A3RT: Attention-Aware AR Teleconferencing with Life-Size 2.5D Video Avatars

Xuanyu Wang*
School of Computer Science and Technology, Xi'an Jiaotong University
School of Creative Media, City University of Hong Kong

Weizhan Zhang†
School of Computer Science and Technology, Xi'an Jiaotong University

Hongbo Fu‡
School of Creative Media, City University of Hong Kong

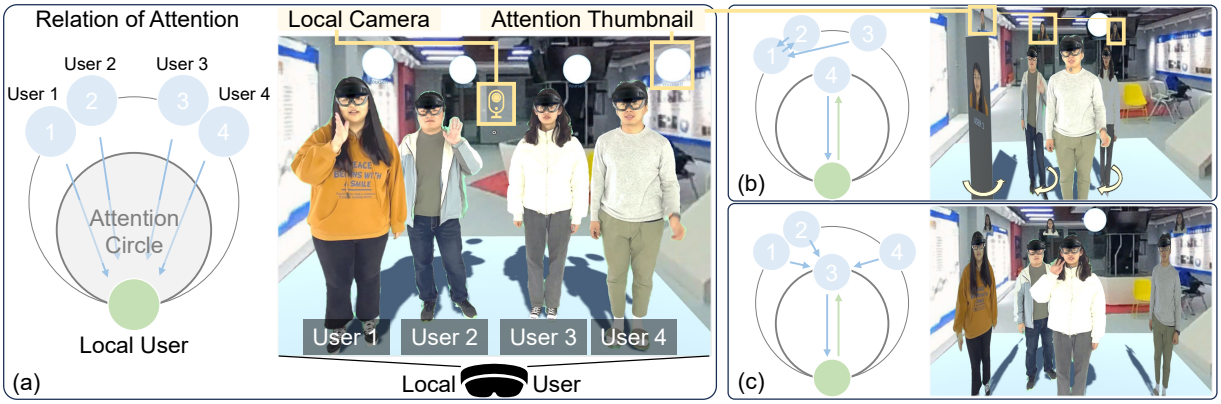


Figure 1: With a self-facing camera, a local AR user is teleconferencing with four remote users (Users 1-4) through their video avatars overlaid in the local space, as shown on the right in (a-c). The remote users have the same setup as the local user. The left images in (a-c) illustrate the layout of conversation groups and each user's attention from a top-down view. The video avatar the local user is looking at will "step up" onto the inner "Attention Circle" in (a-c) to align with the local camera. The "Attention Thumbnail" annotates at whom each user is looking (examples are framed in (a) and (b) and can be found in similar positions in (c)). 2.5D video avatars rotate to face the one being looked at, as annotated in (b). (a) The local user is not looking at anyone while Users 1-4 are looking at the local user (the remote users' Attention Thumbnails show a circle highlighted in white), standing in their initial position on the outer circle. (b) The local user is looking at User 4, whose video avatar transforms onto the Attention Circle. User 4 (Attention Thumbnail highlighted) is looking at the local user, User 1 turns around to look at User 2, and Users 2 and 3 are looking at User 1. Their Attention Thumbnails show the corresponding users' photos and names. (c) The local user and User 3 are looking at each other while others are looking at User 3. (c) is left unannotated.

Abstract

Augmented Reality (AR) teleconferencing aims to enable remotely separated users to meet with each other in their own physical spaces as if they are face-to-face. Among all solutions, the video-avatar-based approach has the advantage of balancing fidelity and the sense of co-presence using easy-to-setup devices, including only a camera and an AR Head-Mounted Display (HMD). However, non-verbal cues indicating "who is looking at whom" are always lost or misdelivered in multiparty teleconferencing experiences. To make users aware of such non-verbal cues, existing solutions explore screen-based visualizations, incorporate additional hardware, or alter to use a virtual avatar representation. However, they lack immersion, are less feasible for everyday usage due to complex installations, or lose the fidelity of remote users' authentic appearances. In this paper, we decompose such attention awareness into *the awareness of being looked at* and *the awareness of attention between other users* and address them in a decoupled process. Specifically, through a user study, we first find an unobtrusive and reasonable layout "Attention Circle" to retarget a looker's head gaze to the one being looked at. We then conduct the second user study to find an effective and intuitive "rotatable 2.5D video avatar with attention thumbnail" visualization to aid users in being aware of other users' attention. With the design choice distilled from the studies, we implement A3RT, a proof-of-concept prototype system that empowers attention-aware 2.5D-video-avatar-based multiparty AR teleconferencing in an easy, everyday setup. Ablation and usability studies on the prototype verify the effectiveness of our proposed components and the full system.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)

1 Introduction

AR teleconferencing creates the illusion of meeting teleported remote users in the local space. With AR HMDs considered a next-generation personal computing platform, AR teleconferencing has the potential to enter people's daily lives and work with its significantly improved sense of co-presence compared to current video conferencing experiences. Existing teleconferencing solutions mainly leverage volumetric capturing and reconstruction [35, 50, 78], virtual avatars [1, 43, 44], and video avatars [5, 6, 8, 17, 27, 69, 75] to represent remote users. Video-avatar-based approaches require only commodity hardware, which is affordable and effortless to set up, and is studied to be able to present high-fidelity remote users with a relatively high level of immersion [69], thus being more feasible for

*E-mail: xwang2247-c@my.cityu.edu.hk
†Corresponding author. E-mail: zhangwzh@xjtu.edu.cn
‡Corresponding author. E-mail: hongbofu@cityu.edu.hk

everyday teleconferencing in interaction-, social-, or user-oriented scenarios [17, 30, 54, 75].

Besides the representation of remote users, attention awareness, i.e., knowing who is looking at, talking to, or referring to whom, is crucial to the engagement of the teleconferencing experience [13,46]. Existing methods explore solutions based on 2D screens [23, 25], special hardware [35, 53, 78], virtual avatars [43, 55], symbolic visualizations [56, 73], and telepresence robots [51, 63]. However, there are still gaps between them and our goal to achieve attention awareness in a mono-camera setup in video-avatar-based multiparty AR teleconferencing. 2D solutions lack immersion, special hardware makes the system less generalizable and accessible, and synthesizing novel views and motions for video avatars is much harder than controlling virtual avatars. While symbolic visualizations are accurate and rotating motions of telepresence robots are intuitive, they are under-explored for life-size video avatars in AR.

To address such an attention awareness problem, we first break it down into three parts. Taking a fundamental three-user (Users A-C) scenario as an example, when an action of looking occurs, e.g., A is looking at B, as shown in Fig. 2 (a), A (looker), B (lookee), and C (onlooker) should all be aware. 1) Looker's awareness is intrinsically ensured. However, when it comes to the 2) lookee's and 3) onlooker's awareness, spatial ambiguity would occur in a static mono-camera setup for video-based solutions, as shown in Fig. 2 (b) and (c). Because A is captured only from the local camera's perspective, B and C would both see A looking to the left front. In this case, B would feel like A is looking somewhere on her right back instead of at her. C might can infer A's attention but the perceived spatial relation is misaligned. To correct the lookee's (B's) awareness of being looked at, A's video avatar should look straight at B. We propose to decouple this process by inserting a head gaze retargeting component. It dynamically transforms the lookee's video avatar to the center of the looker's view, aligned with the local camera. This ensures users are looking straight no matter who they are looking at. This leads to our first question **RQ1: How to adaptively transform the video avatar being looked at to the center in front of the local user unobtrusively and reasonably?** To answer this question, we conduct our first user study and find the "Attention Circle" layout (see Fig. 1 (b) and (c)) to be the most reasonable and preferred, with a high level of co-presence and low task loads.

In addition to the lookee's awareness, we need to further address the onlooker's awareness. Without head retargeting, A's attention in C's view is misaligned (Fig. 2 (c)), while with head retargeting, C would perceive A is looking at him (same as B does) since A's video avatar looks straight at him. This thus leads to our **RQ2: How to visualize the attention between other users accurately, efficiently, and intuitively?** To answer this question, we conduct the second user study to explore design choices and find the "rotatable 2.5D video avatar with attention thumbnail" to meet the needs best.

With the design distilled from the two studies, we then implement A3RT, a proof-of-concept prototype to demonstrate attention-aware multiparty AR teleconferencing based on 2.5D video avatars. Besides the basic video-avatar-based teleconferencing, it contains a head gaze retargeting component and an attention synchronizing component to support the lookee's and onlooker's attention awareness mentioned above. Users can join the teleconferencing with a setup similar to the current video conferencing, including only an AR HMD and a commodity camera. We conduct ablation and usability studies using the prototype to verify each proposed component's effectiveness and evaluate the full system's performance. The results show that both components are indispensable, and our proposed system achieves full attention awareness and has high usability.

The main contributions of this work are threefold.

- We identify elements of full attention awareness in everyday video-avatar-based AR teleconferencing and propose a feasible solution paradigm that retargets users' head gaze and visualizes the attention between them.
- Through two user studies, we find the "Attention Circle" and "rotatable 2.5D video avatar with attention thumbnail" techniques as two key components in the proposed paradigm.
- We build the A3RT proof-of-concept prototype system that implements the proposed techniques to achieve attention awareness. The ablation study and usability study verify the effectiveness of each component and the system.

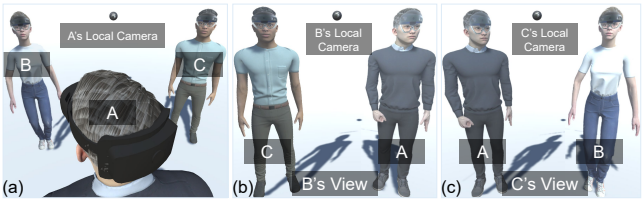


Figure 2: Ambiguity of attention from the lookee's (b) and onlooker's (c) perspectives. (a) A looks to the left front at B's video avatar. (b) B feels A's video avatar is looking to B's right back instead of at B. (c) A's video avatar's looking direction is also misaligned in C's view.

2 Related Work

2.1 Video-Avatar-Based Teleconferencing

A 2D video avatar (referred to as a video avatar) represents a remote user as matted human-only real-time streamed images on a 2D plane in the 3D space. Compared to a complete volumetric representation [35, 50, 78], it requires fewer and simpler instruments to present high-fidelity human images. The ability to present authentic human appearances makes video avatars a better representation than virtual avatars [43, 44] for work [30] and conversation-centered scenarios [17, 54, 69, 71, 75]. Reconstructed avatars tailored for HMDs such as Apple's "Persona" for Vision Pro [1] and Meta's Codec Avatar [41] currently require complicated tracking modules on the HMD to support upper-body-only avatars, making them not applicable to life-size whole-body experiences on future lightweight always-on AR HMDs, especially OST ones. Moreover, unlike the effect of high-end solutions [35], these reconstructed avatars are not clearly free of potential uncanny issues yet. Therefore, the video avatar representation is an essential direction to exploring feasible immersive AR teleconferencing solutions that can truly outperform current video conferencing experiences under a configuration just as simple and accessible, only replacing 2D screens with lightweight, always-on AR HMDs.

Prior arts have explored using video avatars to incorporate real humans into immersive experiences. One group of works embed video avatars in VR environments [8, 14, 27, 57, 70, 76]. For example, Insley et al. [27] present video avatars in a CAVE VR environment. Yura et al. [76] use a video avatar in a collaborative virtual exhibitions. Wang et al. [70] and Cui et al. [14] fuse video avatars into virtual scene models for surveillance systems. In HMD-based VR, researchers use video avatars for social [57], education [8], and teleconsultation [75] scenarios. A 3D mesh [8] or a point cloud [75] captured by a single RGB-D camera presents users mainly from the front view. Thus, they are more like video avatar solutions instead of complete volumetric solutions. Another group of works integrates video avatars into the physical space using AR HMDs [6, 7, 17, 69]. For example, Billinghurst and Kato [6, 7] align half-body video avatars with real objects for teleconferencing. Simone et al. [17] use video avatars to enable users to watch TV together. Wang et al. [69] study the quantitative video avatar placement in AR teleconferencing. However, these works in VR and AR focus on either comparing

different representations or system implementations, but do not address attention awareness for life-size video avatars in multiparty AR teleconferencing.

Another form of using video avatars in teleconferencing and telepresence scenarios is telepresence robots such as Double [59]. Researchers explore using such robotic devices in teleconferencing [51, 63] and academic meetings [47]. To further blend such robots into the surroundings, Furuya and Takashio [21] propose to replace the background of a remote user's video with the local environment. Fuchs et al. propose to display a life-size remote user on a transparent screen [20]. While being a practical solution, the robot form makes it less similar to real humans and less integrated with local surroundings than life-size video avatars in AR. Moreover, the AR solution is more flexible regarding the physical space and more scalable to multi-user scenarios.

2.2 Attention Awareness in Teleconferencing

Attention awareness is essential in offline conversations. It is often naturally delivered by non-verbal spatial cues such as gaze and facing directions [2]. Teleconferencing endeavors have explored solutions to it from the following aspects.

Avatar-based solutions. It is straightforward to control a virtual avatar's head rotation to present attention in a shared virtual space with a fixed coordinate system [43]. For dissimilar spaces, existing solutions retarget avatars' pointing and gazing directions [55] or head gaze directions [34] to correct the attention. However, it does not apply to video avatars because manipulating the viewing angle of videos in real-time, especially for whole-body life-size portraits, is much harder than controlling virtual avatars since it requires novel view synthesis, for which SOTA solutions with relatively simple setups are only able to generate upper-body-only effects [1, 41, 65].

2D-screen-based solutions. To convey attention on 2D screens, the GAZE groupware system [66, 67] adopts a rotatable screen metaphor and GazeChat [23] uses reconstructed profile photos with rotatable eyes. OpenMic [25] indicates the intention to talk using 2D proxemic metaphors. The use of 2D spatial cues in these solutions is inspiring. However, they are less immersive compared to using life-size video avatars in AR. Moreover, 2D spatial cues are constrained in a screen-size 2D space with limited scalability. AR enables presenting intuitive 3D spatial cues similar to offline conversations, which can be further scaled due to the wide interaction space.

Symbolic visualizations. In task-oriented collaborative AR and VR scenarios, existing solutions often use explicit visualizations such as cursors, rays, circles, and cones to represent users' attention [15, 55, 56, 73]. While they can accurately express the attention relation, how such visualizations perform in user-centered conversational AR teleconferencing is unexplored. We thus adopt this kind of visualization for comparison in Sect. 4.

Additional instruments. One group of solutions uses multiple cameras to recreate the spatial relation, e.g., Hydra [62], MultiView [48], MMSpace [51], and others [18, 27, 53, 76]. Some design special displays such as eyeball- and face-shaped [45, 52] ones to present the user's gaze. There are also approaches based on big screens [4, 77, 78], 3D projection [29], and cylindrical displays [33, 53]. These solutions keep the 3D relation unaltered in teleconferencing to maintain spatial faithfulness and, intrinsically, present attention awareness. However, the requirements of many cameras and special displays make them not feasible for everyday usage. We propose decoupling the spatial relation into the looker's and lookee's sides. This allows the exploration of solutions in a mono-camera setup.

Telepresence robots. Actuated screens with swiveling motions are used in teleconferencing [51, 63] and desktop configurations [3, 61] to indicate a remote user's attention, leading to higher engagement in conversations [63]. However, the Monalisa effect [60] makes the rotation inaccurate, leading to errors in the perceived gaze direction from the actual screen orientation [31]. While the rotating motion is intuitive and has been proven efficient in these works on indicating the attention of users on screens, how it performs on life-size video avatars in AR is unexplored. Therefore, we explore this visualization as a potential solution in Sect. 4.

2.3 Adaptive UI in AR

Our goal is related to user-centered UI adaptation research in terms of dynamically adapting the transformation and visualization of virtual content regarding both the local and remote users' head gaze input. Lindlbauer et al. propose Context-Aware adaptation [37] to control UI components' position and visibility according to the user's cognitive load. Lu et al. proposed head-gaze- and eye-gaze-based Glanceable AR [16, 38, 39] to transform applications (e.g., calendar and weather) into the Field of View (FoV) of the HMD to enable information access at a glance. Wysopal et al. proposed level-of-detail AR [72] to adjust the object content based on a visual angle dynamically. Lu and Xu [40] explore the control of UI transitions and find the semi-automated mechanism to be the most preferred.

Unlike these works, we explore the AR teleconferencing application scenario, where the UI components to adapt are life-size video avatars. We adapt the transformation of video avatars semi-automatically and let the attention synchronizing controller take full control to broadcast attention and update visualizations.

3 Head Gaze Retargeting

As mentioned previously, the looker's side-looking image caused by the mono-camera setup would confuse the lookee and reduce the attention awareness of the group (Fig. 2). To address this issue, we conduct the first study to explore adaptation approaches that transform the lookee's video avatar to align with the local camera to retarget the local user's head gaze. The results of this study answer RQ1 and distill design choices for the proof-of-concept prototype.

3.1 Adaptive Video Avatar Layouts

To explore how video avatars should be centered regarding the local user's looking direction, we need to determine 1) the initial position to place teleconferencing video avatars and 2) the center position of a lookee's video avatar. We keep only one video avatar centered at a time. When a video avatar transforms to the center, the previous one at the center will return to its initial position to make room.

There are certain Facing Formations (F-Formations) in physical group conversations [42]. However, for HMD-based AR teleconferencing, the existing quantitative studies in [69] have found specific circular video avatar layouts with larger radii than physical F-Formations. Taking the local user as the origin and the right and front direction respectively as the positive *x-axis* and *y-axis*, we adopt the following video avatar initial placement data from the previous study [69] and set the corresponding center top positions ($/m$): (0.0, 2.2) for 1 video avatar; (-0.7, 2.0) and (0.7, 2.0) for 2 video avatars, leaving the center top of the circle at (0.0, 2.2); (-1.1, 1.8), (-0.5, 2.4), and (1.1, 1.8) for 3 video avatars, center top at (0.0, 2.5); and (-1.4, 1.6), (-0.6, 2.6), (0.6, 2.6), and (1.4, 1.6) for 4 video avatars, center top at (0.0, 2.8). With these data, we develop the Social-Distance-First (**SDF**) layout (Fig. 3 (a)) that naively sets video avatars' initial positions according to the user-generated data and the center position at the center top of the circle, as listed above.

However, it is difficult to keep multiple video avatars in the FoV in **SDF**. It could potentially harm the perception of attention since users need to rotate their heads horizontally more often to keep aware of the periphery due to HMD's small FoV. Targeting this phenomenon, we develop the FoV-First (**FoVF**) layout (Fig. 3 (b)). In this layout, we keep the radius of the circular layout unchanged since it majorly determines the vertical visible range of a video avatar and gather video avatars more toward the center top of the circle to keep all of them within the FoV. With the horizontal FoV of the HMD we used in this work (Microsoft HoloLens 2) being 43°, video avatars'

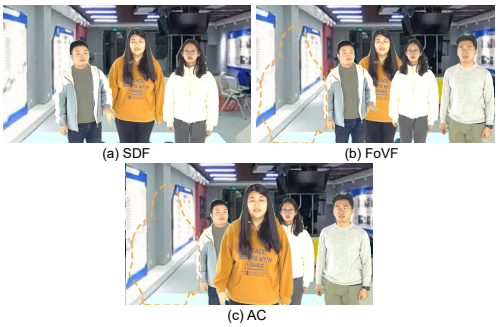


Figure 3: Study conditions in Study 1 (4RU scenario). The images present the local user's view when looking at User 1 (the girl in orange, whose initial position is on the leftmost) in (a) **SDF**, (b) **FoVF**, and (c) **AC** layouts. User 1's initial position is marked as orange dashed silhouettes (not visible to the user) in (b) and (c) while it is out of view in (a) along with the rightmost user. More illustrations can be found in the supplemental file (Fig. 1) and video.

original positions ($/m$) are set to (-0.7, 2.0) and (0.7, 2.0) for 2 video avatars; (-0.9, 2.2), (-0.3, 2.4), and (0.9, 2.2) for 3 video avatars; and (-0.9, 2.4), (-0.4, 2.8), (-0.4, 2.8), and (0.9, 2.4) for 4 video avatars. The center position remains the same as in **SDF**.

Moreover, we consider a hybrid layout that maintains all video avatars within the FoV by setting the initial positions the same as in **FoVF** and keeps the one-to-one social distance between the user and the lookee's video avatar by setting the center position as the 1-video-avatar position. Because the center position in this layout is on a smaller circle instead of the original one, we call it the Attention Circle (**AC**) layout, as shown in Fig. 3 (c) and Fig. 1.

We conduct our first user study with these three layouts as the conditions to explore how they perform in retargeting the local user's head gaze. We show their dynamic effects in the supplemental video.

3.2 Participants

We recruited 16 graduate students (11 males and 5 females; average age: 25.6 (SD = 2.4)) from the local campus as the participants. The sample size (along with that in the subsequent user studies in Sect. 4 and Sect. 6) is in line with the suggestion by relevant research in HCI studies [11] and is thus believed to be sufficient for drawing our conclusions. We recruited young college students as participants since they are sensitive to XR technology and are our main target users. It is also a common sample composition in other AR/VR studies [32, 71, 74]. We mainly focused on the diversity of participants' AR/VR experiences. Among them, 3 had no prior AR/VR experience, 6 had experienced several times, 1 had used HMDs extensively, and 6 were AR/VR application developers.

3.3 Study Scenario and Experiment Setup

We conduct our pilot study in the small-group AR teleconference scenario, where users are located in different spaces and join a teleconference using their AR HMDs as video avatars. We set three scenarios with differences only in group sizes. We denote them as 2RU, 3RU, and 4RU scenarios, including 2, 3, and 4 remote users (a total of 3, 4, and 5 participants, with the local user included), respectively. In each scenario, participants need to finish a task related to a lookee's awareness. In the task, remote users' video avatars wave hello to the local user (the participant) one by one as an explicit non-verbal message expressing their attention to the participant. Participants need to look at the waving avatar and activate the centering effect as soon as they notice the greeting.

We use Microsoft HoloLens 2 to present the AR experience. We pre-record the videos of every remote user greeting the local user and remove the background using MobileNetV3 [24]. To ensure the video avatars are indeed life-size, we calibrate their sizes by referring to the corresponding users through HoloLens.

3.4 Metrics

We record the Reaction Time for each video avatar position (ordered from left to right starting from "Pos 1") as the objective task performance measurement. It is calculated as the period from the start of each greeting action to the moment the participant successfully looks at the greeting video avatar and activates the centering effect. For subjective measurements, we adopt 1) NASA-TLX [22] task load metrics (six in total) and 2) three metrics regarding the teleconference experience referring to previous works [12, 30, 68, 75], namely Reasonability, sense of Co-Presence, and Preference. Reasonability indicates how reasonable the video avatar layout is. We describe Co-Presence according to previous measurements [49]. Preference is an overall subjective rating of the experience. We present the nine metrics to the participants using five-point Likert scales ranging from 1 to 5. For task load metrics, the lower the score, the better, while for teleconferencing experience metrics, the higher, the better.

3.5 Procedure

First, we introduce the participants the basic concept of AR teleconferencing and tell them their role as the local user. We then introduce to them the task described in Sect. 3.3 and the metrics illustrated in Sect. 3.4. Afterward, we walk them through a practice task to get familiar with the operation and the experience. After completing the preparatory work, we proceed to the formal experiment process.

The experiment has 9 trials (3 layout conditions $\times$ 3 scenarios). The layout condition order and the video avatar greeting order are counterbalanced hierarchically using a Balanced Latin Square. After finishing all 9 trials, we have a 5-minute semi-structured interview with each participant, discussing how they feel during the experiment, what they think is the major reason for the described experience, and suggestions on improving the system.

We make the following hypotheses (H2 and H3 are in scenarios where the effects of the three conditions are significant):

H1: The significance of the difference among the conditions will increase with the teleconferencing group size.

H2: **AC** and **FoVF** will have significantly less reaction time and task load than **SDF**.

H3: **AC** will be significantly more preferred than **FoVF** and **SDF**, with significantly higher reasonability and co-presence scores.

3.6 Results

With Shapiro-Wilk normality tests finding non-normal distributions in the sample, we conducted non-parametric Friedman Tests and Post-hoc Nemenyi Tests to compare the objective reaction time and the subjective ratings, respectively. The same analysis process is applied in Sect. 4.4 and Sect. 6.4.1 as well. We will elaborate on specific results under the confidence interval of $p < .05$ in these analysis sections. We mark results with significant difference between conditions at the top of each figure ($* \; p < .05$, $** \; p < .01$, $*** \; p < .001$) in Fig. 4 and the subsequent Fig. 6 and Fig. 7.

Reaction Time We show statistical comparisons of reaction time in the 4RU scenario in Fig. 4 (c) and the comparison of data in the 2RU and 3RU scenarios in Fig. 2 and 3 in the supplemental results. We list the *Means* (*M*s) and *Standard Deviations* (*SD*s) of reaction time on each position in each scenario in Table 1 in the supplemental results. Friedman Tests show significant effects of the three conditions only on position 1 in the 4RU scenario (χ^2 $statistic = 8.69$, $p = .013$), but not in the other positions and scenarios. We discuss this lack of significance in Sect. 3.7. Post-hoc Nemenyi Test results on data from position 1 in the 4RU scenario show that **AC** has significantly less reaction time than **SDF** ($p = .011$). There are no significant differences between **FoVF** and **SDF** ($p = .112$) and **FoVF** and **AC** ($p = .643$).

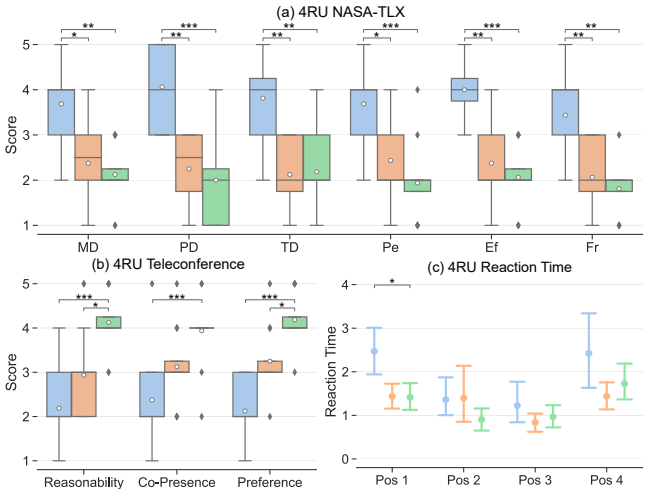


Figure 4: The comparison of **SDF**, **FoVF**, and **AC** layout conditions regarding (a) NASA-TLX scores, (b) teleconferencing experience scores, and (c) reaction times in the 4RU scenario in Study 1. **AC** and **FoVF** result in generally less reaction time and significantly lower task load than **SDF**. **AC** leads to a significantly better teleconferencing experience than **FoVF** and **SDF**. We discuss the results more in Sect. 3.6 and Sect. 3.7.

Subjective Ratings Friedman Tests show significant effects of the three conditions for all metrics in NASA-TLX and teleconferencing experience metrics in 3RU and 4RU scenarios, but not in the 2RU scenario. We summarize the resulting Friedman p values and χ^2 *statistics* (rows 1 and 2) and Post-hoc Nemenyi Test p values (rows 3-5) for the 3RU and 4RU scenarios in Tables 5 and 6 in the supplemental results, respectively. Statistical comparisons of NASA-TLX and teleconferencing experience data from the 4RU scenario are shown in Fig. 4. We list the Ms and SDs of each metric in each scenario in Tables 2, 3, and 4, respectively, in the supplemental results. Statistical comparisons of data from the 2RU and 3RU scenarios are shown in Fig. 2 and 3 in the supplemental results.

NASA-TLX. From the results shown in Fig. 4 (a) in the main paper and Tables 5 and 6 in the supplemental results, we can see that there are no significant differences in all task load metrics between **AC** and **FoVF** in both 3RU and 4RU scenarios. **AC** has significantly less task load in all metrics than **SDF** in the 4RU scenario, and still has significantly lower physical and temporal demand, effort, and frustration than **SDF** in the 3RU scenario. **FoVF** also has significantly less task load in all metrics than **SDF** in the 4RU scenario, but has only significantly less effort in the 3RU scenario.

Teleconference experience. From the results shown in Fig. 4 (b) in the main paper and Tables 5 and 6 in the supplemental results, we can see that there are no significant differences in reasonability, co-presence, and preference scores between **SDF** and **FoVF** in both 3RU and 4RU scenarios. **AC** presents a significantly higher level of co-presence and is significantly more reasonable and preferred than **SDF** in both 3RU and 4RU scenarios. While having no significant difference in the 3RU scenario, **AC** is perceived as more reasonable and is more preferred than **FoVF** in the 4RU scenario.

3.7 Discussion

H1 is supported. The effects of the conditions are not significant in the 2RU scenario, but are significant in the 3RU and 4RU scenarios. The participants "can barely tell the difference when there are only two remote users", and can clearly feel the difference in the 3RU and especially 4RU scenarios. This indicates that in multi-user HMD-based AR teleconferencing, when there are three or more remote users, the design of applications would often require additional considerations compared to scenarios with only one or two remote users. Accounting for the reason, in AR teleconferencing, the main virtual content is the "teleported" remote user, which is commonly presented in life-size. When the number of remote users increases to three or more, the virtual content would be too much to present due to the limitations of the AR HMD such as its limited FoV. This becomes prominent to affect users' attention awareness and teleconferencing experience.

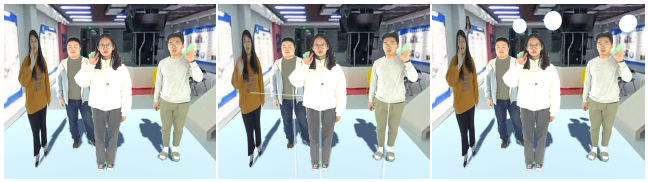
H2 is partially supported. Both **FoVF** and **AC** support faster reaction and have lower task-load-related scores than **SDF** when there are more than two remote users. However, the analysis only shows a significant difference in reaction time on position 1 in the 4RU scenario. We consider the main reason for the lack of significant difference is that occasionally, the participants are already looking at the remote user who will be the next one to wave hello since they are allowed to freely look at anyone during the study. Under this circumstance, the logged reaction time is only a very small number under 0.1s. This happened to nearly every participant several times in random conditions, increasing the similarity of distributions in the sample and thus, reducing the difference. Nonetheless, we can see clearly longer reaction time of **SDF** than **FoVF** and **AC** on the side positions (e.g., positions 1 and 4 in the 4 RU scenario) from Fig. 4. So in line with H2, constraining remote users' video avatars within the FoV could significantly reduce the time and effort to realize "who is looking at me". When there are more than two remote users, **SDF** is quite physically demanding and can cause obvious fatigue and even sickness. This can be seen from the reflection of some participants: "Although I feel they would stand more evenly on a circle (referring to **SDF**) in reality, it is too tiring to rotate my head around to check everyone's status".

H3 is supported. According to the reflections from the participants, placing the center position in a one-to-one conversation distance in AC can provide the "stepping forward or onto a stage to talk" metaphor in real life, the "dialog box switching" metaphor in games, and the "amplification effect in the MacOS dock" 2D UI metaphor. The participants are unaware of the original reason why the system needs to move users to the center, so in **SDF** and **FoVF**, they would ask why remote users need to go to the center. But in **AC**, "its (AC's) real-life metaphor gives such adaptation a meaning, which greatly reduces the visual complexity caused by constantly moving users to the center". This means AC can conceal the true reason for this adaptation (to correct the capturing perspective of the local user for remote users) with its reasonable and natural metaphor, making the process much less obtrusive and noticeable. Meanwhile, AC could make attention information access easy enough because "other users (remote users who are not being looked at by the local user) stand right in the back, waiting to be looked at". This is considered to increase the reasonability and co-presence: "It is very important and natural to be able to quickly notice who is looking at me at a glance, similar to offline conversations".

In conclusion, the results show that AC is the best layout to adaptively transform the video avatar the local user is looking at to the center to align with the local camera, and thus, retarget the user's head gaze. When there are more than two remote users, it leads to significantly easier and faster access to the attention status of remote users and makes the adaptation less obtrusive and more reasonable with its natural metaphor.

4 Attention Visualization

Following the reasoning in Sect. 1, after the head gaze retargeting, onlookers may still be confused about the looker's attention. We thus conducted the second study to further explore visualizations to help onlookers tell the attention relations. The results of this study help us to answer RQ2 and distill visualization techniques for the proof-of-concept.



(a) Rotation (b) Arrow (c) Thumbnail

Figure 5: Study conditions in Study 2 (4RU scenario). The images present the effects of (a) **Rotation**, (b) **Arrow**, and (c) **Thumbnail** visualizations. **Baseline** is not shown here since it only has the "AC" adaptation from Study 1 with no attention visualization. More illustrations can be found in the supplemental file (Fig. 4) and video.

4.1 Techniques to Visualize Attention

Due to the conversational and interaction-oriented feature of our video-avatar-based AR teleconferencing scenario, the design of the attention visualization should follow the principle of being efficient, accurate, and unobtrusive.

As introduced in Sect. 1 and Sect. 2.2, swiveling motions of telepresence robots are intuitive and have been studied to benefit attention awareness in teleconferencing. The effect of AR video avatars is similar to the ultimate teleconferencing robot, which is equipped with life-size transparent screens and no extra mechanical parts. Therefore, we develop the **Rotation** visualization for video avatars following the telepresence robot design to indicate attention by a video avatar's orientation (Fig. 5 (a)). The ability to rotate in the 3D space like billboards makes the 2D video avatar 2.5D. However, as discussed in Sect. 2.2, rotation itself cannot be perceived accurately due to the Monalisa effect, and could potentially be more confusing in multi-user scenarios. It is thus not enough to present more subtle changes in attention. Referring to symbolic visualizations used in task-oriented collaborative scenarios (discussed in Sect. 2.2), we further add annotations to the rotation motion to improve the accuracy while keeping the head rotation metaphor. A representative symbol for orientation indication is the **Arrow**. An arrow explicitly identifies both the starting and ending points of the looking action, i.e., pointing from the looker to the lookee (Fig. 5 (b)). However, this annotation on both sides might introduce extra visual complexity and could bring distraction and discomfort since it is quite obtrusive. Informed by previous work in robotics [10], we develop an **Attention Thumbnail** (simply referred to as **Thumbnail**) that only adds annotation on the starting point to avoid being obtrusive. It denotes who a remote user is looking at by showing the corresponding lookee's photo and name on the upper right to the looker's video avatar, as shown in Fig. 1 and Fig. 5 (c). When the remote user is looking at the local user, we highlight the thumbnail in a white circle to be as explicit as possible to further improve the local user's awareness as the lookee.

We conducted the second user study to explore how these visualizations perform in improving an onlooker's awareness. We set the approach having only the previous head gaze retargeting but no attention visualization as **Baseline** (see Fig. 4 (a) in the supplemental results). We show their dynamic effects in the supplemental video.

4.2 Study Scenario and Experiment Setup

We recruited 16 participants (12 males and 4 females; average age: 25.0 (SD = 2.6)). Among them, 2 had little prior AR/VR experience, 9 had experienced AR/VR several times, and 5 were AR/VR developers. Most (14) of them are from Study 1 and joined this study as a follow-up. We believe there is no significant bias since there is no overlap in the independent variables of these two studies, and the design considerations are clearly distinct.

The scenarios and setup are the same as in Study 1, only the task is different. In the task, remote users' video avatars wave hello one by one to one of the users in the teleconferencing, maybe to the local user (the participant) or other remote users. The participants need to identify the information about who is greeting whom as quickly and accurately as possible and press the space key on a keyboard the moment they have the answer in mind. After pressing the space key, they report their answer to us for subsequent error rate analysis.

We record the Reaction Time and Error Rate as the objective task performance measurement. Reaction Time is calculated as the period from the start of the greeting action to the moment the participants press the space key. We calculate the average Error Rate by comparing the participants' answers with the ground truth. We adopt the subjective metrics used in Study 1 (Sect. 3.4).

4.3 Procedure

The preparatory procedure and the post-study interview are similar to those in Study 1. The experiment has 12 trials (4 attention visualization conditions $\times$ 3 scenarios). The order of the 4 visualization conditions, and the looker's and lookee's order are counterbalanced hierarchically using Balanced Latin Squares.

We make the following hypotheses (H5 and H6 are in scenarios where the effects of the three conditions are significant):

H4: All three visualizations will have significant lower error rates than **Baseline**.

H5: There will be no significant difference among the conditions in the 2RU scenario, but **Arrow** and **Thumbnail** will have significantly lower error rates, shorter reaction times, and lower task loads than **Rotation** in the 3RU and 4RU scenarios.

H6: **Thumbnail** will be significantly more reasonable and preferred, with little harm to the sense of co-presence.

4.4 Results

Objective Performance **Error rate.** The error rates in 2RU, 3RU, and 4RU scenarios respectively are 50.0, 66.7, 75.8 for **Baseline**, 0.0, 18.1, 30.5 for **Rotation**, 0.0, 1.4, 6.3 for **Arrow**, and 0.0, 1.4, 1.9 for **Thumbnail**. We can see that while both **Arrow** and **Thumbnail** have very few errors, **Baseline** leads to very high error rates since it has no attention awareness. Participants feel like all remote users are looking at them all the time in **Baseline**. The error rate of **Rotation** is low in the 2RU scenario, but increases dramatically in the 3RU and 4RU scenarios.

Reaction time. We show statistical comparisons of reaction times in the 4RU scenario in Fig. 6 (c) and the comparison of data in the 2RU and 3RU scenarios in Figs. 5 and 6 in the supplemental results. We list the *M*s and *SD*s of reaction times on each position in each scenario in Table 7 in the supplemental results. Friedman Tests show significant effects of the three conditions on position 2 in the 3RU scenario ($\chi^2 = 9.04$, $p = .011$) and position 2 ($\chi^2 = 16.41$, $p < .001$), position 3 ($\chi^2 = 9.66$, $p = .008$), and position 4 ($\chi^2 = 8.38$, $p = .015$) in the 4RU scenario. Post-hoc Nemenyi Test results show that **Thumbnail** has significantly shorter reaction times than **Rotation** on all these positions ($p = .009$, .001, .010, and .013 for position 2 in the 3RU scenario, and positions 2, 3, and 4 in the 4RU scenario, respectively). There are no significant differences between **Rotation** and **Arrow** on all these positions ($p = .128$, .382, .840, and .126, respectively). **Thumbnail** has significantly shorter reaction times than **Arrow** on positions 2 and 3 in the 4RU scenario ($p = .022$ and .045, respectively) but has no significant difference on position 4 in the 4RU scenario and position 2 in the 3RU scenario ($p = .638$ and .560, respectively).

Subjective Ratings. Since **Baseline** could achieve no attention-awareness, we analyzed the rest three conditions for comparisons. Friedman Tests show significant effects of the three conditions for all metrics in NASA-TLX and teleconferencing experience metrics in the 3RU and 4RU scenarios (except for PD in the 3RU scenario) but not in the 2RU scenario. We summarize the resulting Friedman p values and χ^2 *statistics* (rows 1 and 2) and Post-hoc Nemenyi Test

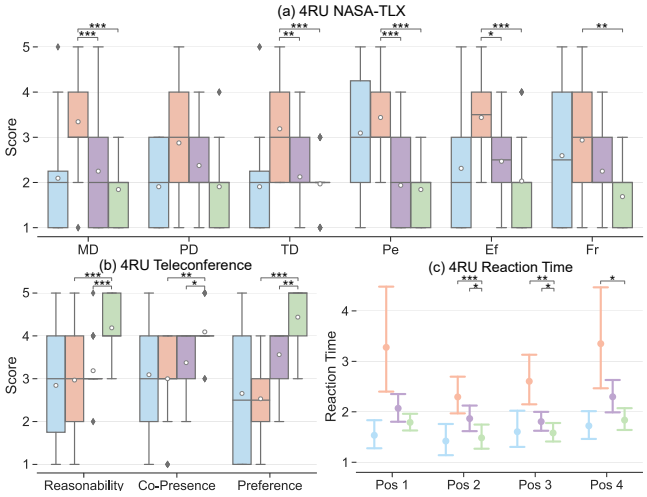


Figure 6: The comparison of **Baseline**, **Rotation**, **Arrow**, and **Thumbnail** conditions regarding (a) NASA-TLX scores, (b) teleconference experience scores, and (c) reaction times in the 4RU scenario in Study 2. **Arrow** and **Thumbnail** result in significantly lower task load than **Rotation**. **Thumbnail** leads to a significantly better teleconferencing experience and less reaction time than **Arrow** and **Rotation**. We discuss the results more in Sect. 4.4 and Sect. 4.5.

p values (rows 3-5) for the 3RU and 4RU scenarios in Tables 11 and 12 in the supplemental results, respectively. Statistical comparisons of NASA-TLX and teleconferencing experience data from the 4RU scenario are shown in Fig. 6. We list the *M*s and *SD*s of each metric in each scenario in Tables 8, 9, and 10, respectively, in the supplemental results. Statistical comparisons of the 2RU and 3RU data are shown in Figs. 5 and 6 in the supplemental results.

NASA-TLX. From the results shown in Fig. 6 (a) in this paper and Tables 11 and 12 in the supplemental results, we can see no significant differences in all task load metrics between **Thumbnail** and **Arrow** in both 3RU and 4RU scenarios. **Thumbnail** has a significantly lower task load in almost all metrics than **Rotation** in both 3RU and 4RU scenarios (except for MD in the 3RU scenario and PD in the 4RU scenario). **Arrow** also has a significantly lower task load in all metrics except PD and Fr than **Rotation** in the 4RU scenario and has significantly better performance with significantly less effort in the 3RU scenario.

Teleconference experience. From the results shown in Fig. 6 (b) in this paper and Tables 11 and 12 in the supplemental results, we can see no significant differences in reasonability, co-presence, and preference scores between **Rotation** and **Arrow** in both 3RU and 4RU scenarios, except that **Arrow** is significantly more preferred than **Rotation** in the 4RU scenario. **Thumbnail** is significantly more reasonable and preferred than both **Rotation** and **Arrow** in both 3RU and 4RU scenarios. The level of co-presence of **Thumbnail** is significantly higher than **Rotation** in both 3RU and 4RU scenarios but is only significantly higher than **Arrow** in the 4RU scenario.

4.5 Discussion

H4 is supported. The participants struggled to identify whom other users were looking at when there were no visual cues. It is necessary to present additional effects to aid attention awareness in AR teleconferences.

H5 is supported. In the 2RU scenario, **Rotation** is clear enough to identify a user's attention. "It would be redundant to show arrows and thumbnails here". In line with H1 and the corresponding discussion in Sect. 3, situations would differ when the number of remote users exceeds two in HMD-based AR teleconferencing. In the 3RU and 4RU scenarios, **Rotation** is not clear. Participants would spend more time and effort to get the right answer ("I need to switch the user's position between center and back several times to determine who he/she is looking at") or give unconfident guesses ("I tend to think he/she is looking at the nearest user in that direction"). **Arrow** has clearly lower error rates and relatively shorter reaction times than **Rotation**. The lack of significant differences in reaction times between them might be due to the common situation where participants give answers with low confidence but in a relatively short time, which can be reflected in error rates. **Thumbnail** has significantly lower error rates and shorter reaction times than **Rotation**, and also significantly outperforms **Arrow**. This is because **Thumbnail** is more intuitive and less visually complex than **Arrow**. "Checking the start and end of the arrow is cumbersome while the thumbnail only requires a simple glance."

H6 is supported. Thumbnail is significantly more reasonable because, similar to real conversations, we can directly infer a person's attention by only checking the person, who is the start point of the "looking" action. It is also a metaphor for presenting a person's thoughts in a bubble above their head in cartoons. "It (thumbnail) is very intuitive and familiar because this kind of visualization is usually used to describe one is thinking about a specific person." It is also not obtrusive and disrupting while **Arrow** "feels quite strange and messy no matter where you put them". Therefore, **Thumbnail** is the most preferred visualization.

In conclusion, the results show that **Thumbnail** is the best visualization to help identify other users' attention. It leads to the best performance, significantly lower task loads, and the most satisfying teleconference experience.

5 Implementation

With the "Attention Circle" layout obtained from Study 1 in Sect. 3 addressing the lookee's attention awareness and the "rotatable 2.5D video avatars with attention thumbnail" visualization obtained from Study 2 in Sect. 4 achieving the onlooker's attention awareness, we implement A3RT, an attention-aware video-avatar-based multiparty AR teleconferencing system. We introduce the three major components of this system namely, the teleconferencing component, the head gaze retargeting component, and the attention synchronizing component. We developed the full interactive system on Microsoft HoloLens 2 OST AR HMD using Unity 2019.4. The full system supports a four-user teleconferencing experience using four gaming laptops (Intel Core i9 CPU and NVIDIA GeForce RTX 3080 Laptop GPU) and four HoloLens 2 HMDs. Each user wears a HoloLens 2 to render the remote users' video avatars, with a laptop and its integrated webcam in the front, capturing and processing his/her real-time image. It is accessible and effortless to set up, similar in the current video conferencing, and can be easily scaled to more users. We introduce the three components of the system below.

5.1 Teleconferencing Component

The teleconferencing component supports the basic video-avatar-based teleconferencing experience. With the video captured by the webcam, a background removal algorithm [24] runs on the PC to separate the human from the background in real-time. We build point-to-point TCP connections to broadcast the matted human-only video from one user to all others frame by frame. Symmetrically, each user receives other users' image frames through another group of TCP connections. Upon receiving the images, the HoloLens updates the texture for corresponding video avatars. The bandwidth requirement for a single video avatar channel is around 200Kb/s.

We manually adjust virtual light sources to match the physical lighting and set a virtual floor to receive shadows. This makes the video avatar's lighting and shadowing effects similar to those of physical objects in the AR environment (as shown in Fig. 1). This process can be further automated following relative research [58].

5.2 Head Gaze Retargeting Component

The head gaze retargeting component implements the Attention Circle design obtained from Study 1. We choose head gaze instead of eye gaze as the input to determine where a user is looking and the video avatar selection technique for two reasons. First, eye tracking on HMD still suffers from jitter problems, which would result in erroneous activation. Second, the rotation of the video avatar billboard is perceived as a metaphor more similar to the head rotation rather than saccades of eyes, which increases behavioral consistency and reasonability of 2.5D video avatars. To determine whether the user is looking at a video avatar (select) or just looking past one (not select), we adopt a speed-based target selection approach to avoid the "Midas Touch Effect" [28]. Specifically, the system determines the user is looking at a video avatar when the head gaze intersects with the video avatar and the current head rotation speed is lower than a threshold ($6^\circ/s$ in our implementation). Since the group size in our current implementation is 4, we adopt the corresponding layout parameters ($/m$) derived from the previous studies in [69] and introduced in Sect. 3.1, i.e., (-1.1, 1.8), (-0.5, 2.4), and (1.1, 1.8) for the 3 video avatars and AC center top at (0.0, 2.2), in the implementation. The retargeting is semi-automatic, with the looking action initiated by the user. The video avatar then automatically transforms to the center in $0.4s$ with an initial velocity of $5D_i\ m/s$ and an acceleration of $-12.5D_i\ m/s^2$, where D_i is the distance between video avatar i's initial position and the center position. Simultaneously, the previous video avatar on the attention circle transforms back to its initial position, keeping only one centered video avatar at a time.

5.3 Attention Synchronizing Component

The attention synchronizing component is used to keep the attention relations between users consistent in all users' views and update the visualization according to the design obtained from Study 2 in Sect. 4. The head gaze retargeting component updates the local user's attention in real-time, and the attention synchronizing component broadcasts it to all others through another group of point-to-point TCP connections. Meanwhile, each user maintains a directed global attention graph to cache all users' attention. It serves as the reference to update local 2.5D video avatars' orientations and attention thumbnails. The synchronizing component updates the corresponding edges in the graph upon receiving the attention message from other users. Due to this decoupled design, our system is robust to the heterogeneity of spaces [32, 68, 74].

6 Evaluation

After developing the A3RT running prototype, we conducted a user study to evaluate the effectiveness of each component and the full system through an ablation study and a usability study, respectively.

6.1 Participants

We recruited 14 graduate students (10 males and 4 females; average age: 27.4 (SD = 2.6)) from the local campus as the participants. Among them, 1 had no prior AR/VR experience, 9 had experienced AR/VR several times, and 4 were AR/VR developers. We invited 4 participants of Study 1 to join the evaluation. We believe this involves no significant bias since the number of reused users is small and video avatars in Study 1 are pre-recorded, but the evaluation here is based on a real-time running system.

6.2 Study Conditions

For the ablation study, we have four conditions. **Full System** is the complete system with all components. **Thumbnail Only** removes the head retargeting component, leaving only the attention visualization. **Centering Only** removes the attention visualization component, leaving only the head retargeting feature. **Basline** removes both components, leaving only the basic video-avatar-based experience. For the usability study, we use the condition selected to be the best by the participants in the ablation study (all participants chose the **Full System**).

6.3 Study Scenario and Experiment Setup

We divide the participants into 2 groups of 4-user teleconferencing and 2 groups of 3-user teleconferencing for the evaluation. The preparatory procedure and the interview are the same as those in Study 1. There are two tasks: one is focused on attention awareness, and the other is a conversational task. In the first task, all participants are asked to randomly choose a remote user to look at, and we ask one of the users to report other users' attention and get feedback from them on whether the corresponding report is correct. Users change the target in each round until they go through every video avatar and then proceed to the next user to report until finished. We record the number of errors for the subsequent calculation of error rates. In the second task, the participants are asked to play a debate-like conversational game referring to previous research [23].

We use Error Rate as the objective task performance measurement. Subjective metrics for the first task are the same as those in Study 1 in Sect. 3.4. We adopt the System Usability Scale (SUS) [9] as the measurement for the second task.

The experiment setup follows the introduction in the implementation in Sect. 5. We designated each participant a user number as the name and the thumbnail for simple reference. The participants first complete the attention-related task and rate the conditions regarding the task load and the teleconferencing experience. After they select the best functioning condition, they play the debate-like game using the system in the selected condition and rate the SUS metrics.

6.4 Results

6.4.1 Ablation Study

The error rate is 58.3% for **Baseline**, 59.4% for **Centering Only**, and 0.0% for both **Thumbnail Only** and **Full System**.

Friedman Tests show significant effects of the four conditions for all metrics in NASA-TLX and teleconferencing experience metrics. We summarized the resulting Friedman p values and χ^2 $statistics$ (rows 1 and 2) and Post-hoc Nemenyi Test p values (rows 3-8) in Table 14 in the supplemental results. Statistical comparisons of NASA-TLX and teleconferencing experience data are shown in Fig. 7. We list the Ms and SDs of each metric in Table 13 in the supplemental results.

NASA-TLX. From the results shown in Fig. 7 (a) in the paper and Table 14 in the supplemental results we can see that there are no significant differences in all task load metrics between **Baseline** and **Centering Only** or between **Full System** and **Thumbnail Only**. **Thumbnail Only** has significantly lower temporal demand, better performance with less effort and frustration than **Baseline** and **Centering Only**. **Full System** is significantly less demanding than **Baseline** and **Centering Only** in all metrics except physical demand with **Centering Only**.

Teleconference experience. From the results shown in Fig. 7 (b) in the paper and Table 14 in the supplemental results we can see that there are no significant differences between **Baseline** and **Centering Only**, **Centering Only** and **Thumbnail Only**, or **Thumbnail Only** and **Full System**. **Full System** is significantly better than **Baseline** and **Centering Only** in all metrics. **Thumbnail Only** has a higher level of co-presence and is more preferred than **Baseline**.

Discussion The results show the necessity of each component for achieving full attention awareness in AR teleconferencing. In the **Thumbnail Only** condition, users would not maintain frontal head orientations due to the ablation of the retargeting component. It did cause confusion to the lookee due to the "looking away" problem, which the retargeting component was designed to tackle. Most participants discovered the ambiguity where the thumbnail shows a remote user is looking at them but he/she is actually looking away. This would make the experience quite unreasonable. However, there

is only a borderline significant difference in reasonability between **Full System** and **Thumbnail Only**. We consider the reason to be the lack of direct eye contact caused by the AR HMD. Although HoloLens 2 is semi-transparent, it still occludes users' eyes, undermining the feeling of eye contact. This is confirmed by the participants: "It would feel more interactive if I could see other users' eyes." and "Because I cannot see their eyes, I would pay more attention to the thumbnail since it is more reactive when someone is looking around". It reduces the difference between **Full System** and **Thumbnail Only**. However, the participants agree that if the AR HMD becomes smaller and more transparent, such ambiguity would certainly become more obvious.

In **Centering Only**, users would no longer be able to tell other users' attention, similar to the Baseline condition in Sect. 4. "It feels like everyone is constantly looking at me with occasional random head rotations". Similar things occur in **Baseline**. "I can only see remote users looking in a certain direction with no meaning".

In conclusion, the ablation study verifies that our full system can truly support full attention awareness in AR teleconferencing, and each one of the two components is indispensable.

6.4.2 Usability Study

Fig. 7 (c) shows the SUS scores. The prototype system can achieve high scores in positive questions: Q1 (M = 4.3, SD = 0.6), Q3 (M = 4.5, SD = 0.6), Q5 (M = 4.6, SD = 0.5), Q7 (M = 4.5, SD = 0.5), Q9 (M = 4.6, SD = 0.5) and low scores in negative questions: Q2 (M = 1.4, SD = 0.6), Q4 (M = 1.7, SD = 0.9), Q6 (M = 1.4, SD = 0.5), Q8 (M = 1.6, SD = 0.6), Q10 (M = 1.4, SD = 0.6). Participants find the interaction intuitive and easy to get used to. "It uses only head rotation as the input to adapt the virtual content, requiring no special training." "Rotating my head to look around is what I would do in real-life conversations." Participants find the system quite responsive and engaging. "It feels good that I can notice someone is looking at me and look back to him/her and start talking." "The rotation makes the video avatar feel quite embodied by the remote user because I know when he/she is looking around". The participants think enabling eye tracking to replace the head gaze with the actual eye gaze as the input would make the system more agile and lively.

7 Conclusion, Limitations, and Future Work

In conclusion, we have explored the solution to achieving attention awareness in video-avatar-based AR teleconferencing in a mono-camera setup. We have decomposed such awareness from the lookee's and onlooker's perspectives. For the lookee's awareness, we have proposed to retarget the looker's head gaze to present a corrected head gaze and distilled the "Attention Circle" layout through the first user study. For the onlooker's awareness, we have proposed to visualize users' attention and distilled the "rotatable 2.5D video avatar with attention thumbnail" visualization through the second user study. Based on the obtained design choices, we have implemented the A3RT attention-aware multiparty AR teleconferencing prototype, consisting of a video-avatar-based teleconferencing component, a head gaze retargeting component, and an attention synchronizing component. It empowers full attention awareness in an unobtrusive, reasonable, intuitive, and efficient manner. We have validated the effectiveness of the system through ablation studies and usability studies. Next, we address the limitations of our work and discuss the corresponding potential future work.

Visibility of 2.5D video avatars. There are occasions where a video avatar rotates to around 90° to the local user and appears as a thin line, undermining the experience. To avoid this, we can further consider such visibility to dynamically optimize the layout. We can also explore designs that help improve the visibility of video avatars such as increasing the thickness of the billboards.

Large-scale and hybrid scenarios. In this work, we only focus on teleconference scenarios with less than 5 separately located users.

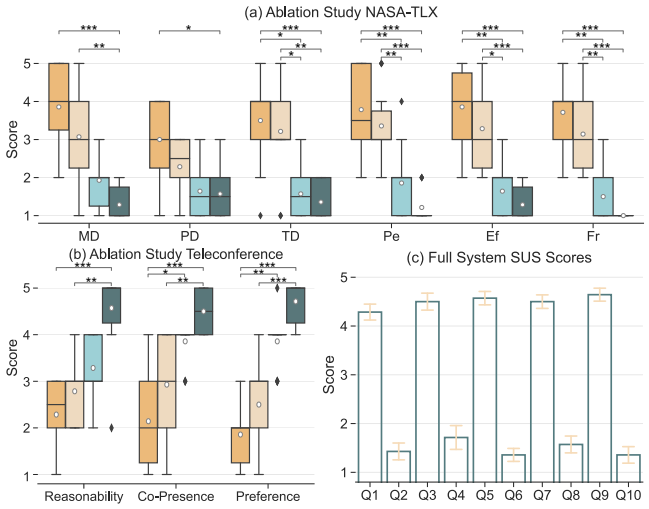


Figure 7: The comparison of **Baseline**, **Centering Only**, **Thumbnail Only**, and **Full System** conditions regarding (a) NASA-TLX and (b) teleconference experience scores in the ablation study. (c) SUS scores of the full system in the usability study. Odd-numbered SUS questions are positive (higher scores are better), while even-numbered ones are negative (lower scores are better). **Full System** results in significantly lower task load and better teleconferencing experiences than **Baseline** and **Centering Only**. **Thumbnail Only** results in significantly lower task load than **Baseline** and **Centering Only** and better teleconferencing experiences than **Baseline**. We illustrate and discuss the results more in Sect. 6.4.1.

We can further adapt our Attention Circle layout and the attention visualization to scenarios involving more users (e.g., tens and hundreds) with more than one co-located user. It will benefit broader applications such as augmented classrooms and mixed-reality events. Potential fatigue and distraction caused by more intense rotating and moving motions in these scenarios should be explored. We are also interested in exploring the system's performance among users with greater background diversity.

Egocentric setup. We explore the problem under a static camera setup. With the miniaturization of commodity 360° cameras and advances in image distortion correction techniques [19, 26], we can explore egocentric setups, e.g., attaching a 360° camera to the HMD or onto the user's body for capturing.

Interaction with the physical environment. We currently only focus on teleconferencing in the standing pose. We can explore more poses (e.g., sitting and leaning) or even allow video avatars to move around in the local space referring to avatar motion adaptation works [36, 68, 74]. We can relight video avatars [64] according to the local lighting condition to better immerse them in the physical environment.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and the interview and user study participants for their time. This work was supported by grants from the National Key Research and Development Program of China (No. 2022ZD0117103), the City University of Hong Kong (No. 9678338, 9667260), Innovation and Technology Commission (No. ITS/106/22), the National Natural Science Foundation of China (No. 62172326, 62137002, and 62192781), and Project of China Knowledge Centre for Engineering Science and Technology.

REFERENCES

[1] Apple. Apple vision pro. https://www.apple.com/apple-vision-pro.

[2] M. Argyle and M. Cook. Gaze and mutual gaze. 1976.

[3] G. Bailly, S. Sahdev, S. Malacria, and T. Pietrzak. Livingdesktop: Augmenting desktop workstation with actuated devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 5298–5310. Association for Computing Machinery, 2016.

[4] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013.

[5] Beem. Beem.me. https://beem.me.

[6] M. Billinghurst and H. Kato. Real world teleconferencing. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, p. 194–195, 1999.

[7] M. Billinghurst and H. Kato. Out and about—real world teleconferencing. *BT technology journal*, 18(1):80–82, 2000.

[8] C. W. Borst, N. G. Lipari, and J. W. Woodworth. Teacher-guided educational vr: Assessment of live and prerecorded teachers guiding virtual field trips. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 467–474, 2018.

[9] J. Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3):189–194, 1996.

[10] L. Brown, J. Hamilton, Z. Han, A. Phan, T. Phung, E. Hansen, N. Tran, and T. Williams. Best of both worlds? combining different forms of mixed reality deictic gestures. *J. Hum.-Robot Interact.*, 12(1), 2023.

[11] K. Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 981–992, 2016.

[12] Y. Choi, J. Lee, and S.-H. Lee. Effects of locomotion style and body visibility of a telepresence avatar. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1–9, 2020.

[13] H. H. Clark. *Using language*. Cambridge university press, 1996.

[14] X. Cui, D. Khan, Z. He, and Z. Cheng. Fusing surveillance videos and three-dimensional scene: A mixed reality system. *Computer Animation and Virtual Worlds*, p. e2129, 2022.

[15] S. D'Angelo and D. Gergle. An eye for design: Gaze visualizations for remote collaborative work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 1–12, 2018.

[16] S. Davari, F. Lu, and D. A. Bowman. Validating the benefits of glanceable and context-aware augmented reality for everyday information access tasks. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 436–444, 2022.

[17] F. De Simone, J. Li, H. G. Debarba, A. E. Ali, S. N. Gunkel, and P. Cesar. Watching videos together in social virtual reality: An experimental study on user's qoe. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 890–891, 2019.

[18] Ò. Divorra, J. Civit, F. Zuo, H. Belt, I. Feldmann, O. Chreer, E. Yellin, W. Ijsselsteijn, R. Van Eijk, D. Espinola, et al. Towards 3d-aware telepresence: Working on technologies behind the scene. *Proc. ACM CSCW: New Frontiers in Telepresence*, 2010.

[19] M. Elgharib, M. Mendiratta, J. Thies, M. Nießner, H.-P. Seidel, A. Tewari, V. Golyanik, and C. Theobalt. Egocentric videoconferencing. *ACM Transactions on Graphics*, 39(6), Dec 2020.

[20] H. Fuchs, A. State, and J.-C. Bazin. Immersive 3d telepresence. *Computer*, 47(7):46–52, 2014.

[21] Y. Furuya and K. Takashio. Telepresence robot blended with a real landscape and its impact on user experiences. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 406–411, 2020.

[22] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988.

[23] Z. He, K. Wang, B. Y. Feng, R. Du, and K. Perlin. Gazechat: Enhancing virtual conferences with gaze-aware 3d photos. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, p. 769–782, 2021.

[24] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[25] 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*, p. 17, 2023.

[26] T. Hu, K. Sarkar, L. Liu, M. Zwicker, and C. Theobalt. Egorenderer: Rendering human avatars from egocentric camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14528–14538, October 2021.

[27] J. A. Insley, D. J. Sandin, and T. A. DeFanti. Using video to create avatars in virtual reality. In *ACM SIGGRAPH 97 Visual Proceedings: The art and interdisciplinary programs of SIGGRAPH'97*, p. 128. 1997.

[28] R. J. Jacob. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in human-computer interaction*, 4:151–190, 1993.

[29] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. *ACM Trans. Graph.*, 28(3), jul 2009.

[30] S. Junuzovic, K. Inkpen, J. Tang, M. Sedlins, and K. Fisher. To see or not to see: A study comparing four-way avatar, video, and audio conferencing for work. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work*, p. 31–34, 2012.

[31] I. Kawaguchi, H. Kuzuoka, and Y. Suzuki. Study on gaze direction perception of face image displayed on rotatable flat display. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 1729–1737. Association for Computing Machinery, New York, NY, USA, 2015.

[32] D. Kim, J.-e. Shin, J. Lee, and W. Woo. Adjusting relative translation gains according to space size in redirected walking for mixed reality mutual space generation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 653–660, 2021.

[33] K. Kim, J. Bolton, A. Girouard, J. Cooperstock, and R. Vertegaal. Telehuman: Effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 2531–2540, 2012.

[34] T. Kim, A. Kachhara, and B. MacIntyre. Redirected head gaze to support ar meetings distributed over heterogeneous environments. In *2016 IEEE Virtual Reality (VR)*, pp. 207–208, 2016.

[35] J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021.

[36] C. Li, W. Li, H. Huang, and L.-F. Yu. Interactive augmented reality storytelling guided by scene semantics. *ACM Trans. Graph.*, 41(4), jul 2022.

[37] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 147–160. Association for Computing Machinery, 2019.

[38] F. Lu and D. A. Bowman. Evaluating the potential of glanceable ar interfaces for authentic everyday uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 768–777, 2021.

[39] F. Lu, S. Davari, L. Lisle, Y. Li, and D. A. Bowman. Glanceable ar: Evaluating information access methods for head-worn augmented reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 930–939, 2020.

[40] F. Lu and Y. Xu. Exploring spatial ui transition mechanisms with head-worn augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

[41] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. De la Torre, and Y. Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 64–73, June 2021.

[42] P. Marshall, Y. Rogers, and N. Pantidi. Using f-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative*

Work, p. 445–454, 2011.

[43] Meta. Horizon workrooms. https://www.oculus.com/workrooms.

[44] Microsoft. Microsoft mesh. https://www.microsoft.com/mesh.

[45] K. Misawa, Y. Ishiguro, and J. Rekimoto. Livemask: A telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, p. 394–397, 2012.

[46] A. F. Monk and C. Gale. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, 33(3):257–278, 2002.

[47] C. Neustaedter, S. Singhal, R. Pan, Y. Heshmat, A. Forghani, and J. Tang. From being there to watching: Shared and dedicated telepresence robot usage at academic conferences. *ACM Trans. Comput.-Hum. Interact.*, 25(6), dec 2018.

[48] D. Nguyen and J. Canny. Multiview: Spatially faithful group video conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, p. 799–808, 2005.

[49] K. L. Nowak and F. Biocca. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(5):481–494, 10 2003.

[50] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, p. 741–754, 2016.

[51] K. Otsuka. Mmspace: Kinetically-augmented telepresence for small group-to-group conversations. In *2016 IEEE Virtual Reality (VR)*, pp. 19–28, 2016.

[52] M. Otsuki, T. Kawano, K. Maruyama, H. Kuzuoka, and Y. Suzuki. Thirdeye: Simple add-on display to represent remote participant's gaze direction in video communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 5307–5312. Association for Computing Machinery, New York, NY, USA, 2017.

[53] Y. Pan and A. Steed. A gaze-preserving situated multiview telepresence system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, p. 2173–2176, 2014.

[54] Y. Pan and A. Steed. A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise. *Frontiers in Robotics and AI*, 3:12, 2016.

[55] T. Piumsomboon, G. A. Lee, J. D. Hart, B. Ens, R. W. Lindeman, B. H. Thomas, and M. Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2018.

[56] T. Piumsomboon, Y. Lee, G. Lee, and M. Billinghurst. Covar: A collaborative virtual and augmented reality system for remote collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies*, SA '17. Association for Computing Machinery, New York, NY, USA, 2017.

[57] M. J. Prins, S. N. B. Gunkel, H. M. Stokking, and O. A. Niamut. Togethervr: A framework for photorealistic shared media experiences in 360-degree vr. *SMPTE Motion Imaging Journal*, 127(7):39–44, 2018.

[58] T. Rhee, L. Petikam, B. Allen, and A. Chalmers. Mr360: Mixed reality rendering for 360° panoramic videos. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1379–1388, 2017.

[59] D. Robotics. Double robotics. https://www.doublerobotics.com.

[60] S. Rogers, M. Lunsford, L. Strother, and M. Kubovy. The mona lisa effect: Perception of gaze direction in real and pictured faces. *Studies in perception and action VII*, pp. 19–24, 2003.

[61] M. Sakashita, E. A. Ricci, J. Arora, and F. Guimbretière. Remotecode: Robotic embodiment for enhancing peripheral awareness in remote collaboration tasks. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22, 2022.

[62] A. Sellen, B. Buxton, and J. Arnott. Using spatial cues to improve videoconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, p. 651–652. Association for Computing Machinery, New York, NY, USA, 1992.

[63] D. Sirkin, G. Venolia, J. Tang, G. Robertson, T. Kim, K. Inkpen, M. Sedlins, B. Lee, and M. Sinclair. Motion and attention in a kinetic videoconferencing proxy. In *Human-Computer Interaction–INTERACT 2011, Lisbon, Portugal, September 5-9*, pp. 162–180. Springer, 2011.

[64] G. Song, T.-J. Cham, J. Cai, and J. Zheng. Real-time shadow-aware portrait relighting in virtual backgrounds for realistic telepresence. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 729–738, 2022.

[65] A. Trevithick, M. Chan, M. Stengel, E. R. Chan, C. Liu, Z. Yu, S. Khamis, M. Chandraker, R. Ramamoorthi, and K. Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023.

[66] 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*, CHI '99, p. 294–301. Association for Computing Machinery, New York, NY, USA, 1999.

[67] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, p. 521–528. Association for Computing Machinery, New York, NY, USA, 2003.

[68] X. Wang, H. Ye, C. Sandor, W. Zhang, and H. Fu. Predict-and-drive: Avatar motion adaption in room-scale augmented reality telepresence with heterogeneous spaces. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3705–3714, 2022.

[69] X. Wang, W. Zhang, C. Sandor, and H. Fu. Real-and-present: Investigating the use of life-size 2d video avatars in hmd-based ar teleconferencing, 2024. arXiv:2401.02171.

[70] Y. Wang, D. M. Krum, E. M. Coelho, and D. A. Bowman. Contextualized videos: Combining videos with environment models to support situational understanding. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1568–1575, 2007.

[71] J. W. Woodworth, N. G. Lipari, and C. W. Borst. Evaluating teacher avatar appearances in educational vr. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1235–1236, 2019.

[72] A. Wysopal, V. Ross, J. Passananti, K. Yu, B. Huynh, and T. Höllerer. Level-of-detail ar: Dynamically adjusting augmented reality level of detail based on visual angle. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 63–71, 2023.

[73] Y. Yan, H. Liu, Y. Shi, J. Wang, R. Guo, Z. Li, X. Xu, C. Yu, Y. Wang, and Y. Shi. Conespeech: Exploring directional speech interaction for multi-person remote communication in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2647–2657, 2023.

[74] L. Yoon, D. Yang, J. Kim, C. Chung, and S.-H. Lee. Placement retargeting of virtual avatars to dissimilar indoor environments. *IEEE Transactions on Visualization and Computer Graphics*, 28(3):1619–1633, 2022.

[75] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth. Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, 2021.

[76] S. Yura, T. Usaka, and K. Sakamura. Video avatar: embedded video for collaborative virtual environment. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, vol. 2, pp. 433–438 vol.2, 1999.

[77] Y. Zhang, Z. Li, S. Xu, C. Li, J. Yang, X. Tong, and B. Guo. Remotetouch: Enhancing immersive 3d video communication with hand touch. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 1–10, 2023.

[78] Y. Zhang, J. Yang, Z. Liu, R. Wang, G. Chen, X. Tong, and B. Guo. Virtualcube: An immersive 3d video communication system. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2146–2156, 2022.