multi-modal direct manipulation systems to inform future interfaces.

Since this topic is significantly broad and rich, we first needed to understand how people present visual material using current VC systems, then explore how people would prefer to directly interact

with visual media during VC presentations, and build a system supporting these proposed interactions. The developed system can then be used to understand how multimodal direct manipulation affects and enriches VC communication. To attenuate this methodology, we conducted an initial formative design study to understand real-world use cases and limitations of existing tools and found a tension between performer expressiveness and media augmentation. Commercial augmented VC systems like mmhmm.ap and Cameo allow presenters to insert graphics into their live video stream, or superimpose their video stream upon backgrounds, but do not allow the two worlds to interact [44]. Performers are limited in their expressive capabilities due to the lack of support for direct interaction with rich graphics, visual media, text, and drawing on the screen. This initial probe revealed that we needed to better understand how performers would want to directly manipulate on-screen media during live presentations.

Through a series of formative surveys, semi-structured interviews, and pantomime studies we found that the participants employed different modes of voice, gesture, mouse, keyboard, and tablet input to perform and control the pantomimed visual effects essential to their use cases. However, presenters expressed concern about being unable to imagine what their effects would look like during a real-time presentation. We also found that enabling direct manipulation usually required technical skills and programming knowledge that many novice users found intimidating or difficult to acquire. Programming direct manipulation to enrich presentations is a substantial burden in simple real-world scenarios like showing vacation photos to loved ones, Q&A sessions during remote presentations, and other use cases where performers might be unable or unmotivated to prepare extensively.

From our observations, we synthesized a research-to-presentation (RtP) workflow for authoring direct manipulations using an immersive approach, and instantiated a system to support this workflow called Clio, a proof-of-concept VC system for augmenting live performances with voice, body and device driven direct manipulation. Clio enables users to easily design and integrate real-time multi-modal visual manipulations without explicit programming by offering a collection of predefined modular primitive operations derived from our formative studies. The effects resemble those that one could do in post-production using applications like Adobe After Effects [1], or through newscaster and weather-reporting support systems that let an external third party control the visuals using a Wizard-of-Oz approach [45]. Our workflow shifts the focus of presentation authoring from content (e.g. slide) authoring to performance authoring, which encourages presenters to focus on preparing the talk itself instead of on the artifact of the talk.

We demonstrate in a series of evaluations with live performers, audiences, and external observers, that letting performers directly interact with media contained in their presentations greatly enriches the communicative and expressive capability of VCs and encourages nonlinear presentation styles. Liberating presentations from linear constraints opens an exciting design space of nonlinear, extemporaneous storytelling and expands the application domain of VCs. In addition, the RtP immersive authoring approach built rapport between performers and machine, even when performers incorporated unfamiliar interaction techniques and machine learning tools such as voice and gesture recognition.

Our contributions of this work include: (1) Analysis of formative studies identifying diverse use cases for media augmented VCs. (2) A rehearsal-to-performance workflow and tool, Clio, for augmenting live VCs with direct manipulation using voice, body, and devices using an immersive authoring approach. (3) A series of user studies with Clio, resulting in insights into how presenters prepare and present live VC performances augmented with interactive direct-manipulation visual effects.

2 RELATED WORK

2.1 Immersive Authoring Tools

Immersive authoring allows users to experience and verify immersive content firsthand, creating it through natural and direct interaction within the same environment [27, 37, 47]. The benefits of an immersive authoring approach is that it provides as much agency and control over system behavior to the user as is possible [57]. Immersive authoring systems for virtual environments have been widely explored in HCI research and usually involve two steps: designing virtual behaviors and content, then mapping interactions between virtual contents and users [38]. While creating virtual content usually involves programming in an environment separate from where the user experiences the behavior, immersive authoring environments blend the authoring and behavior environments [38, 41, 60, 65, 70]. SceneCtrl [69] and Window Shaping [26], for example, enable authors to create in-situ virtual scene assets and static 3D models. Similarly, Calliope supports 3D design idea generation in VR by facilitating communication between users and a creative AI through traditional sculpting techniques [60]. These ideas are extended in other projects [8, 11, 12, 68] to allow virtual contents to be animated in-situ. Visual programming has also been explored as a candidate for democratizing authorship of interactive applications [18, 24, 38, 46, 54, 64, 71]. FlowMatic, for example, allowed a user to build and test virtual interaction models in real time by connecting user input to object parameters [71]. Previous systems were limited, however, by the range of input modalities, often requiring fiduciary markers [30, 38, 49, 55], or operating only using spatial location [23, 46]. Some prior works encourage the use of midair gesture and hand pose recognition as part of the authoring process [56, 63, 65] or the use of voice [20, 31, 50] and tablet [32, 58]. While many of the above approaches use head-mounted displays and explore avatar-based virtual reality immersive authoring, in this work, we explore how VC systems provide a promising alternative venue for immersive authoring. Furthermore, a mixed modality approach that allows authors to define their behavior mappings using a variety of input methods is an unexplored problem that we investigate in this paper.

2.2 Dynamic Media and Performance Interfaces

The proliferation of digital technologies made interactive media an increasingly prevalent, expressive, and powerful medium for communication, art, and design [34]. An integral component of the experience is the colocation of user with their interactive media [61]. Traditional methods for creating rich and dynamic performance-driven graphics either require significant post-processing expertise [1, 4], specialized preprocessing workflows [2], or programming.

However, postprocessing is obviously impossible for real-time performances, and is only appropriate for those who have the time and skill to do complex video editing and compositing. Numerous frameworks for programming dynamic media exist like Processing [52], openFrameworks [48], D3 [15], Flash [3], Unity [59], and ARKit [6]. However, beside requiring significant expertise, they are limited in their support for interactive capabilities. HCI researchers have explored approaches for democratizing the creation of dynamic interactive media by prototyping novel interfaces [35, 39], sketch-based interfaces [7], direct-manipulation interfaces [40, 60], and storytelling through data [36]. Other work explores applications of explanatory illustrations [29, 66, 72] and creating mappings between user-triggered actions and animated effects [67]. Mapping human motion to digital objects [13] and digital characters [17] has also been explored in performance-based systems. Other interfaces such as SketchStudio and Kitty support user-defined relationships and events by directly manipulating elements of an illustration representing an underlying relationship graph [28, 32].

Research and commercial systems which allow manipulation of colocated media (e.g. mmhmm) are most related to our work [44, 53]; it enables users to produce real-time full-body human performances augmented with videos. We extend these ideas to encompass a broader domain of interaction methods, including voice, tablet, midair gesture, and whole body pose estimation. Furthermore, our work examines how enabling users to combine these interaction methods in any way they choose empowers them to create powerful and expressive augmented performances. Finally, our work employs an immersive authoring paradigm allowing presenters to quickly evaluate the visual effects of their authored interactive media.

2.3 Multimodal Direct Manipulation

Systems supporting direct manipulation of content are in common-use but extremely limit performer agency and expression. For example, newscaster and weather-reporting software create the appearance of the performer interacting with their colocated media, but the visual effect is implemented using a Wizard-of-Oz method, meaning all visual effects are controlled by an off-screen person, removing any agency or control from the performer[45]. Early real-time systems that focused on manipulating graphical elements use gestures to communicate to an audience [9]. ChalkTalk [51] and performance-driven tools [53] require users to design graphic assets and other media, then map this media to interactive behaviors. Similarly, GestuAR [63] enables creating custom midair gestures that can be mapped onto behaviors using a head-mounted display. Other works have explored incorporating interactive digital whiteboards as part of the presentation environments [25], or incorporated wearable to assist with communication [16].

Prior work [13, 17, 56] demonstrates that voice commands, tablet interactions, and body movement corresponding to the presentation topic are an integral part of an effective performance, greatly enhancing the audience's understanding of the performer's content. Some research interfaces [33, 43, 62] have explored the potential for supporting improvised presentations and social networking apps with video filters make it simpler to generate real-time effects, but their expressiveness, applicability, and possible use cases are confined by the limited number interaction methods they support [52]. While previous research has investigated integrating direct manipulation using body or voice, these works focus on a single method of manipulation and do not support performer authorship and customizability [51, 60, 71]. However, prior literature has also demonstrated that systems that enable more than a single mode of input are more flexible for users to adopt [57], increase a user's sense of agency and fluidity within the system [10], as well as stimulate performer creativity and audience engagement [43].

Unlike the above systems, Clio enables presenters to author dynamic media behaviors through an RtP workflow using multiple interaction methods including speech, midair gesture, tablet and others. No existing work has explored the direct manipulation of visual effects using multiple modes of input customized by the presenter for real-time video communication. In addition, an increasing number of prototype research systems exploring direct manipulation in VCS are emerging, yet these systems have never been formally studied to understand the benefits, challenges and limits to their expressive capabilities.

3 FORMATIVE DESIGN STUDY

We conducted a formative design study with 8 participants to better understand the use cases and mental models of presenters when preparing and presenting with an immersive VC tool supporting content colocation such as mmhmm or Cameo. We first conducted a semi-structured interview with participants, after which they were asked to select a prepared packet of visual media from 16 topics. Using their visual media packet, they were then asked to prepare and present a presentation using a speaker and content colocation tool. Media was prepared for the participants to reduce the workload for presenters and normalize the conditions across participants. We asked participants to think-aloud during the presentation preparation process, and asked them questions regarding specific choices and preferences while they prepared. We conducted an exit interview after participants presented. 4 participants self-identified as women and 4 as men. All participants were deeply familiar with slide tools for VC presentations and also familiar with at least one immersive speaker/content colocation tool like mmhmm or Cameo.

3.1 Results and Discussion

Iteration: We found that, unlike conventional slide-based tool authoring that requires the presenter to prepare all visual media before presenting, participants continually added media as the presentation progressed. Preparation time was used more to select the appropriate media they may need and create an ordering. Participants playfully experimented with switching backgrounds, incorporating media, and using different system features such as the laser pointer. Playfulness and improvisation in presentation style was consistent across the preparation and performance portion of the study, where additional media was added on an as-needed basis during the presentation. Participants noted that ordering their media prior to presentation acted as a presentation outline, and not a completely concrete formulation of the presentation. Tensions emerged from the inability to display more than one image or piece of text, as well as the limited manipulation afforded by the system. *"I kept blocking the stuff in my slide with my head. It would have been more useful to move the text or other parts of my slide to other*

Table 1: Participant-suggested use cases with accompanying operations and interactions. *Interaction* refers to the input method used by the participant during the pantomime and *Operation* refers to the behavior performed by the interaction.

	Use Case	Operations	Interactions
p1	Portrait Portfolio Consultation	arrange, select, open/close menu, group objects, zoom 2D, highlight, laser pointer, dismiss, make transparent, expand collection	Gesture, Voice
p2	Project Presentation	arrange, select, open/close menu, dismiss, conjure, zoom 2D, laser pointer, next slide, previous slide	Keyboard, Mouse, Gesture
p3	Conference Q&A	conjure, text display, annotate object, annotate air, open/close menu, select, next slide, previous slide	Keyboard, Mouse, Voice
p4	Interactive Demo Session	open/close menu, arrange, dismiss, rotate 3D, zoom 2D, zoom 3D, highlight, select, activate	Keyboard, Gesture
p5	End-of-year Student Presentation	open/close menu, highlight, dismiss, collapse collection, group objects,dismiss, pull audience content into screen, arrange, next slide, previous slide	Gesture, Mouse
p6	TA Session	arrange, conjure, zoom 2D, annotate object, highlight, rotate 3D, dismiss, tangible proxy, select, next slide, previous slide	Tablet, Gesture, Voice
p7	Vacation Photo Presentation	annotate air, display text, arrange, trigger, dismiss, select, open/close menu, conjure, annotate object, highlight, tangible proxy, create virtual copy, group objects, next slide, previous slide, poll/quiz, add shape, screen grab	Gesture
p8	ASL Tutorial and Conversation	text display, conjure, track image (to gesture), screen grab, next slide, previous slide	Keyboard, Gesture

parts of the screen" (P4). We probed into this, and found participants conceptualized the text and visual media in their presentations as individual elements that might coexist in a slide presentation. Manipulating and revealing these elements on command is a limitation in flexibility in the current system.

Barriers Between Worlds: Presenters experimented with various background images native to the system and those contained in the media packets until the limits of presenter and environment interaction was reached, revealing an invisible barrier between the two worlds. When backgrounds were physical places and not static images or color, they were regarded as spaces instead of objects. P2, for example, used an image from the media packet of San Francisco as a virtual background, maneuvering around and interacting with the background image as if it were a physical place in which they were immersed. Similarly, P1 used a virtual background of a coffee shop native to the system, and placed media from their packet on the tables and walls as if they decorated the space. While speaker and content colocation tools allow for the insertion of live video into media content and vice-versa, it does not allow the two worlds to interact. *"The background changes are really fun…I wish I could move [the objects in the background] around like I was really there"* (P4). The inability of participants to interact with their media indicates a gap in immersion when using these systems.

Modalities: Some participants attempted to use different modes of interaction while exploring the system during the presentation preparation phase of the study. *"I don't see why I can't get a robot to change the slide for me when I say 'next'"* (P3). The tension between the presenter and media worlds was also evident in pain points around tool usage in current systems. *"I kept getting confused. I thought [the laser pointing feature] was for drawing and was confused when my pen marks wouldn't stay on the screen. Would love to draw on the screen using this tool and my iPad"* (P3). Some expressed a desire to use different modalities in conjunction with each other in order to manipulate their content. *"It could be like Star Trek where we tell something to 'zoom and enhance' and it enlarges [the image] automatically"* (P1). Others noted that constraining the mode of manipulation to the mouse posed key limitations. Participants also remarked that reliance on using the mouse as a primary modality of control were limited. *"You couldn't use the cool stuff on your phone. You'd have to have tiny fingers to drag things"* (P1). Employing multimodal direct manipulation while maintaining an immersive approach could ease these tension, blending the world of the presenter and their media together and approaching the rich interactive potential evident in augmented and virtual reality.

4 FORMATIVE PANTOMIME STUDY

Understanding how to blend performer and media content is difficult for a variety of reasons, including the wide variety of different embodied interaction possible (e.g. voice, body pose, etc.) and the lack of familiarity that many potential users have with these machine-learning enabled tools. Furthermore, controlling a system that uses a mixture of input modalities presents challenges

because it is unclear which mode of interaction is appropriate for controlling the authoring system. To better understand the need of presenting content in immersive environments using multimodal direct manipulation, we deployed a questionnaire and two-phase semi-structured interview with pantomime study of potential user presentation processes using a think-aloud methodology. The study took place in two 30 minute sessions with 8 participants. 4 of our participants self-identified as women and 4 as men. At the conclusion of the first phase, participants were asked to think of a use case where they might present something in an immersive environment using direct manipulation. The second phase took place on a separate day from the first, and participants were asked to pantomime their use case twice using whatever input modality they felt appropriate. Participants first pantomimed their use case while thinking-aloud, describing their thought processes and what they were trying to accomplish as they proceeded. During this pantomime, participants were encouraged to use any method or tool they deemed necessary to communicate their idea including physical props. After completing their pantomime the first time, participants were asked questions regarding their choices and rationale behind different choices and behaviors. Participants were then asked to perform their pantomime again uninterrupted. Based on our observations and their spoken explanations, we segmented the data into individual interactions that were then analyzed along several dimensions. This approach allowed us to compare the use of gesture, voice, external devices, mouse and keyboard, and identify common usage patterns.

4.1 Results

We refer to the media manipulation behaviors proposed by participants, such as selecting, arranging, highlighting, and zooming an element as *operations*; Figure 2 shows how many participants suggested each operation. They also suggested a variety of input methods, such as voice, gesture, and mouse, and we call these *interactions*. A complete list of the use cases, interactions, and operations evident in these pantomimes can be seen in Table 1. In interviews, 75% of participants indicated that they would prefer to watch a 30 minute presentation that used visible gestures, voice, or other noticeable interactions versus hidden interactions like mouse and keyboard or discreet static hand gestures. Similarly, 63% of participants indicated that they would prefer to perform a 30 minute presentation using visible interactions, and 38% of participants indicated that they would like to both watch and perform this way. This means that while many participants indicated they would prefer to watch presentations using dynamic embodied interactions interactions, they might prefer to give presentations using discreet modalities such as mouse and keyboard or static gestures, and vice-versa.

4.2 Design Considerations

Based on the above observations, we synthesize the following design considerations and goals to guide the developing a system to support for semi-extemporaneous presentations in virtual environments. In the next section we describe a workflow extracted from our formative interviews that supports these design considerations.

Table 2: Occurrences of specific operations presented by participants during their use case in Phase 2 of the study

Operation	Count	Operation	Count
Group Objects	3 (38%)	Expand Object Group	3 (38%)
Next Item	6 (75%)	Previous Item	6 (75%)
Tangible Proxy	2 (25%)	Create Virtual Copy	1 (13%)
Screen Grab	1 (13%)	Push/Pull Content from Chat	1 (13%)
Poll/Quiz	2 (25%)	Laser Pointer	4 (50%)
Add Shape	2 (25%)	Remove Shape	2 (25%)
Alter Shape	2 (25%)	Composite Shape	2 (25%)
Rotate 3D	2 (25%)	Make Transparent	2 (25%)
Activate/Trigger	4 (50%)	Select	8 (100%)
Open Menu	6 (75%)	Close Menu	6 (75%)
Highlight	5 (63%)	Draw on Object	5 (63%)
Draw in Air	5 (63%)	Arrange Objects	6 (75%)
Zoom 3D	2 (25%)	Zoom 2D	6 (75%)
Dismiss	5 (63%)	Conjure	6 (75%)

D1. Maintaining Immersion: To better nurture presenter intuition for how the system will support their presentation, there should be little difference between the interface in which content is prepared and the interface in which it will be presented. Embracing immersive authoring shifts the focus of the presenter from preparing slides and media to the act of delivering the presentation.

D2. Experimentation with Different Interaction Methods: Presenters should be able to experiment with various interaction methods and tools as part of their preparation process. This could alleviate the mistrust expressed by participants when using tools and interaction methods with which they are not familiar. Subsequently, a presenter should be able to rapidly switch between interaction methods to experiment with as many as possible, as well as to adapt their presentation to their current contextual needs. A system that leverages the tools afforded by virtual environments should integrate exploration and experimentation of interaction methods as part of the presentation preparation process.

D3. Nonlinear Presentation Style: As discussed above, one of the weaknesses of traditional virtual presentation tools is their strict sequential nature, making improvisation and extemporizing difficult. One of the benefits of systems that colocate the presenter with their content is the flexibility to support playfulness and improvisation. A system designed to blend the worlds of presenters and their content should support flexible modes of presentation to enable a variety of presentation styles.

D4. Direct Manipulation: Finally, blurring the worlds of the performer and their digital media requires direct manipulation of content. Supporting a variety of methods to perform such manipulations enables presenters to ensure modes of interaction are appropriate for specific presentation media, as well as present material using means most intuitive to them.

5 REHEARSAL-TO-PERFORMANCE WORKFLOW

Speaker/content colocation with systems like mmhmm is one form of immersion, which we extend to incorporate multimodal direct manipulation. Based on the observations from our formative studies, we conceived a workflow that addresses the participants' needs and describes a mental model exhibited by participants while authoring and presenting immersive speaker/author colocation performances during our formative studies. This *rehearsal-to-performance* model replaces traditional approaches to presentation preparation with an immersive authoring approach (D1) that supports direct manipulation of media content (D4). Our model is based on a *rehearsal phase* that supports presenters in iteratively preparing their presentations. The key distinction between the RtP framework and conventional workflows using slides presented within a VC to present content is that that presentation media is authored using the same interface with which it is presented. Performer behavior with the interface changes between the rehearsal and performance phases, but the interface itself does not. During the rehearsal phase, presenters experiment with different interaction methods and operations to identify which behavior mappings between operations and interaction methods are appropriate for the context of their presentation and appropriate for the media they are presenting (D2). During the rehearsal phase, presenters can organize their media and experiment with different behavior mappings using the mapping interface. Once presenters feel sufficiently prepared, they shift to a *performance phase* in which they use the outline as the architecture of their presentation. Our workflow shifts the focus of presentation authoring from content (e.g. slide) authoring to performance authoring, which encourages presenters to focus on preparing the talk itself instead of on the artifact of the talk. Since the presentation media is not bound to any sequential order in this workflow, presenters can discuss their presentation material in any order they may choose (D3) and add additional media or change any element of their presentation as needed. Figure 2 shows this workflow.

5.1 Interaction Design Overview

A presenter begins by importing media, such as images, videos, external web links, and graphs, into the system. After importing the media, the presenter enters the rehearsal phase where they organize their presentation (See Figure 1). The default behavior mapping between operations and interaction methods lets the presenter use a mouse to interact with their imported media. Clicking the menu bar on the left side of the screen makes a tray containing thumbnails of the uploaded media appear. This tray can be moved to any side of the screen by clicking and dragging to the desired location. The user can scroll through media in the open tray and choose to click and drag entries onto the rest of the screen (D3). Here they can perform any supported operation (Section 6.2) using the mouse and keyboard as the controller.

If the presenter clicks the "begin mapping" button in the top right hand corner of the screen, a menu appears that lets them customize the mapping of operations to interactions (D2). After modifying the mapping, the menu closes and the presenter can freely experiment with their new mapping. This enables presenters to quickly evaluate different interaction approaches and find the behavior mapping between interactions and operations appropriate for their performance. An overview of the mapping interface can be seen in Figure 3.

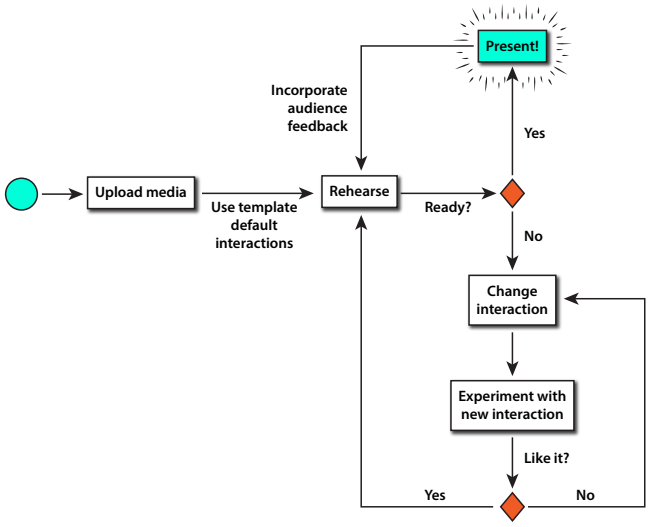

Figure 2: Clio's rehearsal-to-performance workflow derived from the formative study.

6 CLIO SYSTEM DESIGN

To explore the feasibility of a RtF workflow, we created Clio, an immersive authoring and presentation support tool for developing and delivering semi-extemporaneous presentations in virtual environments. Clio is prototyped using a proprietary research system that allows HTML, CSS, and JavaScript code to be overlaid on top of a camera video feed. Much like other VC augmentation systems (e.g. mmhmmm), Clio can be fed directly into the video source of commercial VC systems, using the built-in camera common to contemporary commercial laptops, thus requiring minimal setup. The main interface and logic is coded in Javascript that is rendered in a Chromium browser embedded within the application interface. We refer to the supported input techniques, including voice, mouse, body pose and midair gesture control, as *interactions*, and the resulting behaviors, such as arranging media and drawing on screen, as *operations*. The user is able to create *behavior mappings* that associate an interaction with an operation using a behavior mapping menu.

6.1 Interactions

This section describes the interaction methods supported by our system. We chose them because of their frequent occurrence within our formative study.

Mouse, tablet touch, and keyboard: The mouse is the default mode of interaction with our system since users from our formative study reported the most comfort and familiarity with this interaction. We also provide similar support for tablet touch interactions, which behave similarly, using a finger or stylus. Keyboard interactions can perform operations and provide text for operations that need it.

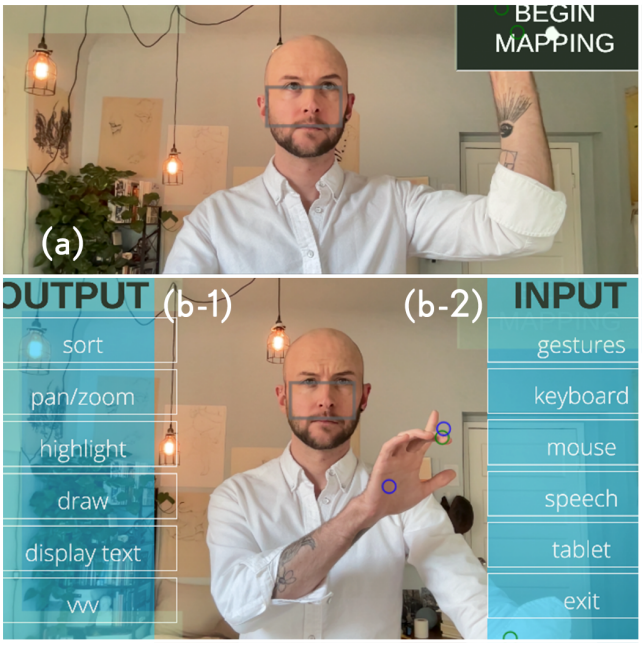

Figure 3: Clio's behavior mapping system. (a) Presenter is able to activate mapping menu at anytime during rehearsal or presentation. (b-1) Inside the menu, the presenter selects one operation from the left side of the menu (b-2) the presenter selects one interaction method from the right side of the screen.

Voice: Presenters can perform specific operations (e.g. displaying text) using their voice. They can specify specific keywords that will perform an operation. Two modes of performing operations are supported, inline and stand-alone. Inline mode parses all detected speech for keywords without requiring the speaker to pause. It is more discrete since the presenter can weave keywords into their presentation text and perform operations in a hidden manner. The stand-alone method requires a break in verbal speech before a keyword can be recognized and perform an operation. Switching between these two keyword modes can be done through the mapping menu during rehearsal or presentation if needed.

Gesture: Midair gesture and body-driven interactions are also supported by Clio. The underlying system provides hand-joint and body-pose data based on the video feed. Our algorithm calculates poses by sampling the positions of fifteen hand-joint angles and averaging these time-series joint angle samples to retrieve the pose of each finger. From here, each finger position is classified open, closed, or bent. We define a hand gesture to be a collection of required states for each finger. If each finger approximates the required state associated with a specific gesture, that gesture is considered active. Our initial implementation of Clio supports 18 static gestures following the example of previous studies which revealed that typical midair gestures are comprised of two components; local properties meaning how the hands are posed, and global properties, indicating the location and movement direction of the palms [63]: *closed fist, open palm, index finger extended, middle finger extended, ring finger*

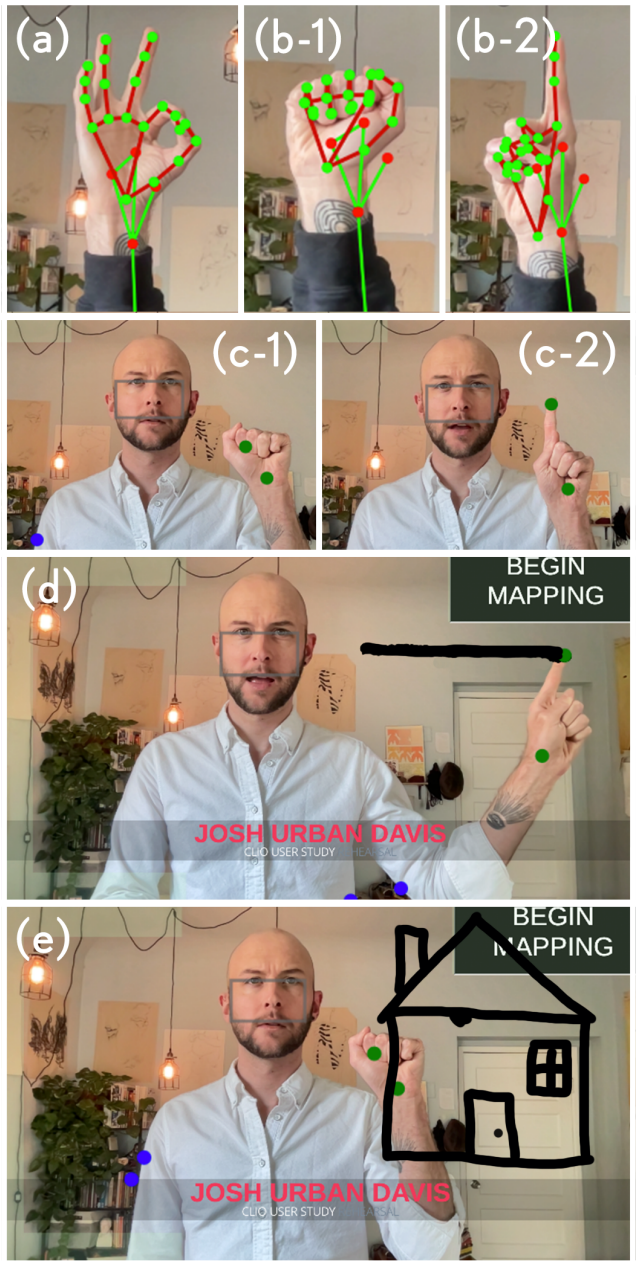

Figure 4: State-based dynamic gesture detection system. (a) Finger-joint angles used for hand pose detection. (b-1) We detect one of eighteen static hand poses. (b-2) We then look for sequences of static hand poses to detect the beginning and ending of dynamic gestures. (c-1) First state gesture in the sequence (closed fist). (c-2) Second state gesture in the sequence (extended index finger).(d) User performs dynamic gesture. (e) User ends the dynamic gesture by returning their hand from the second state pose to the first state pose.

extended, thumb extended, pinkie extended, two-finger pinch, four finger pinch, index and thumb pinch, index and middle extended,

index middle and thumb extended, index and thumb extended, index and pinkie extended, cupped hand (open), cupped hand (closed), index middle and thumb bent. To enable dynamic gesture recognition, we define a dynamic gesture to be a sequence of static gestures. The static gestures must be made in a specific order to perform the operation associated with the gesture. Once performing, the operation remains active until the gesture sequence is repeated in the opposite order. For example, the initial static gesture for 'grab' is an open palm, the second static gestures is a closed fist. Performing these in sequence performs the operation mapped to this gesture by the user in the behavior mapping menu. To end the operation, the user shifts their gesturing hand back to the open palm that began the interaction sequence. Two-handed interactions are only recognized by Clio when both hands are present and performing the requisite gesture sequences.

6.2 Operations

Operations are the ways in which a presenter can manipulate their presentation content using their chosen interactions. We support nine operations, summarized below and described fully in Appendix A (See Figure 5). We chose these because of their frequent occurrence within our formative study as well as their prevalence in existing literature [5].

- Arrange: enables the user to move virtual objects around the screen
- Pan/Zoom: enables participant to enlarge or shrink a virtual object
- Draw: supports annotation anywhere on the screen or on a virtual object. Colors, brush size, and opacity can be changed
- Text Display: User specifies what text they would like to display on the screen and where
- Highlight: Creates a glowing effect around a virtual effect
- Conjure: Makes a specific virtual object appear on the screen
- Dismiss: Makes a specific virtual object disappear
- Next: Cycles to the next virtual object in the media tray (a more general approach to "next slide" in a traditional slide presentation)
- Previous: Reverts to the previous virtual object in the media tray (a more general approach to "previous slide" in a traditional slide presentation)

7 STUDY 1: VALIDATION OF FORMATIVE STUDY USE CASES

To understand whether Clio effectively supports the immersive authoring and multimodal direct manipulation needs expressed by presenters in our formative study, we conducted a follow-up investigation with some of our original participants. Five participants (p1, p2, p3, p5, p8) had pantomimed presentations that were based on operations that we implemented in Clio; the other three (p4, p6, p7) relied heavily on operations like *3D object rotation* and *tangible proxies* that we did not include in our first version. These operations were only used by a few participants, and fully understanding their needs and uses requires a subsequent independent investigation. Excluding these therefore will not affect our findings. Each of the five invited participants used Clio to create and perform an actual presentation that closely resembled their original pantomimed presentation. We gathered statistics on their activities and conducted a post-task interview that included both Likert scale and open-ended questions. Our Likert questions used a scale of 1 to 5 with "1" signifying "not at all" and "5" indicating "very much so".

Results and Discussion: All participants were able to present their use case with Clio. Presentations on average took 2.42 minutes (SD = 0.55) and required an average preparation time of 24.32 minutes (SD = 2.52). This time includes a tutorial session to familiarize participants with the system, as well as time for the participants to experiment with the various capabilities of Clio. Overall, participants found the system easy to use (Q:B.4.17 "It was easy preparing and giving virtual presentations using the prototype" AVG = 4.67 SD = 0.33), intuitive (Q:B.4.19 "It was intuitive to prepare and give my presentation using the prototype" AVG = 4.33 SD = 0.64), and capable of supporting the use case conceived during the formative study (Q:B.4.31 "My presentation closely resembled my pantomime during the previous phase of the study" AVG = 4.25 SD = 0.96). *"It's so fun! I can imagine ways this could be used for art as well as presentations. It's so playful"* (p1). A full list of questions asked can be seen in Appendix B.4.

Commonalities in behavior mappings and interaction preferences were evident among presentations with similar contexts and levels of formality. For example, p2 and p5 both involved presentations in professional environments with more structured presentation requirements, and subsequently used more traditional keyboard/mouse behavior mappings. P1 and P8, however, presented use cases that were more informal and playful, and thus embraced more overt behavior mappings such as arranging using midair hand gestures and pan/zooming using the custom keyword "enlarge". Most participants used behavior mappings similar to those presented using pantomime during the formative study. P3 also presented a formal use case (conference Q&A) and pantomimed mostly mouse/keyboard behaviors during the formative study. However, while using Clio, P3 deviated from their original pantomime, and used hand gestures discreetly to switch presented media. When asked about this change, P3 explained that they were comfortable using hand gestures for small interactions, and that using these gestures instead of the mouse would allow them to potentially move around their physical space while presenting. While most of the use cases suggested by our participants were conceptualized as remote VC presentations, p3's adaptation suggests that the benefits of immersive multimodal mapping could transfer to in-person presentation contexts.

Three participants (p2, p3, p5) noted that the system caused them to rethink their presentation approach. *"I had to rethink my presentation a bit since the system works different than PowerPoint. You can go in order but you can also deviate and that has me rethinking my talk from the ground up"* (p2). Furthermore, participants expressed curiosity about what it would be like to compose a presentation knowing the capabilities and workflow of Clio ahead of time, as opposed to implementing a previously-conceived use case. For this purpose, we performed a second user study to examine how users prepare new presentations.

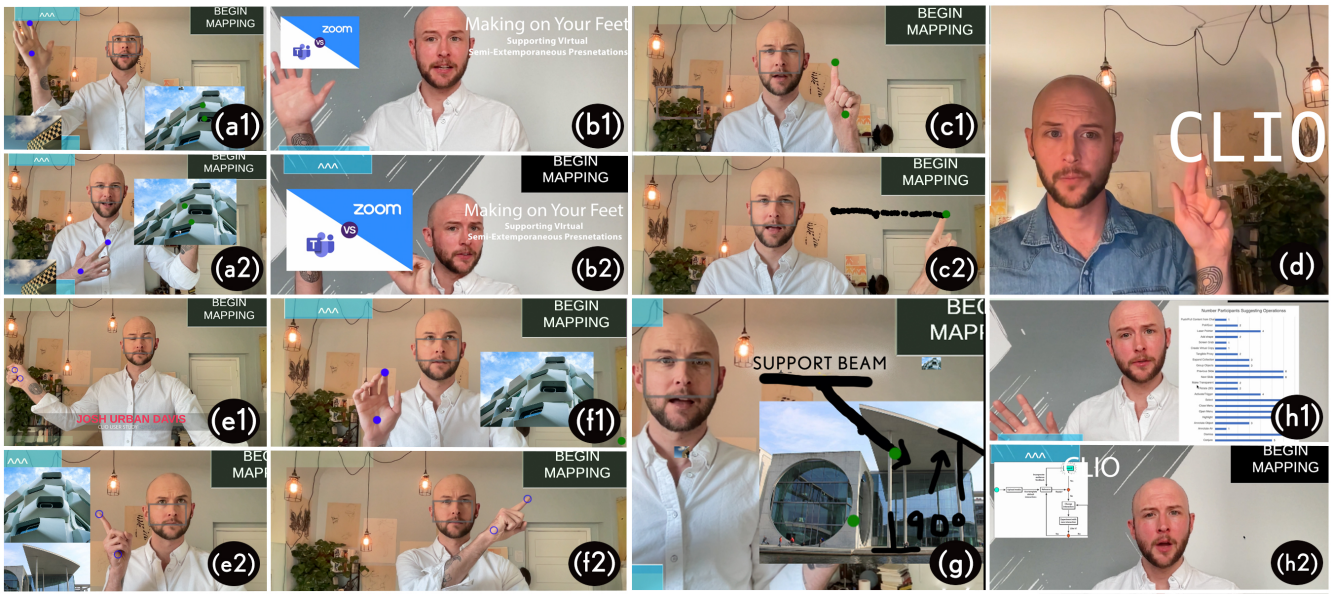

Figure 5: Operations Supported by Clio. (a1) Arrange: the user places their hand over the image and creates the Arrange gesture. (a2) the user moves the image to the desired location and stops making the Arrange gesture to place the image in the desired area. (b1) Pan/Zoom: the user places both hands over the image and makes the Pan/Zoom gesture. (b2) the user moves their hands away from each other to enlarge the image. (c1) Draw: the user makes the draw gesture by raising their index finger. (c2) The user draws on the screen with their index finger. (d) Display Text: the user makes the Display Text gesture (raises two fingers) and speaks to display their spoken words as text on screen. (e1) Conjure: The user makes the Conjure gesture (e2) Images appear on screen where Conjure gesture was initiated. (f1) Dismiss: The user makes a dismiss gesture. (f2) Image disappears when user makes Dismiss gesture over visible image. (g) Highlight: similar to draw gesture, the user makes the highlight gesture (three fingers) to draw highlights on the screen text. (h1) Next/Previous: the user makes the Next or Previous gesture (thumb right and thumb left respectively) to hide all presently visible media and text and move to the "next" or "previous" image in the media library, similar to moving to the "next" or "previous" slide in a slide presentation.

8 STUDY 2: NATIVE IMMERSIVE DIRECT-MANIPULATION PRESENTATIONS

To understand how immersive authoring and multimodal direct manipulation enabled by Clio affects computer-mediated communication, preparation, and presentation for both audiences and presenters, we designed a 3 participant group study methodology. These groups comprised the presenters, the audience, and an external group of commentators. Our study sought to understand how an immersive authoring approach affects presentation preparation and performance, how different modalities used in presentations are received by audiences, does an RtP workflow develop trust between user and machine, and how is the presentation of information different using direct manipulation?

8.1 Participants and Task

Our study comprised 6 presenters who prepared and gave presentations, 12 audience members who watched the presentations, and 5 external commentators who observed the presenters and audience members. Of the 6 presenters, 2 identified as women and 4 as men, and ranged in age from 28–44 with an average of 32. Of the 12 audience members, 7 identified as women and 5 as men, and ranged in age from 24–61, with an average of 35. Of the 5 external commentators, 3 identified as women and 2 as men, and ranged in age from 26–62 with an average of 44. All participants had prior experience with virtual presentations; 78% indicated that this experience was largely with slideshows shared over Zoom, Google Meet, or Microsoft Teams. Additionally, all had experience giving virtual presentations with the exception of one member of the audience group. One member of the audience had previous experience using voice and gestures to interact with virtual presentations through their art-making practice.

8.1.1 Presenters. All presenters were presented with ten potential presentation topics to choose from similar to the prepared media packets used in our formative study (See Section 3.1). After selecting their presentation topic they were given a collection of media that they could use use in a presentation on their topic. No topic was chosen more than once. Presenters were then introduced to Clio and given a brief tutorial on Clio's use. Following this tutorial, presenters were then left alone to rehearse with Clio, changing interaction/operation mappings and experimenting with their assigned media packet. Once all presenters were adequately prepared, our virtual audience was brought in to watch the presentations. Each presenter took a turn presenting their rehearsed presentation

using Clio to the live virtual audience. After all presenters had completed their presentations, they were given an exit questionnaire containing free response and Likert scale questions to probed their experience presenting with the prototype, as well as the overall rehearsal-to-performance workflow.

8.1.2 Audience. Audience members were gathered on Microsoft Teams to watch the virtual presentations. After watching each of the presentations, audience members were given an exit questionnaire with free response and Likert scale questions to examine their experience watching the virtual presentations. Both the virtual presentations and the audience gallery was recorded during the live presentations.

8.1.3 Commentators. After the live virtual presentations were completed, the recordings of the audience gallery and presentations were shown to 5 commentators. We asked these commentators to make observations about the audience and overall impression of the live presentations using a questionnaire.

9 RESULTS

All 6 presenters were able to successfully complete the rehearsal and performance sessions. On average, participants required 31.06 (SD=1.45) minutes to complete the tutorial and rehearsal process for their presentations, which ran 2.52 (SD=.34) minutes on average. Much of this time was spent familiarizing the presenter with the system, as well as allowing the presenter to explore and experiment with as many mappings as they wanted.

9.1 General Impressions and Usability

Presenters indicated that they liked giving their presentation using Clio (AVG=4.67 SD=0.49) and that they would like to experiment giving their talks using different interaction mappings. *"I feel like I played it safe and really would have wanted to use voice or gesture or something more flashy…I think it would be better at keeping people's attention"* (pr5). All presenters indicated that they would rather watch a presentation using Clio than traditional presentation tools (AVG=5.00 SD=0.0). Presenters overall agreed that Clio was easy to use (AVG=4.33, SD=0.67), intuitive (AVG=4.67, SD=0.33), and fun (AVG=5.0, SD=0.0). Audience members were generally excited by the promise of Clio as a presentation system. *"I was intrigued by the use of different presenting tools…It was interesting to see as an audience part of what the presenter was seeing—almost felt like an interesting two-way mirror [where we were] 'in the room' with the presenter"* (a3). Overall, the audience indicated that they enjoyed watching presentations using Clio (AVG=4.75, SD=0.25), felt generally engaged (Avg=4.66, SD=0.47), and would rather watch a virtual presentation using Clio than using traditional presentation tools (AVG=4.66, SD=0.33). While audience members indicated that they had noticed gesture and voice interactions, none had noticed the mouse or tablet being used. Preference was given to inline speech interactions vs stand-alone interactions due to the pause required for the latter to function properly being "distracting" (a12). Similar to remarks made by the presenters, audiences indicated that they felt more engaged because the presentation had a more conversational feel, and thus they were more inclined to ask questions and interact with the speaker (AVG=4.33, SD=0.96). *"It feels like it's more natural to have moments where questions can come up organically and the presenter and audience don't feel like they're interrupting each other"* (a2). Commentators agreed that the audience was engaged (AVG=4.75, SD=0.2) and the presentations were more engaging than traditional presentations (AVG=5.00, SD=0.0). As the presentations progressed, more virtual items were moved onto the screen by the presenter. A few times during the presentations, the images on the screen occluded the presenter's face from the audience, or images consumed the majority of the screen. *"[the audience] start looking off. Checking their phones. They lose interest when there's too many things on the screen"* (c5). Commentators also noted that the gesture interaction seemed to garner the most attention, followed by voice. However, commentators observed that the engagement produced by voice seemed to wane over the course of the presentations. *"The excitement wears off…it feels like a natural part of the presentation."* (c5). Similar to the audience response above, commentators also did not notice the mouse or tablet being used. Additionally, speech interactions using the inline method discussed in section 5.1 also went unnoticed. When asked about this commentators responded *"Sometimes I could follow the person's hands and I knew what was about to happen but sometimes things just moved and I had no idea how…it seemed like magic"* (c1). Full results of Likert study can be seen in Figure 6.

Modality Preferences All interaction methods and operations were used at least once across the 6 presenters and all participants used at least one gesture and one voice interaction. Participants reported that they experimented with at least 3 interaction methods during their rehearsal period, and ultimately gravitated towards using one more predominantly than the others. All presenters experimented with gesture and mouse, while half experimented with speech. 50% of presenters reported that their overall favorite interaction method was midair gestures, 33% favored speech, and 17% preferred the mouse. Similar to Study 1, we observed a correlation between content formality, presentation style, and behavior mappings. For example, p3 presented "architecture" and employed a more formal style and conventional behavior mappings using mouse and keyboard. Compare this to p2 who presented "weird trees", employing mostly voice and overt gesture behaviors. Individual preferences for specific interaction methods varied based on presenter choice of mappings. *"Controlling the presentation with speech and gestures made the presentation feel more 'in person' than the zoom presentation + head screen"* (pr2). While presenters on average suggested they would rather deliver a virtual presentation using the prototype than with traditional presentation tools (AVG=4.0, SD=0.49) they also indicated that different mappings were appropriate for different presentation contexts. *"For something that has to be tightly scripted (time limit, etc.) doing prerecorded video or doing a slide deck with built-in animations would probably be preferable to having the presenter need to move things around"* (pr1).

9.2 Comparison to Existing Tools

Here we outline presenter and audience feedback comparing existing commercial tools with Clio. Participants were asked before and after presentations about their familiarity and experiences using various platforms. Since all participants (audience, presenters,

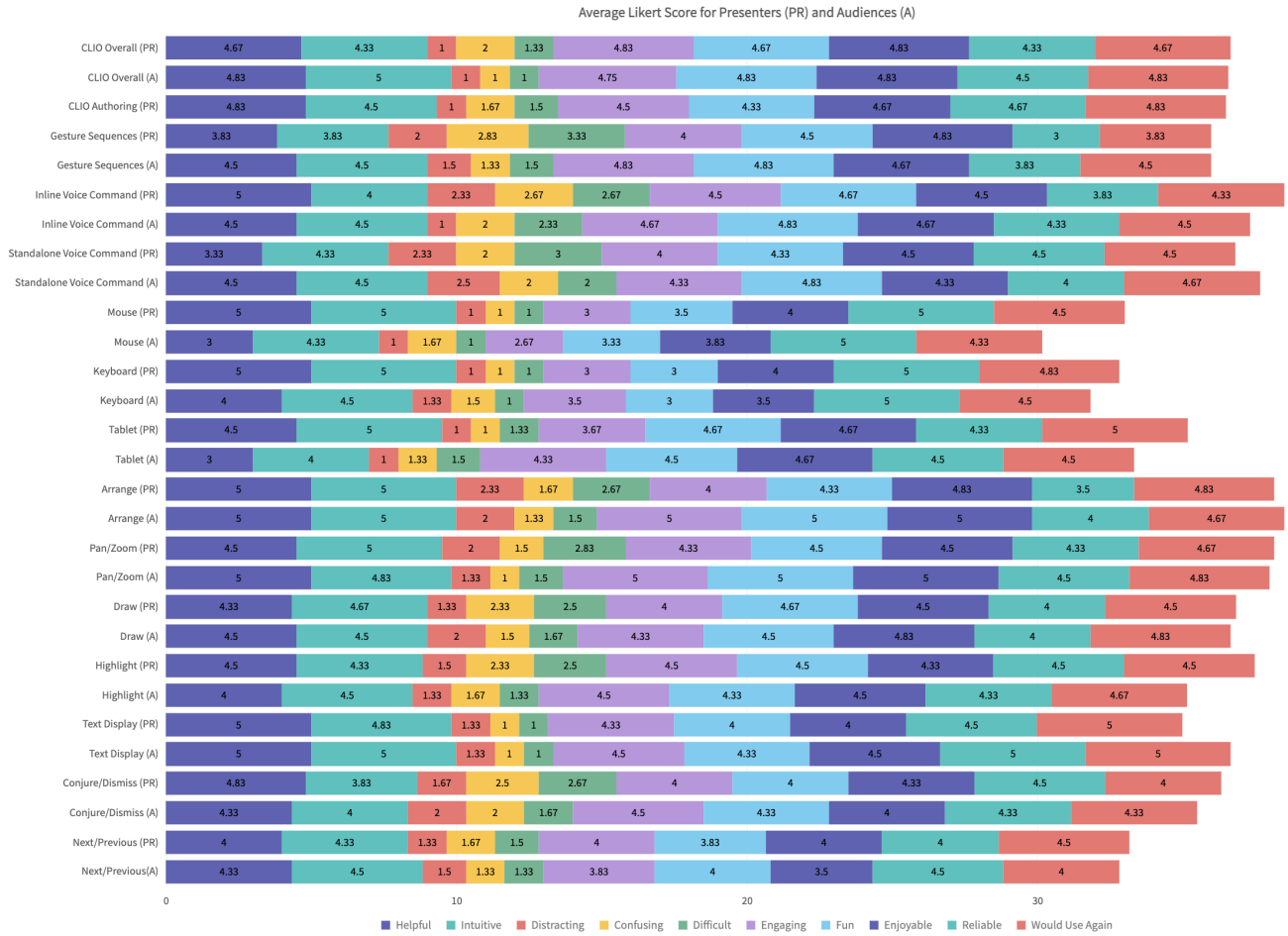

Figure 6: Average Likert Results from Validation Study for Presenters (PR) and Audiences (A) on a 5 point scale where 5 indicates "very much so" and 1 indicates "not at all". Axis labels reflect components of Clio. Color and size of bar indicates average Likert score reported for each dimension evaluated (See Legend)

and commentators) had significant familiarity with one or more commercial VC tool, participants were asked to directly compare their previous experience with these tools to Clio (See Appendix B.1 q6–25, B.2 q23,27, B.3 q3, 28, B.4 q3, 28, B.5 q5–11).

9.2.1 Slideshow w/VC:. All audience and presenters were familiar with traditional VC tools as well as slideshow presentation software used in combination to present graphics and media content. Several key pain points and distinctions were outlined by participants. *"Changing the slides can be clumsy [I have] to say "next slide" on zoom"* (a10). The participant references a context where the person presenting in a VC session is not the same person controlling the slides, or does not have their permissions correctly configured to present their material, and thus must rely on someone else to do so. Since Clio feeds directly into the video stream of commercial VC's, it is possible for Clio to alleviate this pain point by enabling the presenter to change visual media using gestures or other non-vocal modalities similar to how they would change slides. Similar sentiments that sharing the screen and changing slides is clunky was a common theme among both audience and presenters (a1–3, a5–11, p1–4, p6). These participants noted that using voice and gestures allowed them to change media content without touching their computer, potentially alleviating transition clunkiness, and could eliminate needing conference hosts to change slides for them. Additionally, presenters p2–4, and p6 noted that this could allow them to move about the room and still control the media displayed on the screen, untethered to their computers. Others (a1–2, a4–9, a12, p1–2, p5) also expressed that slideshows often made it feel as if there *"was no energy in the room"* (a4) or that there was *"some kind of disconnect between the person talking and everyone else"* (a9). Participants noted that colocating speakers with presentation content increased *"feeling the presence of the speaker in the meeting room... I felt like I couldn't check my email or twitter"* (a2). This result corroborates existing literature on the attention effect of inserting live video feeds into slideshow content [14], demonstrating that live video feeds support increased attention when colocated with presentation media.

9.2.2 Speaker and Content Colocation: Those familiar (a1, a3–5, a9–10, p1, p3, p5–6) with VC platforms that allowed live video footage of the speaker to be inserted into the presentation (e.g. Cameo, mmhmm, etc.) expressed the limitations of engagement resulting from colocation of presenter with content, noting that *"it feels static…[there is] some kind of disconnect between the person talking and the rest of the stuff on the screen"* (a1). Customizable manipulation using voice and gesture mitigates the gap between presenter and content by providing a bodily connection between performer and media behavior. *"Controlling the images and text with speech and gestures made the presentation feel more 'in person' that the zoom presentation plus head screen. It was very interactive."* (p2). All presenters echoed this sentiment, noting that manipulating with their media using different techniques created a stronger sense of agency over their presentation. *"My past virtual presentations were not as dynamic. This type of virtual presentation allowed for a lot of movement and left the door open to improvisation. With audience feedback you could be more fluid with what material you wanted to highlight and what material you could leave to the side."* (p3). Colocation of presenters with their media resulted in a more playful environment, enabling a looser presentation style that could encourage improvisation. Audience members (a1, a3–5, a9–10) highlighted an increased sense of presenter *"presence in the room…[the presenter] moved stuff around the screen like they could move [physical] things around in a real room"* (a10). This relationship between agency over digital objects and the resulting verisimilitude of virtual reality environments is evident in prior literature [22, 27, 47]. While some research suggests that multimodal manipulation of objects enables users to interact with media in the manner most suited to their needs and abilities, little work has documented the impact multimodal manipulation has on the believability of a VC environment.

9.2.3 Nonlinear and Brainstorming: (2 presenters and 4 audience members) Some presenters and audience members (p1, p3, a3–5, a12) were familiar with nonlinear brainstorming tools such as Miro and Whiteboard, and noted how often brainstorming occurred during VC discussion while simultaneously using these tools. Some (a4, a5) suggested that during such brainstorming co-design sessions, Clio could be *"Easier to be more responsive to [the] audience without disrupting flow"* (a4) and thus could support brainstorming and nonlinear tasks where a single person can facilitate the session. Others (a3, a12) were more interested in the multiuser aspect of these nonlinear systems, and suggested that developing methods for allowing multiuser interaction (e.g. Miro) would be an exciting direction for the Clio, and necessary for supporting brainstorming, mindmapping, and other nonlinear tasks. Both presenters, however, envisioned using Clio alongside existing nonlinear systems: *"my dream is to use this with Miro (which I use a lot at work) to teleport me to different parts of the board by voice and use gestures to navigate rather than mouse when presenting or facilitating a workshop"* (p1).

10 DISCUSSION

By employing multiple modes of direct manipulation, presenters blurred the barrier between themselves and their content and transformed static images into interactive characters and responsive environments. Some presenters used voice interaction to summon, move, and dismiss specific images from their packets, which they named and interacted with as if they were characters. Gesture interactions were also used to move and interact with these images, creating small, playful scenes between presenter and characters. *"I loved the dogs! They each had a unique personality and relationship with [the presenter]. Reminds me of playing with my dogs at home"* (a2). All audience members regarded characters animated using Clio as participants in the scene. Perceiving static images as characters interacting with the presenter suggests that the barrier between performer and content is sufficiently blurred. Other presenters held two images at a time, and made them speak to each other as if they were hand puppets. One presenter enlarged an image of a tree, and drew plans for an imaginary tree house to be built on top. *"I kept forgetting that [the presenters] weren't actually holding [physical objects] and that they were just moving images around the screen with their hands. It fooled me!"* (a4). These performances suggest that not only employing bodily direct manipulation blends the gap between presenter and content, but maintains an immersion seamless enough to author interactive skits with digital images.

Audience members remarked that using gestures resulted in more motion being evident during the presentation and this helped keep their attention. *"It was different [than traditional virtual presentations] because you weren't looking at a still image with text. There's a person there with you moving things around so you pay attention. It's cool."* (a4). Presenters indicated that they were skeptical of using interaction techniques that they were unfamiliar with before rehearsing (AVG=4.33, SD=0.33), but that they were much more trusting of these interaction methods afterwards (AVG=4.33, SD=0.67). Part of the reason was that participants felt a greater rapport with the system due to the rehearsal process (AVG=4.67, SD=0.47) and being able to quickly view the interaction mappings they had created during rehearsal (AVG=4.67, SD=0.49). *"I thought that this was pretty vital. Knowing where to stand and what distances were relevant were important. I think in general it's valuable to make practice as close as possible to performance."* (pr2).

User/Data Colocation and Screen Occupancy. Audience members suggested that they were more engaged with the virtual presentation using Clio, partially because they could see the speaker better than in traditional virtual presentations (AVG=4.4 SD=0.64) *"It made the presentation more personalized and allowed me to follow what they were saying. Since they could gesture to the images rather than the images occupying the whole screen meant that I could better associate what they were saying with what I was seeing. Versus having to just use my brain to read words on a screen and not being able to listen to the speaker simultaneously"* (a9). Presenters reported that their attention felt *'split'* (pr6) as more images, text, and annotation were added to the screen. *"I could really only manage three images on the screen at a time. More than that I started feeling overwhelmed and had trouble grabbing a specific image [using gestures]."* (p2). Interestingly, this sentiment was echoed by presenters who primarily used the mouse as opposed to gestures to manipulate objects on the screen. This suggests that interaction method and mapping are not the main factor contributing to the burden caused by screen occupancy. Despite this, all participants still felt more *"connected"* (pr5) to their media due to being colocated with it. *"I could indicated specific parts of the image using my body, or even*

reference common items between two images simultaneously. It was easy and really neat." (pr3).

Improvisation vs Linear Planning. After the rehearsal, all presenters remarked how they felt Clio lent itself to a more improvised mode of speaking. *"It felt less formal, like I was rehearsing how I wanted to walk through my content and ideas and less about structuring presentation materials. I was improvising with support from the software"* (pr1). This sentiment carried through the rehearsal into the performance as well. *"[It] felt more like a conversation in some ways with the audience (even though I only asked one question)—also a bit more like a performance"* (p1). While all presenters could prepare as much as they liked during the rehearsal phase, each indicated that they improvised during their presentations more than anticipated (AVG=4.67, SD=0.33). Presenters reported that this allowed them to follow live feedback and the general mood of the audience to direct their content or even alter their mapping. *"I also felt more freedom to change what I was doing as I went along instead of just reading prepared materials"* (pr3).

A Living Document. One pivotal difference between traditional presentations and Clio is the nature of the final artifact. This was noted by presenters, audience and commentators. *"You're not making slides, which is usually what I focus on when preparing a talk. Instead you're focusing on the presentation and performance. It's not static and can evolve over time. It's a living document"* (pr2). Many participants (pr1–3, pr5, a1–a6, a9, c1, c3) noted Clio made a similar difference in the final product. *"Now that I've seen this, slides seem more like notes that you can give someone. This is different… it's a performance and a presentation. It's alive"* (c3). Clio does not produce a final artifact such as slides that could be given to people for review, but a recording of the talk could be distributed. The semi-ephemeral nature of the final artifact produced by Clio parallels the ephemeral nature of semi-extemporaneous presentations given that both are meant to be experienced in the present moment.

Principal Modality In Section 6 we noted that similar operations were grouped together into the "arrange" operation, and that the interactions used in the mapping process change to match the interactions used to operate the authoring menu. Similarly, we consistently mapped the interaction mapped to the "text display" operation to manipulate text-based manipulations of the authoring menu (e.g. inputting keywords for voice operations). None of the participants commented or even seemed to notice this, suggesting that this adaptation created a transparent, natural authoring experience. Design of future multimodal systems should consider this collection of operations as a "principal modality", meaning that the interaction mapped to these operation is performed by users to control the authoring interface itself. Consistency in mapping of the interaction associated with "arrange" and "text display" with the interaction used to operate appropriate elements of the authoring menu maintains immersion in the authoring environment. Future systems incorporating mixed modes of input should consider consistently mapping these operations.

11 LIMITATIONS AND FUTURE WORK

Scalability of Presentation Style. While we did not explicitly give a time limit for our presentations, participants still opted to give shorter presentations. We collected speculative audience feedback regarding preferences for presentation style if presentations were longer, but knowing exactly how Clio would be used if presentations were scaled to long format is outside the scope of this initial investigation. Understanding how the RtF framework and direct manipulation immersive authoring would affect communication in longer format presentation is an ample avenue for further work.

Novelty Effects. While participants in our studies were generally enthusiastic about the potential for incorporating direct manipulation of media into VC communication, the benefits and preferences expressed during the study may change over long-term use of the system. It's difficult to disentangle how much enthusiasm expressed by participants is accounted by system benefits, and how much is a product of novelty effects. While the findings expressed in this work serve as a foundation for guiding the design of future direct-manipulation VC systems, studying the long-term effects of incorporating such a system into long-term practice remains the subject of future work. One way to approach questions of novelty effects is to deploy a longitudinal study, comparing how behavior and attitudes toward direction manipulation in VC systems shifts with extended use. Conducting such a study would be an ample subject for further research into direct-manipulation VC systems.

Direct Comparison Study. Since all participants in our validation studies had significant prior experience with slide-based VC presentations, we did not perform a direct comparison between Clio and traditional slide VC presentations. Our study instead relied on participants prior experiences with traditional slide-based VC presentations to compare with their experience using Clio. While this approach was sufficient for the purposes of this work and mitigated the demands placed upon our participants, a direct comparison study could yield additional insights into the differences between these two presentation styles. Conducting such a study is a topic for future investigation into VC system design.

Expanded Operation Vocabulary. As mentioned in Section 6, we only supported the most commonly cited operations elicited during our formative study. Subsequently, we were not able to revisit all of our participants from the formative study in our validation study. Many of these suggested operations are so rich that they would require their own study to completely explore. Proposed operations included 3D manipulation using voice and gesture and augmenting physical objects with virtual media (e.g. superimposing a virtual photograph over a blank index card). Future work will explore these avenues and how they interface with our proposed rehearsal-to-performance workflow.

Applications Beyond Presentations. Participants suggested a variety of use cases beyond semi-extemporaneous virtual presentations for employing a system similar to Clio. *"[It] felt like it would be super useful when navigating nonlinear media, like a Miro board, or while reviewing something where people will have questions and might need to jump back to something that was discussed earlier"* (pr1). Other suggested applications included drawing lessons, remote art studio sessions, virtual classrooms for children, remote litigation proceedings (depositions, trials, etc.), client pitches, storyboard and design brainstorming, musicians and composer collaboration or performance, ASL interpretation, and chatting on social

media. Of the future applications suggested by our presenters, audience, and commentators, education (87%) and creative applications (82%) were by far the most commonly suggested.

Artifact Production. It was noted by our participants that systems such as Clio produce a 'living document' which focuses the performer's attention on the presentation itself, instead of the artifact of the production (e.g. slides and written material). While this may be preferable in some contexts, certain use-cases necessitate the production of detailed content, visuals, written materials, or other artifacts to help audience members better understand the topic. For example, students may want a copy of slide material from a class lecture to use in studying. One approach to alleviating this concern is to incorporate artifact production into direct manipulation VC systems. For example, systems such as Clio could produce video content, or transcripts of the presentation to be reviewed by audience members at a later time. Other artifacts such as drawings or notes displayed as on-screen text during a performance could be made available after a performance has concluded. Some of these features are included as accessibility tools in commercial VC systems. Transcripts and post-session recordings are available in Teams, for example, as accessibility feature to help improve access to VC sessions to people who are d/Deaf or hard of hearing. Future work in this domain could explore how artifacts can be generated during performances to increase the accessibility and interoperability of session topics.

Accessibility. This paper explored the potential benefits of direct-manipulation in VC systems for able-bodied people, but many open questions remain regarding how a system like Clio would scale to users with different abilities. Gesture control, for example, could be difficult to perform for a person with motor difficulties, but speech control may be easier, and enable more expressive control of VC systems than mouse and keyboard. Similarly, understanding and placing visual media onscreen presents unique challenges for a person who is blind or visually impaired, but voice commands may be augmented to enable accessible direct manipulation, or wearable devices such as smartwatches could be employed to provide haptic feedback. How people with different abilities and preferences navigate a direct manipulation system such as Clio is an ample and complicated topic demanding a unique investigation to fully understand. We intend to perform a follow-up investigation to this work exploring how direct manipulation of VC media may be enabled by incorporating multi-modality, and how these tools affect the accessibility of VC communication.

12 CONCLUSION

Reimagining remote presentations to take advantage of immersive and directly manipulable environments requires us to rethink not only how we give presentations, but what a presentation should be. Adapting traditional tools like slides for virtual presentations results in a mental model where the artifact of the talk (slides) is the primary focus of preparation, instead of communicating ideas through speech. As demonstrated in our user study, Clio supports creating a 'living document' that centers communication and connection with the audience, as opposed to producing an artifact. We also observed that the playfulness of Clio helped to support extemporization because playfulness leaves room for error without sacrificing audience engagement. This playfulness is partially made possible by leveraging the multimodal capabilities of virtual environments such as machine-learning approaches to gesture and body-pose tracking, as well as voice commands. We also observed that direct manipulation created an immersion so seamless, it enabled presenters to interact with static images using voice and gestures as if they were animated characters. Our rehearsal-to-performance workflow suggests a new approach to presentation authoring and style of presentation deliverance that we believe will inspire future VC environment development.

REFERENCES

[1] Adobe. 2021. AfterEffects. Retrieved 2021-04-01 from https://www.adobe.com/products/aftereffects.html

[2] Adobe. 2021. Animate. Retrieved 2021-04-01 from https://www.adobe.com/products/animate.html

[3] Adobe. 2021. Flash macromedia player. Retrieved 2021-04-01 from https://get.adobe.com/flashplayer/about/

[4] Adobe. 2021. Premiere. Retrieved 2021-04-01 from https://www.adobe.com/products/premiere.html

[5] Roland Aigner, Daniel Wigdor, Hrvoje Benko, Michael Haller, David Lindbauer, Alexandra Ion, Shengdong Zhao, and Jeffrey Tzu Kwan Valino Koh. 2012. Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI. (Nov. 2012). https://www.microsoft.com/en-us/research/publication/understanding-mid-air-hand-gestures-a-study-of-human-preferences-in-usage-of-gesture-types-for-hci/

[6] Apple. 2021. ARkit | Augmented Reality development platform. Retrieved 2021-04-01 from https://developer.apple.com/augmented-reality/arkit/

[7] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. 2018. SymbiosisSketch: Combining 2D and 3D Sketching for Designing Detailed 3D Objects in Situ. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3173574.3173759

[8] Rahul Arora, Rubaiat Habib Kazi, Danny M. Kaufman, Wilmot Li, and Karan Singh. 2019. MagicalHands: Mid-Air Hand Gestures for Animating in VR. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 463–477. https://doi.org/10.1145/3332165.3347942

[9] Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: remote control of objects using free-hand gestures. *Commun. ACM* 36, 7 (July 1993), 28–35. https://doi.org/10.1145/159544.159562

[10] Frederik Brudy, David Ledo, Michel Pahud, Nathalie Henry Riche, Christian Holz, Anand Waghmare, Hemant Bhaskar Surale, Marcus Peinado, Xiaokuan Zhang, Shannon Joyner, Badrish Chandramouli, Umar Farooq Minhas, Jonathan Goldstein, William Buxton, and Ken Hinckley. 2020. SurfaceFleet: Exploring Distributed Interactions Unbounded from Device, Application, User, and Time. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 7–21. https://doi.org/10.1145/3379337.3415874

[11] Yuanzhi Cao. 2019. GhostAR: A Time-space Editor for Embodied Authoring of Human-Robot Collaborative Task with Augmented Reality. https://engineering.purdue.edu/cdesign/wp/ghostar-a-time-space-editor-for-embodied-authoring-of-human-robot-collaborative-task-with-augmented-reality/

[12] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376688

[13] Jiawen Chen, Shahram Izadi, and Andrew Fitzgibbon. 2012. *KinÊTre: Animating the World with the Human Body*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/2380116.2380171

[14] John Joon Young Chung, Hijung Valentina Shin, Haijun Xia, Li-yi Wei, and Rubaiat Habib Kazi. 2021. Beyond Show of Hands: Engaging Viewers via Expressive and Scalable Visual Communication in Live Streaming. 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 109, 14 pages. https://doi.org/10.1145/3411764.3445419

[15] D3. 2021. D3 | Javascript library for data-driven documents. Retrieved 2021-04-01 from https://d3js.org/

[16] Josh Davis. 2022. *PokerFace Mask: Exploring Augmenting Masks with Captions through an Interactive, Mixed-Reality Prototype.* https://doi.org/10.24251/HICSS.2022.398 Accepted: 2021-12-24T17:47:44Z.

[17] Mira Dontcheva, Gary Yngve, and Zoran Popović. 2003. Layered Acting for Character Animation. In *ACM SIGGRAPH 2003 Papers* (San Diego, California) *(SIGGRAPH '03)*. Association for Computing Machinery, New York, NY, USA, 409–416. https://doi.org/10.1145/1201775.882285

[18] Barrett Ens, Fraser Anderson, Tovi Grossman, Michelle Annett, Pourang Irani, and George Fitzmaurice. 2017. Ivy: Exploring Spatially Situated Visual Programming for Authoring and Understanding Intelligent Environments. In *Proceedings of Graphics Interface 2017* (Edmonton, Alberta) *(GI 2017)*. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine, 156 – 162. https://doi.org/10.20380/GI2017.20

[19] Travis Faas, Lynn Dombrowski, Alyson Young, and Andrew D. Miller. 2018. Watch Me Code: Programming Mentorship Communities on Twitch.Tv. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 50 (nov 2018), 18 pages. https://doi.org/10.1145/3274319

[20] Alexander J. Fiannaca, Ann Paradiso, Jon Campbell, and Meredith Ringel Morris. 2018. Voicesetting: Voice Authoring UIs for Improved Expressivity in Augmentative Communication. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173857

[21] C. Ailie Fraser, Joy O. Kim, Alison Thornsberry, Scott Klemmer, and Mira Dontcheva. 2019. Sharing the Studio: How Creative Livestreaming Can Inspire, Educate, and Engage. In *Proceedings of the 2019 on Creativity and Cognition* (San Diego, CA, USA) *(CC '19)*. Association for Computing Machinery, New York, NY, USA, 144–155. https://doi.org/10.1145/3325480.3325485

[22] David M Frohlich. 1993. The history and future of direct manipulation. *Behaviour & Information Technology* 12, 6 (1993), 315–329.

[23] S. Guven and S. Feiner. 2003. Authoring 3D hypermedia for wearable augmented and virtual reality. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.* 118–126. https://doi.org/10.1109/ISWC.2003.1241401 ISSN: 1530-0811.

[24] Valentin Heun, James Hobin, and Pattie Maes. 2013. Reality editor: programming smarter objects. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (UbiComp '13 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 307–310. https://doi.org/10.1145/2494091.2494185

[25] Keita Higuchi, Yinpeng Chen, Philip A. Chou, Zhengyou Zhang, and Zicheng Liu. 2015. ImmerseBoard: Immersive Telepresence Experience using a Digital Whiteboard. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2383–2392. https://doi.org/10.1145/2702123.2702160

[26] Ke Huo, Vinayak, and Karthik Ramani. 2016. Window-Shaping: 3D Design Ideation in Mixed Reality. In *Proceedings of the 2016 Symposium on Spatial User Interaction (SUI '16)*. Association for Computing Machinery, New York, NY, USA, 189. https://doi.org/10.1145/2983310.2989189

[27] Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. 1985. Direct manipulation interfaces. *Human-Computer Interaction* 1, 4 (Dec. 1985), 311–338. https://doi.org/10.1207/s15327051hci0104_2

[28] Rubaiat Habib Kazi, Fanny Chevalier, Tovi Grossman, and George Fitzmaurice. 2014. Kitty: Sketching Dynamic and Interactive Illustrations. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 395–405. https://doi.org/10.1145/2642918.2647375

[29] Rubaiat Habib Kazi, Fanny Chevalier, Tovi Grossman, Shengdong Zhao, and George Fitzmaurice. 2014. Draco: Bringing Life to Illustrations with Kinetic Textures. 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, 351–360. https://doi.org/10.1145/2556288.2556987

[30] Annie Kelly, R. Benjamin Shapiro, Jonathan de Halleux, and Thomas Ball. 2018. ARcadia: A Rapid Prototyping Platform for Real-time Tangible Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3173574.3173983

[31] Hyeongcheol Kim, Shengdong Zhao, Can Liu, and Kotaro Hara. 2020. *LiveSnippets: Voice-Based Live Authoring of Multimedia Articles about Experiences.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3379503.3403556

[32] 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 (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 831–843. https://doi.org/10.1145/3196709.3196736

[33] Everlyne Kimani, Dhaval Parmar, Prasanth Murali, and Timothy Bickmore. 2021. Sharing the Load Online: Virtual Presentations with Virtual Co-Presenter Agents. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 473. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451670

[34] Myron W. Krueger, Thomas Gionfriddo, and Katrin Hinrichsen. 1985. VIDEOPLACE—an Artificial Reality. *SIGCHI Bull.* 16, 4 (apr 1985), 35–40. https://doi.org/10.1145/1165385.317463

[35] James A. Landay and Brad A. Myers. 1995. Interactive Sketching for the Early Stages of User Interface Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 43–50. https://doi.org/10.1145/223904.223910

[36] Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith. 2013. SketchStory: Telling More Engaging Stories with Data through Freeform Sketching. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2416–2425. https://doi.org/10.1109/TVCG.2013.191

[37] Gun A. Lee, Gerard J. Kim, and Mark Billinghurst. 2005. Immersive authoring: What You eXperience Is What You Get (WYXIWYG). *Commun. ACM* 48, 7 (July 2005), 76–81. https://doi.org/10.1145/1070838.1070840

[38] Gun A. Lee, Claudia Nelles, Mark Billinghurst, and Gerard Jounghyun Kim. 2004. Immersive Authoring of Tangible Augmented Reality Applications. In *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR '04)*. IEEE Computer Society, USA, 172–181. https://doi.org/10.1109/ISMAR.2004.34

[39] Sang Won Lee, Yujin Zhang, Isabelle Wong, Yiwei Yang, Stephanie D. O'Keefe, and Walter S. Lasecki. 2017. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping of Interactive Interfaces. 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, 817–828. https://doi.org/10.1145/3126594.3126595

[40] Germán Leiva, Jens Emil Grønbæk, Clemens Nylandsted Klokmose, Cuong Nguyen, Rubaiat Habib Kazi, and Paul Asente. 2021. Rapido: Prototyping Interactive AR Experiences through Programming by Demonstration. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 626–637. https://doi.org/10.1145/3472749.3474774

[41] Germán Leiva, Cuong Nguyen, Rubaiat Habib Kazi, and Paul Asente. 2020. *Pronto: Rapid Augmented Reality Video Prototyping Using Sketches and Enaction.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376160

[42] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2018. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174040

[43] Fabrice Matulic, Lars Engeln, Christoph Träger, and Raimund Dachselt. 2016. Embodied Interactions for Novel Immersive Presentational Experiences. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1713–1720. https://doi.org/10.1145/2851581.2892501

[44] mmhmm inc. 2022. mmhmm. Retrieved 2022-09-15 from https://www.mmhmm.app

[45] NewscastStudio. 2022. Weather Solutions: 4 of today's leading weather software systems. Retrieved 2022-09-15 from https://www.newscaststudio.com/2017/06/21/tv-weather-solutions

[46] Gary Ng, Joon Gi Shin, Alexander Plopski, Christian Sandor, and Daniel Saakes. 2018. Situated Game Level Editing in Augmented Reality. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '18)*. Association for Computing Machinery, New York, NY, USA, 409–418. https://doi.org/10.1145/3173225.3173230

[47] Donald A Norman and Edwin L Hutchins Jr. 1988. *Computation via direct manipulation.* Technical Report. CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

[48] OpenFrameworks Community. 2021. openFrameworks | opensource toolkit for creative processes. Retrieved 2021-04-01 from https://openframeworks.cc/about/

[49] Jong-seung Park. 2011. AR-Room: a rapid prototyping framework for augmented reality applications. *Multimedia Tools and Applications* 55, 3 (Dec. 2011), 725–746. https://doi.org/10.1007/s11042-010-0592-1 Num Pages: 725-746 Place: Dordrecht, Netherlands Publisher: Springer Nature B.V..

[50] Fabio Paternò and Federico Giammarino. 2006. Authoring Interfaces with Combined Use of Graphics and Voice for Both Stationary and Mobile Devices. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Venezia, Italy) *(AVI '06)*. Association for Computing Machinery, New York, NY, USA, 329–335. https://doi.org/10.1145/1133265.1133335

[51] Ken Perlin, Zhenyi He, and Karl Rosenberg. 2018. Chalktalk : A Visualization and Communication Language – As a Tool in the Domain of Computer Science Education. *arXiv:1809.07166 [cs]* (Sept. 2018). http://arxiv.org/abs/1809.07166 arXiv: 1809.07166.

[52] ProcessingFoundation. 2021. Processing | software sketchbook for learning to code in the context of visual art. Retrieved 2021-04-01 from https://processing.

org/
[53] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300852
[54] Scratch Foundation. 2021. Scratch | code, create, share. Retrieved 2021-04-01 from https://scratch.mit.edu/
[55] Hartmut Seichter, Julian Looser, and Mark Billinghurst. 2008. ComposAR: An intuitive tool for authoring AR applications. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*. 177–178. https://doi.org/10.1109/ISMAR.2008.4637354
[56] Jinwook Shim, Yoonsik Yang, Nahyung Kang, Jonghoon Seo, and Tack-Don Han. 2016. Gesture-Based Interactive Augmented Reality Content Authoring System Using HMD. *Virtual Real.* 20, 1 (mar 2016), 57–69. https://doi.org/10.1007/s10055-016-0282-z
[57] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (Nov. 1997), 42–61. https://doi.org/10.1145/267505.267514
[58] 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*. Association for Computing Machinery, New York, NY, USA, 166–181. https://doi.org/10.1145/3379337.3415892
[59] Unity Technologies. 2021. Unity Real-Time Development Platform | 3D, 2D VR and AR Engine. Retrieved 2021-04-01 from https://unity.com/
[60] Josh Urban Davis, Fraser Anderson, Merten Stroetzel, Tovi Grossman, and George Fitzmaurice. 2021. Designing Co-Creative AI for Virtual Environments. In *Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3450741.3465260
[61] Andries van Dam. 1997. Post-WIMP User Interfaces. *Commun. ACM* 40, 2 (feb 1997), 63–67. https://doi.org/10.1145/253671.253708
[62] Andrés Vargas González, Senglee Koh, Katelynn Kapalo, Robert Sottilare, Patrick Garrity, Mark Billinghurst, and Joseph LaViola. 2019. A Comparison of Desktop and Augmented Reality Scenario Based Training Authoring Tools. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 339–350. https://doi.org/10.1109/ISMAR.2019.00032 ISSN: 1554-7868.
[63] Tianyi Wang. 2021. GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications. https://engineering.purdue.edu/cdesign/wp/gesturar-an-authoring-system-for-creating-freehand-interactive-augmented-reality-applications/
[64] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPturAR: An Augmented Reality Tool for Authoring Human-Involved Context-Aware Applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 328–341. https://doi.org/10.1145/3379337.3415815
[65] Matt Whitlock, Jake Mitchell, Nick Pfeufer, Brad Arnot, Ryan Craig, Bryce Wilson, Brian Chung, and Danielle Albers Szafir. 2020. MRCAT: In Situ Prototyping of Interactive AR Environments. In *Virtual, Augmented and Mixed Reality. Design and Interaction: 12th International Conference, VAMR 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 235–255. https://doi.org/10.1007/978-3-030-49695-1_16
[66] Nora S. Willett, Rubaiat Habib Kazi, Michael Chen, George Fitzmaurice, Adam Finkelstein, and Tovi Grossman. 2018. A Mixed-Initiative Interface for Animating Static Pictures. 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, 649–661. https://doi.org/10.1145/3242587.3242612
[67] Nora S. Willett, Wilmot Li, Jovan Popovic, and Adam Finkelstein. 2017. Triggering Artwork Swaps for Live Animation. 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, 85–95. https://doi.org/10.1145/3126594.3126596
[68] Hui Ye, Kin Chung Kwan, Wanchao Su, and Hongbo Fu. 2020. *ARAnimator*: in-situ character animation in mobile AR with user-defined motion gestures. *ACM Transactions on Graphics* 39, 4 (July 2020), 83:83:1–83:83:12. https://doi.org/10.1145/3386569.3392404
[69] 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 (UIST '17)*. Association for Computing Machinery, New York, NY, USA, 427–436. https://doi.org/10.1145/3126594.3126601
[70] Jürgen Zauner, Michael Haller, Alexander Brandl, and Werner Hartman. 2003. Authoring of a mixed reality assembly instructor for hierarchical structures: 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR 2003. *Proceedings - 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR 2003* (2003), 237–246. https://doi.org/10.1145/965400.965448 Publisher: Institute of Electrical and Electronics Engineers Inc..
[71] 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*. Association for Computing Machinery, New York, NY, USA, 342–353. https://doi.org/10.1145/3379337.3415824
[72] Bo Zhu, Michiaki Iwata, Ryo Haraguchi, Takashi Ashihara, Nobuyuki Umetani, Takeo Igarashi, and Kazuo Nakazawa. 2011. Sketch-Based Dynamic Illustration of Fluid Systems. *ACM Trans. Graph.* 30, 6 (dec 2011), 1–8. https://doi.org/10.1145/2070781.2024168

A APPENDIX: OPERATIONS REFERENCE

Below we detail the operations referenced throughout our paper. We will first discuss the operations that we do support:

Previous Slide: presenter is able to cycle to the previous slide in their prepared media. This became the more general operation "Previous" to allow the presenter to cycle through media other than slides like visual media.

Next Slide: presenter is able to cycle to the next slide in their prepared media. This became the more general operation "Next" to allow the presenter to cycle through media other than slides like visual media.

Activate/Trigger: presenter is able to begin specific behaviors using a trigger-action approach to mapping.

Select: presenter is able to select items from a menu of virtual objects on the screen.

Close Menu: presenter closes/collapses the chosen menu.

Open Menu: presenter opens/expands the chosen menu.

Zoom 2D: presenter is able to enlarge or shrink a 2D virtual object such as an image or text.

Laser Pointer: presenter is able to indicate various sections of virtual objects using a small dot which they control, similar using a laser pointer to indicate areas of interest on slides.

Expand Collection: presenter is able to expand a collection of objects away from each other to make them easier to view and manipulate.

Group Objects: presenter is able to cluster or group objects together to clean up the viewspace and make the collection easier to move.

Highlight: presenter is able to draw transparent colored annotations over a selected area to draw attention to it.

Annotate Object: presenter is able to draw on an object and have the resulting drawing track with that virtual object.

Annotation Air: presenter is able to annotate the viewspace itself.

Dismiss: presenter is able to remove virtual objects from the viewspace.

Conjure: presenter is able to make virtual objects appear within the viewspace.

Arrange: presenter is able to move, sort, and manipulate virtual objects around the viewspace.

***

The following operations were pantomimed during the second phase of our formative study but were not implemented:

Push/Pull Content from Chat: presenter is able to drag web links, text, or visual media into the viewspace or push media from the viewspace into the chatbox.

Poll/Quiz: presenter is able to generate and present a quiz for the audience to complete. Results can be displayed and manipulated on screen.

Add Shape: presenter is able to select from a variety of shape primitives such as circles, arrows, and others to appear onscreen.

Screen Grab: presenter is able to capture a section of the screen in an image and then use this image in their viewspace like any other visual media.

Create Virtual Copy: presenter is able to make virtual copies of virtual objects onscreen.

Tangible Proxy: presenter is able to use physical objects in their environment as targets for mapping images, animations, and behaviors. For example, a presenter may use index cards from their physical environment and map images to them. Thus, when they hold the index cards up to the screen, it appears as if they are holding the mapped image.

Make Transparent: presenter is able to lower the opacity of a virtual object.

Rotate (3D): presenter is able to rotate 3D assets in the viewspace.

Zoom 3D: presenter is able to enlarge or shrink a 3D virtual object.

B APPENDIX: QUESTIONNAIRES

Below we include the questionnaires used in our study to collect our qualitative and Likert results.

B.1 Demographics

Making on Your Feet Questionnaire 1 : Demographics

The purpose of this questionnaire is to gather demographic information and prior experience with virtual presentations as part of a larger study. "Virtual presentations" includes any variety of computer communication platforms including Zoom meetings with slides, YouTube videos, or social media live-streaming.

1. Would you like to be contacted about future studies related to novel interfaces?

2. What is your participant number? Ask our research team if you do not know

3. What is your age?

4. What is your gender identity?

5. What is your occupation? Do you have any disabilities that you'd like to disclose or request for accommodation?

If you answered "yes" above, how might we accommodate you in this study?

Experience with Virtual Presentations

"Virtual presentations" includes any variety of computer communication platforms including Zoom meetings with slides, YouTube videos, or social media live-streaming.

6. Do you have prior experience either watching virtual presentations? (virtual presentations include slide shows such as Power Point on telepresence software such as Zoom, video lectures on platforms such as YouTube, or similar media)

7. If so, think of one recent example. Could you describe how the presentation was delivered? What software, media, etc.

Do you have prior experience either giving virtual presentations? (virtual presentations include slide shows such as Power Point on telepresence software such as Zoom, video lectures on platforms such as YouTube, or similar media)

8. If so, think of one recent example. Could you describe how the presentation was delivered? What software, media, etc.

9. What are the strengths of this presentation style?

10. What are the weaknesses of this presentation style?

11. Have you ever watched a virtual presentation that was somewhat improvised? If so, please explain

12. Have you ever given a virtual presentation that was somewhat improvised? If so, please explain

13. Do you have any experience using or witnessing presentations involving gesture recognition, voice activation, or computer-aided visual augmentation? If so, please explain your experience Would you be comfortable using non-mouse and keyboard driven interactions for presentations? Please explain your answer.

Virtual Presentations—Likert Scale

Please rate the following on a scale of 1 to 5 with 1 meaning "not at all" and 5 meaning "very much so". If you have no experience with the questions material or are unsure, please leave the question blank.

"Virtual presentations" includes any variety of computer communication platforms including Zoom meetings with slides, YouTube videos, or social media live-streaming.

14. It is time consuming to prepare virtual presentations ("Virtual presentations" includes any variety of computer communication platforms including Zoom meetings with slides, YouTube videos, or social media live-streaming.)

15. It is difficult to prepare virtual presentations

16. I prefer watching virtual presentations with more images than text

17. I prefer delivering virtual presentations with more images than text

18. I am generally engaged when watching virtual presentations

19. I am more engaged watching virtual presentations in which I can see the speaker

20. When giving virtual presentations, it is easy for me to answer questions using my prepared content (slides etc.)

21. When giving virtual presentations, it is easy for me to answer questions mid-way through the presentation, or encourage interactivity with the audience

22. I feel compelled to ask questions or interact with the speaker when viewing a virtual presentation

23. I often interact or often see highly interactive virtual presentations

24. I like giving virtual presentations using traditional tools

25. I like watching virtual presentations using traditional tools

Closing thoughts

26. Is there anything further you would like to tell us about virtual presentations or any of the topics mentioned in the questions above?

B.2 Audience Post-Performance

Making on Your Feet Questionnaire 3 : Audience Post-Performance

The purpose of this questionnaire is to gather feedback regarding the performance of semi-extemporaneous presentations using a prototype creativity support tool.

Audience Impressions of Presentation

1. Please give an overall impression of the presentations you saw today

2. Were you engaged during the presentations? What factors contributed to your engagement?

3. Please comment on what you liked about this presentation style?

4. Please comment on what you did not like about this presentation style?

5. Please comment on how seeing the presenter co-located with their media affected the presentation

6. Were there certain interaction techniques you wanted to see used more? (interaction techniques include gesture, speech, mouse, etc.)

7. Were there certain interaction techniques you wanted to see used less?

8. Would you rather watch a long (30 minute) presentation using traditional virtual presentation tools (such as powerpoint on Zoom) or the prototype? Why?

9. What use-cases could you see this style of presentation being useful for?

10. Could you imagine ways in which this prototype could be useful for live presentations? Why or why not?

11. Were some presentations better than others? If so, what elements made some presentations better than others?

12. Were any interaction techniques overwhelming or corny? If so, which?

13. Do you think some interaction techniques would be better suited for some presentation contexts than others? If so, please explain

14. Would you want to give a presentation using this prototype? Why or why not

Likert Scale Questions: Please rate the following on a scale of 1 to 5 with 1 meaning "not at all" and 5 meaning "very much so". If you have no experience with the questions material or are unsure, please leave the question blank.

15. I enjoyed watching presentations using the prototype

16. Did you notice any of the following interaction techniques being used?

Mouse Mid-air hand gestures Speech command recognition Tablet Keyboard

17. Did you notice any of the following operations being used?

arranging images text display on-screen drawing pan/zoom

18. If you noticed mid-air gesture interactions, were they engaging and enjoyable

19. If you noticed voice interactions (e.g. commanding elements on the screen using spoken voice) were they engaging and enjoyable?

20. If you noticed on-screen drawing and annotation, were enjoyable and engaging?

21. If you noticed the use of text display, was it enjoyable and engaging?

22. I was generally engaged during virtual presentations using the prototype

23. I was more engaged with the virtual presentation using the prototype partially because I could see the speaker better than traditional virtual presentations

24. I felt compelled to ask questions or interact with the speaker when viewing the virtual presentation using the prototype

25.I would be more interested in interacting with a speaker giving a virtual presentation with the prototype

26. I liked watching virtual presentations with the prototype

27. I would rather watch a virtual presentation using the prototype than traditional presentation tools (such as slides on Zoom)

28. Please rank your preference for seeing the following interactions being used during the virtual presentations with 1 indicating "most favorite interaction" and 5 indicating "least favorite interaction". If you didn't notice one of these interactions, please select N/A.

Gestures Mouse Keyboard Voice Command Tablet

29. Please rank your preference for seeing the following operations being used during the virtual presentations with 1 indicating "most favorite interaction" and 8 indicating "least favorite interaction". If you didn't notice one of these interactions, please select N/A.

Arrange items Pan/Zoom Draw Text Display Conjure Dismiss Next Previous

30. Is there anything further you would like to tell us about virtual presentations or any of the topics mentioned in the questions above?

B.3 Speaker Post-Presentation

Making on Your Feet Questionnaire 4 : Presenter Post-presentation

The purpose of this questionnaire is to gather feedback regarding the performance of semi-extemporaneous presentations using a prototype creativity support tool.

Experience with Prototype (Performance Phase)

1. Please explain your overall impression presenting with the prototype

2. Do you feel it was helpful to prepare in a similar environment to that in which you presented?

3. How would you compare this experience with presenting a traditional virtual presentation?

4. Did the presentation proceed as you expected?

5. How did you manage unexpected incidents during the presentation?

6. Were there any interaction methods you wish you'd used instead of the ones you chose?

7. Which of the following interaction methods did you try during the rehearsal phase?

Gesture Mouse Keyboard Speech Tablet

8. Which of the following interaction methods did you use during your presentation?

Gesture Mouse Keyboard Speech Tablet

9. If you used gesture, did you find it enjoyable, useful, and intuitive? Why or why not?

10. If you used voice, did you find it enjoyable, useful, and intuitive? Why or why not?

11. If you used tablet, did you find it enjoyable, useful, and intuitive? Why or why not?

12. If you used on-screen drawing and annotation, did you find it enjoyable, useful, and intuitive? Why or why not?

13. If you used text display, did you find it enjoyable, useful, and intuitive? Why or why not?

14. Please rank your preference for using the following interactions during the virtual presentation with 1 indicating "most favorite interaction" and 5 indicating "least favorite interaction". If you didn't use one of these interactions, please select N/A.

Gesture Speech Mouse Tablet Keyboard

15. What are the strengths of this presentation style?

16. What are the weaknesses of this presentation style?

17. How did the presentation feel different using the prototype than traditional presentation tools?

18. How did it feel to be co-located with your presentation media (images, etc.)?

19. How would you have prepared differently if asked to present again?

20. Can you think of use-cases in which this tool would be helpful?

Likert Scale Questions: Please rate the following on a scale of 1 to 5 with 1 meaning "not at all" and 5 meaning "very much so". If you have no experience with the questions material or are unsure, please leave the question blank.

21. It was easy giving virtual presentations using the prototype

22. I would use this prototype to prepare virtual presentations in the future

23. I trusted the system more because I prepared my presentation in the same environment in which it was delivered

24. I was skeptical of various interaction techniques (gesture, voice, etc) prior to presenting with the tool

25. I am more trusting of these interaction methods after presenting

26. The rehearsal process built trust between me and the tool

27. It was easy for me to answer questions mid-way through the presentation, or encourage interactivity with the audience using the prototype

28. I would prefer to present future virtual presentations using this tool than traditional presentation tools

29. I would like to see virtual presentations given with this tool

30. I improvised during my presentation more than anticipated

31. It was intuitive to give virtual presentations using the prototype

32. My presentation proceeded in a strict sequential order in which it is to be presented

33. I was able to encourage interactivity during my presentation

34. I liked using this prototype presentation tool

35. It was fun giving presentations using the prototype

36. Is there anything further you would like to tell us about virtual presentations or any of the topics mentioned in the questions above?

B.4 Study 1: Formative Study Follow-Up

Making on Your Feet Questionnaire 6 : Formative Study Speaker Presentations

The purpose of this questionnaire is to gather feedback from presenters who we interviewed during our formative study. Participants will be asked to use our tool to present their proposed use-case. Participants will use our prototype to present their originally proposed use-case instead of pantomiming as they did in the formative study.

Experience with Prototype (Performance Phase)

1. Please explain your overall impression presenting with the prototype

2. Do you feel it was helpful to prepare in a similar environment to that in which you presented?

3. How would you compare this experience with presenting a traditional virtual presentation?

4. Did the presentation proceed as you expected?

5. How did you manage unexpected incidents during the presentation?

6. Were there any interaction methods you wish you'd used instead of the ones you chose?

7. What are the strengths of this presentation style?

8. What are the weaknesses of this presentation style?

9. How did the presentation feel different using the prototype than traditional presentation tools?

10. Were your needs were met for presenting your original use-case?

11. How would you have prepared differently if asked to present again?

12. Can you think of other contexts in which this tool would be helpful?

13. What alterations would you suggest to make this tool better suited for your use case?

14. Were there any interaction techniques you were skeptical to use?

15. How would you say the rehearsal process affected your decision of what interaction methods to use?

16. Overall, was the prototype what you expected? If not, explain

Likert Scale Questions: Please rate the following on a scale of 1 to 5 with 1 meaning "not at all" and 5 meaning "very much so". If you have no experience with the questions material or are unsure, please leave the question blank.

17. It was easy preparing and giving virtual presentations using the prototype

18. I would use this prototype to prepare virtual presentations in the future

19. It was intuitive to prepare and give my presentation using the prototype

20. I was skeptical of various interaction techniques prior to presenting with the tool

21. I am more trusting of these interaction methods after presenting

22. It was easy for me to answer questions using my prepared content with the prototype

23. It was easy for me to answer questions mid-way through the presentation, or encourage interactivity with the audience using the prototype

24. I would prefer to present future virtual presentations using this tool than traditional presentation tools

25. I would like to see virtual presentations given with this tool

26. I improvised during my presentation more than anticipated

27. My presentation was better planned than I anticipated during rehearsal

28. My presentation proceeded in a strict sequential order in which it is to be presented

29. I was able to encourage interactivity during my presentation

30. I liked using this prototype presentation tool

31. My presentation closely resembled my pantomime during the previous phase of the study

32. Is there anything further you would like to tell us about virtual presentations or any of the topics mentioned in the questions above?

B.5 Commentator Post-Presentations

Making on Your Feet Questionnaire 5: Video Feedback

The purpose of this questionnaire is to gather feedback on a series of videos depicting an audience and a presenter interacting using a prototype virtual interaction tool.

1. What is your age?

2. What is your gender identity?

3. What is your occupation?

4. Do you have any disabilities that you'd like to disclose or request for accommodation?

5. Do you have prior experience either giving or watching virtual presentations? (virtual presentations include slide shows such as Power Point on telepresence software such as teams, video lectures on platforms such as YouTube, or similar media)

6. If so, could you describe how the presentation was delivered? What software, media, etc. Please be as specific as possible.

7. How would you compare the prototype presented in the video against your previous experience with virtual presentations?

8. Does the audience seem engaged?

9. How improvised are these presentations?

10. Would you rather present a talk using this prototype or traditional presentation tools? Why?

11. Would you rather watch a talk using this prototype or traditional presentation tools? Why?

12. Can you think of contexts where this prototype would be useful? Explain

13. While the presentations you watched during this study were short, do you think this style of presentation would work well for a longer presentation? Why or why not?

14. Any other observations or comments about the audience and presenter interaction from the video?

Likert Scale Questions: Please rate the following on a scale of 1 to 5 with 1 meaning "not at all" and 5 meaning "very much so". If you have no experience with the questions material or are unsure, please leave the question blank.

15. The audience is engaged while watching virtual presentations using the prototype

16. The speaker is well prepared

17. The presentation is more engaging than traditional presentations

18. Is there anything further you would like to tell us about virtual presentations or any of the topics mentioned in the questions above?