ViGather: Inclusive Virtual Conferencing with a Joint Experience Across Traditional Screen Devices and Mixed Reality Headsets

HUAJIAN QIU, Department of Computer Science, ETH Zürich, Switzerland
PAUL STRELI, Department of Computer Science, ETH Zürich, Switzerland
TIFFANY LUONG, Department of Computer Science, ETH Zürich, Switzerland
CHRISTOPH GEBHARDT, Department of Computer Science, ETH Zürich, Switzerland
CHRISTIAN HOLZ, Department of Computer Science, ETH Zürich, Switzerland

Fig. 1. *ViGather* is a teleconferencing system that integrates users of traditional screen devices (e.g., PCs, laptops, tablets) and Mixed Reality users wearing headsets into a joint experience. ViGather represents *all* users through embodied and animated avatars that reflect the verbal and non-verbal interaction between users in a shared virtual environment. (a) A laptop user emphasizes his statements through hand gestures while videoconferencing with a user of an immersive system. (b) ViGather represents both users as avatars in the same virtual environment, rendering each participant's experience from a first-person perspective on their respective display. (c) To animate avatars' non-verbal communication cues, our system tracks body language and gaze direction using the sensors inside the headsets for Mixed Reality users, while it processes the front-facing camera feed to extract the same behavioral cues on traditional screen devices.

Teleconferencing is poised to become one of the most frequent use cases of immersive platforms, since it supports high levels of presence and embodiment in collaborative settings. On desktop and mobile platforms, teleconferencing solutions are already among the most popular apps and accumulate significant usage time—not least due to the pandemic or as a desirable substitute for air travel or commuting.

In this paper, we present *ViGather*, an immersive teleconferencing system that integrates users of all platform types into a *joint* experience via equal representation and a first-person experience. ViGather renders all participants as embodied avatars in *one* shared scene to establish co-presence and elicit natural behavior during collocated conversations, including nonverbal communication cues such as eye contact between participants as well as body language such as turning one's body to another person or using hand gestures to emphasize parts of a conversation during the virtual hangout. Since each user embodies an avatar and experiences situated meetings from an egocentric perspective no matter the device they join from, ViGather alleviates potential concerns about self-perception and appearance while mitigating potential 'Zoom fatigue', as users' self-views are not shown. For participants in Mixed Reality, our system leverages the rich sensing and

Authors' addresses: firstname.lastname@inf.ethz.ch, Department of Computer Science, ETH Zürich, Switzerland.

2573-0142/2023/9-ART232 $15.00
https://doi.org/10.1145/3604279

reconstruction capabilities of today's headsets. For users of tablets, laptops, or PCs, ViGather reconstructs the user's pose from the device's front-facing camera, estimates eye contact with other participants, and relates these non-verbal cues to immediate avatar animations in the shared scene.

Our evaluation compared participants' behavior and impressions while videoconferencing in groups of four inside ViGather with those in Meta Horizon as a baseline for a social VR setting. Participants who participated on traditional screen devices (e.g., laptops and desktops) using ViGather reported a significantly higher sense of physical, spatial, and self-presence than when using Horizon, while all perceived similar levels of active social presence when using Virtual Reality headsets. Our follow-up study confirmed the importance of representing users on traditional screen devices as reconstructed avatars for perceiving self-presence.

CCS Concepts: • **Human-centered computing** → *Human computer interaction (HCI)*; *Empirical studies in HCI*; *Virtual reality*; **Mixed / augmented reality**; **Collaborative interaction**; *Collaborative and social computing devices*; • **Applied computing** → *Internet telephony*.

Additional Key Words and Phrases: Video Conferencing; Social VR; Embodied Presence; Mixed Reality; Virtual Reality; Teleconferencing; Cross-platform; Collaboration; Avatars; Co-presence; Immersive social interaction

ACM Reference Format:
Huajian Qiu, Paul Streli, Tiffany Luong, Christoph Gebhardt, and Christian Holz. 2023. ViGather: Inclusive Virtual Conferencing with a Joint Experience Across Traditional Screen Devices and Mixed Reality Headsets. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 232 (September 2023), 27 pages. https://doi.org/10.1145/3604279

1 INTRODUCTION

A large driver of (mobile) technology adoption has been its benefit for communication between people, in the consumer space as well as in the business sector [4]. This desire for high-fidelity communication has propelled the development of mobile phones, internet telephony, and, not least, communication apps across platforms and devices, some of which have turned into entire ecosystems [39, 43, 47]. Similar to voice and chat applications, video conferencing has experienced considerable uptake, initially for personal communication [48] that rose with the increased availability of broadband and faster mobile phone networks. With the sudden halt of shared-office use during the pandemic, video conferencing has become a mainstay of communication within and across businesses, now dominating face-to-face meetings in many regards [51]. High-resolution and low-latency video streaming is also capable of capturing and playing back participants' non-verbal cues, such as facial gestures, hand gestures, or attentive behavior—to a level that approaches face-to-face conversations.

As consumer-oriented Mixed Reality (MR) platforms become increasingly affordable, especially those that require no elaborate setup or infrastructure and thus enable mobile use anywhere (e.g., Microsoft Hololens, Meta Quest 2, HTC Vive Flow), video conferencing could again accelerate the adoption of this new platform for mobile communication. Effective video conferencing on immersive platforms can support a diverse set of scenarios, such as remote assistance [52], maintenance and repair [18], business meetings [8], or collaborative work [50]. Previous studies have shown numerous benefits of immersive meetings in a collocated setting in terms of social presence [68], non-verbal communication [34, 35], and education [63, 75, 80], particularly due to the embodied communication and natural gestures the medium enables [58].

However, while chats between users of traditional screen devices (e.g., PCs, laptops, tablets) are compelling as much as chats between immersed MR users are compelling, hybrid solutions that simultaneously integrate both platforms fail to harvest the benefits of either platform. Although current immersive systems insert 2D depictions of desktop participants into immersive spaces (e.g., Horizon [45]), allowing MR users to interact with screen-device users, the user experience on screen devices considerably limits participants' perceived physical presence and reduces their

engagement [49, 68]. This raises the question how traditional 2D video chats may be fused with immersive 3D interfaces in a manner that establishes a *common* experience in a shared setting, comprising screen devices and immersive 3D platforms.

In this paper, we present *ViGather*, a teleconferencing system that represents all participants as avatars in a shared 3D environment, no matter if they join from mobile or stationary screen devices or from immersive devices. ViGather's key benefit is its capability of enabling all avatars and, thus, users to express non-verbal communication cues through body motions, gestures, and attentive behavior. Our system enables this by modeling user representation *independent* of the platform used to join a conversation, reconstructing non-verbal cues from users during interaction and relaying them onto their respective 3D avatars. Thus, users cannot tell apart participants joining from Mixed Reality systems from those on screen devices in ViGather's experience. This core feature of our system puts all participants onto the same level, leading each user to express the same behavior towards all participants as relates to expressiveness, engagement, and attentiveness.

1.1 Hybrid video conferencing with equal representation in *one* shared setting

Figure 1a shows ViGather with two participants during a video conversation. One participant connects to the collaborative setting using a laptop, whereas the other is wearing a mobile Virtual Reality (VR) headset, which establishes their experience in an entirely virtual space. The two engage with other users in the same, shared environment (b) and, independent of the respective client platform, communicate by talking, either into the round or addressing individual participants by looking at them, and supporting their conversation through gestures that their avatar then expresses in the situated 3D experience.

Under the hood, our system establishes this shared 3D experience by integrating input cues across VR headsets and screen devices, mobile as well as stationary. While VR systems already provide animated avatars based on the user's motion (here Meta's Quest 2), ViGather extracts this information from the feed of the front-facing camera for participants joining on traditional screen devices. From the orientation of the user's head and the pose of their body, ViGather detects the face, estimates the direction of head gaze as an approximation of attention, fits an upper-body skeleton into the visible part of the person, and extrapolates the location of the user's hands when outside the camera's view (Figure 1c, left). Our system then maps these tracking cues to each 3D avatar and animates its head and face following detected motions, attention switches, and talking (Figure 1c, right).

Therefore, ViGather allows participants to *turn to* others, wave *at* each other, or point at objects in the scene—in the immersive virtual space as well as in the rendered experience on a traditional 2D screen—which establishes the basis for communication in a similar manner to real-life conversations. Each participant thereby experiences a rendering of their first-person perspective into the shared scene independent of the platform. Unlike in existing screen-device video conferencing where all participants face the camera and, thus, directly see all other participants accordingly, ViGather users obtain a sense of who is talking to them or addressing another participant, as conversations carry a *spatial* meaning and body orientation, attention, and body language is observable at all times. In addition, because our system represents all participants as 3D avatars, even when joining in front of a webcam, a welcome side effect is that concerns around appearance [56] and 'Zoom fatigue' may be alleviated [60].

In our evaluation with 16 participants, we found that ViGather created a higher sense of physical, spatial, and self-presence for participants on screen devices within the immersive meeting compared to participants on screen devices that used Meta's Horizon. At the same time, we found no significant difference in active social presence within the virtual meeting between ViGather's situated conference and Horizon's experience. Our subsequent evaluation focused on the representation of

screen-device users within the virtual environment. The results confirmed that the reconstructed avatars of those participants were essential for enhancing the participants' perceived self-presence.

Overall, our evaluation results support ViGather's ambition to power high-fidelity collaborative social experiences on future commodity devices in shared video conferencing settings, comprising emerging immersive devices, inexpensive mobile devices, as well as traditional desktop platforms.

1.2 Contributions

We make the following contributions in this paper:

- A telepresence experience that integrates users of traditional screen devices, mobile and stationary, and Mixed Reality users into a coherent and joint immersive setting, rendered from a first-person perspective, representing each participant as an embodied 3D avatar that is spatially configured and animated to reflect each participant's non-verbal communication and behavior, such as body language, gestures, and gaze direction as an approximation for attention.
- A prototype system that integrates all client platforms and handles audio and video exchange through a back-end server. ViGather's Virtual Reality clients derive users' non-verbal behavior directly from the headset's tracking, whereas our system reconstructs the same type of body language, gestures, and gaze on screen devices from the feed of the front-facing camera that is integrated into traditional screen devices. This allows such users—despite the 2D screen—to turn to others in the rendered 3D scene, pointing to or addressing each other much like they would when using an immersive system. Our telepresence server relays tracking information between clients to establish a consistent state, such that each platform can render the shared scene and provide the same experience for all participants.
- A 16-participant evaluation of ViGather in groups of four, comparing Meta Horizon within subjects as the baseline, and comparing platform type between subjects (e.g., screen device vs. VR headset). The results revealed ViGather's benefits for screen-device users in physical, spatial, and self-presence by participating in an immersive situated video teleconference.
- A second 16-participant study of ViGather compared to an additional baseline condition where screen-device participants shared ViGather's view and placement configuration but were represented through their actual webcam feed (instead of an avatar). The results emphasized the crucial role of virtual avatar reconstruction for participants' perceived self-presence.

2 RELATED WORK

ViGather is related to research on collaborative teleconferencing systems, webcam-based pose reconstruction, and avatar-based chats.

2.1 Collaborative teleconference systems

Several technologies have been explored on their suitability to augment conventional teleconferencing systems for collaboration. These include displays [17], projected augmented reality [57], augmented table environments [61], and virtual replica of physical environments [19].

Additionally, head-mounted displays (HMDs) for augmented reality (AR) and VR have opened up new possibilities to improve the experience of virtual meetings. Already in 1992, Kuzuoka [36] investigated HMDs for remote collaboration. Billinghurst et al. developed a technology to superimpose meeting participants on marker-specified locations [7]. *VROOM* allows participants to explore a remote environment using the camera of a mobile robot [30]. Participants wearing HMDs in the remote environment can observe a life-size avatar of the remote participant overlaid on the robot. *Holoportation* is an end-to-end system that reconstructs and transmits 3D environments including objects and people in real-time [55]. Other work proposes immersive telepresence systems where a

remote user perceives the environment of another user in first-person view, allowing the latter to share their experience [24, 32, 82].

Other research has investigated adaptations to the appearance of meeting participants to aid collaboration with a focus on gaze [26, 29, 33, 65, 72] and body representation [59]. Prior research found that the collaboration between participants inside environments with varying geometry benefits from mapping techniques that bring all participants into a joint VR experience [53, 67].

2.2 Webcam-based pose reconstruction

Contemporary AR and VR systems track the user's head and hand poses for embodiment and input. For this purpose, a range of pose-tracking systems exist that differ in their underlying sensing modality and the placement of their sensors. The literature generally differentiates between inside-out (e.g., Meta Quest 2) and outside-in (e.g., HTC VIVE) tracking depending on whether the sensors are placed directly on the tracked device or at fixed locations in the external environment.

Human pose estimation systems with surrounding cameras include high-end commercial marker-based systems [54] and marker-less multi-view setups [69]. Other real-time tracking solutions rely on only a single RGB-D camera that benefits from its additional depth channel [66]. On monocular systems, a variety of learning-based algorithms have been introduced that achieve strong results on human body estimation. These algorithms either predict a 2D [74] or a 3D [73] human pose from a monocular image and support real-time inference [42]. *TransforMR* replaces people and vehicles in video feeds from mobile devices with avatars and alternative objects, mapping semantically-consistent real-world behavior to virtual representations in a Mixed Reality experience [31]. For a more detailed survey, we refer to Chen et al.'s overview [11]. More recently, several methods have been introduced that estimate a user's full-body pose representation based on the 6D poses of the tracked VR headset and controllers [1, 28] or have sought to extend body tracking beyond the cameras' field of view [70].

The development of learning-based methods for pose reconstruction has enabled several commercial solutions that make these algorithms available to consumer devices. RGB-based pose estimation is supported by Apple's ARKit [3] and Google's open-source library MediaPipe [40], which incorporates solutions for tracking hands [83] and faces [84].

ViGather incorporates these recent advancements in RGB-based computer vision to afford screen-device users the use of fully animated 3D avatars within a virtual environment, including arm and head motions using the standard webcam built into their devices. To the best of our knowledge, ViGather is the first system that allows screen-device users to appear with the same representation and control over their avatars as provided only by current HMD devices in teleconferencing settings.

2.3 Avatar-based chats

Previous research has investigated the difference between video conferences and VR meetings. For example, Steinicke et al. [68] compared the go-to video conferencing tool *Zoom* [14] with the more immersive VR meeting platform *Mozilla Hubs* [20]. In their pilot study, participants completed a Desert Survival Situation [37] task. Participants that joined the immersive VR platform using a head-mounted device (HMD) felt more socially and more spatially present than participants that joined *Mozilla Hubs* using a standard desktop web browser. When using *Zoom*, participants had a lower sense of social presence compared to *Mozilla Hubs* but felt more spatially present compared to the participants that accessed the VR meeting platform in the desktop setting. When evaluated on their effectiveness for learning, Ryu and Kim [63] showed that students perceived higher learning effects when being able to switch between the immersive VR and the conventional desktop mode. This was also shown in Yoshimura et al.'s work [80], where the authors further report a higher level of presence for students using an HMD to access the *Mozilla Hubs* environment [81]. Students

pointed out that better tracking and a more lively animation of the teacher's avatar would improve the teacher's effectiveness in VR.

The importance of body language was studied by Kurzweg et al. [34, 35]. They concluded that body language enriches virtual meetings and provides cues about users' communication willingness. Further, applying visual transformations that augment social behaviors in VR can significantly increase social presence and influence user behavior [62].

The commercial virtual meeting environment *Horizon Worlds* [45] currently embodies users joining the environment with HMDs via abstracted human-like avatars. The arms are controlled through users' tracked hand positions. The head and facial expressions are animated based on users' voice input and motions. Users that join the environment using the web browser from their computer are displayed in the environment on an embedded 2D screen that shows the captured webcam feed in a video chat setup.

Numerous other commercial virtual meeting platforms exist that support HMDs as well as desktop devices. These platforms include *Frame* [21], *Spatial* [71], *Arthur* [16], *Cluster* [13], *MootUp* [27], and *The Wild* [5]. Typically, desktop users are represented as 3D avatars that they can control using the WASD key setup and the mouse to change their viewing direction with a first-person view (e.g., *Spatial*). Some platforms allow webcam streams to be shown next to the avatar (e.g., *Frame*) or in place of it (e.g., *MootUp*). However, unlike ViGather, these platforms do not translate desktop users' arm and head motions into corresponding movements of their avatar.

Similar to our research is work that uses the front-facing camera of a screen device to detect a user's facial expression, gaze direction, and head rotation in order to animate a virtual avatar. Hart et al. proposed such a system to animate an avatar in cooperative social VR games [22]. To enable non-verbal cues in virtual meetings, the same authors also introduced a hybrid virtual chat system [23]. They found a strong preference for avatars with facial expressions compared to those without. However, they did not find an effect on social presence. Our system and evaluation complement their research, as we focus on enabling non-verbal communication in VR through natural gestures by mapping a user's whole upper body, including hand and fingers, to their avatar.

Most similar to our work, is the system introduced by Woodworth et al. [78], which uses head tracking, eye tracking, and hand tracking to redirect natural human motion onto an avatar. This enables the avatar to perform the most important teaching-related gestures in VR. The authors focused on the naturalness of motion and found that their system was perceived as closer to humanity than mouse-controlled avatars. In contrast, we investigate the effect of avatars animated from screen-device users in the context of VR meetings. Specifically, we were interested in how such a system affects users' feelings of physical and social presence. Our system is also capable of tracking users' whole upper bodies in addition to their hands.

3 SYNTHESIZING SPATIALLY PRESENT AVATARS FOR SCREEN DEVICES

In this section, we describe our avatar synthesis pipeline that allows screen-device users on desktop and mobile platforms to participate in immersive teleconferencing meetings where participants are seated around a virtual table. For this, we create upper-body avatar representations based on users' voice input as well as their 3D body poses estimated from the monocular video streams of their devices' front-facing cameras. In addition, we translate non-verbal cues in the form of eye contact to avatar animations that augment the spatial presence of screen-device users within the virtual meeting room. We implement our pipeline with Python and the Unity Engine.

3.1 Estimation of 3D body-pose motion from monocular video streams

Our 3D-body-pose estimation pipeline receives the video stream of the device's front-facing camera at around 30 fps as input. To standardize the video streams across different camera models, we resize

all input images to 640 × 360 pixels using Unity Textures. The 3D-body-pose estimation module builds on *MediaPipe* [2], a customizable machine learning framework for live and streaming media. For each image of the stream, we use MediaPipe's *Pose* API that integrates a pose detection [6] and pose landmark model [79] to extract 468 facial 3D landmarks, 33 full-body 3D landmarks, and 21 3D landmarks for each hand visible inside the image. MediaPipe offers the advantage of providing the body pose in 3D world coordinates, as opposed to other frameworks that only estimate the 2D pose within image coordinates. Furthermore, unlike Apple's ARKit, MediaPipe is not restricted solely to mobile devices but can be utilized across a wide range of devices.

In case either of the hands is outside the field of view of the camera, ViGather complements MediaPipe's tracking cues to ensure continuous behavior and prevent interrupted avatar motions. We first keep the user's arms and hands at the last tracked pose when they leave the field of view. When a timeout of one second has elapsed, ViGather animates the user's hands to assume a 'natural' resting pose, smoothly placing them in front of the user's belly on top of the table (see Algorithm 1). This resembles how users interact with virtual interfaces in situated settings, using passive affordances to reduce strain and fatigue [12].

For postprocessing, we apply a 1€ [9] filter in order to obtain a temporally smoother pose motion while optimizing for responsiveness. The filter is applied individually to the tracked hands and the head and configured with an intercept frequency of 15 Hz and a slope value of 0.3. It operates on the position and orientation quaternion representation of each joint, maintaining a record of the previous position and orientation values, along with their derivatives.

Algorithm 1 Transition animation to resting pose due to missing hand tracking cues.
t^{clk} is the current time of the system clock in seconds. p_j^{trk} is the jth joint's 3D position estimated by the MediaPipe framework when the jth joint is visible in the image. t_j^{vis} is the time when the jth joint was last visible in the video stream and p_j^{vis} is the joint's respective tracked position. β is a weighting factor that ensures a smooth transition to the resting pose.

for j in [left hand, right hand] **do**
  if j is visible **then**
    $t_j^{vis} = t^{clk}$
    $p_j^{vis} = p_j^{trk}$
  else
    $\beta = \min(t^{clk} - t_j^{vis}, 1)$
    $p = \beta p_j^{res} + (1 - \beta) p_j^{vis}$
  end if
end for

3.2 Generation of spatially present avatar representations from 3D motion and audio

Avatar synthesis. We transform the 3D skeletons to 3D avatars using Meta's *Avatar SDK* [44]. We opted for the Meta Avatar SDK as it provides an avatar representation that aligns with Meta Horizon and facilitates integration in Unity. Moreover, it offers features like lip motion and blinking synthesis. The Avatar SDK's inverse kinematics (IK) module synthesizes upper-body avatar representations based on the 3D pose and 3D orientation of a user's head, wrist and finger joints in real-time. While the MediaPipe pose module only predicts the 3D position of the individual landmarks, the IK module also requires the 3D orientation of the head and the two hands as input. We estimate the head orientation with reference to a canonical calibration pose where the user faces the center of the screen during a calibration step shortly after the launch of the application. We perform

Fig. 2. (a) View of *User A* using the VR client. User A observes the virtual meeting room including User B through stereoscopic images displayed by their VR headset. (b) View of *User B* participating via the screen-device client. User B observes the virtual meeting in a first-person view on their device's 2D screen. The screen also contains a window showing a self-view of their avatar.

Procrustes analysis to obtain the rotation matrix for the head that minimizes the pointwise distance between the facial landmarks of the canonical and currently tracked 3D body pose. We compute the orientation of each wrist with respect to the canonical head orientation via three key points of the hand palm as well as the finger angles using the cosine law. Using the Avatar SDK we also animate the avatar's lip motion from the input audio stream capturing the user's voice.

Motion augmentation. A screen-device user joining the virtual meeting views the meeting through a 2D screen of limited size that usually results in a significant down-scaling of the meeting environment including all participants (see Figure 2b and Figure 4). When conversing with other participants, the user will likely adjust their head orientation to create eye contact depending on the participant they are currently addressing. However, the amount of head tilting needed to face a specific participant is significantly smaller on the 2D screen than what would be necessary for the avatar that is directly seated within the virtual meeting. Thus, only applying the tracked head tilt to a user's synthesized avatar would obscure this non-verbal cue. To create and maintain eye contact within the virtual meeting, we augment the user's motion with an additional body rotation. We estimate the gaze direction of the user based on its head tilting direction facing the screen. Based on the horizontal rotation angle of the user's head, we infer the participant the user is currently addressing. A participant sitting to the left of the user's avatar at the virtual table is also shown on the left side of the screen. We then rotate the avatar's body towards the respective participant's avatar within the virtual meeting.

4 IMPLEMENTATION OF A MULTI-USER CROSS-DEVICE VIRTUAL MEETING ROOM

Using the avatar synthesis pipeline screen-device users can participate in immersive teleconferencing meetings as avatars. Our ViGather teleconferencing system connects participants across VR and screen-device platforms and represents them equally in a joint environment.

ViGather implements a server-client architecture. Depending on the participant's respective device, the participant joins the virtual meeting either through a *VR headset* or a *screen-device* client. Each client runs a local but across-clients identical virtual environment, which is synchronized in real-time across distributed computing platforms (Windows/Android) and hybrid network conditions (WiFi/Ethernet). Central to our implementation are a real-time *relay server* that broadcasts audio as well as motion data, and a *motion server* that handles the computationally intensive 3D body pose estimation for all screen-device users. Collectively, ViGather supports meetings with varying amounts of VR-headset- and screen-device users and reduces computational requirements on participants' device hardware.

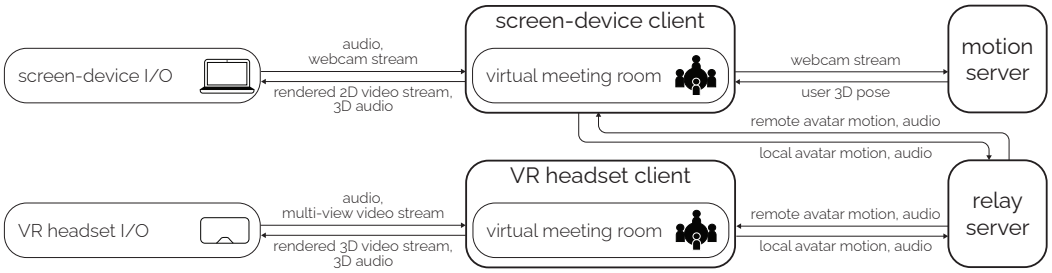

Fig. 3. Overview of the implementation of the ViGather teleconferencing system. Depending on the user's respective device, participants join the virtual meeting via the screen-device or VR headset client. Each client renders a separate local environment of the meeting room. The VR headset client receives the head and hand poses through the Oculus SDK, tracked by the VR headset's four built-in cameras. The screen-device client receives a monocular video stream from the front-facing camera (i.e., webcam) as well as the audio from the microphone. It forwards the video stream to the motion server, which returns the head and hand poses as well as the body orientation. The clients exchange the audio and motion data with each other through a relay server. They then animate the avatars within their virtual meeting room using Meta's Avatar SDK.

Figure 3 shows an overview of our ViGather teleconferencing system. Each client separately records the user's verbal communication and non-verbal behavioral cues using the device's respective sensors: a monocular video stream from the front-facing camera and audio from a microphone in the case of screen-device clients and the body poses reconstructed via a multi-view head-mounted camera setup and audio from the headset's integrated microphones in the case of VR clients. The client then manages the synthesis of the local participant's avatar (i.e., the participant using the device where the client is running on) based on the device's respective sensor stream. It shares the information needed to synthesize the local avatar consisting of the 3D motion and recorded audio data with all other meeting participants via the relay server. In turn, it receives the data to synthesize the other participants' avatars. Each client of our system then renders all meeting participants' avatars in a separate synchronized local virtual environment.

4.1 Relay server

ViGather's relay server facilitates low-latency data exchange across all clients. The server runs in Python 3.8 on an 8-core Intel Core i7-9700K CPU at 3.60 GHz using a static IP address and two ports. Our server communicates with each client through a TCP/IP socket connection. To achieve low-latency, we use two sockets and multi-threaded processes to handle the audio and motion streams separately for each pair of server-client connections. The server constantly listens at two ports for new client connections, which a client application requests after its launch. The server receives audio and 3D body-pose motion data streams from each client and broadcasts them to all other active clients without modifying any of the packets' header or payload. The server also keeps track of the number of successfully connected clients, detects disconnected clients based on the tracked forwarding times and removes them from the broadcasting queue if necessary.

4.2 Motion server

The motion server handles the estimation of the 3D-body-pose motion from the monocular video stream it receives as input from a respective screen-device client. It is implemented in Python 3.8 and runs on a separate machine with an Intel Core i7-9700K CPU and an Nvidia *Quadro RTX 6000* GPU. It processes the individual images of the video stream with the MediaPipe Pose API to

get 3D positions of the head and hand landmarks, and performs the orientation estimation of the individual joints (see section 3). Thus, it takes over a major part of the computational load from the edge devices and due to its powerful back-end GPU reduces the overall latency of the system. For each frame, the motion server returns the 6D pose of the head and the hand as well as the angles of each hand's 15 finger joints to the screen-device client.

4.3 Screen-device client

We implemented the screen-device client in Unity 2020 3.14f1 using C# and compiled it as a Windows executable that runs on various screen devices including Windows-based desktops, laptops and tablets. After launch, the client establishes a connection with the relay server and the motion server. It retrieves a continuous video and audio stream from the front-facing camera (e.g., integrated into display bezel or attached as a separate webcamera to a monitor) and microphone respectively. The client resizes each image frame and compresses it using the JPG binary format. It then forwards the processed frame as a network packet through a TCP/IP connection to the motion server and receives the corresponding pose data in return. Based on the estimated head and hand poses and the recorded voice input, the screen-device client then animates the local participant's avatar using the Meta Avatar SDK. The client further rotates the avatar's body depending on the currently approximated gaze direction.

The screen-device client only forwards the motion and audio data to the relay server but not the raw video stream and receives the corresponding information from all other VR headset and screen-device clients in return. Each client sends the audio and motion data through two independent socket channels at 50 fps. We designed a custom data packet structure as well as an encoding and decoding protocol to transmit the data efficiently. The audio data is captured at 44.1 kHz and split into small chunks. The motion stream contains the pose information for a given frame, which includes the gaze direction, lip pose, facial pose, hand and finger poses, as well as the poses of the other upper body-skeleton joints in the avatar's local coordinate system. It also includes the global position and orientation of the avatar in the global coordinate system. To reduce the network bandwidth overhead due to retransmissions, we designed a lightweight protocol that appends a metadata header of 32 bytes to the raw data package to identify the sender, recording time, size, and type of the packet. Appendix A.1 contains a more detailed description of the packet design.

The client finally evenly positions the animated avatars around a circular table by adjusting their global position and orientation so that the participants are facing each other. The vertical offsets of the avatars are adjusted so that the hands are placed on the table surface during the resting pose.

Rendered scene for screen-device participants. The virtual meeting takes place around a circular table placed inside a living room with static furniture elements for decoration. A speaker is placed on top of the table which broadcasts the voice of the remote moderator during the study. Each participant's avatar is synthesized during runtime. The participants joining the meeting using their screen devices have a first-person view into the scene through a static virtual camera that is positioned between the eyes of their individual avatar in the canonical calibration pose. We adjusted the field of view of the virtual camera so that all participants within the scene are fully visible, enhancing the participants' sense of shared presence with other participants. In addition, at the top of the screen there is a smaller window providing a front-facing view (or self-view) of the participant's avatar through a simulated virtual webcam to enhance the participant's sense of self-presence (see Figure 2b).

4.4 VR headset client

We developed the VR headset client as an Android app to run on a Meta Quest 2 headset using Unity 3D Engine. The client communicates with the relay server through a TCP/IP connection over the local WiFi network. The client application establishes a connection with the relay server, renders the virtual environment, and dynamically generates an avatar for each active participant.

Avatar synthesis. The VR headset client also generates avatars via Meta's Avatar SDK. Contrary to the screen-device client, it does not use the 3D body poses estimated by the motion server, but directly receives information about the 3D head and hand poses of the participant wearing the VR headset in real-time through the Oculus SDK. These are tracked by the device using four built-in cameras. For the lip animation, it processes the voice input captured by the headset's integrated microphones with the Avatar SDK.

Analogous to the screen-device client, the VR headset client shares the 3D motion and audio data via the relay server with all other active clients and uses the received motion and audio data of the other participants to animate all participants' avatars within the scene.

4.4.1 Rendered scene for VR headset participants. The client runs a real-time simulation of the same scene as described for the screen-device client. Participants' avatars are seated together around a table (see Figure 3a). Participants observe the meeting environment including the other participants through dynamic virtual cameras placed at the left and right eye of their avatar. The stereoscopic images are rendered by the VR headset at a rate of 50 fps.

4.5 System latency

We conducted an analysis of our system to determine its latency, throughput, and update framerate. In Table 1, we present the average values obtained under stable running conditions over a one-minute duration. The motion server, relay server, and screen-device client machines are all connected to the same network via Ethernet. The VR headset client connects to the network via WiFi.

The VR headset client performed audio forwarding at a framerate of 36 frames per second (fps) with a throughput of 0.176 MB/s. It recorded the user's avatar motion at a framerate of 55 fps, resulting in a throughput of 0.228 MB/s. The round-trip latency for the VR headset client to send and receive packets from the relay server was around 20 ms.

The screen-device client operated at a framerate of 37 fps for audio, achieving a throughput of 0.176 MB/s. It recorded and transmitted the avatar's motion at a framerate of 45 fps, with a throughput of 0.189 MB/s. The round-trip latency for sending and receiving packets from the relay server was measured around 3 ms. The latency encompassing the transmission of a video frame

Table 1. This table presents the latency, throughput, and update framerate for the screen-device and VR headset clients. We measured the refresh and forward throughput and framerate for motion and audio data for each client. Additionally, we determined the round-trip latency for audio and motion packet transmission between each client and the relay server. We also determined the framerate of the motion server to estimate body poses from the video stream and assessed round-trip latency for video frames to be processed and transmitted between the screen-device client and motion server.

		latency [ms]	throughput [MB/s]	frame rate [fps]
screen-device client	motion	3	0.189	45
	audio	2	0.176	37
motion server	video to motion	59	0.001	17
VR headset client	motion	21	0.228	55
	audio	20	0.176	36

and receipt of the corresponding motion data from the motion server was 59 ms. As for the motion server, it operated at a framerate of 17 fps, with a throughput of 0.001 MB/s.

5 STUDY 1 — VIGATHER VS. HORIZON

The purpose of Study 1 was to determine if our system successfully integrates screen-device users into the VR meeting experience. In addition, we intend to understand how our proposed system affects the meeting experience for VR headset and screen-device users compared to Meta Horizon. We evaluated ViGather with participants in groups of four, who worked on a collaborative task guided by an experimenter.

5.1 Task

In each condition, participants were asked to brainstorm new ideas for a movie and a company or organization. Thus, we designed a task that is inspired by the 635 Method [64] and The Desert Survival Problem [37].

In the first phase, four ideas with respect to a topic were generated. Thus, the experimenter asked each of the four participants a specific question (e.g., for the movie: 1) what is the message of the movie?, 2) how is the message conveyed?, 3) who are the main characters?, 4) Is it a single movie or a series?)[1]. This process was repeated four times such that each participant answered each of the questions once. The result of this phase was four ideas on the respective topic (e.g., four ideas for a movie). The first phase ensured that each participant was sufficiently familiar with how to express themselves in the VR setting, using their respective device.

In the second phase, the experimenter asked the participants to rank the ideas according to their expected impact. Thus, they first defined impact with respect to the topic and then agreed upon the

[1]Questions for the organization or company: 1) What is the mission? 2) How should the mission be realized? 3) Who should be hired? 4) How should the organization be financed?

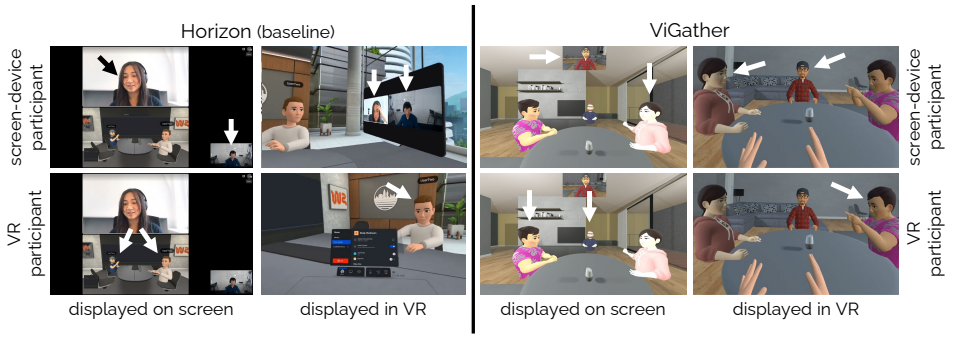

Fig. 4. Illustration of how participants saw other participants in their respective condition in Study 1. Left: In Horizon, participants with VR headsets join the virtual meeting room as avatars. The front-facing camera feed of participants that join with their screen device (e.g., a laptop) is shown on a virtual display within the meeting room. Participants with screen devices have a first-person view into the meeting room from the position of the virtual display and see the camera feed of other screen-device participants in separate windows. They also see their own camera feed in a smaller window at the right lower corner. Right: In ViGather all participants are represented as avatars within the VR meeting, which removes the difference in the representation of participants joining with different devices. Participants with screen devices have a first-person view into the virtual meeting room. Their screen also contains a smaller window showing a self-view of their avatar.

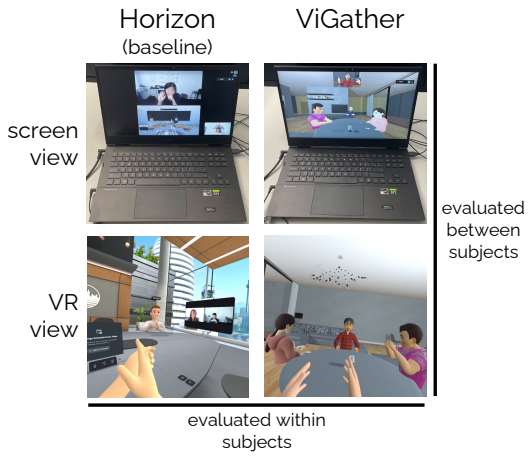

Fig. 5. Conditions for Study 1. Top row: Screen-device participants saw the video conference on a screen as a mix of VR avatars and video feeds in Horizon and as a coherent 3D scene in ViGather. Bottom row: VR participants experienced an immersive video conference with screen-device participants represented through a floating video feed in Horizon and represented through avatars in ViGather.

idea that best fulfils their definition. This phase ensured that participants had to actively discuss and engage with each other.

5.2 Conditions

Our evaluation compared participants' experience between two teleconferencing applications: Meta Horizon and ViGather. All participants experienced both applications during the study (within-subject) and solved an instance of the brainstorming task. Two out of the four participants joined the teleconferencing meetings via a VR headset, whereas the other two were using a screen device (between-subject). This resulted in a 2×2 mixed study design with a total of 4 conditions: 2 meeting applications, 2 device types. Figure 5 shows an overview of the conditions. Figure 4 illustrates how participants saw other participants through the display of their respective devices.

5.3 Apparatus

Participants who joined via a VR headset wore a Meta Quest 2 during the study. The screen-device participants used either a laptop (16", 2560×1440 resolution) or a PC (24", 2560×1440 resolution) to participate in the meeting.

5.4 Procedure

Before the start of the study, each participant was led to a separate room where they signed the consent form and completed the pre-questionnaire. They then solved the brainstorming task with their respective device (headset or screen-device) in the two different teleconferencing applications. The order of teleconferencing applications was fully counterbalanced. The attribution of the four brainstorming task questions to the four participants was counterbalanced using a Latin-square. There were four ideas generated for each topic and each participant had to answer one of the four questions once per idea, so that they answered each of the four questions once.

After each brainstorming task, participants completed a post-questionnaire. We used the Multimodal Presence Scale (MPS) [41] to assess physical presence, social presence, and embodiment and

an adapted version of the Temple Presence Inventory (TPI) [38] to assess spatial presence, social presence of the actors, passive social presence, active social presence, engagement, social richness, social realism, and perceptual realism (we removed the multisensory items related to temperature, haptic feedback, and smell, and the item related to "being there" as it was already present in the MPS). In addition, after the ViGather conditions, participants also rated the perceived naturalness of the other three avatars (each) on a custom 7-points Likert-scale. We did not include these questions after the Horizon conditions as there was no avatar for non-VR participants. Participants reported their overall impression in an open-feedback text-box after each MEETING APPLICATION. Overall, a session took between one-and-a-half and two hours.

5.5 Participants

We recruited 16 participants (5 female, 11 male, ages 20–38, M=24.4, SD=4.2). All participants were students and staff members at a local institution for higher education. Participants received $20 as a gratuity for their time. Some participants knew each other prior to the study, however, each session contained strangers to them.

Participants reported frequency of VR usage, gaming and video conferencing on a 4-point Likert scale (1–every day, 2–several times a week, 3–several times a month, 4–less). We also asked them to rank their level of introversion/extroversion on a 6-point Likert item (from 1–extremely introverted to 6–extremely extroverted). Participants' median responses were VR freq. = 4, gaming freq. = 3, video-conferencing freq. = 2, and level of introversion/extroversion = 3.

5.6 Results

We analyzed the effect of the MEETING APPLICATION across the different DEVICE TYPES on physical, spatial and social presence as well as on embodiment, engagement, richness and realism. For significance testing, we performed an ANOVA as the data was normally distributed (all Shapiro-Wilk $p > .05$) and follow the assumption on equal variance between groups (all Levene's $p > .05$). For each variable, the participant was considered as a random factor, the MEETING APPLICATION as a within-subject factor, and the DEVICE TYPE as a between-subject factor. Pairwise comparisons were performed using t-tests with Bonferroni-adjusted p-values. Figure 6 and 7 show an overview of the results of the TPI and MPS, respectively. Effects of DEVICE TYPE are out of the scope of this paper.

Physical, spatial & self-presence : To analyze physical, spatial, and self-presence, we evaluated the respective dimensions of MPS and TPI. For physical presence (MPS), an ANOVA revealed a statistically main effect of MEETING APPLICATION with ViGather eliciting a higher feeling of physical presence than Horizon [$F_{1,14} = 7.27$, $p = .02$, $\eta_p^2 = .08$]. We found a significant interaction effect between the MEETING APPLICATION and the DEVICE TYPE on physical presence [$F_{1,14} = 14.89$, $p = .002$, $\eta_p^2 = .15$].

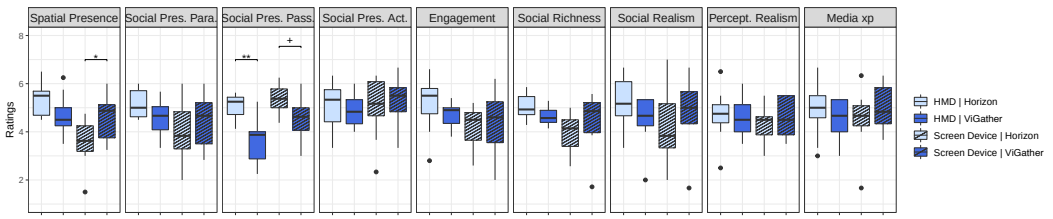

Fig. 6. Results of the Temple Presence Inventory for STUDY 1 - boxplots are shown per dimension for each MEETING APPLICATION × DEVICE TYPE. Significances are indicated per pairs of DEVICE TYPE (paired t-tests with Bonferroni-adjusted p-values): + $p < .1$, * $p < .05$, ** $p < .01$.

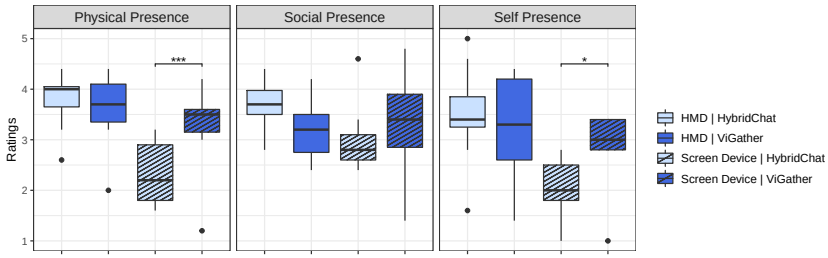

Fig. 7. Results of the Multimodal Presence Scale for Study 1 - boxplots are shown per dimension for each meeting application × device type. Significances are indicated per pairs of device type (paired t-tests with Bonferroni-adjusted p-values): $^{*}p < .05$, $^{***}p < .001$.

The pairwise comparison has shown that ViGather caused a higher sense of physical presence over Horizon in the screen-device condition ($p < .001$). No significant difference between the meeting applications was found in VR ($p = .425$). For spatial presence (TPI), we did not find any main effect of meeting application [$F_{1,14} = .42$, $p = .526$, $\eta_p^2 = .02$]. However, we found a significant interaction effect between the meeting application and the device type [$F_{1,14} = 6.76$, $p = .021$, $\eta_p^2 = .21$]. In the screen-device condition, participants felt a higher sense of spatial presence with ViGather compared to Horizon ($p = .038$). No significant difference between the meeting applications was found in VR ($p = .19$). For self-presence (MPS), we did not find any main effect of meeting application [$F_{1,14} = 1.40$, $p = .19$, $\eta_p^2 = .03$]. However, the interaction effect between the meeting application and the device type was significant [$F_{1,14} = 6.35$, $p = .026$, $\eta_p^2 = .08$]. For screen-device, participants perceived ViGather to cause a higher sense of self-presence compared to Horizon ($p = .019$). The difference between the meeting applications was not significant in VR ($p = .42$).

Social presence, richness & realism: To evaluate the effect of the meeting application on the social aspects of a VR meeting, we analyzed the respective dimensions of the MPS and TPI. The ANOVA revealed no significant differences between the conditions on the social presence dimension of the MPS (meeting application: [$F_{1,14} = .21$, $p = .652$, $\eta_p^2 = .01$]; meeting application × device type: [$F_{1,14} = 2.74$, $p = .120$, $\eta_p^2 = .07$]). The TPI investigates several aspects of social presence. No significant differences were found for active interpersonal social presence (meeting application: [$F_{1,14} = .03$, $p = .861$, $\eta_p^2 = .001$]; meeting application × device type: [$F_{1,14} = .43$, $p = .52$, $\eta_p^2 = .01$]) nor for the aspect of parasocial interaction (meeting application: [$F_{1,14} = .07$, $p = .788$, $\eta_p^2 = .01$]; meeting application × device type: [$F_{1,14} = 2.40$, $p = .143$, $\eta_p^2 = .07$]). However, statistically significant differences were found between the meeting applications in terms of passive interpersonal social presence [$F_{1,14} = 17.66$, $p < .001$, $\eta_p^2 = .36$] where Horizon created a higher interpersonal social presence compared to ViGather, i.e., participants perceived others' expressions (e.g., facial expression, clothes, body language) better with Horizon than ViGather. The interaction between meeting application and device type was not significant for this factor [$F_{1,14} = 1.51$, $p = .239$, $\eta_p^2 = .05$]. The same held for the dimensions of social richness (meeting application: [$F_{1,14} = .01$, $p = .918$, $\eta_p^2 = .001$]; meeting application × device type: [$F_{1,14} = 2.95$, $p = .108$, $\eta_p^2 = .08$]) and social realism (applications: [$F_{1,14} = .14$, $p = .712$, $\eta_p^2 = .003$]; meeting application × device type: [$F_{1,14} = 4.04$, $p = .064$, $\eta_p^2 = .07$]).

Engagement, perceptual realism & experience: To assess differences in engagement, perceptual realism, and media experience, we analyzed the respective dimensions of the TPI. For engagement, the ANOVA did not reveal significant differences between the meeting application [$F_{1,14} = .06$, $p = .807$, $\eta_p^2 = .002$] as well as for their interactions with the device type [$F_{1,14} = .93$, $p = .352$, $\eta_p^2 = .03$].

No significant differences were found for perceptual realism between the MEETING APPLICATIONS [$F_{1,14} = .10$, $p = .757$, $\eta_p^2 = .003$] and their interactions with the DEVICE TYPE [$F_{1,14} = .28$, $p = .607$, $\eta_p^2 = .01$]. The same held for media experience (MEETING APPLICATION: [$F_{1,14} = .07$, $p = .789$, $\eta_p^2 = .002$]; MEETING APPLICATION × DEVICE TYPE: [$F_{1,14} = 2.24$, $p = .156$, $\eta_p^2 = .05$]).

Naturalness: After each ViGather condition, we asked participants to rate the perceived naturalness of the other three avatars on a custom 7-point Likert scale. We analyzed the effect of the "OBSERVED AVATAR'S MEDIUM" (VR or screen-device, within-subject) and "OBSERVERS' MEDIUM" (VR or screen-device, between-subject) on the naturalness score. The Aligned Rank Transform (ART) ANOVA [77] did not show any significant difference on the naturalness score (OBSERVED AVATAR'S MEDIUM: [$F_{1,14} = 3.64$, $p = .066$, $\eta_p^2 = .11$], OBSERVER'S MEDIUM: [$F_{1,14} = 3.83$, $p = .071$, $\eta_p^2 = .21$], OBSERVED AVATAR'S MEDIUM × OBSERVERS' MEDIUM: [$F_{1,14} = 0$, $p = .981$, $\eta_p^2 = .001$]); i.e., with ViGather, participants did not perceive screen-device avatars to be significantly less natural than VR avatars, whether they were themselves in VR or not.

Perceived Qualitative Difference: Participants reported their overall media experience and perceived differences between applications after each MEETING APPLICATION condition. Most participants enjoyed VR teleconferencing with both MEETING APPLICATION (Horizon: P2 "*The experience was joyful and engaging.*"; ViGather: P8 "*It was a fun experience being inside VR and participating as I would actually do in real life.*"). In opposition, one participant mentioned that they still preferred video-conferencing: P12 "*I find it more comfortable and natural joining with a real video.*".

On Horizon, many participants disliked having separate layouts with videos for screen-device users and avatars for VR users: P4 "*I found it weird/confusing having some people in VR and some video-based. It felt like we were not part of the same group.*", P8 "*I thought it could be more engaging if everyone was an avatar. It would allow more expressivity.*", P5 "*Having mixed reality domains and separated layouts made it worse.*"; P9 "*It felt like I was in two separate meetings due to the difference in presentation styles* (between attendees)."; except for one participant: P6 "*Very interesting to see people on a video call in VR.*". One participant suggested to move the video display closer to the avatars: P9 "*Maybe the real video feed could be inside the area where the avatars are sitting*". They appreciated ViGather setting in comparison as it improved non-verbal social interactions: P9 "*I liked that I could look at the other avatars and they could see me looking at them and vice-versa. I could see if others looked at me.*", P4 "*I like the idea very much. It makes it feel a lot more social* (compared to Horizon).", P13 "*I know one of the VR users in real life as well, so seeing his usual responses in VR was very interesting.*".

However, the participants also reported that Horizon better communicates peroples's expressions: P13 "*It was a little better in expressing people's reactions* (compared to ViGather).", P15 "*The screen with the video of another participant felt like I was looking at a screen. This gave a sense of reality.*".

Only few participants perceived a difference in visual fidelity between conditions: P12 "*I feel like the VR simulated person are more vivid* (Horizon) *compared to the previous simulation* (ViGather)."

5.7 Discussion

Our evaluation compared two applications for VR meetings: Meta's Horizon and ViGather. Participants joined the virtual meeting using either a VR headset, a laptop, or a desktop PC.

The results show that ViGather's immersive meeting experience created a sense of active social presence that was not significantly different to the experience in Horizon. Participants felt no difference between applications in terms of how they captured the dynamics of social interaction when they were speaking with others (active interpersonal). Observing the active expressions of other users (parasocial interaction) showed no difference either. The experience for VR users in ViGather

further showed no significant difference to Horizon according to participants' ratings on physical, spatial, and self-presence as well as social richness and realism, engagement, perceptual realism, naturalness, and media experience. In addition, the advantage of a webcam-based representation inside a 3D meeting can be seen in the passive interpersonal aspects of social presence. Participants felt that their own as well as others' facial, gestural, and verbal expressions were better captured by Horizon than by ViGather currently. Participants' qualitative feedback further supports this result.

However, in ViGather, screen-device users felt a significantly higher sense of physical, spatial, and self-presence within the virtual meeting compared to Horizon. This finding is also reflected in participants' positive comments with respect to ViGather.

Yet, it is not clear which design differences between the two applications contribute to these findings. Upon closer analysis, we identified three main design aspects of the applications that could influence users' experience of spatial–, physical–, and self-presence:

(1) Distance between screen– and VR-users: in ViGather, screen-device users are situated on the same table as VR users, while in Horizon, they are positioned on a screen that is several meters away from this table.
(2) Screen-device users' experience of the VR environment: in ViGather, a screen-device user has a game-like full-screen view of the virtual room and an in-frame view of their own avatar. They see themselves as part of the virtual environment. In contrast, in Horizon, the virtual environment is shown in one section of the screen, and the other sections show the video streams of the self-view and the other screen-device user. The screen-device user sees a webcam stream in their own space instead of their avatar in VR.
(3) Coherent representation of screen–device and VR–users: ViGather provides a coherent representation for screen and VR users by giving them an equivalent avatar, whereas Horizon represents screen-device users with their webcam video.

We are particularly interested in determining how the coherent representation of screen–device and VR users affects the perceived user experience (3). This coherent representation is only possible through ViGather's capability of reconstructing screen-device users as virtual avatars.

Therefore, we conducted a second study with an alternative baseline that resembled Horizon but did not differ from ViGather with respect to (1) and (2). This eliminated confounding factors and allowed us to focus on the effect of the avatar reconstruction.

6 STUDY 2 — VIGATHER VS. HYBRIDCHAT

The purpose of our second study was to determine the effect of screen-device users' avatar reconstruction on participants' VR meeting experience. We re-used the task, procedure, and apparatus from Study 1, but we altered the baseline condition to replace Horizon with a custom application, dubbed *HybridChat*.

HybridChat isolates the effect of the coherent avatar representation by ensuring that the distance between screen-device users and VR users is the same as in ViGather. In addition, HybridChat's screen-device UI provides the same game-like first-person full-screen view of the virtual room as ViGather. Also similar, it displays the webcam stream that shows the self-image of a screen-device user as an in-frame view. Figure 8 illustrates how participants saw other participants in the HybridChat application.

6.1 Conditions

The second study compared participants' experience between HybridChat and ViGather, again using groups of four simultaneous participants. Participants experienced both applications during the study, with the application as a within-subject factor. They solved the same instances of the

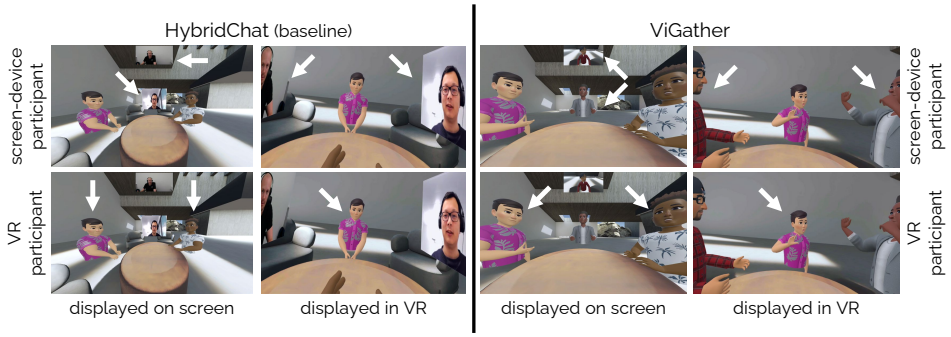

Fig. 8. Illustration of how participants saw other participants in their respective condition in Study 2. Left: In HybridChat, participants with VR headsets join the virtual meeting room as avatars. The front-facing camera feed of participants that join with their screen device (e.g., a laptop) are shown on a virtual display within the meeting room. Participants with screen devices have a first-person view into the meeting room from the position of the virtual display. The camera feed of other screen-device participants is also co-located in the virtual environment. They also see their own camera feed in a smaller window on the upper center of the screen. Right: In ViGather all participants are represented as avatars within the VR meeting.

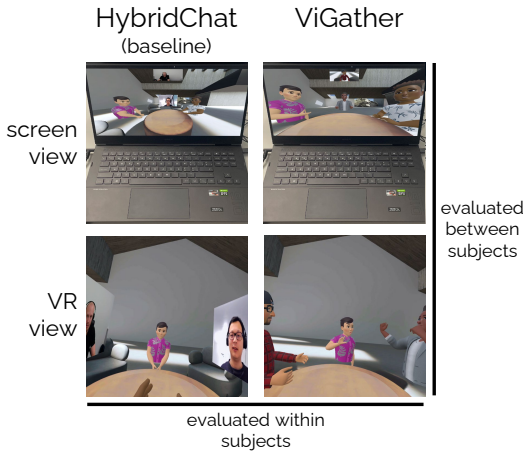

Fig. 9. Conditions for Study 2. Top row: Screen-device participants had a first-person full-screen view of the virtual room in both conditions. In HybridChat, they see themselves and other screen-device participant through their respective webcam stream, while in ViGather, they were displayed as 3D avatars. Bottom row: VR participants experienced an immersive video conference with screen-device participants shown on video feeds placed around the meeting table in HybridChat and represented through avatars in ViGather.

brainstorming task as in Study 1. The study had the interface as a between-subject factor, such that two of the four participants joined the teleconferencing meetings via a VR headset, whereas the other two used a screen device. This resulted in a 2×2 mixed study design with a total of 4 conditions: 2 meeting applications, 2 device types. Figure 9 shows an overview of the conditions.

6.2 Participants

We recruited 16 new participants (8 female, 8 male, ages 20–29, M=24.9, SD=2.3). All participants were students and staff members at a local institution for higher education. Participants received

$20 as a gratuity for their time. While some participants knew each other prior to the study, we assigned groups so that each session contained conversation members they were unfamiliar with.

Participants answered the same pre-questionnaire as in the first study. Participants' median responses were VR frequency = 4, gaming frequency = 3, video-conferencing frequency = 2, and level of introversion/extroversion = 3. One participant's questionnaire response was an outlier and was excluded from the analysis as the responses were inconsistent and partially contradictory as well as clearly outside the distribution of other participants' responses.

6.3 Results

We analyzed the effect of MEETING APPLICATION across the different DEVICE TYPES on physical, spatial and social presence as well as on embodiment, engagement, richness and realism. For significance testing, we performed an ANOVA as the data was normally distributed (all Shapiro-Wilk $p > .05$) and followed the assumption on equal variance between groups (all Levene's $p > .05$). For each variable, participant was considered as a random factor, MEETING APPLICATION as a within-subject factor, and DEVICE TYPE as a between-subject factor. Pairwise comparisons were performed using t-tests with Bonferroni-adjusted p-values. Figures 10 and 11 show an overview of the results of the TPI and MPS, respectively.

Physical, spatial, & self-presence: To analyze physical, spatial, and self-presence, we evaluated the respective dimensions of MPS and TPI. We found a significant interaction effect between MEETING APPLICATION and DEVICE TYPE on the MPS self-presence [$F_{1,13} = 5.80$, $p = .032$, $\eta_p^2 = .13$]. The pairwise comparison showed that ViGather caused a higher sense of self-presence than HybridChat in the screen-device condition ($p = .009$). No significant difference between the MEETING APPLICATIONS was found in VR ($p = .834$). The analysis revealed no significant differences between the conditions involving MEETING APPLICATION for spatial presence (TPI) and physical presence (MPS).

Social presence, richness, & realism: To evaluate the effect of MEETING APPLICATION on the social aspects of a VR meeting, we analyzed the respective dimensions of the MPS and TPI. We found a significant interaction effect between MEETING APPLICATION and DEVICE TYPE on passive interpersonal social presence [$F_{1,13} = 8.98$, $p = .010$, $\eta_p^2 = .15$]. The pairwise comparison revealed that HybridChat created a higher interpersonal social presence than ViGather in the VR condition, i.e., participants perceived others' expressions (e.g., facial expression, clothes, body language) better with HybridChat than with ViGather in VR ($p = .009$). No significant difference between MEETING APPLICATIONS was found in the desktop condition ($p = .549$). The ANOVA revealed no other significant differences between the conditions on the social metrics.

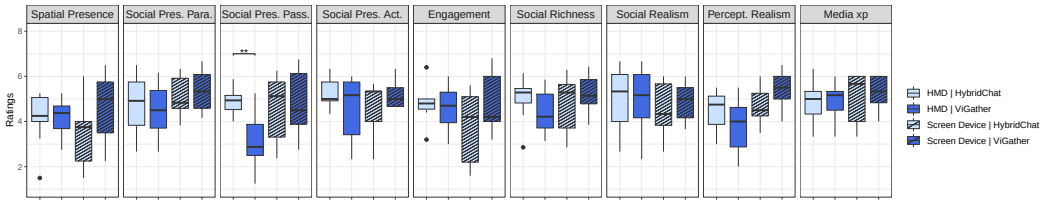

Fig. 10. Results of the Temple Presence Inventory for STUDY 2 - boxplots are shown per dimension for each MEETING APPLICATION × DEVICE TYPE. Significances are indicated per pairs of DEVICE TYPE (paired t-tests with Bonferroni-adjusted p-values): $^{**}p < .01$.

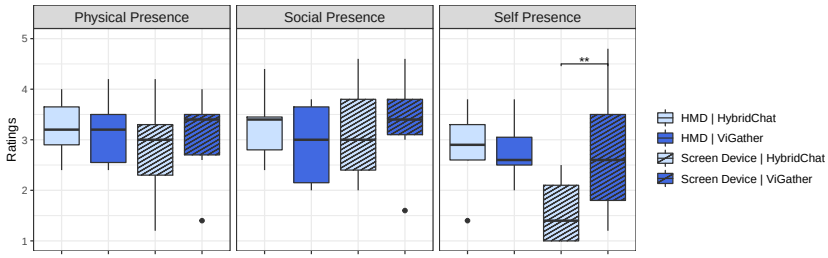

Fig. 11. Results of the Multimodal Presence Scale for Study 2. Boxplots are shown per dimension for each meeting application × device type. Significances are indicated per pairs of device type (paired t-tests with Bonferroni-adjusted p-values): $^{**}p < .01$.

Engagement, perceptual realism & experience: The ANOVA revealed no significant differences between conditions in terms of the engagement, perceptual realism, and media experience dimensions of the TPI.

Naturalness: After each ViGather condition, participants rated the perceived naturalness of the other three avatars on a 7-point Likert scale. We analyzed the effect of the "observed avatar's medium" (VR or screen-device, within-subject) and "observers' medium" (VR or screen-device, between-subject) on the naturalness score. The Aligned Rank Transform (ART) ANOVA [77] showed that ViGather's participants on screen devices perceived the other avatars as more natural than ViGather's participants in VR [$F_{1,13} = 7.32$, $p = .019$, $\eta_p^2 = .38$]. However, the analysis did not show any significant effect of the observed avatar's medium [$F_{1,13} = 0.14$, $p = .710$, $\eta_p^2 = .01$] or interaction effect between the observed avatar's medium and observers' medium [$F_{1,13} = 0.88$, $p = .357$, $\eta_p^2 = .03$] on the naturalness scores; i.e., using ViGather, participants did not perceive screen-device avatars to be significantly less natural than VR avatars.

Perceived Qualitative Difference: We asked participants for their overall media experience and the perceived differences between applications after each meeting application condition. Again, most participants enjoyed the VR meeting experience with both meeting application (HybridChat: P8 "*It is really an interesting experience.*"; ViGather: P4 "*It's an impressive experience.*"; P13 "*It's nicer than a normal meeting*").

Several participants highlighted the importance of seeing facial expressions during a meeting and, hence, appreciated HybridChat (P10 "*I personally liked looking at real people in the VR environment.*") or missed the capability in ViGather (P5 "*I can't see their facial changes.*"). One participant highlighted that, although the virtual experience broke due to the introduction of screens, they felt most connected to the other screen user: P13 "*The video of the user in front completely broke the virtual experience but I felt more connected to her than the others because I could see clearly her facial expressions.*"

6.4 Discussion

Our second evaluation compared two VR meeting experiences that differed in how screen-device users were represented. In HybridChat, screen-device users appeared in the virtual meeting environment as their webcam stream. In contrast, in ViGather, screen-device users joined the virtual meeting as 3D avatars that replicated the motions of their heads and hands captured by a webcam.

Although the view of screen-device users into the virtual meeting was similar for both conditions, their in-frame view of themselves differed. In ViGather, participants saw themselves as their avatar representations, whereas in HybridChat, they saw their actual webcam stream.

In ViGather, screen-device users reported a significantly higher level of self-presence compared to HybridChat. This finding supports ViGather's premise that by mimicking users' body movements through animating their avatars, our system created a stronger sense of embodiment within the VR meeting. Interestingly, there was no significant difference in physical presence or social presence between the two conditions. This suggests that the difference between the Horizon and ViGather conditions we observed in Study 1 is more likely due to the difference between how screen-device participants view the virtual meeting, i.e., either with separate windows for the camera feeds of other screen-device participants (in Horizon) or through a single first-person full-screen view (in ViGather). Our results highlight the potential of first-person full-screen views for future VR meeting environments, where users are placed closer to other participants in order to increase their perception of being present within the meeting.

In Study 2, we observed similar trends between the HybridChat and ViGather conditions for VR participants as we did between the Horizon and ViGather conditions in Study 1. Specifically, VR participants reported a higher passive social presence when interacting with other participants that appeared as a webcam stream. We attribute this effect to the additional and more nuanced expressions that webcam streams capture, whereas avatar representations necessarily rely on the fidelity of computer vision and audio interfaces for detecting and representing body motions and facial expressions, in ViGather's case Google MediaPipe and Meta Avatar SDK. Participants' qualitative comments confirmed this impression, as they mentioned the importance of conveying facial expressions in VR meeting settings for a holistic conversation.

7 LIMITATIONS

While our evaluation showed ViGather's benefits and its promise for future teleconferencing, we outline several limitations of our current implementation as the basis for future research.

Facial expressions. Since the face of ViGather's avatars is animated using only the user's voice as input, the range of facial expressions is limited and may not fully capture or accurately represent the user's current expression, both on screen devices as well as VR devices. For VR users, future implementations of ViGather can benefit from Meta's most recent introduction of the Quest Pro [46], a high-end HMD device that embeds sensors for capturing users' facial expressions and eye gaze in the headset. For screen-device users, MediaPipe is already capable of reconstructing the 3D face mesh and tracking eye gaze from the front-facing camera. Future iterations of our system could directly map these features onto avatar faces by leveraging Meta's most recent Avatar SDK, which was released with Quest Pro. Therefore, both platforms will support avatars with fully animated facial expressions and upper-body motions across screen devices and VR headsets, as more and more vendors introduce facial tracking to their platforms (e.g., VIVE Facial Tracker [15]). This will also enable future evaluation of our system to further investigate the effect of even higher fidelity animations on participants' perceived immersion in VR meetings.

Avatar personalization. While we ensured matching appearances between participants and their avatars in all conditions, ViGather so far does not automatically personalize 3D avatars based on a user's appearance. Future versions of our system could personalize avatars, for example, through neural volumetric representations [10] to instantly reconstruct realistic appearance on smartphones [76]. We also consider analyzing the effect of highly realistic avatars on participants' perceived social presence in VR meetings as an interesting direction for future work.

Non-stationary meeting settings. We evaluated ViGather in a static seated virtual meeting environment. As VR meetings are likely to become more mobile through the wider availability of mobile VR devices, future research could investigate how to better situate avatar-based chats in

such settings. In fully mobile scenarios, we foresee challenges in capturing users' body poses in the context of their current 3D environment and its physical constraints. An interesting future research direction would be the integration of ViGather on mobile devices that already offer advanced capabilities for environmental understanding (e.g., Apple ARKit [3]).

Collaboration tools. While Horizon implements a set of tools (e.g., sketching) that ease collaboration between users, ViGather's current implementation does not support such tools. For this, future research could explore unifying the set of collaboration tools across device types without limiting each device type's inherent advantages [25].

8 CONCLUSION

We have presented ViGather, an immersive teleconferencing system that integrates all participants into a joint 3D meeting experience, independent of the particular device type they join from, be it a traditional screen device (e.g., PCs, laptops, or tablets) or an immersive Virtual Reality headset. ViGather represents all participants as embodied avatars in a shared scene to enable natural behavior during collocated conversations. Participants are not just able to verbally communicate with each other but also make eye contact, turn to each other for directly addressing a conversation partner, and use hand gestures to support their conversations. For VR participants, we leverage the sensing and reconstruction capabilities of today's headsets. For screen-device users, ViGather reconstructs the 3D upper-body pose from the front-facing camera inside the display and estimates eye contact with other participants. Our system then relates these non-verbal cues to 3D avatar animations in the shared scene and broadcasts this to all clients.

In our evaluation study, we compared participants' behavior in virtual meetings between ViGather and Meta Horizon as a baseline. Using ViGather, participants on screen devices reported a higher sense of physical, spatial, and self-presence within the virtual meeting than when using Horizon. In our second study, we provided a detailed investigation into the experience of screen-device users and highlighted the crucial role of virtual avatar reconstruction in shaping their perceived self-presence. Our findings suggest that future VR meeting environments could benefit from a first-person full-screen view for screen-device users, situated closer to VR participants, to increase the sense of physical presence.

ViGather's use cases are not limited to Virtual Reality, as connecting users' virtual experiences to their physical setting (i.e., one's chair and surrounding desks) may make for even more compelling Mixed Reality applications. We therefore believe that ViGather's approach generalizes to the wider ecosystem of Augmented Reality and Mixed Reality platforms.

Taken together, we believe that ViGather demonstrates a viable direction to fuse immersive telepresence chats across screen devices and MR platforms to further people's experience during virtual conversations no matter where they are or what device type they use.

Beyond live meetings, ViGather could also facilitate meeting replays and inspections from arbitrary vantage points, allowing spectators to be immersed in 3D meetings from a first-person perspective even after the fact [19].

REFERENCES

[1] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. 2021. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23.

[2] Alphabet. 2022. MediaPipe. https://mediapipe.dev/

[3] Apple. 2022. ARKit. https://developer.apple.com/alp-reality/

[4] Sara Atske. 2021. 1. How the internet and technology shaped Americans' personal experiences amid COVID-19. https://www.pewresearch.org/internet/2021/09/01/how-the-internet-and-technology-shaped-americans-personal-

experiences-amid-covid-19/
[5] Autodesk. 2023. The Wild. https://thewild.com/
[6] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047* (2019).
[7] Mark Billinghurst and Hirokazu Kato. 1999. Real world teleconferencing. In *CHI'99 extended abstracts on Human factors in computing systems*. 194–195.
[8] Abraham G Campbell, Thomas Holz, Jonny Cosgrove, Mike Harlick, and Tadhg O'Sullivan. 2019. Uses of virtual reality for communication in financial services: A case study on comparing different telepresence interfaces: Virtual reality compared to video conferencing. In *Future of information and communication conference*. Springer, 463–481.
[9] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2527–2530.
[10] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. 2022. gDNA: Towards Generative Detailed Neural Avatars. In *Computer Vision and Pattern Recognition (CVPR)*.
[11] Yucheng Chen, Yingli Tian, and Mingyi He. 2020. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding* 192 (2020), 102897.
[12] Yi Fei Cheng, Tiffany Luong, Andreas Rene Fender, Paul Streli, and Christian Holz. 2022. ComforTable User Interfaces: Surfaces Reduce Input Error, Time, and Exertion for Tabletop and Mid-air User Interfaces. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 150–159.
[13] Cluster. 2023. Cluster. https://cluster.mu/
[14] Zoom Video Communications. 2022. zoom. https://zoom.us/
[15] HTC Corporation. 2022. HTC Vive Flow. https://www.vive.com/us/product/vive-flow/overview/
[16] Arthur Digital. 2023. Arthur. https://www.arthur.digital/
[17] Scott Elrod, Richard Bruce, Rich Gold, David Goldberg, Frank Halasz, William Janssen, David Lee, Kim McCall, Elin Pedersen, Ken Pier, et al. 1992. Liveboard: a large interactive display supporting group meetings, presentations, and remote collaboration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 599–607.
[18] Ben Elzendoorn, Marco De Baar, Rene Chavan, Timothy Goodman, Cock Heemskerk, Roland Heidinger, Klaus Kleefeldt, Jarich Koning, Stephen Sanders, Peter Späh, et al. 2009. Analysis of the ITER ECH Upper Port Launcher remote maintenance using virtual reality. *Fusion Engineering and Design* 84, 2-6 (2009), 733–735.
[19] Andreas Rene Fender and Christian Holz. 2022. Causality-preserving asynchronous reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
[20] Mozilla Foundation. 2022. mozilla:hubs. https://hubs.mozilla.com/
[21] Frame. 2023. Frame. https://learn.framevr.io//
[22] Jonathon D Hart, Thammathip Piumsomboon, Louise Lawrence, Gun A Lee, Ross T Smith, and Mark Billinghurst. 2018. Emotion sharing and augmentation in cooperative virtual reality games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 453–460.
[23] Jonathon Derek Hart, Thammathip Piumsomboon, Gun A Lee, Ross T Smith, and Mark Billinghurst. 2021. Manipulating Avatars for Enhanced Communication in Extended Reality. In *2021 IEEE International Conference on Intelligent Reality (ICIR)*. IEEE, 9–16.
[24] Jeremy Hartmann, Christian Holz, Eyal Ofek, and Andrew D Wilson. 2019. Realitycheck: Blending virtual environments with situated physical reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[25] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. Collabovr: A reconfigurable framework for creative collaboration in virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 542–554.
[26] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-aware 3D Photos. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 769–782.
[27] Hyperspace. 2023. MootUp. https://mootup.com/
[28] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*. Springer, 443–460.
[29] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–8.
[30] Brennan Jones, Yaying Zhang, Priscilla NY Wong, and Sean Rintel. 2020. Vroom: virtual robot overlay for online meetings. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.
[31] Mohamed Kari, Tobias Grosse-Puppendahl, Luis Falconeri Coelho, Andreas Rene Fender, David Bethge, Reinhard Schütte, and Christian Holz. 2021. Transformr: Pose-aware object substitution for composing alternate mixed realities. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 69–79.

[32] Shunichi Kasahara and Jun Rekimoto. 2014. JackIn: integrating first-person view with out-of-body vision generation for human-human augmentation. In *Proceedings of the 5th augmented human international conference*. 1–8.
[33] 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*. 105–112.
[34] Marco Kurzweg, Jens Reinhardt, Wladimir Nabok, and Katrin Wolf. 2021. Using Body Language of Avatars in VR Meetings as Communication Status Cue. *Proceedings of Mensch und Computer 2021* (2021).
[35] Marco Kurzweg and Katrin Wolf. 2022. Body Language of Avatars in VR Meetings as Communication Status Cue: Recommendations for Interaction Design and Implementation. *i-com* 21 (2022), 175 – 201.
[36] Hideaki Kuzuoka. 1992. Spatial workspace collaboration: a SharedView video support system for remote collaboration capability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 533–540.
[37] J Lafferty and P Eady. 1974. *The Desert Survival Problem Manual*.
[38] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring presence: the temple presence inventory. In *Proceedings of the 12th annual international workshop on presence*. 1–15.
[39] Tencent Holdings Ltd. 2022. Wechat. https://www.wechat.com/
[40] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
[41] Guido Makransky, Lau Lilleholt, and Anders Aaby. 2017. Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior* 72 (2017), 276–285.
[42] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)* 36, 4 (2017), 1–14.
[43] Inc Meta Platforms. 2022. Facebook. https://facebook.com
[44] Inc Meta Platforms. 2022. Meta Avatar SDK. https://developer.oculus.com/documentation/unity/meta-avatars-overview/
[45] Inc Meta Platforms. 2022. Meta Horizon Workrooms. https://www.oculus.com/workrooms
[46] Inc Meta Platforms. 2022. Meta Quest Pro. https://www.meta.com/ch/en/quest/quest-pro/
[47] Inc Microsoft. 2022. Microsoft Teams. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software
[48] Inc Microsoft. 2022. Skype. https://www.skype.com/en/
[49] Sebastian Molinillo, Rocío Aguilar-Illescas, Rafael Anaya-Sánchez, and María Vallespín-Arán. 2018. Exploring the impacts of interactions, social presence and emotional engagement on active collaborative learning in a social web-based environment. *Computers & Education* 123 (2018), 41–52.
[50] Teresa Monahan, Gavin McArdle, and Michela Bertolotto. 2008. Virtual reality for collaborative e-learning. *Computers & Education* 50, 4 (2008), 1339–1353.
[51] Thanh Khuong Nguyen and Thi Hong Tham Nguyen. 2021. The Acceptance and Use of Video Conferencing for Teaching in Covid-19 Pandemic: An Empirical Study in Vietnam. *AsiaCALL Online Journal* 12, 5 (2021), 1–16.
[52] Ohan Oda, Carmine Elvezio, Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual replicas for remote assistance in virtual and augmented reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 405–415.
[53] Kenton O'hara, Jesper Kjeldskov, and Jeni Paay. 2011. Blended interaction spaces for distributed team collaboration. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 1–28.
[54] OptiTrack. 2022. Motion Capture Systems. http://optitrack.com/
[55] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*. 741–754.
[56] Juyeon Park and Jennifer Paff Ogle. 2021. How virtual avatar experience interplays with self-concepts: the use of anthropometric 3D body models in the visual stimulation process. *Fashion and Textiles* 8, 1 (2021), 1–24.
[57] 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*. 1716–1725.
[58] Zhongling Pi, Yi Zhang, Fangfang Zhu, Ke Xu, Jiumin Yang, and Weiping Hu. 2019. Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures. *Comput. Educ.* 128 (2019), 345–352.
[59] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. 2018. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[60] Rabindra Ratan, Dave B Miller, and Jeremy N Bailenson. 2022. Facial appearance dissatisfaction explains differences in zoom fatigue. *Cyberpsychology, Behavior, and Social Networking* 25, 2 (2022), 124–129.
[61] Holger Regenbrecht, Michael Haller, Joerg Hauber, and Mark Billinghurst. 2006. Carpeno: interfacing remote collaborative virtual environments with table-top interaction. *Virtual Reality* 10, 2 (2006), 95–107.
[62] Daniel Roth, Constantin Kleinbeck, Tobias Feigl, Christopher Mutschler, and Marc Erich Latoschik. 2018. Beyond Replication: Augmenting Social Behaviors in Multi-User Virtual Realities. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2018), 215–222.
[63] Jiwon Ryu and Gerard Jounghyun Kim. 2020. Interchanging the Mode of Display Between Desktop and Immersive Headset for Effective and Usable On-line Learning. In *IHCI*.
[64] Bernd Schroeer, Andreas Kain, and Udo Lindemann. 2010. Supporting creativity in conceptual design: Method 635-extended. In *DS 60: Proceedings of DESIGN 2010, the 11th International Design Conference, Dubrovnik, Croatia*.
[65] Keisuke Shiro, Atsushi Okada, Takashi Miyaki, and Jun Rekimoto. 2018. Omnigaze: A display-covered omnidirectional camerafor conveying remote user's presence. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 176–183.
[66] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.
[67] Misha Sra, Aske Mottelson, and Pattie Maes. 2018. Your place and mine: Designing a shared VR experience for remotely located users. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 85–97.
[68] Frank Steinicke, Nale Lehmann-Willenbrock, and Annika Luisa Meinecke. 2020. A first pilot study to compare virtual group meetings using video conferences and (immersive) virtual reality. In *Symposium on Spatial User Interaction*. 1–2.
[69] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*. IEEE, 951–958.
[70] Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. 2023. HOOV: Hand Out-Of-View Tracking for Proprioceptive Interaction using Inertial Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[71] Spatial Systems. 2023. Spatial. https://www.spatial.io//
[72] Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2500–2509.
[73] Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2500–2509.
[74] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1653–1660.
[75] Hongyan Wang, Zhongling Pi, and Weiping Hu. 2019. The instructor's gaze guidance in video lectures improves learning. *J. Comput. Assist. Learn.* 35 (2019), 42–50.
[76] Stephan Wenninger, Jascha Achenbach, Andrea Bartl, Marc Erich Latoschik, and Mario Botsch. 2020. Realistic virtual humans from smartphone videos. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*. 1–11.
[77] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.
[78] Jason W Woodworth, David Broussard, and Christoph W Borst. 2022. Redirecting Desktop Interface Input to Animate Cross-Reality Avatars. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 843–851.
[79] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6184–6193.
[80] Andrew Yoshimura and Christoph W. Borst. 2020. Evaluation and Comparison of Desktop Viewing and Headset Viewing of Remote Lectures in VR with Mozilla Hubs. In *ICAT-EGVE*.
[81] Andrew Yoshimura and Christoph W. Borst. 2020. Evaluation of Headset-based Viewing and Desktop-based Viewing of Remote Lectures in a Social VR Platform. *26th ACM Symposium on Virtual Reality Software and Technology* (2020).
[82] Jacob Young, Tobias Langlotz, Matthew Cook, Steven Mills, and Holger Regenbrecht. 2019. Immersive telepresence and remote collaboration using mobile and wearable devices. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 1908–1918.
[83] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).
[84] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer graphics forum*, Vol. 37. Wiley Online Library, 523–550.

A APPENDIX

A.1 Application-level packet design

Our custom communication protocol divides the motion and audio stream into discrete packets. The flexible size of the packets is dynamically adjusted depending on the information they carry. Our packet design allows for easier logging of the data as well as aids the debugging of our multi-user system.

Our motion packet contains information about the global position and orientation of the body, the finger angles as well as the position and orientation of the two hands and the head.

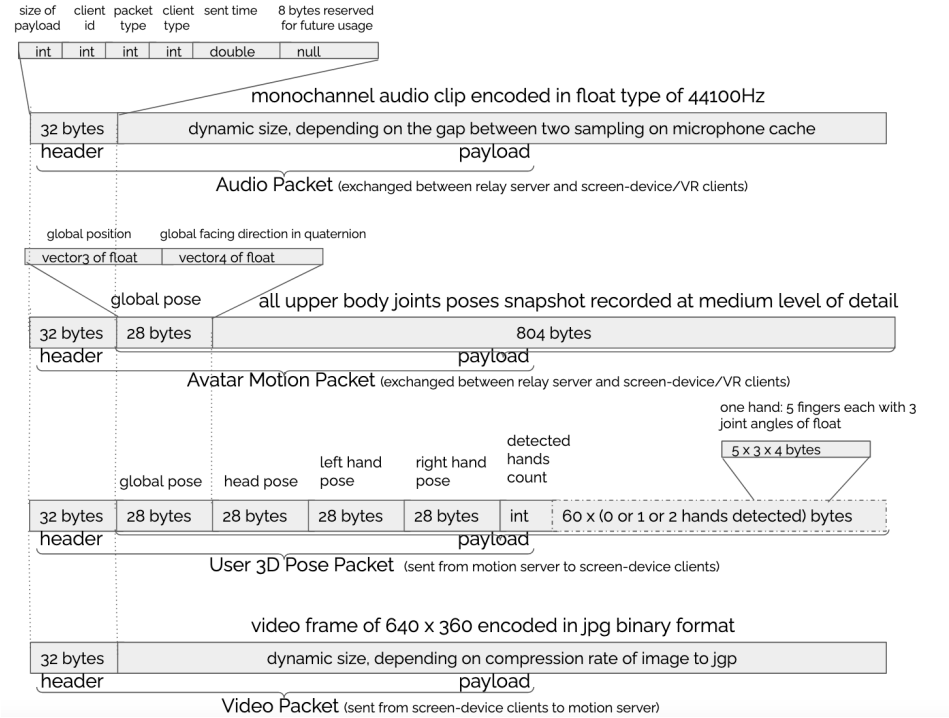

Fig. 12. Packet design for our custom application-level real-time protocol.

A.2 Illustration of avatar synthesis pipeline

The avatar synthesis pipeline handled by the motion server is depicted in Figure 13, providing an overview of the process.

Received January 2023; revised May 2023; accepted June 2023

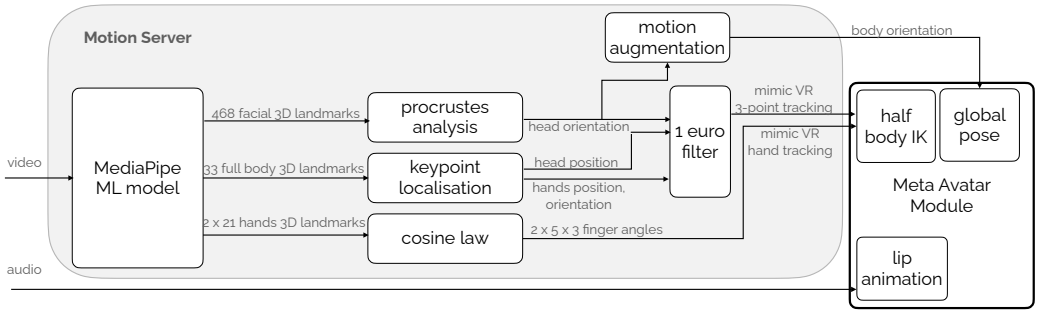

Fig. 13. Overview of the steps performed by the motion server for the avatar synthesis.