# Since U Been Gone: Augmenting Transcriptions for Re-engaging in Immersive VR Meetings

ANONYMOUS AUTHOR(S)*

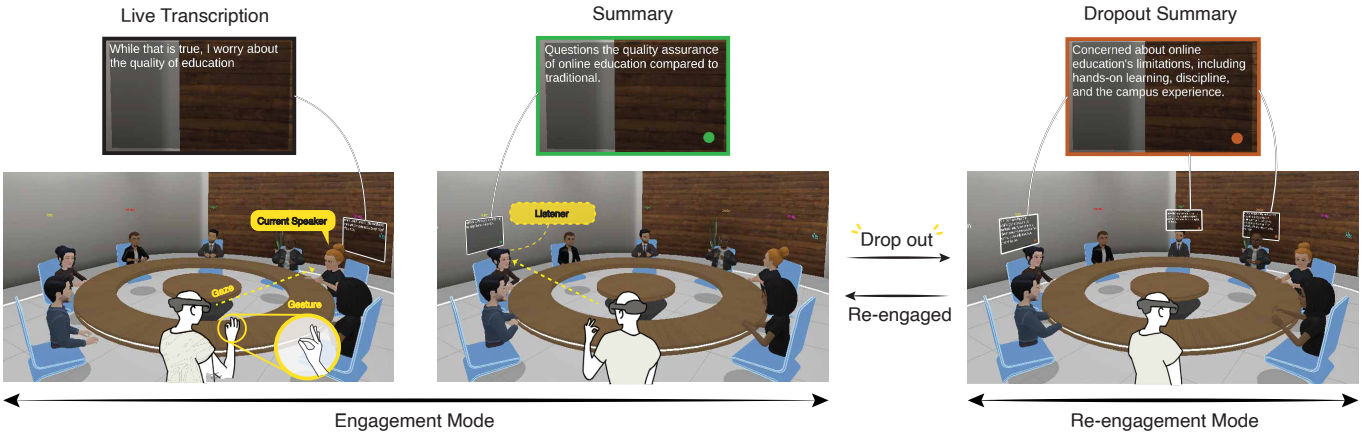


Fig. 1. **EngageSync** is a context-aware transcription panel designed to help users in immersive VR meetings catch up on conversations while maintaining social presence. It operates in two modes: **Engagement Mode** and **Re-engagement Mode**. In Engagement Mode, users can view live transcriptions of the speaker or a summary of the previous utterance of a listener by gazing at the person of interest and performing a pinch gesture. In Re-engagement Mode, summaries of conversations that occurred during the user's absence are displayed for all relevant participants. Once all summaries are read, the user is considered "re-engaged" and caught up with missed context, and the system returns to Engagement Mode.

Maintaining engagement in immersive meetings is challenging, particularly when users must catch up on missed content after disruptions. Traditional transcription interfaces, like table-fixed panels, have the potential to distract users from the group, diminishing social presence, while avatar-fixed captions fail to provide past context. We developed EngageSync, a context-aware avatar-fixed transcription interface that adapts based on user engagement, offering live transcriptions and summaries to enhance catching up while preserving social presence. We implemented a live VR meeting setup for a 12-participant formative study. This guided our design, leading to two user studies with small (3 avatars) and mid-sized (7 avatars) groups, our method significantly improved social presence ($p < .05$) and time spent gazing at others in the group instead of the interface over table-fixed panels. Also, it increased information recall ($p < .05$) and faster re-engagement time over avatar-fixed interfaces, with greater improvements in mid-sized groups ($p < .01$).

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**.

Additional Key Words and Phrases: Immersive VR Meeting; Social VR; Virtual Reality; Teleconferencing; Co-presence; Re-engagement; Group Conversations

**ACM Reference Format:**
Anonymous Author(s). 20XX. Since U Been Gone: Augmenting Transcriptions for Re-engaging in Immersive VR Meetings. In *Proceedings of the XX, Nov 26–Dec 1, 20XX, XX, XX*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*XX '2x, Nov 26–Dec 1, XX, XX*


## 1 INTRODUCTION

Remote work and virtual collaboration have become integral to how people communicate and collaborate, with video-mediated meetings rising significantly in the wake of the COVID-19 pandemic [25, 30]. As the demand for seamless remote interaction continues to grow, researchers and developers have turned their attention not only to video conferencing platforms but also to more immersive technologies like Virtual Reality (VR) and Mixed Reality (MR) [32]. These technologies offer new opportunities to reimagine how we engage in virtual environments, providing enhanced social interaction and presence that video conferencing alone cannot achieve. [31, 32, 39]

While remote meetings offer significant convenience, they also make users more susceptible to distractions, both internal (e.g., multitasking or loss of focus) and external (e.g., notifications, people entering the room) [8]. Research in remote video meeting interfaces has made strides in keeping users engaged through meeting summaries, post-meeting transcripts [46], and more recently, live interactive transcription tools designed to maintain participation during discussions [9, 48]. These tools not only help users stay on track but also enable them to catch up after disruptions, which can occur frequently in remote work settings.

Although immersive VR platforms promise heightened focus and social presence through their immersive nature, users in VR environments are not immune to external distractions. Interruptions such as device notifications, someone physically entering the room, or technical difficulties can pull users out of their virtual space. Previous work has attempted to address these challenges in VR by mitigating disruptions during experiences [12, 14]. However, when

it comes to applying interactive transcription interfaces, proven effective in video meetings, the question arises whether we can directly use them without any consideration to adapt them to the more spatial and immersive nature of VR meetings.

One possible solution might be to replicate video meeting interfaces by providing a spatially fixed transcription panel in front of the user, allowing them to scroll through transcripts. However, this approach often forces users to look away from other meeting participants, breaking social engagement and reducing the sense of presence in the group. Current social VR platforms, such as VRChat [55], address this issue by attaching captions directly to the speaking avatar, maintaining eye contact and social interaction. Yet, these avatar-fixed captions, do not provide missed content, particularly when users return from a disruption and need to catch up.

This raises the question of how we can bring the advantages of transcription interfaces from video meetings into immersive VR environments while balancing the need for both social presence and effective re-engagement after interruptions.

To address these challenges, we propose **EngageSync**, a novel, context-aware avatar-fixed transcription interface that dynamically adapts to the user's engagement state. EngageSync supports two key use cases: (1) providing real-time transcription and summaries during active participation, and (2) offering context-aware summaries to support re-engagement after disruptions. By adapting based on whether users are engaged or re-engaging, EngageSync maintains social presence while ensuring efficient catch-up on missed content.

We evaluate EngageSync through two user studies, comparing it to existing interfaces such as table-fixed and always-on avatar-fixed transcription panels. Our evaluation investigates its effectiveness in enhancing social presence and information recall across different group sizes and scenarios.

This work makes the following contributions:

- Design insights for context-aware transcription in immersive VR environments: Based on a formative study with 12 participants, we identify key design challenges for supporting re-engagement and enhancing social presence in immersive group discussions.
- EngageSync, a novel adaptive avatar-fixed transcription interface: EngageSync provides live transcriptions for active conversations and context-aware summaries for re-engagement, helping users balance social presence with effective conversation catch-up.
- Comprehensive evaluation across group sizes: Through two user studies, we evaluate EngageSync's performance compared to table-fixed and avatar-fixed interfaces, demonstrating significant improvements in social presence, information recall, and faster re-engagement times ($p < .05$).
- Exploration of group size effects on interface performance: Our results show how EngageSync's context-aware features become more beneficial in larger groups, providing important insights for the design of future immersive meeting platforms.

## 2 RELATED WORKS

### 2.1 Meeting Engagement, Distractions, and Eye-contact

*2.1.1 Meeting Engagement and Participation.* Meeting effectiveness is often linked to participants' engagement and inclusiveness in both physical and virtual settings, as highlighted by Schwartzman [47]. Tools that support inclusivity, like Hosseinkashi et al.'s system for detecting failed speech interruptions, aim to ensure that all voices are heard, especially those who struggle to engage [21].

*2.1.2 Distractions in Remote Meetings and VR.* Maintaining focus is a known challenge in remote meetings, with multitasking occurring in roughly 30% of meetings, significantly reducing engagement [8, 54]. Common distractions include personal obligations, household chores, and non-meeting-related activities like checking emails [8, 25, 30]. Technical issues such as connectivity problems and background noise further exacerbate these distractions [15]. Eye-tracking studies show participants often look away from their screens, particularly in smaller groups [11].

Although VR reduces visual distractions from the physical world, external interruptions like notifications or environmental sounds can disrupt users [12, 14, 45]. While VR environments enhance spatial presence [39], external disruptions become more jarring and break immersion [14]. Unlike diegetic solutions for managing distractions [14], our focus is on helping users re-engage after disruptions.

*2.1.3 Awareness of Attention and Mutual Presence in Remote Communication.* Awareness of others' focus is crucial for fostering trust and collaboration in meetings [20, 24]. In video-mediated communication, the absence of gaze cues reduces social presence, though tools like Gazechat address this by visualizing gaze [18]. In immersive VR, mutual presence is key [28]. While gaze visualization has been used in task-oriented collaborations [4, 41], our approach avoids such disruptions, focusing instead on keeping attention on avatars to maintain engagement and facilitate re-engagement after disruptions.

### 2.2 Interactive Transcription Interfaces for Video Meetings

Maintaining engagement and focus in remote meetings is a well-documented challenge. While much of the attention has been on fostering awareness of participants' attention and mutual presence, another line of research focuses on how interactive transcription interfaces can help users stay engaged with meeting content, particularly in situations where they may become distracted or miss parts of the conversation.

*2.2.1 Transcription Interfaces.* Transcription interfaces have evolved significantly over the years to support participants in engaging with and reflecting on meeting content. Whittaker et al. demonstrated that reading a transcript is often more efficient than rewatching video or audio recordings of meetings, as transcripts allow users to quickly skim and locate relevant information [57]. Moreover, studies have shown that enabling interaction with transcripts, such as marking or highlighting key points, facilitates information recall and imposes a lower cognitive load compared to traditional note-taking methods [23].

Post-meeting transcription tools has evolved, focusing on generating summaries and providing feedback after meetings have ended. For example, Banerjee et al. explored methods for generating summaries post-meeting to help users efficiently digest key takeaways [6]. Moreover, Wang et al.'s MeetingCoach helps users reflect on past meetings by offering personalized feedback on their performance and participation [46].

*2.2.2 Real-time Interactive Transcription Interfaces.* In recent years, the focus has also shed light on real-time interactive transcription tools that keep participants engaged during the meeting itself. For instance, Zhang et al. found that generating summaries of chat-based meetings in real-time helps participants quickly catch up on missed discussions and enhances overall engagement [58]. Tucker et al. developed and evaluated a "Catchup audio player" designed specifically for participants who join meetings late. The system automatically identifies the gist of what was missed, allowing latecomers to quickly catch up and participate effectively without needing to process the full transcript [52].

MeetScript is another example of a real-time transcription interface that allows participants to actively interact with the transcript as the conversation unfolds, marking significant moments and revisiting key sections in real-time [9]. This approach shifts the focus from post-meeting reflection to active participation. Similarly, Iijima et al. introduced interactive text clouds to assist Deaf and Hard of Hearing (DHH) users in staying engaged with live video conferences [22]. Son et al. introduced OPARTs, a meeting interface that allows users to seamlessly switch between live transcription, summary modes, and keyword extracts from each speaker's utterance [48] to support users catching up with missed context due to internal and external distractions.

Although these advancements have been made in video-mediated meetings, adapting interactive transcription and summarization tools to immersive environments like VR presents new challenges and opportunities which we will address more in detail in the section below.

## 2.3 Transcription and Engagement in Immersive VR Meetings

Immersive VR meetings, often viewed as part of the broader metaverse, are emerging as a powerful tool for remote collaboration. These platforms offer opportunities to replicate and enhance traditional office interactions, with features such as spatial audio, natural gestures, and shared virtual spaces designed to improve communication and social presence [32, 39]. As remote and hybrid work models continue to evolve, immersive VR meetings are increasingly seen as a solution for creating engaging, collaborative workspaces that foster both social presence and productivity [50].

Several VR meeting platforms, including Mozilla Hubs [36], Spatial [2], and Meta Horizon Workrooms [34], have gained popularity by providing immersive environments where participants can interact in ways that closely simulate face-to-face meetings. These platforms report increased engagement and interaction due to the immersive nature of VR, but they also present unique challenges that need to be addressed for effective user experiences [3].

*2.3.1 Challenges in Immersive VR Meetings.* Immersive VR environments, while offering many advantages for remote work, come with technical and social challenges. Issues such as hardware limitations, network latency, and avatar fidelity can hinder the overall user experience [39]. Furthermore, users must navigate new social norms and behaviors in these virtual environments, which can impact professional interactions [10].

To overcome these challenges, researchers have proposed various solutions. For example, Qian et al. introduced ChatDirector, which converts RGB video streams into 3D portrait avatars, overcoming the need for VR-specific equipment while enhancing the realism of VR meetings [42]. Similarly, ViGather bridges the gap between traditional devices (laptops, desktops) and VR environments, reconstructing users' poses and conveying non-verbal cues like eye contact to enhance social presence [43]. To overcome the limited field of view of VR headsets, Lee et al., implemented a multi-modal attention guidance system that leverages light and spatial audio so users can notice new speakers in a VR meeting [29].

*2.3.2 Maintaining Engagement through Non-verbal Cues.* Non-verbal cues like eye contact and spatial behavior are crucial for maintaining social presence in VR. Wang et al. [56] found that proxemics and mutual gaze awareness enhance social connection and perceived attention during group interactions.

However, visualizing gaze or non-verbal cues in VR can sometimes distract users and pull them away from the conversation [19, 27]. To avoid this, our approach focuses on promoting natural engagement with avatars without explicitly visualizing gaze, ensuring the flow of conversation remains uninterrupted.

*2.3.3 Transcription, Captions, and Accessibility in VR.* Live transcription and captions have been widely used in video conferencing and Augmented Reality (AR) to improve accessibility and comprehension. Systems like Wearable Captioning [37] provide real-time captions to assist DHH users. In AR, gaze-assisted interfaces such as StARe reveal relevant information progressively during conversations, allowing for smoother interactions [44]. These systems have proven valuable for enhancing accessibility, but they also introduce challenges in immersive settings where prolonged focus on captions can detract from the overall sense of presence [53].

In social VR platforms like VRChat, captions are often used to enhance communication. However, captions can disrupt immersion if they dominate the user's visual attention for too long. Uber et al. found that Automatic speech recognition (ASR) captions improve accessibility but can lead to disengagement from the conversation and reduce social connection if exposure time is prolonged [53]. Thus, there is a need for careful consideration in designing immersive transcription interfaces that maintain a balance between accessibility and engagement.

# 3 FORMATIVE STUDY

In immersive VR meetings, where participants rely heavily on environmental and spatial cues, disruptions such as late arrivals or temporary absences present unique challenges. Traditional video

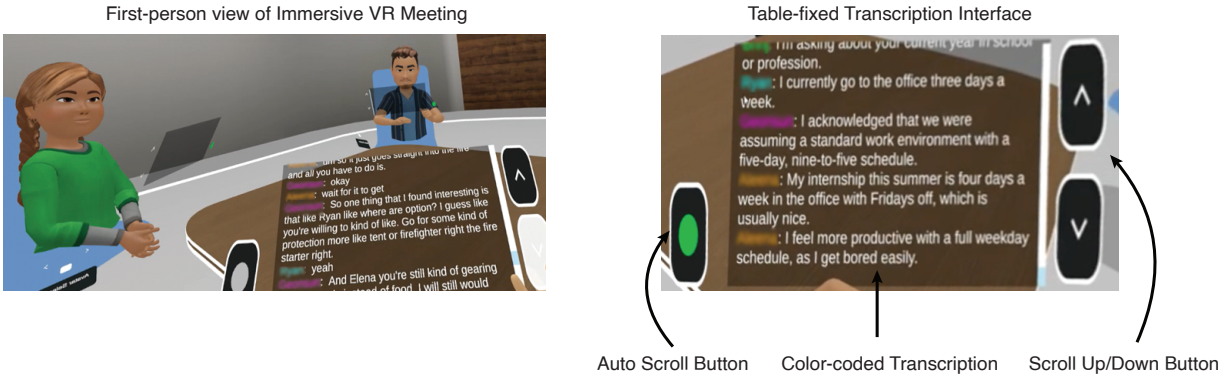


Fig. 2. The formative study setup of a four-people meeting. A screenshot of the immersive meeting environment from a first-person view (Left). The text panel interface was used in the formative study. The interface consists of a text panel where the participants' names are color-coded for readability, an auto-scroll button to the left that follows the newest lines in the panel, and a scroll-up and down button to the right (Right).

meetings offer features like live transcriptions to support multitasking and re-engagement, but it remains unclear how these mechanisms function in the more immersive and cognitively demanding context of VR. To explore this, we conducted a formative study to investigate how users interact with live transcriptions in immersive VR environments and to derive design implications for improving re-engagement support.

Our goal is to understand how participants manage disruptions and whether live transcription tools can effectively aid in *re-engagement* within VR meetings. Specifically, we define re-engagement as catching up on missed content and being ready to engage in the current conversation. We examine how users achieve this reintegration after disruptions.

Participants in the study wore Oculus Quest Pro HMDs and engaged in a multi-user networked meeting setup, where four users were seated in separate rooms. To simulate common disruptions, participants were asked to remove their VR headsets and rejoin the meeting after a set interval (see Section 3.3 for procedural details). This allowed us to observe how participants used transcription tools to regain focus and re-enter the conversation.

The study focused on two key scenarios: i) participants' reliance on live transcriptions during discussions, and ii) their ability to catch up on missed content after a period of absence. We compared two transcription interfaces: a full transcription mode, where all speech was transcribed, and a summarized version, where each utterance—defined as non-stop sentences spoken by a single user—was condensed to 10 words. This builds on prior research emphasizing the need for concise transcriptions for quick comprehension in meetings [48], with a specific focus on investigating the use cases of each transcription mode in our drop-out and rejoin scenario. By alternating between "dropping out" and rejoining, we evaluated how effectively each transcription method supported participants' re-engagement with the conversation

## 3.1 Implementation of Immersive VR Meeting Setup

In this section, we detail the setup of our immersive VR meeting environment, focusing on the transcription interface and multi-user networking. Participants, represented as avatars, communicated in real-time and were assisted by live transcriptions and summaries displayed on a table-fixed transcription interface. We implemented a VR meeting room environment similar to Meta Horizon Workrooms, with participants seated around a virtual table, using Oculus Quest Pro HMDs.

*3.1.1 Immersive Meeting Room and Transcription Interface Setup.* We used Unity 3D Engine 2022.3.8f1 to create a virtual meeting room modeled after Meta Horizon Workrooms, aligning the virtual environment with the physical experimental room setup [34]. Meta's Avatar SDK [33] enabled upper-body tracking, lip-sync animation, and eye-blink synthesis using Oculus Quest Pro's eye-tracking data for natural avatar interactions.

Participants could customize their avatars through a mirrored avatar selection feature, with the option to hide the mirror for reduced distractions.

The transcription interface, fixed to the table in front of each participant, mirrored the UI panel setup in Meta Horizon Workrooms' remote desktop feature [35]. It displayed either full transcripts or summaries, with color-coded participant names for clarity. Users navigated using up-and-down buttons or an auto-scroll option, with hand interactions enabled through collision detection for smooth navigation.

*3.1.2 Speech Transcription and Summarization.* We transcribed participants' speech using the Google Speech-to-Text (STT) API [13] to achieve high transcription accuracy and a low word error rate. Since no pre-existing Unity package was available for long-duration (>20 min) live transcription, we developed a custom Unity plugin compatible with both Windows and Android platforms. The plugin captured 48 kHz audio from VR HMD microphones, split it into segments based on pauses between words, and asynchronously processed the speech data via the API. The average word error rate from the three participant groups is 15%.

Once transcriptions were generated, we used OpenAI GPT-4-Turbo [38] to summarize each utterance within a 10-word limit. The model's large context window (128k) ensured that it could maintain conversational context, improving summary coherence and accuracy. We evaluated the accuracy of the summarization by comparing

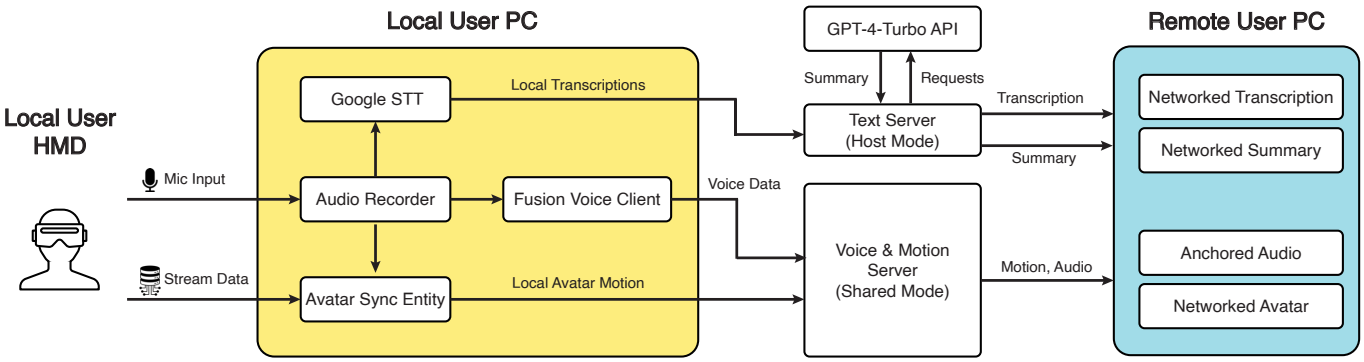


Fig. 3. Overview of the real-time multi-user networking pipeline for immersive VR meeting setup used in our formative study. Local users' mic inputs are captured and processed using Google STT for transcription. Audio data is streamed to a shared server via the Fusion Voice Client, which synchronizes voice and avatar motions. Transcriptions and summaries are handled by the Text Server (host mode), which sends requests to the GPT-4-Turbo API for summarization. Networked transcription, summaries, and audio are synchronized across remote users through the server to maintain a seamless, real-time meeting experience.

the results with manually verified transcripts, categorizing the summaries as accurate, relevant, or inaccurate. The average rates from the three participant groups are 89%, 9%, and 2%.

*3.1.3 Real-Time Multi-User Networking.* Our networking pipeline, shown in Figure 3, was built using Photon Fusion 2, Photon Voice, and Meta Avatar SDK to create a real-time environment for Meta avatars. We used a Photon Fusion shared mode server to synchronize avatar body movements, reducing latency across clients compared to host mode. Photon Voice handled voice streaming, while Meta Avatar SDK animated avatars using body tracking data from Oculus Quest Pro HMDs.

To handle captions and summaries, we set up a separate Photon host mode server, with the first user (moderator) acting as the host. The host server received transcriptions, generated summaries using the GPT-4-Turbo API, and shared them across clients. This separate server reduced client-side computational load, maintaining system stability over low-latency demands for caption generation. The research moderator also facilitated the session, ensuring smooth conversation flow and guiding participants during silences.

When a user's speech was transcribed locally via Google STT, an RPC signaled the host server to generate a concise summary via GPT-4-Turbo. A data processing queue managed the order of utterance inputs, ensuring accurate token generation. The server matched each user's Photon network ID with their respective captions and summaries, displaying them on the tabletop panel in the correct color and position, ensuring a synchronized, real-time meeting experience.

## 3.2 Participants

We recruited 12 participants (5 female, 7 male) for our formative study, ranging in age from 20 to 39 years ($M = 27.78$, $SD = 6.91$). Participants were organized into four groups, with three participants per group. On average, participants reported a moderate familiarity with VR experiences, with a mean score of $M = 3.5$ on a 7-point Likert scale (1: not familiar at all, 7: extremely familiar). None of the participants reported color blindness, and all had normal or corrected-to-normal vision. They were compensated with a $15 e-gift card for their participation.

## 3.3 Procedure

Participants were grouped by availability, with three per group plus the experiment coordinator. Each participant was escorted to separate rooms on the same floor, equipped with a PC and Oculus Quest Pro, all connected to the same Wi-Fi. They calibrated their eyes using the Oculus Quest Pro's eye calibration tool.

The study involved two trials per group with fixed debate topics: the first trial was the desert survival task [26, 51], where participants selected the top two items to take to a desert from seven options. The second trial asked participants to choose the top two workplace perks a company should implement, also from seven options. These topics were designed to encourage differing opinions and consensus-building. The experiment coordinator facilitated the conversation and acted as an agitator if consensus was reached too quickly.

Participants experienced both the full-transcript and summarized transcript interfaces, with the order counterbalanced across groups. Two groups began with the full-transcript interface, and two began with the summarized interface.

Prior to the trials, participants familiarized themselves with the transcription interface. The trial started after participants confirmed their opinions on the debate topic. Each trial lasted about 30 minutes, with participants taking turns "dropping out" for four minutes before rejoining as depicted in Figure 4. The coordinator guided the drop-out/rejoin process, and participants were asked to catch up on the conversation upon rejoining. Drop-out order was randomized to avoid interaction biases.

Quantitative data included automatically logged gaze behavior to assess focus and interaction with the transcription panel. Post-trial questionnaires included the NASA-TLX for cognitive load [17] and selected Networked Minds Social Presence Inventory (NMSPI) factors: *Co-presence, Attentional Allocation, Perceived Message Understanding,* and *Perceived Affective Understanding* [16]. Participants also completed a recall task and rated the interfaces on effectiveness, ease of use, and preference using a 5-point Likert scale.

Semi-structured interviews were conducted after the trials to gather insights into challenges with keeping up with meetings and preferences for the transcription interfaces. The entire experiment lasted approximately 90 minutes.

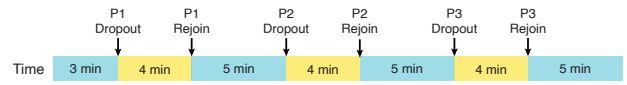


Fig. 4. An example of dropouts and rejoins of participants during the formative study. Note that the order of drop-out was randomized between trials.

## 3.4 Findings

In this section, we report qualitative and quantitative findings from our formative study.

*Gaze Patterns and Interactions.* Analysis of gaze patterns and interaction frequency revealed notable differences between full-transcript and summary modes. The frequency of interactions with the panel was lower in summary mode (M = 20) compared to full-transcript mode (M = 34.67). Participants spent more time gazing at the text panel in full-transcript mode, often at the expense of engaging with other participants. In contrast, summary mode reduced gaze time on the tabletop, allowing for more interaction with the group. One participant (P7) noted, "It can be a little distracting when there's so many words... Sometimes, I just look at the texts instead of looking at the people." Two researchers double-coded the gaze and interaction data, cross-checking with video recordings to confirm or adjust for hardware errors.

*Cognitive Load.* Cognitive load, measured using the NASA-TLX, showed no significant differences between the modes in mental demand, physical demand, effort, frustration, or overall score. However, a paired t-test revealed a significant difference in temporal demand, with summary mode outperforming full-transcript mode. This suggests that participants felt less time pressure using the summary interface. Detailed scores are shown in Figure 5 left.

*Social Presence.* Although no significant differences were found in co-presence, perceived message understanding, or perceived affective understanding between the two modes, summary mode had higher average scores across all factors. The only statistically significant difference was in attentional allocation, with summary mode scoring higher (M = 40.5, SD = 6.08) than full-transcript mode (M = 30.0, SD = 6.43), indicating participants felt their attention was better distributed in summary mode (see Figure 5 right).

*User Preferences.* Both modes were rated similarly in perceived usefulness for keeping up with meetings (summary: 4.44, full-transcript: 4.0 on a 5-point scale). However, 9 out of 12 participants preferred the summary mode when choosing between the two. Those who preferred the full-transcript mode wanted more detail and to follow the entire conversation, while those favoring summary mode felt it was more helpful after dropping out. Some noted that transcription errors, compounded by summarization, could occasionally result in loss of context.

*Challenges.* Participants highlighted several challenges with the transcription panel. Many found that the inability to capture tones and gestures hindered their understanding of nuanced conversations (P1, P6, P10, P12). Others found it difficult to catch up when asked for input while still reviewing missed content (P1, P2, P5). Frequent gaze switching between the panel and other participants was also a common issue, with some noting it was more disruptive than in traditional platforms like Zoom (P2). Several participants mentioned that scrolling through long transcripts felt overwhelming (P4, P8), leading some to skim for keywords rather than fully reading the text (P4). Concerns were raised that these issues might be more pronounced in larger groups or more complex discussions (P2, P5, P7, P8, P9).

*Desired Features.* Participants also provided valuable insights into how the transcription panel could be improved to better support their needs during immersive meetings. Their suggestions focused on enhancing the usability of the interface, allowing for more efficient re-engagement with the conversation, and better integrating the panel with the spatial dynamics of the VR environment.

Key areas for improvement highlighted by participants include:

- The ability to catch up on missed content at a glance, enabling quicker re-engagement (P1, P2, P9).
- An option to attach opinions or key points directly next to the respective participant's avatar, enhancing contextual understanding (P1, P2, P7, P8).
- Participants highlighted the need for greater utility in larger group meetings, where the interface could help manage more complex discussions (P1, P2, P7).
- There was an emphasis on improving the spatial integration of the transcription panel within the immersive environment to enhance the overall meeting experience (P2, P4, P10, P12).
- A feature to disable the transcription panel when not needed, as some participants found it distracting (P5, P7, P8, P10, P11).
- The ability to switch between full-transcript and summary modes based on the context and user preference (P2, P3, P5, P7, P11, P12).

## 3.5 Design Considerations

Drawing from our formative study findings, we propose three design considerations (DCs) to guide the development of the transcription panel for immersive meetings. These aim to address the challenges participants identified and improve user engagement and interaction.

***DC1: Attach transcripts to avatars to reduce distraction.*** Participants often found the fixed transcription interface distracting, which reduced social presence and attentional allocation. The full-transcript mode, in particular, increased temporal demand as users felt pressure to keep up with the continuous text, leading to disengagement. P11 noted, "I felt the rush to read through it quickly… I noticed others were not paying attention, just interacting with the transcription interface in front of them." Eight participants expressed a preference for avatar-fixed transcripts, which would help them balance attention between the interface and group conversation.

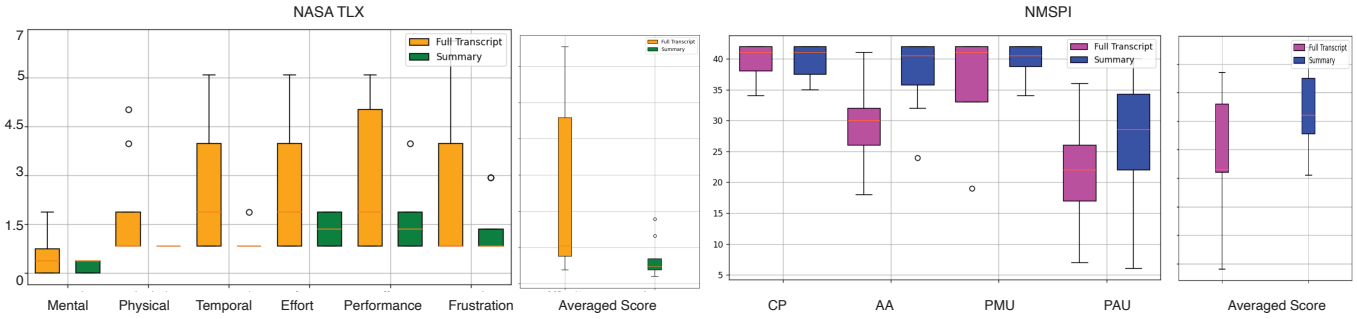


Fig. 5. NASA TLX and NMSPI scores from the formative study comparing Full Transcript and Summary interfaces. The NASA TLX results show lower cognitive load for the Summary interface across multiple subscales (Mental, Physical, Temporal, Effort, Performance, and Frustration). The NMSPI scores highlight higher Co-presence (CP) and Attention Allocation (AA) for the Summary interface, with relatively smaller differences in Perceived Message Understanding (PMU) and Perceived Affective Understanding (PAU). The boxplots indicate the median, quartiles, and outliers across participants, with overall lower cognitive load and higher social presence scores associated with the Summary interface.

Therefore, **DC1** emphasizes dynamic, avatar-fixed transcriptions to reduce distractions and support ongoing social engagement.

***DC2: Adapt transcription modes based on user context.*** Preferences for transcription modes varied depending on the meeting context. When rejoining after a disruption, participants preferred full transcripts to catch up, but for reviewing past discussions, they favored summaries for quicker re-engagement. They wanted the system to be aware of their disengagement periods and tailor summaries accordingly. This aligns with gaze pattern analysis, showing varied interaction based on engagement level. Thus, **DC2** suggests adaptive transcription modes that switch between full and summarized views based on user context to optimize information delivery.

***DC3: Provide on-demand access to the transcription interface.*** Several participants found the constant presence of the transcription interface distracting during conversations. P3 and P12 mentioned that being able to toggle the interface on and off would reduce cognitive load and help maintain focus. On-demand access would allow users to manage distractions and avoid interference with the conversation flow. Therefore, **DC3** focuses on designing the interface to offer on-demand access, letting users control when to engage with the transcription information and ensuring it remains a supportive rather than intrusive element.

## 4 ENGAGESYNC: CONTEXT-AWARE AVATAR-FIXED TRANSCRIPTION INTERFACE

Building on the design considerations (DCs), we developed **EngageSync**, an interface aimed at enhancing social presence and information recall in immersive meetings by dynamically adapting the information displayed on avatar-fixed text panels. EngageSync specifically addresses the challenges identified during the formative study, namely the need for reducing distractions (**DC1**), providing context-aware transcription modes (**DC2**), and enabling on-demand access to transcription information (**DC3**).

### 4.1 System Overview and Adaptive Transcription

EngageSync adapts transcription content based on the user's **context**, which refers to the state of user engagement and attention in the virtual meeting. From our formative study and interviews, we observed that user interaction with the transcription panel varied depending on whether they were focused on a speaker, a listener, or re-engaging after disengagement. These insights informed our definition of context for the system.

The system defines three key contexts based on user behavior:

- Focused on a Speaker: When the user directs their gaze at a speaking avatar, this indicates active engagement with the conversation.
- Focused on a Listener: When the user focuses on a listening avatar (i.e., an avatar that is not actively speaking), it reflects the user's attention on non-verbal aspects of the meeting.
- Re-engaging after Disengagement: If the user's gaze is not focused on any avatar or object in the virtual environment, the system detects disengagement. When the user refocuses on the avatars, the system recognizes a re-engagement scenario.

**Color-Coded Mode Differentiation:** To ensure users can easily differentiate between the modes, each text panel is marked with a colored circle in the bottom-right corner. Live-transcription panels do not feature a circle, engagement summaries display a green circle, and re-engagement summaries display an orange circle. This visual indicator helps users quickly identify the current mode, reducing confusion and improving interaction flow (see Figure 1 for visual examples).

**Gaze Tracking and Speech Detection:** EngageSync uses gaze tracking and speech activity detection (SAD) to monitor user engagement and context. The Meta Movement SDK's OVREyeGaze script captures the user's gaze direction, determining which avatar they are focused on. Simultaneously, SAD detects whether the avatar being observed is speaking or listening. Based on this input, the system toggles between live transcription and summary modes to ensure users receive the appropriate content. For instance:

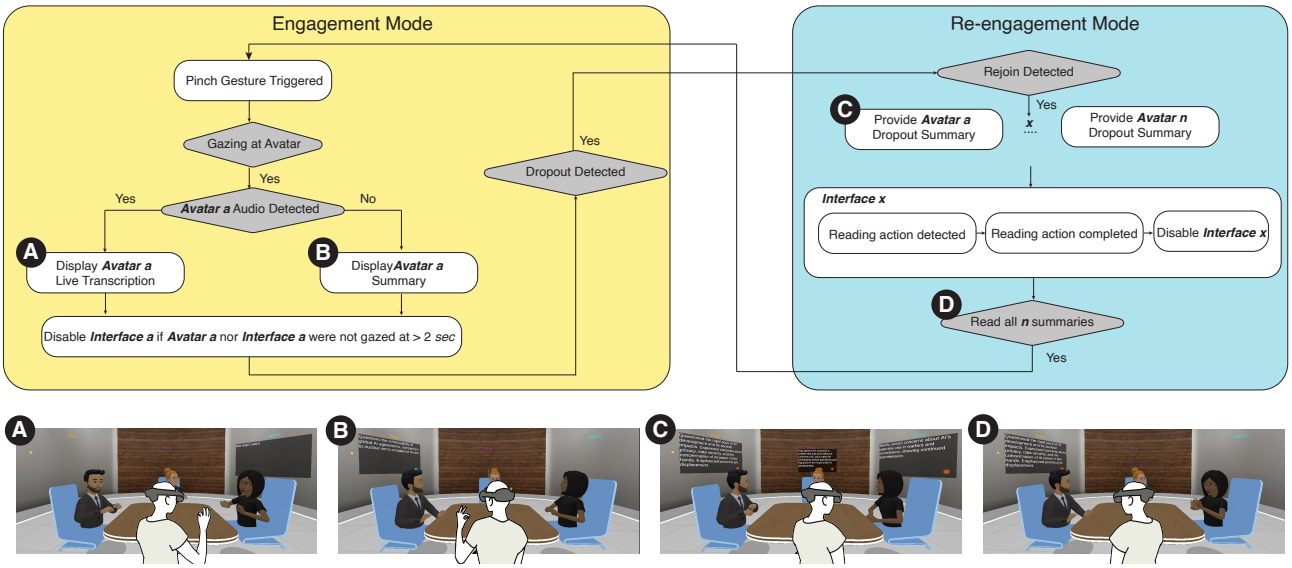


Fig. 6. An overview of **EngageSync** flowchart and demonstrational screen shots of key features. In Engagement Mode, (a) if the user performs a pinch gesture while looking at a speaker, the panel attached to the avatar displays a live transcription; (b) if the avatar is a listener (no audio detected), a summary of their previous utterance is shown. Upon rejoining after a dropout, (c) summaries of what each avatar said during the dropout are displayed. (d) Once it is 'read', the interface disappears, and when all the summary panels are read, the system returns to Engagement Mode.

- When the user focuses on a speaking avatar, EngageSync displays a live transcription panel attached to that avatar, providing real-time updates of the conversation.
- When the user focuses on a listening avatar, the system displays a summary of that avatar's last utterance, offering a brief overview rather than continuous text.

This approach to context-driven transcription directly addresses user feedback from the formative study, where participants noted that continuously receiving transcription while not focusing on the speaker could become overwhelming.

**Gesture Interaction:** To provide more intuitive control over transcription access, EngageSync leverages gesture-based interaction through the OVR Interaction SDK. Users can perform a pinch gesture while looking at an avatar to activate the corresponding transcription panel. This interaction method allows the user to summon or dismiss the transcription interface as needed, reducing cognitive load when it's not necessary.

Once activated, the transcription panel remains visible until the system detects a lack of gaze focus (i.e., if the user looks away or is inactive for over two seconds), at which point the transcription panel automatically fades, addressing the need for on-demand access (**DC3**). This gesture-based control enables users to balance information retrieval and social presence during the meeting.

By combining gaze tracking, speech detection, and gesture interaction, EngageSync dynamically adapts the information presented to users based on the current context. This context-aware approach allows users to stay engaged in the conversation without being distracted by unnecessary or overwhelming information. The system's flexibility in presenting full transcriptions during active engagement or concise summaries during re-engagement ensures that users can maintain their social presence in the meeting, even after disruptions.

## 4.2 Interaction Workflow: Engagement Mode

Engagement Mode manages the system's behavior when users are visually engaged in the meeting, either by actively contributing or passively observing. By tracking gaze and detecting speech, EngageSync adapts transcription content in real time, ensuring relevant information is displayed without causing distractions.

When a user focuses on a speaking avatar, a live transcription panel attached to the avatar provides real-time updates on the ongoing conversation. If the user's gaze shifts to a listening avatar, the system adapts by showing a concise summary of the avatar's last utterance instead of continuous transcription, helping the user remain informed without overwhelming them with unnecessary text.

To prevent visual clutter, the transcription panel automatically disappears if the user looks away for more than two seconds, in line with **DC1**. Additionally, users can control when the transcription panels appear or disappear by performing a pinch gesture, allowing them to summon or dismiss the information on demand, which reduces cognitive load.

Overall, Engagement Mode offers a dynamic, context-aware experience, ensuring that transcription content is available when necessary while minimizing distractions. For a visual overview, refer to Figure 6.

## 4.3 Interaction Workflow: Re-engagement Mode

Re-engagement Mode activates when users temporarily disengage from the meeting, such as by looking away or becoming inactive.

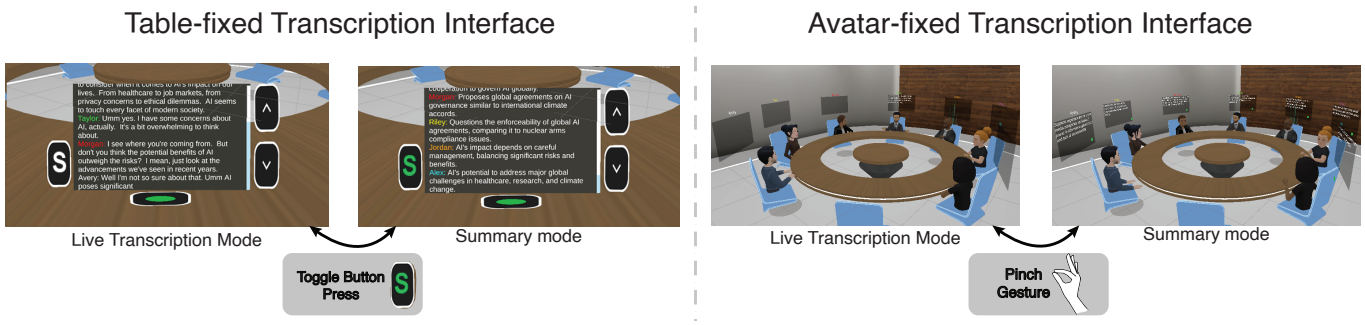


Fig. 7. The two interface compared with EngageSync in the user study. Users can toggle between live transcription mode and summary mode for both interfaces. For Table-fixed Transcription Interface (Left), this is triggered by simply pushing the 'S' button on the interface. In Avatar-fixed Transcription Interface (Right), users toggle between mode by doing a pinch gesture regardless of their gaze.

The system detects disengagement based on the absence of gaze on avatars or objects in the room, logging conversation contributions during this period to ensure users are quickly updated upon returning.

When disengagement is detected, the system records ongoing conversations, capturing full utterances from speaking avatars to provide users with complete context. Upon re-engagement—, the user's gaze refocuses on avatars or virtual objects—concise 15-word summaries of the missed conversations are generated using GPT-4 Turbo. These summaries are attached to the corresponding avatars, helping the user quickly catch up on key points.

The system tracks whether users engage with these summaries by monitoring their gaze. If the user's gaze remains on a summary panel for more than 1.5 seconds, it indicates that they have begun reading the panel. Once a summary is read, the panel disappears, unless the user looks back within two seconds, allowing the information to be reviewed without overwhelming the user with excessive content.

After all summaries have been reviewed, the system automatically transitions back to Engagement Mode, resuming real-time transcription and summaries based on the user's focus. This adaptive interaction flow ensures that users remain informed and engaged, even after temporary disruptions, maintaining their social presence in the meeting without unnecessary distractions.

## 5 USER STUDY

To evaluate the effectiveness of EngageSync, we conducted a user study comparing it against two other transcription interfaces in immersive VR meetings.

### 5.1 Compared Interfaces and Hypotheses

In our study, we compare EngageSync to two other transcription panel configurations: Table-fixed Transcription Interface and Avatar-fixed Transcription Interface. See Figure 7 for visual reference.

**Table-fixed Transcription Interface (TableTI)**: This interface, expanding from the version used in our formative study, positions the transcription panel in front of the user in a fixed position on the table. We introduced a toggle button to the left of the text panel to switch between real-time transcription and a summary of each speaker's previous utterances. The interface retains the familiar scrolling controls with up and down buttons and an auto-scroll option to the right and bottom. When a user rejoins the meeting after a disruption, the text display resumes from where they left off, ensuring continuity.

**Avatar-fixed Transcription Interface (AvatarTI)**: This interface attaches transcription panels directly above each avatar, similar to live captions or subtitles used in social VR platforms like VRChat [55]. These panels are constantly visible, and the user can toggle between live transcription mode and summary mode using a pinch gesture. In live transcription mode, only the currently speaking avatar has its text displayed, while in summary mode, all avatars display a summary of their previous utterance. This interface does not adapt to disengagement and simply shows previous summaries, as common avatar-attached interfaces typically lack context awareness.

On the other hand, EngageSync adapts the display of transcription panels based on the user's engagement. Instead of always showing the panels, users must look at an avatar and make a pinch gesture to activate the panel. The content displayed depends on the user's context, providing live transcriptions when engaged and summarizing missed conversations when re-engaging after a disruption. The differences and common aspects between these interfaces are summarized in Table 1.

Our study aims to test the following hypotheses, based on findings from the formative study and previous research:

- **H1: EngageSync enhances both social presence and the ability to keep up with the conversation.** We hypothesize that EngageSync, which provides live transcription and context-based summaries, will lead to improved social presence and more effective re-engagement with the conversation, compared to TableTI and AvatarTI.
- **H2: Avatar-fixed transcription interfaces increase social presence compared to baseline conditions.** We expect that all avatar-fixed transcription interfaces (AvatarTI and EngageSync) will provide a greater sense of social presence than the baseline tabletop condition, as they allow users to maintain focus on the avatars during the conversation.
- **H3: Table TI and EngageSync support better re-engagement after disengagements.** We hypothesize that both TableTI

| Feature | Table TI | Avatar TI | ES (EngageSync) |
|---|---|---|---|
| **Panel Position** | Fixed on Table | Fixed to Avatar | Fixed to Avatar |
| **Missed Content Handling** | Yes (Manual Scroll) | No | Yes (Automatic Summary) |
| **Activation** | Always Visible | Always Visible | Gaze and Gesture-Triggered |
| **Transcription Mode Switching** | Manual (Button) | Manual (Gesture) | Automatic (Context-based) |

Table 1. Comparison of transcription panel configurations. EngageSync shares common aspects with Avatar TI in being avatar-fixed, and with Table TI in handling missed content. However, EngageSync stands out by offering on-demand access (gaze and gesture-triggered) and automatic transcription mode switching, making it more adaptable and less intrusive in immersive meetings.

and EngageSync will enable participants to catch up more effectively after disengagements, improving their ability to re-engage with the conversation compared to AvatarTI.

- **H4: Avatar-fixed transcription interfaces are more effective in larger groups.** Based on the formative study feedback, we hypothesize that AvatarTI and EngageSync will show greater usability and effectiveness in larger groups, where the conversation is distributed among more participants compared to TableTI.

## 5.2 Experiment Design

To test these hypotheses, we designed the user study to consist of two groups of immersive meeting setups of different sizes. Here, participants join pre-recorded group conversations as listeners, allowing for consistent analysis across all participants.

***Group Size Conditions.*** The evaluation of these interfaces will be conducted under two distinct group sizes.

- **Small Group Conversation (3 speakers):** Following the setup of our formative study, this condition will consist of ***three*** virtual agents conversing. This will result in four avatars placed in the room, including the participants.
- **Mid-sized Group Conversation (7 speakers):** In this condition, the group size is scaled up with participants observing a conversation among ***seven*** virtual agents. This design choice was driven by formative study feedback, where users expressed that group size would affect the usability of the interface and that they would find avatar-fixed panels even more useful in larger groups.

To clarify, the study involves comparing the three interfaces within each group size condition, resulting in a within-subjects design for each group size. Additionally, the comparison between the two group sizes will follow a between-subjects design, assessing how group size influences the usability and effectiveness of the interfaces.

***Dropout Simulation.*** To simulate real-world disruptions, the participant hears a phone ringing sound 3 minutes into the experiment, signaling a “drop out" situation. Immediately after, the participant’s avatar is relocated to a separate virtual environment where a simple math quiz is displayed on a large text panel (see Figure 8). A countdown timer is also shown, indicating the time remaining until the participant is automatically returned to the meeting. Participants are instructed to solve as many math problems as possible within the remaining time. This design follows

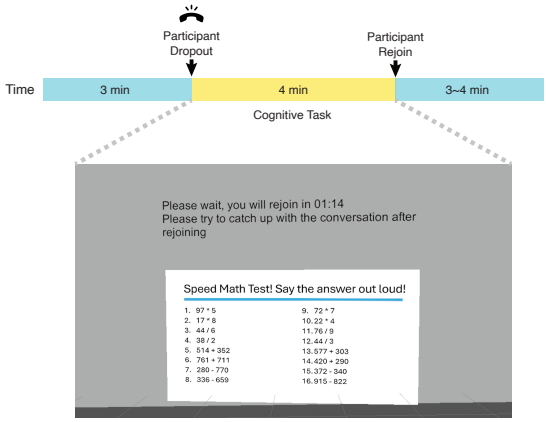


Fig. 8. Timeline of participant dropout and rejoin during the user study. Three minutes into the conversation, participants are interrupted by a phone chime and teleported to a different virtual environment, where they complete a speed math test. A timer displayed above the test indicates the time remaining until they rejoin the meeting.

prior work [49] that employs cognitive tasks that interrupt participants’ engagement from the original task; in this case, distracting them from the previous conversation context. After 4 minutes, the participant’s avatar is relocated back to the ongoing meeting, with the conversation continuing for another 3 to 4 minutes, as each conversation script is designed to last between 10 and 11 minutes.

| Role | Small Group (SA1-3) | Mid-sized Group (MA1-7) |
|---|---|---|
| Pro-topic | SA1 | MA1, MA4, MA7 |
| Against-topic | SA2 | MA2, MA5, MA6 |
| Less talkative | SA3 | MA3 |

Table 2. An example of script distribution in small and mid-sized groups. Each participant in the small group corresponds to multiple participants in the mid-sized group with similar viewpoints, while one participant in each group represents a less talkative role.

***Conversation Script.*** It is important to note that the script for both group size conditions is exactly the same. The speech content is divided among three avatars in the small group and seven avatars in the mid-sized group. We designed the script carefully so that a script for one speaker in the small group was distributed to three speakers in the mid-sized group. Additionally, one speaker in both groups speaks less than the others, this design was made to evaluate

whether users can recall or remember what that speaker said using the different interfaces. An example of the script distribution can be found in Table 2.

We designed three different topics/scripts for the conversations: "*Is Online Education as Effective as Traditional In-Person Education?*", "*Should Social Media Platforms be Regulated by the Government?*", "*Is Artificial Intelligence More Beneficial or Harmful to Society?*".

***Measured Items.*** In the study, we collected both quantitative and qualitative data to assess the effectiveness of the transcription methods in supporting user engagement and understanding in immersive meetings.

For quantitative measures, we measured the following metrics:

- **Re-engagement Time:** After a trial, participants were shown a screen recording of their trial and were asked to mark the time they thought they were 'caught up' with the conversation.
- **Information Recall:** We measure information recall by twofold. One is a 6-question quiz about the content of the conversation. The question consists of two questions about the conversation before drop-out, 2 questions during dropout, and 2 questions after dropout. Second, participants were asked to recall and write down as much as they could remember from the least talkative member in the conversation after the session, we measured whether the participant remembered what the least talkative speaker's main point was in a binary manner. The accuracy will be quantified to assess the effectiveness of the transcription method in supporting memory retention.
- **Gaze on Interface:** The Gaze tracking data were analyzed to measure the time a participant spent gazing at the interface versus avatars in percentage out of the trial duration excluding dropout time. This was automatically logged based on gaze-tracking and was cross-checked with two researchers compared with screen recordings of the trial.

For qualitative measures, we will collect the following data:

- **Social Presence:** Assessed using the NMSPI questionnaire, the same one used in the formative study, focusing on selected components: co-presence, attention allocation, PMU, and PAU.
- **Cognitive Load:** Measured using the NASA-TLX questionnaire in a 7-point Likert scale.
- **Usability:** Measured using the System Usability Scale (SUS) [7] questionnaire in a 5-point Likert Scale.
- **Utility Questions** We further asked participants how much the perceived interface supported them in (i) keeping up with the conversation (ii) and after re-joining in the conversation on a 5-point Likert scale. Finally, we asked participants to rate their how they prefer having this interface in an immersive meeting setup in a 5-point Likert Scale.

## 5.3 Setup

The setup for this user study largely mirrors that of our formative study, leveraging the same foundational technologies and equipment. We used Unity running on a PC equipped with 32GB RAM, an Intel i9 processor, and an NVIDIA GeForce RTX 3080 Ti. Participants wore a Meta Quest Pro headset, which supports both gesture and eye tracking. The Meta Avatar SDK was again utilized to represent avatars, maintaining consistency with the formative study.

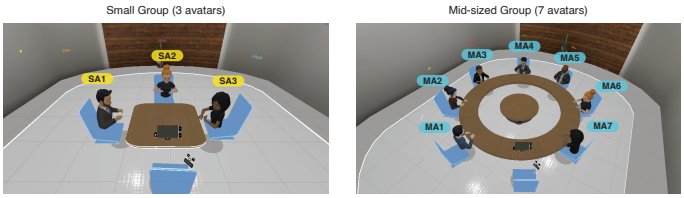


Fig. 9. The virtual environment setup for small group with three speakers (Left) and mid-sized group with seven speakers (Right).

The avatars in this user study were pre-recorded with voice actors reciting scripts specifically designed for the two group sizes: three actors for the small group and seven actors for the mid-sized group. The motion capture was performed using the Avatar Recording tool [1], and all conversation sequences were synchronized using Unity Timeline to ensure consistency across sessions.

The virtual environment remained a conference room setup, with avatars seated around a table. The configuration of the environment for each group size is illustrated in Figure 9. Each speaker's name was virtually generated, color-coded, and attached above their avatar to aid identification. Additionally, for AvatarTI and EngageSync conditions, the text panels were placed above each corresponding avatar, as was done in the formative study.

## 5.4 Participants

We recruited 30 participants (14 female, 16 male) from a university sample, aged between 19 and 35 years ($M$ = 26.70, $SD$ = 4.19). Participants were randomly assigned to one of two group size conditions: the small group study (3 avatars) or the mid-sized group study (7 avatars), with 15 participants in each condition. None of the participants reported color blindness, and all had normal or corrected-to-normal vision.

On average, participants reported a moderate familiarity with virtual reality (VR), with a mean score of $M$ = 4.77 ($SD$ = 1.45) on a 7-point Likert scale (1 = not familiar at all, 7 = extremely familiar). They also indicated their frequency of attending online meetings ($M$ = 3.87, $SD$ = 0.92) on a scale from 1 (never) to 5 (daily). Additionally, participants reported their experience with either personally dropping out of meetings ($M$ = 2.75, $SD$ = 1.06) or observing others drop out ($M$ = 3.10, $SD$ = 1.05) on a scale from 1 (never) to 5 (always). Participants were compensated with a $15 e-gift card for their participation.

## 5.5 Procedure

Participants first reviewed and signed a consent form, after which they completed a demographic survey. They were briefed on the study's objective, which was to evaluate how effectively they could keep up with the conversation during the discussion and after experiencing a simulated drop-out. Eye-tracking calibration was performed using Meta Quest Pro's internal eye-tracking calibration software to ensure accuracy throughout the experiment.

| Measurement | | Small Group | Mid Group | Combined Group |
|---|---|---|---|---|
| Social Presence | CP | ES > Avatar > Table<br>$\chi^2 = 6.739$ $p = .032$ | ES > Avatar > Table<br>$\chi^2 = 15.500$ $p = .004$ | ES > Avatar > Table<br>$\chi^2 = 14.1261$ $p > .001$ |
| | AA | ES > Avatar > Table<br>$\chi^2 = 6.0370$ $p = .049$ | Avatar > ES > Table<br>$\chi^2 = 12.171$ $p = .002$ | Avatar > ES > Table<br>$\chi^2 = 19.9818$ $p < .001$ |
| | PMU | ES > Avatar > Table<br>$\chi^2 = .1509$ $p = 0.9273$ | Avatar > ES > Table<br>$\chi^2 = 1.7818$ $p = 0.4103$ | ES > Avatar > Table<br>$\chi^2 = 1.9205$ $p = .2516$ |
| | PAU | Avatar > Table> ES<br>$\chi^2 = .6250$ $p = .2691$ | ES > Avatar > Table<br>$\chi^2 = 1.7193$ $p = .4233$ | Avatar > Table> ES<br>$\chi^2 = 1.734$ $p = .4204$ |
| Gaze Time at Avatar (%) | | ES > Avatar > Table<br>$\chi^2 = 11.831$ $p = 0.003$ | ES > Avatar > Table<br>$\chi^2 = 9.552$ $p = 0.008$ | ES > Avatar > Table<br>$\chi^2 = 21.0427$ $p < .001$ |
| Information Recall | Quiz | ES > Table> Avatar<br>$\chi^2 = 2.2593$ $p = .323$ | ES > Table> Avatar<br>$\chi^2 = 6.882$ $p = .018$ | ES > Table> Avatar<br>$\chi^2 = 3.406$ $p = .182$ |
| | Remember | ES > Table> Avatar<br>$\chi^2 = Nan$ $p = Nan$ | ES > Table> Avatar<br>$\chi^2 = 6.0479$ $p = .046$ | ES > Table> Avatar<br>$\chi^2 = 4.4783$ $p = 0.1066$ |
| Re-engage Time | | ES > Table> Avatar<br>$\chi^2 = 6.7733$ $p = 0.047$ | ES > Table> Avatar<br>$\chi^2 = 19.733$ $p < .001$ | ES > Table> Avatar<br>$\chi^2 = 22.200$ $p < .001$ |
| NASA TLX | | Table> ES > Avatar<br>$\chi^2 = 3.333$ $p = .189$ | Table> ES > Avatar<br>$\chi^2 = 13.1189$ $p = .002$ | Table> ES > Avatar<br>$\chi^2 = 14.721$ $p = .002$ |
| SUS | | ES=Table> Avatar<br>$\chi^2 = 1.750$ $p = .196$ | ES > Avatar > Table<br>$\chi^2 = 3.263$ $p = .159$ | ES > Avatar > Table<br>$\chi^2 = 4.266$ $p = 0.119$ |
| Facilitated Catching Up | | ES > Avatar=Table<br>$\chi^2 = .0571$ $p = .972$ | ES > Avatar > Table<br>$\chi^2 = 1.792$ $p = 0.342$ | ES > Avatar > Table<br>$\chi^2 = 1.942$ $p = .679$ |
| Facilitated Re-engagement | | ES > Table> Avatar<br>$\chi^2 = 8.509$ $p = .014$ | ES > Table> Avatar<br>$\chi^2 = 19.304$ $p < .001$ | ES > Avatar > Table<br>$\chi^2 = 26.00$ $p < .001$ |
| Preference | | ES > Table> Avatar<br>$\chi^2 = 2.167$ $p = .339$ | ES > Avatar > Table<br>$\chi^2 = 4.00$ $p = .135$ | ES > Avatar > Table<br>$\chi^2 = 4.356$ $p = .113$ |
| Ranked Preference | | ES > Table> Avatar | ES > Avatar > Table | ES > Table> Avatar |

Table 3. Statistical analysis results for different groups and measures (EngagementSync = ES, TableTI = Table, AvatarTI = Avatar). Results with significant differences are highlighted with a light pink color for $p < .05$, a pink color for $p < .01$, and a darker pink color for $p < .001$. Overall, EngageSync provides higher co-presence and attention allocation with faster re-engagement time with a more significant difference with mid-sized group.

Each participant completed three trials, with each trial corresponding to a different transcription interface and discussion topic. While the topic order remained fixed, the order in which the interfaces were presented was counterbalanced across participants to mitigate any ordering effects. Before each trial, participants received a 5-minute training session on the interface they would be using for that particular trial. Once the participant felt comfortable with the interface, the trial commenced.

Each trial consisted of a 10-minute conversation, including a 4-minute "drop-out" phase. During the drop-out phase, participants were relocated to a different virtual environment, where they were tasked with solving a simple math quiz. Participants were instructed to focus on solving the quiz while keeping an eye on the timer to be aware of the time remaining before rejoining the conversation.

After each trial, participants completed the NMSPI, NASA TLX, SUS, and a custom set of utility questions designed to assess their experience with the interface. Upon completing all three trials, participants were asked to rank their preference among the three transcription interfaces. A semi-structured interview followed, aimed at gaining deeper insights into their experiences and preferences.

Participants were offered short breaks between trials if needed, and the entire study lasted approximately 60 to 90 minutes.

## 5.6 Results

We present both qualitative and quantitative results for small, mid-sized, and combined group conditions. Differences between these group sizes are also reported.

The Shapiro–Wilk test indicated that the data were not normally distributed ($p < 0.05$), so we applied non-parametric tests throughout. Friedman tests were used to assess differences within each group condition, followed by Dunn's post-hoc tests with Bonferroni correction for pairwise comparisons. Additionally, Mann-Whitney U tests were used for between-group comparisons.

All statistical results, including $\chi^2$ and $p$-values, can be found in Table 3. In the following sections, we focus on reporting the significant findings with an emphasis on the key insights.

*5.6.1 Social Presence and Gaze Time on Avatar.* For social presence, both EngageSync and AvatarTI, avatar-fixed transcription interfaces, yielded significantly higher CP and AA compared to TableTI across all group sizes (see figure 10. This supports the hypothesis that avatar-fixed transcription helps users maintain a greater sense of social presence. The effect was more pronounced in mid-sized groups, where EngageSync ($p < .05$) and AvatarTI ($p < .05$) had significantly higher CP and AA than TableTI. In small groups, EngageSync and AvatarTI still outperformed TableTI in both CP and AA. These results support **H2**.

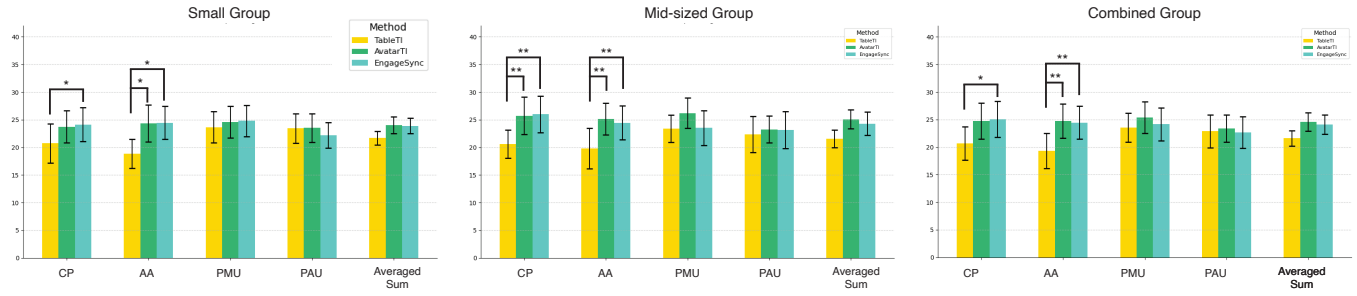


Fig. 10. Results for social presence across three group conditions (small, mid-sized, combined). Social presence was measured using four subfactors: Co-presence (CP), Attentional Allocation (AA), Perceived Message Understanding (PMU), and Perceived Affective Understanding (PAU), along with an averaged sum of all factors. Significant differences are indicated by * ($p < 0.05$) and ** ($p < 0.01$). Across all conditions, EngageSync consistently resulted in higher co-presence and attentional allocation scores compared to both AvatarTI and TableTI, particularly in the mid-sized group. In the combined group analysis, significant differences were observed between EngageSync and the other interfaces for CP and AA, demonstrating its effectiveness in enhancing social presence and attention management.

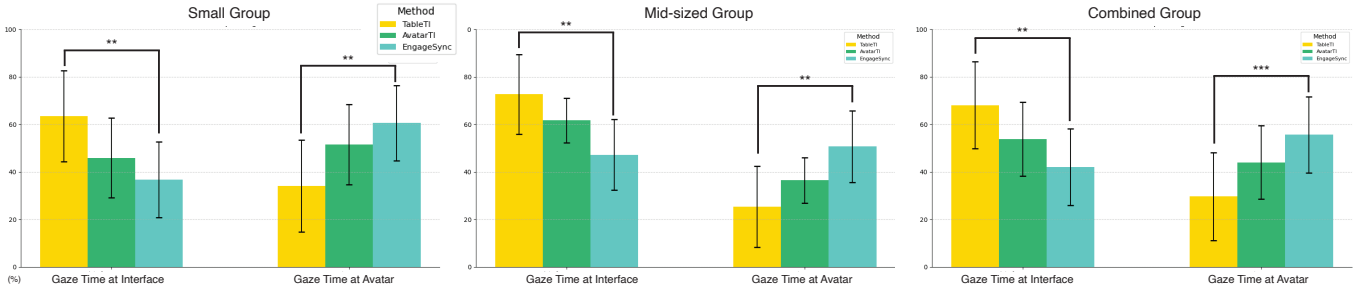


Fig. 11. Gaze time distribution for both avatars and interfaces across different group conditions. The left panel shows that in the small group condition, participants spent significantly more time gazing at the interface when using TableTI compared to both AvatarTI and EngageSync, whereas gaze time on avatars was higher for both avatar-fixed interfaces. A similar trend is observed in the mid-sized and combined groups, with EngageSync and AvatarTI leading to significantly more gaze time at avatars and less at the interface compared to TableTI. Significant differences are indicated by ** ($p < 0.01$) and *** ($p < 0.001$). This suggests that avatar-fixed interfaces reduce user focus on the interface, enabling more natural engagement with other participants.

While no significant differences were found in PMU and PAU ($p > .05$ for both), this might be due to the non-participatory nature of the study, as participants were observers, and the use of cartoonish avatars, which may have limited the expression of nuanced facial cues. Despite this, the overall improvement in CP and AA confirms that avatar-fixed systems encourage users to remain more engaged with the group rather than focusing on a table-fixed interface.

In terms of gaze time, participants spent significantly more time looking at avatars in EngageSynccompared to TableTI, especially in mid-sized groups (EngageSync vs. TableTI: $p = 0.003$). In small groups, TableTI also drew more attention, with participants spending significantly more time looking at the interface compared to and EngageSync ($p = .032$). No significant difference was found in gaze time between EngageSync and AvatarTI in either group size condition.

No significant differences were found between small and mid-sized groups for CP ($U = 110.5, p = .21$) or AA ($U = 115.0, p = .65$), suggesting that the benefits of avatar-fixed transcriptions remain consistent regardless of group size. However, participants in mid-sized groups spent significantly more time gazing at the table-fixed interface ($U = 177.0, p = .0079$) compared to small group participants, suggesting that as the conversation becomes more complex, static transcription panels become more distracting.

These findings suggest that EngageSync and AvatarTI improve social presence by allowing users to focus more on avatars rather than transcription interfaces. The results also suggest that TableTI may lead to increased distractions in larger group meetings, supporting the need for adaptive, context-aware systems. While EngageSync shows promising potential, further research is needed to fully understand how minimizing interface-gazing affects social engagement over time.

*5.6.2 Information Recall and Re-engagement Time.* Information recall and re-engagement time are closely related measures, both assessing how well participants recover from interruptions. Information recall focuses on participants' memory retention of key discussion points, while re-engagement time measures how quickly they rejoin the conversation flow after being disengaged. Together, these metrics provide insight into how each interface supports users in regaining conversational context (see figure 12). This supports **H3**.

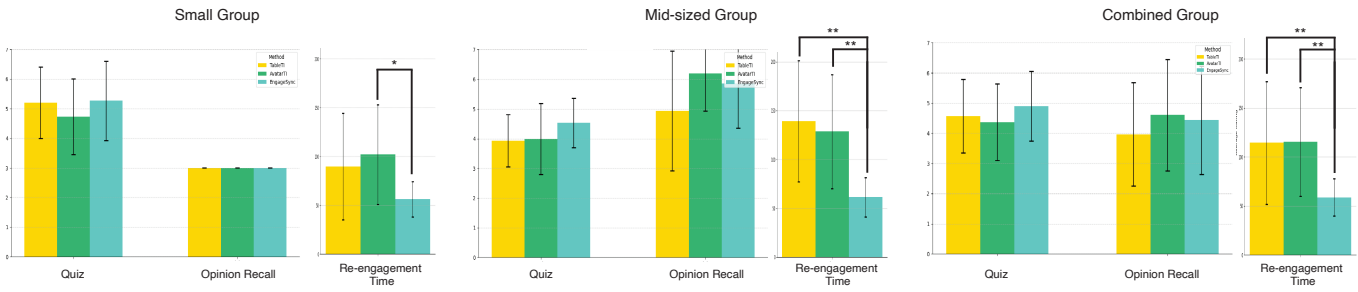


Fig. 12. Information recall results, including quiz scores, opinion recall, and re-engagement time across small, mid-sized, and combined groups. In the small group, no significant differences were found in quiz scores or opinion recall across the methods, but EngageSync led to significantly faster re-engagement times compared to AvatarTI. In the mid-sized group, significant differences emerged for both re-engagement time and opinion recall, with EngageSync outperforming AvatarTI and TableTI. Across the combined group, EngageSync consistently led to faster re-engagement times and better opinion recall. Significant differences are indicated by * ($p < 0.05$) and ** ($p < 0.01$).

**Information Recall:** EngageSync consistently outperformed both TableTI and AvatarTI in supporting memory retention, particularly in mid-sized groups. While no significant differences were found in the small group condition, participants using EngageSync in the mid-sized group recalled more information, particularly from quieter speakers. This is likely due to the dynamic summaries generated during their re-engagement period, which allowed users to focus on key points. Specifically, EngageSync participants scored higher in quiz results and written recall tasks compared to the other interfaces, confirming that the context-aware summaries helped users retain more detailed information. This effect was particularly strong in the mid-sized group, where tracking individual contributions became more challenging, highlighting the advantage of adaptive summaries in larger, more complex conversations. For recalling the least talkative person, no significant differences were found. However, in the small group, all participants remembered the least talkative person. In the mid-sized group, 6 participants using Table T1 and 3 participants using avatarTI were unable to recall the opinion of the least talkative person. In contrast, no such recall issues were reported in any of the EngageSync trials.

**Re-engagement Time:** Across both group sizes, participants using EngageSync re-engaged with the conversation faster than those using TableTI or AvatarTI. In the mid-sized group, the differences were more pronounced, with EngageSync users catching up significantly faster than those using AvatarTI. The flexibility of EngageSync's automatic summary display upon rejoining helped participants quickly digest what they had missed. In contrast, users of AvatarTI often reported taking longer to catch up, particularly in larger groups, where more speakers made it harder to stay oriented. These findings suggest that in more complex conversational environments, EngageSync's ability to summarize missed content dynamically led to more efficient re-engagement.

The difference in performance between small and mid-sized groups reinforces the importance of adaptive systems in managing re-engagement. In smaller groups, all three interfaces performed relatively similarly, as fewer speakers made it easier to catch up. However, in mid-sized groups, the advantage of EngageSync became clear, as it significantly shortened re-engagement times compared to both TableTI and AvatarTI. This suggests that as group size and conversational complexity increase, the ability to quickly summarize missed content becomes crucial for efficient re-engagement.

These results underline EngageSync's strength in both helping users recall information and facilitating faster re-engagement, especially in larger groups. The context-aware nature of EngageSync seems to play a key role in minimizing cognitive load, enabling users to quickly catch up and retain more information after disruptions.

*5.6.3 Cognitive Load and Usability.* **Cognitive Load** The NASA-TLX scores revealed no significant differences in the small group condition across the three interfaces (see Figure 13). This suggests that in smaller groups, the cognitive effort required to use EngageSync, AvatarTI, and TableTI was comparable.

In the mid-sized group, however, TableTI induced significantly higher cognitive load compared to both EngageSync ($p < .05$) and AvatarTI ($p < .05$). A similar trend was observed across all participants in the combined group, where TableTI ($Mdn = 22.0$) led to significantly higher cognitive load than EngageSync ($Mdn = 18.0$, $p < .01$) and AvatarTI ($Mdn = 19.0$, $p < .05$). Notably, there was no significant difference between the two avatar-fixed panels ($p = .19$), indicating that the avatar-fixed solutions maintained a relatively low cognitive load in all conditions.

In terms of group size, participants in the mid-sized group reported significantly higher cognitive load when using TableTI compared to the small group ($U = 50.0, p = 0.0084$). This suggests that as the group size increases, the cognitive demands of processing full transcripts on the tabletop interface become more pronounced, while EngageSync and AvatarTI remain more manageable.

**Usability** Despite these differences in cognitive load, no significant differences in usability (SUS scores) were observed across the interfaces in either group size (small or mid-sized) or when combining all participants (see Figure 14). This suggests that, even though TableTI imposed higher cognitive load, particularly in mid-sized groups, participants did not report a negative impact on overall usability. There were no significant group size effects for usability scores either ($U = 103.5, p = .29$).

This finding indicates that while TableTI requires more cognitive effort, particularly in larger groups, it does not detract from its

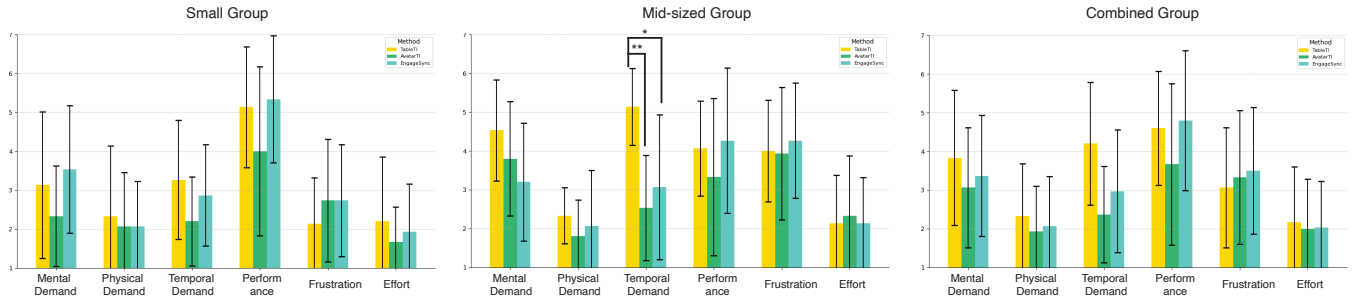


Fig. 13. NASA TLX results across group conditions. Temporal demand was significantly higher for TableTI compared to both AvatarTI and EngageSync in the mid-sized group, while no significant differences were observed for other subscales in the small group. The combined group analysis similarly showed higher temporal demand for TableTI, with EngageSync and AvatarTI resulting in lower mental and physical demand, frustration, and effort across all group sizes. Significant differences are indicated by * ($p < 0.05$) and ** ($p < 0.01$).

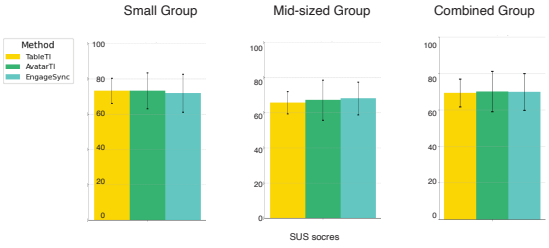


Fig. 14. SUS results across group conditions. The scores did not show any significant differences in any group, indicating that participants found all interfaces to have comparable usability, regardless of group size or interface type.

usability. Participants may still find it helpful for full transcripts, despite the increased effort, while EngageSync and AvatarTI offer similar usability with less cognitive load.

*5.6.4 Utility Questions.* We asked participants to rate the utility of each interface on a 7-point Likert scale, focusing on how well the interfaces facilitated catching up and re-engagement after disruptions (see Figure 15).

**Facilitated catching up:** Participants were asked how well each interface helped them catch up with the conversation after being disrupted. Across all conditions (small, mid-sized, and combined groups), there were no significant differences in participants' ratings of the interfaces' ability to help them catch up. This suggests that participants felt equally capable of catching up using EngageSync, AvatarTI, and the tabletop interface, regardless of group size or interface type ($U = 101.6$, $p = .28$).

**Facilitated re-engagement:** When evaluating how well each interface helped participants re-engage after dropping out, significant differences were found across group sizes. In the small group condition, both EngageSync and the tabletop baseline were rated as more helpful for re-engagement compared to AvatarTI. Participants found it easier to return to the conversation using EngageSync (*Mdn* = 4.0) and the tabletop interface (*Mdn* = 3.0) than with AvatarTI (*Mdn* = 2.0), although no significant difference was observed between EngageSync and the tabletop baseline.

In the mid-sized group, EngageSync performed notably better than both AvatarTI and the tabletop interface. Participants rated EngageSync significantly higher (*Mdn* = 5.0) for helping them re-engage compared to AvatarTI (*Mdn* = 3.0) and the tabletop baseline (*Mdn* = 4.0). This suggests that the adaptive features of EngageSync were particularly effective in facilitating re-engagement in larger group settings, likely due to its ability to present summarized content after periods of disengagement.

Across all participants, EngageSync (*Mdn* = 4.0) was rated as the most helpful interface for re-engagement, outperforming both AvatarTI (*Mdn* = 3.0) and the tabletop interface (*Mdn* = 4.0). These results indicate a clear preference for the context-aware interface, which provided more effective support for rejoining the conversation after disruptions. Additionally, participants in the mid-sized group rated EngageSync as significantly more helpful compared to those in the small group, suggesting that the interface's adaptive features are even more valuable in larger group settings.

***Ranked Preference.*** Participants ranked their preferences for the three interfaces, and results showed a clear favoring of EngageSync. In the small group condition, 7 participants ranked EngageSync as their first choice, closely followed by 6 participants preferring TableTI, while only 2 chose AvatarTI. In the mid-sized group, preference for EngageSync increased, with 10 out of 15 participants ranking it first, while AvatarTI saw a slight rise in preference (3 participants), and TableTI fell to 2 participants ranking it first, more frequently receiving third-place rankings.

Across all participants, EngageSync was the most preferred interface (17 out of 30), followed by TableTI (8 out of 30) and AvatarTI (5 out of 30). Interestingly, while AvatarTI's preference increased in mid-sized groups, TableTI remained consistently less favored, suggesting that the on-demand nature of EngageSync was better suited for both group sizes, with AvatarTI becoming slightly more acceptable in larger groups.

## 6 DISCUSSION

Our study offers important insights into the balance between maintaining social engagement and efficiently keeping up with conversations in immersive meetings. The results and participant feedback reveal key themes that shed light on how different transcription

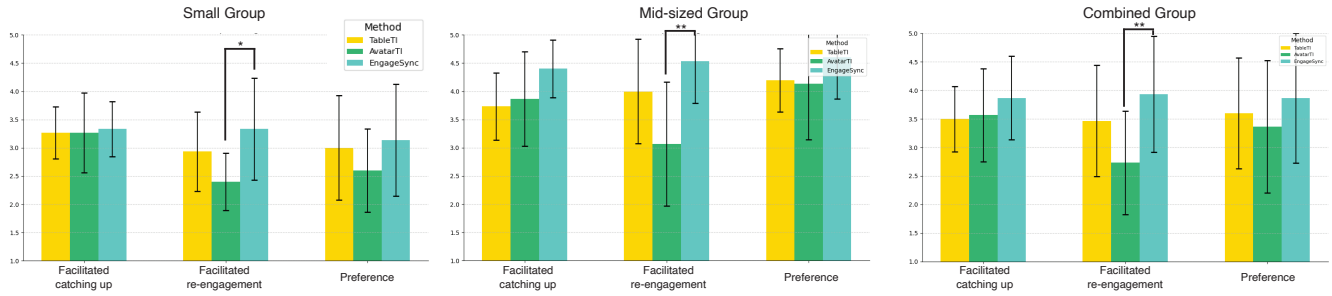


Fig. 15. Statistical results of utility-related questions. Results showed that participants perceived EngageSync to facilitate re-engagement compared to AvatarTI. No statistical differences were observed for perceived catch-up facilitation and preference.

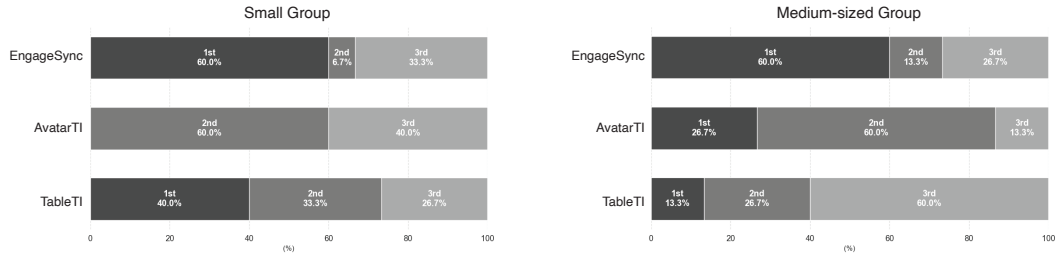


Fig. 16. Ranked Preference Scores across different group conditions. EngageSync was consistently ranked between small and mid-sized group conditions with over half of the participants preferring it.

interfaces, particularly EngageSync, influence the user experience. While participants valued EngageSync's adaptive nature, their preferences varied depending on factors such as group size, ease of catching up, and the cognitive effort required to manage the conversation.

In the following sections, we explore the perceived trade-offs between social engagement and information retention, as highlighted by participants. We also examine the strategies participants used to catch up on missed content, the specific challenges associated with recalling contributions from less talkative participants, and how group size influenced the usability and effectiveness of the interfaces. These themes underscore the intricate relationship between interface design, user behavior, and the overall effectiveness of immersive meeting tools, providing valuable directions for future design and research. For participant quotes, we refer to those from the small group as P1-P15 and those from the mid-sized group as P16-P30.

## 6.1 The Re-engagement and Social Presence Trade-off

One of the most significant findings in our study was the trade-off between maintaining social engagement and efficiently keeping up with the conversation in immersive meetings. This tension was particularly evident when comparing the always-on avatar-fixed panels, EngageSync, and TableTI. The results underscore the complexity of balancing real-time awareness of the group dynamic with the cognitive demands of catching up on missed content.

Several participants expressed a preference for TableTI due to its familiarity and the convenience of viewing all the information in one place. As P3 remarked, *"I'm used to seeing everything in one place, it feels natural to me, like how I use Zoom."* However, this convenience came at the expense of social presence. The constant focus on a single, centralized panel distracted participants from the group, supporting **H2**, which predicted that avatar-fixed panels would provide higher social presence than TableTI. Participants frequently reported that their attention shifted from the group to the text panel, resulting in a loss of immersion in the meeting itself. P7 reflected, *"I liked having everything in front of me, but I realized I was just staring at the text, and at that point I didn't feel like I was in a meeting. For a second, I forgot I was in VR as well."*

This detachment is further reaffirmed by our gaze tracking data, which revealed that participants using TableTI spent significantly more time focusing on the transcription panel rather than on the avatars, compared to those using avatar-fixed panels. This shift in attention not only reduced social presence but also led to users feeling disconnected from the conversation. As P11 noted, *"I caught up with the conversation, but I didn't really feel like I was part of it."* This behavior aligns with previous research on remote meetings, where a centralized focus on text often diminishes interpersonal connection and interaction.

In contrast, the always-on avatar-fixed panels mitigated some of this issue by aligning users' gaze more closely with the group, thus enhancing social presence. However, these panels lacked the ability to provide context-sensitive summaries, which became particularly valuable in situations where users needed to re-engage after disengagement. EngageSync addressed this gap by dynamically adapting to the user's context, offering summaries that allowed participants to efficiently catch up on missed content without sacrificing their connection to the group. This finding supports **H1**, showing that

EngageSync not only improved users' ability to keep up with conversations after disruptions but also preserved social presence more effectively than both TableTI and always-on avatar-fixed panels.

The trade-off between re-engagement and social presence highlights the need for more nuanced transcription interfaces in immersive settings. While traditional interfaces like TableTI offer familiarity and comprehensive information at once, they risk alienating users from the social dynamics of the meeting. Conversely, spatially distributed interfaces like EngageSync strike a more delicate balance, enabling users to seamlessly re-engage with conversations without losing their sense of immersion and presence. This suggests that future interface designs should prioritize flexibility and adaptability, ensuring that users can maintain social connections while efficiently managing the flow of information.

### 6.2 Remembering the Least Talkative Person's Comments

Although not statistically significant, the difficulty participants experienced in recalling the least talkative person's comments provides valuable insights into how different interface designs impact attention and memory, particularly in mid-sized groups where conversational dynamics become more complex. This issue was most evident with the TableTI setup, where participants frequently shifted their focus from the group to the transcription panel, losing track of quieter voices. P5 remarked, "It was hard to keep track of what everyone was saying, especially the person who didn't speak much," highlighting how traditional table-fixed transcription panels may detract from the cohesion of social presence, drawing attention away from individual speakers.

This observation aligns with prior research on meeting inclusivity, an important concept of ensuring that all participants, regardless of how often they contribute, feel heard and visible in group discussions [21]. TableTI's design may inadvertently marginalize quieter voices by focusing users' attention on the transcription panel rather than the individuals speaking, particularly in larger groups where the flow of conversation is less predictable. This tendency to overlook less vocal participants could lead to diminished equity in the meeting, where only the most frequent contributors are remembered and acknowledged.

Conversely, the avatar-fixed transcription panels appeared to mitigate this issue by spatially anchoring the transcription to the speaking avatar, thereby reinforcing the connection between the speaker's location and their verbal contributions. P7 noted, "Having the panel attached to the avatar really helped me remember who said what. I could recall both their position and their comments more easily," demonstrating how spatialized interfaces help maintain memory continuity and social presence. This finding is consistent with prior work on spatial memory in immersive environments, where positioning information in context with the user's surroundings improves recall and focus.

The ability of avatar-fixed panels to enhance recall of even the least vocal participants has important implications for inclusivity in virtual meetings. As attention is often drawn toward more vocal participants, ensuring that quieter individuals maintain a visible and persistent presence is essential for fostering equitable participation. This spatialized transcription interface ensures that all participants, regardless of speaking frequency, are given equal representation, supporting the notion of inclusivity in meetings. In larger groups, where maintaining attention across multiple participants can be challenging, this design helps prevent the natural tendency to overlook quieter voices.

While further validation is needed, these initial findings suggest that spatialized transcription interfaces hold promise for promoting equity in meetings by distributing attention more evenly across all participants. By anchoring transcriptions to avatars, EngageSync could help foster a more inclusive environment, where every voice is remembered and valued, contributing to a more balanced and collaborative meeting dynamic. Future research could explore the long-term impact of such interfaces on meeting inclusivity, particularly in more diverse and larger group settings.

### 6.3 Re-engagement Strategies

Participants employed varied strategies to catch up on missed conversations across different interfaces. With EngageSync, two distinct approaches emerged. The majority of participants read all the summary panels before re-engaging with the ongoing discussion, while others, particularly in both small and mid-sized groups, preferred to listen to the current speaker first before referring back to the summaries. P8 explained, *"It was useful as I listened and tried to match that with what they had said while I was gone."* This strategy helped participants integrate both past and present contexts more effectively.

For those who chose to read the summaries first, some expressed a sense of urgency. P3 noted, *"Since I knew I had to read everything to return to the default mode, I tried to comprehend the summaries as fast as I could."* This sense of pressure to quickly process the summaries highlights potential areas for improvement in how the system manages the transition between summary and live conversation modes.

Participants using TableTI tended to scroll through the text panels to skim for key points before switching back to the live conversation. However, some, particularly in the mid-sized group (P18, P23, P26), reported feeling overwhelmed by the volume of text and chose to "give up" on catching up entirely, citing the difficulty of keeping up with the conversation flow.

For AvatarTI, the majority of participants (23 out of 30) opted to listen to the current speaker and attempt to fill in the gaps, rather than relying solely on summary panels. Although some still tried to catch up through the summaries, they found it less effective than simply re-engaging with the live discussion.

These findings suggest a possible future direction to accommodate more diverse re-engagement strategies for transcription interfaces.

### 6.4 Group Size Effects on Interface Effectiveness

Our findings provide strong evidence that avatar-fixed transcription tools are more effective in larger groups, supporting **H4**. Participants in the mid-sized group expressed a clear preference for avatar-fixed panels, particularly EngageSync, as they had to manage more speakers and conversation threads. The increased complexity of managing conversations in larger groups made the advantages of spatially fixed transcription panels more apparent.

In the mid-sized group, EngageSync significantly outperformed TableTI in terms of attention allocation and re-engagement time. Participants reported that transcription panels attached to avatars made it easier to stay engaged with the conversation. As P7 noted, *"With the avatar-fixed panel, I felt more in sync with the conversation—it was easier to follow who was saying what."* This spatial alignment of transcription with speakers helped users track the conversation flow and remain engaged with the group, highlighting the effectiveness of this interface in larger settings.

Cognitive load was also notably higher in the mid-sized group for TableTI, compared to both avatar-fixed interfaces. Participants emphasized the added difficulty of managing multiple speakers when using the TableTI setup. As P9 explained, *"It took extra effort to keep track of who was speaking and where they were, especially when there were so many people."* This feedback points to the cognitive burden of non-spatial transcription tools in larger groups, where the need for mentally mapping speakers to their contributions becomes more pronounced.

EngageSync also showed clear advantages in reducing re-engagement time after participants dropped out and rejoined the conversation, particularly in the mid-sized group. As P12 shared, *"I liked how the summary panels stayed up when I rejoined, making it easier to catch up without losing track of the ongoing conversation."* This demonstrates how the context-sensitive adaptation of the interface allowed participants to quickly catch up, a critical need in managing larger, more complex group conversations.

In smaller group settings, although avatar-fixed panels were still preferred over TableTI, the differences in attention allocation and cognitive load were less dramatic. In these scenarios, participants faced fewer challenges in managing conversation threads, as the group size made it easier to maintain focus on the speakers without the need for avatar-attached transcription panels. However, in larger groups, where the conversation dynamics are more complex, avatar-fixed panels—especially EngageSync, proved essential for maintaining social presence and supporting re-engagement.

## 7 LIMITATIONS AND FUTURE WORK

One limitation of this study is the fixed placement of text panels above avatars. While this placement was chosen to minimize interference, alternative configurations (e.g., left, right, or below avatars) may affect attention, cognitive load, and social presence differently. Future work could explore how these variations impact user experience in various meeting contexts.

The study also used a four-minute dropout to simulate real-world interruptions, but different types of disengagements (e.g., shorter or longer absences) could produce varying effects on re-engagement. Future studies should examine different interruption scenarios, building on previous work [48].

Our basic presentation of missed conversations aimed to prevent clutter. However, future research could improve the order and presentation of missed content, balancing clarity, cognitive load, and UI complexity.

Another limitation is the use of pre-recorded conversations in the evaluation study. As participants were passive observers, their recall and social presence may differ in live meetings. Future studies should investigate these interfaces in real-time interactive settings to assess their impact during active participation.

Future work could also improve how missed content is structured during re-engagement, possibly through adapting spatiotemporal visualizations [5]. Additionally, exploring hybrid interfaces combining the tabletop and avatar-fixed panels, as suggested by [40], could provide more flexibility.

This study focused on avatar-fixed panels to understand re-engagement and social presence, but future research should explore adaptive UI placements and modalities that allow seamless switching between interfaces. Further investigation into gesture or gaze-triggered modalities could enhance usability.

Finally, the accuracy of the ASR-to-LLM-generated summaries remains a challenge. While this study focused on how to display the summaries, future work could address the issue of improving summary generation.

## 8 CONCLUSION

In this paper, we introduced EngageSync, a context-aware avatar-fixed panel, and demonstrated its effectiveness in enhancing both social presence and information recall in immersive meetings. Our study showed that EngageSync outperformed traditional Tabletop and always-on Avatar-fixed panels, particularly in mid-sized groups, where maintaining engagement and catching up after disruptions posed greater challenges. These results reinforce the findings from our formative study, which identified the need for context-sensitive transcription methods in VR environments.

The simplicity of adapting transcription panels based on user engagement is key to EngageSync's effectiveness. Our findings suggest that by providing live transcriptions and summaries when necessary, users are better equipped to re-engage with ongoing discussions without sacrificing social presence. This adaptive approach can be seamlessly integrated into current VR meeting platforms, providing a user-friendly solution to a persistent challenge in immersive meetings.

Our work also opens up new avenues for exploration in the broader HCI field. Future research could examine how context-aware interfaces like EngageSync can be further refined or expanded. For example, combining avatar-fixed panels with more flexible, user-controlled interactions, or exploring the role of different panel placements, could yield even more efficient designs. We hope this study inspires further innovation in adaptive VR interfaces, promoting more effective and natural interactions in immersive environments.

## REFERENCES

[1] [n. d.]. Avatar Recording Tool , author = Meta Platforms, Inc , year = 024, howpublished https://www.meta.com/experiences/avatar-recording/6145925042141671/.
[2] [n. d.]. Spatial - 3D Workspaces to Meet and Collaborate. https://spatial.io/. Accessed: 2024-09-11.
[3] Katarzyna Abramczuk, Zbigniew Bohdanowicz, Bartosz Muczyński, Kinga H Skorupska, and Daniel Cnotkowski. 2023. Meet me in VR! Can VR space help remote teams connect: a seven-week study with horizon workrooms. *International Journal of Human-Computer Studies* 179 (2023), 103104.
[4] Judith Amores, Xavier Benavides, and Pattie Maes. 2015. Showme: A remote collaboration system that supports immersive gestural communication. In *proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*. 1343–1348.
[5] Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. 2003. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages*

*& Computing* 14, 6 (2003), 503–541.

[6] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Generating abstractive summaries from meeting transcripts. In *Proceedings of the 2015 ACM Symposium on Document Engineering*. 51–60.

[7] John Brooke. 1996. SUS: A "quick and dirty" usability scale. *Usability Evaluation in Industry* (1996), 189–194.

[8] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla NY Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large scale analysis of multitasking behavior during remote meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[9] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. MeetScript: Designing Transcript-based Interactions to Support Active Participation in Group Video Meetings. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.

[10] Ersin Dincelli and Alper Yayla. 2022. Immersive virtual reality in the age of the Metaverse: A hybrid-narrative review based on the technology affordance perspective. *The journal of strategic information systems* 31, 2 (2022), 101717.

[11] J George, A Mirsadikov, M Nabors, and Kent Marett. 2022. What do users actually look at during videoconference calls? Exploratory research on attention, distraction effects and gender. In *Proceedings of the 55th Hawaii international conference on system sciences*. 4779–4787.

[12] Sarthak Ghosh, Lauren Winston, Nishant Panchal, Philippe Kimura-Thollander, Jeff Hotnog, Douglas Cheong, Gabriel Reyes, and Gregory D Abowd. 2018. Notifivr: Exploring interruptions and notifications in virtual reality. *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1447–1456.

[13] Google. 2023. Cloud Speech-to-Text API. https://cloud.google.com/speech-to-text. Accessed: 2024-09-12.

[14] Matt Gottsacker, Nahal Norouzi, Kangsoo Kim, Gerd Bruder, and Greg Welch. 2021. Diegetic representations for seamless cross-reality interruptions. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 310–319.

[15] Camila R Guetter, Maria S Altieri, Marion CW Henry, Elizabeth A Shaughnessy, Sadia Tasnim, R Yu Yangyang, and Sanda A Tan. 2022. In-person vs. virtual conferences: Lessons learned and how to take advantage of the best of both worlds. *The American Journal of Surgery* 224, 5 (2022), 1334–1336.

[16] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, Vol. 2004. Universidad Politecnica de Valencia Valencia.

[17] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). North-Holland, Amsterdam, 139–183.

[18] 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.

[19] Fernanda Herrera, Soo Youn Oh, and Jeremy N Bailenson. 2020. Effect of behavioral realism on social interactions inside collaborative virtual environments. *Presence* 27, 2 (2020), 163–182.

[20] Jari K Hietanen. 2018. Affective eye contact: An integrative review. *Frontiers in psychology* 9 (2018), 1587.

[21] Yasaman Hosseinkashi, Lev Tankelevitch, Jamie Pool, Ross Cutler, and Chinmaya Madan. 2024. Meeting effectiveness and inclusiveness: large-scale measurement, identification of key features, and prediction in real-world remote meetings. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–39.

[22] Ryo Iijima, Akihisa Shitara, Sayan Sarcar, and Yoichi Ochiai. 2021. Word cloud for meeting: A visualization system for dhh people in online meetings. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[23] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the benefit of highlighting key words in captions for people who are deaf or hard of hearing. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 43–55.

[24] Niclas Kaiser, Kimberly Henry, and Hanna Eyjólfsdóttir. 2022. Eye contact in video communication: Experiences of co-creating relationships. *Frontiers in Psychology* 13 (2022), 852692.

[25] Katherine A Karl, Joy V Peluchette, and Navid Aghakhani. 2022. Virtual work meetings during the COVID-19 pandemic: The good, bad, and ugly. *Small group research* 53, 3 (2022), 343–365.

[26] Kangsoo Kim, Celso M de Melo, Nahal Norouzi, Gerd Bruder, and Gregory F Welch. 2020. Reducing task load with an embodied intelligent virtual assistant for improved performance in collaborative decision making. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 529–538.

[27] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* 2 (2022), 786665.

[28] Geonsun Lee, HyeongYeop Kang, JongMin Lee, and JungHyun Han. 2020. A user study on view-sharing techniques for one-to-many mixed reality collaborations. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 343–352.

[29] Geonsun Lee, Dae Yeol Lee, Guan-Ming Su, and Dinesh Manocha. 2024. "May I Speak?": Multi-modal Attention Guidance in Social VR Group Conversations. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[30] Minha Lee, Wonyoung Park, Sunok Lee, and Sangsu Lee. 2022. Distracting moments in videoconferencing: A look back at the pandemic period. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[31] Joshua McVeigh-Schultz and Katherine Isbister. 2021. The case for "weird social" in VR/XR: a vision of social superpowers beyond meatspace. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–10.

[32] Joshua McVeigh-Schultz and Katherine Isbister. 2022. A "beyond being there" for VR meetings: envisioning the future of remote work. *Human–Computer Interaction* 37, 5 (2022), 433–453.

[33] Meta Platforms, Inc. 2021. Meta Avatars SDK. https://developer.oculus.com/downloads/package/meta-avatars-sdk. Accessed on September 10, 2024.

[34] Meta Platforms, Inc. 2021. Meta Horizon Workrooms. https://www.meta.com/experiences/meta-horizon-workrooms-beta/2514011888645651/. Accessed: 2024-09-11.

[35] Meta Platforms, Inc. 2024. Use your computer in VR in Meta Horizon Workrooms. https://www.meta.com/help/quest/articles/horizon/getting-started-in-horizon-workrooms/use-computer-in-VR-workrooms/. Accessed: 2024-09-12.

[36] Mozilla. 2018. Mozilla Hubs. https://hubs.mozilla.com/. Development ended May 31, 2024.

[37] 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.

[38] OpenAI. 2023. GPT-4 Turbo. https://openai.com/gpt-4. Announced at OpenAI DevDay.

[39] Hyanghee Park, Daehwan Ahn, and Joonhwan Lee. 2024. Lessons From Working in the Metaverse: Challenges, Choices, and Implications from a Case Study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[40] Siyou Pei, David Kim, Alex Olwal, Yang Zhang, and Ruofei Du. 2024. UI Mobility Control in XR: Switching UI Positionings between Static, Dynamic, and Self Entities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2019. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers in Robotics and AI* 6 (2019), 5.

[42] 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*. 1–16.

[43] Huajian Qiu, Paul Streli, Tiffany Luong, Christoph Gebhardt, and Christian Holz. 2023. ViGather: Inclusive Virtual Conferencing with a Joint Experience Across Traditional Screen Devices and Mixed Reality Headsets. *Proceedings of the ACM on Human-Computer Interaction* 7, MHCI (2023), 1–27.

[44] Radiah Rivu, Yasmeen Abdrabou, Ken Pfeuffer, Augusto Esteves, Stefanie Meitner, and Florian Alt. 2020. Stare: gaze-assisted face-to-face communication in augmented reality. In *ACM Symposium on Eye Tracking Research and Applications*. 1–5.

[45] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. 2019. Notification in vr: The effect of notification placement, task and environment. In *Proceedings of the annual symposium on computer-human interaction in play*. 199–211.

[46] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] Helen B Schwartzman. 1989. *The meeting*. Springer.

[48] Seoyun Son, Junyoug Choi, Sunjae Lee, Jean Y Song, and Insik Shin. 2023. It is okay to be distracted: how real-time transcriptions facilitate online meeting with distraction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[49] Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the effects of reading interruptions by providing reviews and previews. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[50] Frank Steinicke, Nale Lehmann-Willenbrock, and Annika Luisa Meinecke. 2020. A first pilot study to compare virtual group meetings using video conferences and (immersive) virtual reality. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction*. 1–2.

[51] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The effect of multimodal emotional expression and

agent appearance on trust in human-agent interaction. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 1–6.

[52] Simon Tucker, Ofer Bergman, Anand Ramamoorthy, and Steve Whittaker. 2010. Catchup: a useful application of time-travel in meetings. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 99–102.

[53] Sunday D Ubur, Naome A Etori, Shiva Ghasemi, Kenneth King, Denis Gračanin, and Maria Gini. 2024. EasyCaption: Investigating the Impact of Prolonged Exposure to Captioning on VR HMD on General Population. In *International Conference on Human-Computer Interaction*. Springer, 382–403.

[54] Alexander Vedernikov, Zhaodong Sun, Virpi-Liisa Kykyri, Mikko Pohjola, Miriam Nokia, and Xiaobai Li. 2024. Analyzing Participants' Engagement during Online Meetings Using Unsupervised Remote Photoplethysmography with Behavioral Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 389–399.

[55] VRChat Inc. 2014. VRChat. https://hello.vrchat.com/.

[56] Portia Wang, Mark R Miller, Anna CM Queiroz, and Jeremy N Bailenson. 2024. Socially Late, Virtually Present: The Effects of Transforming Asynchronous Social Interactions in Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[57] Steve Whittaker, Simon Tucker, Kumutha Swampillai, and Rachel Laban. 2008. Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing* 12 (2008), 197–221.

[58] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009