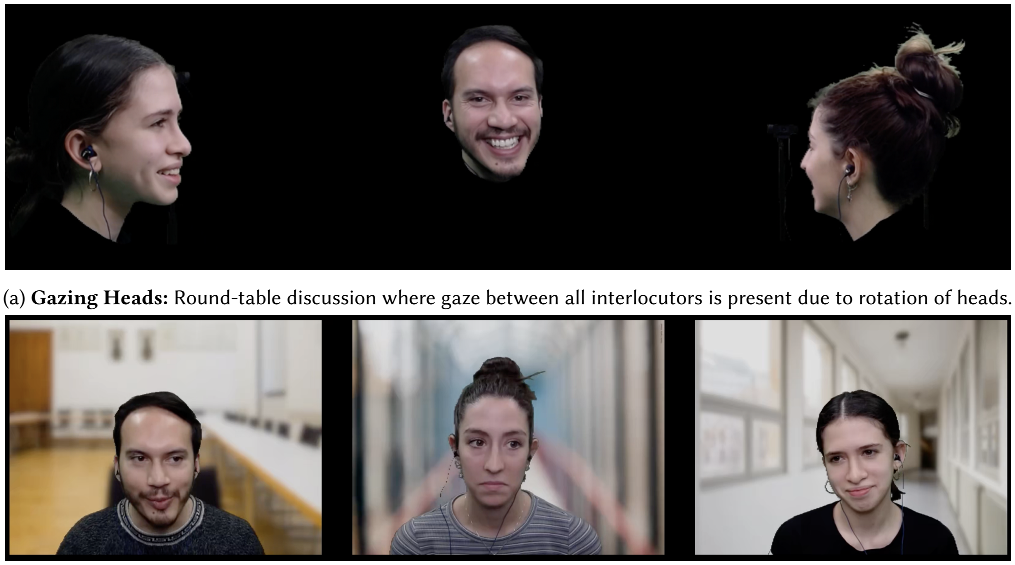
Gazing Heads: Investigating Gaze Perception in Video-Mediated Communication

MARTIN SCHUESSLER, University of Heidelberg, Germany
LUCA HORMANN, University of Heidelberg, Germany
RAIMUND DACHSELT, Technische Universität Dresden, Germany
ANDREW BLAKE, University of Cambridge, UK
CARSTEN ROTHER, University of Heidelberg, Germany



(a) **Gazing Heads:** Round-table discussion where gaze between all interlocutors is present due to rotation of heads.

(b) **Tiled View:** Traditional video conferencing where gaze cues are absent.

Fig. 1. **Snapshots of the Gazing Heads (a) and Tiled View (b) simulations taken during our user study**: They show the view of the fourth participant in a virtual discussion. We found that with Gazing Heads it is effortlessly apparent who is looking at whom. Gazing Heads proved beneficial for social presence and user engagement.

Videoconferencing has become a ubiquitous medium for collaborative work. It does suffer however from various drawbacks such as zoom fatigue. This paper addresses the quality of user experience by exploring an enhanced system concept with the capability of conveying gaze and attention. Gazing Heads is a round-table virtual meeting concept that uses only a single screen per participant. It

Authors' addresses: Martin Schuessler, research@mschuessler.de, University of Heidelberg, Germany; Luca Hormann, University of Heidelberg, Germany; Raimund Dachselt, raimund.dachselt@tu-dresden.de, Technische Universität Dresden, Germany; Andrew Blake, ab@ablake.ai, University of Cambridge, UK; Carsten Rother, carsten.rother@iwr.uni-heidelberg.de, University of Heidelberg, Germany.

Manuscript submitted to ACM

enables direct eye contact, and signals gaze via controlled head rotation. The technology to realise this novel concept is not quite mature though, so we built a camera-based simulation for four simultaneous videoconference users.

We conducted a user study comparing Gazing Heads with a conventional “Tiled View” video conferencing system, for 20 groups of 4 people, on each of two tasks. The study found that head rotation clearly conveys gaze and strongly enhances the perception of attention. Measurements of turn-taking behaviour did not differ decisively between the two systems (though there were significant differences between the two tasks). A novel insight in comparison to prior studies is that there was a significant increase in mutual eye contact with Gazing Heads, and that users clearly felt more engaged, encouraged to participate and more socially present.

Overall, participants expressed a clear preference for Gazing Heads. These results suggest that fully implementing the Gazing Heads concept, using modern computer vision technology as it matures, could significantly enhance the experience of videoconferencing.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

ACM Reference Format:
Martin Schuessler, Luca Hormann, Raimund Dachselt, Andrew Blake, and Carsten Rother. 2024. Gazing Heads: Investigating Gaze Perception in Video-Mediated Communication. *ACM Trans. Comput.-Hum. Interact.* 1, 1 (March 2024), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Since the beginning of the Covid-19 pandemic, video-conferencing has seen an unprecedented scale of adoption. Despite the benefits, “zoom fatigue” has become a major concern. One cause is the lack of non-verbal communication cues [4, 53], including gaze. Gaze cues serves crucial functions in regulating turn-taking, providing feedback, signalling attention, and conveying intimacy and emotions [2, 35, 39]. They increase group engagement, collective performance, and creativity [10, 54]. In today’s typical video conferencing systems, each user views the other users only frontally, confined to a small screen. We refer to this kind of layout as a “Tiled View”. Because of gaze misalignment, and because all users receive the same view, users cannot perceive who is gazing at whom. Conversation is measurably and palpably different from a face-to-face encounter [41, 58]. Alternative communication cues are needed [4] to help perceive others’ communicative acts and avoid misunderstandings [54].

We envision a system, *Gazing Heads*, in which each user sees the others displayed on their single-screen display, with gaze and attention conveyed through synthesised head rotation. Gazing Heads would use hardware which all video-conference users already own — a single screen, single camera, and microphone, with no need for additional displays or wearables. We anticipate that head rotation could soon be synthesisable in real-time, with sufficient realism, using software only [23, 71], once issues such as graphical realism [26], the uncanny valley effect, delays and synchronisation issues in high-quality video transmission are solved.

We built a simulation of Gazing Heads for four users using seven cameras placed around each user. The illusion of head rotation is created by transitioning between cameras. Our within-groups user study (N=80) compared Gazing Heads with Tiled View (Figure 1), with 20 groups of 4 participants, all tackling two different tasks. The first task was a group discussion about a controversial topic; the second was a game where participants were assigned specific roles with conflicting objectives. We used a wider variety of measures than prior gaze studies [27, 38, 52, 58, 64, 69] to gain more detailed insights. Objective measures of gazing behaviour and turn-taking were recorded, together with subjective ones, via questionnaires inspired by previous work. There were interviews at the end of each session to obtain qualitative insights.

Results show that simulated head rotation in Gazing Heads indeed conveys gaze and attention. Participants knew better when they or others were being addressed. Compared with Tiled View, users experienced a higher degree of social presence and engagement. We observed a significant increase in mutual gaze but there was no significant difference in turn-taking.

This summarise our work:

- we conceptualise Gazing Heads which provides higher levels of gaze realism and is anticipated to be feasibly implementable on standard laptop computers in the near future;
- we built a four-party experimental rig to test the idea that synthetic head-rotation enhances videoconferencing;
- we ran a study with more participants ($N = 80$) and using a wider range of measures than earlier work;
- making gaze perceivable (Gazing Heads) is found to improve the perception of attention in videoconferences;
- control over head rotation significantly enhances mutual eye contact, social presence and user engagement;
- interviews and questionnaires show that users prefer Gazing Heads over Tiled View, suggesting that by tackling missing gaze cues a major factor of zoom fatigue could be alleviated, once the technicalities of real-time synthesised head rotation are solved.

2 BACKGROUND AND RELATED WORK

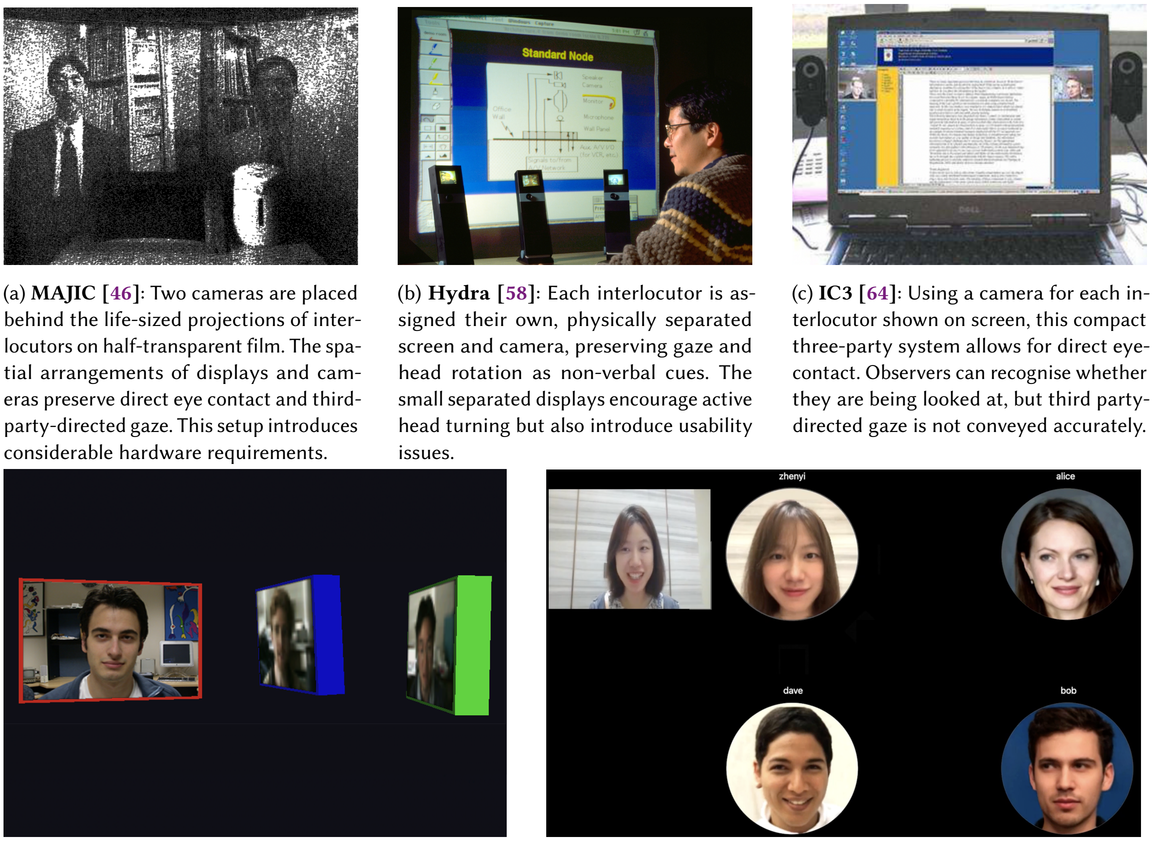
Gazing Heads would operate on a single screen without additional hardware while providing live, gaze-aware video, including third-party gaze. But what are the alternatives? There are four classes of telepresence systems that make gaze perceivable. Group-to-group systems (e.g. [45, 48, 49]) enable video-mediated communication for spatially separated groups of people. One-to-many systems (e.g. [34, 37, 62]) represent one remotely located interlocutor to a group. One-to-one systems (e.g. [40]) focus on the video-mediated conversation between two interlocutors. Virtual meeting room systems (e.g. [58, 70]) have each spatially separated interlocutor joining individually, and Gazing Heads is in this class. In standard virtual meeting room systems, the same view of each user, taken from that user's single camera, usually placed above the screen, is transmitted to all other participants. This introduces misalignments hindering gaze-awareness. *Direct eye contact misalignment* occurs when a user is being looked at but that gaze is misaligned. *Third-party gaze misalignment* occurs when a user is being looked at by another user, but the observing third party does not perceive this respective gaze. Many methods only focus on direct eye contact while ignoring third-party gaze [19, 27, 29, 33, 38, 74]. Gazing Heads addresses both issues.

2.1 Technology for virtual meeting rooms

Gaze-aware Virtual meetings can also be created using Virtual Reality (VR) (e.g. [63]) or Augmented Reality (AR) (e.g. [50]) but they demand substantial hardware: tracking devices, head-mounted displays, and powerful GPUs. Moreover, avatars lack realism [5, 6, 12, 55], obscure social cues [43, 47], and uncanny valley effects are evident [43, 47].

Given that AR/VR is expensive and unrealistic, an alternative is camera-based systems, and we found five precedents, as depicted in Figure 2. Their technical capabilities are compared with standard video-conferencing in Figure 3.

The first three systems use ante-hoc correction — avoiding gaze misalignment before images are recorded. A 1:1 physical-virtual space mapping, with dedicated cameras and displays for each interlocutor, is arranged to coincide with the virtual mapping and foster active head turning. The *MAJIC* system by Okada et al. [46] places cameras behind two life-sized projections of interlocutors (Figure 2a) — three users in all. The *Hydra* system by Sellen [58] uses three small screens with integrated cameras instead of projections (Figure 2b), supporting four users. Both systems require special



(a) **MAJIC** [46]: Two cameras are placed behind the life-sized projections of interlocutors on half-transparent film. The spatial arrangements of displays and cameras preserve direct eye contact and third-party-directed gaze. This setup introduces considerable hardware requirements.

(b) **Hydra** [58]: Each interlocutor is assigned their own, physically separated screen and camera, preserving gaze and head rotation as non-verbal cues. The small separated displays encourage active head turning but also introduce usability issues.

(c) **IC3** [64]: Using a camera for each interlocutor shown on screen, this compact three-party system allows for direct eye-contact. Observers can recognise whether they are being looked at, but third party-directed gaze is not conveyed accurately.

(d) **GAZE-2** [70]**:** Three cameras are placed behind interlocutor video tiles. Third-party-directed gaze is conveyed by rotating 2D video tiles in 3D, which is inferior to any actual 3D rotation of the head due to distorted visual gaze cues.

(e) **GazeChat** [27]**:** A single camera per user is used for eye-tracking. Gaze information is used to animate the user profile image. In this screenshot, all interlocutors are gazing at user zhenyi. Besides the modification of gaze, the images remain inanimate. They convey fewer cues than live video. Zhenyi is smiling (top left mirror view) but that is not conveyed by her avatar.

Fig. 2. **Screenshots of five gaze-aware meeting room systems**: different approaches to preserve gaze in video conferencing have been pursued, each introducing their own issues with usability or barriers to adoption. All images reproduced with the kind permission of their respective authors.

hardware, and are hard to extend for more users. The *IC3* system by Sun and Regenbrecht [64] addresses some of these issues by using a single display. The three-party video conference system places the two interlocutors to the far left and far right of a screen with a camera mounted next to them (Figure 2c). The setup is simple and compact but cannot be extended to more than three users, nor does it address the third-party gaze problem, given that it has only two views. The four-party system *GAZE-2* by Vertegaal et al. [70] uses a single semi-transparent display with three cameras behind it (Figure 2d). In principle, more cameras could be added to support further users. The 2D mages are rotated in a 3D virtual space to attempt to convey third party gaze direction, which largely fails because of the well-known “Mona Lisa effect”: the eyes of Mona Lisa gaze towards the observer, rather than rotating with the 2D display.

	Direct eye contact method	Third-party gaze-method	Users	Displays per User	Cameras per User	Live video
Zoom	None	None	N	1	1	Yes
MAJIC [46]	ante-hoc (camera behind)	ante-hoc (camera behind)	3	2	2	Yes
Hydra [58]	ante-hoc (camera close)	ante-hoc (camera close)	4	3	3	Yes
Gaze-2 [70]	ante-hoc (camera behind)	post-hoc (approximating gaze)	4	1	1	Yes
IC3 [64]	ante-hoc (camera close)	None	3	1	2	Yes
GazeChat [27]	post-hoc (synthesised gaze)	post-hoc (synthesised gaze)	N	1	1	No

Fig. 3. **Key properties of prior gaze-aware systems in relation to standard video conferencing**: Systems that use a single display (Zoom, Gaze-2, IC3 and GazeChat) have limited or no gaze-awareness support or lack live video. Systems that offer full gaze correction and live video (Hydra and MAJIC) rely on multiple displays and cameras. Note that Zoom was included as a representative of common video conferencing solutions, of which it has the highest market share [8].

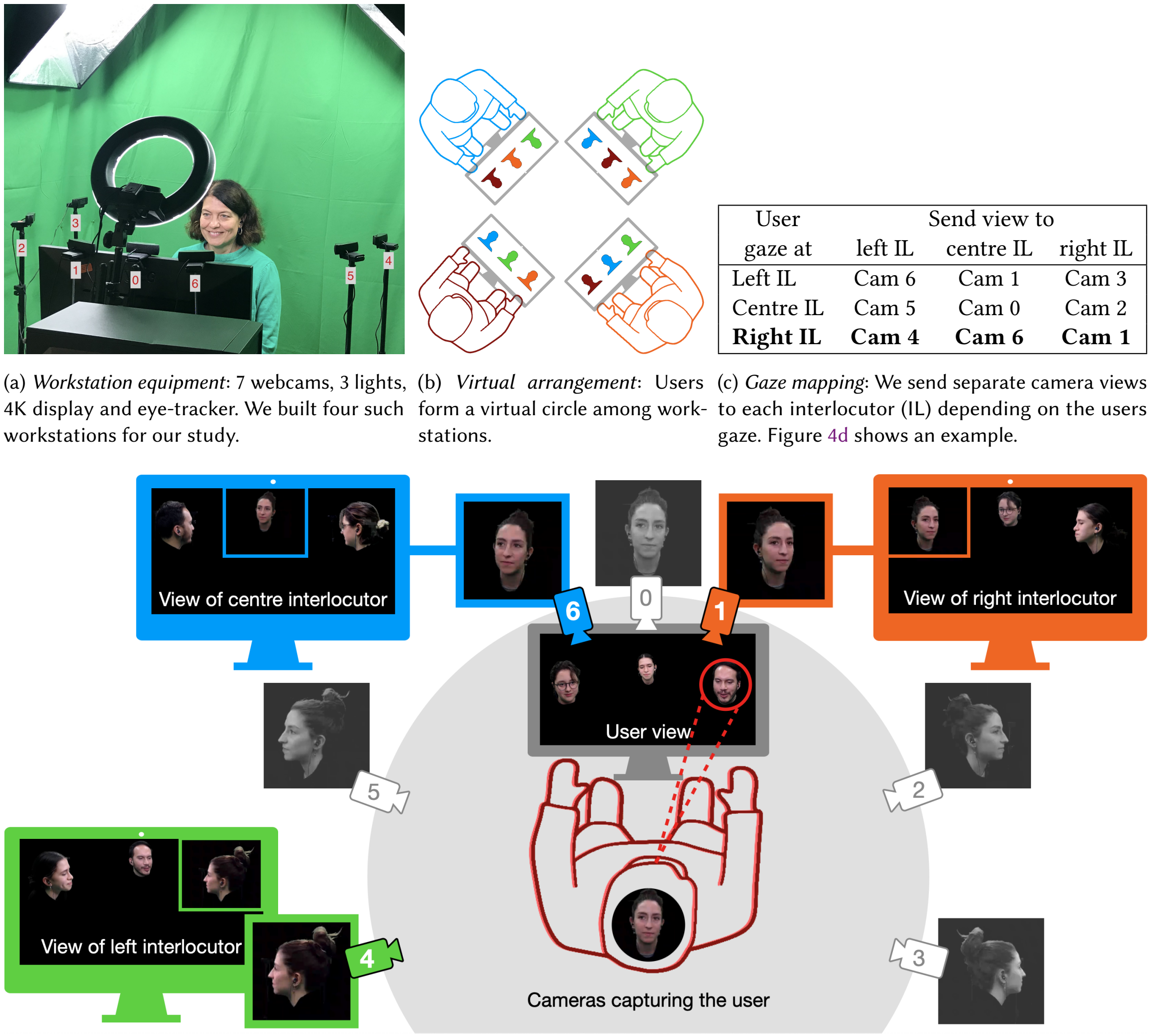
GazeChat by He et al. [27] is the only system which can be used on a conventional laptop without additional hardware. However it transmits animated 3D profile photos (created by neural rendering) (Figure 2e), without head rotations or live video. It animates users' gaze but other parts of the face are inanimate — verbal cues and facial expressions are lost.

2.2 Related prior user studies

Only two of the five related systems mentioned in the previous section have been evaluated in an extensive user study. GazeChat [27] was evaluated by four groups of four people (N=12), each having a group discussion. Each group tested four conditions: two variations of GazeChat, an audio-only meeting, and Tiled View with live video (within-group design). Questionnaires were employed to measure social presence, user engagement, and general user experience. GazeChat proved superior to the audio-only interface in some ways but, compared with TileView, GazeChat was worse at signalling direct eye contact and did not provide any other improvements – presumably because there was no live video. Turn-taking and gaze behaviour were not investigated in that study. It was not the objective of GazeChat to compete with live video conferencing but rather for use where live video is not an option due to privacy concerns or bandwidth limitations.

The user study conducted with the Hydra system [58] is the most relevant for our work. Twelve groups of four people (N=48) had a discussion in three conditions: face-to-face, Tiled View, and the Hydra (within-group design). A questionnaire was used to measure user experience and some aspects of social presence. Turn-taking behaviour was measured by processing participants' voice recordings but no significant differences were found between Hydra and Tiled View. Nonetheless Hydra was preferred by participants and was rated as superior for perceiving gaze and attention. The study also investigated the difference between the two video-mediated systems and face-to-face. It found that video-mediated conversations were significantly less dynamic. Hydra did show that gaze-awareness and head rotation could make attention perceivable in video conferencing and served as an inspiration for our study.

Our Gazing Heads study aims to determine whether future systems using synthetic head rotation, once the technology is mature, are actually likely to improve the video-conferencing experience. Given current technical limitations, we have built a simulation of Gazing Heads using additional cameras. The study provides more detailed insights than the prior work. It represents a substantial update in the light of modern developments in hardware and processing. It tested 4 users together, where some of the prior studies had only 3. It included a larger number of participants (N=80), employed a more comprehensive questionnaire, and also analysed eye-tracking data.


2
3
1
0
6
5
4
View of centre interlocutor
View of right interlocutor
View of left interlocutor
User view
Cameras capturing the user
0
1
2
3
4
5
6

User gaze at	Send view to left IL	centre IL	right IL
Left IL	Cam 6	Cam 1	Cam 3
Centre IL	Cam 5	Cam 0	Cam 2
Right IL	**Cam 4**	**Cam 6**	**Cam 1**

(a) *Workstation equipment*: 7 webcams, 3 lights, 4K display and eye-tracker. We built four such workstations for our study.

(b) *Virtual arrangement*: Users form a virtual circle among workstations.

(c) *Gaze mapping*: We send separate camera views to each interlocutor (IL) depending on the users gaze. Figure 4d shows an example.

(d) *Camera angles and mapping example*: This user gazes at her right interlocutor (IL). Based on the mapping shown in Figure 4c, we send the following camera views: Direct eye contact via camera 1 to right IL, slight head rotation via camera 6 to centre IL and strong head rotation via camera 4 to left IL.

Fig. 4. **Setup of Gazing Heads**: We use eye-tracking to convey gaze by sending different camera views to each interlocutor. Note that in all subfigures, cameras are consistently numbered (0 – 6), and colours denote interlocutors and their respective views.

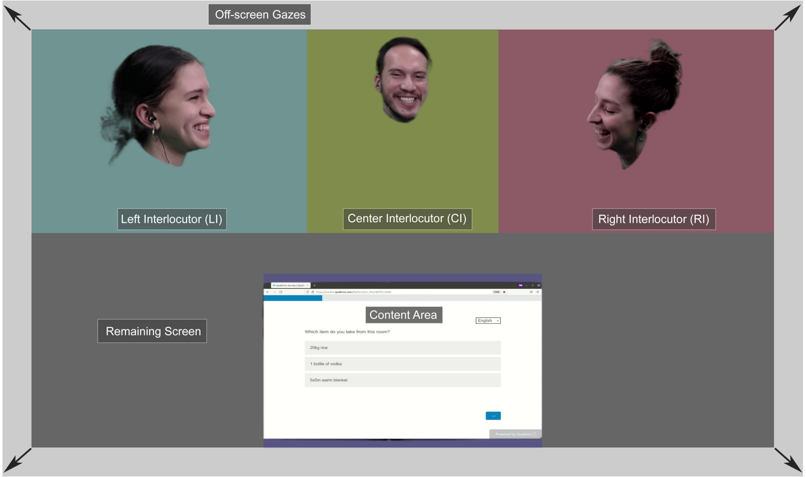
3 SIMULATING GAZING HEADS

Gazing Heads was developed through a sequence of pilot studies, and is illustrated in Figures 1a and 4. Since technology is not yet mature enough and the goal was to conduct a user study to evaluate the concept, rather than providing a full implementation, we did not yet build a system that uses only a single camera.

3.1 Sitting in a circle

Interlocutors were arranged on a single screen to enable ready perception of gaze. We created the illusion of a circular arrangement which encourages the exchange of gazes [32] by displaying the middle one of three interlocutors slightly smaller and higher than the others (Figure 1 top). On each user's workstation, the other users are displayed on the left, centre, and right, in a way that is consistent with four users around a table. The user in each of these three positions can turn to any of the other three. Hence, nine rotation angles for every user are needed: three positions on the screen, each with three unique head rotations. Note that we did not include a self-view as they are believed to have numerous negative effects [4], such as absorbing visual attention [20] and reducing the perception of others' emotional responses [59].

Available computer vision methods to synthesise rotation angles from a single camera perspective face substantial technical challenges: low realism [61], artefacts [18, 36, 77], and limited rotation angles [18, 71]. Some more advanced methods do not work with participants wearing glasses or having long hair [12]. We therefore used several cameras to obtain the required views (ante-hoc gaze correction) and switch between them. In principle nine cameras are needed — one for each of three screens, then one for each of three head rotations. By careful selection of viewing angles, we reduced nine cameras to seven, with two doing double duty — cameras 1 and 6 in the illustration of Figure 4d.



Which item do you take from this room?

20kg rice

1 bottle of vodka

5x5m warm blanket

Fig. 5. **The six focus areas of the Gazing Heads prototype in the game task**: We used these areas to decide which camera view to send to the other interlocutors. We also used them in our eye-tracking analysis of our user study. Participants changed their gaze from one area to another roughly every second. Note that in the discussion task, the content area was removed.

3.2 Gaze switching

Dedicated hardware (Tobii Eye Tracker 5) tracks gaze on a 4k 27-inch display (see also additional material). Each user's screen splits into focus areas to map gaze to camera views (Figure 5). When a user changes gaze, a camera transition is triggered appropriately. Users typically switch gaze about every second, often just glancing briefly which often goes unnoticed in physical settings [2]. When a gaze switch is detected, two criteria are continuously assessed: gaze duration on the same user for at least 750*ms* (dwell time) and a 2000*ms* lapse since the last transition (refractory period). The moment both are met, a new transition is initiated. This approach avoids too frequent transitions, yet allows quicker gaze switches following a period without transitions. Both values were determined through pilot studies. Views are faded into one another using alpha-blending (Appendix A.3).

Consistent diffuse lighting from all angles was obtained from two softboxes (85W CFL light bulb, 5500K) mounted overhead and an LED Ring for facial illumination (35W, 5500K). Each camera captures a different background so Chroma-keying with U-shaped green screens removes the backgrounds. That requires additional hardware, unrealistic in a commodity implementation but acceptable in a simulation. A black background strengthens the illusion of 3D head rotations since heads have a rough ellipsoidal shape. It also masks small errors in background segmentation. Interlocutors are centred and scaled to similar size in their respective views by software that tracks faces after background removal, and crops them, with temporal filtering to suppress jitter. This is done gracefully over time to allow for some margin for meaningful natural head movements (e.g. leaning into a conversation). A typical user shifts gaze by rotating the head and changing eye-gaze but alters upper body posture only slightly. Switching between camera views creates a "stiff-necked" illusion of substantial upper body rotation, and that looks unnatural. Therefore participants wore green turtlenecks so the chroma-keying background segmentation omitted neck and shoulders.

The system is carefully engineered to keep audio and video latency within acceptable limits (Appendix A.2). We used Rode Lavalier GO microphones with in-ear headphones and a feedback loop from microphone to headphones to ensure high quality spatial audio.

4 USER STUDY METHODOLOGY

We conducted a user study to test our concept and investigate whether head rotation can be used for conveying gaze and attention in video conferencing. We also wanted to understand how Gazing Heads influences users' experience and communication behaviour. Concretely, we expected Gazing heads to influence turn-taking behaviour due to the regulatory function of gaze as a turn-taking cue [15, 16, 35, 39]. We also expect it to create a more personal, intimate and immersive experience [39] leading to an increase in social presence [14, 41, 60]. Lastly, we were interested whether Gazing Heads influences users engagement [13] and gazing behaviour. Two tasks — discussion and game — were used with a wider variety of measures than previous gaze-awareness studies [27, 38, 52, 58, 64, 69], and with the largest number of participants ($N = 80$) used in studies of this kind to date. Participants used the two different systems (Gazing Heads and Tiled View) in a within-group design.

4.1 Experimental setup

The Gazing Heads simulation as described above served as the *treatment system*, with the Tiled View as a comparative *baseline system*. We did not compare against other gaze-aware solutions as no empirical evidence exists that they would outperform the Tiled View (e.g. [58]) but instead may perform worse (e.g. [27]). In Tiled View, three equally sized video tiles were placed at screen locations similar to those used in Gazing Heads. Head rotation was disabled and only central cameras were used, ignoring the 6 off-centre cameras. Background segmentation remained active but without the green turtlenecks, so the upper body was not segmented out — see Figure 1 on page 1.

We recruited 42 male, 35 female, and 3 non-binary or non-conforming gender subjects, and relatively young with 76 % younger than 25 and only 6 % 35 years or older. Participants were accustomed to video conferencing; 96 % of them used it at least once a month and 81 % at least every week. One group conducted the experiment in Russian, two groups in Spanish and the remainder in German (all native speakers). Thereby we ensured proficiency in the respective language selected. All participants received a 30€ Amazon voucher for their participation.

4.2 The group discussion and the survival game

Prior studies have mostly used group discussion [27, 38, 49, 52, 58, 64] – though a few used collaborative problem-solving [67–69] – and it is evident that task type can significantly affect behaviour. Group discussion as in Sellen [58] resulted in higher turn-taking frequencies ($3.9 - 4.3\,\mathrm{min}^{-1}$) than the problem-solving task used by Vertegaal et al. [69] ($1.0 - 1.3\,\mathrm{min}^{-1}$). In a problem-solving task there is reduced need for eye contact which may well reduce the likelihood of a measurable effect [64].

Group discussion has good external validity because it represents a real-world situation. One drawback is reduced internal validity as participants' prior knowledge, individual traits such as extroversion, and group dynamics, can influence the conversation. In extreme cases, pilot studies showed one or two participants holding the floor most of the time, masking the effects being measured. We nonetheless included group discussion, for comparability with other studies [27, 38, 49, 52, 58, 64]. We also included the problem-solving task for increased internal validity.

Controversial group discussion. Five controversial statements were tested, for example: "Industrial livestock farming should be progressively banned" (details in Appendix B.1). In each group, prior to the main experiment, participants rated agreement with each of five statements. The two statements whose ratings varied the most were selected as the two topics for discussion. Then for each topic, participants were instructed to find consensus as a group within five minutes. They were not stopped dead at five minutes, to avoid lowering engagement for later tasks. Instead, we interrupted the task when the current speaker(s) finished their turn or the group reached an agreement.

Game: surviving in the wild. Our second task was designed to:

- incentivise participants to take the floor;
- make turn-taking more dynamic and more evenly distributed;
- encourage participants to pay attention to non-verbal communication and understand others' intentions.

Players travelling to a remote island have to reach consensus on the choice of essential items shown on screen, to bring with them (Figure 5). They were given seven minutes to agree on as many items as possible. One player was randomly selected to be a clandestine saboteur so they would be incentivised to focus even more intensely on one another (details in Appendix B.2). Again we allowed players to reach agreement rather than stopping the game dead at seven minutes.

4.3 Three Questionnaires: Semantic Differential, UX, and Comparative

We reviewed similar studies for potential questions [11, 25, 44, 58, 60, 67, 76, 76] and categorised them by the concept or property dimension they measure. The most relevant were: presence, turn-taking, engagement, user satisfaction and usability. No existing questionnaire covered all dimension we were interested in so that we selected a few questions for each dimension to create two custom questionnaires. The *comparative questionnaire* asks about seven aspects of system preference relating to turn-taking, perceived attention and naturalness of interaction (see Tables 1 and 2, in Appendix B.4 for details). The *UX questionnaire* asks about direct eye contact, directed third-party gazes, and "off-gazes" directed at no one. In addition a standard *Semantic Differential* questionnaire measures social presence [60]. The UX and Semantic Differential questionnaires were filled out twice, once after using each system.

Presence. Lombard and Ditton [41] refer to "presence as social richness" which we call *social presence* which is measured by the semantic differential questionnaire. They term "presence as transportation [to a virtual room]" which we call *virtual presence* and which is measured in the UX questionnaire via a question about the feeling of "being in the same room". Two statements were included about the perceptibility of interlocutors' reactions and becoming acquainted with them. The comparative questionnaire asked which system participants would use for persuading others. These questions are known to be correlated with social presence [14, 60]. We also added a question about which system was considered to be more social.

Engagement. Two statements about participants' excitement and the interactiveness of the conversation were included in the UX questionnaire. We also added a question to the comparative questionnaire measuring participants' satisfaction with their contribution to problem-solving [67], and one addressing engagement/excitement.

Usability. During pilots, some participants complained that they felt excluded when interlocutors turned away from them. Others found the head rotation and camera transitions distracting. We added two questions addressing these issues to the custom UX questionnaire.

Overall preference and willingness to adopt. The comparative questionnaire asked which system allowed for a more natural interaction — this can be seen as a measure of presence [73], and also as a high-level quality metric for video-mediated communication. As indicators of overall user experience, participants were asked which system they would recommend to others and which system they wanted to use for a final interview.

4.4 Procedure

A full factorial within-group design was used, each group using both Gazing Heads and Tiled View systems for the group discussion, and for the game. The system to use first was distributed evenly across groups. Each session began with the group discussion, followed by the game. Then, switching systems, there was a game session followed by another discussion on a different topic. The discussion lasted about 6 minutes (M = 5:58, SD = 1:27) and the game took around 8 minutes (M = 8:22, SD = 2:16). A final interview of roughly 14 minutes (M = 13:37, SD = 6:49) was conducted using the system favoured by the majority. In case of a tie, the group was asked to reach a consensus on which system should be used. Altogether the experiment took 90 – 120 minutes, which included calibration, answering questionnaires and extensive Covid-19 protection measures (Appendix B.3).
(Additional details can be found in Appendix B.)

4.5 Data analysis

Speech. We recorded participants separately in each station. Quantitative measures for video-mediated communication are generally based on simultaneity of speech and on turn-taking [66]. We used the definitions of terms and measures from Sellen [58], which are sensitive enough to detect difference between a physical and virtual conversation. We converted any absolute measures to frequency or proportion measures as our session duration varied in length (Appendix B.5). Audio streams were analysed using a Voice Activity Detector [65], based on a deep learning transformer model. Pure laughter, detected by a residual neural network [22, 56], was distinguished from speech activity. Manual annotation separated pure laughter from any utterances made while laughing.

Gaze times. We defined six layered rectangular gaze areas as in Figure 5 corresponding to: the three participants, the display area for the game task, the rest of the screen, and off-screen. Once again, dependent variables for duration were converted into frequency of occurrence or proportional duration.

Shared eye contact. Shared eye contact is perceivable in Gazing Heads (but not in Tiled View) and we investigated whether this affects how long and how frequently participants gaze at one another. A mutual gaze event begins when participant A focuses on the area where participant B is displayed on screen A, and vice-versa. It ends as soon as either is focusing on a different area.

Significance tests. For ordinal self-reported data, we used a single-tailed proportions test for comparative questions and Wilcoxon signed-rank test for all others. We used a factorial repeated measures ANOVA for the metric speech and gaze measures, assessing the assumptions of normality and sphericity with Shapiro-Wilk and Mauchly's tests, respectively.

Interview insights. Thematic analysis [9] was applied to recordings of group interviews, and captured using a coding scheme. A scheme was devised with two major domains of technology and experience and a total of 11 topics described by 52 codes. Codes were applied, checked and resolved by two of the authors.

5 RESULTS

Here are the main results from the three sources of data: recorded speech, gaze data, and user experience questionnaires.

5.1 Speech activity

Only one significant difference between systems was found in speech activity, that group turns occur 21% more frequently in Tiled View than in Gazing Heads, as Figure 6 shows. However they are rare and so are not a strong indicator of difference in turn-taking behaviour. No significant differences are observed between systems for the other 10 turn-taking measures, refuting our assumption that Gazing Heads would make speech activity more akin to physical rather than virtual interaction. Between tasks, there are significant differences in 10 out of the 11 measures, indicating that the game was a more dynamic task (see also Appendix C.1). Those differences confirm that the measures are indeed sensitive to changes in communication behaviour.
(Note 1 of the 20 sessions didn't produce speech data usable for analysis due to recording issues.)

5.2 Gaze and eye contact

Eye contact occurs just a bit more frequently (Figure 7) with Gazing Heads, both during the discussion (+8.9 % more frequent) , and the game (+3.6 %) tasks ($p < 0.01$). However such mutual gaze as *did* occur in Tiled View was also not so useful — participants were "somewhat not" able to distinguish whether gaze was actually directed at them (Figure 11). Change of focus would be another potentially interesting variable, happening on average every second. However there is no significant difference across tasks or systems. As an aside, there are marked differences between tasks. For example mutual gaze is significantly reduced in the game task, for both systems ($p < 0.001$), because more time is spent in the game gazing at the content area which contains the information needed to play (see also Appendix C.2).

5.3 Overall system preference

Users prefer Gazing Heads for most aspects of interaction, as Figure 8 shows. It is more engaging (87%) and makes it easier to sense the attention of others (91%). There is a clear tendency to choose Gazing Heads for the final discussion session (72%). Participants tended to find Gazing Heads more natural (62%) and social (62%), suitable for a persuasive discussion (51%, $p = 0.015$), and to provide a better turn-taking experience (49%, $p = 0.052$). Gazing Heads is favoured for facilitating interactive conversation, getting to know people ($p = 0.01$) and for making the discussion more exciting.

		Discussion Task		Game Task		p-Value	
	Hydra [58]*	Gazing Heads	Tiled View	Gazing Heads	Tiled View	Task	System
Turn Frequency/Minute	4.29	2.69 (1.2)	2.75 (0.9)	6.31 (1.1)	6.32 (1.2)	< **0.001**	0.869
Turn Duration	16.62 s	24.48 s (9.8)	22.97 s (10.5)	8.04 s (1.8)	7.74 s (2.0)	< **0.001**	0.535
Group Turn Freq./Minute	0.24	0.22 (0.2)	0.23 (0.2)	1.43 (0.6)	1.76 (0.8)	< **0.001**	**0.037**
Turn Distribution (H)	1.83	1.90 (0.1)	1.90 (0.1)	1.95 (0.0)	1.97 (0.0)	**0.001**	0.620
Time one Person spoke	74.70 %	90.5 % (4.4)	90.3 % (3.3)	73.0 % (6.0)	70.9 % (5.8)	< **0.001**	0.300
Simultaneous Speech	5.40 %	2.5 % (3.0)	3.3 % (2.3)	11.1 % (5.4)	13.2 % (6.5)	< **0.001**	0.062
non-Int. Simult. Speech	10.22 %	2.4 % (2.5)	3.1 % (1.9)	10.9 % (4.7)	11.9 % (5.1)	< **0.001**	0.187
Interruptive Simult. Speech	3.50 %	1.0 % (1.3)	1.1 % (1.0)	4.0 % (1.8)	4.5 % (2.1)	< **0.001**	0.276
Sim. Speech Taking Control	41.60 %	30.9 % (21.5)	22.8 % (14.0)	31.3 % (8.5)	30.5 % (6.9)	0.113	0.212
Speaker Switches Overlaps	43.50 %	23.7 % (14.9)	25.6 % (17.9)	45.0 % (12.2)	44.3 % (9.3)	< **0.001**	0.810
Switching Time	0.25 s	1.00 s (0.8)	0.69 s (0.7)	0.18 s (0.4)	0.14 s (0.4)	< **0.001**	0.171

Fig. 6. **Speech analysis.** The middle four columns compare our two systems within the two tasks. Differences between systems are significant only for the group turns. Differences between tasks are significant for most measures, indicating that the measures are sensitive. Figure 12 in Appendix C has additional descriptive statistics along with F values.

[*] The first column shows the results of Sellen's study with the Hydra system for comparison [58]. Interestingly, even after 30 years, measures are broadly consistent with our measurements, lying between the value for the discussion and the game.

	Discussion Task		Game Task		p-Value	
	Gazing Heads	Tiled View	Gazing Heads	Tiled View	Task	System
Focus Changes per Minute	58.25 (24.0)	58.34 (22.9)	59.28 (19.9)	62.49 (22.5)	0.109	0.053
Eye Contact per Minute	30.71 (13.2)	28.21 (10.9)	13.60 (6.9)	13.13 (7.0)	< **0.001**	**0.008**
Eye Contact (% of Session)	27.8 % (10.9)	26.0 % (9.6)	10.1 % (5.3)	9.3 % (4.9)	< **0.001**	**0.035**
Eye Contact Duration	0.58 s (0.2)	0.58 s (0.2)	0.46 s (0.1)	0.45 s (0.1)	< **0.001**	0.548

Fig. 7. **Eye-gaze analysis.** Four columns comparing the average measures obtained in the two tasks with our two systems. There was a small but significant system effect: participants had more eye contact when using Gazing Heads.

Overall, however, participants were undecided about recommending Gazing Heads, in its current state, over Tiled View. *(All reported differences are significant at $p < 0.001$, unless otherwise stated.)*

5.4 Social presence, user experience, and awareness

The four attributes used to measure social presence are reliable with Cronbach $\alpha = 0.85$ — they are plotted in Figure 9, together with an overall score. The first conclusion is that social presence is rated more highly (62%) for Gazing Heads ($p = 0.02$). Figure 10 shows us exactly which aspects of social presence are different: Gazing Heads made it easier to get to know people ($p = 0.01$); and also made them a little more aware of others' presence ($p < 0.01$). With both systems, participants found it exciting to follow the discussion and perceived the conversation as highly interactive; still, they significantly preferred Gazing Heads ($p = 0.04$, $p < 0.001$). Gazing Heads makes participants feel significantly more in the same room as one another ($p < 0.001$), and although it is perceived as "somewhat" safe from distraction, it is not felt to be as safe as Tiled View ($p < 0.001$). The two systems are similar in several respects: the ability they give an

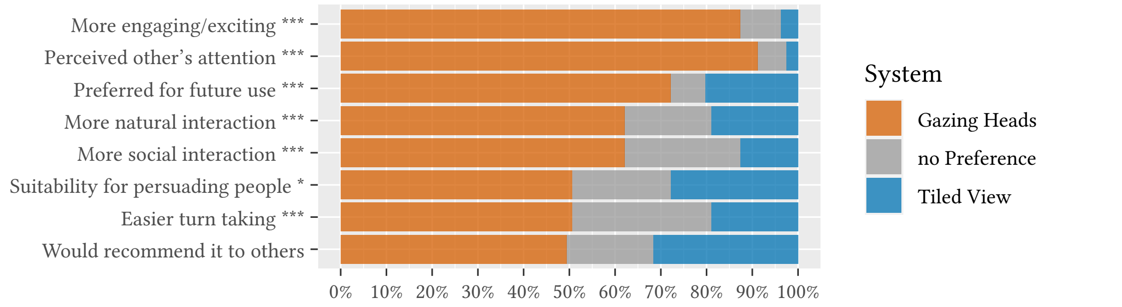


Fig. 8. **Participants' system preference ratings**: Participants generally preferred Gazing Heads (significance levels: p < 0.05 *, p<0.01 **, p<0.001 ***).

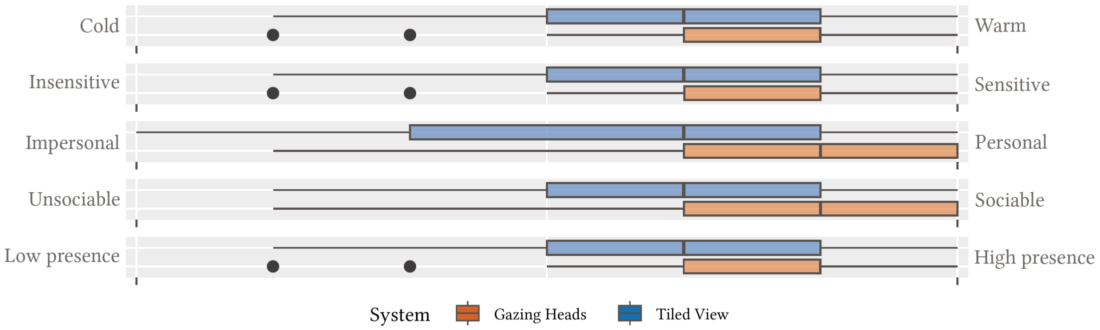


Fig. 9. **Participants' social presence ratings**: The social presence score was based on participants' median ratings of four attributes (sociable, personal, sensitive, warm) for the two systems. Scores are significantly higher for Gazing Heads. (Dots outside the whiskers indicate outlier scores.)

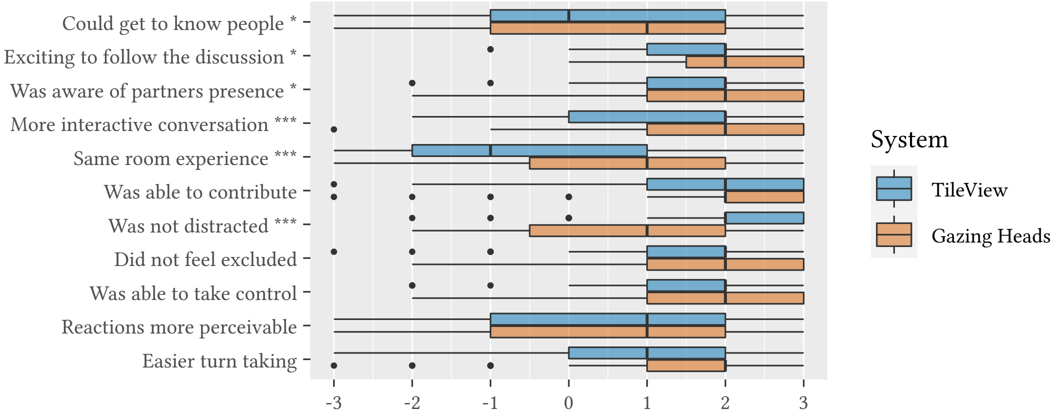


Fig. 10. **Participants' user-experience ratings**: Ratings for several aspects of the user experience ratings. The scale ranges from strongly disagree (−3) to strongly agree (3). (Significance levels: p < 0.05 *, p<0.01 **, p<0.001 ***.)

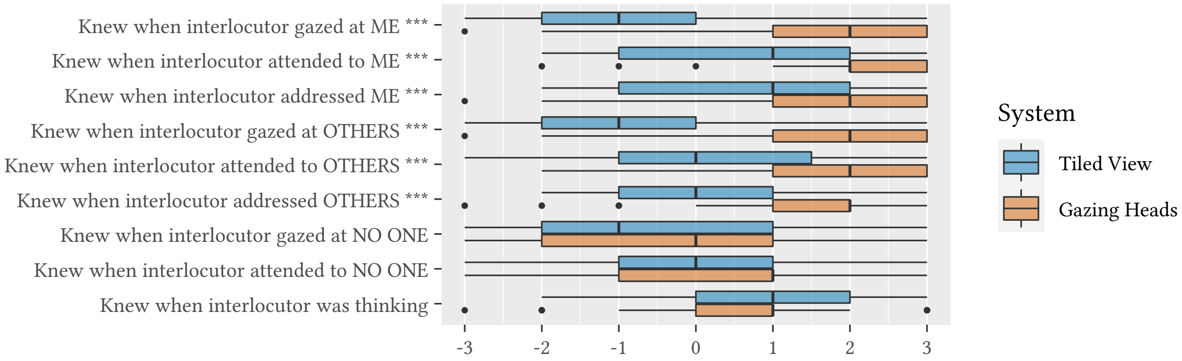


Fig. 11. **Participants' awareness ratings**: Ratings for perceiving gaze, perceiving attention, and awareness of being addressed. (Significance levels: p < 0.05 *, p<0.01 **, p<0.001 ***).

individual to contribute to their team's solution ($p = 0.246$); degree of exclusion from the conversation ($p = 0.790$); and ability to take control of it when they want to ($p = 0.176$). The reactions of others are only "somewhat" perceivable ($p = 0.263$) in both systems. Turn-taking is "easy" with Gazing Heads, and "somewhat easy" with Tiled View, but the difference is not strongly significant ($p = 0.084$). Gazing Heads also performs significantly better on awareness of gaze and attention, and awareness of who is being addressed (Figure 11), with 91% of participants reporting that they perceived attention more easily (Figure 8, $p < 0.001$). They also found it easier to reason about who is being addressed or attended to (Figure 11, $p < 0.001$ for all six ratings). It was unclear whether either system could convey disengagement or "gazing at no one".

(Note: The responses of 1 group were excluded from the analysis due to a procedural error. One participant lost his replies by accidentally logging out of the system.)

6 DISCUSSION

Our study has shown that the Gazing Heads concept improves video conferencing, providing gaze awareness via synthesised head rotation. This section explores the implications of results from the previous section, making particular use of insights gained from the exit interviews.

6.1 Conveying attention

Participants (all but 2) understood intuitively, within the first two minutes of using Gazing Heads, that head rotation conveys visual attention. The questionnaires showed already that they perceived attention more readily in Gazing Heads, and during interviews they often mentioned that Gazing Heads conveyed attention better ($N = 36$), made it more perceivable ($N = 69$), and made it easier to gain attention ($N = 16$). *(Here N denotes the number of participants for which a code was applied at least once.)* One **implication** is that correcting a user's frontal view without adding head rotation as in [19, 27, 29, 33, 38, 74] is insufficient to achieve gaze awareness. The user study for Gaze Chat [27] which only used such a correction showed that it did not improve users ability to perceive attention. Hydra [58] did use head rotation, and did find an improvement. Users' accuracy in estimating the gaze of others may be low [39] so head rotation provides an additional cue for attention [26].

6.2 Subjectively higher engagement

In the last section, we saw that Gazing Heads felt more engaging and this was a common topic in interviews ($N = 39$). Interviewees occasionally even felt social pressure to participate because the signalling of attention was so clear ($N = 32$):

> "If your goal is that everyone is taking part in a discussion, you HAVE TO use this system. Because you are simply forced to stay engaged." "[Especially] when everyone starts looking at one person."

Our findings contrast with the studies of Hydra [58] (and GazeChat [27]) where no improvement of engagement over Tiled View was found. With Hydra, users needed to actively turn their heads and were reluctant to do that. Gazing Heads translates almost every gaze into a head rotation and makes them salient by placing them in users central vision rather than in peripheral vision on small screens. We believe this increased users engagement.

6.3 Conveying disengagement and gazing-away

During interviews, users took issue with the camera selection rule which showed them turned towards another participant, even when they were not gazing at anyone ($N = 11$):

> "The only thing that is missing, perhaps, is that you cannot look at nobody. So even if I am just staring in front of me, then I am still [being displayed as] looking at someone."

They also want to be able to disengage, which Gazing Heads does not facilitate:

> "if people could like watch me very closely ... when I am looking at my phone or something like that [...] in larger groups when you want to disengage from the discussion ... then it's almost creepy."

To address these concerns, a future system could add a neutral object like a table, as suggested in interviews ($N = 10$), similar to the content area which for us was present only during the game. However, this would risk reducing mutual eye contact and attenuating the effects of gaze cues [3] (Appendix C.2). One might also show disengaged users as greyed, inanimate or separated, allowing them to be listeners only, or to work on something in parallel. Systems could offer different layouts depending on whether high engagement is a major objective, and future work might look into that.

6.4 Increased social presence in a virtual space

We saw in the previous section that Gazing Heads increases the feeling of social presence. Also in the interviews, an increase of social presence was frequently ($N = 58$) mentioned as an advantage of Gazing Heads. It was usually described as a feeling of "being closer" to the other participants or having a more personal experience. Similarly, interviewees described improved virtual presence ($N = 50$) but were divided as to whether the improvement was substantial enough to feel "in the same room" ($N = 23$), or not ($N = 14$), for example:

> "With heads floating in a black room it is totally unclear in which real distance we are actually located to one another [...] we are all floating in this empty black thing."

or

> "Well I still think the situation ... uhm ... is still video-telephony .. and that's just not so immersive that it completely detaches you [from your current surroundings]."

It was also evident that the concepts of social and virtual presence are considered as related (see also Lombard and Ditton [41]), for example:

> “Here you feel closer, because it’s [rather] simulating a room and not just [a] screen.”

As for other design aspects that may contribute to physical and social presence, participants frequently praised the consistent placement of interlocutors in a virtual circle ($N = 15$). Participants also commented that the separate backgrounds in Tiled View were a strong visual indicator that they were not in a shared space ($N = 16$). For Gazing Heads our design choice of using a plain black background was a frequent topic. Several participants disliked it ($N = 21$) because it was perceived as cold, providing too little context. Others preferred it in plain black ($N = 8$) for increasing contrast and reducing distraction. Our study leaves open how virtual backgrounds impact physical and social presence. Future research could examine whether a photorealistic scene with a unified background and elements like a table or a bonfire enhance presence and convey virtual togetherness.

6.5 Subjective effect of eye contact

Although participants spent only *a little* more time gazing at one another in Gazing Heads, eye contact changed their perception substantially. Since gaze is hardly noticeable in Tiled View, according to questionnaires, any "eye contact" was probably guided by auditory perception not visual cues. In interviews ($N = 7$) eye contact in Gazing Heads revealed a clear, positive effect on participants’ experience:

> “For a discussion, the second system [Gazing Heads] is much more comfortable. You just have more of a feeling of being part of a group. For me, in seminars, [...] others were just sitting inside of their tiles simply looking [somewhere], not really taking part. Here you have the feeling of being integrated, even if you are not saying anything, especially when you are being looked at.”

Interviews suggested, as also indicated by prior work [54], that eye contact in Gazing Heads facilitated a stronger feeling of engagement and social presence:

> “I think it’s more personal because, when you are looking at me, I have the feeling you are actually looking in[to] my eyes, and then I want to [...] explain my point of view to you [...] and I know [...] if you are looking at my face you will see also my eyes.”

Our observations from recordings suggest that synthesised head rotations make eye contact and attention selective and salient. They *amplify* associated positive effects such as increased social presence and higher engagement.

6.6 Inconclusive results regarding turn-taking

The previous chapter’s speech and turn-taking analysis saw no significant difference between the two systems for 10 of the 11 measures suggested by Sellen [58], nor did the questionnaire. This may be explained by video recordings of sessions, in which participants seemed to rely more heavily on auditory cues for turn-taking. Participants gave partly contradictory answers about turn-taking: 51% found Gazing Heads easier with 19% for Tiled View, and 30% had no preference, the largest degree of ambiguity we observed in any direct comparison question. Participants made several comments in interviews about turn-taking being easier with Gazing Heads ($N = 28$):

> “Who is talking next or who is generally talking right now ... emerged organically.” “Yeah right [...] when you started talking and you notice the others are looking at you or are turning towards you, then you knew ok I can talk right now.”

and

> “You could directly address people. When I asked [other participant] his name, I just looked at him and that worked.”

6.7 Camera transitions can be distracting

In the previous chapter we saw Gazing Heads rated as less protected from distraction. Interviews indicated that this was due to the transitions between cameras (N = 27), for example:

> “You see this short fade between perspectives [...] and that instantly distracted me for half a second and I thought: Oh cool where is he looking right now? And then it was like ... ok what just happened [in the discussion]?”

Frequently, participants described the animated transitions not as head rotation but as fading, vanishing, switching or flickering (N = 41) , and others perceived transitions as too slow or lagging (N = 8). Some complained that not every gaze switch resulted in a transition, but that was a deliberate design choice made to reduce the number of transitions, as they are visually very salient. With one group that suggested the dwell and animation time was too slow, we tried a Gazing Heads configuration where head rotations occurred more promptly and frequently. After four minutes of testing the group agreed it was “more natural” and preferable. This raises an open question about the optimal frequency for transmitting gaze among users. While transmitting every gaze would be overwhelming, as participants shift focus about every second, our chosen dwell and recovery times might have been too conservative.

6.8 Realism and nonverbal cues

Another prevalent interview topic was the design choice to show only heads, without clothing or shoulders (N = 52). Participants noticed that this strengthened the 3D illusion (N = 13). It also made the experience somewhat artificial (similar to a game) and gave the impression of a “work in progress” (N = 25). The most common concern was the absence of non-verbal communication from posture, shoulder shrugging and hand gestures (N = 46).

> “At some point during the game I shrugged my shoulders and I thought – ok – do they even see that? Then I thought ... oh no ... now I have to say something like “I don’t care” or something like that.”

At the same time it was noticed that reducing non-verbal communication to just the head increased the saliency of facial expressions (N = 15). There was a variety of opinions about showing the upper body and clothing. Arguments against showing gestures or posture were rare, but not all participants wanted the actual appearance of their upper body to be visible since, amongst other things, it meant having to dress appropriately for a meeting:

> “Think about fat-shaming for example, or women who are often exposed to unpleasant gazes.”

We tried 13 interview groups using the system without green turtlenecks, adding shoulders back into view, at the expense of less natural camera switches and exaggerated upper body movement. They were unsure whether this experience was more (N = 27) or less natural (N = 30) – some described this configuration as a “hybrid”, that means, easier to adapt to, given stronger similarity to Tiled View, yet providing gaze awareness benefits (N = 12). An open question is the mapping of upper body movement which needs to be consistent with head rotation ("inverse kinematics") while still appearing natural.

6.9 Implementation on commodity hardware

Overall, our interviews confirmed that technical maturity and realism are the main obstacles to a full implementation of Gazing Heads. Even under sub-optimal lighting conditions, with users wearing glasses or having long hair, systems need to realistically convey users' gaze, unique facial expressions and facial features. Current solutions may degrade the experience to a point where standard single-view video conferencing is preferred over unrealistic gaze-aware solutions. There are several implications for future research. Computer vision methods which modify users gaze (by using GANs or creating avatars) need to tackle two issues: realism [5, 6] and latency [7, 17, 17, 75]. For example, the head rotations shown by He et al. [27] appear fairly realistic but incidentally affect facial expression[1]. Several recent works [23, 51, 57] have claimed higher levels of realism since our study was performed in 2022, and it remains to be seen how effective they are in a system like Gazing Heads. Latency needs to be held down to an acceptable level (cf. Gazing Heads: 133.33 ± 33.33 ms — Appendix A.2.) Gazing Heads also needs webcam eye-tracking with 4.86° of accuracy (Appendix A.1.) This is achievable with modern methods under ideal conditions [28] but accuracy under realistic system conditions remains to be confirmed.

6.10 Finding more sensitive measures

We saw earlier that the nature of the task had a dominant influence on quantitative behavioural measures, consistent with prior work [3, 39, 64]. One implication is that quantitative results on gaze awareness may not generalise well across different studies with varying task scenarios and interface layouts. It also raises the question of exactly what alternative measures could be used. The measures of turn-taking [58] seem sufficient only to detect gross changes between tasks or comparing a physical to a virtual setting [58]. One possible interpretation is that artificially introduced gaze leaves video-mediated communication behaviour unchanged. Alternatively they may be the wrong measures [30]: "how many turns people take, how long those turns are, how many pauses people take [..] doesn't reflect people's real experiences of what those conversations are like." Compared to other studies, our wider range of measures did capture some differences in users' experiences. It remains for future work to find better quantitative measures to avoid relying on self-reported data. Since it is known that "People generally get along better and communicate more effectively when they look at each other" [39], one potential approach in the game setting may be to adapt measures of user cooperation and responsibility from social psychology and behavioural game theory.

6.11 Limitations

We are aware of four areas in which our experiments have limitations.

Realistic setting. The tasks (game and discussion) may not entirely represent typical video conferencing sessions in domestic or business settings. We chose unusually dynamic tasks. However this was done to help participants engage and get socially comfortable as quickly as possible, given that they generally did not know each other beforehand.

Questionnaires neglected differences between tasks. Questionnaires asked about differences between the two systems, but did not explore differences between the game and the discussion. This was to avoid overburdening participants with questions and consuming more time, but admittedly may have caused some relevant and interesting effects to be missed. However, no differences in experience between tasks were mentioned during interviews.

[1] https://youtu.be/dGY8NbG11Ng

Novelty effect. The novelty of the systems, especially Gazing Heads, may have influenced participants' behaviour and ratings. During interviews, many participants ($N = 50$) emphasised that Gazing Heads was novel, though they clearly understood that Tiled View represented familiar video-conferencing applications.

Participant demographics. Most participants were young students (Section 4.1), accustomed to video conferencing, so results may not generalise well to older populations less familiar with technology. We conducted experiments in three languages for variety but the speech analysis results may not entirely generalise across languages. Lastly, gaze is believed to be influenced by socio-cultural norms [24, 72], and our population was predominantly Northern European and Western Caucasian, so that is another potential limitation to generality.

7 CONCLUSION

The Tiled View layout for video-conferencing is well-established, and our user study and prior work [27, 58] have all shown how challenging it is to improve upon it. There are two obstacles to realising that improvement: convincing realism [6] and low latency [7, 17, 17] — without which the user experience is severely compromised. In this study we have addressed the realism issue.

We have found that Gazing Heads represents a clear advance over present day video-conferencing in its capacity for conveying gaze and attention. In contrast to earlier studies by Sellen [58] and by He et al. [27], which also took Tiled View as baseline, Gazing Heads has been found to increase social presence, mutual eye contact, and user engagement. It unequivocally enhances the experience of users. We attribute these results to the amplifying effect of head rotations for conveying gaze. In its current design, Gazing Heads enhances highly interactive small group meetings. For other communication scenarios, like collaborative content editing or presentations with a large audience, alternative designs may be beneficial — an idea that is somewhat supported by our interviews. For meetings however, conveying attention to content, and to other participants, and reducing distractions for the presenter, seem to be particularly important.

Human communication in virtual space is a topic of considerable and growing significance because remote working has become a permanent and prominent feature of working life, but Zoom fatigue is a challenge. Any technical progress that may mitigate it could substantially impact the effectiveness, health and well-being of users. We believe that this study, and the Gazing Heads concept in particular, could represent an important step towards that goal.

ACKNOWLEDGEMENTS

We thank Radek Mackowiak and Titus Leistner for their technical support in setting up our experimental setup, Tianling Yang for her help with the initial literature review and Ihtisham Ahmad for implementing parts of the eye-tracking analysis and generating Figure 5. The authors are grateful for the feedback of the anonymous reviewers that was very helpful to improve this article. This work was co-funded by the Federal Ministry of Education and Research of Germany in the programs "Weizenbaum Institute for the Networked Society - The German Internet Institute" (project ID: 16DII113), "Souverän. Digital. Vernetzt.", joint project 6G-life, (project ID 16KISK001K) and "Berlin Center for Machine Learning". We also acknowledge financial support from from the ERC Consolidator Grant RSM (647769) and by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy – EXC 2050/1 – 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) and DFG grant 389792660 as part of TRR 248 – CPEC (see https://perspicuous-computing.science).

REFERENCES

[1] John P. Abraham, Brian D. Plourde, and Lijing Cheng. 2020. Using heat to kill SARS-CoV-2. *Reviews in Medical Virology* 30, 5 (2020), e2115. https://doi.org/10.1002/rmv.2115 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rmv.2115

[2] Michael Argyle and Mark Cook. 1976. *Gaze and mutual gaze.* Cambridge University Press, Cambridge, Eng. ; New York.

[3] Michael Argyle and Jean A. Graham. 1976. The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology & Nonverbal Behavior* 1, 1 (1976), 6–16. https://doi.org/10.1007/BF01115461 Place: Germany Publisher: Springer.

[4] Jeremy N. Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior* 2, 1 (Feb. 2021). https://doi.org/10.1037/tmb0000030

[5] Gary Bente, Sabine Rüggenberg, Nicole C. Krämer, and Felix Eschenburg. 2008. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Human Communication Research* 34, 2 (04 2008), 287–318. https://doi.org/10.1111/j.1468-2958.2008.00322.x arXiv:https://academic.oup.com/hcr/article-pdf/34/2/287/22325251/jhumcom0287.pdf

[6] Nienke Martine Bierhuizen, Wendy Powell, Tina Mioch, Omar Niamut, and Hans Stokking. 2022. Influence of Photorealism and Non-Photorealism on Connection in Social VR. In *Annual Review of Cybertherapy and Telemedicine*. Interactive Media Institute. https://www.interactivemediainstitute.com/cypsy25/ 25th Annual CyberPsychology, CyberTherapy and Social Networking Conference, CYPSY 2021, CYPSY25 ; Conference date: 13-09-2021 Through 15-09-2021.

[7] G. Blakowski and R. Steinmetz. 1996. A media synchronization survey: reference model, specification, and case studies. *IEEE Journal on Selected Areas in Communications* 14 (1996), 5–35. Issue 1. https://doi.org/10.1109/49.481691

[8] Joel Bruckstein and Bob Veres. 2022. Videoconferencing software market share 2022. https://www.statista.com/statistics/1331323/videoconferencing-market-share/

[9] Emeline Brulé. 2020. Thematic analysis in HCI. https://sociodesign.hypotheses.org/555

[10] Martin Böcker and Lothar Mühlbach. 1993. Communicative Presence in Videocommunications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 37, 3 (1993), 249–253. https://doi.org/10.1177/154193129303700308 arXiv:https://doi.org/10.1177/154193129303700308

[11] Martin Böcker and Lothar Mühlbach. 1993. Communicative Presence in Videocommunications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 37, 3 (Oct. 1993), 249–253. https://doi.org/10.1177/154193129303700308 tex.ids= bocker1993a.

[12] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.* 41, 4, Article 163 (jul 2022), 19 pages. https://doi.org/10.1145/3528223.3530143

[13] Valerie Caproni, Douglas Levine, Edgar O'neal, Peter McDonald, and Gray Garwood. 1977. Seating position, instructor's eye contact availability, and student participation in a small seminar. *The Journal of Social Psychology* (1977).

[14] Brain G. Champness. 1973. The assessment of user reactions to confravision: I. Design of the questionnaire. (1973).

[15] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology* 23, 2 (1972), 283. https://osf.io/45krv/download

[16] Starkey Duncan Jr and George Niederehe. 1974. On signalling that it's your turn to speak. *Journal of experimental social psychology* 10, 3 (1974), 234–247.

[17] ITU-T Recommendation G.114. 2003. One-way transmission time. (2003), 9–9.

[18] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8649–8658.

[19] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 311–326.

[20] Joey George, Akmal Mirsadikov, Misty Nabors, and Kent Marett. 2022. What do users actually look at during 'zoom'meetings? Discovery research on attention, gender and distraction effects. (2022). https://hdl.handle.net/10125/79919

[21] Agostino Gibaldi, Mauricio Vanegas, Peter J. Bex, and Guido Maiello. 2017. Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods* 49 (6 2017), 923–946. Issue 3. https://doi.org/10.3758/s13428-016-0762-9

[22] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. 2021. Robust Laughter Detection in Noisy Environments. In *Proc. Interspeech 2021*. 2481–2485. https://doi.org/10.21437/Interspeech.2021-353

[23] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

[24] Jennifer X. Haensel, Tim J. Smith, and Atsushi Senju. 2022. Cultural differences in mutual gaze during face-to-face interactions: A dual head-mounted eye-tracking study. *Visual Cognition* 30, 1-2 (Feb. 2022), 100–115. https://doi.org/10.1080/13506285.2021.1928354 Publisher: Routledge.

[25] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW '06)*. Association for Computing Machinery, New York, NY, USA, 413–422. https://doi.org/10.1145/1180875.1180937

[26] Muchen He, Beibei Xiong, and Kaseya Xia. 2021. Are you looking at me? Eye gazing in web video conferences. *CPEN 541 HIT'21* 27 (4 2021), 28.

[27] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-Aware 3D Photos. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 769–782. https://doi.org/10.1145/3472749.3474785

[28] Melanie Heck, Christian Becker, and Viola Deutscher. 2023. Webcam Eye Tracking for Desktop and Mobile Devices: A Systematic Review. In *Proceedings of the 56th Hawaii International Conference on System Sciences*. 6820–6829. https://hdl.handle.net/10125/103459

[29] Chih-Fan Hsu, Yu-Shuen Wang, Chin-Laung Lei, and Kuan-Ta Chen. 2019. Look at Me! Correcting Eye Gaze in Live Video Communication. 15, 2, Article 38 (jun 2019), 21 pages. https://doi.org/10.1145/3311784

[30] Alyssa Hughes. 2021. New Future of Work: Meeting and collaborating in a remote and hybrid world with Jaime Teevan and Abigail Sellen. https://www.microsoft.com/en-us/research/podcast/new-future-of-work-meeting-and-collaborating-in-a-remote-and-hybrid-world-with-jaime-teevan-and-abigail-sellen/

[31] ITU-R. 1998. Rec. ITU-R BT.1359-1 1 RECOMMENDATION ITU-R BT.1359-1 RELATIVE TIMING OF SOUND AND VISION FOR BROADCASTING.

[32] Jerald M Jellison and William John Ickes. 1974. The power of the glance: Desire to see and be seen in cooperative and competitive situations. *Journal of Experimental Social Psychology* 10, 5 (1974), 444–450.

[33] Jason Jerald and Mike Daily. 2002. Eye gaze correction for videoconferencing. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '02*. ACM Press, New Orleans, Louisiana, 77. https://doi.org/10.1145/507072.507088

[34] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System. 28, 3, Article 64 (jul 2009), 8 pages. https://doi.org/10.1145/1531326.1531370

[35] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

[36] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM transactions on graphics (TOG)* 37, 4 (2018), 1–14.

[37] Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. 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 (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2531–2540. https://doi.org/10.1145/2207676.2208640

[38] Jesper Kjeldskov, Jacob H. Smedegård, Thomas S. Nielsen, Mikael B. Skov, and Jeni Paay. 2014. EyeGaze: enabling eye contact over video. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14*. ACM Press, Como, Italy, 105–112. https://doi.org/10.1145/2598153.2598165

[39] Chris L. Kleinke. 1986. Gaze and eye contact: A research review. *Psychological Bulletin* 100, 1 (1986), 78–100. https://doi.org/10.1037/0033-2909.100.1.78

[40] Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. 2021. Project Starline: A High-Fidelity Telepresence System. 40, 6, Article 242 (dec 2021), 16 pages. https://doi.org/10.1145/3478513.3480490

[41] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication* 3, 2 (09 1997). https://doi.org/10.1111/j.1083-6101.1997.tb00072.x JCMC321.

[42] Andrew Monk, John McCarthy, Leon Watts, and Owen Daly-Jones. 1996. Measures of process. In *CSCW requirements and evaluation*. Springer, 125–139.

[43] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (June 2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811 Conference Name: IEEE Robotics & Automation Magazine.

[44] Lothar Muhlbach, Martin Bocker, and Angela Prussog. 1995. Telepresence in Videocommunications: A Study on Stereoscopy and Individual Eye Contact. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 2 (June 1995), 290–305. https://doi.org/10.1518/001872095779064582

[45] David T. Nguyen and John Canny. 2007. Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1465–1474. https://doi.org/10.1145/1240624.1240846

[46] Ken-Ichi Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. 1994. Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*. ACM Press, Chapel Hill, North Carolina, United States, 385–393. https://doi.org/10.1145/192844.193054

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

[48] Kazuhiro Otsuka. 2016. MMSpace: Kinetically-augmented telepresence for small group-to-group conversations. 19–28. https://doi.org/10.1109/VR.2016.7504684 ISSN: 2375-5334.

[49] Kazuhiro Otsuka. 2018. Behavioral Analysis of Kinetic Telepresence for Small Symmetric Group-to-Group Meetings. *IEEE Transactions on Multimedia* 20, 6 (June 2018), 1432–1447. https://doi.org/10.1109/TMM.2017.2771396 Conference Name: IEEE Transactions on Multimedia.

[50] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1716–1725. https://doi.org/10.1145/2818048.2819965

[51] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. *arXiv preprint arXiv:2312.02069* (2023).

[52] H. Regenbrecht, L. Müller, S. Hoermann, T. Langlotz, M. Wagner, and M. Billinghurst. 2014. Eye-to-eye contact for life-sized videoconferencing. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. ACM, Sydney New South Wales Australia, 145–148. https://doi.org/10.1145/2686612.2686632

[53] René Riedl. 2022. On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets* 32, 1 (March 2022), 153–177. https://doi.org/10.1007/s12525-021-00501-3

[54] Giuseppe Riva, Brenda K. Wiederhold, and Fabrizia Mantovani. 2021. Surviving COVID-19: The Neuroscience of Smart Working and Distance Learning. *Cyberpsychology, Behavior and Social Networking* 24, 2 (Feb. 2021), 79–85. https://doi.org/10.1089/cyber.2021.0009

[55] Shane L Rogers, Rebecca Broadbent, Jemma Brown, Allan Fraser, and Craig P Speelman. 2022. Realistic motion avatars are the future for social interaction in virtual reality. (2022).

[56] Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. 2018. Capturing, Representing, and Interacting with Laughter. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173932

[57] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2023. Relightable Gaussian Codec Avatars. (2023). arXiv:2312.03704 [cs.GR]

[58] Abigail J. Sellen. 1992. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*. ACM Press, Monterey, California, United States, 49–59. https://doi.org/10.1145/142750.142756

[59] Soo Yun Shin, Ezgi Ulusoy, Kelsey Earle, Gary Bente, and Brandon Van Der Heide. 2022. The effects of self-viewing in video chat during interpersonal work conversations. *Journal of Computer-Mediated Communication* 28, 1 (11 2022). https://doi.org/10.1093/jcmc/zmac028 arXiv:https://academic.oup.com/jcmc/article-pdf/28/1/zmac028/47194912/zmac028.pdf

[60] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications*. Toronto; London; New York: Wiley.

[61] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

[62] David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. In *Human-Computer Interaction – INTERACT 2011 (Lecture Notes in Computer Science)*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer, Berlin, Heidelberg, 162–180. https://doi.org/10.1007/978-3-642-23774-4_16

[63] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. 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* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 197–200. https://doi.org/10.1145/1460563.1460593

[64] Jian Sun and Holger Regenbrecht. 2007. Implementing three-party desktop videoconferencing. In *Proceedings of the 2007 conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artifacts and environments - OZCHI '07*. ACM Press, Adelaide, Australia, 95. https://doi.org/10.1145/1324892.1324910

[65] Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[66] Katherine M. Tsui, Munjal Desai, and Holly A. Yanco. 2012. Towards measuring the quality of interaction: communication through telepresence robots. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems - PerMIS '12*. ACM Press, College Park, Maryland, 101. https://doi.org/10.1145/2393091.2393112

[67] Rick van der Kleij, Jan Maarten Schraagen, Peter Werkhoven, and Carsten K. W. De Dreu. 2009. How Conversations Change Over Time in Face-to-Face and Video-Mediated Communication. *Small Group Research* 40, 4 (Aug. 2009), 355–381. https://doi.org/10.1177/1046496409333724 Publisher: SAGE Publications Inc.

[68] Roel Vertegaal and Yaping Ding. 2002. Explaining effects of eye gaze on mediated group conversations:: amount or synchronization?. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02*. ACM Press, New Orleans, Louisiana, USA, 41. https://doi.org/10.1145/587078.587085

[69] Roel Vertegaal, Gerrit Van Der Veer, and Harro Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. *Proceedings of Graphics Interface 2000* Montréal (2000), 8 pages, 366.40 KB. https://doi.org/10.20380/GI2000.14

[70] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. 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)*. Association for Computing Machinery, New York, NY, USA, 521–528. https://doi.org/10.1145/642611.642702

[71] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[72] Oscar Michael Watson. 1970. *Proxemic behavior: a cross-cultural study.*

[73] Bob G. Witmer, Christian J. Jerome, and Michael J. Singer. 2005. The Factor Structure of the Presence Questionnaire. *Presence* 14, 3 (June 2005), 298–312. https://doi.org/10.1162/105474605323384654 Conference Name: Presence.

[74] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2018. GazeDirector: Fully Articulated Eye Gaze Redirection in Video. *Computer Graphics Forum (CGF)* 37, 2 (2018), 217–225. https://doi.org/10.1111/cgf.13355

[75] Yang Xu, Chenguang Yu, Jingjiang Li, and Yong Liu. 2012. Video Telephony for End-Consumers: Measurement Study of Google+, IChat, and Skype. In *Proceedings of the 2012 Internet Measurement Conference* (Boston, Massachusetts, USA) *(IMC '12)*. Association for Computing Machinery, New York, NY, USA, 371–384. https://doi.org/10.1145/2398776.2398816

[76] Svetlana Yarosh and Panos Markopoulos. 2010. Design of an instrument for the evaluation of communication technologies with children. In *Proceedings of the 9th International Conference on Interaction Design and Children (IDC '10)*. Association for Computing Machinery, New York, NY, USA, 266–269. https://doi.org/10.1145/1810543.1810587

[77] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4578–4587.

A IMPLEMENTATION DETAILS OF OUR SIMULATION

A.1 Eye-tracking accuracy

We tracked participants' gaze using a Tobii Eye Tracker 5. It uses a near-infrared and RGB sensor operating at an interlaced sampling rate of 133Hz and a non-interlaced rate of 33Hz. The maximum supported field of view in each direction is 40 degrees. It can recover lost gaze using neural head tracking. Gibaldi et al. [21] tested an older version of the Tobii eye tracker. They found it has at least an accuracy of 0.6° and an end-to-end latency of around 47 ± 4 ms when receiving the data via a local UDP connection. We found these specifications suitable for our design, and since the gaze elements in question were rather large, we encountered no issues with the eye-tracker's accuracy.

For the implementation of our concept with web cam eye tracking, we calculated a minimum accuracy using triangulation and the smallest margin for error in our interface. A gaze focused on centre interlocutor (Figure 5) can horizontally deviate by 6.75 cm on a 24 inch desk screen, and by 4.25 cm on 13 inch laptop before it leads to an error. Assuming a viewing distance of 70 cm for the desk monitor, and 50 cm for the laptop screen the required accuracy is 5.51° for the desktop, and 4.86° for the laptop.

A.2 Latency

We explored several design alternatives to reduce latency and skew while maintaining a frame rate of 30 Hz. Ultimately we chose an architecture were each client would send an separate audio and H265 video stream to the other three clients without any synchronisation. Measured latency was 80.0 ± 0.2 ms for audio and 133.33 ± 33.33 ms for video (See Appendix **??** for details). Consequently, skew is 53.33 ± 33.53 ms, falling within ITU recommendations G.114 and BT.1359-1 [17, 31]. It also outperforms common video conferencing solutions for which Xu et al. [75] reported delays of 130 to 270 ms for audio and 230 to 270 ms for video latency, for common multiparty video conferencing solutions.

A.3 Fading between camera views

Instant transitions between camera views are visually distracting, so scaled alpha-blending was used to blend the current view with the subsequent view:

$$g(x) = (1 - \alpha) f_0(x) + \alpha f_1(x), \tag{1}$$

where g is the new image and f_0 and f_1 are the images that are blended. Here α is increased from 0 to 1 as time t increases from 0 to T:

$$\alpha = (e^{\frac{3t}{T}} - 1)/(e^3 - 1). \tag{2}$$

This scaled exponential function for α gives a strong visual hint initially of the transition and then smoothly fades out. We tested several blending functions and found this one to be an effective compromise between rapid initial signalling of change in attention and maintaining an illusion of smooth head rotation.

B USER STUDY DETAILS

B.1 Controversial statements for group discussion

- Research and development of brain-machine interfaces, such as Elon Musk's Neuralink, should be prohibited or at least placed under strict regulation, as reading one's thoughts has dramatic ethical implications.
- Covid vaccination should be compulsory for all those who are not expected to suffer long-term adverse health effects from vaccination.
- Industrial livestock farming should be progressively banned.
- Short-distance flights should be banned or taxed heavily.
- Physically healthy people should have the right to euthanasia (e.g. by taking a deadly pill under a doctor's supervision) if this is their own explicit wish.

B.2 Details of game design

The survival game involves four players. Based on a majority vote, they have to decide on one out of three items to bring to the island. Once they reach a consensus, we present the next set of items. We instructed them that several items would be crucial for survival (For handout and recorded instructions, see supplemental material). Wanting to bring as many items as possible within the seven minutes they are allowed to play provided an incentive to reach an agreement quickly on each item. However, before commencing the game, each player was assigned a role undisclosed to the others. Three players are cooperative. Each received information about a different set of two items necessary for survival (six items in total.) We were inspired by Vertegaal and Ding [68], who also distributed information necessary for success among participants. In addition to these crucial items, we also assigned them a non-crucial item. Their secondary task was to convince the group to choose it at least once. They win the game if, at the moment time runs out, the group has chosen this item at least once in addition to all other items necessary for survival. The fourth player is uncooperative with the goal of jeopardising the survival plan and is given a lot of information about crucial items. In addition, we made them aware of a doom item. The uncooperative player wins if the group is convinced to take this item or fails to select all necessary items.

We chose this game design because the screen needed to show very little information, allowing players to still focus on their interlocutors. They were incentivised to do so since facial expressions might reveal who the uncooperative player is. At the same time, every player would eventually suggest an item non-crucial for survival to win the game, adding distrust to the social dynamics. We decided against using the survival of the team as a performance measure. Such measures are typically only sensitive to substantial manipulations of the experimental factors [42, 69], which we did not expect.

B.3 Covid-19 protection measures

We took extensive measures to comply with local Covid regulations. The selection of possible participants was restricted to vaccinated or negative tested members or guests of the university. They were required to wear an FFP2 mask until they were alone in their assigned room. We ventilated all rooms prior to any experimental session. Further, we disinfected all materials and surfaces that participants interacted with. Based on the finding of Abraham et al. [1], turtlenecks were ensured to be SARS-CoV-2 virus free by treating them for at least 5 min with 65°C hot air or 20 min with 40° hot water. Further, we used waterproofed in-ear headphones cleaned and disinfected in a 15 min ultrasonic bath using a 4% Instrusol AF+ solution. Our governing authority approved our documented safety measures before conducting the experiment.

B.4 Questionnaire items

The questions asked are listed in table 1, together with a record of prior publications from which they were derived. The UX questionnaire included statements assessing participants' perception and interpretation of three types of gazes: direct eye contact, directed third-party gazes, and "off-gazes" directed at no one. For each of those three gaze types, three statements were included. One asked whether such gazes were perceivable; another asked whether they helped to notice interlocutors' attention; and a last one inquired if they aided turn-taking. There were also two statements about the difficulty of turn-taking. The comparative questionnaire asked about turn-taking, perceiving attention and naturalness of interaction. Social presence was measured here with the *Semantic Differential* questionnaire. Its questions were based on the smallest adequate set of four attribute pairs [60]: *cold-warm*, *impersonal-personal*, *insensitive-sensitive* and *unsociable-sociable*.

Statement	based on	measures
It was exciting to follow the discussion.	Sellen [58]	Engagement
Turn-taking was difficult.	Böcker and Mühlbach [11]	turn-taking
I was able to take control of the conversation when I wanted to.	[58]	turn-taking
The conversation was highly interactive.	[58]	Engagement
Sometimes I had the feeling I was excluded from the conversation.	Pilot study	UX issue
I could not contribute anything to the solution we came up with.	Hauber et al. [25]	Satisfaction
The system was distracting me from the conversation.	Pilot study	UX issue
During the experiment, I had the feeling we were all in the same room.	[25]	Virtual presence
One does not get a good enough idea of how people at the other end are reacting.	Champness [14]	Social presence
I couldn't get to know people very well if I only met them over this system.	[14]	Social presence
I was always aware of my partner's presence.	[25]	Social presence
It was easy for me to notice when my conversation partners looked at me.	[11, 25]	Perceivable direct eye contact
I knew when I was being addressed by someone.	[11]	non-verbal cues for turn-taking
I knew when someone was listening to me or paying attention to me.	[11, 58]	non-verbal cues for attention
It was easy to notice when my conversation partners looked at someone else (other than me).	[25]	Perceivable third-party gaze
I knew when I was not addressed (but instead, someone else was.)	[11]	non-verbal cues for turn-taking
I knew when someone was listening or paying attention to someone else (other than me.)	[11, 58]	non-verbal cues for attention
It was easy to notice when my conversation partners looked at no one (not at me and not at anybody else.)	-	Perceivable third-party gaze
I knew when someone was not following the conversation, was thinking about something or became distracted.	[25]	non-verbal cues for attention
I knew when someone was thinking about something.	[25]	non-verbal cues

Table 1. **UX questionnaire.** Participants rated these 20 statements on a 7-point Likert scale after using each system. The scale ranges from "strongly disagree" to "strongly agree."

Question	based on	measures
Which system would you recommend to your friends and colleagues?	-	Overall satisfaction
For which system was turn-taking easier?	[11]	turn-taking
Which system facilitated a more natural interaction with your conversation partners?	Witmer et al. [73]	Social presence and overall satisfaction
With which system was the interaction more engaging/exciting?	[58]	Engagement
Which system was better for noticing if your conversation partners were paying attention to you or someone/something else?	[11, 58]	non-verbal cues for attention
With which system was the interaction more social?	-	Social presence
Which system would you choose for a meeting where you intend to persuade other people?	Champness [14]	Social presence

Table 2. **Comparative questionnaire.** Participants indicated for these questions which system they prefer or whether they prefer both equally. That allows for a focused comparison between systems while still revealing participants' uncertainty.

B.5 Statistical Analysis Details

Comparative Questions. Participants had the option, in the questionnaire, to like both systems equally. In those cases votes were omitted from analysis, treated solely as an indicator of uncertainty about that question. A single-tailed proportions test was used, with the null hypothesis that each system would be liked equally ($p_0 = 0.5$).

Social presence semantic differential. We assessed the reliability of the attributes used to measure the underlying concept of social presence using Cronbach α. Treating participants' ratings as ordinal values, each participant's social presence score was calculated as the median score of the four attributes. Significance was assessed using a Wilcoxon signed-rank test, with Tiled View serving as the baseline, and Gazing Heads as the treatment.

User experience questionnaire. Participants' ratings were treated as ordinal, assessing significant differences between the two systems again using a Wilcoxon signed-rank test and Tiled View as the baseline.

Definitions of terms for speech analysis. Since the experimental design and baseline condition (Tiled View) of Sellen [58] bears significant similarity with our work, we use their definitions of terms stated in their paper:

> A *turn* consists of the sequence of talk-spurts and pauses by a speaker who "has the floor". A speaker gains the floor when they begin speaking to the exclusion of everyone else and when they are not interrupted by anyone else for at least 1.5 seconds. The duration of a turn begins with the first unilateral sound, and ends when another individual turn or a "group turn" begins. Note that turns therefore include periods of mutual silence at the end of utterances, when no one else has yet taken the floor.
>
> A *group turn* begins the moment an individual turn taker has fallen silent and two or more others are speaking together; the group turn ends the moment any individual is again speaking alone.
>
> *Simultaneous speech* is speech by one or more speakers who do not have the floor. [We] further distinguish between overlaps and simultaneous speech which do not lead to a speaker switch. Simultaneous speech which does not precede a speaker switch is called non-interruptive simultaneous speech.

For their analysis Sellen [58] used several absolute measures based on a fixed session length of 16 minutes. In our case, sessions varied in length, therefore requiring conversion to relative measures. The number of turns was converted to turn frequency, and the amount of simultaneous speech to the proportion of session time during which a given type of speech occurred.

C ADDITIONAL DETAILS FOR RESULTS OF STATISTICAL ANALYSIS

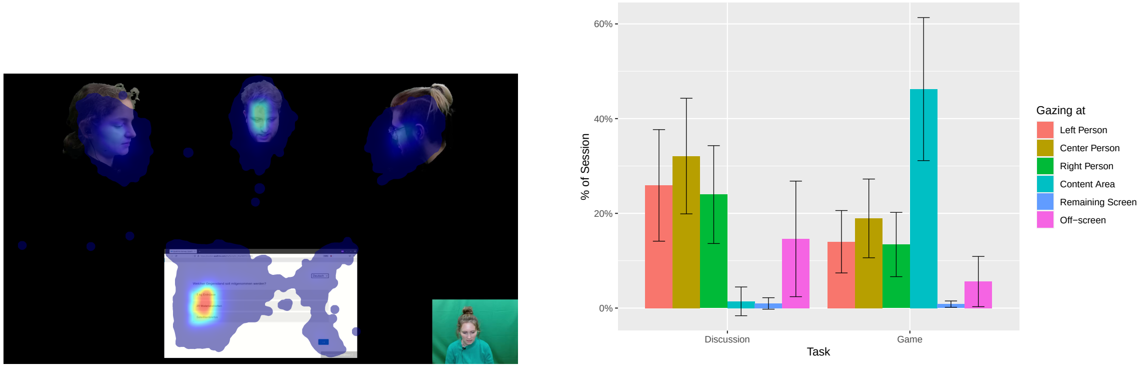
Figure 12 shows the F-values and descriptive statistics aggreagted for each factors of the two factorial within-sibject ANOVA. Figure 13 shows the quantiles and median for all rating of the ux questionnaire items. Figure 14 shows the quantiles and median for the semantic differential. Figure 15 shows all percentages of ratings of the comparative questionnaire.

C.1 More dynamic turn taking during game

The results of our speech activity analysis shown Figure 12 indicate that turning was more dynamic during the game given the following significant effects:

(1) Turn switches occurred more frequently
(2) Turns were shorter
(3) Group turns occurred more frequently
(4) Less time with only one person speaking
(5) More Simultaneous speech
(6) More none interruptive simultaneous speech
(7) Turns were more evenly distributed.
(8) Percentage of time only one person spoke was lower
(9) - not significant -
(10) Higher overlaps in speaker switches
(11) Lower switching time

C.2 Focus areas received different amounts of visual attention



(a) **Heatmap of visual attention:** In the game task, the content area at the bottom of the screen received significantly more attention than the faces.

(b) **Distribution of visual attention:** The proportion of session time participants spent gazing at the different focus areas was distributed differently during the game and the discussion.

Fig. 16. Results of eye tracking for the different tasks.

	Factor: System			Factor: Task		
	Gazing Heads	Tiled View	F and p-Value	Discuss	Game	F and p-Value
Turn Frequency/Minute	4.50 (1.1)	4.54 (0.7)	$F(1, 18) = 0.028$ $p = 0.869$	2.72 (1.1)	6.32 (1.1)	$F(1, 18) = 291.247$ $\mathbf{p < 0.001}$
Turn Duration	16.26s (10.9)	15.36 (10.7)	$F(1, 18) = 0.400$ $p = 0.535$	23.73 (10.1)	7.89 (1.9)	$F(1, 18) = 82.083$ $\mathbf{p < 0.001}$
Group Turn Freq./Minute	0.82 (0.8)	1.00 (1.0)	$F(1, 18) = 5.046$ $\mathbf{p = 0.037}$	0.23 (0.2)	1.60 (0.7)	$F(1, 18) = 123.869$ $\mathbf{p < 0.001}$
Turn Distribution (H)	1.93 (0.1)	1.93 (0.1)	$F(1, 18) = 0.255$ $p = 0.620$	1.90 (0.1)	1.96 (0.0)	$F(1, 18) = 14.431$ $\mathbf{p = 0.001}$
Time one Person spoke	81.72% (10.3)	80.62% (10.9)	$F(1, 18) = 1.140$ $p = 0.300$	90.40% (3.8)	72.0% (5.9)	$F(1, 18) = 242.338$ $\mathbf{p < 0.001}$
Simultaneous Speech	6.78% (6.1)	8.23% (6.9)	$F(1, 18) = 3.949073$ $p = 0.062$	2.90% (2.7)	12.12% (6.0)	$F(1, 18) = 101.303$ $\mathbf{p < 0.001}$
non-Int. Simult. Speech	6.64% (5.7)	7.50% (5.8)	$F(1, 18) = 1.883$ $p = 0.187$	2.76% (2.2)	11.39% (4.9)	$F(1, 18) = 135.425$ $\mathbf{p < 0.001}$
Interruptive Simult. Speech	2.48% (2.2)	2.83% (2.4)	$F(1, 18) = 1.263$ $p = 0.276$	1.04% (1.2)	4.28% (2.0)	$F(1, 18) = 157.325$ $\mathbf{p < 0.001}$
Sim. Speech Taking Control	31.08% (16.1)	26.67% (11.6)	$F(1, 18) = 1.677$ $p = 0.212$	26.85% (18.4)	30.90% (7.7)	$F(1, 18) = 2.771$ $p = 0.113$
Speaker Switches Overlaps	34.34% (17.2)	34.96 (17.0)	$F(1, 18) = 0.059$ $p = 0.810$	24.65% (16.3)	44.64% (10.7)	$F(1, 18) = 67.024$ $\mathbf{p < 0.001}$
Switching Time	0.59s (0.8)	0.42s (0.6)	$F(1, 18) = 2.037$ $p = 0.171$	0.85s (0.8)	0.16s (0.4)	$F(1, 18) = 37.438$ $\mathbf{p < 0.001}$
Focus Changes per Minute	58.76 (22.0)	60.42 (22.7)	$F(1, 75) = 3.874$ $p = 0.053$	58.29 (23.4)	60.88 (21.2)	$F(1, 75) = 2.632$ $p = 0.109$
Eye Contact per Minute	22.15 (13.6)	20.67 (11.8)	$F(1, 75) = 7.480$ $\mathbf{p = 0.008}$	29.46 (12.1)	13.36 (6.9)	$F(1, 75) = 195.247$ $\mathbf{p < 0.001}$
Eye Contact (% of Session)	18.9% (12.3)	17.6% (11.3)	$F(1, 75) = 4.630$ $\mathbf{p = 0.036}$	26.88% (10.32)	9.69% (5.12)	$F(1, 75) = 304.942$ $\mathbf{p < 0.001}$
Eye Contact Duration	0.52s (0.2)	0.51s (0.2)	$F(1, 75) = 0.365$ $p = 0.365$	0.58s (0.2)	0.46s (0.1)	$F(1, 75) = 83.649$ $\mathbf{p < 0.001}$

Fig. 12. **Speech and eye-gaze analysis.** The first three columns compare our two systems, while averaging over the two tasks. The remaining three columns compare tasks, while averaging over the two systems. Note that one session was interrupted due to network problems. Hence, it needed to be excluded from the voice and eye-tracking analysis.

We analysed what proportion of a session participants spent looking at the different focus areas. The focus area was added as the third predictor to the repeated measures ANOVA. Since Mauchly's test indicated that the assumption of sphericity had been violated for this predictor and all its interactions, the degrees of freedom were Greenhouse-Geisser corrected. Interaction effects were analysed using a Bonferroni post-hoc test. As expected, the proportion of time spent gazing was different among focus areas ($F(3.21, 241.09) = 169.495$, $p < 0.001$). More importantly, there was an interaction effect between the task participants worked on and the proportion of time they gazed at different focus areas ($F(3.21, 241) = 415$, $p < 0.001$). As shown in Figure 16b, for the discussion task, participants spend most of their

Question	System	25st Quantile	Median	75st Quantile	p-Value
It was exciting to follow the discussion	GH	1.5	2	3	0.044
	TV	1	2	2	
Turn-taking was difficult *	GH	1*	2*	2*	0.084
	TV	0*	1*	2*	
I was able to take control of the conversation when I wanted to	GH	1	2	3	0.176
	TV	1	2	2	
The conversation was highly interactive	GH	1	2	3	<0.001
	TV	0	2	2	
Sometimes I had the feeling I was excluded from the conversation *	GH	1*	2*	3*	0.790
	TV	1*	2*	2*	
I could not contribute anything to the solution we came up with.	GH	2*	2*	3*	0.246
	TV	1*	2*	3*	
The system was distracting me from the conversation *	GH	-0.5*	1*	2*	<0.001
	TV	2*	2*	3*	
During the experiment I had the feeling we were all in the same room	GH	-0.5	1	2	<0.001
	TV	-2	-1	1	
One does not get a good enough idea of how people at the other end are reacting *	GH	-1*	1*	2*	0.263
	TV	-1*	1*	2*	
I couldn’t get to know people very well if I only met them over this system *	GH	-1*	1*	2*	0.010
	TV	-1*	0*	2*	
I was always aware of my partner’s presence	GH	1	2	3	0.002
	TV	1	2	2	
It was easy for me to notice when my conversation partners looked at me	GH	1	2	3	<0.001
	TV	-2	-1	0	
I knew when I was being addressed by someone	GH	1	2	3	<0.001
	TV	-1	1	2	
I knew when someone was listening to me or paying attention to me	GH	2	2	3	<0.001
	TV	-1	1	2	
It was easy to notice when my conversation partners looked at someone else (other than me)	GH	1	2	3	<0.001
	TV	-2	-1	0	
I knew when I was not addressed (but instead someone else was)	GH	1	2	2	<0.001
	TV	-1	0	1	
I knew when someone was listening or paying attention to someone else (other than me)	GH	1	2	3	<0.001
	TV	-1	0	1.5	
It was easy to notice when my conversation partners looked at no one (not at me or anybody)	GH	-2	0	1	0.708
	TV	-2	-1	1	
I knew when someone was not following the conversation, was thinking about something or became distracted	GH	0	1	1	0.289
	TV	0	1	2	
I knew when someone was thinking about something	GH	-1	1	1	0.993
	TV	-1	0	1	

Fig. 13. **Detailed results of the UX questionnaire**: Quantiles and median ratings. p-Values were obtained using a Wilcoxon signed-rank test. Note: * indicates that the rating had been multiplied with -1 to account for the inverse phrasing of the question. The scale ranges from “strongly disagree” (−3) to “strongly agree” (3).

Attribute	System	25st Quantile	Median	75st Quantile
Sociable	GH	1	2	3
	TV	0	1	2
Personal	GH	1	2	3
	TV	-1	1	2
Sensitive	GH	1	1	2
	TV	0	1	2
Warm	GH	1	2	2
	TV	0	1	2
Social presence	GH	1	2	2
	TV	0	1	2

Fig. 14. **Detailed results of the semantic differential**: Quantiles and median ratings

	Preferred	Preferred	Preferred	p-Value
	Gazing Heads	both equally	Tiled View	
Which system would you recommend to your friends and colleagues?	49.37%	18.99%	31.65%	**0.052**
For which system was turn-taking easier?	50.63%	30.38%	18.99%	< **0.001**
Which system facilitated a more natural interaction with your conversation partners?	62.03%	18.99%	18.99%	< **0.001**
With which system was the interaction more engaging/exciting?	87.34%	8.86%	3.80%	< **0.001**
Which system was better for noticing if your conversation partners were paying attention to you or someone/something else?	91.14%	6.33%	2.53%	< **0.001**
With which system was the interaction more social?	62.03%	25.32%	12.66%	< **0.001**
Which system would you choose for a meeting where you intend to persuade other people?	50.63%	21.52%	27.85%	**0.015**
Which system would you like to use for the interview?	72.15%	7.59%	20.25%	< **0.001**

Fig. 15. **Detailed results of the compartive questionaires**: Participants preferences regarding the two systems.

time looking at the centre interlocutor (M = 32.1%, SD = 12.2%). The second and third largest proportions of time were spent looking at the left (M = 25.9%,SD = 11.8%) and the right interlocutor (M = 24.0%, SD = 10.3%) with no significant difference between them. The fourth largest proportion of time was spent looking off-screen (M = 14.6%, SD = 12.2%). The significantly smallest amount of time was spent looking at the empty content area (M = 1.42%,SD = 3.04%) and the remaining screen (M = 0.98%, SD = 1.21%) with no significant difference between them.

For the game, the main difference is that participants looked at the content area (M = 46.2%, SD = 15.1%) significantly more than any other focus area, which is unsurprising since it contained game-relevant information. The second, third and fourth largest proportion of time was spent looking at the centre (M = 18.9%, SD = 8.31%), left (M = 14.0%, SD = 6.59%) and right interlocutor (M = 13.4%, SD = 6.79%). Again the centre interlocutor was gazed at significantly more than the left and right one. The left and right interlocutors received similar gaze times. Participants looked off-screen (M = 5.60%, SD = 5.30%) significantly less than any face or the content areas but significantly more frequently than on the remaining screen showing no content (M = 0.83%, SD = 0.67%).

As a result the frequency of mutual eye was significantly ($F = 195, p < 0.01$) reduced by 54.6% during the game ($M = 13.380 min^{-1}$) in comparison to the discussion ($M = 29.460 min^{-1}$). The duration of mutual eye contact was also significantly ($F = 83.6$, $p < 0.001$) reduced by 20% during the game ($M = 0.458s$, $SD = 0.144s$) in comparison to the discussion ($M = 0.577s$, $SD = 0.193s$). The overall share of session spend gazing with mutual gazes was significantly ($F = 305$, $p < 0.001$) lower the game ($M = 9.69\%$, $SD = 5.12\%$) in comparison to the discussion ($M = 26.9\%$, $SD = 10.3\%$).