A Live Speech Driven Avatar-mediated Three-party Telepresence System: Design and Evaluation

Aobo Jin[1] Qixin Deng[2] * Zhigang Deng[2] †

1 University of Houston at Victoria, TX

2 University of Houston, Houston, TX

Abstract

In this paper, we present a live speech-driven, avatar-mediated, three-party telepresence system, through which three distant users, embodied as avatars in a shared 3D virtual world, can perform natural three-party telepresence that does not require tracking devices. Based on live speech input from three users, this system can real-time generate the corresponding conversational motions of all the avatars, including head motion, eye motion, lip movement, torso motion, and hand gesture. All motions are generated automatically at each user side based on live speech input, and a cloud server is utilized to transmit and synchronize motion and speech among different users. We conduct a formal user study to evaluate the usability and effectiveness of the system by comparing it with a well-known online virtual world, *Second Life*, and a widely-used online teleconferencing system, *Skype*. The user study results indicate our system can provide a measurably better telepresence user experience than the two widely-used methods.

*Equal contribution

†Correspondence E-mail: zdeng4@central.uh.edu

1 INTRODUCTION

With the recent rapid advances in Internet technologies, telepresence has been increasingly used for meeting people located in different places. Several telepresence types have emerged such as telemeeting through audio/video streams or virtual meetings in a shared online virtual world. Researchers have attempted various efforts to increase the telepresence experience by improving the quality of audio/video streams (Goolcharan & Karunasiri, 2000; Yoo et al., 2004; Williams & Libove, 1999), or by developing novel telepresence or telecommunication algorithms (Chia, 1992; Vannucci, 1995). Popular commercial systems, such as Skype, Microsoft Teams, Zoom, and Google+ hangout, can provide a low-cost (or even free) solution to audio/video telecommunication. Such systems essentially record and transfer remote participants' face/body video streams in real-time. Nevertheless, they cannot truly immerse remote participants into an extended perception space (i.e., a shared meeting environment) due to the lack of multiparty conversation situated gestures. Meanwhile, immersive telepresence in shared online virtual worlds (also called *avatar-mediated multi-party telepresence*) has attracted increasing attention in recent years. In these systems, each user can control an embodied avatar in the virtual world to communicate with other avatars (users). The main advantages of such immersive telepresence systems, compared to off-the-shelf video-based telemeeting (e.g., Skype), include: having a high freedom of interaction (e.g., view selection and manipulation), selecting stylized or customized outlook of the embodied avatars, and providing the users with a sense of natural interaction that closely mimics real-world multiparty conversations.

To improve the user experience of avatar-mediated multi-party telepresence, researchers and practitioners have proposed various schemes including enhancing the modeling quality (Nichol & Wong, 2005; Achenbach, Waltemate, Latoschik, & Botsch, 2017; Ma & Deng, 2019), enriching avatar animations (Singh, Ohya, & Parent, 1995), and simplifying user control (Le, Ma, & Deng, 2012; Le, Zhu, & Deng, 2013). Despite these progresses, current avatar-mediated multi-party telepresence methods, even in the most widely-used systems (e.g., the Second Life), suffer from the following limitations. First, users are required manually select *when* and *which* of the pre-

created conversational gesture animations needs to be played (e.g., type a command or select a menu item to trigger the play of a pre-created hand-waving animation) during the conversation, which is an unnatural user interaction in the middle of multi-party conversations. More importantly, because of such manual involvements, the natural synchronization between conversational gesture and conversational content (e.g., speech or texts) cannot be preserved. Second, all the conversational gesture animations need to be pre-created. Therefore, the variety of pre-built gesture animations is often limited, and the natural flow between different gesture animations is missing (Kelly, Özyürek, & Maris, 2010). Because of the above limitations, existing avatar-mediated multi-party telepresence methods are rudimentary and generally fall short of producing life-like presence experience that is anticipated to closely mimic real-world multiparty conversations.

(a) (b)

Figure 1: (a): The first person view of a user when he/she is using our system. (b): The third person view of our runtime system, where three avatars stand at the three vertices of a equilateral triangle in the virtual environment.

Inspired by the above challenges, we propose a light-weight, live speech-driven, avatar-mediated, three-party telepresence system (Figure 1), which is aimed to faithfully mimic real-world natural three-party conversations in a shared virtual environment, without requiring any tracking devices (e.g., motion capture systems). Our system can take the live speech of all the three users as the input and then real-time generate their conversational gestures simultaneously (including head movement, eye movement, hand movement, and torso movement). In this way, from the first-person view, each user would be immersed in the shared virtual meeting environment for telepresence. It is noteworthy that, instead of focusing on general multi-party telepresence, this

work is specifically focused on three-party telepresence, since, as the simplest form of multiparty telepresence, three-party telepresence contains many ingredients of general multi-party telepresence and thus it would inspire further research along this direction. The design of our system also aims to be easily deployable. In other words, it can be set up quickly and only requires commodity computing devices (e.g., an off-the-shelf PC) with a microphone.

Specifically, our system employs a client-server architecture (refer to Figure 2), where three users located in different places communicate with each other, with the aid of a Unity3D Photon Networking cloud server, by transmitting and synchronizing gesture motion and speech data (blue arrows). At each user (client) side our method generates his/her motion based on his/her speaking or listening status as well as live speech input (if speaking), as illustrated in Figure 2. We determine whether a user is speaking or listening at any moment by thresholding the speech signal. The details of our motion synthesis algorithms at the user side are described in Section 3. To evaluate the effectiveness and usability of our system, we conducted a formal study to directly compare our method with both the Skype system, one of the widely-used teleconferencing tools, and the Second Life (SL) system, which is one of the widely-used avatar-mediated telepresence systems. The user study results indicate that our system can measurably outperform both the Skype method and the SL method in terms of user experience, effectiveness, and enjoyability.

2 Related Work

In this section, we briefly review recent related efforts on telepresence systems, studies on avatar representations, and computer-mediated communication, instead of virtual human technologies. Readers of interest are referred to recent surveys on facial animation (Deng & Noh, 2008; Lewis et al., 2014), character/skinning animation (Jacobson, Deng, Kavan, & Lewis, 2014), eye animation (Ruhland et al., 2014), presence in virtual reality (VR) (Schuemie, Van Der Straaten, Krijn, & Van Der Mast, 2001), and social VR platforms.

Telepresence Systems. Over the past decades, many high-end commercial telepresence solutions, such as Cisco's TelePresence, Polycoms TPX, and HPs HALO, have been developed. Still, besides their high cost, these systems are in general video-based, which limits the natu-

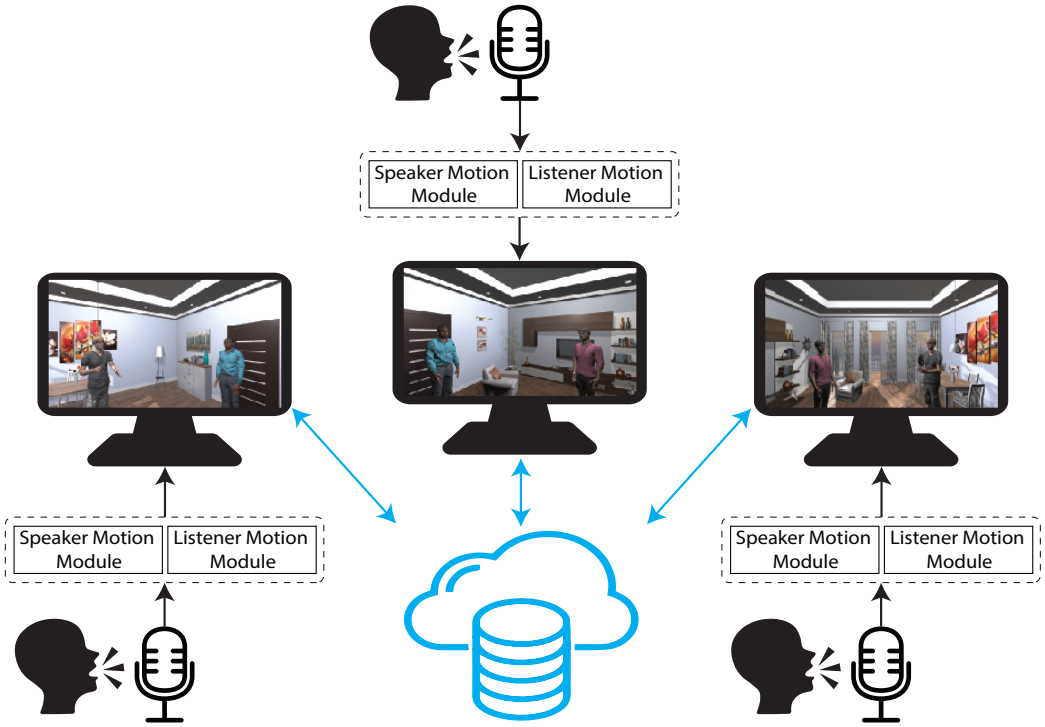

Figure 2: The architecture illustration of our system.

ralness and thereby the sense of telepresence. In academia, with the aid of off-the-shelf cameras, projectors, and trackers, various video-based teleconferencing research prototype systems including the tele-cubicles system (Gibbs, Arapis, & Breiteneder, 1999), the Office of the Future system (Raskar et al., 1998; Wen, Towles, Nyland, Welch, & Fuchs, 2000; Johnson, Gyarfas, Skarbez, Towles, & Fuchs, 2007; Yang, Kurashima, Towles, Nashel, & Zuffo, 2007; Johnson, Welch, Fuchs, La Force, & Towles, 2009; Lincoln et al., 2009), the virtual team user environment (VIRTUE) (Kauff & Schreer, 2002, 2005; Divorra et al., 2010), and room-size informal telepresence system (Dou et al., 2012) have been proposed for telepresence and tele-collaboration. Besides the relatively high cost, these systems are ideally suitable for only a fixed number (2-3) of participants since adding more means the rearrangement of physical space, and their applications are limited to certain indoor environment with delicately pre-configured and calibrated hardware setup. In addition, pseudo-3D video conferencing (Harrison & Hudson, 2008) and

applying augmented reality technology to video conferencing (Barakonyi, Fahmy, & Schmalstieg, 2004) had also been explored. Another category of approaches in this area is focused to seamlessly integrate 2D video or even 3D video avatars into collaborative virtual graphical environments to increase realism (Ståhl, 1999; Rauthenberg, Graffunder, Kowalik, & Kauff, 1999). A simplified yet portable immersive virtual environment that leverages mobile communication platforms, 3D virtual humans, motion trackers and displays to facilitate ad-hoc virtual collaboration has also been proposed (Basu, Raij, & Johnsen, 2012). In addition, there are room-sized virtual environment systems such as CAVE (Cruz-Neira, Sandin, & DeFanti, 1993) and BLUE-C system (Gross et al., 2003).

Studies on Avatar Representations. Numerous studies have been conducted to investigate the effect of avatar representations in virtual environment. For example, researchers studied the effect of using a human-like face as a computer interface (Sproull, Subramani, Kiesler, Walker, & Waters, 1996; Pandzic, Ostermann, & Millen, 1999; Bonito, Burgoon, & Bengtsson, 1999; Cassell, Sullivan, Churchill, & Prevost, 2000; Rizzo, Neumann, Enciso, Fidaleo, & Noh, 2001; Gulz, 2005; Ku et al., 2005; Yee, Bailenson, & Rickertsen, 2007; Rincón-Nigro & Deng, 2013). In general, people like to do interactions with humans instead of with avatars. To investigate the effect of reduced social information and behavioral channels in immersive virtual environments with full-body avatar embodiment, Roth et al. compared physically-based and verbal-based social interactions in real world (RW) and virtual reality (VR) and concluded that the reduced social information and behavioral channels of participants is due to the shifting their attentions to other behavioral channels (Roth et al., 2016). Latoschik et al. explored the effect of avatar realism on embodiment and social interactions in VR by comparing abstract avatar representations based on a wooden mannequin with high fidelity avatars generated from photogrammetric 3D scan methods (Latoschik et al., 2017). Aseeri and Interrante investigated the influence of avatar representations and behavior on communication in an immersive, multi-user, same-place virtual environment by comparing three conditions of avatar representation: video see-through, scanned realistic avatar, and no-avatar representations (Aseeri & Interrante, 2018). Cho et al. studied the

effect of volumetric capture avatars on social presence in immersive virtual environments, by comparing the volumetric capture avatar of an actor with the actor captured in 2D video and another 3D avatar obtained by pre-scanning the actor (Cho, Kim, Lee, Ahn, & Han, 2020). Maloney et al. showed empirical evidence on the role of non-verbal communication and its influence on interactive experience in virtual environment (Maloney, Freeman, & Wohn, 2020).

Computer-Mediated Communication (CMC). With the rapid development of VR technologies, various avatar-mediated communication systems have emerged, reintroducing non-verbal features into mediated communication. Avatars are thus more similar to real-time Face-to-Face interactions than traditional text-based CMC with regard to communication bandwidth. However, avatar-mediated communication provides wider control over the representations of the actors. Experimental comparison of different CMC modalities confirms that avatar-mediated communication is better than text-based CMC in perceived intimateness, co-presence, and trust (Bente, Rüggenberg, Krämer, & Eschenburg, 2008). Empirical research further found that nonverbal cues from avatars tend to elicit the same socio-emotional responses from human participants (Yee & Bailenson, 2007; Yee, Bailenson, Urbanek, Chang, & Merget, 2007), and careful engineering of avatar-mediated communication would result in transformed interaction outcomes (Blascovich & Bailenson, 2011).

Social VR Platforms. With the rapid advances of VR technologies and systems, researchers utilized VR environment for social interaction and collaboration for people at different geographical locations. In recent years a number of such social VR platforms have been developed. For example, the Rec Room system (Rec Room, 2016) is a platform for players to chat, hang out, and explore millions of player-created rooms, or build something new together. The Rec Room supports various types of VR devices including Oculus, PlayStation, Xbox, and iOS. The AltspaceVR platform (Microsoft Inc., 2022) enables players to build virtual worlds together with VR controllers. The VRChat system (VRChat Inc., 2022) shares a similar idea with the Rec Room system, but it provides more customization freedoms on avatars including avatar modeling and animation. In the vTime XR system (vTime Holdings Limited, 2022), people can easily customize their avatars

as well as virtual gestures to improve social interaction experience. Researchers also developed the Horizon Worlds (Meta Platforms, Inc., 2022) to enhance the capability for creativity in VR environment. People can even do their own real-life work in the VR environment by creating a virtual office space. Similarly, the ImersedVR system (Immersed Inc., 2022) provides a virtual working environment for people to improve the work efficiency. The above social VR platforms and applications have been mainly focused on the design, modeling, and creation of avatars or virtual spaces, instead on efficient communication (Neos VR Metaverse, 2022; Ad Alternum Game Studios, 2022; Mozilla Foundation, 2022; Wild Technology Inc., 2022; Bigscreen inc., 2022). For instance, these VR platforms either require VR controllers or use pre-designed animations for avatars to represent conversational gestures at runtime.

3 System Design

To develop a live speech-driven, multi-party telepresence system, the main challenge is to real-time generate socially-engaging non-verbal behaviors for both speakers and listeners, with live speech signal as the input. As illustrated in Figure 2, in our system each user's side generates his/her avatar motion based on his/her status (e.g., listening or speaking) and the live speech signal (if speaking). The synthesized motions of different avatars are transmitted and synchronized via a Photon Networking cloud server. In the following, we describe how our system real-time generates avatar motion at each user's side.

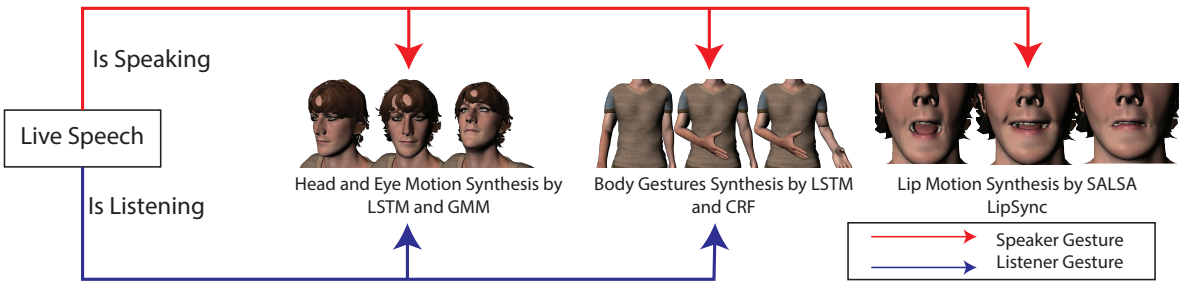

Figure 3: The pipeline of the motion synthesis components at each user's side in our system. We use a pre-defined intensity threshold to determine a user is speaking or listening at any moment and then call the corresponding motion synthesis modules. Here GMM stands for Gaussian Mixture Models.

As illustrated in Figure 3, we utilize a divide-and-conquer strategy to divide the avatar motion synthesis into several relatively independent parts; meanwhile, we use the live speech signal,

as the synchronizer, to drive their motion synthesis. These parts include: (i) the head-and-eye motion synthesis for both speakers and listeners (Section 3.1), (ii) the body gesture synthesis (including torso movement and hand movement) for speakers (Section 3.2), (iii) the body gesture synthesis (including torso movement, and hand movement) for listeners (Section 3.3), and (iv) lip motion generation for speakers. Note that the synthesis of finger motion is not considered in our current system. One major advantage of our system design is that, even when speech overlap occurs (e.g., multiple users are speaking simultaneously), our system can handle it naturally due to the distributed design of avatar motion synthesis (refer to Figure 2). Our system directly employs the SALSA LipSync tool (Unity Asset Store, 2020) to automatically generate lip motion based on live speech input. Below we describe the main algorithms used in the other three parts.

3.1 Head-and-Eye Motion Synthesis

In our system, we employ the recent deep learning based framework (Jin, Deng, Zhang, & Deng, 2019) to generate head-and-eye motion of the avatars in a three-party conversation, based on live speech input. For the sake of readability, below we briefly describe its head-and-eye motion synthesis algorithm. For more algorithm details, please refer to (Jin et al., 2019).

Given a pre-collected three-party conversational motion dataset, the method in (Jin et al., 2019) trains a one-layer Long Short-term Memory (LSTM) model, called "speaker LSTM", to generate the Direction-of-Focus (DFocs) of the speaker frame by frame based on speech features, and trains another one-layer LSTM model, called "listener LSTM", to generate the DFocs of listeners frame by frame. Then, from the predicted DFocs at each frame, a two-steps refinement method is designed to extract the head and eye rotation angles for this frame. We obtain pre-trained models using the work of (Jin et al., 2019) and build the same architecture mentioned in (Jin et al., 2019) to infer the head-and-eye motion of avatars with speech input. Since both the pre-trained one-layer LSTM model and the refinement method are highly efficient at runtime, the head-and-eye motion synthesis module in our system can real-time generate motion based on live speech input. Since the positions of all the avatars are fixed (as shown in Figure 1(b)), the horizontal rotation angle for head-and-eye is in the range of [-30, 30] degree.

3.2 Body Gesture Synthesis for Speakers

We employ the "gesture controller" (Levine, Krähenbühl, Thrun, & Koltun, 2010) to generate both torso movement and hand movement of a speaking avatar based on live speech input. This method consist of two layers: *an inference layer*, which analyzes vocal prosody and produces a distribution over a set of learned hidden states, and *a control layer*, which uses the inferred hidden state distribution and other available inputs to select the most appropriate gesture segments from a pre-created library of motion data using a pre-computed optimal policy. Specifically, the inference layer trains a HMM (Hidden Markov Models) on the motion signal to obtain a distribution over a sequence of hidden states. Then, the fixed distribution is utilized to train CRF (Conditional Random Fields) to maximize the probability of the hidden state distribution given the input signal which is live speech in this case. We utilize a pre-recorded, audio-synchronized motion capture dataset as the training data to train both the HMM and CRF models. The length of the used dataset is about 10 minutes, where one participant stands naturally and talks about a specified topic (i.e., campus life in our experiment).

After the training, the online control layer, which uses a Markov decision process to synthesize an animation stream using an optimal gesture selection policy. During online synthesis, given the previously selected frame f from a segment a, the cost of selecting a frame f' from a segment a' at time step t is given by:

$$C(a', f', a, f, \alpha_t) = \alpha_t \cdot \delta_{a'} + S(a, f, a', f'), \tag{1}$$

where α_t is the forward probability, $\delta_{a'}$ is the pre-computed vector for the segment a', and S is defined as:

$$S(a, f, a', f') = \begin{cases} 0 & if a_i = a_j \text{ and } f_i = f_{i+1} \\ D_{int}(a_i, f_i, a_j) & if f_i \in R_{a_i} \text{ and } f_j = 0 \\ D_{s-p}(a_i, a_j) & if f_i = n_{a_i} \text{ and } f_j = 0 \\ \infty & \text{otherwise} \end{cases} \tag{2}$$

The value function that optimizes this cost over an infinite time horizon with a discount factor η is given by:

$$\begin{aligned} V_c(a', f', a, f, \alpha_t) &= C(a', f', a, f, \alpha_t) \\ &+ \eta \min_{a'', f''} V_c(a'', f'', a', f', \alpha_{t+1}). \end{aligned} \tag{3}$$

At run-time, the optimally selected frame and segment at time t are given by

$$(a^*, f^*) = \arg\min_{a', f'} \sum_i \alpha_t(i) V_d(a', f', a, f, s_i). \tag{4}$$

After some pruning operations, this method can work in real time. Note that, in our current system, the synthesized body gesture only contains upper-body gesture, without finger motion.

3.3 Body Gesture Synthesis for Listeners

We extend the texture synthesis based eye motion synthesis approach (Deng, Lewis, & Neumann, 2005) to generate torso movement and hand movement for listeners. Specifically, from a pre-collected listener motion dataset, we select some sequences as motion samples in our synthesis process. Each texel in this case includes all the relevant joint angles J of the upper body. We randomly choose a patch from qualified candidate patches in the provided motion samples. The distance between two motion patches (i.e., motion sample blocks) is calculated as follows:

$$d(B_{in}, B_{out}) = (\frac{1}{A} \sum_{k=1}^{A} d\_tex(t_{in}^k, t_{out}^k))^{\frac{1}{2}} \tag{5}$$

$$d\_tex(t^1, t^2) = \sum_{j \in J} ((\alpha_j^1 - \alpha_j^2)^2 / V_j^\alpha + (\beta_j^1 - \beta_j^2)^2 / V_j^\beta + (\gamma_j^1 - \gamma_j^2)^2 / V_j^\gamma), \tag{6}$$

where A is the size of the boundary zone that functions as a search window, α, β, and γ denote the yaw, pitch, and roll angles of one joint, respectively, and V_j is the variance of the joint angle in the acquired motion dataset. Here t_{in} represents the given motion samples and t_{out} represents the synthesized motion sample. Note that any length of body gesture for listeners can be generated with an online fashion using the above method. Since finger motion and lower-body gestures

of listeners are generally less directly correlated with conversational context, our system ignores the synthesis of these motions for listeners.

3.4 Other Implementation Details

Speaking/listening determination. With live speech signal input, we first extract its pitch and intensity features with a sampling rate of 44,100 Hz. Since the pitch feature is undefined for unvoiced periods, in this case the pitch feature is generated for unvoiced periods through a uniform distribution. We also experimentally choose a threshold value for the intensity feature to determine whether a user is speaking or listening. If the current intensity feature is higher than the threshold, we assume the user is speaking; otherwise, we assume the user is listening.

Motion transition for role change. Role change (e.g., switching from speaking to listening; or vice versa) could bring a discontinuous motion for avatars, so we utilize a motion buffer to linearly interpolate and thus smooth discontinuous motion. For example, if the switching is from speaking to listening, we first create a buffer by linearly interpolating the last predicted motion frame of the speaker and the current predicted motion frame of the listener.

System modes. We design two modes for our system: Control model and Auto mode. In the Control mode, each avatar stands at the same location, but each user can use the keyboard to control the viewing direction of his/her avatar during the telepresence. In the Auto mode, the first-person perspective is fixed and users cannot control it. The Auto mode directly uses the method described in Section 3.1 to synthesize the head-and-eye motion for avatars. Users have flexibility to select one from the two modes to participate in the telepresence. Also, the users can switch the mode between the two at any time during the runtime of our system.

System performance. We used several measures to characterize our system performance, including motion synthesis delay, and motion synchronization delay. *Motion synthesis delay* is the average time elapsed between the feeding of live speech and synthesis of corresponding motion. *Motion synchronization delay* is the average time for frame-level motion synchronization (i.e., the average time for sending a frame of motion to the server and then distributing it to all the users). We show the recorded average values of the two measures with system bandwidth in Table 1.

Table 1: The recorded performance measures of our runtime system.

Motion synthesis delay	Motion synchronization delay	Download	Upload
0.031996 s	0.717670 s	100 Mbps	100 Mbps

4 USER STUDY

To evaluate the effectiveness and usability of our system, we conducted a formal comparative user study, described below.

4.1 Comparison Conditions

In our comparison study, participants use 6 different conditions for three-party conversation experiments. We describe these conditions below. First, depending on (i) whether the participant wears a VR HMD Headset and (ii) whether the participant can manually control avatar head movement at runtime, we generate 4 different experimental conditions based on our method: (1) Control Mode + w/ HMD, (2) Control Mode + w/o HMD, (3) Auto Mode + w/ HMD, and (4) Auto Mode + w/o HMD. Note that we created two different modes (Auto/Control) of our system in the study since the two modes can be treated as two slightly different systems in terms of user control/interaction, which may provide different user experiences and telepresence immersion for participants.

(1) **Control Mode + w/ HMD**. In this condition, the participant wears an Oculus VR Headset to participate in three-party conversations, and the participant can freely control the head motion of his/her avatar by directly transferring his/her real-world head movement, with the aid of the HMD tracker.

(2) **Control Mode + w/o HMD**. In this condition, the participant does not wear an HMD headset to participate in three-party conversations, and the participant can use keyboards to control the left/right/up/down head movement of his/her avatar at runtime.

(3) **Auto Mode + w/ HMD**. In this condition, the participant wears an Oculus VR Headset, but he/she cannot manually control the movement of his/her avatar at runtime. All the motions of the avatar are automatically generated by our algorithms, described in Section 3.

Figure 4: Runtime screenshots of the experimental conditions in this study, including a screenshot of the Control Mode of our system (a), a screenshot of the Auto Mode of our system (b), a screenshot of the Second Life method (c), and a screenshot of Skype (d).

(4) **Auto Mode + w/o HMD**. In this condition, the participant does not wear an HMD headset, and he/she also cannot manually control the movement of his/her avatar at runtime.

Meanwhile, we add the Second Life method and the Skype method as the fifth and sixth experimental conditions in our comparative study, described below. Through the comparison with the Skype method, we aim to evaluate whether our system can rigorously mimic three-party conversations in virtual worlds and provide sufficient telepresence to participants. Through the comparison with the Second Life method, we aim to evaluate if the automated conversational gesture by our system can significantly facilitate the telepresence experience for participants.

(5) **SL (Second Life) + w/o HMD**. We choose the SL (Second Life) as one of the comparison methods since one of its wide uses is to do avatar-mediated multiparty telepresence in an online virtual world. In this condition, the participant does not wear an HMD headset. In the SL system, the first-person view is fixed (similar to the above Auto Mode of our

system) and the participant does not have control over it. Note that the SL environment allows the participant to control the movement of the whole avatar, but it does not allow the participant to only control the head orientation.

(6) **Skype + w/o HMD**. We also choose the Skype method as another comparison method due to its wide use for multi-party teleconferencing in the real world. Note that the Skype system generally requires a PC with both a webcam and a microphone, while both the SL and our system only need a PC with a microphone. In this condition, the participant does not wear an HMD headset.

To make a fair comparison, we position three avatars in the three vertices of an equilateral triangle in the virtual environment, and the 3D environment is rendered with the first-person view in both our method (i.e., the above (1)-(4) conditions) and the SL + w/o HMD condition. Also, with the SL + w/o HMD condition, participants are asked to use their speech (not typing) during the experiments, and the participants can choose to manually trigger SL-provided, pre-created gesture animations during the telepresence. Figure 4 shows the runtime screenshots of the conditions in our study. Note that in Figure 4, both (a) and (b) are runtime screenshots of our method, but the difference is: (a) is in the Control mode and (b) is in the Auto mode. Also, the UI of our system in both (a) and (b) is invisible during the experiments.

4.2 Participants

A total of 30 participants (24 males and 6 females, 20 to 35 years old) who are naive in this research were recruited from a university campus for our study. They were randomly divided into 10 groups. For each group, the three participants are instructed to stay in three different rooms to do the experiments.

4.3 Study Design

The three participants in each group (10 groups in total) are randomly assigned to three different physical rooms with the controlled environment. In one of the rooms, VR environment with an

Oculus HMD headset (with the embedded audio/microphone support) is set up, and the participant assigned to this room is asked to wear the HMD to participate in our experiments when our avatar-mediated telepresence system is used. However, the participant is not asked to wear the HMD when the Skype or SL method is used. In the other two rooms, each participant sits in front of a monitor and wears earphones connected with a desktop computer. Therefore, when our system (both the Control Mode and the Auto mode) is used, one participant with HMD telecommunicates with two other participants who do not wear any HMD headsets. In other words, among the 30 participants, a total of 10 participants use our system by wearing the HMD headset, while the remaining 20 participants use our system with naked eyes (i.e., not wearing any HMD headsets). Note that, we intentionally utilize one HMD headset for one of the three participants in each group, instead of using three HMD headsets for all the three participants in one group, since we also aim to compare and measure the telepresence difference of participants when they use our system with/without an immersive HMD headset. During the experiments the participants physically sit in front of a computer, while their embodied avatars stand in the virtual environment. The main reason is that, we recorded the training data from standing participants and thus it is more natural and easier to synthesize the motions of standing avatars than those of sitting avatars. Figure 5 shows the environment and equipment used in our user study.

There are four sessions for each group. In each session, a different experimental condition is used. Specifically, for the participant in the room with an HMD headset, he/she has following four experimental conditions: (i) Skype + w/o HMD, (ii) SL + w/o HMD, (iii) Control Mode + w/ HMD, and (iv) Auto Mode + w/ HMD. Meanwhile, for the two participants in the rooms without an HMD headset, they have the following four experimental conditions: (i) Skype + w/o HMD, (ii) SL + w/o HMD, (iii) Control Mode + w/o HMD, and (iv) Auto Mode + w/o HMD. Furthermore, during the experiments, we instruct the participants not to move their avatars substantially to keep triangular positioning in three-party conversations. For each group, the order of the 4 sessions is randomized to eliminate the potential context effect. To make a fair comparison, we inform participants the two modes of our system as two slightly different systems, and we also ensure

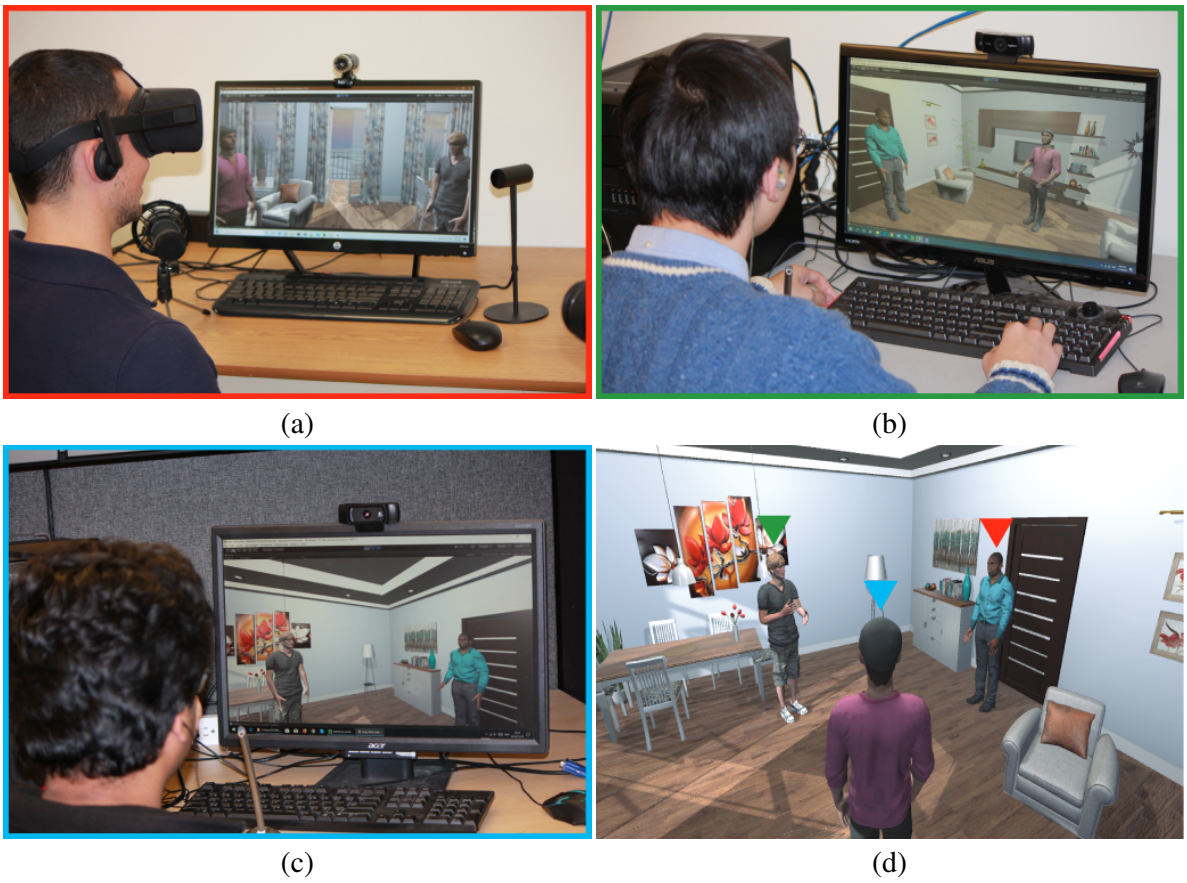

Figure 5: Snapshots of our user study. One participant wears the HMD (a) to communicate with the other two participants (b) and (c). Their avatars are situated in the same online virtual room (d), which is rendered from the third person view. The boundary colors of (a), (b) and (c) are consistent with the colors of the inverted triangles in (d) to represent the specific avatar of each participant.

not to arrange the two modes of our system consecutively during our study.

Each session lasts about 10 minutes, and there is a 2-minute break between two sessions. To increase the willingness of talking for participants, before each session, participants are not given any topics or tasks to mimic real-world scenarios. We only instruct participants to talk with each other naturally. Since all participants are from the same university campus, they can easily discuss some common campus topics, and none of the groups shows communication barriers. After each session, participants are asked to respond to a questionnaire with six questions (Table 2). Specifically, they need to give a score to each question regarding the telepresence condition they just experienced in the last session. The same questionnaire is used for different sessions in our study. Inspired by the well-known SUS questionnaires (Slater, Sadagic, Usoh, & Schroeder,

2000; Slater, McCarthy, & Maringelli, 1998), six questions in our questionnaire are designed to cover six different aspects: (1) enjoyment, (2) isolation, (3) meet individuals again, (4) comfort within environment, (5) embarrassment, and (6) visually comfort with the other participants, as shown in Table 2. Except the question No. 2, other questions are rated with a scale from 1 to 7, where 1 denotes the lowest level and 7 denotes the highest level. Some questions have their own scale explanations to help participants consistently rate them.

Table 2: Questionnaire used in our user study.

#	Question
Q1 (enjoyment)	Think about a previous time when you enjoyed meeting with others. To what extent have you enjoyed the meeting experience just now?
Q2 (isolation)	To what extent was yourself 'isolated' compared to the other two people in this meeting environment? Give a score out of 100, where a person scores 100 if they were completely isolated from the other two.
Q3 (meet individuals again)	Would you like to meet any of the other two people again in this meeting environment? (Explanation. 1: I would not like to meet this person in this meeting environment; 4: No preference either way; 7: I would very much like to meet this person in this meeting environment.)
Q4 (comfort within environment)	The extent to which I felt comfortable with each of the other two persons was using this meeting environment. (Explanation. 1: I felt very uncomfortable; 4: Neither comfortable/nor uncomfortable; 7: I felt very comfortable.)
Q5 (embarrassment)	Did this meeting environment make you feel self-conscious or embarrassed? (Explanation. 1: It did not make me feel this way; 7: It did make me feel this way very much.)
Q6 (visually comfort with the other participants)	The extent to which I felt comfortable with the other two persons' faces (including head and eye motion). (Explanation. 1: I felt very uncomfortable; 4: Neither comfortable/nor uncomfortable; 7: I felt very comfortable.)

4.4 Materials and Setup

Three desktop computers are put in three different rooms at a university campus. Each desktop computer has an LCD monitor with 1920×1080 resolution and the distance between the head of the participant and the monitor is approximately 0.6 meter. Each desktop computer is connected

to a computer stand microphone and an HD USB webcam with 1080P streaming. The desktop computer connected with HMD has the following configuration: Intel i7-6700 CPU @3.4 GHz, 16GB Memory, NVIDIA Geforce GTX 1070 GPU, and a 64-bit Operating System (x64-based processor) Windows 10 Home. The used HMD is Oculus Rift with a resolution 2160 $\times$ 1200 (1080 $\times$ 1200 per eye), refresh rate 90Hz, field of view (FOV) 110°, and weight of 480 grams. The configurations of the other two desktop computers in our study are: (i) Intel i7 CPU 860 @2.8 GHz, 16GB Memory, NVIDIA Geforce GTX 1060 GPU, and a 64-bit Operating System (x64-based processor) Windows 10 Home; and (ii) Intel i5-6400 CPU @2.7 GHz, 16GB Memory, NVIDIA Geforce GTX 1060 GPU, and a 64-bit Operating System (x64-based processor) Windows 10 Home, respectively. All the computers are connected to the campus wireless network through WI-FI. Figure 6 shows the experimental setup without (left) and with (right) HMD.

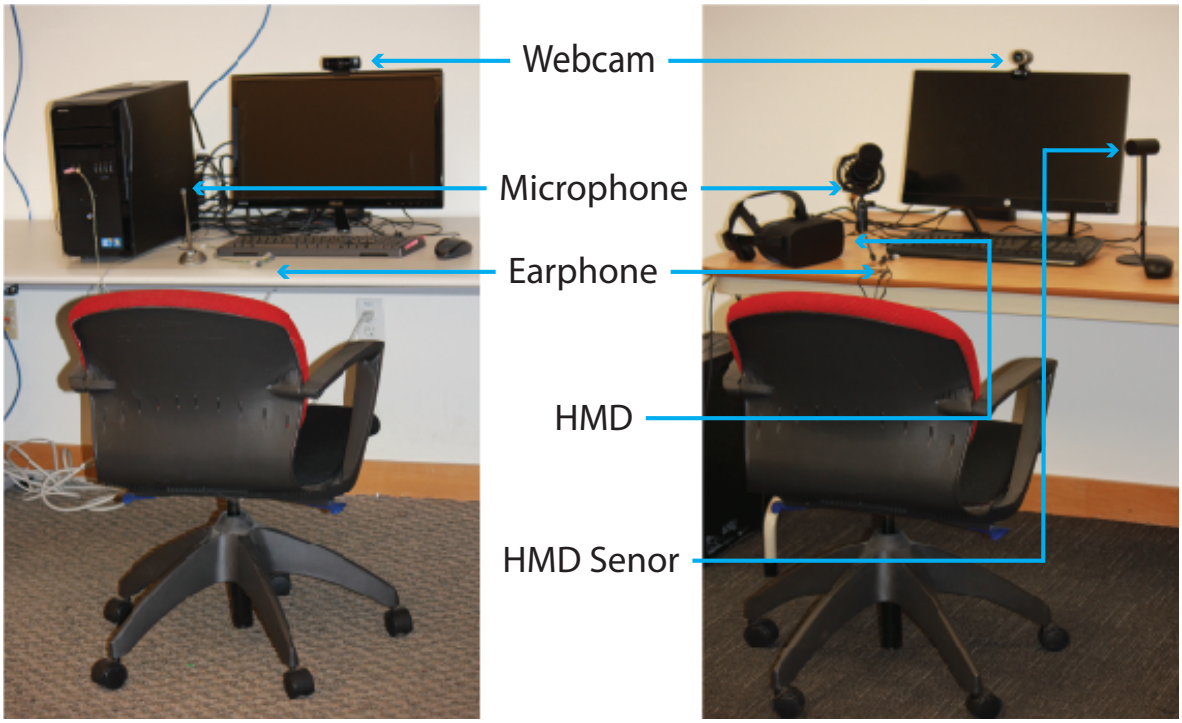

Figure 6: The devices used for participants without HMD (left) and with HMD (right)

4.5 Results and Discussion

After the participants' responses to the questionnaire are collected, we perform a quantitative analysis of our received score data, as described in this section. We utilize the Friedman test to analyze whether there is a statistically significant difference among the four w/o HMD conditions, based on the scores of the 20 participants w/o HMD as shown in Table 3. As shown in the table, there are statistically significant differences among the four w/o HMD conditions for all the 6 questions. In the following, we describe quantitative analysis of the user scores on each question.

Table 3: *p*-values of the user scores on the all questions with the Friedman test.

Question #	Q1	Q2	Q3	Q4	Q5	Q6
p-value	0.000002	0.004362	0.000001	0.000002	0.000046	0.000599

4.5.1 Enjoyment

The No. 1 question in our questionnaire concerns the enjoyment experience of the participants. Figure 7 illustrates the box plots of the enjoyment scores for four conditions (i.e., Control Mode + w/o HMD, Auto Mode + w/o HMD, SL + w/o HMD, and Skype + w/o HMD). We consider 20 participants in total since each of them scored all the four conditions during the experiments. As shown in Figure 7, the Auto Mode + w/o HMD condition (of our system) obtains the highest average score, and the SL + w/o HMD condition obtains the lowest average score. Also, the average score of our system (i.e., averaging the Auto Mode + w/o HMD condition and the Control Mode + w/o HMD condition) are slightly higher than the average score of the Skype + w/o HMD condition and significantly higher than the average score of the SL + w/o HMD condition.
The average scores and standard deviations of Question #1 are reported in Table 4. We performed a non-parametric statistical analysis because the collected data are not normally distributed (Shapiro-Wilk test). The results by the Nemenyi post-hoc test are shown in Table 5, in which we found statistically significant differences for 3 pairs, p-value < 0.05: Auto Mode + w/o HMD vs. Skype + w/o HMD, Auto Mode + w/o HMD vs. SL + w/o HMD, and Control Mode + w/o

HMD vs. SL + w/o HMD. Specifically, in terms of enjoyment, our system (both Control Mode + w/o HMD and Auto Mode + w/o HMD conditions) is statistically significantly better than the SL + w/o HMD condition, and the Auto Mode + w/o HMD condition is statistically significantly better than the Skype + w/o HMD condition. However, we cannot find the statistically significant difference between the Skype + w/o HMD condition and the SL + w/o HMD condition in terms of enjoyment.

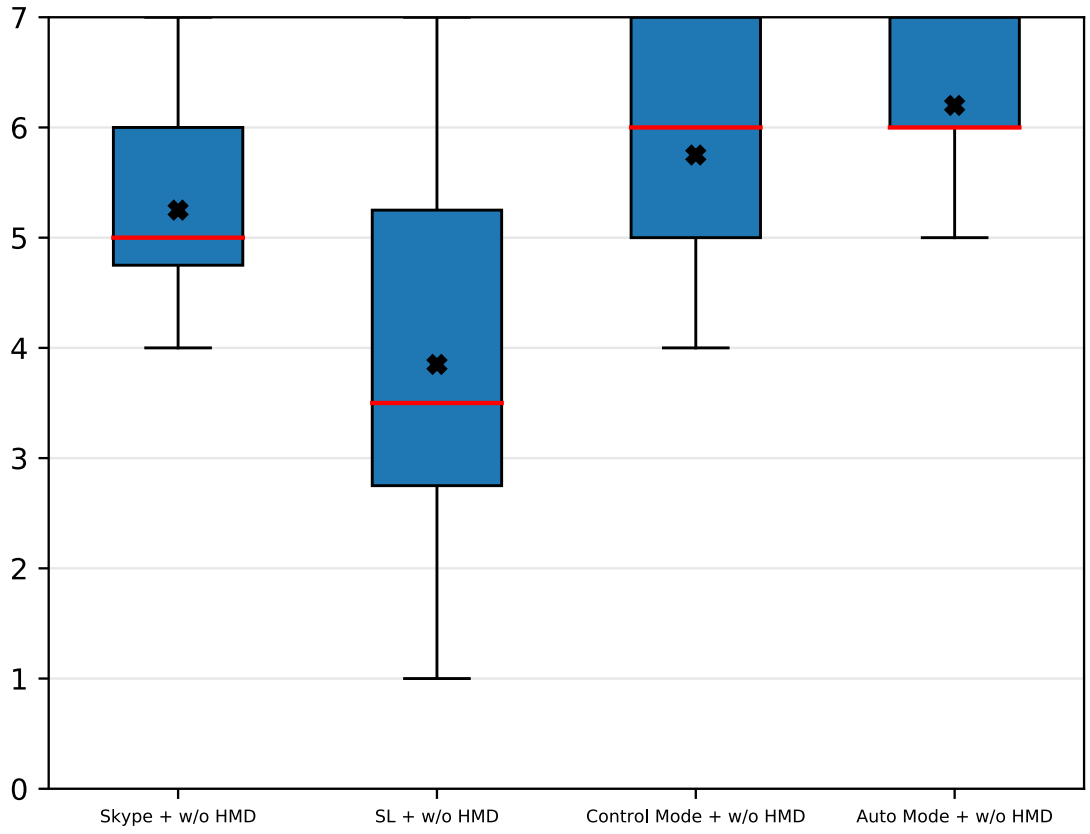

Figure 7: Box plots of the obtained scores on Question No. 1 (enjoyment). In each box plot, the x marker denotes the mean value and the red line denotes the median value.

The above results indicate that our system (in particular, the Auto Mode) can provide more enjoyment during three-party telepresence, which is better than the Skype method that directly transmits high fidelity audio/video streams for telecommunication. The participants feel more enjoyable at group-based telepresence in a shared virtual world. Finally, not surprisingly, due to the lacking of automated conversational gestures on avatars, the SL method did not perform well

with respect to the enjoyment aspect.

Table 4: The average scores and standard deviations of Question #1 (Enjoyment). The best performance is highlighted bold. For Question #1, a larger score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	5.3 ± 1.0	3.9 ± 1.8	5.8 ± 1.2	$\mathbf{6.2 \pm 0.7}$

Table 5: *p*-values of the user scores to Question #1 (Enjoyment). The statistically significant difference pairs are highlighted bold.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.384103	0.285158	**0.014335**
SL + w/o HMD	0.384103	-1.000000	**0.004217**	**0.001000**
Control Mode + w/o HMD	0.285158	**0.004217**	-1.000000	0.597695
Auto Mode + w/o HMD	**0.014335**	0.001000	0.597695	-1.000000

4.5.2 Isolation

The No. 2 question in our questionnaire concerns the isolation perception of the participants. Similar to the work of (Slater et al., 2000), we first normalized the original user responses, on a scale from 1 to 100, of question No. 2 to a scale from 1 to 7. Figure 8 illustrates the box plots of the isolation scores for four conditions (i.e., Control Mode + w/o HMD, Auto Mode + w/o HMD, SL + w/o HMD, and Skype + w/o HMD). From Figure 8, we can see that the Auto Mode + w/o HMD condition (of our system) obtains the lowest average score of isolation, and the SL + w/o HMD condition obtains the highest average score.

Table 6: The average scores and standard deviations of Question #2 (Isolation). The best performance is highlighted bold. For Question #2, a smaller score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	1.4 ± 1.1	2.8 ± 2.1	1.3 ± 1.2	$\mathbf{1.1 \pm 1.2}$

The average scores and standard deviations of Question #2 are reported in Table 6. We performed a non-parametric statistical analysis because the collected data are not normally distributed

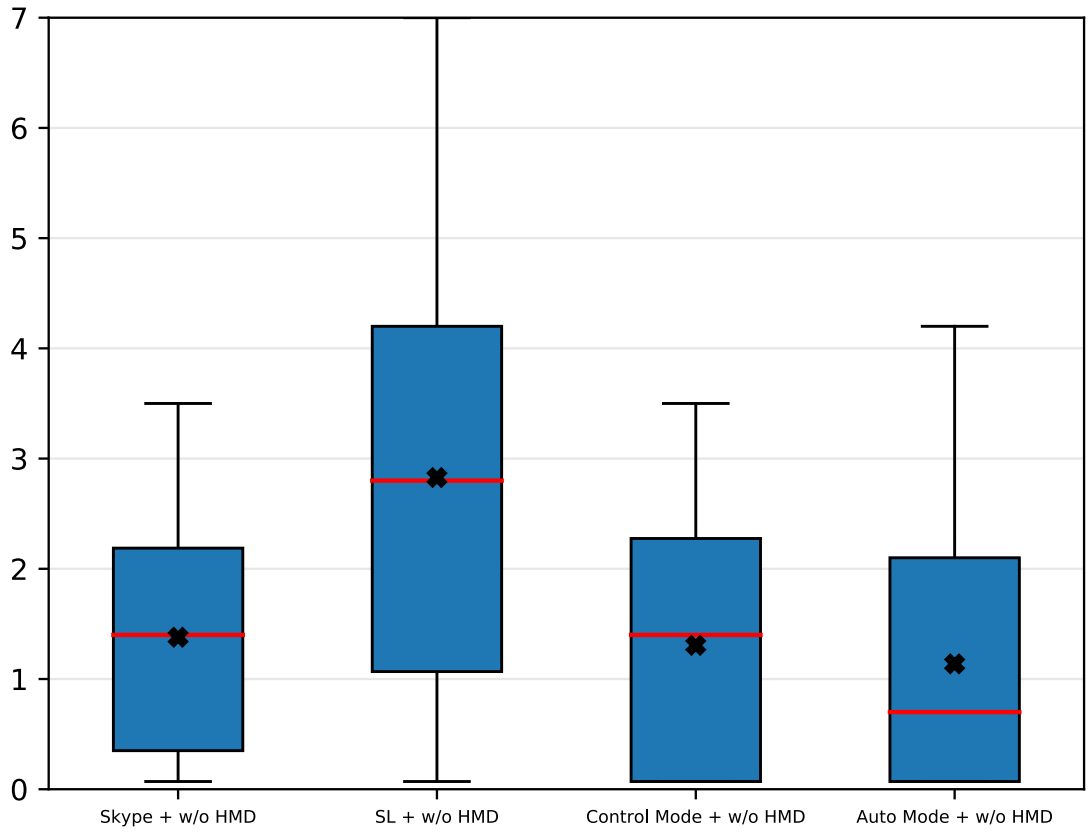

Figure 8: Box plots of the obtained scores of Question No. 2 (isolation). In each box plot, the x marker denotes the mean value and the red line denotes the median value.

Table 7: *p*-values of the user responses to Question #2 (Isolation). The statistically significant difference pairs are highlighted bold.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.316064	0.900000	0.666442
SL + w/o HMD	0.316064	-1.000000	0.091841	**0.025015**
Control Mode + w/o HMD	0.900000	0.091841	-1.000000	0.900000
Auto Mode + w/o HMD	0.666442	**0.025015**	0.900000	-1.000000

(Shapiro-Wilk test). The results by the Nemenyi post-hoc test are shown in Table 7. We found statistically significant differences for 1 pair, p-value < 0.05: Auto Mode + w/o HMD vs. SL + w/o HMD. In terms of isolation perception (i.e., lower isolation scores are better, and vice versa), the Auto Mode + w/o HMD condition (of our system) is statistically better than the SL + w/o HMD condition. However, we did not find statistically significant differences between our system (including both the Control Mode and the Auto mode) and the Skype method, and between the

Skype method and the SL method, in terms of isolation perception.

4.5.3 Meet Again

The No. 3 question in our questionnaire concerns the meet-again perception of the participants. Figure 9 illustrates the box plots of the meet-again scores for four conditions (i.e., Control Mode + w/o HMD, Auto Mode + w/o HMD, SL + w/o HMD, and Skype + w/o HD). As shown in Figure 9, in terms of the average meet-again score, our system (both Control Mode + w/o HMD and Auto mode + w/o HMD conditions) are significantly higher than that of the SL + w/o HMD condition. Also, the average score of the Skype + w/o HMD condition is higher than that of the SL + w/o HMD condition.

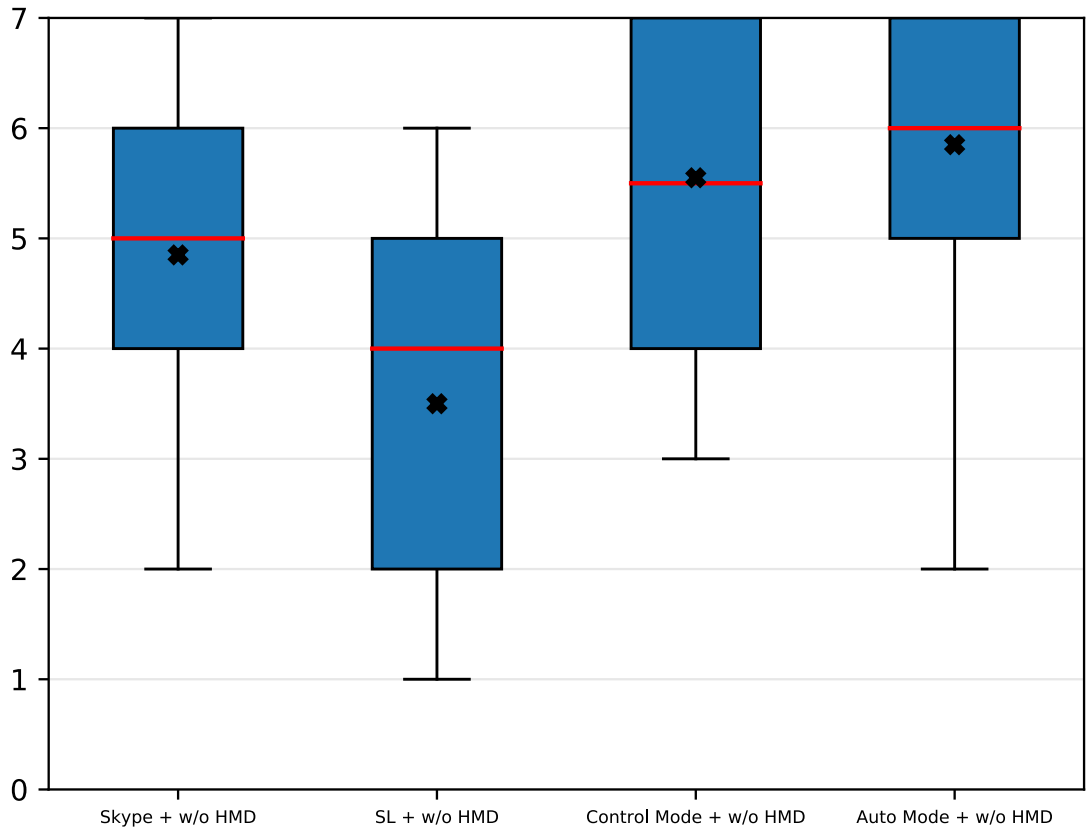

Figure 9: Box plots of the obtained user responses to Question No. 3 (meet again). In each box plot, the x marker denotes the mean value and the red line denotes the median value.

Table 8: The average scores and standard deviations of Question #3 (Meet Again). The best performance is highlighted bold. For Question #3, a larger score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	4.9 ± 1.4	3.5 ± 1.7	5.6 ± 1.4	$\mathbf{5.9 \pm 1.2}$

Table 9: *p*-values of the user responses to Question #3 (Meet Again). The statistically significant difference pairs are highlighted bold.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.159034	0.420347	0.058299
SL + w/o HMD	0.159034	-1.000000	**0.001721**	**0.001000**
Control Mode + w/o HMD	0.420347	**0.001721**	-1.000000	0.735188
Auto Mode + w/o HMD	0.058299	**0.001000**	0.735188	-1.000000

The average scores and standard deviations of Question #3 are reported in Table 8. We performed a non-parametric statistical analysis because the collected data are not normally distributed (Shapiro-Wilk test). We also performed the Nemenyi post-hoc test on the data, as shown in Table 9. From this table, we found statistically significant differences for 2 pairs, p-value < 0.05: Auto Mode + w/o HMD vs. SL + w/o HMD, and Control Mode + w/o HMD vs. SL + w/o HMD. In other words, our system (including both the Control Mode and the Auto mode) is statistically significantly better than the SL method in terms of the "meet again" effect. Also, a statistically significant difference cannot be found between the SL + w/o HMD and the Skype + w/o HMD. Compared to both our system and the Skype method, the SL method does not provide realistic conversational gestures (except manually playing back pre-created animations). This shows realistic conversational gestures can improve participants' desire to meet again in virtual environment.

4.5.4 Comfort

The No. 4 question in our questionnaire concerns the comfort perception of the participants. Figure 10 illustrates the box plots of the comfort scores for four conditions (i.e., Control Mode + w/o HMD, Auto Mode + w/o HMD, SL + w/o HMD, and Skype + w/o HMD). As shown in this figure, our system (including both the Auto Mode and the Control mode) has significantly higher

average scores than both the SL method and Skype method.

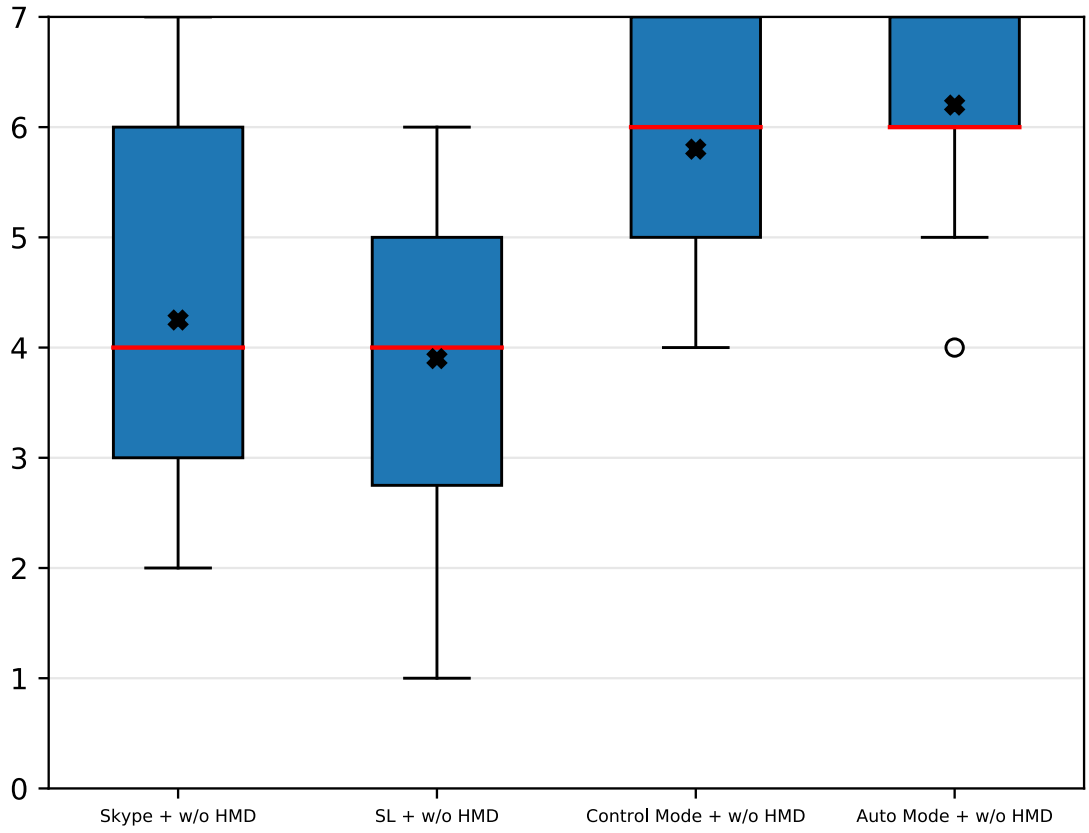

Figure 10: Box plots of the obtained response scores of Question No. 4 (comfort). In each box plot, the x marker denotes the mean value and the red line denotes the median value.

Table 10: The average scores and standard deviations of Question #4 (Comfort). The best performance is highlighted bold. For Question #4, a larger score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	4.3 ± 1.4	3.9 ± 1.6	5.8 ± 1.1	$\mathbf{6.2 \pm 0.9}$

The average scores and standard deviations of Question #4 are reported in Table 10. We performed a non-parametric statistical analysis because the collected data are not normally distributed (Shapiro-Wilk test). We also performed the Nemenyi post-hoc test on the collected user response data, as shown in Table 11. There are statistically significant differences for 4 pairs, p-value < 0.05: Control Mode + w/o HMD (of our system) vs. Skype + w/o HMD, Control Mode

Table 11: *p*-values of the user responses to Question #4 (Comfort). The statistically significant difference pairs are highlighted bold.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.900000	**0.029881**	**0.001000**
SL + w/o HMD	0.900000	-1.000000	**0.014335**	**0.001000**
Control Mode + w/o HMD	**0.029881**	**0.014335**	-1.000000	0.666442
Auto Mode + w/o HMD	**0.001000**	**0.001000**	0.666442	-1.000000

+ w/o HMD vs. SL + w/o HMD, Auto Mode + w/o HMD vs. SL + w/o HMD, and Auto Mode + w/o HMD vs. SL + w/o HMD. In other words, in terms of comfort perception, our system (including both the Control Mode and the Auto Mode) is statistically significantly better than both the SL method and the Skype method, while a statistically significant difference cannot be found between the SL method and the Skype method.

The above finding is interesting but not surprising, since, unlike the Skype method, our system does not require users to show their identities, expressions, and even their true gestures during interactions. This would make them more comfortable. By contrast, although the SL method is also one type of avatar-mediated telepresence systems, it lacks automated conversational gestures. Its average comfort score is the lowest among all the methods in the comparison.

4.5.5 Embarrassment

The No. 5 question in our questionnaire concerns the embarrassment perception of the participants. Figure 11 illustrates the box plots of the embarrassment scores for four conditions (i.e., Control mode + w/o HMD, Auto mode + w/o HMD, SL + w/o HMD, and Skype + w/o HMD). The average scores and standard deviations of Question #5 are reported in Table 12. We performed a non-parametric statistical analysis because the collected data are not normally distributed (Shapiro-Wilk test). We also performed the Nemenyi post-hoc test on the collected response scores, as shown in Table 13, and found statistically significant differences for 4 pairs, p-value < 0.05. As clearly shown in Figure 11 and Table 13, our system (including both the Control Mode and the Auto Mode) has statistically significantly lower embarrassment scores than both the SL method and the Skype method (in this case, smaller embarrassment scores are bet-

ter). We argue that the participants may feel more embarrassed at their avatars if the avatars do not have automated conversational gestures (like in the SL method). Furthermore, we argue that virtual environments can better protect the privacy of participants during telepresence, which is an advantage over video-based telecommunication methods like Skype.

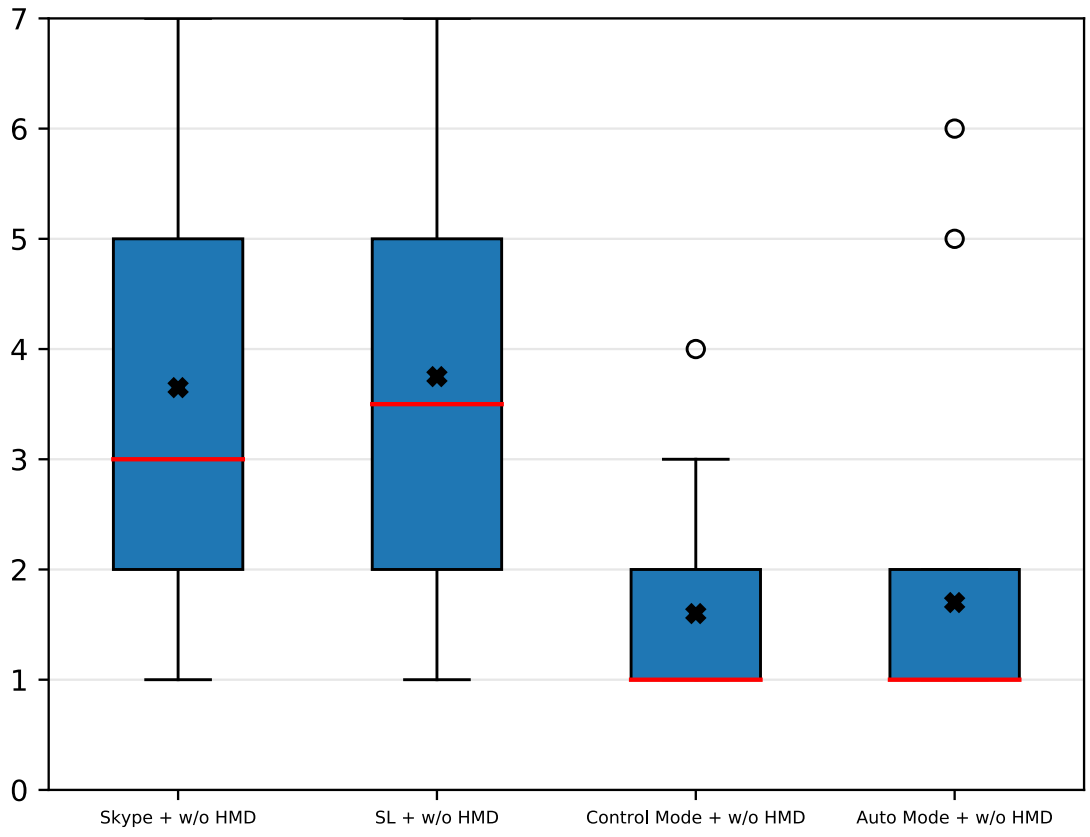

Figure 11: Box plots of the obtained response scores of Question No. 5 (Embarrassment). In each box plot, the x marker denotes the mean value and the red line denotes the median value.

Table 12: The average scores and standard deviations of Question #5 (Embarrassment). The best performance is highlighted bold. For Question #5, a smaller score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	3.7 ± 1.6	3.8 ± 2.0	$\mathbf{1.6 \pm 0.9}$	1.7 ± 1.3

Table 13: *P*-values of the user responses to Question #5 (Embarrassment). The statistically significant difference pairs are highlighted in bold.

Condition	Skype + w/o HMD	sL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.900000	**0.009685**	**0.011807**
SL + w/o HMD	0.900000	-1.000000	**0.017331**	**0.020867**
Control Mode + w/o HMD	**0.009685**	**0.017331**	-1.000000	0.900000
Auto Mode + w/o HMD	**0.011807**	**0.020867**	0.900000	-1.000000

4.5.6 Visual Comfort

Figure 12 shows the box plots of the visual comfort scores of the participants (i.e., the responses to Question No. 6 in our questionnaire). The average scores and standard deviations of Question #6 are reported in Table 14. We also performed a non-parametric statistical analysis since the collected data are not normally distributed (Shapiro-Wilk test). The Nemenyi post-hoc test was used to compute the statistically significant differences between different conditions, p-value < 0.05. As shown in Table 15, there are statistically significant differences between our system (including both the Control Mode and the Auto Mode) and the SL method. However, we cannot find a statistically significant difference between the Skype and the SL method in terms of visual comfort.

Table 14: The average scores and standard deviations of Question #6 (Visual Comfort). The best performance is highlighted bold. For Question #6, a larger score means a better performance.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Size	20	20	20	20
mean $\pm$ std	4.4 ± 1.6	3.6 ± 1.6	5.6 ± 1.3	$\mathbf{5.8 \pm 1.2}$

Table 15: *P*-values of the user responses to Question #6 (Visual Comfort). The statistically significant difference pairs are highlighted in bold.

Condition	Skype + w/o HMD	SL + w/o HMD	Control Mode + w/o HMD	Auto Mode + w/o HMD
Skype + w/o HMD	-1.000000	0.666442	0.384103	0.068212
SL + w/o HMD	0.666442	-1.000000	**0.035540**	**0.002169**
Control Mode + w/o HMD	0.384103	**0.035540**	-1.000000	0.803933
Auto Mode + w/o HMD	0.068212	**0.002169**	0.803933	-1.000000

As shown in Figure 12, in terms of the visual comfort, without counting the HMD factor, the four

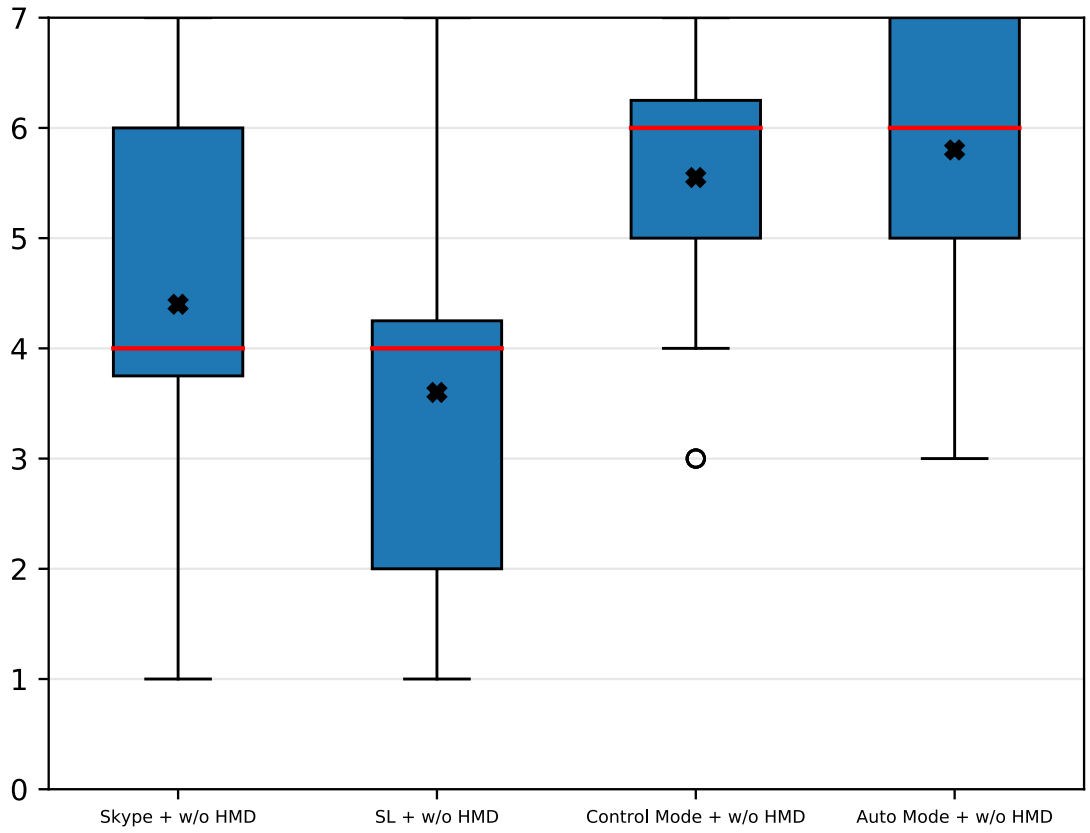

Figure 12: Box plots with the obtained response scores of Question 6 (visual comfort) for different conditions. In each box plot, the x marker denotes the mean value and the red line denotes the median value.

different conditions are sorted in the following way (from the highest to the lowest): Auto mode + w/o HMD (of our system), Control mode + w/o HMD (of our system), Skype + w/o HMD, and SL + w/o HMD. Question #6 was posed in such a way that the participants respond to this question mainly based on the face parts (in particular, head and eye motion). In our system (both the Control Mode and the Auto mode), conversational head and eye motions are automatically generated by algorithms, which facilitates to increase the visual comfort of the participants. In the Skype method, the face video streams of the other two participants are arranged horizontally; therefore, participants typically just need to have small eye movements to look at face video, with negligible head movements. This could affect their visual comfort. In the SL method, the head and the eyes are not animated unless pre-created animations are manually triggered.

4.5.7 HMD VR versus Naked Eye VR (w/o HMD)

We also compared the user experiences between w/ HMD and w/o HMD (i.e., naked eye VR) when using our system. Specifically, we compared the Control Mode + w/ HMD condition with the Control Mode + w/o HMD condition, and compared the Auto mode + w/ HMD with the Auto mode + w/o HMD condition, because the participants w/ HMD (a total of 10 participants) are independent of the participants w/o HMD (a total of 20 participants). Table 16 shows the average scores and standard deviations of the 6 questions in the Control Mode. In the Control Mode of our system, the w/o HMD option received better user responses on Q4, Q5, and Q6 than the w/ HMD option, while the w/ HMD option received better user responses on the other questions (Q1, Q2, and Q3) than w/o HMD option. Table 18 shows the average scores and standard deviations of the 6 questions in the Auto Mode. Interestingly, in the Auto Mode of our system, the w/o HMD option (i.e., naked eye VR) received better user responses than the w/ HMD option for all the six questions. This indicates that in the Auto Mode of our system, participants clearly prefer the w/o HMD option over the w/ HMD option. Also, by comparing Table 16 and Table 18, we also can see that in general the participants favored the Auto Mode of our system over its Control mode, and they also favored the naked eye VR (i.e., w/o HMD) over the HMD VR condition when using our system for telepresence.

Table 16: The average scores and standard deviations of the six questions for the Control Mode condition of our system. The bold score in each column indicates the best performance. Note that, for Q2 and Q5, a smaller score means a better performance, while for the other questions, a larger score means a better performance.

Condition	Size	Q1	Q2	Q3	Q4	Q5	Q6
Control Mode + w/o HMD	20	5.8 ± 1.2	1.3 ± 1.2	5.6 ± 1.4	$\mathbf{5.8 \pm 1.1}$	$\mathbf{1.6 \pm 0.9}$	$\mathbf{5.6 \pm 1.3}$
Control Mode + w/ HMD	10	$\mathbf{6.1 \pm 0.9}$	$\mathbf{1.2 \pm 0.9}$	$\mathbf{6.1 \pm 0.8}$	5.4 ± 1.1	2.0 ± 1.3	4.8 ± 1.3

Table 17: *P*-values of the user responses to all questions between Control Mode + w/o HMD and Control Mode + w/ HMD.

Question #	Q1	Q2	Q3	Q4	Q5	Q6
p-value	0.697313	0.927847	0.334131	0.362685	0.461690	0.113886

We also performed a non-parametric statistical analysis since the collected data are not normally

distributed (Shapiro-Wilk test). The Kruskal-Wallis H-test did not find a statistically significant difference between Control mode + w/ HMD versus Control mode + w/o HMD, and between Auto mode + w/ HMD versus Auto mode + w/o HMD, as shown in Table 17 and Table 19. These statistical results indicate that despite the participants prefer using our system w/o HMD, their user experience difference regarding with/without HMD is not statistically significant.

Table 18: The average scores and standard deviations of the six questions for the Auto Mode condition of our system. The bold score in each column indicates the best performance. Note that, for Q2 and Q5, a smaller score means a better performance, while for the other questions, a larger score means a better performance.

Condition	Size	Q1	Q2	Q3	Q4	Q5	Q6
Auto Mode + w/o HMD	20	$\mathbf{6.2 \pm 0.7}$	$\mathbf{1.1 \pm 1.2}$	$\mathbf{5.9 \pm 1.2}$	$\mathbf{6.2 \pm 0.9}$	$\mathbf{1.7 \pm 1.3}$	$\mathbf{5.8 \pm 1.2}$
Auto Mode + w/ HMD	10	5.6 ± 0.9	1.2 ± 0.7	5.8 ± 0.9	5.8 ± 1.1	2.2 ± 1.4	5.4 ± 1.0

Table 19: *p*-values of the user responses to all questions between Auto Mode + w/o HMD and Auto Mode + w/ HMD.

Question #	Q1	Q2	Q3	Q4	Q5	Q6
p-value	0.704476	0.929723	0.342968	0.371895	0.471384	0.115352

One interesting yet not completely surprising finding is that participants clearly prefer using our system over the Skype method for telecommunication, although Skype can offer stable high-quality audio/video streams. One of the main reasons could be, for various reasons participants may prefer not showing their faces, expressions, physical environments, clothes, etc. in telecommunication applications. Another possible reason is that the Skype method is not designed to construct a shared virtual meeting environment for users. By contrast, our avatar-mediated system can automatically generate plausible conversational gestures on avatars based on their real-time statuses, besides virtually putting them into a shared meeting environment. This is also consistent with the previous finding that non-verbal cues on avatars facilitate to elicit the same socio-emotional responses from human participants (Yee & Bailenson, 2007; Yee, Bailenson, Urbanek, et al., 2007).

4.5.8 Other User Feedback

After each group finished all of its sessions, we interviewed all the participants in the group to solicit their free-form comments regarding the user experience, usability, and effectiveness of all the experimental conditions. From their comments, we found that (i) most of them were more interested in our live speech-driven avatar-mediated telepresence system than other methods in the comparison. (ii) Also, some of them suggested the needed improvements of real-time generated avatar animations, including the addition of eyelid motion, finger motion, and facial expressions. (iii) Some of them also suggested the improvement of our system by allowing users to move around the avatars in the virtual environment, instead of staying at fixed locations in the current system.

5 Conclusion

In this paper, we present a live speech-driven, three-party, avatar-mediated telepresence system, where conversational gestures are real-time generated on avatars in a shared virtual meeting environment. Through a formal comparative user study, we evaluated our system by comparing it with the Skype method and the SL method, both of which have been widely used for telepresence or group meetings. Our user study results indicate that our system can measurably outperform the selected two methods in terms of enjoyability, comfort, and other user experience aspects, based on the subjective responses and feedback from participants. We argue that the real-time automated conversational gestures in our system play an important role for the improvement of user experience.

Despite the encouraging user study results, our current system still has the following limitations: (i) It is limited to three-party conversations or telepresence. Therefore, it cannot handle the cases involved with an arbitrary number of participants. (ii) To ensure the real-time performance of our system, our animation synthesis method sacrifices the quality of the synthesized motion, to a certain extent. (iii) The avatar motion by our current system does not include finger motion, eyelid motion, and facial expressions, since we directly use acoustic features to drive the motion

generation, not utilizing emotional or semantic information enclosed in human speech. Therefore, how to improve the real-time generation of high-quality conversational avatar animations is one of our future works. We also plan to explore the incorporation of avatar customization into our current system, for example, allowing users to quickly create their virtual selves with photorealistic 3D faces/bodies for avatar-mediated telepresence or teleconferencing. Finally, we are also interested in investigating how to extend or improve the current framework for more general multiparty telepresence or teleconferencing applications, such as more than three parties involved and arbitrary standing formulations.

ACKNOWLEDGMENTS

This work is in part supported by NSF IIS-1524782 and NSF IIS-2005430. Zhigang Deng was a consulting professor at the East China Jiaotong University, China.

REFERENCES

Achenbach, J., Waltemate, T., Latoschik, M. E., & Botsch, M. (2017). Fast generation of realistic virtual humans. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (p. 12:1 - 12:10). ACM.

Ad Alternum Game Studios. (2022). *Orbusvr,* `https://orbusvr.com`.

Aseeri, S. A., & Interrante, V. (2018). The influence of avatar representation and behavior on communication in social immersive virtual environments. In *Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces 2018* (p. 823-824). IEEE.

Barakonyi, I., Fahmy, T., & Schmalstieg, D. (2004). Remote collaboration using augmented reality videoconferencing. In *Proceedings of Graphics interface 2004* (p. 89-96). Canadian Human-Computer Communications Society.

Basu, A., Raij, A., & Johnsen, K. (2012). Ubiquitous collaborative activity virtual environments. In *CSCW'12: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (p. 647-650). ACM.

Bente, G., Rüggenberg, S., Krämer, N. C., & Eschenburg, F. (2008). Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human Communication Research*, *34*(2), 287-318.

Bigscreen inc. (2022). *Bigscreen,* `https://www.bigscreenvr.com`.

Blascovich, J., & Bailenson, J. (2011). *Infinite reality: Avatars, eternal life, new worlds, and the dawn of the virtual revolution*. William Morrow & Co.

Bonito, J. A., Burgoon, J. K., & Bengtsson, B. (1999). The role of expectations in human-computer interaction. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work 1999* (p. 229-238). ACM.

Cassell, J., Sullivan, J., Churchill, E., & Prevost, S. (2000). *Embodied conversational agents*. MIT press.

Chia, S. (1992). The universal mobile telecommunication system. *IEEE Communications Magazine*, *30*(12), 54–62.

Cho, S., Kim, S.-w., Lee, J., Ahn, J., & Han, J. (2020). Effects of volumetric capture avatars on social presence in immersive virtual environments. In *Proceedings of IEEE Conference on Virtual Reality and 3D User Interfaces 2020* (p. 26-34). IEEE.

Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the cave. In *SIGGRAPH'93: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques siggraph '93* (p. 135-142). ACM.

Deng, Z., Lewis, J. P., & Neumann, U. (2005). Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications*, *25*(2), 24–30.

Deng, Z., & Noh, J. (2008). Computer facial animation: A survey. In *Data-Driven 3D Facial Animation* (pp. 1–28). Springer.

Divorra, O., Civit, J., Zuo, F., Belt, H., Feldmann, I., Chreer, O., ... others (2010). Towards 3d-aware telepresence: Working on technologies behind the scene. In *Proceedings of ACM Conference on Computer Supported Cooperative Work 2010, New Frontiers in Telepres-*

ence. ACM.

Dou, M., Shi, Y., Frahm, J.-M., Fuchs, H., Mauchly, B., & Marathe, M. (2012). Room-sized informal telepresence system. In *Proceedings of IEEE Virtual Reality 2012 Workshops* (p. 15-18). IEEE.

Gibbs, S. J., Arapis, C., & Breiteneder, C. J. (1999). Teleport-towards immersive copresence. *Multimedia Systems*, *7*(3), 214-221.

Goolcharan, B., & Karunasiri, T. R. (2000, May 16). *Telecommunication system for broadcast quality video transmission.* Google Patents. (US Patent 6,064,422)

Gross, M., Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., ... others (2003). Blue-c: a spatially immersive display and 3d video portal for telepresence. *ACM Transactions on Graphics*, *22*(3), 819-827.

Gulz, A. (2005). Social enrichment by virtual characters-differential benefits. *Journal of Computer Assisted Learning*, *21*(6), 405-418.

Harrison, C., & Hudson, S. E. (2008). Pseudo-3d video conferencing with a generic webcam. In *Proceedings of Tenth IEEE International Symposium on Multimedia 2008* (p. 236-241). IEEE.

Immersed Inc. (2022). *Imersedvr,* `https://immersed.com`.

Jacobson, A., Deng, Z., Kavan, L., & Lewis, J. P. (2014). Skinning: real-time shape deformation. In *ACM SIGGRAPH 2014 Courses* (p. 24:1 - 24:95). ACM.

Jin, A., Deng, Q., Zhang, Y., & Deng, Z. (2019). A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *2*(2), 9:1 - 9:19.

Johnson, T., Gyarfas, F., Skarbez, R., Towles, H., & Fuchs, H. (2007). A personal surround environment: Projective display with correction for display surface geometry and extreme lens distortion. In *Proceedings of IEEE Virtual Reality Conference 2007* (p. 147-154). IEEE.

Johnson, T., Welch, G., Fuchs, H., La Force, E., & Towles, H. (2009). A distributed cooperative

framework for continuous multi-projector pose estimation. In *Proceedings of IEEE Virtual Reality Conference 2009* (p. 35-42). IEEE.

Kauff, P., & Schreer, O. (2002). An immersive 3d video-conferencing system using shared virtual team user environments. In *Proceedings of the 4th International Conference on Collaborative Virtual Environments 2002* (p. 105-112). ACM.

Kauff, P., & Schreer, O. (2005). Immersive videoconferencing. In *3D Videocommunication (O. Schreer, P. Kauff, T. Sikora, Editors).* John Wiley and Sons.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260-267.

Ku, J., Jang, H. J., Kim, K. U., Kim, J. H., Park, S. H., Lee, J. H., ... Kim, S. I. (2005). Experimental results of affective valence and arousal to avatar's facial expressions. *CyberPsychology & Behavior*, *8*(5), 493-503.

Latoschik, M. E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., & Botsch, M. (2017). The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (p. 39:1 - 39:10). ACM.

Le, B. H., Ma, X., & Deng, Z. (2012). Live speech driven head-and-eye motion generators. *IEEE Transactions on Visualization and Computer Graphics*, *18*(11), 1902–1914.

Le, B. H., Zhu, M., & Deng, Z. (2013). Marker optimization for facial motion acquisition and deformation. *IEEE Transactions on Visualization and Computer Graphics*, *19*(11), 1859–1871.

Levine, S., Krähenbühl, P., Thrun, S., & Koltun, V. (2010). Gesture controllers. *ACM Transactions on Graphics*, *29*(4), 124:1 - 124:11.

Lewis, J. P., Anjyo, K., Rhee, T., Zhang, M., Pighin, F. H., & Deng, Z. (2014). Practice and theory of blendshape facial models. In *Proceedings of Eurographics 2014 STAR (State of the Art Reports)* (p. 199-218). Eurographics Association.

Lincoln, P., Nashel, A., Ilie, A., Towles, H., Welch, G., & Fuchs, H. (2009). Multi-view lenticular display for group teleconferencing. In *Proceedings of the 2nd International Conference*

on Immersive Telecommunications 2009 (p. 1-8). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Ma, L., & Deng, Z. (2019). Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games* (pp. 1–10). ACM.

Maloney, D., Freeman, G., & Wohn, D. Y. (2020). "talking without a voice": Understanding non-verbal communication in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 175:1 - 175:25.

Meta Platforms, Inc. (2022). *Horizon worlds.* https://www.oculus.com/facebook-horizon/.

Microsoft Inc. (2022). *Altspacevr.* https://altvr.com.

Mozilla Foundation. (2022). *Mozilla hub,* https://www.mozilla.org/en-US.

Neos VR Metaverse. (2022). *Nerovr,* https://neos.com.

Nichol, J., & Wong, M. S. (2005). Modeling urban environmental quality in a tropical city. *Landscape and Urban Planning*, *73*(1), 49–58.

Pandzic, I., Ostermann, J., & Millen, D. (1999). Users evaluations: synthetic talking faces for interactive. *The Visual Computer*, *15*, 330-340.

Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., & Fuchs, H. (1998). The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *SIGGRAPH'98: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (p. 179-188). ACM.

Rauthenberg, S., Graffunder, A., Kowalik, U., & Kauff, P. (1999). Virtual shop and virtual meeting point-two prototype applications of interactive services using the new multimedia coding standard MPEG-4. In *Proceedings of the International Conference on Computer Communication 1999* (p. 1-3).

Rec Room. (2016). *Rec room.* https://recroom.com.

Rincón-Nigro, M., & Deng, Z. (2013). A text-driven conversational avatar interface for instant messaging on mobile devices. *IEEE Transactions on Human-Machine Systems*, *43*(3),

328–332.

Rizzo, A. A., Neumann, U., Enciso, R., Fidaleo, D., & Noh, J. (2001). Performance-driven facial animation: basic research on human judgments of emotional state in facial avatars. *CyberPsychology & Behavior*, *4*(4), 471-487.

Roth, D., Lugrin, J.-L., Galakhov, D., Hofmann, A., Bente, G., Latoschik, M. E., & Fuhrmann, A. (2016). Avatar realism and social interaction quality in virtual reality. In *Proceedings of IEEE Virtual Reality 2016* (p. 277-278). IEEE.

Ruhland, K., Andrist, S., Badler, J., Peters, C., Badler, N., Gleicher, M., ... Mcdonnell, R. (2014). Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Proceedings of Eurographics 2014 STAR (State of the Art Reports)* (pp. 69–91). Eurographics Association.

Schuemie, M. J., Van Der Straaten, P., Krijn, M., & Van Der Mast, C. A. (2001). Research on presence in virtual reality: A survey. *CyberPsychology & Behavior*, *4*(2), 183–201.

Singh, K., Ohya, J., & Parent, R. (1995). Human figure synthesis and animation for virtual space teleconferencing. In *Proceedings Virtual Reality Annual International Symposium'95* (pp. 118–126). IEEE.

Slater, M., McCarthy, J., & Maringelli, F. (1998). The influence of body movement on subjective presence in virtual environments. *Human Factors*, *40*(3), 469–477.

Slater, M., Sadagic, A., Usoh, M., & Schroeder, R. (2000). Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, *9*(1), 37–51.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*(2), 97-124.

Ståhl, O. (1999). Meetings for real—experiences from a series of vr-based project meetings. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology 1999* (p. 164-165). ACM.

Vannucci, G. (1995, October 17). *Wireless telecommunication system.* Google Patents. (US Patent

5,459,727)

VRChat Inc. (2022). *Vrchat.* https://hello.vrchat.com.

vTime Holdings Limited. (2022). *vtime xr.* https://vtime.net.

Wen, W.-C., Towles, H., Nyland, L., Welch, G., & Fuchs, H. (2000). Toward a compelling sensation of telepresence: Demonstrating a portal to a distant (static) office. In *Proceedings of IEEE Visualization 2000* (p. 327-333). IEEE.

Wild Technology Inc. (2022). *The wild,* https://thewild.com.

Williams, W., & Libove, J. (1999, March 16). *Method and apparatus for increased quality of voice transmission over the internet.* Google Patents. (US Patent 5,883,891)

Yang, R., Kurashima, C. S., Towles, H., Nashel, A., & Zuffo, M. K. (2007). Immersive video teleconferencing with user-steerable views. *Presence: Teleoperators and Virtual Environments*, *16*(2), 188-205.

Yee, N., & Bailenson, J. N. (2007). The proteus effect: Self transformations in virtual reality. *Human Communication Research*, *33*(3), 271-290.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2007* (p. 1-10). ACM.

Yee, N., Bailenson, J. N., Urbanek, M., Chang, F., & Merget, D. (2007). The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *CyberPsychology & Behavior*, *10*(1), 115-121.

Yoo, S. K., Kim, D., Jung, S., Kim, E., Lim, J., & Kim, J. (2004). Performance of a web-based, realtime, tele-ultrasound consultation system over high-speed commercial telecommunication lines. *Journal of Telemedicine and Telecare*, *10*(3), 175–179.