Challenges and Design Considerations for Multimodal Asynchronous Collaboration in VR

KEVIN CHOW, University of British Columbia, Canada
CAITLIN COYIUTO, University of British Columbia, Canada
CUONG NGUYEN, Adobe Research, USA
DONGWOOK YOON, University of British Columbia, Canada

Studies on collaborative virtual environments (CVEs) have suggested capture and later replay of multimodal interactions (e.g., speech, body language, and scene manipulations), which we refer to as *multimodal recordings*, as an effective medium for time-distributed collaborators to discuss and review 3D content in an immersive, expressive, and asynchronous way. However, there exist gaps of empirical knowledge in understanding how this *multimodal asynchronous VR collaboration (MAVRC)* context impacts social behaviors in mediated-communication, workspace awareness in cooperative work, and user requirements for authoring and consuming multimedia recording. This study aims to address these gaps by conceptualizing MAVRC as a type of CSCW and by understanding the challenges and design considerations of MAVRC systems. To this end, we conducted an exploratory need-finding study where participants (N = 15) used an experimental MAVRC system to complete a representative spatial task in an asynchronously collaborative setting, involving both consumption and production of multimodal recordings. Qualitative analysis of interview and observation data from the study revealed unique, core design challenges of MAVRC in: (1) coordinating proxemic behaviors between asynchronous collaborators, (2) providing traceability and change awareness across different versions of 3D scenes, (3) accommodating viewpoint control to maintain workspace awareness, and (4) supporting navigation and editing of multimodal recordings. We discuss design implications, ideate on potential design solutions, and conclude the paper with a set of design recommendations for MAVRC systems.

CCS Concepts: • **Human-centered computing → Virtual reality**; **Collaborative interaction**; *Computer supported cooperative work*; • **Applied computing →** *Multi / mixed media creation*;

Keywords: Multimodal recording, virtual reality, asynchronous collaboration, spatial task, 3D, speech, gesture, body language, pointing, presence, immersion, proxemics

ACM Reference Format:
Kevin Chow, Caitlin Coyiuto, Cuong Nguyen, and Dongwook Yoon. 2019. Challenges and Design Considerations for Multimodal Asynchronous Collaboration in VR. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 40 (November 2019). ACM, New York, NY. 24 pages. https://doi.org/10.1145/3359142

1 INTRODUCTION

Collaborative virtual environments (CVEs) are emerging as a promising CSCW platform, through which distributed stakeholders can create, discuss, and review spatial content (e.g., 3D models,

Authors' addresses: Kevin Chow, kchowk@cs.ubc.ca, University of British Columbia, Vancouver, British Columbia, Canada; Caitlin Coyiuto, coyiutoc@students.cs.ubc.ca, University of British Columbia, Vancouver, British Columbia, Canada; Cuong Nguyen, cunguyen@adobe.com, Adobe Research, San Francisco, California, USA, 94103; Dongwook Yoon, yoon@cs.ubc.ca, University of British Columbia, Vancouver, British Columbia, Canada.

2573-0142/2019/11-ART40 $15.00
https://doi.org/10.1145/3359142

room layouts, animations, as in [7, 28, 71, 79]). Such content is becoming increasingly important in creative industries such as architecture firms, video game companies, and design agencies. Simultaneously, in these domains, the modern workforce is also becoming more globalized and time-distributed, increasing the need for supporting *asynchronous* collaboration between teams working across different time-zones or working hours. In these settings, successful collaboration depends on support for asynchrony, which has several unique advantages over synchronous communication, such as: work parallelism, flexible time-coordination, reviewability, and reflection [40, 58]. Several contexts where asynchronous collaboration becomes crucial in the evolution of a shared artifact were discussed in-depth by Tam and Greenberg [76].

With the advancement of virtual reality technology, *asynchronous* collaboration in immersive 3D environments has become an active research area [14, 32, 44, 50, 60, 67, 80, 85]. *Multimodal recording*—the capture and later playback of multiple, dynamic interaction modes, such as speech, locomotion, body language, and object manipulations—is at the foundation of asynchronous collaboration in these VR systems. In these studies, however, multimodal recording was regarded only as one of many features of a system, designed for a specific purpose such as training [85] or 3D design [80]. The state of the literature leaves room for more empirical research [84] on conceptualizing multimodal recording in VR and producing a generalizable understanding of it.

We aim to address the knowledge gap in conceptualizing and understanding this *multimodal asynchronous VR collaboration (MAVRC)* as a type of CSCW, backed by an empirical understanding of its challenges and design considerations. VR, CSCW, and HCI literature reveals three aspects of MAVRC: social behaviors in mediated-communication, awareness in cooperative work, and authoring and consuming multimedia. Using these aspects as an intellectual lens, we begin to identify and tackle these knowledge gaps by formulating the following research questions:

- *How are social behaviors transferred from face-to-face to MAVRC?* In VR, where collaborators can leverage embodied avatars with multimodal communication capacities, different aspects of social norms (e.g., proxemics, bias, anxiety, etc.) can be transferred to, or even amplified or diminished in CVEs [4, 26, 73]. However, MAVRC is fundamentally different from those of the previous studies in that asynchronous communication is one-way—the message sender cannot respond to the viewer's inquiry in real-time. It is an open question to what extent, or if at all, users feel the social presence of an embodied 3D representation of their *asynchronous* collaborator and display social behaviors towards it (e.g., what do they feel if the recorded avatar breaches their intimate space?).
- *What are the challenges of maintaining workspace awareness and coordination in MAVRC?* It is critical for the success of any CSCW system to offer proper support for establishing and maintaining different types of awareness about others' activities: workspace awareness is the knowledge of what is going on and what others are doing [17, 25, 36]; asynchronous change awareness is the knowledge of how shared artefacts (e.g., 3D models, source code, documents) have evolved via the contributions of time-distributed collaborators [76]. The MAVRC setting is at the intersection of 3D and asynchrony, where the challenges of awareness may be exacerbated by the combination of spatial occlusion, limited viewpoint, and lack of real-time feedback mechanisms (e.g., dynamic activities of one's asynchronous collaborator can occur out of the user's view, and the collaborator would not know how to draw one's attention to it).
- *What are the challenges of navigating and creating multimodal recordings in MAVRC?* Benefits and challenges of multimodal recording have been deeply studied in non-VR CSCW contexts, such as document annotation [86], design prototyping [49], and video production [62]. On the one hand, the rich, communicative capacities of recorded multimodal interactions help

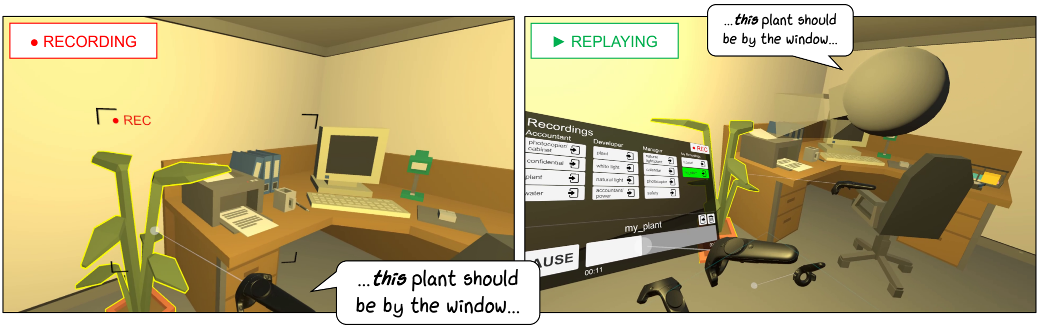


Fig. 1. Capturing and replaying speech, pointing, and head movement for a spatial collaborative task using our experimental system; (left, in the view of a recording producer) a remote collaborator makes a request about moving the position of the plant; (right, in the view of a recording consumer) an asynchronous listener replays the recording, watching the avatar of the remote collaborator as they express themselves through multimodal interactions.

> collaborators express and understand nuanced, complex ideas effectively [13, 55, 87]. On the other hand, as Grudin pointed out, when consuming, browsing recorded speech can be tedious and slow for the consumer [34], and, when producing, it is known that multimodal recordings are harder to revise than text [72] and can provoke self-consciousness [1, 52]. When it comes to supporting *spatial tasks in immersive CVEs*, however, there is a significant gap of knowledge regarding the challenges and design considerations of authoring or consuming multimodal recording in VR.

To answer these questions, we conducted a qualitative, exploratory need-finding study based on our experimental system, in support of multimodal recording in immersive VR. Our participants were tasked to perform asynchronously collaborative 3D design tasks by viewing and recording multimodal interactions. Key findings from our observations and interviews include: (1) participants felt the social presence of asynchronous collaborators when viewing recordings, leading to proxemic behaviors and empathy, (2) immersion in VR can cause challenges in viewing multimodal recordings, such as viewpoint disorientation or confusion from different versions of 3D scenes, (3) recording (and viewing) co-expressive speech and body language enabled participants to express (and understand) nuanced ideas effectively but highlighted needs for editing (and browsing) tools. We discuss implications of these findings for designing MAVRC systems and present a set of design recommendations. Demonstration and preliminary evaluation of four proof-of-concept interfaces provide support for the feasibility of our design implications.

The contributions of this paper are three-fold: (1) a conceptualization of multimodal recording-based asynchronous collaboration in VR; (2) empirical findings on the user challenges of respecting proxemics, maintaining awareness, and consuming/producing multimodal recordings in MAVRC; and (3) design implications for multimodal recording for asynchronous collaboration: proactive proxemics management, animating changes in 3D, viewpoint display for awareness, and 3D navigational cues. We also touch on the transferability and generalizability of our findings on MAVRC to non-VR asynchronous 3D interactions (e.g., AR, mobile 3D, etc.). Through the exploration of asynchronous collaboration in VR, this work opens up new opportunities for richer forms of collaboration, lessening the challenges presented by geographical and temporal barriers.

2 RELATED WORK

Our study combines themes of asynchronous collaboration, CVEs, and multimodal annotation from HCI, CSCW, and VR literature.

2.1 Tools to support asynchronous collaboration

Asynchronous collaboration refers to cooperative scenarios where participants interact at different times. CSCW researchers established theoretical frameworks to help identify benefits and challenges of asynchronous collaboration. Olson and Olson theorized advantages of time-distributed collaboration as work parallelism, flexibility in coordination, reviewability, and reflection [58]. Tam and Greenberg suggested *change awareness*, the ability of the user to trace the non-real-time changes made to the shared workspace, as a critical consideration in designing asynchronous collaboration systems for shared documents and graphical artifacts [76]. One of the main foci in HCI, informed by these CSCW frameworks, is to develop asynchronous collaboration tools for a variety of applications: video meetings [77], data visualizations [39], web search [54], and review & feedback [63, 64, 86].

However, since these frameworks and tools are developed for applications in a 2D environment, it is still unclear how the challenges and design considerations for asynchronous collaboration reveal themselves differently in VR. Our study extends past research by identifying issues unique to VR collaboration.

2.2 Multimodal asynchronous collaboration in virtual reality

Virtual reality systems with digitally constructed environments and robust spatial tracking can offer compelling opportunities in aiding collaboration. In the field of collaborative virtual environments (CVEs), researchers have focused on harnessing VR to provide more effective means in supporting communication and information sharing [74]. Most research, however, focuses on applications of *synchronous* collaboration, such as joint-tasks [24, 65], giving instructions via a ghost-hand metaphor [78, 85], tele-presence [59], or review & feedback [56]. Our study aims to discover the design considerations specific to *asynchronous* VR collaboration.

The premise of MAVRC in our research is grounded on many successful previous studies on VR systems for asynchronous collaboration. V-Mail [44] and MASSIVE-3 [32] are the most relevant to MAVRC in that they supported the capture and replay of rich, multimodal interactions. Several applications were built based on the notion of multimodal recording in different domains including architectural review [35], creative feedback [56, 80], training [85], and tele-communication [14, 60, 67]. The primary foci of these studies were more on developing novel interaction techniques/concepts or designing a new technical pipeline. In contrast, we aim to provide extensive *empirical evidence* that help us understand the range of user needs, challenges, and design considerations for MAVRC.

Very recently, Lindlbauer and Wilson [50] conceptualized several types of time manipulations, including pause, loop, and replay of a captured 3D scene, as part of the taxonomy of possible interactions in mixed reality. Our study adds to this by lending a fresh task-centric view to the time manipulation that takes place in a collaboration context.

2.3 Multimodal recording and annotation

Central to asynchronous collaboration is the exchange of messages across time. Early examples of asynchronous communication are typically textual (e.g., email or text messages). With the ubiquity of audio capture devices, use of recorded speech (e.g., voicemail) for expressive messages have been extensively explored in HCI and CSCW. With a voice recording, message producers can leverage paralingual cues (e.g., inflection, pause timing, energy) for conveying equivocal and complex ideas [13]. Speech annotation/commenting, thanks to its rich, nuanced expressivity, has been widely adopted in different applications including online discussions [52] or document review [55]. Inspired by media richness theory [18] and the conceptual framework of deictic gestures [17], several multimodal annotation systems have tried to enrich communication and collaboration by combining multiple face-to-face inspired modalities, such as inking + speech [75, 83], inking + gesture + speech [86], and video + hand gesture overlay [38].

The benefit of expressive richness comes at a cost in browsing and editing. Grudin, in his seminal work on CSCW applications [34], insisted that the major reason why voicemail is not popular is because its *browsing and navigation* features are very slow and tedious. To address this challenge, HCI researchers suggested various approaches such as using semantic navigation cues like captions [82], ink strokes [75], gesture trace [86], video thumbnails [66], viewers' collective navigational trace (e.g., skips and jumps) [47], or directly manipulating an in-video visual entity along its motion trajectory [57]. Besides the browsing issue, in Yoon et al.'s study where a speech-based document annotation tool was deployed to classrooms, it was found that the support for efficient interfaces for *editing and revising* audio recording is necessary for the success of speech commenting tools [87]. For this problem, researchers devised efficient multimodal editing solutions such as using time-aligned captions as a proxy for editing audio and video [8, 68].

In contrast to these previous studies on 2D media, our empirical setting is in 3D where the recording of body movement and scene manipulation is a multidimensional data stream. Our study identifies users' specific needs for viewing, skimming, and revising such data.

2.4 Proxemic interactions in VR and HCI

Proxemics, originally introduced in Hall's seminal work [37], describes the way interpersonal distance manifest social context. As modern computing systems became sensitive to the physical presence and social context of users, designing proxemic interactions attracted the attention of HCI researchers [5]. Greenberg, Marquardt, and their colleagues—the pioneers of proxemics in HCI—proposed proxemics as a new genre of HCI [31]. Due to the embodied nature of VR, many researchers studied the social implication of interpersonal distance in CVEs. The most relevant work is Bailenson et al.'s study on proxemics in immersive VR [4]. They found that human proxemic behaviors, such as moving away from an embodied agent who invades one's personal space, can be transferred or even amplified in collaborative VR environments. Similarly, Schroeder and colleagues showed that VR users exhibit such behaviors to an interactive avatar with high behavioral realism [70]. Our study adds to this literature by suggesting that proxemic behaviors can be transferred to *asynchronous* CVEs.

3 METHOD

Our mode of inquiry was geared toward an exploration and elicitation of ways in which multimodal recording can support and impact spatial tasks in MAVRC. Answering such questions requires qualitative data collection and analysis procedures. Hence, we collected interview and observation data where subjects experienced a representative spatial task in an asynchronously collaborative setting. First, we describe our experimental system, which enabled subjects to create and consume multimodal recordings in a CVE, and then detail our task, procedure, and analysis methods.

3.1 Experimental system: CVE for creating and viewing 3D multimodal recordings

To study the nature of MAVRC, we needed to have a representative instantiation of MAVRC systems[1] where participants' user experiences can be grounded and the task scenarios can be executed for the experiment. We were inspired by the common denominators of existing MAVRC systems [14, 50, 60, 67, 80, 85], which include the capture and replay of multiple interaction modalities, known to be conducive for communication in spatial tasks in VR.

3.1.1 Multimodal recording. In MAVRC systems, a multimodal recording encapsulates time-synced streams of interactions that can be captured and then replayed, akin to streams of pixels and audio in a video recording, but in 3D. For our experimental system, we selected (1) speech, (2) body gestures (e.g., head and hand movements, locomotion, and pointing), and (3) manipulations on scene objects as core interaction modalities. These modalities were chosen as they enabled rich expression for spatial tasks in a CVE and were easy to capture using most off-the-shelf VR systems.

On recording playback, the body gestures and movements of the producer are represented through an embodied avatar in VR. The producer's speech during the capture process is also spatially synced to the changing location of the avatar, forming a soundscape. We chose a simple, identity-neutral avatar design (Figure 1 (right)). A modeled view frustum indicates the avatar's head orientation, and two controllers represent their hands. The controllers can emit laser pointers that highlight targeted objects, in support of deictic referencing. Integrating a highly realistic avatar design showing the full range of emotions and gestural interactions can pose complex challenges (e.g., tracking instrumentation, computational complexity) that is beyond our scope [30].

In asynchronous collaboration, multimodal recording is a medium for time-distributed collaborators to exchange immersive and expressive messages. Recordings in our system were akin to annotations, which are anchored to some master document—or, in the case of a VR-based multimodal recording, to a given master 3D scene. Just like an annotation that doesn't change body text, scene changes made by producers while creating a recording do not change the master scene. The solution for version control in 3D is an open question beyond the scope of this paper, and thus not implemented. Instead, we focus on discovering the specific challenges of tracking changes between 3D scenes in MAVRC.

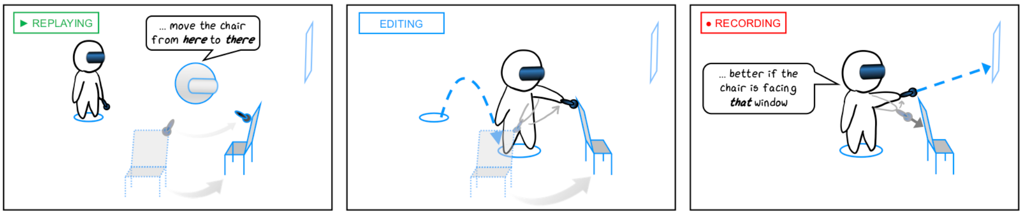
3.1.2 Creating and viewing multimodal recordings. Our system uses a simple and familiar interface for creating and viewing recordings, modeled after typical screen recording or video player apps.

To create recordings, users can click on the "REC" button on the world-stabilized VR menu. An audible beep goes off and a viewfinder is overlaid on the user's view to indicate capture in progress (Figure 1 (left)). Clicking on the same button again stops and saves the recording. In addition to the user's synced speech, hand and head gestures, and scene manipulations—indicators for certain types of input, such as teleportation arcs or laser pointer traces—are also saved for visual consistency.

To view, users browse through saved recordings in the menu and select one. A standard timeline-based widget allows users to control playback. Although advanced spatial playback mechanisms exist [50], we purposely adopt a simple and familiar baseline interface. Clicking on the exit icon button stops recording playback, returning users to the master scene. During playback, an embodied avatar will speak, gesture, and manipulate the 3D scene, based on the captured interaction streams.

3.1.3 Apparatus. We developed the experimental system in Unity for the HTC Vive, which includes a VR headset (FOV: 110°, refresh rate: 90 Hz, resolution: 1080 × 1200 pixels per eye) and a

[1] It is worth noting that MAVRC is a type of CSCW, not the name of an experimental system we built for this study. Hence, existing systems from previous research can be referred to as MAVRC systems.


▶ REPLAYING
... move the chair from *here* to *there*
EDITING
● RECORDING
... better if the chair is facing *that* window

(a) Viewing a recording of client requests (b) Editing the layout as requested by the client (c) Producing a recording as a response

Fig. 2. An illustration depicting the steps participants completed in the experimental task, starting from (a) viewing recordings to (b) editing the layout appropriately, and (c) producing a "response" recording.

pair of 6-DOF controllers [42]. The system ran on an Alienware 15 R3 Laptop with an Intel Core i7-7700HQ CPU @ 2.80GHz processor and an NVIDIA GeForce GTX 1070 graphics card.

3.2 Experimental task: reviewing and designing VR environments

In our study, participants performed three types of activities critical to collaborative 3D content design: (1) listening to input from collaborators, (2) revising 3D content based on their input, and (3) responding to and justifying changes made (see Figure 2). Real asynchronous collaboration is composed of multiple interleaving and evolving chains of such activities. As an initial exploration, and for cost and feasibility, we looked only at a single chain of these three activity types. In doing so, we could simulate the essential elements of MAVRC that constitute long-term collaborative turn-taking, without significant loss of generalizability.

3.2.1 Interior design task context. To put participants' mindset into a concrete context for asynchronous collaboration, we asked them to act as an *interior designer* working overseas at an international agency to plan a local company's newly renovated office layout, *simulating a realistic remote work scenario.* As interior designers, participants collaborated with their clients—employees of the local company—by exchanging multimodal recordings about the requirements of the office layout. The task context of interior design was carefully chosen to examine the space where multimodal recordings may have the most impact. Because of its 3D, spatial nature, interior [10] or architecture design [28] has been commonly identified as a key application for VR. Work in interior design has also been extended to evaluating collaborative VR systems in previous research [43, 61].

3.2.2 Task materials. Multimodal recordings of client requirements were pre-generated by pilot users of the experimental system who recorded themselves acting as clients based on scripts created by the research team. In these recordings, each client detailed their suggestions on the office layout to the interior designer, as if they were collaborating in person, face-to-face, despite being distributed physically and temporally in the task context.

Three clients each had five different requirements which varied in importance. Requirements were purposefully designed to be *dilemmatic* so that the layout task would be non-trivial, and generate richer response recordings. Examples of client requirements include:

- "I work with confidential information—my screen shouldn't be easily visible to bystanders."
- "I need to access the photocopier occasionally." *(a dilemmatic request from multiple clients)*

3.2.3 Procedure and data collection. First, participants were briefed through a short tutorial, which familiarized them with VR and the production and consumption features of our system.

For the task (Figure 2), participants carried out the following steps in order: (Step 1) view the clients' multimodal recordings to understand their suggestions, (Step 2) edit the layout by moving furniture models around, and finally (Step 3) create a "response" recording back to the clients to describe how they balanced the different needs of each client in the final layout. Participants were given the option to rewatch recordings viewed in Step 1 while editing the layout in Step 2, if needed. They were not given a time limit, and continued with the task until they were satisfied with their final office layout and response recording. We did not assign a strict measure of task completion, as completion criteria may vary between individuals due to the dilemmatic and open-ended nature of our task.

Participants took a short break from VR after each step. To check for possible adverse motion-sickness effects, we collected Simulator Sickness Questionnaire (SSQ) data [46] before and after the task. We refer to each questionnaire as the pre-exposure and post-exposure SSQ, respectively.

At the end of the entire task, we conducted a 30-minute semi-structured interview where we asked participants about their needs and challenges experienced while using the system. While interviewing participants, we also reviewed their point of view (POV) video data with them to both guide our questions and ground their answers. The interview was audio-recorded and transcribed for later analysis. During the task, participant head orientation, spatial position, controller input, and timeline navigational patterns were logged by our system. POV video data was captured using the OpenVR input plugin [45] for supplementary video analysis.

3.3 Analysis

A systematic thematic analysis [51] approach was used to identify relevant patterns and themes in the interview and observation data. In our analysis, we invested a conscious effort to maintain theoretical sensitivity to patterns regarding a range of issues, motivated by previous work and our research questions. An inter-coder reliability test yielded a Cohen's kappa score of 0.87 indicating strong agreement between two independent coders. This score was calculated using 64% of the entire dataset (719 out of 1123 total sentences) while the other 36% was set aside for coder training; it is common to evaluate inter-coder reliability based on a dataset sample [12].

3.4 Participants

The study was conducted with 15 (7 female, 8 male, mean age = 22 years, SD = 1.55) university students or recent graduates in a Western society. They had various academic backgrounds: forestry, psychology, media studies, statistics, biology, and computer science. Of note are P_5 and P_6—students with school and work experience in construction and architecture, respectively. 6 participants have never used VR nor AR, 5 have used only VR, and 4 have used both VR and AR. Each participant was compensated with $15 for their time (1.5 hours).

4 RESULTS

All but one participant successfully completed all task steps; P_7 only completed the viewing and editing steps, as they did not follow instructions when re-watching recordings. Hence, P_7 was excluded from any descriptive statistics, but not from qualitative data analysis as their comments corroborated other participants and helped us characterize the nature of MAVRC. On average, the study took 67.5 minutes (SD = 5.81), not including breaks or briefing time. Average total task time was 35.1 minutes (SD = 7.55), and average time for the post-task interview was 32.4 minutes (SD = 6.97).

While acting as an interior designer, participants said that they saw the benefits of multimodal recording for collaborative, spatial tasks. 8 participants reported that they felt the emotions and social presence of their asynchronous collaborators, showing signs of empathy towards them. "It's almost like the real person is there, so it's much easier to relate to the client (P_6)." Their remarks are supported by the known benefits of multimodality for complex collaborative tasks, such as

Table 1. Summary of our thematic analysis, which included 8 sub-themes organized into 4 unique themes. Participant count (rightmost column) is out of 15.

Theme	Sub-theme	Count
Proxemics of asynchronous, embodied interactions in VR	Asynchronous avatars violate one's personal space and disorient viewers	11
	Viewers chose to remain at a "social distance" from the avatar	5
Version tracing and change awareness in different VR scenes	Feeling disoriented when transitioning between different VR scenes	11
	Awareness of changes between multiple MAVRC recordings anchored to different timelines	4
Challenges and workarounds in establishing workspace awareness	Consumer-side challenges in maintaining awareness of the avatar's intention, action, and location in an immersive, 3D recording	10
	Producer-side challenges in accommodating and directing the awareness of viewers	6
Challenges in navigating and creating multimodal recordings for MAVRC	Immersive recordings need appropriate space-time navigational cues	11
	Performance anxieties and needs for editing features when creating asynchronous recordings	9

expressivity [13] and ease of deictic referencing [29]. Watching the combined co-expression of speech and body gestures helped them "gauge the magnitude of how much someone cares about something (P_4)". The laser pointer was a powerful deictic device for associating speech to target spatial objects; when producing the "response" recordings, participants used deictic keywords, such as "here", "this", "there", and "that", on average 3.13 times (SD = 3.20). 57.4% of such utterances were accompanied with a co-expressive pointing gesture. SSQ metrics indicated no significant difference in VR sickness before and after using our system ($t(14)$ = -1.602, p = 0.13, paired t-test).

Our thematic analysis revealed four novel design considerations (see Table 1) for MAVRC systems: (1) coordinating proxemic behaviors between time-distributed collaborators, (2) tracing changes across different versions of 3D scenes, (3) maintaining workspace awareness via viewpoint control in 3D, and (4) navigating and editing 3D multimodal recording.

4.1 Proxemics of asynchronous, embodied interactions in VR

A major theme from our analysis was indicative of the nature of proxemic social protocols in MAVRC. Users exhibited proxemic behaviors that are *structurally* similar to that of face-to-face scenarios, due to the embodied nature of MAVRC interactions. However, asynchrony can lead to breaches in proxemic norms, provoking negative emotional responses from users. Motivators for proxemic behaviors were both socio-cognitive (e.g., interpersonal attitudes) and functional (e.g., visibility, ease of communication), analogous to that of real-world proxemics [37]. In the following subsections, we focus on findings unique to MAVRC.

4.1.1 Asynchronous avatars violate one's personal space and disorient viewers. 11 out of 15 participants reported feeling "disoriented", "annoy[ed]", and "jarr[ed]" when avatars in recordings teleported too close to them or even "into" their body. P_1 explains: "It disoriented me because suddenly it was right in my face, and that messed me up." In Hall's theory of proxemics, managing interpersonal distance has both functional and social connotations [37]. Reports from our participants were indicative of both. The functional aspect was trivial—when the avatar's head was too close, it blocked their vision.

A fresh finding was that many participants perceived this behavior from the avatar as an invasion of their space. 5 participants described that they felt "violated" and "invaded". This indicates that the social norm of proxemic interactions—in this case, a respect for one's intimate or personal space—was expected from the recorded behaviors of the avatar. This expectation is not met as recorded avatars cannot adjust their interpersonal distance and circumvent the participant's position. Several participants, such as P_2, seemed to be aware of the cause behind this issue, saying: "...you can't tell them to move, you just have to navigate around them." However, they still felt that a breach of their personal space by the embodied avatar was invasive.

4.1.2 Viewers chose to remain at a "social distance" from the avatar. A relative of the above avatar-to-viewer effect is the way viewers adjust their *own* distance *to* the avatar's position. To unpack this, we created a histogram (Figure 3) showing the distribution of viewer-to-avatar distances for every teleport action made by participants in the study, an approach akin to previous work on synchronous CVEs [27]. Because participants rarely walked in our study, we only counted teleportations, an aim-and-shoot style mode of locomotion widely used in modern off-the-shelf VR platforms. All distances are Euclidean and in metres. Units of distance in our VR environment have an approximate 1-to-1 mapping to metres in the real world. We count repeated teleports that don't vary significantly (>1 metre) from the users' previous position as one action to capture deliberate teleportations only.

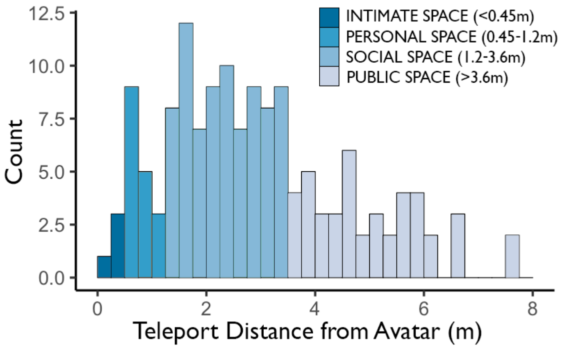


Fig. 3. Histogram showing viewer-to-avatar distances (in meters) for each deliberate teleport action made when watching a recording. Bars are colored according to Hall's classification of interpersonal distances. [37].

The histogram showed structural similarity to the classification of proxemic spaces in face-to-face [37] or synchronous CVEs [27]. Participants chose to teleport most frequently (83 out of 144 total teleports) to the avatar's 'social space' (a distance range of 1.2 to 3.6 metres), which is known to be conducive to communication among acquaintances. They rarely teleported to distances closer than .45 metres, the 'intimate space', reserved only for those whom one has close, intimate relationships. This triangulates our findings about the avatar-to-viewer proxemic breach effects, showing that participants respect the embodied avatar's intimate space, and expect the same likewise. Some teleports were to the personal or public spaces, but they were not as common as the social space.

Reports from participants support this finding. 5 participants stated that they "followed" the avatar, maintaining an appropriate distance. Their primary motivation was to "get a better view (P_{15})" of the shared workspace for better awareness of "what they're doing (P_8)", or what they are working on: "they were trying to point to things that were sort of far away in the office space, so that is why I had to teleport closer (P_9)." Many related their proxemic behaviors to similar real-life scenarios. P_6 says: "If I have anything to say to them they would be within speaking distance. I feel like if they're standing far away I have to shout. [...] Kind of like if someone was giving you a tour of their house, you wouldn't stand opposite them and look in the other direction, you'd probably be standing near them, next to them."

4.2 Version tracing and change awareness in different VR scenes

Challenges in managing different workspace versions are inevitable in asynchronous collaboration, where collaborators contribute changes to the shared workspace in a time-distributed manner. We found that the unique beneficial traits of MAVRC, such as immersion and the capture of scene manipulations in recordings, can engender new challenges that existing version control systems cannot handle.

4.2.1 Feeling disoriented when transitioning between different VR scenes. 11 participants reported that they struggled with understanding the *transition* between different recordings in VR. As multimodal recordings are made on different versions of the same, shared 3D workspace, many recordings have similar-looking VR environments, but also contain changes (e.g., objects added, removed, or re-positioned). These changes caused confusion: "I was trying to move things around then I would switch back to the recording to see what they wanted, and the layout suddenly changed and kind of made me disoriented—took me a couple of seconds to figure out where things were (P_1)", "...it is the same space, but things are moved around (P_2)." In a way, the transition of switching between scenes is similar to a teleportation, which tend to trigger disorientation [53]. However, the disorientation effect might have been amplified when the transition to a new scene is not apparent, such that viewers might think they are looking at a slightly different scene, but could not figure out what had been changed.

4.2.2 Awareness of changes between multiple MAVRC recordings anchored to different timelines. 4 participants reported difficulties in maintaining asynchronous change awareness of the 3D scenes [76]. They articulated the need for tracking and understanding changes made in scenes (e.g., what object was moved, where it was moved from/to, and when the change was made). For example, P_1 and P_6 wanted to compare two different scenes: "Having a side-by-side comparison of seeing how things are changing at the moment would be helpful (P_1)" and "Maybe if I could compare it to the original layout, because I already had forgot about what the original one looked like, so I'm not sure where she made changes, because I'm only focusing on what she's saying and not able to remember what it looked like beforehand (P_6)."

Tracing the origin and destination of changes in multiple branching timelines were deemed to be difficult. P_7 even tried to compromise potentially conflicting visual information for focused attention to the avatar's verbal description by closing their eyes when watching the recordings, saying: "I was imagining the visuals in my head, instead of the default layout that I was seeing." This problem applies to many MAVRC systems that contain 3D scene manipulation in exchanged messages (e.g., the avatar moves an object in the scene), as scenes are changed temporally within a recording. As illustrated in Figure 4, scene manipulation within a recording can exacerbate challenges in understanding asynchronous changes. The version of a 3D scene can change not only when the user moves the playhead to various points within a recording's timeline, but also when the user switches to an entirely different recording. For example, in our experimental system (see Figure 4), the M_4-to-${R^{M1}}_0$ transition will make the drawer and monitor seem to disappear, while the ${R^{M1}}_2$-to-${R^{M2}}_0$ transition will remove the lamp and add the drawer to the scene.

Although our depiction of the challenges in tracing changes was situated in the particular implementation of our experimental system, such problems can be easily generalized to any MAVRC system that employs a similar versioning or scene manipulation feature.

4.3 Challenges and workarounds in establishing workspace awareness

Workspace awareness refers to the understanding of the activities of other collaborators [36]. In typical asynchronous collaboration systems (e.g., collaboration systems for document editing or software version control), there is little need for real-time workspace awareness because there are no *present activities* of one's collaborators [76]. What is unique about MAVRC is that there still are needs for establishing and maintaining components of real-time workspace awareness. Because MAVRC recordings capture interactions of collaborators in an expressive VR environment, viewing

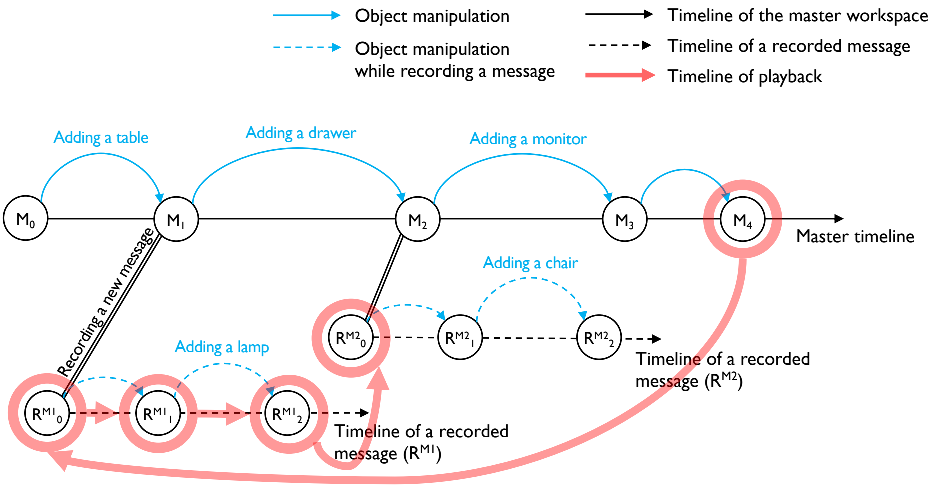


Fig. 4. In our version of a MAVRC system, a multimodal recording is anchored to a version of the master scene (e.g., R^{M1} is a recording anchored to M_1). A recording can contain scene manipulations as part of the captured interaction streams. Each scene manipulation within a recording generates a version of the 3D scene (e.g., $R^{M1}{}_0$, $R^{M1}{}_1$, and $R^{M1}{}_2$ in R^{M1}). When the user navigates between multiple timepoints of different recordings (e.g., the red arrows in the figure), it can be challenging to maintain awareness of the asynchronous changes.

these recordings can simulate aspects of real-time face-to-face collaboration *as if* one is talking to someone being there or watching others’ ongoing activity. We document the unique user needs regarding such pseudo real-time workspace awareness below.

4.3.1 Consumer-side challenges in maintaining awareness of the avatar’s intention, action, and location in an immersive, 3D recording. Unlike 2D video, relevant content in VR-based multimodal recording is not always in the consumer’s view. 10 participants reported challenges and workarounds in establishing workspace awareness of the avatar’s activity. We found several key design factors.

First, when viewing an avatar’s recorded activity, participants reported that they want to see and understand more of what the avatar is doing. For example: “where they are (P_{13})”, “where they go (P_3)”, “ what they’re doing (P_8)”, “the thing they’re pointing at (P_{11})”, and what they can see. Some participants wanted to see exactly what the avatar sees: "I kind of want to be locked and just see it from their view, you don’t have to teleport or move around, you don’t have to follow him, [you can] just become him (P_3)." These requests are similar to those found in previous studies about real-time workspace awareness [36].

Second, participants reported difficulties in locating the avatar’s position in the scene. This was particularly difficult when a recording was loaded up for the first time (“I just couldn’t find that person at first (P_9).”) and right after avatars teleport (“when the employees [clients] use the teleport function, you will lose track of where they go (P_3)”). Sometimes, the avatar might appear into view, disappear from view, or even be completely behind the participant: “...they would be off or behind me somewhere, so I’d look to the side and not know where they were (P_2).” As a result, they needed to spend time and effort to locate the avatar, causing them to miss out on viewing key content. Some participants had to rewind to rewatch missed segments, or potentially even restart the recording entirely.

Third, participants put in a great deal of effort to maintain a good viewpoint—both position and view orientation in the 3D scene—from which they can best see what the avatar is doing and looking at (P_3 and P_5), as well as what they are pointing at (P_3 and P_{11}). Participants reported two types of barriers in establishing an optimal 3D viewpoint: occlusion and apparent scale. Participants often moved in response to occlusion, as depicted by P_{13}: "I had to go closer when changes are being made in the developer section because it was blocked off by a wall, so I didn't even know that there was a small couch there until I actually went there." To increase visibility, participants adopted a surveillance-like strategy: they first observe the scene from afar to gain spatial context (P_3 and P_{11}) and then when needed, move toward the avatar to observe details (P_3, P_4, and P_5). P_4's quote best illustrates this pattern: "I think I would go with the wide angle perspective first, but if they were to focus in on a certain thing, if they talk about the water cooler, for a few seconds, and they keep on going, then I would probably go closer to them and then think about it."

4.3.2 Producer-side challenges in accommodating and directing the awareness of viewers. When looking at the production step of the task (e.g., when participants created a "response" recording), we found that 6 participants, when creating a multimodal recording, tried to bring the attention of an imaginary listener to their ongoing activities, but were missing interactive grounding devices, such as feedback through acknowledgements and relevant conversational turn-taking, for confirmation of established common ground [17].

Producers felt like they were giving a walk-through to an imaginary listener, with the expectation that the listener would follow them around. They predicted the expected locomotion of *a prospective viewer* and tried to accommodate for the viewer's awareness of their activity. P_9 says, "If I want to move a plant to a corner of the room, then I need to teleport there. So, sometimes I need to go through the far end of the office to do certain tasks and explain to people. I think it's best for them to know and follow me around." To engage the viewer of their action and intent, several participants employed a number of gestures to call for attention. Some examples include teleporting close to spatial areas of interest, orienting head direction towards a particular position, and casting the laser pointer onto an object for deictic referencing. These are known to be effective mechanisms for offering awareness information.

A significant challenge arises from the one-way communication nature of asynchronous collaboration. In MAVRC, there is no way to know where the viewer will be, whether or not the viewer is seeing what they are seeing, and whether or not they understand the producer's message. Participants suggested a potential solution of directly sharing their viewpoint with the viewer: "I thought more about where I am specifically, because I want to show my point of view to them. That doesn't really matter though, because it's going to be in another point of view [for the viewer]. But I'm just stuck in this mindset that what I'm looking at is what the viewer is looking at, but that's not really the case (P_{14})."

4.4 Challenges in navigating and creating multimodal recordings for MAVRC

Participants reported different types of challenges in consuming and producing multimodal recordings. Some of the challenges share similarities with those in consuming and producing non-VR multimedia content (e.g., video or voice annotation). For example, navigating/skimming multimodal recordings was slower and reviewing/editing was more tedious than textual media. On average, participants spent less time on creating the "response" recordings (M = 2.62 minutes, SD = 1.32) than viewing the "request" recordings (M = 7.98 minutes, SD = 1.15). This result conforms to Grudin's remarks on the workload imbalance between the consumption and production of a voice-laden message [34]. In the following subsections, we highlight findings unique to MAVRC settings (e.g., lack of space-time navigation cues and performance anxiety when creating a live recording).

4.4.1 Immersive recordings need appropriate space-time navigational cues. 11 participants mentioned that they had to rewatch the recordings in order to get a firm grasp of the client's needs. They wanted to have a skimmable form of the recorded content (such as captions of the client's speech) because re-watching the same recording was tedious and also remembering and finding key points in the timeline was difficult. A comment from P_6 summarizes this point aptly: "...I had to remember so much of it and I didn't want to have to replay the recordings all over again because it's a lot watching it, I wish I could have taken notes..."

As detailed in Section 4.3.1, a unique challenge of navigating multimodal recordings is viewpoint management. In a recording, an avatar can move around different spatial positions of the 3D environment. Hence, when viewers navigate between different time points, there's no guarantee that the avatar's activity will be visible in the viewer's viewpoint. P_{15} elaborates on this issue: "I think the only part that is different, is that if you scroll to different points, and you're not quite sure what you're looking for, it looks like the exact same thing, because you're in the office, so they might be in a different position and talking about something else but then, when you're looking for the [point in the] video, you might not remember that they're in that position." As presented in 4.3.1, when watching a recording, users tend to teleport around the scene multiple times to follow client activity, which made it harder for them, as they had to remember and associate both *where* and *when* an event happened: "It's like, I need to remember the spatial - where I am spatially over time, it's kinda confusing when I'm teleporting to all these spaces, and I forget, where things are [in the recording] and stuff (P_{14})."

4.4.2 Performance anxieties and needs for editing features when creating asynchronous recordings. Producing multimodal recordings can be taxing. At least 7 participants felt performance pressures in creating asynchronous, recorded content, because re-recording the same message can be time consuming. P_{10} says "If I mess up I have to go back, delete it, rerecord, so I am conscious of messing up." Hence, more than half (8 of 14 participants, excluding P7) recorded their "response" in one go. 2 recorded twice, 3 recorded three times, and 1 recorded five times (mean = 1.86, SD = 1.23). Of the participants that recorded more than once, 4 made mistakes they deemed big enough to warrant restarting, such as losing their train of thought or picking up the wrong object.

Since participants knew that recorded content can be time-consuming to listen to and understand, they tried to keep their produced recordings concise and clear. P_6 states: "...when I was recording I was more aware of time, because nobody really wants to watch that long of a recording because I didn't want to either. So I want to be really concise." P_9's remark on the importance of clarity: "I didn't really want to confuse people. So I became a little more tense."

In contrast to in-person presentations to real-time audiences (i.e. public speaking), producers didn't feel as anxious as they knew that they were alone in MAVRC and did not feel the pressures of a live audience. P_8 says "...it wasn't as bad as if I was presenting something directly to people, I knew that I could start over if I screwed up, so it was less scary." In addition, our simple avatar representation allowed participants to detach their identity from the embodied avatar, which remedied their social anxiety (different from performance anxiety). For instance, P_{15} said "I wasn't self-conscious at all when making this as I knew people wouldn't see my face. So then, I didn't really feel anxiety or anything."

9 participants asked for appropriate editing capabilities, in order to fix mistakes without excessive recovery costs. They asked for standard features such as clipping or stitching recording segments together: "I would like the ability to edit a recording - kind of like when you're doing it in Windows Movie Maker, you just want to clip out certain parts, and make it coherent. I don't want it to be a really, really long list of gibberish, things that I could have said in five sentences instead of twenty (P_{13})" A lack of effective editing tools would have also contributed to the performance

pressures participants when recording. However, manipulating multimodal interactions can be difficult especially in VR where a recording captures expressive data such as head movement and gestural references (e.g., "adding motion to whatever you're doing just makes things more complicated (P_{14}).").

4.5 Closing remarks on the results

There are three core concepts, each of which cross-cut multiple themes presented earlier. First, adjustment of viewer's viewpoint in 3D, evident in Sections 4.1.2 and 4.4.1, was critical for understanding visual events in a recording. Second, multimodality was both a blessing in communicative expressivity, but also a curse in navigation and editing challenges, as in Section 4.4. Lastly, the lack of real-time feedback hampered MAVRC users from leveraging interactive communication devices that would have enabled them to recover from challenges such as breach of proxemic protocols and common ground breakdowns.

5 DISCUSSION

Upon further reflection on our methods and the results, we discuss and present a set of novel design implications and recommendations. To demonstrate the feasibility of these insights in generating solutions for the discovered challenges, we also illustrate four proof-of-concept interfaces and conduct an informal evaluation.

5.1 Design implications

5.1.1 Embodied asynchronous social presence: impact to social norms and anxiety. In our study, our participants tended to humanize the avatar, which was recorded asynchronously, as if it was a flesh-and-bone social agent. Although it's known that collaborators may exhibit proxemic behaviors in synchronous settings [4], it is a fresh finding in our study that aspects of social norm can occur in *asynchronous* VR interactions as well. This finding highlights a rich design space regarding social dynamics that a designer must consider when designing MAVRC systems.

One potential direction to explore is what instances of social norm, other than proxemics, can be transferred to MAVRC. Since our avatar design was simple rather than realistic and sophisticated, interpersonal distance was one of the most apparent behavioral patterns observed. However, researchers have found that people using asynchronous communication tools feel a stronger social presence of collaborators as the expressivity and interactivity of the media gets richer (e.g., audio over text and video over audio) [11]. Hence, it's worth studying whether increasing the fidelity of the avatar could help transfer more nuanced social interactions such as trust, deception, self-consciousness, or vanity to MAVRC.

The task design is also an important aspect that affects social behaviors in MAVRC. The task context in our study was moderately sociopetal, meaning that collaborators are invited to stand nearby so they can see and understand what the other is doing. Also, our task only involved one asynchronous collaborator and no synchronous one. In a hybrid collaboration setting where multiple synchronous and asynchronous collaborators are intermixed, users will show different uses of interpersonal space. The theories of proxemics and F-formation will be viable sources of inspiration to study such topics [16, 37].

Yet, researchers may need to be vigilant about the over-interpretation of social nuances insinuated in MAVRC interactions. We want to reinforce our finding that people depicted *functional* needs (e.g., visibility, workspace awareness) as the major driver or their proxemic behaviors. Hence, our design solution idea focuses on remedying the apparent breakdown of appropriate proxemic behaviors: the violation of personal space by the recorded avatar. The design challenge is that the asynchronous avatar is incapable of adjusting its locomotion on its own to avoid collision with the viewer. One possible solution involves proactive management of such violations. Given the asynchronous nature of a recording, positions of the avatar are known at all time points. When a user is about to teleport into a location where the avatar will also teleport to in the near future, the user can be prevented or warned against doing so.

5.1.2 *Asynchronous change awareness across multiple 3D scenes.*

The results from 4.2 indicate that users in MAVRC systems may face difficulties in keeping track of the complex changes made across and within different versions of a 3D/spatial scene, which evolves as collaborators contribute changes over time. This is an instance of challenges in asynchronous version tracking, previously documented as asynchronous change awareness in CSCW [76] or as traceability management and recovery in the software engineering field [19].

Existing version control systems offer some insights on what kind of version management features a fully fledged MAVRC system should support, such as diffing, merging, branching, history tracing, and pull-request/discussion. Some studies that focus on 3D model version control [22, 23, 69] exist, but they fall short of such versioning features. These are still open questions in Graphics, VR, HCI, and CSCW.

A newly found challenge in MAVRC is the jarring sensation when the viewer sees sudden changes at the transition between two different 3D scenes. Animated transition is a popular technique to remedy such confusion in navigating multiple versions of a shared artefact, as demonstrated in the context of versioning systems for documents [15] and graphs [3]. MAVRC systems can help users trace changes in versions of a 3D scene by animating objects that are moved, created, and removed, or the person who made such changes. In VR it is important to animate scene differences only when the changes are visible in the user's field-of-view because the user may fail to notice visual signifiers out of view.

5.1.3 *Support for establishing workspace awareness.*

MAVRC users were missing support for awareness display mechanisms, when acting as both a consumer and as a producer. One particular consumption feature needed was a way to reveal "what the avatar can see right now" without excessively changing the viewer's viewpoint for optimal visibility. In particular, avatars hiding "behind a wall" was one of the major reasons why viewers had to follow them to keep them in view. Hence, solutions should employ a disocclusion technique, such as X-ray [2, 21] or multiperspective [81]. Following the avatar's position was so tedious that some participants even wanted to take the avatar's position and orientation as their point-of-view (POV). Slaving a viewpoint in VR is prone to motion sickness; however, methods exist that help to mitigate sickness effects [20], including ones which look at transitions between a 1st- and 2nd-person POV [48].

When recording their activities as a producer, users wanted to accommodate awareness of their prospective viewers. However, there's no way for them to be assured if their multimodal performance will successfully build up a robust common ground about their communicative intent when viewed by the consumer. As a potential solution, the grounding process can be delegated to an intelligent MAVRC system that is capable of analyzing the clarity of messages in a multimodal recording (e.g., the quality of narration, the semantic alignment between the producer's verbal content and their body gestures), so that the system can give feedback to the producer about their message even before the consumer views it.

The nature of this asynchronous workspace awareness problem is similar to that of real-time collaboration [36]. What makes this problem more challenging in asynchronous interactions is the lack of interactive or communicative mechanisms to signal to collaborators so that they respond to the viewer's needs for workspace awareness in real-time. Participants had to pause, locate the avatar themselves, and rewind recordings when they missed out on important information.

5.1.4 Navigating and authoring space-time recordings. Participants wanted to edit the recorded content for clarity and conciseness. Creating multimodal recording in VR shares many challenges in authoring different types of space-time recording such as VR/360 video [56]. Identifying proper time-points to make a cut or transition in a recording is still an open research problem. Previous work on editing linear streams of multimedia suggests that automatically recognizing semantic editing handles, such as auto-captions or key-frame [8, 72], can be a potential solution. However, interfaces from previous work were primarily designed for desktop input. A follow up study can approach this problem from a perspective of designing spatial UIs for VR-based multimedia editing.

On the consumption side, prior work on skimming and browsing multimedia content offer insights to address an analogous problem in MAVRC. Existing solutions extract and present semantic navigation cues that viewers can time-index to jump or skim through linear multimedia content [33]. Such cues include captions from speech recognition, key-frames, or artefacts where meaningful events have occurred. These solutions can be transferred from 2D media to 3D VR with ease.

A unique challenge of space-time navigation in MAVRC is assisting the viewer in directing their viewpoint to the location where important events are taking place. For example, navigating in a recording should not only move the time point, but also the user's viewpoint to the optimal position to see the avatar's activity and the surrounding context. At the beginning of a recording, the user's view can also automatically point towards the avatar's location so the initial load of locating the avatar is reduced. Also, recordings can be paused when the avatar is out of view.

5.2 Design recommendations

To help guide the design of effective MAVRC systems, we offer the following design recommendations (*DR<number>*) derived directly from the results and design implications of our study.

- *(DR1) Prevent proxemic violations*: Alert the users when a breach of intimate space is expected.
- *(DR2) Animate changes at transition*: Provide traceable visual indicators of how 3D scenes will change.
- *(DR3) Highlight what others see*: Display objects relevant to the collaborator's activity with higher visual salience.
- *(DR4) Adjust viewer's viewpoint on time indexing*: Position and orient the viewer's point-of-view so they can follow the collaborator's activity without excessive locomotion.
- *(DR5) Provide rich navigational cues*: Structure the timeline with visual or semantic delimiters.
- *(DR6) Provide effective editing features*: Give users ways to clip, cut & paste, or re-record part of a multimodal recording for revision and editing.

5.3 Proof-of-concept interfaces

To concretize our design recommendations, we developed and evaluated four proof-of-concept (PoC) techniques (see Figure 5). Their key contribution lies in helping to demonstrate the feasibility of our design implications and to explore the possibilities of this design space, rather than presenting the only, or optimal, design solutions.

- ***Proxemic-sensitive teleportation*** (from *DR1*): If the user moves to a location where the avatar will soon pass by, it is likely that the avatar will trespass the user's personal or intimate space. We adopt a preemptive approach to help users avoid this collision. Since, in our study, the primary mode of locomotion was to teleport, we modified the teleportation technique such that it can detect and highlight potential collisions before teleportation. During the teleport aiming stage, our technique detects potential collisions in the near future by testing if the recorded movement of an avatar will overlap with the teleportation target the user is aiming at. If a potential collision is detected, our technique will warn the user with an animated

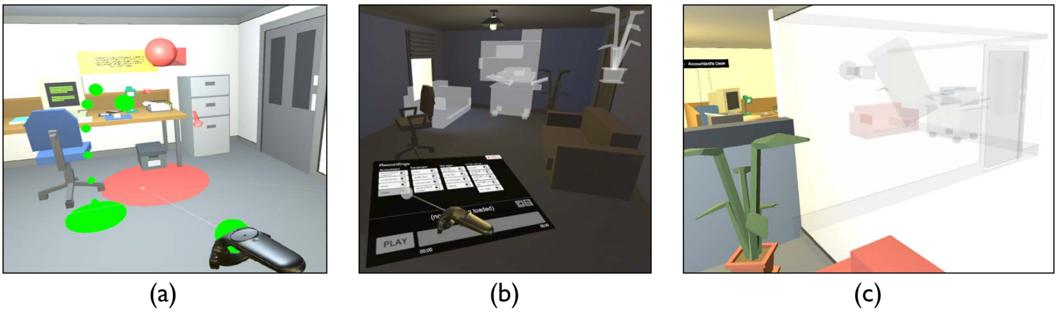


Fig. 5. Screenshots of three (out of four) of our proof-of-concept interfaces: (a) Proxemic-sensitive teleportation; (b) Animated transitions between 3D scenes; (c) Collaborator-aware X-ray vision. (not depicted: Enhanced timeline for space-time navigation)

snippet of the avatar (Figure 5 (a)), and redirect the target endpoint of the teleportation arc away from the avatar to a distance that matches the upper limit of Hall's classification of personal space (1.2 m). This puts users just outside of the avatar's personal space, maintaining proxemic protocols.

- ***Animated scene transitions*** (from *DR2*): A VR environment can abruptly change when the user jumps to a different time using the timeline or when she opens a new recording. To remedy disorientation, we added a transition mechanism into our system to render visual changes in VR. Inspired by Mnemonic Rendering [9], our system shows the upcoming changes with animation and alpha blending. The animation shows the ghosting of an object moving from the previous position to the current position (see Figure 5 (b)). A unique aspect of VR is that changes can happen all around the user. Thus, we buffer changes outside the user's current viewpoint and animate these changes when updated objects are in view.
- ***Collaborator-aware X-ray vision*** (from *DR3*): When viewing a recording, a user may miss activities of a collaborator avatar if they are occluded by the 3D scene (e.g., behind a wall or desk). To enhance workspace awareness about the avatar's activity, we developed an X-ray vision technique that selectively disoccludes important visual entities, including the avatar's body parts and objects that the avatar is interacting with (see Figure 5 (c)). To achieve this effect, each object in the scene was assigned a shader that controls visibility. The scene is rendered conventionally from the user's viewpoint. When occluded objects come into the view frustrum of the avatar, the shader modifies the visibility of these objects based on the depth buffer to create the X-ray vision effect.
- ***Enhanced timeline for space-time navigation*** (from *DR4* and *DR5*): Our timeline slider shows key moments of a recording as clickable bookmarks that help the viewer quickly identify and navigate key moments in a recording. Each bookmark records two data points: the time of the moment and the camera pose (position + orientation) in the 3D scene for viewing that moment. For the sake of testing this technique, we reused the recordings and the participants' behavior from the main user study to estimate the bookmarks. Specifically, we manually chose the key moments and the optimal camera pose based on the participants' recorded teleportation location and interactions with objects in the scene. Because of the asynchronous nature of recordings, more advanced techniques may automatically organize the data for easier consumption by identifying key moments in a VR recording based on spoken content or recorded behaviors (e.g., interacting with objects), or to plan optimal viewpoints [6].

To test the feasibility of our design implications and gauge user reactions toward our suggested interfaces, we conducted an informal evaluation comparing each of the four PoC interfaces against the original experimental system described in Section 3.1. Eight participants, none of which experienced our main study, were convenience sampled by the authors. After an on-boarding session similar to Section 3.2.3, participants conducted 4 types of tasks, one task using each of the new interfaces and another using the original interface (in total, 4 × 2 tasks). Each of the four tasks replicated a challenging situation described in the relevant design implication: (Proxemics) teleporting to a location where the avatar can violate the viewer's private space, (Scene-diff) seeing a scene transition with sudden changes and describing the changes, (Awareness) viewing and describing the avatar's activity hidden behind a wall, and (Navigation) time-navigating a complex recording. We collected subjective ratings (e.g., questionnaires on user frustration at the proxemic violation for the Proxemics task and at the scene transition for the Scene-diff task, and weighted NASA TLX ratings for Awareness and Navigation tasks) and qualitative responses after each of the 8 sessions. The order of the two systems and task materials were balanced.

Overall, the participants' rating and comments were in favor of each of our PoC interfaces over the original. (Proxemics) Participants rated their experience with the proxemic teleportation interface to be less intrusive ($M(SD)^{PoC} = 3.00(1.51), M(SD)^{Original} = 4.00(1.31)$) and more comfortable ($M(SD)^{PoC} = 3.50(1.20), M(SD)^{Original} = 2.89(1.13)$) than the original teleportation interface. (Scene-diff) The animated transition interface got higher ratings, on average, than the original with respect to the ease of anticipating what will happen ($M(SD)^{PoC} = 4.13(.83), M(SD)^{Original} = 3.00(1.51)$) and adjusting to the scene ($M(SD)^{PoC} = 3.88(0.99), M(SD)^{Original} = 3.50(1.07)$) after the transition. The X-Ray and Timeline interfaces yielded lower TLX workload ratings than the original interfaces ((Awareness) $M(SD)^{PoC} = 28.8(20.1), M(SD)^{Original} = 37.8(28.9)$; (Navigation) $M(SD)^{PoC} = 30.7(19.3), M(SD)^{Original} = 35.1(18.6)$). Although no ratings showed statistical significance due to the small sample size, the qualitative comments were in line with our design intent and expected user responses, confirming the feasibility of our design implications: "(Proxemics) it would prevent me from teleporting to a place where the avatar would suddenly jump inside my body," "(Scene-diff) it was a lot easier to identify moved objects in the animation," "(Navigation) it allowed me to [...] teleport to the area of the scene that has the best viewing location," and "(Awareness) being able to see how the avatar was trying to interact and tell me what to do."

6 CONCLUSION AND FUTURE WORK

In this paper, we investigated asynchronous VR collaboration based on multimodal recording for spatial tasks. Using our experimental system, which affords the capture and replay of multiple interaction modalities, we conducted a qualitative, exploratory study where participants consumed and produced recordings to collaborate on interior design. Participants displayed social behaviors, such as proxemics and empathy, even when viewing a replayed representation of an asynchronous collaborator. Viewpoint management and awareness were critical for coordinating varying perspectives without real-time feedback in spatial VR. While immersion in CVEs can help build up 3D context in spatial tasks, it may disorient users when switching between different versions of scenes.

Although interior design is a popular representative spatial task, not all of our results are necessarily transferable to other types of asynchronous tasks in VR. They may vary by spatial scale (e.g., urban planning), number of collaborators (e.g., virtual conferences), and the complexity of the required multimodal interactions. Applying our method to different task contexts will help us triangulate and contrast findings, further inspiring future work.

We are aware of some limitations of the study. First, our depiction of a versioning mechanism shown in Figure 4 is specific to our implementation of the experimental system and cannot be generalized to all possible implementations of MAVRC systems. Another implementation-specific example is that, as an initial exploration into characterizing MAVRC, our system only supported viewing a single recording at a time. Future work could investigate the utility of a different implementation that affords viewing multiple recordings or scenes in parallel. Second, the choice of interaction modalities are only representative of what is affordable in the current VR infrastructure. For example, given that eye-tracking is beginning to be commonly integrated into the latest VR headsets [41], one might consider replicating our study with an implementation that can capture user gaze as a way to register awareness and attention of the collaborators. Third, participants only exchanged a single chain of multimodal recordings. Tasks that involve back-and-forth chains of asynchronous messages are costly, but important to study. Tracking across a long history of past scene changes is likely more challenging to understand and manage.

Extending our design implications to *augmented reality* (AR) may be a promising angle. However, regarding version control issues, reality-virtuality conflicts [88] are more difficult to resolve than virtuality-virtuality conflicts in VR. An AR user's viewpoint is strictly bounded to their real body, so addressing the time-space navigation problem will be a non-trivial problem. Exploring how heterogeneous collaborators (e.g., multiple co-located/distant or synchronous/asynchronous collaborators) might work together in a similarly immersive system could also be challenging but rewarding.

7 ACKNOWLEDGEMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-04591). This work was supported in part by a generous gift from Adobe Research.

REFERENCES

[1] Ian Arawjo, Dongwook Yoon, and François Guimbretière. 2017. TypeTalker: A Speech Synthesis-Based Multi-Modal Commenting System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1970–1981. https://doi.org/10.1145/2998181.2998260

[2] B. Avery, C. Sandor, and B. H. Thomas. 2009. Improving Spatial Perception for Augmented Reality X-Ray Vision. In *2009 IEEE Virtual Reality Conference*. 79–82. https://doi.org/10.1109/VR.2009.4811002

[3] B. Bach, E. Pietriga, and J. Fekete. 2014. GraphDiaries: Animated Transitions and Temporal Navigation for Dynamic Networks. *IEEE Transactions on Visualization and Computer Graphics* 20, 5 (May 2014), 740–754. https://doi.org/10.1109/TVCG.2013.254

[4] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2003. Interpersonal Distance in Immersive Virtual Environments. *Personality and Social Psychology Bulletin* 29, 7 (2003), 819–833. https://doi.org/10.1177/0146167203029007002 PMID: 15018671.

[5] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic Interaction: Designing for a Proximity and Orientation-aware Environment. In *ACM International Conference on Interactive Tabletops and Surfaces (ITS '10)*. ACM, New York, NY, USA, 121–130. https://doi.org/10.1145/1936652.1936676

[6] William Bares, Scott McDermott, Christina Boudreaux, and Somying Thainimit. 2000. Virtual 3D Camera Composition from Frame Constraints. In *Proceedings of the Eighth ACM International Conference on Multimedia (MULTIMEDIA '00)*. ACM, New York, NY, USA, 177–186. https://doi.org/10.1145/354384.354463

[7] Steve Benford, Chris Greenhalgh, Tom Rodden, and James Pycock. 2001. Collaborative Virtual Environments. *Commun. ACM* 44, 7 (July 2001), 79–85. https://doi.org/10.1145/379300.379322

[8] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Trans. Graph.* 31, 4, Article 67 (July 2012), 8 pages. https://doi.org/10.1145/2185520.2185563

[9] Anastasia Bezerianos, Pierre Dragicevic, and Ravin Balakrishnan. 2006. Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 159–168. https://doi.org/10.1145/1166253.1166279

[10] Anastasiia Bobeshko. 2017. The Future of Interior Design: Virtual Showrooms and Augmented Catalogs. (2017). http://www.masonrydesignmagazine.com/future-interior-design-virtual-showrooms-augmented-catalogs/

[11] Jered Borup, Richard E. West, and Charles R. Graham. 2012. Improving online social presence through asynchronous video. *The Internet and Higher Education* 15, 3 (2012), 195 – 203. https://doi.org/10.1016/j.iheduc.2011.11.001 Emotions in online learning environments.

[12] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320. https://doi.org/10.1177/0049124113500475

[13] Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. 1991. Expressive Richness: A Comparison of Speech and Text As Media for Revision. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*. ACM, New York, NY, USA, 21–26. https://doi.org/10.1145/108844.108848

[14] Henry Chen, Austin S. Lee, Mark Swift, and John C. Tang. 2015. 3D Collaboration Method over HoloLens™ and Skype™ End Points. In *Proceedings of the 3rd International Workshop on Immersive Media Experiences (ImmersiveME '15)*. ACM, New York, NY, USA, 27–30. https://doi.org/10.1145/2814347.2814350

[15] Fanny Chevalier, Pierre Dragicevic, Anastasia Bezerianos, and Jean-Daniel Fekete. 2010. Using Text Animated Transitions to Support Navigation in Document Histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 683–692. https://doi.org/10.1145/1753326.1753427

[16] T. Matthew Ciolek and Adam Kendon. 1980. Environment and the Spatial Arrangement of Conversational Encounters. *Sociological Inquiry* 50, 3–4 (1980), 237–271. https://doi.org/10.1111/j.1475-682X.1980.tb00022.x

[17] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.

[18] Richard L Daft and Robert H Lengel. 1986. Organizational information requirements, media richness and structural design. *Management science* 32, 5 (1986), 554–571.

[19] A. De Lucia, F. Fasano, and R. Oliveto. 2008. Traceability management for impact analysis. In *2008 Frontiers of Software Maintenance*. 21–30. https://doi.org/10.1109/FOSM.2008.4659245

[20] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of Sharing Physiological States of Players in a Collaborative Virtual Reality Gameplay. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4045–4056. https://doi.org/10.1145/3025453.3026028

[21] Arindam Dey and Christian Sandor. 2014. Lessons learned: Evaluating visualizations for occluded objects in handheld augmented reality. *International Journal of Human-Computer Studies* 72, 10 (2014), 704 – 716. https://doi.org/10.1016/j.ijhcs.2014.04.001

[22] Jozef Doboš and Anthony Steed. 2012. 3D Diff: An Interactive Approach to Mesh Differencing and Conflict Resolution. In *SIGGRAPH Asia 2012 Technical Briefs (SA '12)*. ACM, New York, NY, USA, Article 20, 4 pages. https://doi.org/10.1145/2407746.2407766

[23] Jozef Doboš and Anthony Steed. 2012. 3D Revision Control Framework. In *Proceedings of the 17th International Conference on 3D Web Technology (Web3D '12)*. ACM, New York, NY, USA, 121–129. https://doi.org/10.1145/2338714.2338736

[24] Ciro Donalek, S. G. Djorgovski, Alex Cioc, Anwell Wang, Jerry Zhang, Elizabeth Lawler, Stacy Yeh, Ashish Mahabal, Matthew Graham, Andrew Drake, Scott Davidoff, Jeffrey S. Norris, and Giuseppe Longo. 2014. Immersive and collaborative data visualization using virtual reality platforms. In *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 609–614. https://doi.org/10.1109/BigData.2014.7004282

[25] Paul Dourish and Victoria Bellotti. 1992. Awareness and Coordination in Shared Workspaces. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work (CSCW '92)*. ACM, New York, NY, USA, 107–114. https://doi.org/10.1145/143457.143468

[26] Jesse Fox, Dylan Arena, and Jeremy N Bailenson. 2009. Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology* 21, 3 (2009), 95–113.

[27] Doron Friedman, Anthony Steed, and Mel Slater. 2007. Spatial Social Behavior in Second Life. In *Intelligent Virtual Agents*, Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 252–263.

[28] P. Frost and P. Warren. 2000. Virtual reality used in a collaborative architectural design process. In *2000 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics*. 568–573. https://doi.org/10.1109/IV.2000.859814

[29] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/358916.358947

[30] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. 2003. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the conference on Human factors in computing systems - CHI '03*. ACM Press, New York, New York, USA, 529. https://doi.org/10.1145/642700.642703

[31] Saul Greenberg, Nicolai Marquardt, Till Ballendat, Rob Diaz-Marino, and Miaosen Wang. 2011. Proxemic Interactions: The New Ubicomp? *Interactions* 18, 1 (Jan. 2011), 42–50. https://doi.org/10.1145/1897239.1897250

[32] C. Greenhalgh, M. Flintham, J. Purbrick, and S. Benford. 2002. Applications of temporal links: recording and replaying virtual environments. In *Proceedings IEEE Virtual Reality 2002.* 101–108. https://doi.org/10.1109/VR.2002.996512

[33] Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2010. Chronicle: capture, exploration, and playback of document workflow histories. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology.* ACM, 143–152. http://doi.acm.org/10.1145/1866029.1866054

[34] Jonathan Grudin. 1988. Why CSCW Applications Fail: Problems in the Design and Evaluationof Organizational Interfaces. In *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work (CSCW '88).* ACM, New York, NY, USA, 85–93. https://doi.org/10.1145/62266.62273

[35] João Guerreiro, Daniel Medeiros, Daniel Mendes, Maurício Sousa, Joaquim Jorge, Alberto Raposo, and Ismael Santos. 2014. Beyond Post-It: Structured Multimedia Annotations for Collaborative VEs. *International Conference on Artificial Reality and Telexistence Eurographics Symposium on Virtual Environments* (2014), 55–62. https://doi.org/10.2312/ve.20141365.055-062

[36] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (01 Sep 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[37] E.T. Hall. 1992. *The Hidden Dimension.* Anchor Books. https://books.google.ca/books?id=p3g0ngEACAAJ

[38] Steve Harrison, Scott Minneman, and Joshua Marinacci. 1999. The DrawStream Station or the AVCs of Video Cocktail Napkins. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems - Volume 2 (ICMCS '99).* IEEE Computer Society, Washington, DC, USA, 9543–. https://doi.org/10.1109/MMCS.1999.779259

[39] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. 2007. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, Vol. 52. ACM Press, New York, New York, USA, 1029. https://doi.org/10.1145/1240624.1240781

[40] Jim Hollan and Scott Stornetta. 1992. Beyond Being There. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92).* ACM, New York, NY, USA, 119–125. https://doi.org/10.1145/142750.142769

[41] HTC. 2019. Vive Pro. (2019). https://enterprise.vive.com/us/product/vive-pro-eye/

[42] HTC. 2019. Vive Virtual Reality System. (2019). https://vive.com/

[43] Hikaru Ibayashi, Yuta Sugiura, Daisuke Sakamoto, Natsuki Miyata, Mitsunori Tada, Takashi Okuma, Takeshi Kurata, Masaaki Mochimaru, and Takeo Igarashi. 2015. Dollhouse VR: A Multi-view, Multi-user Collaborative Design Workspace with VR Technology. In *SIGGRAPH Asia 2015 Emerging Technologies (SA '15).* ACM, New York, NY, USA, Article 8, 2 pages. https://doi.org/10.1145/2818466.2818480

[44] T. Imai, A. E. Johnson, J. Leigh, D. E. Pape, and T. A. DeFanti. 1999. Supporting transoceanic collaborations in virtual environment. In *Fifth Asia-Pacific Conference on ... and Fourth Optoelectronics and Communications Conference on Communications,*, Vol. 2. 1059–1062 vol.2. https://doi.org/10.1109/APCC.1999.820446

[45] Kegetys. [n. d.]. OpenVR input plugin | Open Broadcaster Software. https://obsproject.com/forum/resources/openvr-input-plugin.534/. ([n. d.]). Accessed: 2019-03-27.

[46] Robert Kennedy, Norman Lane, Kevin Berbaum, and Michael Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. (1993), 203–220 pages. https://doi.org/10.1207/s15327108ijap0303_3

[47] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven Interaction Techniques for Improving Navigation of Educational Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14).* ACM, New York, NY, USA, 563–572. https://doi.org/10.1145/2642918.2647389

[48] Ryohei Komiyama, Takashi Miyaki, and Jun Rekimoto. 2017. JackIn Space: Designing a Seamless Transition Between First and Third Person View for Effective Telepresence Collaborations. In *Proceedings of the 8th Augmented Human International Conference (AH '17).* ACM, New York, NY, USA, Article 14, 9 pages. https://doi.org/10.1145/3041164.3041183

[49] Guang Li, Xiang Cao, Sergio Paolantonio, and Feng Tian. 2012. SketchComm: A Tool to Support Rich and Flexible Asynchronous Communication of Early Design Ideas. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12).* ACM, New York, NY, USA, 359–368. https://doi.org/10.1145/2145204.2145261

[50] David Lindlbauer and Andy D Wilson. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3173574.3173703

[51] John Lofland and Lyn H Lofland. 1971. Analyzing social settings. (1971).

[52] Philip Marriott. 2002. Voice vs text-based discussion forums: An implementation of Wimba Voice Boards. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education.* Association for the Advancement of Computing in Education (AACE), 640–646.

[53] Kasra Rahimi Moghadam, Colin Banigan, and Eric D Ragan. 2018. Scene Transitions and Teleportation in Virtual Reality and the Implications for Spatial Awareness and Sickness. *IEEE transactions on visualization and computer graphics* (2018). https://doi.org/10.1109/TVCG.2018.2884468

[54] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology - UIST '07*. ACM Press, New York, New York, USA, 3. https://doi.org/10.1145/1294211.1294215

[55] Christine M. Neuwirth, Ravinder Chandhok, David Charney, Patricia Wojahn, and Loel Kim. 1994. Distributed Collaborative Writing: A Comparison of Spoken and Written Modalities for Reviewing and Revising Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 51–57. https://doi.org/10.1145/191666.191693

[56] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. CollaVR: Collaborative in-headset review for VR video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17*. ACM Press, New York, New York, USA, 267–277. https://doi.org/10.1145/3126594.3126659

[57] Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2013. Direct Manipulation Video Navigation in 3D. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1169–1172. https://doi.org/10.1145/2470654.2466150

[58] Gary M. Olson and Judith S. Olson. 2000. Distance Matters. *Human–Computer Interaction* 15, 2-3 (sep 2000), 139–178. https://doi.org/10.1207/S15327051HCI1523_4

[59] Sergio Orts-Escolano, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Christoph Rhemann, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, Shahram Izadi, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, and Sameh Khamis. 2016. Holoportation : Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*. ACM Press, New York, New York, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[60] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[61] Yun Suen Pai, Benjamin I. Outram, Benjamin Tag, Megumi Isogai, Daisuke Ochi, Hideaki Kimata, and Kai Kunze. 2017. CleaVR: Collaborative Layout Evaluation and Assessment in Virtual Reality. In *ACM SIGGRAPH 2017 Posters (SIGGRAPH '17)*. ACM, New York, NY, USA, Article 20, 2 pages. https://doi.org/10.1145/3102163.3102186

[62] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: Video-based Asynchronous Video Review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 517–528. https://doi.org/10.1145/2984511.2984552

[63] Amy Pavel, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. 2016. VidCrit: Video-Based Asynchronous Video Review. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*. ACM Press, New York, New York, USA, 517–528. https://doi.org/10.1145/2984511.2984552

[64] Julien Phalip, Ernest a. Edmonds, and David Jean. 2009. Supporting remote creative collaboration in film scoring. In *Proceeding of the seventh ACM conference on Creativity and cognition - C&C '09*. ACM Press, New York, New York, USA, 211. https://doi.org/10.1145/1640233.1640266

[65] Lauriane Pouliquen-Lardy, Isabelle Milleville-Pennel, François Guillaume, and Franck Mars. 2016. Remote collaboration in virtual reality: asymmetrical effects of task distribution on spatial processing and mental workload. *Virtual Reality* 20, 4 (nov 2016), 213–220. https://doi.org/10.1007/s10055-016-0294-8

[66] Gonzalo Ramos and Ravin Balakrishnan. 2003. Fluid Interaction Techniques for the Control and Annotation of Digital Video. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (UIST '03)*. ACM, New York, NY, USA, 105–114. https://doi.org/10.1145/964696.964708

[67] H. Regenbrecht, K. Meng, A. Reepen, S. Beck, and T. Langlotz. 2017. Mixed Voxel Reality: Presence and Embodiment in Low Fidelity, Visually Coherent, Mixed Reality Environments. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 90–99. https://doi.org/10.1109/ISMAR.2017.26

[68] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based Tools for Editing Audio Stories. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 113–122. https://doi.org/10.1145/2501988.2501993

[69] Gabriele Salvati, Christian Santoni, Valentina Tibaldo, and Fabio Pellacini. 2015. MeshHisto: Collaborative Modeling by Sharing and Retargeting Editing Histories. *ACM Trans. Graph.* 34, 6, Article 205 (Oct. 2015), 10 pages. https://doi.org/10.1145/2816795.2818110

[70] Ralph Schroeder. 2012. *The social life of avatars: Presence and interaction in shared virtual environments.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4471-0277-9

[71] Abhishek Seth, Judy M. Vance, and James H. Oliver. 2011. Virtual reality for assembly methods prototyping: a review. *Virtual Reality* 15, 1 (01 Mar 2011), 5–20. https://doi.org/10.1007/s10055-009-0153-y

[72] Venkatesh Sivaraman, Dongwook Yoon, and Piotr Mitros. 2016. Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 1045–1054. https://doi.org/10.1145/2858036.2858416

[73] M. Slater, D. . Pertaub, and A. Steed. 1999. Public speaking in virtual reality: facing an audience of avatars. *IEEE Computer Graphics and Applications* 19, 2 (March 1999), 6–9. https://doi.org/10.1109/38.749116

[74] Dave Snowdon, Elizabeth F Churchill, and Alan J Munro. 2000. Collaborative Virtual Environments: Digital Spaces and Places for CSCW. *Collaborative Virtual Environments* (2000), 1–34. https://doi.org/10.1.1.114.9226

[75] Lisa Stifelman, Barry Arons, and Chris Schmandt. 2001. The Audio Notebook: Paper and Pen Interaction with Structured Speech. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01).* ACM, New York, NY, USA, 182–189. https://doi.org/10.1145/365024.365096

[76] James Tam and Saul Greenberg. 2006. A framework for asynchronous change awareness in collaborative documents and workspaces. *International Journal of Human-Computer Studies* 64, 7 (2006), 583 – 598. https://doi.org/10.1016/j.ijhcs.2006.02.004 Theoretical and empirical advances in groupware research.

[77] John Tang, Jennifer Marlow, Aaron Hoff, Asta Roseway, Kori Inkpen, Chen Zhao, and Xiang Cao. 2012. Time Travel Proxy: Using Lightweight Video Recordings to Create Asynchronous, Interactive Meetings. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12.* ACM Press, New York, New York, USA, 3111. https://doi.org/10.1145/2207676.2208725

[78] Franco Tecchia, Leila Alem, and Weidong Huang. 2012. 3D helping hands: a gesture based MR system for remote collaboration. *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry - VRCAI '12* 1, 212 (2012), 323. https://doi.org/10.1145/2407516.2407590

[79] Víctor Theoktisto and Marta Fairén. 2005. Enhancing collaboration in virtual reality applications. *Computers & Graphics* 29, 5 (2005), 704–718. https://doi.org/10.1016/j.cag.2005.08.023

[80] Michael Tsang, George W. Fitzmaurice, Gordon Kurtenbach, Azam Khan, and Bill Buxton. 2002. Boom Chameleon: Simultaneous Capture of 3D Viewpoint, Voice and Gesture Annotations on a Spatially-aware Display. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology (UIST '02).* ACM, New York, NY, USA, 111–120. https://doi.org/10.1145/571985.572001

[81] L. Wang, J. Wu, X. Yang, and V. Popescu. 2019. VR Exploration Assistance through Automatic Occlusion Removal. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. https://doi.org/10.1109/TVCG.2019.2898782

[82] Steve Whittaker, Julia Hirschberg, Brian Amento, Litza Stark, Michiel Bacchiani, Philip Isenhour, Larry Stead, Gary Zamchick, and Aaron Rosenberg. 2002. SCANMail: A Voicemail Interface That Makes Speech Browsable, Readable and Searchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02).* ACM, New York, NY, USA, 275–282. https://doi.org/10.1145/503376.503426

[83] L. Wilcox. 1998. Dynomite: a dynamically organized ink and audio notebook. *IET Conference Proceedings* (January 1998), 1–1(1). http://digital-library.theiet.org/content/conferences/10.1049/ic_19980381

[84] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-computer Interaction. *Interactions* 23, 3 (April 2016), 38–44. https://doi.org/10.1145/2907069

[85] Ungyeon Yang and Gerard Jounghyun Kim. 2002. Implementation and Evaluation of "Just Follow Me": An Immersive, VR-Based, Motion-Training System. *Presence: Teleoperators and Virtual Environments* 11, 3 (2002), 304–323. https://doi.org/10.1162/105474602317473240

[86] Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: blending ink, speech, and gesture to support collaborative document review. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14.* ACM Press, New York, New York, USA, 481–490. https://doi.org/10.1145/2642918.2647390

[87] Dongwook Yoon, Nicholas Chen, Bernie Randles, Amy Cheatle, Corinna E. Löckenhoff, Steven J. Jackson, Abigail Sellen, and François Guimbretière. 2016. RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16).* ACM, New York, NY, USA, 195–205. https://doi.org/10.1145/2818048.2819951

[88] 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).* ACM, New York, NY, USA, 427–436. https://doi.org/10.1145/3126594.3126601

Received April 2019; revised June 2019; accepted August 2019