“May I Speak?”: Multi-modal Attention Guidance in Social VR Group Conversations

Geonsun Lee, Dae Yeol Lee, Guan-Ming Su, and Dinesh Manocha

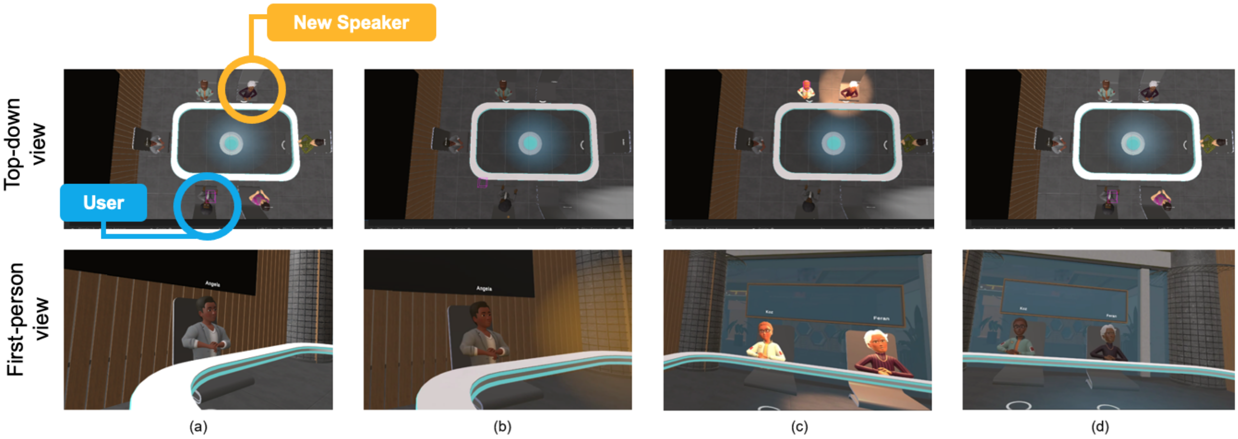


Fig. 1: We present a diegetic multi-modal attention guidance method designed for group conversations in social VR. The top row illustrates a top-down view of a virtual conference room, while the bottom row presents the first-person view. When (a) an agent (target) positioned out of the user's view signals a desire to speak; (b) The environment light dims, a point light affixed to the user's right periphery of the screen instantiates, and a spatial sound effect with a chiming sound commences; (c) As the user turns their head towards the target, the environment light brightens, the point light vanishes, a spotlight illuminates the target, and the chiming sound emanates from the target's location; (d) Once the user's gaze aligns with the target, the spotlight and chiming sound dissipate. Our approach demonstrated statistically significant enhancements in users' satisfaction with the conversation, faster response times, and higher preference compared to conventional methods.

Abstract—In this paper, we present a novel multi-modal attention guidance method designed to address the challenges of turn-taking dynamics in meetings and enhance group conversations within virtual reality (VR) environments. Recognizing the difficulties posed by a confined field of view and the absence of detailed gesture tracking in VR, our proposed method aims to mitigate the challenges of noticing new speakers attempting to join the conversation. This approach tailors attention guidance, providing a nuanced experience for highly engaged participants while offering subtler cues for those less engaged, thereby enriching the overall meeting dynamics. Through group interview studies, we gathered insights to guide our design, resulting in a prototype that employs light as a diegetic guidance mechanism, complemented by spatial audio. The combination creates an intuitive and immersive meeting environment, effectively directing users' attention to new speakers. An evaluation study, comparing our method to state-of-the-art attention guidance approaches, demonstrated significantly faster response times ($p < 0.001$), heightened perceived conversation satisfaction ($p < 0.001$), and preference ($p < 0.001$) for our method. Our findings contribute to the understanding of design implications for VR social attention guidance, opening avenues for future research and development.

Index Terms—Social VR, Attention Guidance, Multi-modal Interaction, Group Conversations, Turn-taking

1 Introduction

Social virtual reality has gained traction as a compelling solution for enabling people to interact within shared virtual environments with head-mounted displays (HMDs). This field encompasses an array of commercially available applications, including VRChat, Spatial, Rec-Room, Mozilla Hubs, Glue, Horizon Worlds, and others. While initially geared towards casual socializing, these platforms have progressively evolved to cater to diverse use cases, extending into professional domains such as conferences and business-focused meetings [49]. The ability of VR to provide an immersive and shared spatial experience adds a layer of depth to interactions, setting the stage for a paradigm shift in remote collaboration and communication.

In the context of remote communication, the issue of *turn-taking* has emerged as a persistent challenge in effectively supporting remote meetings. This challenge, which has been the subject of scholarly inquiry for decades [54, 61], has become especially pronounced during the widespread adoption of remote work necessitated by the COVID-19 pandemic [45]. The intricacies of managing turn-taking in the context of video-mediated communication have become a focal point, with numerous contributing factors [5, 11, 41]. The inherent difficulty in discerning non-verbal cues, such as gestures and head/body movements, leads to disruptions and inefficiencies in the communication flow.

While VR offers a promising alternative by simulating face-to-face meetings and providing cues based on spatial layout and proximity [62, 63, 65], it needs to deal with many other challenges. The limitations of immersive VR, such as a restricted field of view and imperfections in face and gesture tracking, pose unique obstacles that impact

- *Geonsun Lee and Dinesh Manocha are with University of Maryland, College Park. E-mail: [gsunlee|dmanocha]@cs.umd.edu.*
- *Dae Yeol Lee and Guan-Ming Su are with Dolby Laboratories. E-mail: [DaeYeol.Lee|guanming.su]@dolby.com.*

Manuscript received 4 October 2023; revised 17 January 2024; accepted 24 January 2024. Date of publication 7 March 2024; date of current version 15 April 2024.
Digital Object Identifier no. 10.1109/TVCG.2024.3372119

the effective identification of new speakers in a conversation. Additionally, individuals with conditions such as autism or those with reduced social skills may face difficulties participating in social group conversations [16, 58]. Despite VR's potential to address these challenges by serving as a training ground for social skill development [2, 28] and enabling novel interactions that are beyond simply imitating reality [23, 52], there remains an unmet need for comprehensive solutions that facilitate natural signaling of the desire to speak and enhance the perceptibility of such signals by other users.

Main Results: In this paper, we present improved methods related to improving user experiences related to VR turn-taking. Our goal is to facilitate the recognition of new speakers in a mid-sized group conversation, thereby guiding users' attention effectively. Our approach uses attention guidance methods, which have been extensively studied in the context of VR, and seek to direct users' attention with rapid responsiveness while ensuring minimal disruption to their immersion [30]. To this end, we introduce a diegetic multi-modal attention guidance approach that utilizes both lights and spatial audio within the virtual meeting environment. Our approach is designed to enhance the user experience by enabling the identification of new speakers, all while maintaining the seamless flow of ongoing conversations and sustaining immersion within the VR environment. Additionally, our formulation strives to enhance social presence within these virtual interactions.

Our methodology involves a group interview study with experienced users in VR meeting platforms to elucidate key design considerations. Based on these considerations, we propose an approach where attention guidance dynamically adjusts based on a user's engagement in an ongoing conversation. We present various components of our approach and illustrate their interactions within the virtual environment.

Subsequently, we conduct a user study simulating participants engaging in conversations within a social VR setting, employing avatars with pre-recorded interactions. We perform an evaluation corresponding to when a new virtual speaker signals their desire to speak in two scenarios: (i) while the user is already engaged in conversation and (ii) while the user is in listening mode. We compare our approach against two existing methods: a text window and icon indication used in the VR meeting application Horizon Workrooms, and Subtle Gaze Direction (SGD) [17]. Our evaluation results demonstrate the effectiveness of our method in guiding users' attention, positively impacting perceived conversation quality, and high preference. Our novel contributions include:

- Insights into the design of attention guidance methods for social VR derived from a comprehensive study involving user group interviews with experienced participants.
- The novel concept of an engagement-based attention guidance approach for turn-taking in social VR group conversations.
- A prototype of a diegetic multi-modal attention guidance method utilizing light and spatial audio within the virtual environment.
- Highlighting the significant impact of our multi-modal attention guidance approach on response time, perceived conversation quality, and achieving the highest overall preference when compared to traditional methods.

2 Related Work

2.1 Social VR

Social virtual reality (Social VR) has recently gained considerable attention from the human-computer interaction (HCI) and VR communities [27] exploring spatial navigation and social mechanics [37, 40]. A longitudinal study by Moustafa et al. [40] revealed the transferability of existing social group dynamics to VR interactions. Other studies have delved into social interactions on Social VR platforms, addressing challenges such as mitigating harmful behavior [27, 36].

Prior studies have demonstrated that self-embodiment and nonverbal cues play pivotal roles in establishing social presence within social VR applications [64]. To enhance user experiences in social interactions, endeavors have been made to amplify these non-verbal cues [12]. Gaze direction emerges as a prominent method for social signaling, facilitating the seamless transition of users from the awareness of others' presence to interactive engagement. Consequently, researchers have been actively engaged in the development of gaze-tracking VR solutions [3, 64]. Roth et al.'s work [46] involved augmenting social behaviors within a multi-user virtual museum environment. Their experimentation included visualized eye contact represented by bubbles, highlighted joint attention through visual cues, and color-coded group affiliations.

2.2 Social Attention in Group Conversations

Turn-taking describes the dynamic flow of participation among speakers in a conversation over time. Conversations involve a constant reshuffling of participation, and it is defined that participation shifts occur in four different types: *turn-receiving, turn-claiming, turn-usurping, and turn-continuing* [15]. When a person speaks after being addressed, it is termed turn-receiving; if a person speaks after someone else is addressed, it is called turn-usurping. When a person speaks after someone addresses the group, it is termed turn-claiming. Finally, when someone who is already talking changes targets, it is termed turn-continuing. While turn-taking in conversation is often unconstrained and unplanned, in institutional settings, the system has been described as more restricted and specialized [21].

Common indicators of the desire to speak in a group include raising hands and nodding heads. In some settings, participants need to be selected by a moderator and queued to be the next speaker [39]. In face-to-face conversations, the next speaker is usually indicated by social attention or eye contact [22, 31, 35]. Dawson and Foulsham [9] investigated how shifts in attention between speakers depend on visual or auditory cues. They found that eye-tracked participants often fixated on the person speaking and shifted their gaze in response to changes in the speaker, even when sound was removed or the video was freeze-framed.

The underlying medium used also influences conversation patterns and social attention [1]. In video conferencing, users can use the 'raise-hand' feature to signal their intention to speak and wait to be addressed by others in the group [45]. Hu et al. [24] developed OpenMic, an interface that visualizes conversational floor transitions by incorporating proxemic metaphors in a videoconferencing system. Steptoe et al. [59] introduced one of the first avatar gaze tracking systems and provided preliminary evidence that it improves communication. In VR, Li et al. [32] introduced a shared VR environment visualization to aid in conversational turn-taking, employing cylinders that expand over time to represent the duration of each speaker's turn, thereby facilitating balanced participation in the conversation.

Within the realm of turn-taking, our approach focuses on the participation shifts of turn-claiming and turn-usurping. This occurs when users are perceived to primarily use the 'raise-hand' feature or when the person who wants to speak needs to grab the social attention of the group because they are not directly addressed to speak.

2.3 Guiding User's Attention

The exploration of guiding user attention spans across diverse mediums, from the realm of 2D images to the immersive experiences of VR and AR [47]. The aim is to steer users toward specific focal points intended by creators. In the VR community, users often confront the challenge of significant scene details lying beyond their field of view, instigating a concern of missing out on crucial elements [34, 60].

A technique to address this involves direct modifications to images. Smith et al. [56] employed a combination of blurred and non-blurred areas in videos, directing viewers towards regions with minimal spatial blur when the rest of the image is intentionally blurred. Hata et al. [19] explored thresholds where blur effects can be subtly applied, effectively guiding users visually. Stylistic rendering, encompassing control over depth of field, colors, brightness, and sharpness, represents another avenue. Additionally, the simulation of brightness contrast through lights, a technique well-established in film, has been leveraged for effective visual attention [8]. El et al. [10] implemented ALVA (Adaptive Lighting for Visual Attention), dynamically adjusting lighting color and brightness to enhance visual attention within gaming environments.

A prevalent category in attention guidance involves diegetic methods [42, 48, 55], where cues seamlessly integrate into the scene, encompassing elements like characters, lights, or sounds. This concept, rooted in film theory as diegesis, entails elements belonging to the narrative world. Diegetic visual cues, such as a person looking in a certain direction, significantly influence the viewer’s gaze, as observed with moving objects [55]. Auditory cues within the diegetic framework leverage sound to prompt users to search for the source, prompting a change in their viewing direction [48]. Noteworthy for their subtlety, diegetic cues afford viewers the freedom to follow them naturally. More recently, Lange et al. [30] introduced Hive Five, a particle effect emulating the swarm motion of bees as a cursor, showcasing its ability to guide users’ attention while preserving immersion.

Various techniques have been explored to accentuate crucial content in VR, ranging from methods utilizing different shapes like arrows [7, 33] to halos [18]. Subtle Gaze Direction (SGD), leveraging eye tracking, subtly guides users’ gaze without their explicit awareness [4]. This method modulates a target region in the peripheral area, encouraging viewers to direct their gaze while discontinuing modulation when the viewing direction aligns to a certain degree. Luminance modulation and warm-cool modulation were developed, employing flickering at 10 Hz within a circular region. Grogorick et al. [17] later adapted this method for VR environments.

Multi-modal attention guidance represents an advanced approach. Reyes et al. [44] developed a guiding method integrating both visual and auditory cues, with their study demonstrating the superior efficacy of incorporating an auditory cue alongside visual cues.

While a myriad of methods exists, none have systematically addressed their adaptation to social meeting settings, characterized by pre-existing social attention dynamics and diverse individual social engagements. Our method takes these nuanced considerations into account, dynamically adjusting the intensity and components of the attention guidance framework presented to the user.

2.4 Notifications in VR/AR

A notification is a proactive delivery of information to users through visual, auditory, or haptic alerts designed to attract their attention [26,43]. Ghosh et al. [14] investigated notifications in VR, comparing visual, audio, and haptic modalities, as well as their pairwise combinations. The study found that both audio and haptic modalities effectively elicited reactions from participants in VR. However, haptic notifications faced challenges such as confusion or being missed due to interactions with existing objects in the VR environment. In a related study, George et al. [13] conducted an exploratory lab study comparing three notification types: text, spotlight, and global light. Text notifications prompted quick responses but exhibited the lowest presence, while ambient light showed the lowest attention-grabbing but the highest presence. This trade-off highlights the importance of considering both aspects in notification design. To minimize disruption, Chen et al. [6] identified opportune times for delivering notifications in VR, allowing for their scheduled presentation. Rzayev et al. [51] investigated efficient notification presentation in VR by comparing different placements of notifications in various tasks. They also explored the position of notifications in AR glasses and how they would be perceived during face-to-face communication [50].

It’s noteworthy that existing works on notifications in VR primarily focus on conveying information from the external world to users who are obscured by wearing a VR HMD. Similarly, our work aligns with the concept of notifications as we aim to convey information to users engaged in tasks that are susceptible to "interruption." Users may revisit a signal indicating a new speaker’s intention to speak, similar to the way notifications are revisited after initially noticing them once the current speaker is done.

3 Formative Study: Group Interview

Our primary goal was to precisely identify the issue at hand and develop a corresponding interaction strategy. We conducted a group interview study, focusing on two main objectives: firstly, to pinpoint the gap between user needs for turn-taking in social VR and current attention guidance methods; and secondly, to determine the desired features for attention guidance in this context. This study involved seven experienced VR users, all actively participating in VR meetings.

3.1 Participants

For the group interview, we conducted a pre-screening process and selected participants with substantial VR application experience and engagement in regular or irregular VR meetings. Seven participants were recruited, each with 2 to 10 years of VR experience and over a year of professional VR meeting attendance. The experienced meeting sizes ranged from small (up to 5 attendees) to medium-sized groups (5 to 14 attendees). Participants had prior exposure to VR platforms such as Spatial (5 participants) and Horizon Workrooms (3 participants), along with others like Glue or BigScreen VR. All had access to an Oculus Quest 2 or Oculus Quest Pro.

3.2 Study Setup and Procedure

3.2.1 Prototype of Existing Attention Guidance Methods

We captured videos of prototypes featuring various existing attention guidance methods within a VR meeting setting. Four methods from previous work were implemented: arrow, SGD [4], Hive Five [30], and conventional text-based notifications akin to video conference platforms. We use the Unity game engine for developing the prototype, with each video lasting around 15 seconds. These videos showcased the user looking at a virtual avatar, the application of the attention guidance method, and the user’s view being directed to an avatar on the far right. Due to computational constraints faced by some participants, we presented existing methods using videos, ensuring uniform exposure and circumventing technical limitations. This approach facilitated immediate, collaborative discussion and analysis of specific method aspects with visual aids.

3.2.2 Procedure

The group interview was conducted in two sessions to accommodate participant schedules, with three and four participants per session. The interviews took place remotely via Meta Horizon Workrooms to introduce newcomers to Horizon Workrooms’ features and enable instant demonstration of comments and ideas within VR. The Workrooms’ layout was set to “Meeting." Participants were given a 5- to 10-minute platform orientation before the actual interview to prevent distractions during the session. The interview comprised three phases: (1) Initial questions focused on general VR meeting experiences, benefits and limitations compared to in-person and video calls, methods of grabbing attention to speak, instances of cues being missed, and hand-raising behavior; (2) Prototype videos of existing attention guidance methods were presented using screen share within Horizon Workrooms. This was followed by a try-out of Horizon Workrooms’ “raise hand" feature, which participants explored within different seat positions for 3 to 5 minutes; (3) Participants shared opinions on the methods from phase two, suggested additional features, and brainstormed. The study duration was approximately 75 minutes.

3.3 Findings

Positive Aspects of VR Meetings Participants in the interview highlighted intuitive interactions, like head-turning to see others and discerning directional sounds, which enhance the "social presence" in VR meetings, setting them distinctly apart from traditional video calls. They also noted the benefits of customizable environments tailored to attendee count and meeting type, and an increased focus enforced by wearing VR headsets, offering clear advantages over in-person meetings.

Limitations in VR Meetings: Despite their strengths, current VR-based meetings have many limitations. Foremost among these is the limited field of view in VR headsets, which impairs peripheral awareness and the ability to notice distant users and their nonverbal cues for speaking. While recent advancements have improved body and facial tracking, participants noted a disparity between real-world gestures and their VR counterparts, resulting in less expressive and sometimes ambiguous non-verbal cues. Technical issues like network lag and tracking errors disrupt the fluidity of conversation, causing significant inconvenience.

Social Attention in VR Meetings: Drawing parallels to in-person meetings, participants signaled their intent to speak in similar ways: i) emitting sounds like throat clearing or table tapping, and ii) using gestures such as head turns, nods, or virtual hand raises. Notably, participants highlighted scenarios where such cues are missed. Instances included complex discussions that deter opportune contributions, participants positioned out of others' views while concentrating on someone else, and challenges in noticing others within larger groups. These turn-taking issues were particularly prominent in formal meetings where participants are less acquainted and no dedicated moderator exists.

Feedback on Existing Attention Guidance Methods: Participant feedback on the presented attention guidance methods within a social context was strikingly uniform. A consensus emerged among participants that the demonstrated methods proved overly distracting. The utilization of screen-fixed UI elements like arrows or text-based notifications, despite their intention to guide attention, was deemed overly intrusive by most. Among the four methods, SGD was acknowledged as the most subtle, yet its constant flickering in the user's peripheral vision was deemed distracting. One participant likened the SGD's flickering effect to an "alert" rather than a guidance mechanism, attributing this to its design for on-screen visual targets rather than off-screen objects as is common in social VR. HiveFive's swarm motion was considered distracting and out of sync with the meeting room environment, undermining its usability. Participants collectively expressed that existing methods prioritize directing attention to a target, overlooking the nuances of group conversations.

As for the "raise hand" feature in Horizon Workrooms, participants acknowledged its usefulness mainly because they found it similar to those in video-conference platforms like Zoom or Microsoft Teams. However, they noted its potential to be overlooked, especially in larger groups or by those lacking social skills to naturally draw attention.

Desired Characteristics in Social Attention Guidance: Participants' brainstorming suggestions revealed recurring themes aligning with their preferred attention guidance characteristics, emphasizing context-dependent effectiveness. For example, informal gatherings might not require guidance, in contrast to formal or large meetings without a clear moderator. After transcribing and coding these inputs, we identified five key characteristics, informed by prior discussions on VR meetings and existing attention guidance methods.

- **Diegetic:** Participants advocated for attention guidance cues that seamlessly integrate into the environment, avoiding excessive user distraction.
- **World-Referenced:** Echoing the "diegetic" principle, participants preferred cues that are fixed to the virtual world's elements rather than to the screen.
- **Subtlety:** Desired descriptors for the method included "subtle," "ignorable," and minimally disruptive. Interviewees stressed this quality's importance for speakers, listeners, and those interjecting, with a balance between unobtrusiveness and efficient signaling seen as crucial.
- **Control of Urgency:** Four participants articulated a desire for nuanced control over the degree of attention they receive. They envisioned situations where they might prefer to go unnoticed or, conversely, urgently seek recognition based on the content of their contribution. This nuanced approach contrasts with existing "raise hand" features that typically represent user intention in binary terms.
- **Multi-Modality:** Participants consistently highlighted the significance of audio cues. One participant underscored that VR meetings stand out due to their spatial audio dimension, which aids users in identifying sound direction and speaker location. Participants proposed augmenting this auditory spatial awareness with corresponding visual cues, such as modulating a speaker's volume or introducing non-intrusive chime sounds from the new speaker's direction. Vocal narration was generally discouraged due to its potential to disrupt the conversation.

4 Multi-modal Attention Guidance for Social VR

Our multi-modal attention guidance method for Social VR, informed by a literature review in group conversation psychology and insights from the formative study (Section 3), integrates key factors to determine user engagement in group conversations.

Engagement, defined as the degree of active involvement in the conversation, also reflects perceived social proximity to new speakers. For example, a participant deeply engaged in a dialogue might have a higher social distance, requiring more effort for a new speaker to attract their attention.

It is important to note that in our method, the term 'user' refers to a participant already engaged in the conversation (speaker or listener), distinct from the 'new speaker' who is attempting to interject. The method employs several parameters for gauging engagement: (1) New Speaker Coordinate, representing the 3D position of a participant intending to interject, defined in the world coordinate system; (2) Head-Body Rotation, reflecting the user's head position; and (3) Gaze Direction, indicating the user's viewing direction.

These parameters form the foundation of our heuristic approach for determining the optimal level of guidance intrusiveness. Utilizing these, we developed two key modules: the **Light Manipulator Module** and the **Spatial Audio Control Module**.

4.1 Light Manipulator Module

Following the design characteristics drawn from the group interview study, we chose to adopt lighting effects as the core element of our attention guidance method. This choice was informed by several factors: (1) the ubiquitous nature of light in any environment, lending a diegetic quality to its usage, and (2) the potential for nuanced control of light effects, enabling us to finely tune the subtlety of the effect.

We manipulate three types of lighting sources in our approach: environmental light, point light, and spotlight, each strategically used for out-of-view and within-view attention guidance scenarios. This manipulation involves precise control over parameters such as intensity and color warmth, using the Unity Light object's parameters for accurate integration. We define 'out-of-view' as when the target is outside the user's viewport angle and 'within-view' as the opposite.

Environment Light: Environment light management involves dimming its brightness and restoring original luminance upon user acknowledgment of the new speaker, applicable in both out-of-view and within-view scenarios. This process entails setting the environment light's minimum intensity to a pre-defined parameter, based on the desired subtlety, with the unaltered brightness as the maximum intensity.

When a new speaker signals their intention to speak, the environmental light's intensity decreases over two seconds to a set minimum, ensuring a subtle, non-disruptive environmental shift for the user.

At the same time, angular ranges $[\theta_{min}, \theta_{max}]$ are determined for adaptively controlling light brightness. Here, θ_{max} represents the angular deviation from the user's gaze direction to the new speaker's coordinate at the signal's moment, and θ_{min} denotes the aligned angle. Ideally, θ_{min} would be 0, but we allow non-zero magnitude for flexibility.

Let θ be the angle between the user's current gaze direction and the gaze direction at the moment the new speaker's signal is received. The environment light brightness L adjusts according to the function $f^{env}(\theta)$, as shown below:

$$L = L_{\min} + (L_{\max} - L_{\min}) \cdot \left(\frac{\min(\theta_{\max}, \max(\theta, \theta_{\min}))}{\theta_{\max} - \theta_{\min}} \right)^{\gamma} \quad (1)$$

where γ is a curvature parameter adjusting the intensity change rate relative to θ (note that $\gamma > 0$). Figure. 2-(a) depicts how L changes in accordance with θ. The min and max functions in Equation 1 ensure the brightness remains within desired limits, preventing excessive darkening or brightening outside the range $[\theta_{min}, \theta_{max}]$. This mechanism maintains appropriate lighting even if the user's gaze deviates significantly, supporting ongoing conversation engagement.

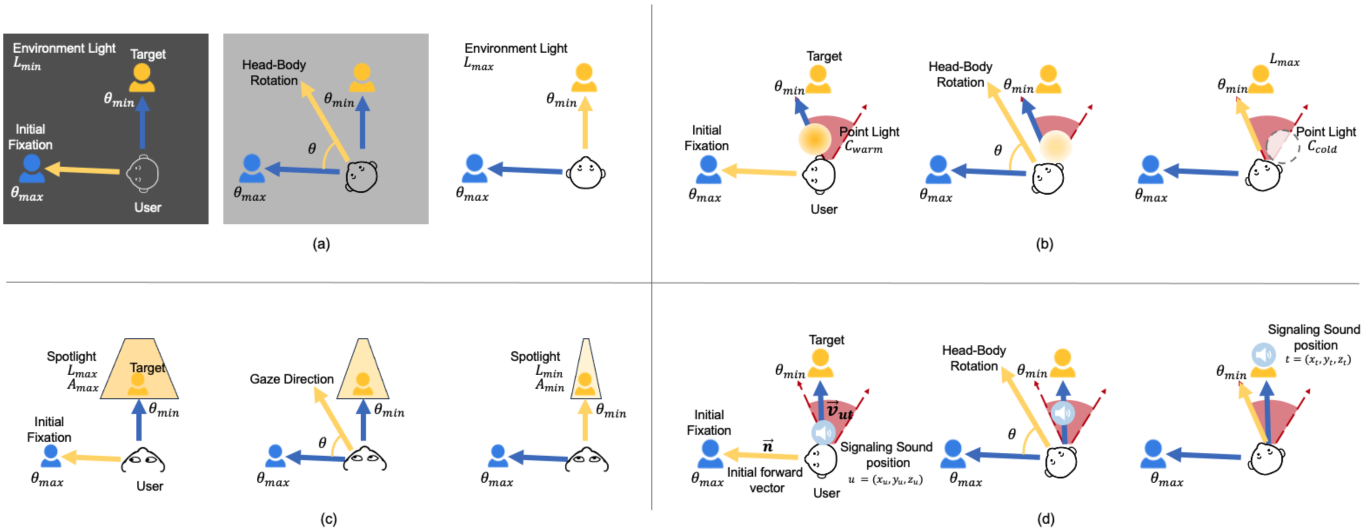
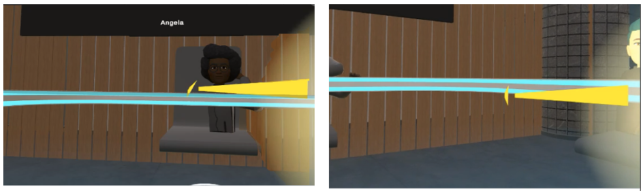


Fig. 2: Progressive adjustments in the light manipulator and spatial audio control module components relative to the angular distance (θ) between the user’s head-body rotation or gaze direction and the new speaker’s coordinate, with subfigures (a) environment light, (b) point light, (c) spotlight, and (d) signaling sound demonstrating the range from maximum to minimum angular thresholds. The area colored in red represents θ_{viewport}, the range that the new speaker coordinates is within the user’s viewport.



(a) (b)

Fig. 3: The point light gradually interpolates from (a) a warm color(yellow) to (b) a cold color (white) as θ gets closer to θ_{min}

Point Light: When the new speaker is entirely outside the user’s field of view, offering a directional cue becomes imperative to guide the user’s head orientation. To fulfill this purpose, we utilize a point light affixed to the user’s head position, placed at 75° to remain within peripheral vision. The point light’s direction, left or right, is contingent on the new speaker’s position, offering balanced guidance. We further adjust the color warmth (C) of the point light depending on the angular deviation (θ) from the target, following the function:

$$C = C_{\text{org}} + (C_{\text{warm}} - C_{\text{cold}}) \cdot \left(\frac{\min(\theta_{\max}, \max(\theta, \theta_{\min}))}{\theta_{\max} - \theta_{\min}} \right)^{\gamma} \tag{2}$$

Influenced by the visual attention model [10], which favors warm colors for attracting attention, we vary the RGB value of the point light based on the angular divergence between the user’s current view and the target. This transition from a ‘warm’ to a ‘cold’ color is contingent on θ. Note that our ‘warm’ color (C_{warm}) and ‘cold’ color (C_{cold}) were heuristically set as ‘yellow’ and ‘white’. The point light deactivates when the new speaker coordinate is within the user’s viewport. The parameter γ controls the rate of change in color warmth with respect to θ, with $\gamma = 1$ as the default setting for a linear relationship, though it can be adjusted as needed. As depicted in Figure. 2-(b), a higher θ yields a warmer point light, which decreases in warmth as the user’s gaze aligns with the target. The area colored in red, θ_{viewport}, represents when the new speaker coordinate is within the user’s viewport, in which the point light deactivates.

Spotlight: The spotlight only activates when the target is within the user’s viewport. Hence, here θ is defined as the angular deviation of the gaze direction, not the head-body rotation. The angle of the spotlight’s cone and the intensity of the brightness will dynamically adjust based on the user’s gaze direction relative to the new speaker coordinate. Regarding the intensity control, the same angular distance equation as described in Equation. 1 is applied. The cone’s angle (A) defines the size of the area in which the spotlight covers and can be formulated as a function of θ. The equation is as follows:

$$A = A_{\min} + (A_{\max} - A_{\min}) \cdot \left(\frac{\min(\theta_{\max}, \max(\theta, \theta_{\min}))}{\theta_{\max} - \theta_{\min}} \right)^{\gamma} \tag{3}$$

When the user’s gaze is directed further away from the new speaker coordinate (θ_{max}), the brightness of the spotlight increases up to its maximum value L_{max}, and the angle of the spotlight widens up to A_{max}. Conversely, when the user’s gaze focuses directly on the target (θ_{min}), the spotlight’s brightness and range decrease (L_{min}, A_{min}), potentially even deactivating the spotlight. Figure. 2-(c) depicts how the parameter changes. The sensitivity at which the spotlight range (A) changes on varying angular deviation (θ) can again be controlled through γ. The $\gamma = 1$ will lead to a linear relation of A and θ. $\gamma > 1$ will lead to a steeper decrease of R as θ gets smaller. The $\theta < \gamma < 1$ will lead to a gradual decrease of A as θ gets smaller. We use $\gamma = 1$ as the default but the value can be configured as needed.

4.2 Spatial Audio Control Module

Audio plays a crucial role in signaling and notifying users about new information. Spatial audio is particularly advantageous as it allows us to not only provide audible cues but also direct users’ attention to the source of the sound. It has been shown that in hybrid video calls spatializing participants’ voices was preferred to an increased speech stream identification [25].

In the Spatial Audio Control Module of our system, we manage two types of sound sources: a signaling sound to indicate a user who wishes to speak and adjustments to the volume of the current speaker in the group conversation.

Signaling Sound: To notify users of a participant awaiting their turn to speak, we employ an arbitrary beeping sound akin to a chiming tone. Based on their role as a speaker or the predetermined subtlety weight, the sound source will be projected to the user’s head position. This positioning amplifies the sound and makes it appear closer to the user, indicating the need to turn their head. When the user shifts their gaze and the new target speaker enters their field of view, the sound source returns to its original position—coincident with the new speaker’s coordinates. This transition of sound source coordination is proportionately controlled when it is outside the user’s viewport. Figure. 2-(d) visualizes the sound source location scenario. We denote the user’s position in 3D space as $u = (x_u, y_u, z_u)$ and the target speaker’s

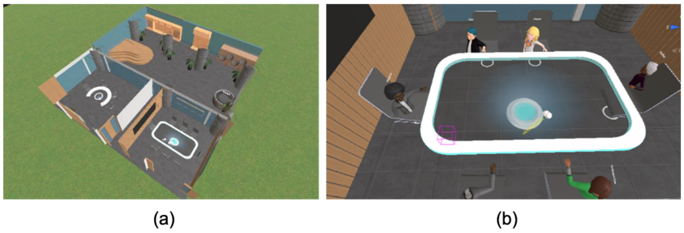

(a) (b)

Fig. 4: We illustrate the virtual environment from (a) the top-down view and (b) with the virtual agent avatars

position as $t = (x_t, y_t, z_t)$. At the moment when the target speaker signals their intention to speak, the angular distance between the user's initial forward vector $\vec{n}$ (i.e. where they are facing) and the vector pointing from u to t, denoted as $\vec{v}_{ut}$ form angle of θ_{max}. If we denote o as the 3D position of the sound source location, it can be expressed as follows:

$$o' = \begin{cases} u, & \text{if } \theta \geq \theta_{\max} \\ u + |\vec{v}_{ut}| \cos(\theta_{\max} - \theta), & \text{if } \theta_{\max} > \theta > \theta_{\min} \\ t, & \text{if } \theta \leq \theta_{\min}, \end{cases} \tag{4}$$

where the equation ensures that the attached object is initially projected to the user's head position. As the user's rotation shifts from θ_{max} to θ_{min}, the sound source follows the path along the vector t connecting the user and the new speaker coordinate. Finally, when θ becomes less than or equal to θ_{min}, the sound source returns to the original position of the target speaker.

Speaker Volume: In addition, we lower the volume (intensity) of the current speaker for 2 seconds as the chiming sound plays. Afterward, the volume returns to its normal level. This design choice ensures that participants do not feel pressured to halt their speech. Notably, speaker volume adjustments are exclusive to listener users, as speakers do not hear their own voices through the VR headset but in their real-life surroundings. This constraint precludes us from controlling their audio levels.

5 Evaluation Study

5.1 Study Design

The evaluation of our approach is undertaken from the perspective of the signal *receiver*, i.e., the user actively engaged in an ongoing conversation, who should be made aware of a new speaker within the group intending to contribute. This section assesses the performance of our approach using various attention guidance methods aimed at directing attention to a new speaker.

In our study, we compare two primary attention guidance methods: the *Text-Icon* method, which is inspired by Horizon Workrooms—a state-of-the-art application [38]. It features a text notification window displaying the user's name who is 'raising their hand', accompanied by a hand icon above the new speaker's avatar. We chose this method as a baseline for comparison due to its resemblance to the "raise-hand" feature prevalent in video call platforms, ensuring user familiarity and its current implementation in commercial VR conference systems. Additionally, we assess the *SGD* method [17], noted for its subtlety and favorable evaluation in section 3, to evaluate how attention guidance methods not specifically tailored for social VR contexts are received.

Our investigation also includes testing our proposed method with (*Light-Audio*) and without audio (*Light*) to determine the effectiveness of integrating light and audio modules. The methods *Text-Icon* and *SGD* are depicted in Figure 6. Furthermore, we present a table that outlines the characteristics of each method based on a taxonomy from prior work [47], as shown in Table 1.

Our study uses a within-subjects design, ensuring the counterbalanced presentation of each method through a Latin square design. Participants utilized an Oculus Quest Pro headset, which supports eye tracking.

Method	Diegesis	Senses	Placement
Text-Icon	Non-diegetic	Visual	World-fixed
SGD	Non-diegetic	Visual	Screen-fixed
Light	Diegetic	Visual	World/Screen-fixed
Light-Audio	Diegetic	Visual/Auditory	World/Screen-fixed

Table 1: Comparison of attention guidance methods in Section 5

Considering the influence of a user's role (speaker or listener) on gaze patterns [35], we designed two scenarios: one where the user receives a signal while speaking and another while listening. A simulated multi-user VR meeting featured five pre-recorded virtual agents engaged in simple small talk topics. Eight distinct topics were utilized, divided evenly between the listener and speaker scenarios. Pre-recorded scripts were crafted to simulate interaction, with virtual agents asking the participant's viewpoint on the topic and responding when the participant posed a pre-determined question to the group. It should be noted that participants were instructed beforehand on the specific question to direct to the group.

In a scenario, turn-taking occurred among the user, a virtual agent within the user's field of view, and another virtual agent positioned outside the user's field of view. The turn-taking order and timing of the new speaker's signal for each scenario are depicted in Figure 5. To prevent user anticipation of the turn-taking order, which could lead to predictive identification of the next speaker, we employed a randomized approach. Specifically, we randomized the name and avatar graphic skinning for each virtual agent before every topic, and alternated the users' positions between two seats, ensuring an equal distribution along the topics.

5.2 Measurements

The study collected both quantitative and qualitative metrics. Quantitative measures included response time, defined as the time from signal issuance to the user aligning their gaze with the target; and a "missed" count if the user failed to turn their head towards the target within 5 seconds. A brief quiz assessing users' attention to the conversation was also administered.

For qualitative evaluation, we administered a communication satisfaction questionnaire [20] to investigate how the turn-taking method employed could influence the overall communication experience. The Igroup Presence Questionnaire (IPQ) [53] was utilized to determine the impact of our attention guidance method on users' sense of immersion. Additionally, the Notification questionnaire [14] was employed to assess the method's effectiveness in signaling a new speaker's intervention. User preferences were quantified using a 5-point Likert scale, providing a comparative metric of user favorability across the tested methods.

5.3 Implementation Setup

For the virtual environment, a virtual meeting room environment was implemented with Unity version 2021.3.11f1. We utilized several SDKs for our implementation: Movement SDK for eye tracking to discern user saliency and viewing direction, Meta Avatar SDK for rendering avatars, gestures, and lip-syncing, and the Steam Audio Plugin to incorporate spatial audio features. The resulting virtual scene is shown in Figure 4. The voiceover recordings of the avatars were generated through Speechify [57]. Avatar animation and lip sync were manually recorded by motion capture to ensure there were no socially odd gestures.

A preliminary study was conducted to test the functionality of the prototype and to set the values of parameters described in Section 4 for the evaluation study. As a result, we empirically set the parameters as; $L_{min} = 0.5$, $L_{max} = 1.1$ for environment light; $C_{warm} = (1, 0.902, 0.259)$ (close to yellow), $C_{cold} = (1, 1, 1)$ for point in RGB color 0 to 1 scale, $L_{min} = 0.8$, $L_{max} = 1.5$ and $A_{min} = 30$, $A_{max} = 60$ for spotlight. Note that these parameters are directly applied to Unity's light object component.

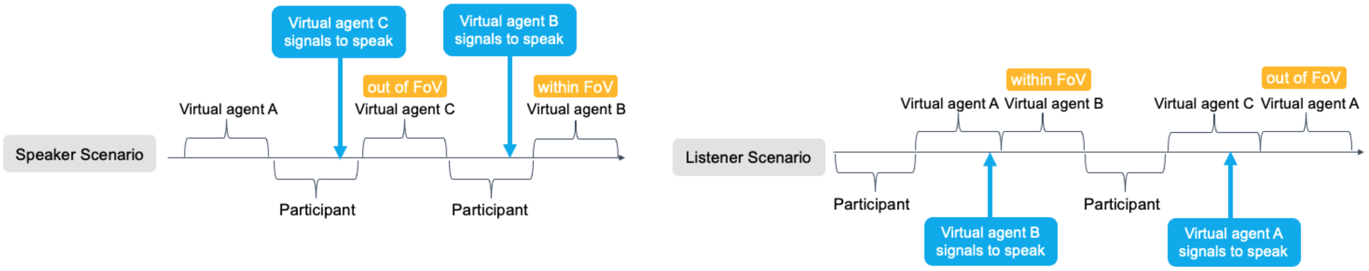


Fig. 5: Sequential turn-taking order for speaker and listener scenarios, with virtual agents' animations and voiceovers executed accordingly. The signal for the new speaker is dispatched 5 seconds following the current speaker's turn. Each scenario includes instances of the attention guidance method activation, both when the new speaker is out of and within the participant's field of view.

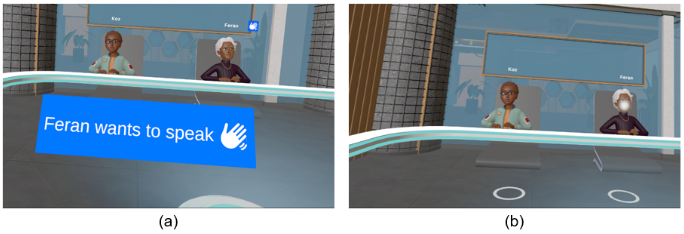


Fig. 6: We compared our approach to (a) *Text-Icon*, a text window fixed to the user's desk space and an icon appearing next to the agent's name tag, and (b) *SGD* a flickering effect on the target at user's peripheral view.

5.4 Participants

We recruited 20 volunteers, comprising eleven females and nine males, with ages ranging from 19 to 36 years ($\mu = 28.45$, $\sigma = 5.25$). All participants exhibited normal color vision, with ten having unaided normal vision, and the remaining individuals possessing vision corrected to normal. Among the participants, eleven had prior experience with VR, while none had experience specifically in social VR applications.

5.5 Procedure

Upon welcoming the participants, we presented an overview of the study's procedures, obtained their consent through a signed form, and gathered demographic information. Seated in swivel chairs, participants were introduced to the Oculus Quest Pro and underwent eye tracker calibration. To acclimate to the virtual environment, participants briefly explored their surroundings. The nature of the study was explained, emphasizing that participants would engage in a VR conversation with a group of people without specifying that these individuals were prerecorded avatars, aiming to enhance participant engagement.

Each participant was assigned the username “*Charlie*" for the virtual meeting, which other virtual agents would use as a reference. Participants were instructed to respond when addressed by the agents. In the listener scenario, where the participant initiated the conversation, individuals were given a topic and instructed to commence the group discussion by posing a question to the entire group without specifying an agent to answer immediately. The conversation script was executed in a manner where the experimenter manually triggered a specific voiceover and animation at the correct timing, adhering to the predetermined turn-taking order, of which participants remained unaware. The signal, or the presentation of the method, was dispatched 5 seconds after the current speaker spoke, ensuring that participants spoke for at least 5 seconds guaranteed sufficient time to transmit the signal when the participant was the speaker. If the participant's gaze direction aligned with the new speaker and this alignment lasted for over 1.5 seconds, the subsequent speaker's animation would be triggered. If the previous speaker's animation was still playing, it would be interrupted. Alternatively, should the participant fail to notice the new speaker within 5 seconds, the new speaker's animation would commence.

Each trial's conversation spanned two to three minutes, and participants underwent one trial per method, resulting in four trials for each speaker/listener scenario and eight trials in total with distinct conversation topics. The entire experiment lasted approximately 70 minutes. After each trial, participants were required to fill out the questionnaires detailed in Section 5.2. After the entire experiment, a semi-structured interview was conducted to gather additional comments and qualitative feedback. The experimenter regularly checked if participants needed a break, allowing breaks as necessary.

5.6 Results

We conducted a statistical analysis of our results. It is important to note that all participants correctly answered all three questions in the post-trial quiz, indicating their attentiveness to the conversational task. Additionally, no presented methods were missed by any participant in any trial. This may be attributed to the brevity of the conversation and the limited number of method presentations in each trial.

In our analysis, we consider the impact of several factors: the presence or absence of the target within the participant's view (denoted as *View*), the role of the participant (speaker or listener, denoted as *Role*), and the presented method (denoted as *Method*). We follow the analytical approach outlined by Rzayev et al. [51], which applied the Aligned Rank Transform (ART) using the ARTool toolkit and applied paired-sample t-test with Tukey correction and for ANOVAs used paired-sample t-test with Bonferroni correction. We depict statistical results in this section in Figure. 7.

Response Time: There were statistically significant main effect for all *Method*, *View*, and *Role*. A two-way interaction effect of *Method*×*View* was observed. The pairwise comparison of *Method* revealed that all comparisons were statistically significant to each other ($p < .001$, except for between *Light* and *SGD* with $p = .013$ and between *SGD* and *Text-Icon* with $p = .017$)

Pairwise comparisons for *View* revealed that response times within view were statistically significantly shorter than those out of view ($p < .001$). Similarly, for *Role*, listener scenarios exhibited statistically significantly shorter response times compared to speaker scenarios ($p < .001$).

Light-Audio led to a significantly faster response time than other methods for within-view new speakers (with *Light-Audio* $p = .001$, with *SGD* $p < .001$, and with *Text-Icon* $p = .010$) and out-of-view new speakers (all $p < .001$). Compared to other methods, *Text-Icon* resulted in the least response time for out-of-view signals (all $p < .001$). See Figure. 7-(a).

Communication Satisfaction We found a statistically significant main effect of *Method* on participants' communication experience. Pairwise comparisons showed that participants had significantly lower communication satisfaction with *SGD* compared to *Light* ($p < .001$) and *Light-Audio* ($p < .001$).

Presence: While *Light-Audio* and *Light* showed a higher IPQ score than *Text-Icon* and *SGD*, there were no statistically significant main and interaction effects found in the statistical analysis ($p > .05$).

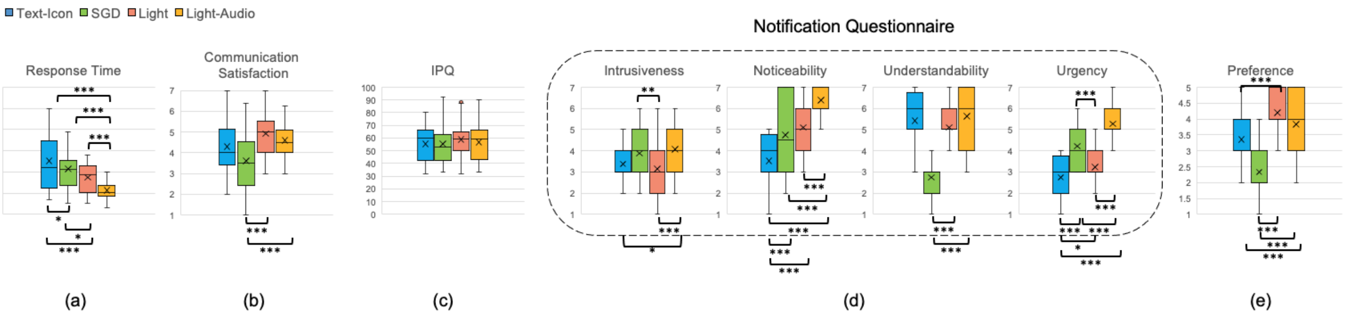


Fig. 7: Statistical results for (a) response time (b) communication satisfaction (b) presence (c) preference, and (d) Notification questionnaire. The post hoc analysis revealed that *Light-Audio* has a significantly faster response time, better-perceived conversation satisfaction, and preference. It is also shown that participants reported *Light-Audio* to have high scores in all Notification subscales. Asterisk (*) indicates a statistically significant difference between conditions: $p < 0.05(*); p < 0.01(**); p < 0.001(***)$.

	Response Time					Conversation Quality					Presence					Preference				
Factor	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2
M	3	28.013	41.954	$< .001$	.293	3	12.578	9.545	$< .001$	0.159	3	105.966	.473	.701	.009	3	25.342	29.607	$< .001$	.369
V	1	28.301	42.385	$< .001$	.122	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R	1	9.041	13.540	$< .001$	.043	1	.018	.014	.906	.000	1	.165	.001	.978	0.000	1	.100	.117	.733	.001
M×V	3	14.975	22.427	$< .001$	.181	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
M×R	3	.860	1.288	.279	.013	3	.023	.018	.997	.000	3	15.860	.071	.975	.001	3	.183	.214	.886	.004
V×R	1	1.986	2.974	0.086	0.010	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
M×V×R	3	.161	.242	.867	.002	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

	Intrusiveness					Noticeability					Understandability					Urgency				
Factor	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2	df_{effect}	MS	F	p	Partial η^2
M	3	7.373	5.042	.002	0.91	3	55.940	32.410	$< .001$	.937	3	71.217	30.146	$< .001$	.375	3	50.273	49.928	$< .001$	.496
R	1	6.006	4.108	.044	.026	1	.156	.091	.764	.001	1	.000	.000	1.000	.000	1	.056	.056	.813	.000
M×R	3	.456	.312	.817	.006	3	.056	.033	.992	.001	3	.117	.050	.985	.001	3	.106	.106	.957	.002

Table 2: ANOVA main effects and interactions for the Notification questionnaire (M: *Method*, V: *View*, R: *Role*). Note that *View* were not evaluated for the subjective questionnaires as a trial encompasses both within-view and out-of-view cases.

Notification: The Notification questionnaire consists of four scales: Intrusiveness, Noticeability, Understandability, and Urgency. Statistical analysis revealed a significant main effect of *Method* across all scales. Notably, Intrusiveness demonstrated both a significant main effect of *Role*. No significant interaction effect was observed.

In terms of Intrusiveness, pairwise comparisons indicated that the speaker role yielded significantly higher intrusiveness scores than the listener role ($p = .044$). Regarding *Method*, significant differences were observed between *Light-Audio* and *Light* ($p < .001$), *Light-Audio* and *Text-Icon* ($p = .011$), and *Light* and *SGD* ($p = .008$).

For Noticeability, the pairwise comparison showed that *Light-Audio* was significantly more noticeable than all other methods ($p < .001$) while *Text-Icon* was significantly less noticeable than all other methods ($p < .001$).

For Understandability, the pairwise comparison revealed that *SGD* was significantly less 'understandable' than all other methods ($p < .001$). No other pair showed significant differences.

For Urgency, the pairwise comparison revealed statistically significant differences between all methods, with *Light-Audio* having the highest score ($p < .001$ for all pairs except for *Light* and *Text-Icon*, $p = .036$). See Figure. 7-(e).

Preference: There was a statistically significant main effect of *Method* on participants' preference scores. No other statistically significant main or interaction effects were found. Pairwise comparisons revealed significant differences ($p < .001$) in preference scores between *Light-Audio* and *SGD*, *Light* and *SGD*, *Light-Audio* and *Text-Icon*, and *Light* and *Text-Icon*. See Figure. 7-(d).

6 Discussion

In this section, we explore the results and discuss possible design implications derived from them.

6.1 Effect of Audio Cues in Our Approach

The results analysis unveiled intriguing effects of the Audio module in our approach. The outcomes suggest that the audio cue is effective in alerting users and conveying information, as evidenced by significantly faster response times and higher noticeability. The fact that it scored significantly higher in Urgency implies the potential to dynamically add or detach the audio module based on the urgency of user conversations. This could be an interesting avenue for further evaluation, especially considering the perspective of new speakers. However, the significantly high perceived intrusiveness suggests that there may be cases where users dislike the intrusive nature of audio cues. Nonetheless, it is promising that users rated a high preference for this method.

Expanding on this, it would be valuable to explore user preferences concerning the customization of the audio module. For instance, investigating whether users would prefer adjustable volume levels or the ability to choose specific sounds could contribute to a more tailored and user-friendly experience.

6.2 Perceived Interruption in Conversational Flow

Interestingly, we observed an effect related to whether the participant was a speaker or a listener when receiving the new speaker signal. Participants generally perceived a better conversational experience when they were listeners compared to when they were the speakers. A participant noted, "*I felt slightly interrupted when there was a cue while I was talking. I wasn't completely bothered by some of the methods, but I guess it's natural to feel that way.*" This is consistent with findings from previous research indicating that users are more likely to visually engage with others when they are listeners in a group conversation, as opposed to when they are the speakers [35].

Six participants verbally expressed lower satisfaction with the *SGD* method compared to others, citing interruptions to their conversational flow as the primary reason for its reduced scoring. This aligns with the higher perceived intrusiveness scores of *SGD*. Extending this observation, further investigation into user reactions during different stages of conversation, such as during critical points or pauses, could provide deeper insights into how these cues impact the natural flow of communication.

6.3 Preferred Subtlety in Attention Guidance

Participants highlighted reasons for favoring *Light*, emphasizing its ability to blend into the environment. A participant stated, "*I liked that* *the lighting effect provided just enough cues to make me notice that there is a change in the environment, but also I can easily ignore the fact and revisit the direction the signal came from whenever I want without feeling interrupted in between.*” While users expressed a preference for *Light-Audio*, concerns were raised regarding real-world applications. A participant commented, “*I was fine with the chiming sound here since this was a short small talk among strangers, but I will be quite annoyed by the frequency of the chiming sound if it constantly comes up. It would be great if we can turn off the sound or the frequency of playback can be controlled.*” Surprisingly, *SGD* was disliked as users found the constant modulation effect in their peripheral vision annoying.

Building on this, exploring user-friendly controls for adjusting the subtlety of cues, such as the intensity of light or the frequency of audio, could enhance the applicability of these methods in various real-world scenarios.

Aligning with our group interview study, users consistently expressed a preference for characteristics of subtlety and diegeticness in attention guidance methods in a conversational setting. Further research into the nuanced aspects of these preferences could contribute to the development of more user-centric and adaptable communication systems.

7 Conclusions, Limitations, and Future Work

We present a multimodal approach for guiding user attention in a social VR group conversation setting. Our method aims to offer tailored attention guidance, providing more pronounced cues for highly engaged participants and subtler cues for those less engaged, ultimately enhancing the overall meeting experience. Leveraging light and spatial audio as diegetic guidance methods within the virtual environment, our approach demonstrated significantly reduced response times while maintaining high perceived conversation quality and preference. Moving forward, we envision extending our work to diverse VR social scenarios, including presenter-audience relationships and dynamic party-like settings with multiple small groups forming and disbanding continuously.

Although our work has achieved notable results, it has some limitations. Firstly, in our formative study, interviewees were presented with existing methods using video forms, not experiencing them within VR themselves, due to computational constraints among the interviewees. This approach suggests that different findings might emerge if the methods were presented in an immersive VR environment. In our evaluation study, a pre-recorded setup was employed, necessitated by the requirement for multiple actors to represent different avatars in each scenario. This approach was taken to avoid participant familiarization with specific virtual agents and to maintain consistency. Additionally, it mitigated potential network connectivity issues that could adversely affect the conversational experience. However, it is important to note that communication satisfaction scores might vary in real-life settings and may present more nuanced outcomes. In future work, we plan to incorporate real-user conversations using the Desert Survival Task (DST [29]), as employed in prior studies [32].

The brevity and informal nature of the conversations conducted in this study may also have impacted the results. Additionally, the brevity and informality of the conversations. Further investigation in more formal or presentation-like settings, where participants engage in longer and structured discourse, could yield diverse outcomes.

Furthermore, our virtual environment’s default lighting setups, optimized for typical scenarios, may not suit environments with different lighting conditions, such as dark rooms. A promising avenue for future work involves exploring user perceptions of ‘subtle’ and ‘intrusive’ lighting in varied settings and defining parameter thresholds accordingly. Another potential research direction is the integration of automatic engagement detection, utilizing non-verbal cues like nodding or physical gestures. This would provide a more comprehensive understanding of user engagement dynamics in virtual social interactions.

As we progress, it is essential to consider the challenge of managing multiple participants who wish to speak using our approach. The visualization and signaling of the order of users and queuing mechanisms demand careful consideration. Future work should delve into how to effectively represent and manage user queues, ensuring a smooth and organized communication flow.

References

[1] A. Abdullah, J. Kolkmeier, V. Lo, and M. Neff. Videoconference and embodied vr: Communication patterns across task and medium. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021. 2

[2] A. Adjorlu, A. Hussain, C. Mødekjær, and N. W. Austad. Head-mounted display-based virtual reality social story as a tool to teach social skills to children diagnosed with autism spectrum disorder. In *2017 IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual & Augmented Reality (KELVAR). IEEE*, 2018. 2

[3] P. R. K. Babu, P. Oza, and U. Lahiri. Gaze-sensitive virtual reality based social communication platform for individuals with autism. *IEEE Transactions on Affective Computing*, 9(4):450–462, 2017. 2

[4] R. Bailey, A. McNamara, N. Sudarsanam, and C. Grimm. Subtle gaze direction. *ACM Transactions on Graphics (TOG)*, 28(4):1–14, 2009. 3

[5] B. Buxton. Mediaspace–meaningspace–meetingspace. In *Media space 20+ years of mediated life*, pp. 217–231. Springer, 2009. 1

[6] K.-W. Chen, Y.-J. Chang, and L. Chan. Predicting opportune moments to deliver notifications in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2022. 3

[7] L. Chittaro and S. Burigat. 3d location-pointing as a navigation aid in virtual environments. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 267–274, 2004. 3

[8] F. Cole, D. DeCarlo, A. Finkelstein, K. Kin, R. K. Morley, and A. Santella. Directing gaze in 3d models with stylized focus. *Rendering Techniques*, 2006:17th, 2006. 2

[9] J. Dawson and T. Foulsham. Your turn to speak? audiovisual social attention in the lab and in the wild. *Visual Cognition*, 30(1-2):116–134, 2022. 2

[10] M. S. El-Nasr, A. Vasilakos, C. Rao, and J. Zupko. Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2):145–153, 2009. 2, 5

[11] K. E. Finn. *Video-mediated communication*. L. Erlbaum Associates Inc., 1997. 1

[12] G. Freeman and D. Acena. Hugging from a distance: Building interpersonal relationships in social virtual reality. In *ACM international conference on interactive media experiences*, pp. 84–95, 2021. 2

[13] C. George, M. Demmler, and H. Hussmann. Intelligent interruptions for ivr: investigating the interplay between presence, workload and attention. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*, pp. 1–6, 2018. 3

[14] S. Ghosh, L. Winston, N. Panchal, P. Kimura-Thollander, J. Hotnog, D. Cheong, G. Reyes, and G. D. Abowd. Notifivr: Exploring interruptions and notifications in virtual reality. *IEEE transactions on visualization and computer graphics*, 24(4):1447–1456, 2018. 3, 6

[15] D. R. Gibson. Participation shifts: Order and differentiation in group conversation. *Social forces*, 81(4):1335–1380, 2003. 2

[16] L. E. Girolametto. Improving the social-conversational skills of developmentally delayed children: An intervention study. *Journal of Speech and Hearing Disorders*, 53(2):156–167, 1988. 2

[17] S. Grogorick, M. Stengel, E. Eisemann, and M. Magnor. Subtle gaze guidance for immersive environments. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 1–7, 2017. 2, 3, 6

[18] U. Gruenefeld, A. E. Ali, S. Boll, and W. Heuten. Beyond halo and wedge: Visualizing out-of-view objects on head-mounted virtual and augmented reality devices. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*, pp. 1–11, 2018. 3

[19] H. Hata, H. Koike, and Y. Sato. Visual guidance with unnoticed blur effect. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 28–35, 2016. 2

[20] M. L. Hecht. The conceptualization and measurement of interpersonal communication satisfaction. *Human Communication Research*, 4(3):253–264, 1978. 6

[21] J. Heritage and P. Drew. Talk at work. *Interaction in institutional settings*, 1992. 2

[22] S. Ho, T. Foulsham, and A. Kingstone. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905, 2015. 2

[23] J. Hollan and S. Stornetta. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 119–125, 1992. 2

[24] E. Hu, J. E. S. Grønbæk, A. Houck, and S. Heo. Openmic: Utilizing proxemic metaphors for conversational floor transitions in multiparty video meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2023. 2

[25] J. Hyrkas, A. D. Wilson, J. Tang, H. Gamper, H. Sodoma, L. Tankelevitch, K. Inkpen, S. Chappidi, and B. Jones. Spatialized audio and hybrid video conferencing: Where should voices be positioned for people in the room and remote headset users? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023. 5

[26] S. T. Iqbal and B. P. Bailey. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(4):1–28, 2010. 3

[27] M. J. Bietz, N. Goyal, N. Immorlica, B. MacIntyre, A. Monroy-Hernández, B. C. Pierce, S. Rintel, and D. Y. Wohn. Social presence in virtual event spaces. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–5, 2022. 2

[28] W. Jarrold, P. Mundy, M. Gwaltney, J. Bailenson, N. Hatt, N. McIntyre, K. Kim, M. Solomon, S. Novotny, and L. Swain. Social attention in a virtual public speaking task in higher functioning children with autism. *Autism Research*, 6(5):393–410, 2013. 2

[29] J. C. Laferty and A. W. Pond. Desert survival situation. https://www.humansynergistics.com/docs/default-source/defaultdocument-library/desert_prod_info_sheet_v2-0_np.pdf, 1928. Accessed: [Insert Date Here]. 9

[30] D. Lange, T. C. Stratmann, U. Gruenefeld, and S. Boll. Hivefive: Immersion preserving attention guidance in virtual reality. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020. 2, 3

[31] S. R. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000. 2

[32] J. V. Li, M. Kreminski, S. M. Fernandes, A. Osborne, J. McVeigh-Schultz, and K. Isbister. Conversation balance: A shared vr visualization to support turn-taking in meetings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–4, 2022. 2, 9

[33] Y.-C. Lin, Y.-J. Chang, H.-N. Hu, H.-T. Cheng, C.-W. Huang, and M. Sun. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2535–2545, 2017. 3

[34] A. MacQuarrie and A. Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*, pp. 45–54. IEEE, 2017. 2

[35] T. Maran, M. Furtner, S. Liegl, T. Ravet-Brown, L. Haraped, and P. Sachse. Visual attention in real-world conversation: Gaze patterns are modulated by communication and group size. *Applied Psychology*, 70(4):1602–1627, 2021. 2, 6, 8

[36] J. McVeigh-Schultz, A. Kolesnichenko, and K. Isbister. Shaping pro-social interaction in vr: an emerging design framework. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–12, 2019. 2

[37] J. McVeigh-Schultz, E. Márquez Segura, N. Merrill, and K. Isbister. What's it mean to" be social" in vr? mapping the social vr design ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, pp. 289–294, 2018. 2

[38] Meta. Horizon workrooms, 2023. 6

[39] L. Mondada. Embodied and spatial resources for turn-taking in institutional multi-party interactions: Participatory democracy debates. *Journal of pragmatics*, 46(1):39–68, 2013. 2

[40] F. Moustafa and A. Steed. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2018. 2

[41] C. Nash, M. H. Jarrahi, and W. Sutherland. Nomadic work and location independence: The role of space in shaping the work of digital nomads. *Human Behavior and Emerging Technologies*, 3(2):271–282, 2021. 1

[42] L. T. Nielsen, M. B. Møller, S. D. Hartmeyer, T. C. Ljung, N. C. Nilsson, R. Nordahl, and S. Serafin. Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pp. 229–232, 2016. 2

[43] M. Pielot, A. Vradi, and S. Park. Dismissed! a detailed exploration of how mobile phone users handle push notifications. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services*, pp. 1–11, 2018. 3

[44] G. Reyes and A. Alles. Multi-modal multi-scale attention guidance in cyber-physical environments. In *26th International Conference on Intelligent User Interfaces*, pp. 356–365, 2021. 3

[45] S. Rintel, A. Sellen, A. Sarkar, P. Wong, N. Baym, and R. Bergmann. Study of microsoft employee experiences in remote meetings during covid-19 (project tahiti). microsoft research, 2020. 1, 2

[46] D. Roth, C. Klelnbeck, T. Feigl, C. Mutschler, and M. E. Latoschik. Beyond replication: Augmenting social behaviors in multi-user virtual realities. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 215–222. IEEE, 2018. 2

[47] S. Rothe, D. Buschek, and H. Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1):19, 2019. 2, 6

[48] S. Rothe and H. Hußmann. Guiding the viewer in cinematic virtual reality by diegetic cues. In *Augmented Reality, Virtual Reality, and Computer Graphics: 5th International Conference, AVR 2018, Otranto, Italy, June 24–27, 2018, Proceedings, Part I 5*, pp. 101–117. Springer, 2018. 2, 3

[49] L. Rubinger, A. Gazendam, S. Ekhtiari, N. Nucci, A. Payne, H. Johal, V. Khanduja, and M. Bhandari. Maximizing virtual meetings and conferences: a review of best practices. *International orthopaedics*, 44:1461–1466, 2020. 1

[50] R. Rzayev, S. Korbely, M. Maul, A. Schark, V. Schwind, and N. Henze. Effects of position and alignment of notifications on ar glasses during social interaction. In *Proceedings of the 11th Nordic conference on human-computer interaction: shaping experiences, Shaping Society*, pp. 1–11, 2020. 3

[51] R. Rzayev, S. Mayer, C. Krauter, and N. Henze. Notification in vr: The effect of notification placement, task and environment. In *Proceedings of the annual symposium on computer-human interaction in play*, pp. 199–211, 2019. 3, 7

[52] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro, and M. Elsherief. Challenges of moderating social virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2023. 2

[53] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3):266–281, 2001. 6

[54] A. J. Sellen. Remote conversations: The effects of mediating talk with technology. *Human-computer interaction*, 10(4):401–444, 1995. 1

[55] A. Sheikh, A. Brown, Z. Watson, and M. Evans. Directing attention in 360-degree video. 2016. 2, 3

[56] W. S. Smith and Y. Tadmor. Nonblurred regions show priority for gaze direction over spatial blur. *The Quarterly Journal of Experimental Psychology*, 66(5):927–945, 2013. 2

[57] Speechify. Speechify, turn any reading into audio: Tts ai voice, 2016. 6

[58] T. L. Stanton-Chapman and M. E. Snell. Promoting turn-taking skills in preschool children with disabilities: The effects of a peer-based social communication intervention. *Early Childhood Research Quarterly*, 26(3):303–319, 2011. 2

[59] W. Steptoe, R. Wolff, A. Murgia, E. Guimaraes, J. Rae, P. Sharkey, D. Roberts, and A. Steed. Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 197–200, 2008. 2

[60] A. Tse, C. Jennett, J. Moore, Z. Watson, J. Rigby, and A. L. Cox. Was i there? impact of platform and headphones on 360 video immersion. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pp. 2967–2974, 2017. 2

[61] R. Vertegaal. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 294–301, 1999. 1

[62] J. R. Williamson, J. O'Hagan, J. A. Guerra-Gomez, J. H. Williamson, P. Cesar, and D. A. Shamma. Digital proxemics: Designing social and collaborative interaction in virtual environments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2022. 1

[63] B. Xu, J. Ellis, and T. Erickson. Attention from afar: simulating the gazes of remote participants in hybrid meetings. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pp. 101–113, 2017. 1

[64] A. Yassien, P. ElAgroudy, E. Makled, and S. Abdennadher. A design space for social presence in vr. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pp. 1–12, 2020. 2

[65] X. Zhang and G. W. Furnas. Social interactions in multiscale cves. In *Proceedings of the 4th international conference on Collaborative virtual environments*, pp. 31–38, 2002. 1