ChatDirector: Enhancing Video Conferencing with Space-Aware Scene Rendering and Speech-Driven Layout Transition

Xun Qian*
Google Research &
Purdue University
Mountain View, CA, USA
xunqian@google.com

Feitong Tan
Google Research
Mountain View, CA, USA
feitongtan@google.com

Yinda Zhang
Google Research
Mountain View, CA, USA
yindaz@google.com

Brian Moreno Collins
Google Research
San Francisco, CA, USA
briancollins@google.com

David Kim
Google Research
Zurich, SwitzerLand
kidavid@google.com

Alex Olwal
Google Research
Mountain View, CA, USA
olwal@acm.org

Karthik Ramani
Purdue University
West Lafayette, IN, USA
ramani@purdue.edu

Ruofei Du†
Google Research
San Francisco, CA, USA
me@duruofei.com

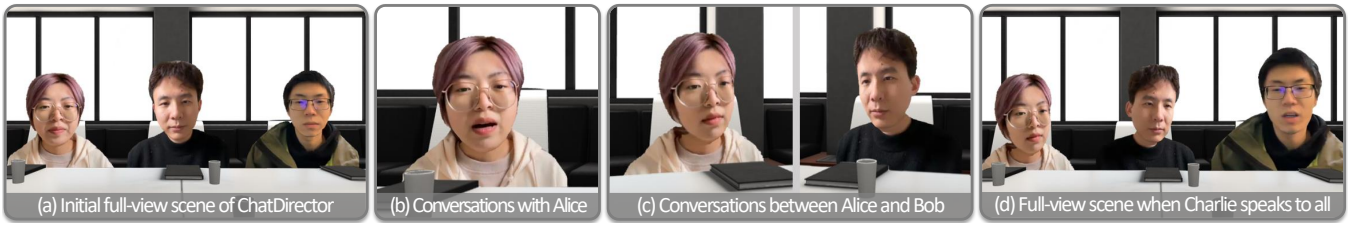

Figure 1: Screenshots of ChatDirector, captured from the local user, Sean's laptop during a remote meeting with Alice (left), Bob (center), and Charlie (right). (a) Using an off-the-shelf laptop or workstation equipped with an RGB camera, ChatDirector depicts remote participants as 3D portrait avatars and renders them in a shared virtual meeting environment. Sean starts his progress update to the team. (b) When Sean inquires about a feature update from Alice, ChatDirector recognizes the speech activity and automatically focuses the camera on Alice, facilitating a more personal one-on-one discussion. (c) Later, Bob steps in and asks Alice further questions, ChatDirector arranges their avatars in a pairwise layout and simulates direct eye contact by orienting their 3D avatars towards each other. (d) When Charlie updates his progress to everyone, the camera is zoomed out with other avatars turning to Charlie, to provide Sean with a visual cue of the speech transition.

ABSTRACT

Remote video conferencing systems (RVCS) are widely adopted in personal and professional communication. However, they often lack the co-presence experience of in-person meetings. This is largely due to the absence of intuitive visual cues and clear spatial relationships among remote participants, which can lead to speech interruptions and loss of attention. This paper presents ChatDirector, a novel RVCS that overcomes these limitations by incorporating space-aware visual presence and speech-aware attention transition assistance. ChatDirector employs a real-time pipeline that converts participants' RGB video streams into 3D portrait avatars and renders them in a virtual 3D scene. We also contribute a decision tree algorithm that directs the avatar layouts and behaviors based on participants' speech states. We report on results from a user study (N=16) where we evaluated ChatDirector. The satisfactory algorithm performance and complimentary subject user feedback imply that ChatDirector significantly enhances communication efficacy and user engagement.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

video conferencing, 3D portrait avatar, tele-presence, attention transition, depth map, depth estimation, machine learning, video-mediated communication, collaborative work, augmented communication

ACM Reference Format:
Xun Qian, Feitong Tan, Yinda Zhang, Brian Moreno Collins, David Kim, Alex Olwal, Karthik Ramani, and Ruofei Du. 2024. ChatDirector: Enhancing Video Conferencing with Space-Aware Scene Rendering and Speech-Driven Layout Transition. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642110

*Project conducted when the first author interned and worked at Google.
†Corresponding author.

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642110

1 INTRODUCTION

Remote video conferencing systems (RVCS) have become indispensable tools in facilitating virtual group meetings across various domains, including work [3, 61] (*e.g.*, weekly stand-ups, city council meetings, group interviews), education [40, 41] (*e.g.*, office hours, parent-teacher meetings, language classes), and social interactions [19] (*e.g.*, family gatherings, conversational games). Prevalent RVCS, such as Google Meet [24], Zoom [86], and Microsoft Teams [48], commonly adopt a grid layout on 2D screens to render remote participants' video streams, enabling open and unrestricted conversations during virtual meetings. While these products have introduced features that extend RVCS capabilities (*e.g.*, screen sharing and hand raising), leveraging RVCS primarily for speech-focused conversations remains a prevalent usage scenario for common users. In common scenarios, like the ones mentioned above, small groups of people engage in virtual gatherings to share insights and exchange opinions in back-and-forth discussions. However, prior research has shown that a traditional 2D-based RVCS often fail to replicate the visual cues present in face-to-face conversations, such as head movements and eye contact, which leads to numerous issues. For example, loss of attention [56, 69] and speech disruptions [11, 52] may impact communication efficiency and engagement. In this paper, we propose solutions to enhance spatial awareness and speech fluency in RVCS for small group conversations. Our approach requires no special equipment beyond a typical computing environment with a common 2D display and an RGB camera.

Recently, there has been a lot of attention in strategies to reproduce in-person visual cues in RVCS. Commercial applications [24, 48, 86] offer features such as dynamic borders and icons to highlight the current speaker, as well as the ability to resize and rearrange windows of remote participants. Meanwhile, Human-Computer Interaction (HCI) researchers have proposed innovative solutions, including visually illustrating eye contact [21, 31, 77] and attention [10, 83], and dynamically adjusting the 2D layouts and visual representations of remote participants according to conversational states [28, 36]. However, these designs are still constrained to a 2D space, adhering to the grid-layout paradigm prevalent in mainstream RVCS. Consequently, users may exert unnecessary mental effort to interpret the presented information. Additionally, the 2D representations lack many of the co-presence attributes that make in-person meetings fluent and engaging. To address these limitations, we introduce an RVCS that leverages the spatial awareness inherent in face-to-face meetings through 3D rendering of both remote participants and environments.

3D capture and display technologies have been explored as potential avenues for simulating face-to-face meetings. Prior art has introduced depth-enabled displays and devices capable of reconstructing the visual representations [22, 63], spatial layouts [39, 54, 85], and head movements [58, 73] of remote participants within a 3D environment. These advances enable users to experience a sense of co-presence with remote attendees, effectively preserving the visual cues inherent in offline conversations. While these promising solutions offer high-fidelity and spatial awareness in visual representations of remote participants, their scalability is hindered because of requiring specialized hardware. This dependency restricts users from starting remote meetings on-the-go, thereby limiting the widespread adoption and scalability of such solutions. Meanwhile, existing research has primarily centered on technical contributions. Yet, the attention [56, 69] and speech [11, 52] issues have not been well addressed in practical multi-user remote meeting settings.

In light of these challenges and opportunities, we introduce ChatDirector, an RVCS that facilitates spatially preserved and speech-fluent remote conferencing on standard computing devices, such as laptops with a front-facing camera. ChatDirector employs a lightweight rendering pipeline that reconstructs 3D portrait avatars from a single RGB webcam, and renders a virtual 3D conference scene. This scene's viewport dynamically adjusts in response to the user's head movements. Additionally, we have designed an algorithm that modulates the layout and poses of remote participants based on speech activity, emulating the natural eye contact and attention shifts of face-to-face conversations. A user evaluation with 16 participants revealed that ChatDirector significantly enhances both communication efficacy and user engagement over traditional RVCS. In summary, our contributions are:

- **A formative study** (N=10) that informs the design considerations to address the challenges in existing RVCS.
- **A web-based RVCS with space-aware scene rendering and speech-driven layout adjustment**, providing a video conferencing experience that resembles the co-presence and fluidity of in-person meetings.
- **A novel real-time RGB video to 3D avatar reconstruction pipeline.** We introduce a pipeline that reconstructs 3D portrait avatars from RGB-webcams via a lightweight depth estimation model, and dynamically renders them in a virtual meeting scene.
- **A speech-driven layout transition algorithm.** We contribute a decision tree algorithm to dynamically adjust the scene layout and avatar poses based on participants' speech states, facilitating natural transitions of attention.
- **A lab study** (N=16). We report on findings from a lab study where the algorithm performance and user feedback imply significant improvements in communication efficacy and user engagement over traditional RVCS.

2 RELATED WORK

After decades of development since the debut of the first video streaming prototype in the 1960s [14], remote video conferencing systems (RVCS) have become ubiquitous in our daily life and work. RVCS serve as a crucial tool for enabling video-mediated communication among geographically dispersed parties. However, previous studies have identified two major issues that hinder the user experience of RVCS when compared to traditional face-to-face meetings: ***lack of visual cues*** and ***lack of spatial relationships***. In RVCS, it is not intuitive for participants to use visual cues such as eye contact and head rotation to effectively draw other users' attention [25, 77] or indicate speech handovers [11, 52], which leads to interruptions and conversation delays. Additionally, the absence of remote participants' spatial relationship further challenges communication grounding and reduces the fluency of remote conferencing [7, 35, 66]. In this section, we review prior works that have endeavored to address these concerns through technical prototypes and user-centered interaction designs.

2.1 RVCS with Visual Augmentation

Commercial RVCS [24, 48, 86] operate on computing devices with RGB cameras, including cellphones and laptops. They allow users to initiate remote meetings anytime and anywhere. These systems have incorporated several features to address the above-mentioned issues by providing visual assistance through highlighted boarders and enlarged windows to indicate the active speaker. Using the same screen-camera setup, Human-Computer Interaction (HCI) researchers have proposed several approaches to further enhance the user experience of RVCS with advanced visual assistance.

Eye contact and head rotation serve as critical visual cues that can implicitly indicate attention transitions in face-to-face meetings [56]. Prior works have utilized gaze detection techniques to perceive and represent mutual eye contact in RVCS by rotating remote users' 2D video windows [78] and synthesizing users' visual appearance with different gaze behaviors [21, 31]. Additionally, eyeView [36] resizes the 2D video windows of remote users to indicate eye contact states, while LookAtChat [30] rearranges and tilts remote users' windows based on ongoing conversations. Furthermore, DeVincenzi et al. [10] and Yao et al. [83] propose blurring irrelevant elements when multiple remote participants appear in one window, to help local users focus on the speaker.

However, these approaches are constrained by the 2D grid-layout form, where visual assistance is limited to adjusting users' live videos and layouts. The 2D window layout frequently changes as remote users join and leave the virtual meeting room, requiring users to expend additional mental effort to interpret the inconsistent changes of the 2D layout. Hence, our goal is to create an RVCS that emulates the benefits of in-person meetings by immersing 2D-screen-based users in a 3D space, granting a spatial perception of the virtual meeting scenarios with intuitive delivery of 3D visual augmentation.

2.2 RVCS with Spatial Awareness

One approach to preserving the spatial awareness is through visualizing each remote user on a separate 2D display and place the displays in front of local users [68]. Eye contact have also been integrated by mapping remote users' head movements onto the rotations and movements of the displays [58, 59, 73]. Yet, the scalability of such systems is limited due to the requirement of additional displays to represent the remote users.

In other perspectives, the metaphor of the 'shared virtual space' [66] (*i.e.*, all participants are co-present in a shared virtual environment, while the spatial relationships are preserved in each participant's local view) has gained attraction in remote conferencing. Commercial applications such as ohyay [53] places 2D live video streams in a virtual 3D background with seats and tables to create a sense of remote participants sitting together. Furthermore, with recent advances in depth cameras and displays, researchers have proposed the concept of immersive conferencing. Using stereo cameras, remote participants can be reconstructed as volumetric avatars with rectification [39], 3D reconstruction [47, 84], and 3D display [37, 63] technologies. These immersive conferencing systems then construct a 3D virtual meeting scene, where the reconstructed avatars are rendered around local users [22, 50, 54]. VirtualCube [85] proposed multiple spatial layout designs that further improve the co-presence and collaboration efficiency with room-scale displays. When looking at the 3D avatar representations of remote participants, users feel immersed in a shared virtual environment with their spatial relationships preserved, akin to traditional face-to-face meetings.

Most of these works focus on technical contributions and system deployment with a maximum of three users (one local and two remote). However, when more participants join the shared 3D environment, whether the system can achieve the same level of visualization and whether the issue of speech interruptions and delays [11, 52] could be resolved remain unclear due to the limited display size. Recently, Meta Horizon Workrooms [32] and Spatial [72] have leveraged Extended Reality (XR) to immerse users into a shared 3D virtual meeting environment using head-mounted devices. However, the visual representations of participants in these systems are either cartoon avatars or pre-set profile photos, which may not be preferable in application scenarios where high-fidelity live visual representation of meeting participants and their facial expressions is required, such as formal meetings or press conferences, for example. Last but not least, all the above-mentioned systems require external hardware setups (*e.g.*, depth cameras, large displays, and head-mounted devices), which significantly limits user mobility and flexibility to collaborate with other PC-based tools and services. In ChatDirector, we fully recognize the spatial awareness brought by prior immersive conferencing systems, and we endeavor to develop a solution that exploits such benefits using widely available setups (*e.g.*, laptops with webcams) to achieve higher scalability.

3 FORMATIVE STUDY

Inspired by prior exploration on the visual augmentation and spatial awareness approaches, we aim to address the main issues of RVCS, *lack of visual cues* and *lack of spatial relationships*, by proposing an integrated solution from both the technical and human-centered perspectives, so that participants can experience a video conferencing that includes the advantages of both in-person conversations and online meetings.

3.1 Procedure

Prior works have identified major drawbacks in basic RVCS, and proposed diverse solutions, as discussed in the Related Works. We aim to advance insights on what the key factors are that would impact the 3D-based experience of RVCS on 2D screens and the corresponding design considerations, to guide us in designing a novel RVCS. Hence, we conducted a brainstorming session with 10 participants (recruited from Google), who had various technical backgrounds including software engineers, HCI researchers, and UX designers. All participants have more than five years in designing and developing computer applications. Moreover, to collect ideas more effectively, the brainstorm discussion was designed to be based on concrete virtual meeting scenarios. We recruited participants who used commercial RVCS in diverse scenarios multiple times per day. The participants reported scenarios including work meetings, family gatherings, language classes, conversational games, and local community and kids' school meetings. As discussed in §2, prior works have addressed various challenges with RVCS from different perspectives and proved their effectiveness

accordingly. While this paper aims to fill the gap between enabling engaging and fluent remote meetings on a 2D device, prior findings are still valuable resources for inspiring the design process. Therefore, the one-hour brainstorming session started with a 10-minute presentation of prior research and systems, accompanied by videos and brief explanations (all the works discussed in §2 were covered). Next, we asked participants to brainstorm specific examples and corresponding concerns on a digital whiteboard to address two prompts in 25 minutes:

(1) If you were to design a new RVCS that is expected to be comparable with face-to-face meetings, what matters and attributes would you consider? And what features would you design?
(2) From a user's perspective, when consuming the features you proposed, what may reduce the overall user experience? And how would you mitigate it?

Finally, each participant presented their ideas, followed by open-ended discussion aiming to achieve a series of agreements. The entire brainstorming session was recorded for post-analysis.

3.2 Design Considerations

Two researchers organized the participants' responses with the affinity diagram approach. By analyzing the user-proposed concerns and addressing findings and suggestions from prior works, we propose five design considerations (DCs) that serve as a guide when designing ChatDirector, to address the conversation fluency and engagement research problems in video-mediated communication.

DC1: Enable spatial awareness in RVCS. All participants proposed at least one design that mimics typical in-person meeting scenarios, addressing the need of co-presence in RVCS [35, 66]. *"The very first idea came to my mind was constructing digital replica of offline scenes where all other people stand in front of me in a meeting room, a bar, or at home. Being present in a same environment would largely increase the feeling of co-presence. (P3)" "I totally agree with [P3]. I always think the virtual background feature of [commercial RVCS] is trying to emphasize that we are not in the same place. (P10)"* P10 also proposed a design of adding reference objects in the scene to improve the feeling of co-presence. *"Think about our offline chats, we always have an unchanged physical environment, like, we sit together at a bar table or meeting room. But in [commercial RVCS], the grid layout changes if someone joins or leaves, which really distracts my attention. (P10)"* Following this design, P7 proposed an idea of anchoring remote participants' window frames at chairs in a meeting room, which led to further discussion: *"[P7], I also thought about it. And from UX design perspective, I was then considering the consistency of the visualization. Now that we want to create a feeling of 3D, we need to stick to it. Showing 2D assets, especially here, not 2D UI buttons, but human faces, in a 3D environment may introduce a perceptive gap, reducing user experience. (P2)", "I would point out the advantage of providing depth-perception in a 3D place. Like what [P2] said, we need to give users an illusion as they are in real world with other participants, tables, and the entire scene. (P1)"* Eventually, the participants reached a consensus that the feeling of co-presence would significantly improve the user experience and we should build a shared environment across all meeting participants while presenting remote participants and assets in a fully 3D manner. Such visualization would underpin a seamless integration of additional features addressing other design considerations discussed below.

DC2: Provide speech-driven assistance. Seven participants raised designs that provide additional assistance rather than a pure reconstruction of a physical scene. The discussion was initiated by P8: *"Originally, I thought just duplicating what we have in offline meetings. Place everyone around me or behind a table. But later, I realized, well, we are already facing some drawbacks because people are not face to face. But we have a computer, and at the end of the day, we are doing an online meeting. We should leverage the computational power here to compensate the reduced experience. (P8)" "Agree with [P8]. I suggest breaking down a meeting scenario into something that a computer can understand. (P5)"* Essentially, the participants dived into the characteristics of group meetings, and achieved an agreement that the assistance should be driven by user speeches. *"I imagine the system always knows who is talking to whom. Only this way can it provide timely assistance such as visual adjustments and hints. In my opinion, group chats are the matters of temporal sequence of speeches. (P11)" "I would let the system detect each user's speech activity as a discrete output, and leverage it to provide proper assistance. And I agree with [P11]'s temporal sequence idea. Because if you think about a group chat, no matter a casual chat, or a company meeting, there are always someone talks to everybody, some people talk with each other, and some people as audience at different moments. (P1)"* The discussion regarding speech awareness was also aligned with prior studies regarding turn-taking and speech fluency [11, 77], and it revealed a consideration to provide digital assistance in RVCS utilizing all participants' speech activities. Meanwhile, the discussion immediately shifted to the next design consideration about what assistance the system should provide.

DC3: Replicate visual cues in offline meetings. In the seven participants' designs, we observed a strong consensus to replicate visual cues such as eye contact and head rotation involved in typical offline meetings, which was used to resolve a key issue in RVCS, loss of attention [56, 69]. *"I was thinking why in offline meetings, I felt so natural and engaged. It might because I could keep track of the ongoing conversations. How? I follow their head movements and eye contact. I know [Bob] is talking to [Alice] because they are looking at each other. So, I unconsciously transit my focus to their talks. And this happens all the time. (P8)" "I play a video game, Danganronpa. In that game, it rotates the camera to different characters when they start to talk. I would add that feature in my design. Just like there is a camera man transit your focus during the meeting. (P9)" "My design target other users. In [commercial RVCS], I often get confused about who is talking to whom by looking at those 2D grids. I believe it's because the absence of direct eye contact between that two users. So, I would add dynamic behaviors to remote participants just like what they will do in offline meetings, such as rotating their head towards each other. (P1)"* Following this consideration, we are motivated to design visual assistance that help users shift their focus properly to the ongoing meeting contents from both the local and remote perspectives.

DC4: Reduce users' mental load. As the participants discussed more complicated features, P1 raised a thought-provoking comment: *"I came up with an idea of placing remote participants in adjacent tiles and rotating their representations based on their speech activities. But, I realized this would increase the mental load, right? And am I*

going back to existing [RVCS] layout? (P1)" "[P1] you are right. And we should not let users do too much. Especially in our scope, speech fluency should not be broken by additional user inputs. I think our system needs to deliver the assistance in an unobtrusive manner. (P3)" All the participants agreed that providing assistive features could help users keep track of the conversations. Yet, we should not make the system over-complicated — not provide too dense information simultaneously. Keeping the meeting fluency and engaging should be prioritized over complex features.

DC5: Maintain a high scalability. While the last point was not explicitly raised by the participants, we believe it serves as another key concern. Existing commercial RVCS allow users to initiate remote meetings using laptops or cellphones, providing sufficient freedom and seamless access to other tools such as text chats and screen sharing on the same device. While we acknowledge the benefits of spatial awareness granted by technical solutions [63, 84, 85], our aim is to create an RVCS with augmented assistance that can be democratized to all common users with the same setup used by existing RVCS (*e.g.*, a laptop with a webcam).

4 CHATDIRECTOR

Following the design considerations, we developed ChatDirector, a remote video conferencing system that depicts users as 3D portrait avatars in a virtual environment with automatic speech-driven layout transitions and avatar rotations, running on *a common off-the-shelf laptop*. In this section, we provide a high-level overview of ChatDirector and then describe the rendering pipeline that enables space-aware visualization of both the shared meeting scene and remote participants. Specifically, we detail the process of reconstructing a 3D portrait avatar of a participant using the live RGB video stream as input, and building a shared meeting environment with space-aware visualization and real-time data communication. Then, we introduce a decision tree algorithm that utilizes the speech states of remote participants as inputs, and visually adjusts the layout and behavior of the remote avatars to help users keep track of the ongoing conversations.

4.1 System Overview

Let's review the example user journey of ChatDirector shown in Figure 1. *Alice, Bob, Charlie, and Sean (the local user) attend an online meeting using ChatDirector to discuss their team project. They join the same remote meeting room using their personal laptops, and turn on their cameras and audio.* Figure 1 shows four screenshots from the local user, Sean's laptop. In Figure 1a, the 3D portrait avatars of the other three co-workers are rendered in a pre-selected virtual conference room. *Sean proceeds to update everyone on his progress, while ChatDirector renders a full view of the scene, decided by the layout transition algorithm, to give Sean a sense of talking to everyone. Later, Sean asks Alice several detailed questions, where ChatDirector zooms in the camera on Alice's avatar (Figure 1b), allowing Sean to concentrate on the one-on-one conversation with Alice. When Bob interjects to ask Alice follow-up questions, ChatDirector adjusts the layout to pairwise focus on Alice and Bob, and turns their avatars towards each other to simulate eye contact during their back-and-forth conversations (Figure 1c). After Sean completes his progress update and the related discussion, Charlie starts his turn. The system zooms out Sean's camera to a full view, enlarging Charlie's avatar, and turns Bob's and Alice's avatars to Charlie, so that Sean's attention transitions to Charlie's speech.* All the layout transitions occur simultaneously on participants' devices, while the resulting scene and avatar behaviors may vary based on the speech activities as perceived from their individual viewpoints. In summary, ChatDirector enables users to engage in remote conferencing with dynamic visual assistance for attention transition, creating a sense of co-presence in the shared 3D meeting scene.

4.2 Space-Aware Scene Rendering Pipeline

4.2.1 Portrait Depth Estimation Model. The high-fidelity visualization of meeting participants is crucial for enabling a space-aware perception of the meeting scene, allowing for more intuitive delivery of implicit visual cues such as eye contact, similar to in-person meetings. Prior systems [5, 63] have shown that 3D representation of remote participants can extensively improve the user experience in terms of immersiveness and engagement. However, these approaches typically require cumbersome external devices. To address DC1 and DC5, in ChatDirector, we aim to reconstruct a user's portrait as a 3D avatar in real-time using only RGB video streams as inputs.

We contribute a real-time portrait depth estimation model that takes a single RGB image and predicts a depth image in the same resolution. We first crop the raw input with a face detection model [16], and segment the foreground using a body segmentation module [6]. In order to optimize computational efficiency, we adopt a light-weight U-Net architecture with short-cut connections. As shown in Figure 2a, the encoder gradually downscales the image, and the decoder increases the feature resolution back to the same as the input. Deep learning features from the encoder are concatenated to the corresponding layers with the same spatial resolution in the decoders to bring high-resolution signals that would benefit the recovering of geometrical details, *e.g.*, object boundary and thin structures. To train the depth estimation model, we use a combination of synthetic data and portrait photos captured from a large group of people. Specifically, for the synthetic training data, we randomly place virtual cameras from a portrait-like view points (*e.g.*, 30-50cm from camera to the person, 0-35 degrees relative to the canonical facing direction) and render 5M pairs of color images and ground truth depth images from the high-fidelity human captures provided by Guo et.al. [27]. To improve the generalization on real photos, we use a state-of-the-art photo relighting method [60] to augment the illumination on the face. For the real images, we collect video scans of the upper body from 200 subjects that rotated their head or used moving mobile phone cameras to capture multiple perspectives. We empirically found that training with the combination of synthetic and real images achieves the best depth prediction quality. The depth estimation network is trained with a scale-invariant loss [15]. During training, we force the decoder to produce depth predictions with increasing resolutions at each resolution, and add a loss for each of them with the ground truth. This helps the decoder predict accurate depth by gradually adding details. When having virtual meetings, users would look at the 2D screen with a webcam equipped closely above the screen (*e.g.*, a laptop setup). Hence, in most cases, the input image is a front-facing

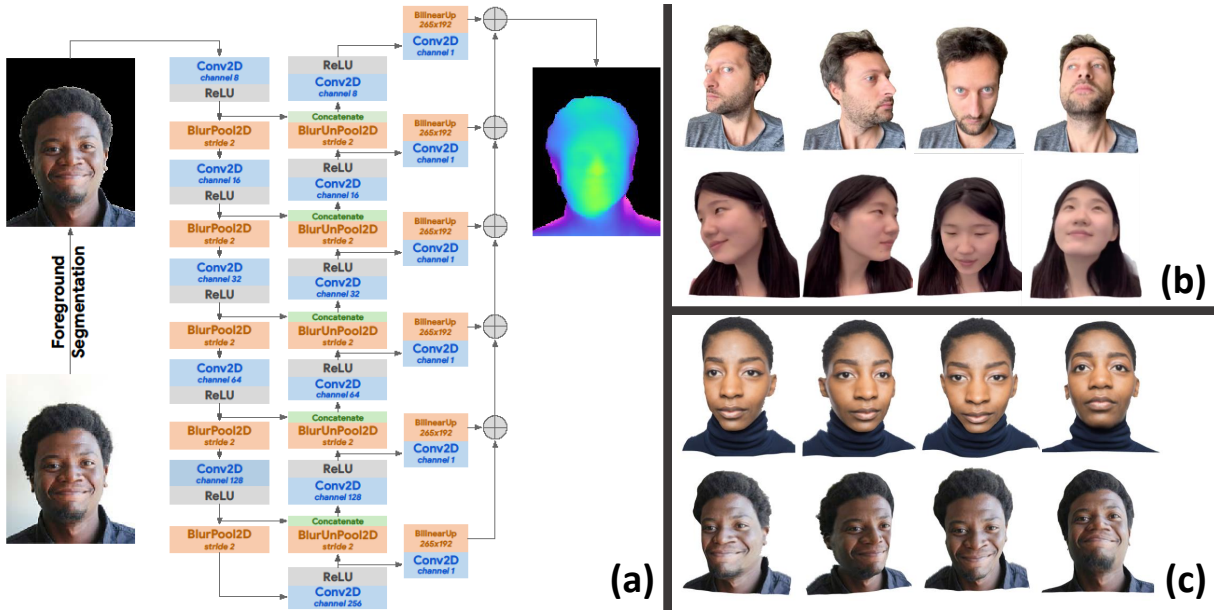

Figure 2: (a) The portrait depth estimation pipeline of ChatDirector. The model takes in a real-time RGB video stream of a local user as the input, crops its portrait region based on a face detection model, then segments the foreground with a body segmentation module, feeds the image to a customized lightweight U-Net, and generates the estimated depth image. (b) Examples of the rendering outputs when the user does not look at the webcam. (c) Examples of the rendering outputs viewing from different perspectives while the user looks at the webcam. Viewing angle=35 degrees.

head. Given the training setting, the depth model also supports the uncommon scenarios where the user does not look directly into the webcam. Examples are shown in Figure 2b. In Figure 2c, we also illustrate more examples of the 3D avatars from different viewing angles while the user looks at the webcam.

4.2.2 Construction of the Space-Aware Shared Virtual Meeting Scene. Figure 3 depicts the comprehensive pipeline that empowers each user to (1) stream their visual representation and speech, (2) receive remote participants' visual presence and speech, (3) reconstruct remote participants as 3D portrait avatars, and (4) render the virtual meeting scene with depth perception. This pipeline also enables ChatDirector to recognize remote participants' real-time speech states and independently control the behavior of each portrait avatar, addressing DC2. These capabilities are crucial for the speech-driven layout transition algorithm detailed in §4.3.

We leverage WebRTC [80] for data communication among all participants, where the peer connections are set using a back-end server [71]. On each user's device, the depth estimation model continuously infers depth images, and our system streams the horizontally stacked RGB and depth images out via the video channel. Meanwhile, the local user's speech, together with the recognized transcriptions (detected by the Web Speech API [79]) are streamed out via the audio and data channels respectively. On the receiving end, ChatDirector renders a space-aware virtual environment, mimicking in-person meeting scenarios from the local user's first-person view. Visually, a custom shader is used to reconstruct all remote participants as high-fidelity 3D portrait avatars from the remote video channels. The avatars are then placed at the pre-designated positions in a virtual room asset. In large-scale remote meetings, commercial RVCS [24, 86] only show a subset of the participants in the main meeting grid to avoid excessive mental load and distraction of the overall visualization, which is also raised in DC4. Following this concern, ChatDirector only visualizes a certain number of remote participants (6 in the current design) as 3D portrait avatars in the virtual scene, while hiding others and listing their names in a drop-down menu (note that the audio of hidden users are still available to the local user). We will discuss in §7 how to address large-scale meeting scenarios in future work. In order to further improve the spatial awareness of the 3D meeting scene (DC1), we adopt the idea proposed by prior immersive conferencing systems [63, 84] in our camera-screen setup. We detect the local user's head movement using a facial landmark detection module [16], and slightly adjust rendering camera's pose to achieve a depth perception effect. The final 3D virtual meeting scene is shown in Figure 1.

4.3 Speech-Driven Layout Transition

In this section, we delve into the details of the layout transition algorithm that offers speech-responsive support (orange block in Figure 3), addressing DC2 and DC3.

4.3.1 Algorithm Inputs. In the example shown in Figure 1, from Sean's viewpoint, when Alice speaks to him, his attention is focused on Alice's face. When Bob and Alice converse with each other, Sean's attention encompasses both individuals simultaneously. Previous studies [11, 52, 69] have highlighted the importance of being

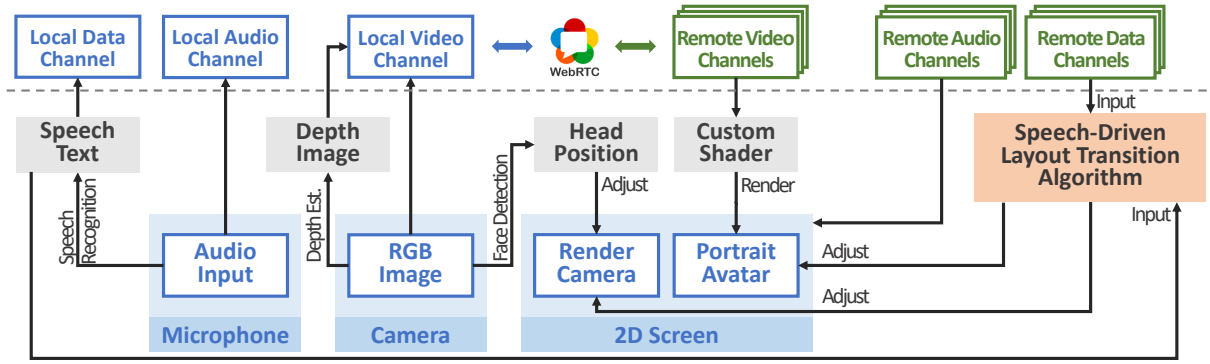

Figure 3: The end-to-end workflow of ChatDirector for space-aware scene rendering. The blue blocks reside in the local user's domain, while the green blocks are incoming remote channels. The gray blocks represent the intermediate outputs and modules. The orange block indicates that the speech-driven layout transition algorithm uses the detected speech transcriptions of both local and remote participants to adjust both the camera pose and the avatars' behaviors in the local user's screen.

cognizant of every participant's speech activities for sustaining smooth and engaging face-to-face conversations. Similarly, in the formative study, participants raised the same concern (DC2). Hence, we propose to leverage all participants' speech activities as inputs to the layout transition algorithm in our system. Generally, we consider three ***Speech State***s, inspired by the user quotes discussed in DC2.

- ***Quiet*** $\{i\}$ represents the state when the individual (i) is not speaking. This frequently occurs in situations like formal presentations and weekly group meetings, where the audience remains silent, attentively listening to other speakers.
- ***Announce*** $\{i\}$ represents the state when the individual (i) is making an announcement to the other participants, or generally speaking to everyone, *e.g.*, a presentation, or a teacher lecturing.
- ***Talk-To*** $\{i \rightarrow j\}$ represents the state when the individual (i) seeks to engage in a dialogue with a specific remote user (j). For instance, a person may ask a presenter questions in team meetings and project presentations after the progress update. We further propose a ***Pair*** $\{i \leftrightarrow j\}$ state as a subset of the *Talk-To* state, which indicates that there exists two users who are talking to each other. For instance, after person i initiates a question (*Talk-To* $\{i \rightarrow j\}$), the presenter j enters *Talk-To* $\{j \rightarrow i\}$ as well, which forms a continuous back-and-forth conversation between two participants.

The *Speech State*s of both local and remote participants are inferred from transcribed speech. Typically, a user enters the *Quiet* state when the system has not received any speech transcription for 1.5 seconds (empirically set). After the system has detected speech for over 0.5 seconds, it implies either an *Announce* or *Talk-To* state. We adopt the keyword detection method to distinguish these two states. If the system detects the user Id (user j) in the live transcription of the user i, the *Speech State* of the user i is set to *Talk-To* $\{i \rightarrow j\}$. We will show the GUI for users to enter their user Ids in §4.4. We use a keyword dictionary ({ "all", "everyone", "everybody"}) to indicate the *Announce* state, with only the first 3 words in every incoming speech transcription will be examined to eliminate potential ambiguity. Note that if user i was in *Talk-To* state, and the system detects speech again, the state remains unchanged unless one keyword for another *Talk-To* or *Announce* state is detected. When using ChatDirector, users are directed to investigate the *Announce* keywords and add the ones they feel natural and preferred to use in their personal announcement speech. We will analyze user feedback on this novel feature in §6.

4.3.2 Algorithm Outputs. Following DC3, we propose two algorithm outputs to help users infer ongoing speech activities, and shift their focus promptly, thereby enhancing overall conversation fluidity and engagement. First, the algorithm replicates the behavior of one who gazes at different people as the conversations go on. It outputs one of the three ***Layout State***s that shows different field-of-views (FOVs) and scene layouts by rotating and zooming the virtual camera. Moreover, recalling the participants' comments in DC4, we avoid designing an over-complicated visualization or grid-layout-like design that may increase user's mental load. In offline group chat live streaming and video editing areas, researchers and developers have investigated how directors control the presentation based on the ongoing conversations [38, 42, 65], hence, leading the audience's attention transition throughout the conversations. In this paper, we follow these works and in-person meet scenarios and propose three *Layout State*s.

- ***One-On-One*** $\{i\}$ renders one single remote avatar (i) on the 2D screen (Figure 1b), which mimics the in-person scenarios where one hopes to maintain eye contact with another person during one-on-one conversations such as post-presentation Q&A, intense back-and-forth discussions in casual exchanges, and conversational games. Such a design is also aligned with the speech-turns ideas in prior elicitation studies [52] that in a common group conversation, only one person speaks up at one time and speech turns should happen seamlessly to ensure a smooth conversation experience.
- ***Pairwise*** $\{i, j\}$ places two remote users (i, j) horizontally in two split viewports (Figure 1c), which mainly addresses the *Pair Speech State*, representing scenarios of listening to a one-on-one conversation between two individuals. Note that this design is also adopted in the above-mentioned live streaming works

[42, 65], which has been proved to be an effective way to present one-on-one conversations to an audience.
- ***Full-View*** renders the entire virtual meeting environment with all available remote participants (Figure 1a and d). This state aims to address the needs when the conversation involves multiple participants (*e.g.*, a general announcement to all participants in a group meeting, or multiple pairs of one-on-one conversations).

As one shifts the gaze at different people, each remote participant also switches the eye contact target by slightly rotating the head in face-to-face conversations. With the help of the 3D portrait avatar representation and the spatial awareness inherent in the virtual scene, we could replicate such behavior in ChatDirector in a more natural manner than rotating the 2D windows [78] or displays [28, 58] adopted by prior works. We propose two ***Avatar State***s that rotate each remote avatar in the local user's virtual scene to indicate remote participants' attention transition.

- ***Local*** $\{i\}$ rotates the remote participant i towards the rendering camera as if looking at the local user.
- ***Remote*** $\{i \rightarrow j\}$ indicates the remote participant i is looking at participant j with the corresponding rotation.

4.3.3 Decision Tree Algorithm. The decision tree algorithm is shown in Figure 4a. The algorithm starts from examining the local user's *Speech State*. The first two straightforward cases shown in Figure 4a reflect the scenarios when the local user's *Speech State* is either *Announce* or *Talk-To*. When the local user is *Quiet*, which means the local user is engaged in other conversations as a listener, the algorithm starts to check the *Speech State*s of all other remote participants in sequential order for the existence of: (1st) *Announce* $\{i\}$, (2nd) *Pair* $\{i \leftrightarrow j\}$, and (3rd) *Talk-To* $\{i \rightarrow j\}$. In order to guarantee that the algorithm has the potential to be utilized in more complicated scenarios with more participants, we then consider the number of each *Speech State* during the decision process. Considering that in offline scenarios, an individual's attention and gaze are constrained, the algorithm also aims to prevent rendering an over-complicated virtual scene. Hence, when there are multiple engaging *Talk-To* or *Pair*, the algorithm switches the layout back to *Full-view* rather than multiple *Pairwise* viewports. Eventually, the algorithm outputs one of the 9 available cases with both a *Layout State* and *Avatar State*s for all remote avatars (Figure 4a). Now, the system starts to adjust the virtual meeting scene by manipulating the render camera to reflect the *Layout State* and the corresponding 3D portrait avatars for the *Avatar State*s. In *Pairwise*, we leverage the spatial relationship among the remote avatars to ensure the avatars with the *Remote Avatar State* can properly rotate towards each other. When there are more than one remote participants in *Announce*, we rotate each avatar with the *Remote Avatar State* to the closest *Announce* avatar. Moreover, when the *Layout State* is *Full-View*, we slightly enlarge the remote avatar who is in either the *Announce* or *Talk-To Speech State*s to inform the local user to pay specific attention.

We detail the decision process of the algorithm with concrete examples (Figure 1). First, as shown in Figure 1a, the local user, Sean, initiated the remote conferencing with a general announcement. As Sean's *Speech State* was *Announce*, the algorithm went to case 1, where a *Full-View* layout was used to render all the other participants, while all looked at Sean as each remote participant holds the *Local Avatar State*. In Figure 1b, Sean and Alice had one-on-one conversations. This led to case 2 since Sean was in the *Talk-To* $\{Local \rightarrow Alice\}$ *Speech State*. Therefore, the system only rendered Alice's avatar in Sean's view. In Figure 1c, Sean was *Quiet*. Meanwhile, both Bob and Alice were having conversations with each other (*Talk-To*), which indicated the existence of a *Pair* $\{Alice \leftrightarrow Bob\}$. As a result, the algorithm output case 4 and drove the system to enter the *Pairwise* $\{Bob, Alice\}$ *Layout State*, and rotated both avatars towards each other (*Remote* $\{Bob \rightarrow Alice\}$ and *Remote* $\{Alice \rightarrow Bob\}$). Last but not least, when Charlie started to make an announcement to every one (*Announce*), the layout was changed back to *Full-View*, and all the other avatars rotated towards Charlie as if all the participants were paying attention to Charlie's speech (case 3).

Moreover, as shown in Figure 4, the algorithm is deployed on each participant's device, which leads to distinct and tailored outputs for each participant in every moment. For instance in Figure 4b, we show the screenshots taken from the four participants' devices when Alice and Bob are discussing. According to the algorithm, since Sean and Charlie are *Quiet*, the algorithm chooses to the *Pairwise Layout State* together with two *Remote Avatar States*. On the other hand, for Alice and Bob, since they are talking, the system renders *One-On-One* respectively. The design of the algorithm also ensures reasonable speech-visualization coordination on each participant's device. For instance, when two users are talking to each other with *Talk-To Speech State*s in *One-On-One Layout State*s, a *Quiet* user is listening to the conversations using *Pairwise Layout State*. Later, when another user *Talk-To* the *Quiet* user, the layout will be immediately changed to *Full-view* to make sure that the *Quiet* user is aware of the newly initiated conversation.

4.4 Implementation

We trained the depth estimation model on 16 NVIDIA V100 Tensor Core (32GB) [51] for 72 hours. The avatar rendering pipeline was validated using Rapsai [13]. The resolution of the RGB live video streamed via WebRTC is 360×480 pixels while the resolution of the depth image, as mentioned in the model description, is 192×256 pixels. Currently, ChatDirector supports rendering 6 remote participants using an Apple MacBook Pro (M1 with 32GB unified memory) at 30FPS. As described before, more remote participants are supported but will be listed in a drop-down menu. We will discuss future improvements in the Limitation section. As shown in Figure 5a, we develop a website for users to join a shared meeting room. The previously mentioned back-end socket server will help construct WebRTC peer connections among the users who enter the same meeting id. Meanwhile, we provide a GUI (Figure 5b) that provides users with necessary capabilities including toggling on and off audio and video, adjusting the sensitivity of head-movement detection for the spatial awareness, changing room assets with pre-set avatar placements, and adding custom keywords for triggering the *Announce Speech State*. During the execution of the layout transition algorithm output, a time threshold of two seconds is implemented to avoid fluctuating transitions of the *Layout State* and *Avatar State*.

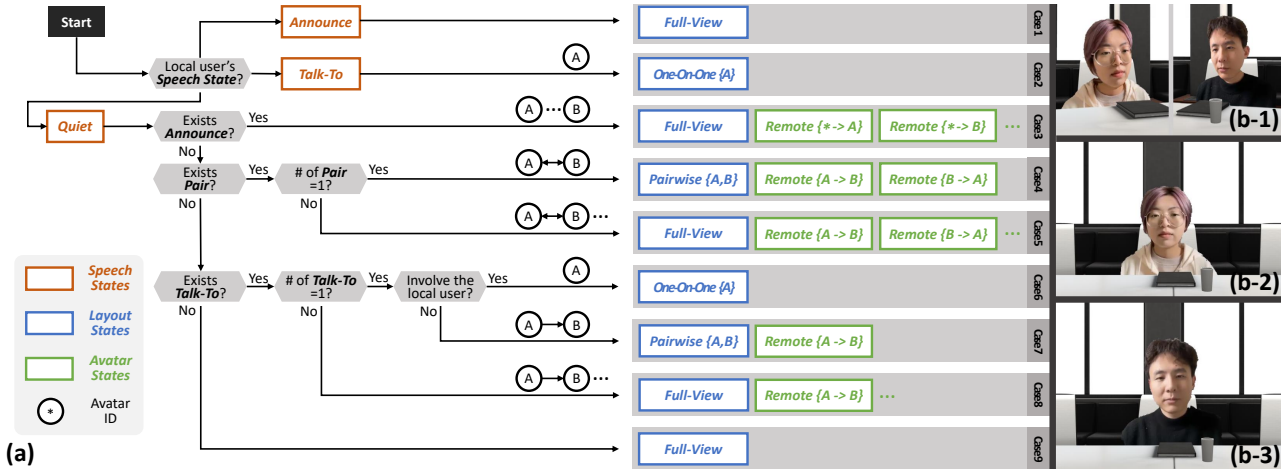

Figure 4: (a) The decision tree algorithm of ChatDirector. The gray blocks indicate the decision process. The algorithm outputs one out of the 9 available cases with one *Layout State* for the entire virtual scene and the *Avatar State*s for all remote avatars. Note: The *Avatar State* is *Local* for all non-depicted remote avatars. (b) The screenshots taken from the four participants' devices when Bob and Alice are discussing, indicating the distinct algorithm output for each participant at the same moment. (b-1) Sean's and Charlie's views: Case 4. (b-2) Bob's view: Case 2. (b-3) Alice's view: Case 2.

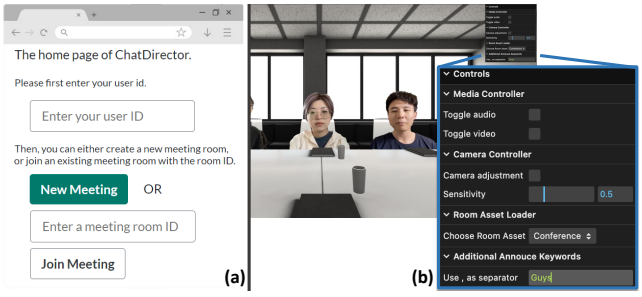

Figure 5: (a) The home page that enables participants to enter user ID and meeting room ID. (b) The GUI that provides proactive control over the system functionalities.

5 APPLICATION SCENARIOS

Since the COVID pandemic, virtual meetings have become a popular norm for a variety of purposes, including one-on-one online consultations [18], office meetings [4], and large-scale online classes [75]. In this section, we aim to illustrate the significance of spatial awareness and attention transition assistance facilitated by ChatDirector through multiple application scenarios where more than two participants engage in intense conversations.

Brainstorming. Brainstorming is a creative problem-solving technique that encourages open and free-flowing discussion among participants to generate new ideas or approaches to a given topic or challenge. The process typically involves frequent turn-taking for idea grounding and sudden announcements with inspiring thoughts. In Figure 6a-1 and a-2, five students are having a brainstorming session to come up with an idea for a toy design class project. ChatDirector renders different *Pairwise Layout State*s as the meeting progresses. The layouts indicate that the female student who sits in the middle turns to different students in different *Pairwise* layouts, so that the local user (coordinator) can easily keep track of the current discussion between different students.

Debates. A debate is a structured form of discussion involving participants arguing for or against a specific topic, statement, or proposition. Debates typically feature two opposing sides, each presenting well-reasoned arguments and evidence to support their respective positions. In this application scenario, we mainly focus on the viewpoint of the audience to demonstrate that with ChatDirector, debate can be more engaging and interesting to watch. For instance, when a team member (the second left person) makes an announcement, a *Full-View* is used to replicate an in-person debate scene from the audience's viewpoint where all the other participants look at the speaker (Figure 6b-1). In Figure 6b-2, a *Pairwise* layout better sets the atmosphere of an intense debate between two opposite members.

Conversation games. Online conversation games are entertaining activities that stimulates fun conversation, creativity, and icebreakers among friends and strangers. Typically, participants are expected to actively listen and react to each other's contributions and announcements, which leads to frequent attention transition and back-and-forth communication with different players. Here, six people are playing a conversation game, Fact or Fiction [17]. When there are discussions between one or more *Pair*s, the system automatically helps the local user transit to the proper layout, so that the local user can better collect useful information from the conversations (Figure 6c-1 and c-2).

Remote office hour sessions. The COVID-19 pandemic has significantly impacted the educational landscape [75], prompting a rapid shift to remote learning and online platforms. As a result, online office hours have become welcomed by both students and instructors. In this example, we show, from an instructor's perspective, how ChatDirector improves the online office hour experience when explaining homework problems. Typically, a *One-On-One*

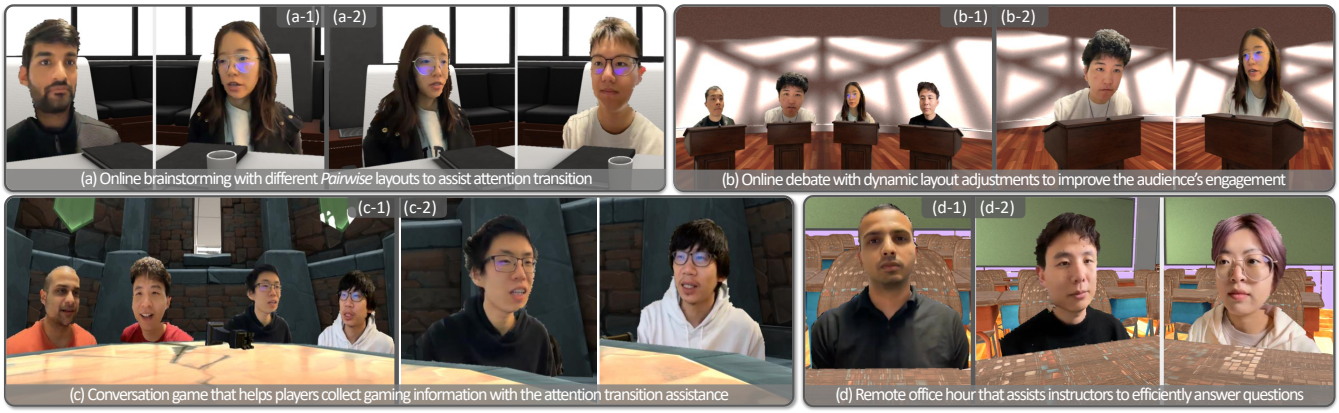

Figure 6: Application scenarios enabled by ChatDirector: (a-1 and a-2) A remote brainstorming session. (b-1 and b-2) An online debate. (c-1 and c-2) An online conversation game. (d-1 and d-2) A remote office hour.

layout helps the TA concentrate on answering each student's questions (Figure 6d-1). Meanwhile, the TA is also willing to engage in the discussions among students to ensure they have digested the knowledge (Figure 6d-2).

6 USER STUDY

In this section, we describe a systematic user study that was conducted to evaluate how ChatDirectoraddresses the research questions identified in this paper. One key contribution of ChatDirector lies in the integrated system design with ML technical support that offers attention transition assistance with 3D-like visualization in 2D-screen-based RVCS. Thus, we first investigated whether the space-aware scene rendering and the speech-driven layout transition performs to participants' expectations and facilitates fluid conversations. Further, we evaluated how ChatDirector impacted conversation engagement and overall virtual meeting experience from a system-level contribution's perspective. We envisioned the findings of this paper would enlighten future research in democratizing 3D-based assistance in RVCS, and inspire future studies on how to leverage the designs of ChatDirector in more virtual meeting scenarios. With commercial RVCS platforms [24, 48, 86] continuing to predominate in this field, we chose a commercial RVCS, Google Meet [24], as a benchmark to explore how our system could offer performance on par with, or superior to, these established commercial systems.

6.1 Participants

We invited 16 participants (4 females and 12 males), with an average age of 25.75 (SD=2.77) from our institution. All participants had previous experience with commercial RVCS (*e.g.*, Zoom [86], Google Meet [24], and Microsoft Teams [48]) in both personal and professional scenarios: 13 participants had used RVCS for attending online classes and formal project presentation; 6 had used RVCS for group discussions with classmates and friends. 13 participants had used RVCS for more than once per week, while 7 had used RVCS for more than once per day. None of the participants had prior experience with our system before participating in the study.

6.2 Procedure

We conducted a 60-minute group user study with four participants per group in a controlled lab setting. The four participants were asked to sign a consent form upon their arrival. Afterward, the researcher provided a brief introduction to the study's purpose and procedures. For each group, we provided the participants four laptops installed with ChatDirector, and ensured everyone wore headphones or were physically dispersed. The study consisted of two 20-minute sub-sessions, with each group conducting a remote conferencing task using either ChatDirector or Google Meet [24] (labeled as "Video" in the questionnaire results). The arrangement of the tasks and systems was shuffled to counterbalance the data. For the sub-session with ChatDirector, the researcher also provided a tutorial including instructions on how to join a shared meeting room and how to use the GUI, including asking the participants to add custom *Announce* keywords if necessary. Additionally, the participants were instructed to report any unexpected performance of the layout transition algorithm by clicking two buttons displayed on the GUI: one for the *Layout State* and one for the *Avatar State*. This allowed for a quantitative assessment of the algorithm's performance.

As mentioned in §1, while RVCS have been adopted in diversified conversation-intense virtual group meetings, the similarities in speech interactions across them allow researchers to conduct elicitation studies and develop systems that address common limitations. In this user study, we mainly targeted usability evaluation of ChatDirector. Considering data counter-balancing of the study setup, we selected two virtual meeting scenarios, a group debate and a conversational game, that not only represented the typical virtual meeting scope focused in this paper, but also contained adequate complexity that could extensively trigger system features to help assess effectiveness. Specifically, these two tasks consisted of conversational interactions raised in prior works [38, 42, 52, 65] such as one-to-all announcements and back-and-forth speech turns. In the debate task, each participant was instructed to present either a supporting or opposing claim on the debate topic with evidence, followed by open discussions and counterexamples by other participants. In the conversation game task, each participant was asked

to provide a word for others to guess, providing basic information about the word, followed by more questions from other participants until the word was successfully guessed. The researcher did not provide any guidance during the two sub-sessions except for time-up warnings and technical issues. Screens and error logs were recorded for verifying participant-reported errors and to ensure that there was no other unexpected performance of ChatDirector, and additionally also to contextualize participant quotes during post-study analysis. After each sub-session, the participants were asked to complete a 7-point Likert-scale questionnaire regarding the user experience. After completing both sessions, the participants were asked to complete the Temple Presence Inventory (TPI) [45] questionnaire that was designed to measure dimensions of presence. Additionally, an open-ended verbal interview was conducted by the researchers to collect subjective feedback on ChatDirector.

6.3 Results

All 16 participants across the four groups successfully completed the two remote meeting scenarios using the corresponding RVCS. We report the results of the user study based on the research question we aim to address: whether ChatDirector succeeds in improving the overall conversation flow and engagement by the speech-driven layout and avatar transitions within the space-aware shared meeting environment. We analyzed the resultsusing the Wilcoxon Signed-Rank Test [81] for the Likert-scale questions to examine potential statistical differences between ChatDirector and commercial RVCS. Following the recommendations from previous research [70], we ensured that the sample size for the Wilcoxon Signed-Rank Test exceeded 15 pairs. Yet, considering the limited sample size, we hold a conservative opinion on these test results, and provide user feedback as supplementary evidence to support our findings. We summarize the key takeaways of the user study as follows, and elaborate on the study results in the following sub-sections.

- ChatDirector effectively addresses speech-related issues involved in RVCS [52] given the high accuracy of the layout transition algorithm outputs, as well as the preferable user ratings on the attention transition assistance.
- ChatDirector enhances co-presence and engagement when compared with commercial 2D-based RVCS, which is supported by the TPI ratings and constructive participant feedback.

6.3.1 Attention Transition and Speech Fluency. From prior studies, we identified the importance of assisting remote participants to keep track of the ongoing conversations. Following the DC2 and DC3 distilled from the formative study, we contribute a speech-sensitive algorithm to dynamically change the layout and spatial behaviors of remote avatars in a shared meeting environment. During the sub-sessions that used ChatDirector, the layout transition algorithm output on average *Layout State* change 10.44 (SD=3.10) times and *Avatar State* change 66.50 (SD=16.65) times. Regarding the *Layout State* behavior, the participants reported 0.50 (SD=0.63) unexpected mistakes, which led to a 94.79% (SD=6.88%) accuracy. For the *Avatar State*, the accuracy was 98.80% (SD=1.26%) with 0.88 (SD=0.96) unsatisfactory avatar behavior. *"The layout and avatar behaved very naturally and smoothly. I didn't need to pay additional attention and it already showed me what I wanted. (P11)"* The participants also acknowledged the need for customizing the *Announce* keywords. *"It makes sense to me to use some keywords for announcement. In group conversations, you really need to make some claims to let everybody pay attention to you. It was very natural and didn't break the overall speech fluency at all. (P4)" "When you asked me to add some [Announce keywords], I realized I always say 'alright' or 'awesome' when I want to conclude one-on-one conversations and come back to an announcement speech. I found ChatDirector did a good job detecting my habit and showed [Full-view] accurately." (P6)*

The participants also welcomed the improvements to attention transition brought by our system, as shown in Figure 7a. When participants were speaking, they appreciated that ChatDirector gave them an explicit feeling that the remote participants started to pay attention to them using *One-On-One Layout State* (Q1: M=6.13, SD=0.81). In contrast, the commercial RVCS received a significantly less preferable result (Q1: M=4.06, SD=1.18) with Z=-3.30,p<.01. Similarly, our visual assistance also enables the remote participants to rapidly react to the local participant so that the local participant had a significantly better feeling of the responsiveness (Q4-ChatDirector: M=6.25, SD=0.77; Q4-Video: M=3.81, SD=1.17; Z=-3.54, p<.001). *"The zoom-in effect when I started to talk to someone reminded me of a face-to-face conversation, where I had a direct eye contact with that person. It was really cool to get that feeling on my laptop to help me focus on our discussion." (P4) "In [commercial RVCS], the layout was always unchanged. I could feel my partner didn't notice I was talking to him at sometime. But I felt ChatDirector helps us be more responsive. Not only me, but also my partners." (P1)*

Furthermore, when the participants were not speaking, the *Layout State* helped the participants immediately respond to the other participants (Q2-ChatDirector: M=6.06, SD=0.85; Q2-Video: AVG: 3.13, SD=0.81; Z=-3.46, p<.01). *"Before I realized someone was talking to me, the system already helped me focus on that person. [Commercial RVCS] could only let me know who was speaking, but would never let me know who was speaking to me." (P1)* The combination of the layout transition and the animations of the remote avatars enabled the participants to shift the attention to the right conversations on time (Q3-ChatDirector: M=6.00, SD=0.63; Q3-Video: AVG: 3.19, SD=1.11; Z=-3.54, p<.001). *"I liked the [Pairwise] the most. It gave me a very realistic feeling just like they were sitting there to do the discussion in front of me." (P8) "I play conversational games a lot on Zoom with my friends. It's always a big problem for me to extract useful information when they start to have intense conversations. I could definitely imagine how ChatDirector helps improve that situation." (P16)* Furthermore, we observed that some participants did not contribute much during the open discussion, but still found that they appreciated the system features. *"I didn't know the others well, but I still felt quite interesting when I could see two people debating against each other in those two tiles. I enjoyed it just like watching a TV show." (P2) "I gave this system a higher rating than commercial systems. The dynamic transition gave this meeting more energy. Everybody was like standing in front of me and walk around when they talk." (P8)* Using ChatDirector, the feeling of engagement was significantly better than using traditional grid-layout 2D RVCS (Q5-ChatDirector: M=6.00, SD=0.82; Q5-Video: M=2.19, SD=0.83; Z=-3.56, p<.001). *"ChatDirector provided me with more energy. I think if I used this system to take virtual classes, I would like to raise more discussions with the instructor." (P11) "I felt like being driven by*

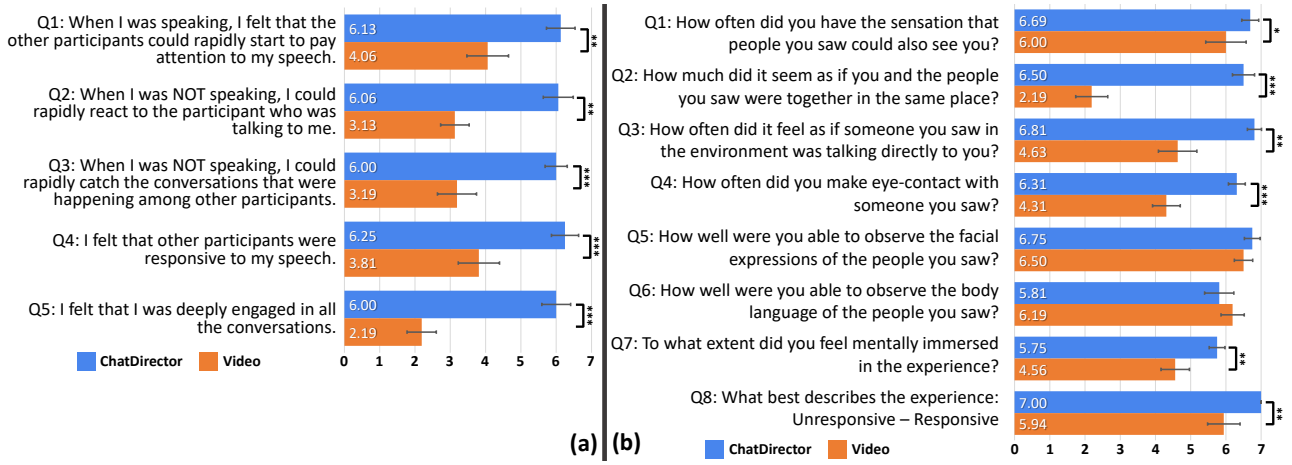

Figure 7: (a) The comparison results between ChatDirector and commercial RVCS in terms of the attention transition experience. The participants agreed that the layout and avatar adjustments driven by the layout transition algorithm could help keep concentrated on the ongoing conversations as well as improve the overall engagement. (b) The results of the TPI. Overall, ChatDirector received significantly higher feedback on the co-presence experience by immersing remote participants in the virtual meeting scene with 3D portrait avatar representations and spatial-sensitive layout and avatar adjustments. ($* : p < .05, ** : p < .01, *** : p < .001$)

an invisible camera man, leading me into a story, which was super engaging to me." (P10)

6.3.2 Spatial Presence and Overall User Experience. We further identified the needs of combining the visual assistance [30, 31] and the spatial awareness enabled by 3D virtual environment rendering [39, 58, 85] so that participants have a natural feeling of co-presence, and the attention transition can be delivered to end-users in a non-obtrusive manner. The space-aware rendering pipeline and the 3D shared virtual meeting environment are then designed following DC1 and DC4. In order to evaluate the presence-oriented experience of ChatDirector, we pulled the questions that were related to RVCS from the Temple Presence Inventory (TPI) [45], which were designed to qualitatively measure the media experience in spatial and social presence. The results are shown in Figure 7b.

In terms of the visual fidelity of the 3D portrait avatar, as shown in Q5 and Q6, the participants could clearly observe both the facial expression and body language of remote participants without any significant difference when compared with commercial RVCS (Q5-ChatDirector: M=6.75, SD=0.45; Q5-Video: M=6.50, SD=0.52; Z=-1.27, p=.206; Q6-ChatDirector: M=5.81, SD=0.83; Q6-Video: M=6.19, SD=0.66; Z=-1.39, p=.165). These comparative results indicated the feasibility of our avatar reconstruction technique.

We built a 3D virtual meeting scene that provided the participants with a significantly higher feeling of social presence. When using our system, the feeling of sitting together with each other was much higher than commercial RVCS (Q2-ChatDirector: M=6.50, SD=0.63; Q2-Video: M=2.19, SD=0.91; Z=-3.55, p<.001). *"The most dominant reason that I would use ChatDirector is the feeling of being together with my friends. This reminds me of the virtual background feature we always use in [commercial RVCS]. People choose different background, which hugely reduced the feeling of being together." (P10)* Meanwhile, such co-presence together with the attention transition enhanced the mutual speech awareness among the participants (Q1-ChatDirector: M=6.69, SD=0.48; Q1-Video: M=6.00, SD=1.15; Z=-2.37, p<.05). *"When I used [commercial RVCS], I was used to confirming with my partners that they heard my speech. But when I used ChatDirector, the dynamic visual feedback gave me a higher confidence because I knew they also had the same layout transition features." (P3)*

Similarly, the participants were more clear about that someone was talking to them with the help of the layout transition (Q3-ChatDirector: M=6.81, SD=0.40; Q3-Video: M=4.63, SD=1.09; Z=-3.43, p<.01). *"I'm a TA, and I would like to use this system instead of [commercial RVCS] when I do virtual classes because the [One-On-One] layout really help me remember who actively interacts with me." (P11)* Moreover, the eye contact was successfully preserved with the help of our system (Q4-ChatDirector: M=6.31, SD=0.48; Q4-Video: M=4.31, SD=0.79; Z=-3.58, p<.001). *"I really like the rotation of the avatar. On [commercial RVCS], it's super difficult for me to recognize who is talking to whom. But now, I can even feel that they are having some direct eye contacts in that [Pairwise] layout." (P5)* Enabled by the automatic transition of the layout and avatars, the participants felt that ChatDirector was much more responsive than commercial RVCS (Q8-ChatDirector: M=6.38, SD=0.50; Q8-Video: M=5.94, SD=0.93; Z=-2.97, p<.01). *"[Commercial RVCS] is too stable. But with those dynamic assistance of ChatDirector, I feel like every time when I need an assistance, the system is responsive to my need." (P10)*

To sum up, with all the features provided by ChatDirector, the participants felt much more engaged in the meeting (Q7-ChatDirector: M=5.75, SD=0.45; Q7-Video: M=4.56, SD=0.81; Z=-3.44, p<.01). *"It was a lot of fun to use ChatDirector. In the past when I used [commercial RVCS] to do group discussions, I felt bored when others have*

conversations. But now, I feel like ChatDirector is trying to push me to join those conversations." (P9)

6.4 Discussion

In this section, we discuss the study results, to provide insights and opportunities for current system improvement and future RVCS research.

6.4.1 Unexpected behavior of the layout transition algorithm. In the post-study interview, some participants raised concerns about the unexpected behavior of the attention transition assistance. *"When I listened to others' discussion about what I said, sometimes the layout jumps between [Full-View] and [Pairwise]." (P13).* Although we added a time threshold to avoid frequent transitions between sequential *Layout State*s, the current algorithm does not understand the semantics of the conversations. In some scenarios, the local participant may not care about the detailed turn-taking and handovers. Instead, the participant expects to enjoy the discussion from a high-level perspective. Hence, we believe that when designing future speech-aware assistance systems, there is potential benefit in interpreting semantics from different levels of detail.

While the accuracy of the *Avatar State* output and the corresponding qualitative results were generally satisfactory, the unexpected errors mostly came from the cases in *Full-View Layout State* where the two remote participants were not next to each other. *"I reported an [Avatar State] error when I thought [the left-most participant] turned to the avatar next to him. But latter, I realized I was wrong. [The left-most participant] was talking to [the right-most participant]." (P2)* On one hand, such spatial ambiguity could often be resolved as the continued conversation provided more context. Meanwhile, combining speech with visual cues, such as shrinking or moving unrelated participants and adding visualizations of *Talk-To Speech State*s, may be promising directions for future RVCS design.

6.4.2 Impact of individual differences on system ratings. In Figure 7a, we noticed that the baseline video-based RVCS received much lower scores than ChatDirector, especially, in the engagement (Q5) and responsiveness (Q2, Q3) questions. One reason was because the participants we recruited did not know each other. When some participants used the commercial RVCS that they have been quite familiar with, they were not impressed and did not show high enthusiasm. Hence, we observed that some participants did not talk much during the tasks, and some participants didn't react timely as the commercial RVCS did not provide hints for attention transition. *"Well, when I used [commercial RVCS] in the first session, I really didn't get the point of this study. I didn't see any interesting point there. After I tried ChatDirector, I realized the difference there. Honestly speaking, ChatDirector was new to me, and I really enjoyed trying out new things. It was pretty cool! (P13)"* This feedback suggests a potential additional benefit that we did not consider during the design process. ChatDirector may facilitate ice-breaking scenarios in helping with inclusion and facilitating connections in a social setting. By dynamically adjusting the layouts, ChatDirector has the potential to act as a director or host, facilitating each participant's participation. Furthermore, while we received positive feedback regarding the *Pairwise* layout from the participants who did not tend to express much in group conversations, people who are more active may want to be considerate to those quiet participants. Therefore, we are motivated to conduct a larger-scale user study to investigate system performance when people with different personalities (*e.g.*, extroverted vs. introverted) use our system. We could also track longitudinal user feedback as they become more familiar with the system.

6.4.3 Effects of the types of virtual meetings. In our user study, we designed two tasks with two topics: a casual one and a formal one. As a system-oriented work, the study of ChatDirector was mainly designed to verify the reliability of the novel technical features and overall usability. However, both tasks included complex speech-turns and announcements, representing common application scenarios within the scope of this paper. The study findings align with the prior works that require complicated hardware setups [63, 85]. Hence, we believe that the features of ChatDirector fit well within the current ecosystem of conversation-oriented RVCS. This paper suggests that participants would welcome spatial awareness on 2D screens, provided the system properly integrates 3D-driven features, (*e.g.*, layout and avatar transitions). In addition to the technical aspects, we also received user feedback related to the conversation topics. *"I used ChatDirector to do the guess-the-word-game, and it was super fun, I really enjoyed looking at people's faces with the zoom-in effect because it made me feel like we were laughing together. (P12)" "I really liked the [Pairwise] design when I did the debate. It was exactly what I expected when looking at two people having intense back-and-forth chats. (P9)"* We realized that the participants showed slightly different preferences for the system features under different conversation topics, which motivates us to conduct further user studies across diverse virtual meeting scenarios such as the ones discussed in Application Scenarios.

7 LIMITATIONS AND FUTURE WORK

The satisfactory accuracy of the attention transition algorithm, coupled with positive feedback on the speech fluency and spatial co-presence suggests promising potential for improved usability with ChatDirector. In this section, we further discuss the issues we observed and which were raised by study participants, and suggest potential solutions and avenues for improvement.

Avatar representation. Using an accelerated portrait depth prediction neural network together with a mesh rendering approach, we enable a real-time reconstruction of a participant's upper-body using a single RGB video. However, due to the limited field-of-view of the camera, the side and back of the participant remain unaccounted for. Most participants felt that the current visualization provided a clear visual hint for attention transition, but would prefer if the visual artifacts were addressed. One potential solution could involve asking users to take photos of their faces from multiple angles and utilizing 3D object reconstruction [9, 82] and rendering techniques [49, 64, 74] to complete the missing side mesh. Alternatively, real-time facial expressions of the local user could be mapped onto a given 3D head mesh model [8, 20, 34]. Yet, it still needs extensive technical validation to prove the feasibility of implementing these state-of-the-art modules in real-time using a single RGB camera.

Inputs of the layout transition algorithm. We received positive feedback on the attention transition assistance. However, unexpected layout behaviors occurred during certain scenarios. One of the primary reasons is the limitation of the algorithm's input. Specifically, semantic-level information may play a crucial role in complex discussion scenarios. For instance, in group conversations, a summary of a series of opinion exchanges may describe the ongoing semantics more precisely. Additionally, emotion may also be revealed from the speech and influence the user attention transition. By leveraging Natural Language Processing (NLP) and Large Language Model (LLM) techniques, such as dialogue summarization and emotion recognition [23, 62, 76], we envision incorporating semantic perception to expand the definition of the *Speech State* and expand the capability of the speech-driven algorithm in future work.

Large-scale meeting scenarios. We considered the number of *Speech State*s as a critical factor in the decision tree algorithm, which allows the system to handle scenarios with more participants. Additionally, following commercial RVCS, we avoided visualizing too much information simultaneously (*e.g.*, using a grid-like *Pairwise* layout when more than two *Pairs* exist) to reduce mental load. However, we could leverage the spatial awareness enabled by the system for visualizations of large-scale remote meetings. One straightforward add-on feature follows the 'pin a user' idea in commercial RVCS. We could allow the user to select which subset of participants to visualize as 3D avatars for very large meetings. The concept of the break-out rooms could be another improvement. During the user study, P11 mentioned: *"One application scenario I could imagine is an online group discussion with many students. I, as a TA, would like to join different groups to check their progress."* By leveraging the 3D meeting scene and the attention transition assistance, we would be interested in future work to place spatial anchors as groups in the entire virtual environment for the hidden remote avatars, and enable either the local user or the algorithm to translate the rendering camera to focus on different user groups.

Automation *vs.* customization. The user study showed that the automatic attention transition was effective in improving the remote meeting experience. Most participants recognized its effectiveness with the design of *Layout State* and *Avatar State*. P5 suggested a human-in-the-loop approach: *"I was wondering could the system use my feedback to improve the transition effects?"* P4 raised a concern that *"What if I want to always show my mom's avatar in our family chats?"* How to balance between automation and customization is always a non-trivial issue. We believe that potential improvements could involve allowing users to manually toggle on/off specific features, providing real-time feedback to fine-tune the algorithm to suit their preferences, and incorporating unsupervised approaches such as rule-based machine learning and regressions [1]. Inspired by commercial RVCS [24, 53, 86] that allow users to actively pin specific users and adjust the grid layout, we envision future spatialized RVCS to provide 3D anchors to allow users to manually place important participants in the virtual scenes or split views of groups to address the customization needs and concerns.

Integration with more meeting elements. In this paper, we limit our research scope to conversation-oriented remote conferencing scenarios and propose ChatDirector to address speech-sensitive issues such as loss of attention and speech interruptions [11, 25, 52]. As mentioned in the beginning of the paper, digital assistance enabled by commercial RVCS (*e.g.*, presentation sharing) has been adopted in many virtual meeting cases. Since our system enables a 3D shared meeting environment, exploiting the advantages of spatial awareness becomes an attractive development direction. This includes integrating meeting elements commonly used in commercial RVCS into our system. Examples include placing chats and relevant visuals [43], physical objects [33, 67], live captions [44, 55], and shared screens next to corresponding users for intuitive spatial reference, popping up emojis and raising hand icons above users to attract the presenter's attention, and enabling private chats with spatially-aware audio. However, given the limited size of the 2D screen, further research and study are required to identify the most practical designs for such integration.

Integration with extended reality. Extended Reality (XR) has witnessed a rapid growth recently. It has also been leveraged in remote conferencing [26, 29, 57], and social media platforms [12]. Meta Horizon Workrooms [32] and Spatial [72] allow users join a virtual shared environment by wearing XR headsets. In this paper, we target the more commonly used computing devices (*e.g.*, laptop) as they have a higher accessibility and flexibility to be integrated with other office tools. Meanwhile, XR-based works adopt either profile photo or cartoon avatars as users' visual representations. Yet, in many formal scenarios such as product pitches, debates, and press conferences, a high-fidelity facial presence would be required. Pixel codec avatars [46] and Apple Vision Pro [2] have shown both research and commercial exploration in enabling real-time photorealistic avatar driving while end-users wear XR headsets. Following this trend, we envision the integration between our system and XR-based conferencing systems so that cross-platform users can join the same virtual meeting with high-fidelity self representation.

8 CONCLUSION

In this paper, we introduce ChatDirector, a novel RVCS designed to make video meetings more engaging and to enhance the feeling of co-presence through speech-driven visual cues in a space-aware scene. We introduce a real-time technical pipeline that makes it possible to render remote participants as 3D portrait avatars in a shared 3D virtual environment without requiring specialized equipment. We contribute a speech-aware algorithm that dynamically adjusts the on-screen arrangement of remote participants and their behaviors, offering visual assistance to aid attention tracking and shifting. We further demonstrate various application scenarios, highlighting the potential for ChatDirector to enhance common remote conferencing experiences, such as brainstorms, debates, games, and office hours. We share our findings from a comparative user study with a baseline RCVS, which suggests that ChatDirector provides a more engaging remote conferencing experience. Our study participants provided valuable insights about increased co-presence from the space-aware virtual scene and how the dynamic scene transitions encourage engagement, but also pointed to important future work for improving avatar representation and integrating semantic perception. We hope that ChatDirector will inspire continued work on

everyday computing platforms that leverage state-of-the-art perception and interaction techniques to increase the sense of co-presence and engagement.

ACKNOWLEDGMENTS

We wish to express our gratitude to Zhengzhe Zhu, Liyun Xia, Xiangyu Qu, Rahul Jain, Jingyu Shi, Fengming He, Moiz Rasheed, and Dizhi Ma for their assistance in recording the footage for videos and figures. Our appreciation also extends to Eric Turner for his constructive suggestions regarding the manuscript. Furthermore, we are grateful to our anonymous reviewers for their perceptive and valuable feedback.

REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. 207–216. https://doi.org/10.1145/170036.170072

[2] Apple Vision Pro 2024. Set Up Your Persona (Beta) on Apple Vision Pro. https://support.apple.com/en-us/HT214002.

[3] Knut Magne Augestad and Rolv Ole Lindsetmo. 2009. Overcoming Distance: Video-Conferencing as a Clinical and Educational Tool Among Surgeons. *World Journal of Surgery* 33 (2009), 1356–1365. https://doi.org/10.1007/s00268-009-0036-0

[4] Maral Babapour Chafi, Annemarie Hultberg, and Nina Bozic Yams. 2022. Post-Pandemic Office Work: Perceived Challenges and Opportunities for a Sustainable Work Environment. *Sustainability* 14, 1 (2022), 294. https://doi.org/10.3390/su14010294

[5] Stephan Beck, Andre Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive Group-to-group Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625. https://doi.org/10.1109/TVCG.2013.33

[6] Body Segmentation 2023. Body Segmentation with MediaPipe and Tensorflow.js. https://blog.tensorflow.org/2022/01/body-segmentation.html.

[7] Erin Bradner and Gloria Mark. 2001. Social Presence with Video and Application Sharing. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*. 154–161. https://doi.org/10.1145/500286.500310

[8] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars From a Phone Scan. *ACM Transactions on Graphics* (Jul. 2022). https://doi.org/10.1145/3528223.3530143

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3D-R2N2: A Unified Approach for Single and Multi-View 3D Object Reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 628–644.

[10] Anthony DeVincenzi, Lining Yao, Hiroshi Ishii, and Ramesh Raskar. 2011. Kinected Conference: Augmenting Video Imaging with Calibrated Depth and Audio. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. 621–624.

[11] Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'malley, Steve Langton, Simon Garrod, and Vicki Bruce. 1997. Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance. *Journal of Experimental Psychology: Applied* 3, 2 (1997), 105. https://doi.org/10.1037/1076-898X.3.2.105

[12] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI, 685)*. ACM, 13. https://doi.org/10.1145/3290605.3300915

[13] Ruofei Du, Na Li, Jing Jin, Michelle Carney, Scott Miles, Maria Kleiner, Xiuxiu Yuan, Yinda Zhang, Anuva Kulkarni, Xingyu Liu, et al. 2023. Rapsai: Accelerating Machine Learning Prototyping of Multimedia Applications through Visual Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23. https://doi.org/10.1145/3544548.3581338

[14] Carmen Egido. 1988. Video Conferencing as a Technology to Support Group Work: A Review of Its Failures. In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*. 13–24. https://doi.org/10.1145/62266.62268

[15] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image Using a Multi-scale Deep Network. *Advances in Neural Information Processing Systems* 27 (2014). https://doi.org/10.48550/arXiv.1406.2283

[16] Face Detection 2023. Face Detection Task Guide, MediaPipe. https://developers.google.com/mediapipe/solutions/vision/face_detector.

[17] Fact or Fiction 2024. Fact or Fiction: A Fun Icebreaker Game to Get to Know Each Other. https://www.icebreakerspot.com/activities/fact-or-fiction.

[18] Francisco Figueroa, David Figueroa, Rafael Calvo-Mena, Felipe Narvaez, Natalia Medina, and Juan Prieto. 2020. Orthopedic Surgery Residents' Perception of Online Education in Their Programs During the COVID-19 Pandemic: Should It Be Maintained After the Crisis? *Acta Orthopaedica* 91, 5 (2020), 543–546.

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

[20] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8649–8658.

[21] Jim Gemmell, Kentaro Toyama, C Lawrence Zitnick, Thomas Kang, and Steven Seitz. 2000. Gaze Awareness for Video-Conferencing: A Software Approach. *IEEE MultiMedia* 7, 4 (2000), 26–35. https://doi.org/10.1109/93.895152

[22] Simon J Gibbs, Constantin Arapis, and Christian J Breiteneder. 1999. Teleport–Towards Immersive Copresence. *Multimedia Systems* 7, 3 (1999), 214–221. https://doi.org/10.1007/s005300050123

[23] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. *arXiv preprint arXiv:1911.12237* (2019).

[24] Google Meet 2023. Google Meet. https://meet.google.com.

[25] David M Grayson and Andrew F Monk. 2003. Are You Looking at Me? Eye Contact and Desktop Video Conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 10, 3 (2003), 221–243.

[26] Jens Emil Sloth Grønbæk, Ken Pfeuffer, Eduardo Velloso, Morten Astrup, Melanie Isabel Sønderkær Pedersen, Martin Kjær, Germán Leiva, and Hans Gellersen. 2023. Partially Blended Realities: Aligning Dissimilar Spaces for Distributed Mixed Reality Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3544548.3581515

[27] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. *ACM Transactions on Graphics (ToG)* 38, 6 (2019), 1–19. https://doi.org/10.1145/3355089.3356571

[28] Chris Harrison and Scott E Hudson. 2008. Pseudo-3D Video Conferencing With a Generic Webcam. In *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, 236–241. https://doi.org/10.1109/ISM.2008.12

[29] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. CollaboVR: A Reconfigurable Framework for Multi-user to Communicate in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 542–554. https://doi.org/10.1109/ISMAR50242.2020.00082

[30] Zhenyi He, Ruofei Du, and Ken Perlin. 2021. LookAtChat: Visualizing Gaze Awareness for Remote Small-Group Conversations. *arXiv preprint arXiv:2107.06265* (2021). https://doi.org/10.48550/arXiv.2107.06265

[31] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 769–782. https://doi.org/10.1145/3472749.3474785

[32] Horizon Workrooms 2023. Meta Horizon Workrooms Virtual Office and Meetings. https://www.meta.com/work/workrooms.

[33] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. 2023. ThingShare: Ad-hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM. https://doi.org/10.1145/3544548.3581148

[34] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-Held Video Input. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–14.

[35] Tomoo Inoue, Ken-ichi Okada, and Yutaka Matsushita. 1997. Integration of Face-to-Face and Video-Mediated Meetings: HERMES. In *Proceedings of the 1997 ACM International Conference on Supporting Group Work*. 405–414.

[36] Tracy Jenkin, Jesse McGeachie, David Fono, and Roel Vertegaal. 2005. eyeView: Focus+ Context Views for Large Group Video Conferences. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1497–1500. https://doi.org/10.1145/1056808.1056950

[37] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving Eye Contact in a One-to-many 3D Video Teleconferencing System. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–8.

[38] Rene Kaiser, Wolfgang Weiss, Manolis Falelakis, Spiros Michalakopoulos, and Marian F Ursu. 2012. A Rule-based Virtual Director Enhancing Group Communication. In *2012 IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 187–192. https://doi.org/10.1109/ICMEW.2012.39

[39] Peter Kauff and Oliver Schreer. 2002. An Immersive 3D Video-conferencing System Using Shared Virtual Team User Environments. In *Proceedings of the 4th International Conference on Collaborative Virtual Environments*. 105–112. https://doi.org/10.1145/571878.571895

[40] Md Saifuddin Khalid and Md Iqbal Hossan. 2016. Usability Evaluation of a Video Conferencing System in a University's Classroom. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 184–190.

[41] Chandra Bhushan Kumar, Anjali Potnis, and Shefali Gupta. 2015. Video Conferencing System For Distance Education. In *2015 IEEE UP Section Conference on Electrical Computer and Electronics (UPCON)*. IEEE, 1–6.

[42] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational Video Editing for Dialogue-driven Scenes. *ACM Transactions on Graphics* 36, 4 (2017), 130–1. https://doi.org/10.1145/3072959.3073653

[43] Xingyu Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–20. https://doi.org/10.1145/3544548.3581566

[44] Xingyu Liu, Jun Zhang, Leonardo Ferrer, Susan Xu, Vikas Bahirwani, Boris Smus, Alex Olwal, and Ruofei Du. 2023. Modeling and Improving Text Stability in Live Captions. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA, 208)*. ACM, 1–9. https://doi.org/10.1145/3544549.3585609

[45] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring Presence: The Temple Presence Inventory. In *Proceedings of the 12th Annual International Workshop on Presence*. 1–15.

[46] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 64–73. https://doi.org/10.48550/arXiv.2104.04638

[47] Andrew Maimone and Henry Fuchs. 2011. Encumbrance-free Telepresence System with Real-time 3D Capture and Display Using Commodity Depth Cameras. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 137–146. https://doi.org/10.1109/ISMAR.2011.6092379

[48] Microsoft Teams 2023. Video Conferencing, Meetings, Calling, Microsoft Teams. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software.

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[50] David Nguyen and John Canny. 2005. Multiview: Spatially Faithful Group Video Conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 799–808. https://doi.org/10.1145/1054972.1055084

[51] NVIDIA V100 2023. NVIDIA V100. https://www.nvidia.com/en-us/data-center/v100.

[52] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human-Computer Interaction* 8, 4 (1993), 389–428. https://doi.org/10.1207/s15327051hci0804_4

[53] ohyay 2023. ohyay, the Best Virtual Events. https://ohyay.co.

[54] Ken-Ichi Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. 1994. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. 385–393. https://doi.org/10.1145/192844.193054

[55] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1108–1120.

[56] Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce. 1996. Comparison of Face-to-Face and Video-Mediated Interaction. *Interacting with Computers* 8, 2 (1996), 177–192. https://doi.org/10.1016/0953-5438(96)01027-2

[57] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 741–754. https://doi.org/10.1145/2984511.2984517

[58] Kazuhiro Otsuka. 2016. MMSpace: Kinetically-augmented Telepresence for Small Group-to-group Conversations. In *2016 IEEE Virtual Reality (VR)*. IEEE, 19–28. https://doi.org/10.1109/VR.2016.7504684

[59] Ye Pan and Anthony Steed. 2014. A Gaze-preserving Situated Multiview Telepresence System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2173–2176. https://doi.org/10.1145/2556288.2557320

[60] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–21. https://doi.org/10.1145/3450626.3459872

[61] Niki Panteli and Patrick Dawson. 2001. Video Conferencing Meetings: Changing Patterns of Business Communication. *New Technology, Work and Employment* 16, 2 (2001), 88–99.

[62] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access* 7 (2019), 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

[63] Project Starline 2023. Project Starline: Feel Like You Are There, Together. https://blog.google/technology/research/project-starline.

[64] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic Head Avatars With Rigged 3D Gaussians. *arXiv preprint arXiv:2312.02069* (2023). https://doi.org/10.48550/arXiv.2312.02069

[65] Abhishek Ranjan, Jeremy Birnholtz, and Ravin Balakrishnan. 2008. Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 227–236. https://doi.org/10.1145/1357054.1357095

[66] Shaker Sabri and Birendra Prasada. 1985. Video Conferencing Systems. *Proc. IEEE* 73, 4 (1985), 671–688. https://doi.org/10.1109/PROC.1985.13192

[67] Mose Sakashita, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View Dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–14. https://doi.org/10.1145/3613904.3642945

[68] Abigail Sellen, Bill Buxton, and John Arnott. 1992. Using Spatial Cues to Improve Videoconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 651–652. https://doi.org/10.1145/142750.143070

[69] Abigail J Sellen. 1995. Remote Conversations: The Effects of Mediating Talk with Technology. *Human-Computer Interaction* 10, 4 (1995), 401–444. https://doi.org/10.1207/s15327051hci1004_2

[70] Siegel Sidney. 1957. Nonparametric Statistics for the Behavioral Sciences. *The Journal of Nervous and Mental Disease* 125, 3 (1957), 497.

[71] socket.io 2023. Socket.IO. https://socket.io.

[72] Spatial 2023. Spatial Create Share and Experience Your Creativity in 3D. https://www.spatial.io.

[73] Md Tahsin Tausif, RJ Weaver, and Sang Won Lee. 2020. Towards Enabling Eye Contact and Perspective Control in Video Conference. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 96–98. https://doi.org/10.1145/3290605.3300530

[74] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2387–2395. https://doi.org/10.1145/3292039

[75] Tuul Triyason, Anuchart Tassanaviboon, and Prasert Kanthamanon. 2020. Hybrid Classroom: Designing for the New Normal After COVID-19 Pandemic. In *Proceedings of the 11th International Conference on Advances in Information Technology*. 1–8. https://doi.org/10.1145/3406601.3406635

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[77] Roel Vertegaal. 1999. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 294–301. https://doi.org/10.1145/302979.303065

[78] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. Gaze-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Directionn. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 521–528. https://doi.org/10.1145/642611.642702

[79] Web Speech API 2023. Web Speech API. https://wicg.github.io/speech-api.

[80] WebRTC 2023. WebRTC. https://webrtc.org.

[81] Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*. Springer.

[82] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. 2016. Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction Without 3D Supervision. *Advances in Neural Information Processing Systems* 29 (2016). https://doi.org/10.48550/arXiv.1612.00814

[83] Lining Yao, Anthony DeVincenzi, Anna Pereira, and Hiroshi Ishii. 2013. FocalSpace: Multimodal Activity Tracking, Synthetic Blur, and Adaptive Presentation for Video Conferencing. In *Proceedings of the 1st Symposium on Spatial User Interaction*. 73–76. https://doi.org/10.1145/2491367.2491377

[84] Cha Zhang, Qin Cai, Philip A Chou, Zhengyou Zhang, and Ricardo Martin-Brualla. 2013. Viewport: A Distributed, Immersive Teleconferencing System with Infrared Dot Pattern. *IEEE MultiMedia* 20, 1 (2013), 17–27. https://doi.org/10.1109/MMUL.2013.12

[85] Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong, and Baining Guo. 2022. VirtualCube: An Immersive 3D Video Communication System. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2146–2156. https://doi.org/10.1109/TVCG.2022.3150512

[86] Zoom 2023. One Platform to Connect, Zoom. https://zoom.us.