frontiers in Virtual Reality

ORIGINAL RESEARCH
published: 16 August 2021
doi: 10.3389/frvir.2021.698523

# Spatial Perception Enhancement in Assembly Training Using Augmented Volumetric Playback

*Prasanth Sasikumar*, Soumith Chittajallu, Navindd Raj, Huidong Bai and Mark Billinghurst*

*The University of Auckland, Auckland, New Zealand*

Conventional training and remote collaboration systems allow users to see each other's faces, heightening the sense of presence while sharing content like videos or slideshows. However, these methods lack depth information and a free 3D perspective of the training content. This paper investigates the impact of volumetric playback in a Mixed Reality (MR) spatial training system. We describe the MR system in a mechanical assembly scenario that incorporates various instruction delivery cues. Building upon previous research, four spatial instruction cues were explored; "Annotation", "Hand gestures", "Avatar", and "Volumetric playback". Through two user studies that simulated a real-world mechanical assembly task, we found that the volumetric visual cue enhanced spatial perception in the tested MR training tasks, exhibiting increased co-presence and system usability while reducing mental workload and frustration. We also found that the given tasks required less effort and mental load when eye gaze was incorporated. Eye gaze on its own was not perceived to be very useful, but it helped to compliment the hand gesture cues. Finally, we discuss limitations, future work and potential applications of our system.

**Keywords: MR training, remote collaboration, presence enhancement, augmented reality, volumetric playback, hand gestures, eye gaze**

OPEN ACCESS

***Edited by:***
*Youngho Lee,*
*Mokpo National University,*
*South Korea*

***Reviewed by:***
*Myungho Lee,*
*Pusan National University, South Korea*
*SeungJun Kim,*
*Gwangju Institute of Science and Technology, South Korea*

****Correspondence:***
*Prasanth Sasikumar*
*prasanth.sasikumar.psk@gmail.com*

***Specialty section:***
*This article was submitted to Augmented Reality, a section of the journal Frontiers in Virtual Reality*

***Received:*** *21 April 2021*
***Accepted:*** *19 July 2021*
***Published:*** *16 August 2021*

***Citation:***
*Sasikumar P, Chittajallu S, Raj N, Bai H and Billinghurst M (2021) Spatial Perception Enhancement in Assembly Training Using Augmented Volumetric Playback. Front. Virtual Real. 2:698523. doi: 10.3389/frvir.2021.698523*

## 1 INTRODUCTION

The fourth industrial revolution is fundamentally changing the way we live, work, and relate to one another (Xu et al. (2018)). With the advent of the Internet of Things, robots and automation are being used to enhance product quality and reduce the cost of training and maintenance procedures. However, even though robots are used to automate mundane, repetitive tasks, there is still a need for human expert intervention to work on challenging problems. Collaborative Mixed Reality (MR) systems are one of the technologies that can address these challenges (Billinghurst and Kato (1999)).

Unlike using Virtual Reality (VR) solutions where the view of the physical surroundings is blocked, one of MR tool's main benefits is that people can train in and interact with the real world while referring to overlaid 3D information. MR techniques could be employed to improve key operations performed in a factory, including training or maintenance activities.

All techniques on the MR continuum (Milgram and Kishino (1994)) share the basic properties of being computer-generated, interactive, three-dimensional, and rendered in real-time, and they allow for the development of applications with an enhanced sense of presence (Regenbrecht et al. (2017)).

Most MR training systems have explored the use of visual cues in the forms of graphic objects or text to deliver instructions. For example, Yang et al. (2020) compared the impact of spatial audio in remote collaboration among various visual cues like hand gestures, annotation and eye gaze. There

has also been some research using point-clouds to visualize instructional cues from trainers, but it is not well studied (Schwald and de Laval (2003); Ke et al. (2005); Olwal et al. (2008); Alem and Li (2011)). Regenbrecht et al. (2017) used voxels for the visual representation of collaborators in MR and showed that better visualization of spatial data could improve MR remote collaboration.

This paper explores how different representations of MR instructional cue's can affect performance and user experience in real-world assembly tasks. We present a MR remote collaboration system that features four different visual cues: 3D annotations, hand gestures, full-body avatars, and full-body volumetric playback. We conducted a user study to compare and evaluate the effectiveness and usability of these cues.

Simulating a real-world assembly task in the industry, we designed a motorcycle engine assembly task for participants to complete. The actions of an expert demonstrating the assembly tasks were pre-recorded: The annotations, hand gestures, skeletal movements, and volumetric data of the expert were recorded and replayed. The participants used a VR head-mounted display (HMD) to watch instructions for assembly tasks with different spatial visual cues, and were asked to perform the tasks themselves. We expanded the system to incorporate eye gaze and conducted a second user study to evaluate this system.

The main contributions of this paper include: 1) A MR training system that uses volumetric playback to deliver instructions; 2) A user study that explores the usability of the volumetric playback cues in an MR system compared to traditional visual cues; 3) A user study that compares the usability of eye gaze and hand gestures in an MR training system.

## 2 RELATED WORK

Experienced service technicians who have accumulated invaluable knowledge over years of work can precisely handle machine faults. Unfortunately, this knowledge is often restricted to only a few people and is rarely documented. Even when there is documentation, it is often incomplete and does not demonstrate the skills and knowledge accumulated by the technician. However, Augmented Reality (AR) and Virtual Reality (VR) can be used to capture the actions of experienced technicians and play them back in a way that enables a novice to learn from them.

One of the main advantages of AR/VR training systems is that they are capable of facilitating independent viewpoints into a collaborative task space (Billinghurst and Kato (2007)). This way, the trainee can interact with real-world objects while simultaneously accessing virtual information for guidance, thereby creating a mapping between the training and the real task. This adds another dimension to remote training support that is traditionally being provided via video conferencing software like Zoom.[1] or Skype.[2].

[1]https://zoom.us/

[2]https://www.skype.com/en/

Numerous studies (Schwald and de Laval (2003); Ke et al. (2005); Olwal et al. (2008); Webel et al. (2013); De Pace et al. (2019)) have investigated the potential of AR-based training systems in the context of guidance and maintenance tasks. These systems have employed 3D-reconstruction, user-behavior modeling and tracking, multi-modal interaction, and 3D-interactive graphics for distributed training. In traditional MR remote systems, the instructions are often delivered using visual cues like Annotation, Hand Gestures, and Avatar representation. These are discussed in detail below:

### 2.1 Annotation

Incorporating spatial references into remote collaboration or telepresence systems has been an actively investigated topic. Many researchers have studied bringing remote pointers into a collaborators display space either on a screen (Fussell et al. (2004)), via projection (Gurevich et al. (2012)) or in an HMD(Bauer et al. (1999); Sasikumar et al. (2019)). Annotating physical objects is an important user interaction task in AR (Wither et al. (2009)). Although annotation is widely used in AR, there is no general agreed-upon definition of what precisely constitutes an annotation. In this context, we take computer-generated lines drawn in 3D space as an example of annotation. Rose et al. (1995) used AR for annotation to provide information for engine mechanics. This desktop-AR system allowed the mechanic to freely rotate the virtual engine, and the parts of the engine visible to the viewer were annotated by text. Chang et al. (2017) evaluated gesture-based AR annotation and found a preference among participants for beautified annotations. Since our work focuses on comparing the instruction cues instead of comparing annotations with other visualization forms, we did not provide beautified annotations in our system.

### 2.2 Hands

To evaluate the effectiveness of using an augmented tele-pointer in wearable video conferencing system, Bauer et al. (1999) conducted an experimental study involving pairs of users performing a set of artificial tasks. Analysis of verbal communication behavior and pointing gestures in the experiment determined that experts pointed substantially more than verbal instructions for guiding workers through the physical tasks. The use of pointers reached 99%, while in 20% of cases, experts did not use verbal instructions at all. Pairs relying almost exclusively on tele-pointing displayed the fastest completion times. This provides a solid ground for more extensive research into the field. It is very clear that gesturing is an effective means for enhancing remote collaboration between users for physical tasks. Previous studies (Ou et al. (2003); Alem and Li (2011); Kirk et al. (2007)) have also found that participants experienced a higher quality of collaboration using overlaying hands than a cursor pointer, describing their interactions as "more transparent" when seeing their partners hands. Our system builds upon this research to provide hand representation as a visual cue for training tasks.

## 2.3 Avatars

Previous studies have found that representing body information improved social presence in remote collaboration (Smith and Neff (2018); Wang et al. (2020)). Further studies have also indicated that such a body, even if simple, heightens the sense of presence (Slater and Usoh (1994)). The use of Avatars for real-time communication in computer-generated environments has been used in many applications (Kalra et al. (1998); Yang et al. (2002)). For example, Carrasco et al. (2017) compared avatar representation preference across age groups and found that older adults prefer attractive avatars with expressive features. This influenced the age group of the avatar model that was chosen in the study.

Considerable research has been done regarding how the appearance of avatars influences communication and interaction Heidicker et al. (2017). A study by Mohler et al. (2008) concluded that a full-body avatar improves egocentric distance judgments in an immersive virtual environment, however using an avatar that consists only of head and hands was not significantly worse than using a complete avatar body with pre-defined animations.

## 2.4 Volumetric Playback

The role of volumetric playback in enhancing the sense of presence in immersive virtual environments have been studied (Cho et al. (2020)). With the availability of affordable depth cameras, a large number of volumetric capture and playback solutions like KinectFusion have emerged (Izadi et al. (2011); Newcombe et al. (2015)). Machine learning techniques have been used to predict skeletal joint positions from single RGB video (Dou et al. (2017); Habermann et al. (2019)). Use of Voxels (gap-less volumetric pixels in a regular grid in space) is another technique that was studied to enhance visual coherence. Researchers have tried several methods to capture and reconstruct the human body in 3D space (Hasenfratz et al. (2004); Jung and Bajcsy (2006)). Using Holoportation, users wearing virtual or augmented reality displays could interact with remote participants in 3D, almost as if they were present in the same physical space (Orts-Escolano et al. (2016b)).

## 2.5 Eye Gaze

Several researchers have explored the impact of sharing eye gaze for collaboration (Gupta et al. (2016); Bauer et al. (1999); Yang et al. (2020); Piumsomboon et al. (2017)). For example, Špakov et al. (2019) investigated sharing of visual focus between two players in a collaborative VR game so that one player would know where the other one was looking. The study aimed to determine whether there is an added value of eye gaze in the context of collaborative games. They found that teamwork ratings were higher from pairs using eye gaze in the VR environment than those using head gaze. This suggests that eye gaze provides a better collaborative game experience than head gaze, but Spakov states that further investigation is required to answer this question properly. We build upon this as the motivation for the second user study.

As this research shows, various techniques have been studied to enhance different aspects of remote collaboration, like visual coherence, co-presence and usability. However, there has been no study comparing the visual cues of annotations, hands, avatars, eye gaze, and volumetric reconstruction in an AR training environment. Our research addresses this important gap. In the next section, we describe the prototype training system that we developed to conduct the user study.

# 3 SYSTEM OVERVIEW AND IMPLEMENTATION

We developed a system capable of recording and replaying instructions using four different visual cues. The prototype system was built with: an HMD (HTC Vive Pro Eye .[3]), three depth cameras (Azure Kinect), and Unity.[4] running on a windows PC. For the assembly task, we used a motorcycle engine from a 2008 Hyosung GT 250R.[5]. The tasks were inspired by general engine maintenance procedures described in the service manual of the motorcycle.

We opted to use the see-through video capabilities of the HTC VIVE instead of an optical see-through AR headset as we required a very high level of precision in tracking over a large area. Additionally, portable HMD devices were ruled out as volumetric playback requires a large amount of computational resource that is not available in most mobile devices.

For recording instructions, the instructor performed real-world tasks while wearing the HMD, and the actions were saved onto the PC. Annotations were recorded by moving the HTC Vive handheld controller along the desired path while pressing the trigger button. The controller's time and position were saved and played back in time-space synced format to recreate the instruction. Similarly, for recording gestures, the instructor's hand movements were captured using the front-facing cameras of Vive HMD, which were played back to recreate the instruction. To record the Avatar representation, the instructor wore the HMD while holding the left and right controllers. The position and rotation of HMD and controllers were saved along with time while the instructions were performed. These values were applied to a skeleton rigged with inverse kinematics to create a virtual avatar representation. For volumetric capture, the Azure Kinect.[6] cameras were used. **Figure 1** shows an overview of the training system framework. The instructor performs the task while being captured by three Azure Kinect cameras placed 1.2 m apart along the vertices of an equilateral triangle to provide optimum coverage, as shown in **Figure 2**.

The annotations and hand gestures were captured and recorded using the SR run-time library support that HTC provides for the VIVE Pro. After performing each task, the recordings were saved to the local file system as a Unity asset.

[3]https://www.vive.com/eu/product/vive-pro-eye/overview/
[4]https://unity.com/
[5]https://en.wikipedia.org/wiki/Hyosung_GT250/
[6]https://azure.microsoft.com/en-us/services/kinect-dk/

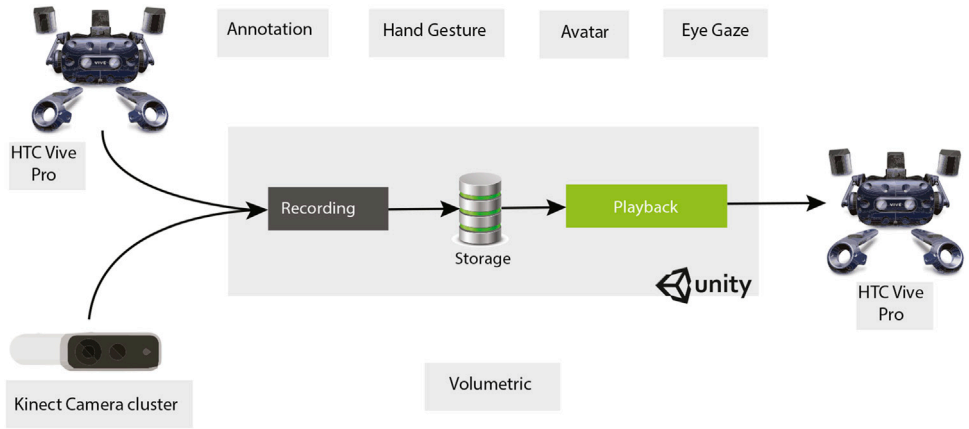


**FIGURE 1 |** The overview of our MR training system. Volumetric Capture is facilitated by Kinect Camera cluster and all other visual cues are captured and played back using HTC Vive.

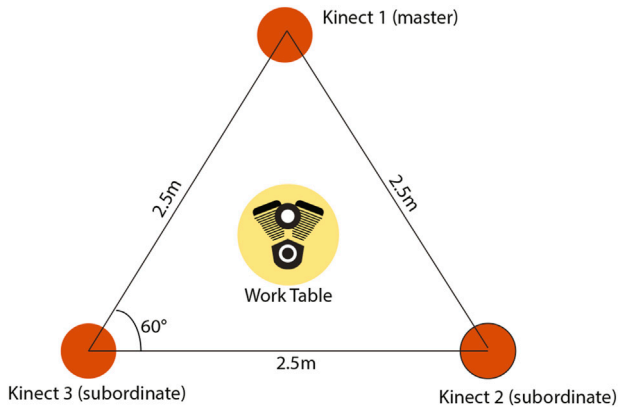


**FIGURE 2 |** Studio setup for volumetric capture. Three Azure Kinect cameras are kept 60° apart in the form of an equilateral triangle.

The recordings stored information regarding the position and rotation of corresponding coordinates and timestamp in milliseconds. Similarly, for Avatar capture and playback, the system saved the position and rotation information of the expert's HMD and controllers. We used inverse kinematics to provide natural movement to the avatar representation. For the volumetric capture and playback, one Kinect acted as the master, and the rest as subordinates. This hardware synchronization ensured a smooth playback when the recordings were stitched together.

For the study, we used a PC equipped with Intel Core i7-8,700 3.2 GHz CPU with 6 Cores, 32 GB DDR4 RAM, NVIDIA GeForce GTX 1080 GPU, running Windows 10. We converted the color and depth frames into a point-cloud using the Unity's HDR pipeline and visual graph library. This point cloud data from three individual sources was merged together under one coordinate system to reconstruct the work-space as a live 3D panorama using the calibration method mentioned in Section 3.1.

The software was developed using the Unity 3D game engine (2019.3.2f1), and the official Azure Kinect unity plugin was used to stream the Kinect data feed to the Unity system. Additionally, an image processing Unity plugin was coded in C++ for processing and stitching the dense point-cloud data. This framework allowed us to rapidly prototype an MR training system that supported various visual communication cues. However, one limitation is that the depth camera's stitched depth image was slightly noisy around the stitched surfaces.

To perform the assembly task, the participant wore the HMD and followed the pre-recorded set of instructions using either of the four visual communication cues. Annotations appeared as spatially aligned drawings starting at the object to be picked up and ending where it needed to be placed, as shown in **Figure 3**. A directional arrow would be drawn where an action was to be performed. In the case of hand gestures, the instructor expressed the action that needed to be performed - expressing how to pick and place the object of interest, as shown in 4. Similarly, for the avatar representation, the character would start at the object that needed to be picked up, move towards the location of placement and perform the action using hand movements, as shown in 5. Volumetric capture was spatially aligned for playback as shown in **Figure 4**.

For the assembly task, the motorcycle engine was placed at the center of a round table, and the required tools were placed around the engine, as shown in **Figure 5**. We kept additional tools around the table to increase task complexity and mimic the real-world scenario. The participant was able to move around the table to perform the tasks.

A total of 15 basic motorcycle engine maintenance tasks were selected from the service manual. We conducted an informal trial with three participants. The participants performed all fifteen tasks in all four conditions and the completion time was measured for each of the tasks. Participants also rated the tasks based on difficulty level. Analyzing the difficulty and completion time, we found that the difficult tasks tended to take the longest time. Based on this, we narrowed the task set down to nine tasks that could be classified into three hard, three

**FIGURE 3 | (A)** Annotation and **(B)** Hand Gesture from the user's perspective.

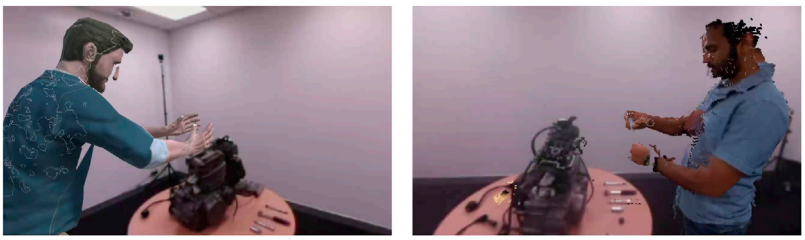

**FIGURE 4 | (A)** Avatar and **(B)** Volumetric visual cue from the user's perspective.

medium, and three easy tasks. This classification was based on the number of individual actions required to complete the task. For example, in a medium difficulty task, the user takes the oil filler cap from the table and places it in its corresponding location. Whereas in one of the hard tasks, the user has to remove the banjo bolt that connects the radiator feed to the engine and thread it through the radiator feed, tightening it back to the engine. The final list of tasks is described in **section 4.2**. In any given condition, participants would receive a random subset of six instructions–two from each of the three categories. After completing an instruction, audio and visual confirmation was provided, after which the system moved on to the next instruction. This process was repeated until all of the instructions were completed and the completion time was logged at the end of each condition. In the next section, we present the user study conducted with our prototype system.

# 4 USER STUDY

## 4.1 Experimental Design

We were interested in evaluating the impact of full-body representation in the training system. We conducted a formal user study to explore the usability of visual cues described in Section 3.2. We used the annotation and hand gestures as visual cues for non-body communication and avatar and volumetric capture for full-body representation. In this case, our primary independent variable was the type of natural communication cues shared from the remote to the local user. Time was the dependent variable between the conditions. Altogether, we had the four communication conditions A1–A4 as follows:

1) Annotation (A1). Instructions appear like a drawing in 3D space aligned to the component that needs to be interacted with. In an assembly task, the drawing typically starts from the location of the component. The path drawn is the ideal path of moving and would terminate by pointing to the destination where it needs to be assembled.
2) Hand Gesture (A2). The virtual hand gestures of the instructor appear to guide the user. The instructor takes care to make sure the gestures are expressive.
3) Avatar (A3). A virtual avatar representation that imitates the movements involved in the task is present next to the user.
4) Volumetric Playback (A4). Spatially aligned virtual volumetric playback of the instructor performing the instruction.

In the user study, we investigated the following two research questions:

- RQ1: Would volumetric playback increase the sense of co-presence in instruction delivery?
- RQ2: Would volumetric playback reduce the completion time for a task and how it compares with simple (A1 and A2) and full-bodied instruction (A3 and A4) delivery cues?

Our research hypotheses were:

- H1. Volumetric playback would provide a better sense of social presence in a remote training system.

- H2. Volumetric playback would enable faster completion of tasks in a remote training system.

It is important to note that full-bodied augmented visual cues have different properties compared to non-embodied cues. Full-bodied cues increase the sense of presence (Adcock et al. (2013); Pejsa et al. (2016); Orts-Escolano et al. (2016a); Joachimczak et al. (2017); Piumsomboon et al. (2018)) but at the same time occupy a larger volume of instruction space. The effect of this has not been studied before in an augmented visual instruction delivery setting.

## 4.2 Study Environment and Experimental Task

As shown in **Figure 5**, a closed office space was used for the user study. The participant performed tasks in a (3 m × 3 m) marked area. The engine was placed on top of a round desk with a diameter of 0.8 m (**Figure 5**). Although the studies were carried out during office hours, the space was isolated to avoid any distractions. The prototype, as described in **section 3**, was used for the user study. We gathered participants from around the university and they did not need any skills or knowledge of mechanical assembly.

For the experiment, we chose three hard, three medium, and three easy tasks. The participant would be doing six tasks (two tasks each from a category) in all the conditions. The tasks were as follows:

1) Assembling the ignition coil on top of the front spark plug [Medium]
2) Disassembling the rear ignition coil from the rear spark plug [Hard]
3) Removing the rocker rear rocker cover [Medium]
4) Assembling the oil filler cap [Hard]
5) Assembling rear top engine mount screw [Easy]
6) Assembling rear bottom engine mount screw [Easy]
7) Assembling the three-piece spark plug socket wrench (socket + extension + wrench) [Easy]
8) Removing the rear spark plug using the spark plug wrench [Medium]
9) Removing the banjo bolt that connects the radiator feed to the engine and threads it through the radiator feed, followed by tightening it back to the engine [Hard]

Based on the criteria discussed in **section 3**, the first three tasks were classified as easy, the next three medium, and the last three were classified as hard tasks.

## 4.3 Experimental Procedure

The experiment began with the participants signing a consent form, answering demographic questions, and describing their experience with VR/AR. Participants were then shown how to perform each task before beginning the experiment. This was done to compensate for any advantage that the mechanically minded participants would have. A random instruction set was picked for each condition. After finishing each trial and at the end of the experiment, they evaluated their experience and provided qualitative feedback about the user experience and system in general. The study took about 1 hour on average to complete.

## 4.4 Measurements

We used a within-subject design between four trials of different cue conditions, as described above. Both objective and subjective measures were collected from each condition. The time for completing the tasks was recorded to measure task performance quantitatively. The error rate was not measured in the user study. At the end of each trial, the participants were asked to complete several subjective questionnaires. We used the NMM Social Presence Questionnaire (Chad Harms (2004)) for measuring Social Presence and the NASA Task Load Index Questionnaire (Hart and Staveland (1988)) for measuring mental and physical load. We also measured the usability of the system using the System Usability Scale (SUS) (Brooke (1996)). Finally, After completing all four trials, participants were asked to rank the four conditions in terms of advantages and disadvantages of each condition, and they provided qualitative feedback from open questions in a post-experiment questionnaire.

# 5 RESULTS

In this section, we report on the results of the user study regarding the performance and usability of all communication cue conditions and summarize the subjective feedback collected from the participants. The mean difference was significant at the 0.05 level, and adjustment for multiple comparisons was automatically made with the Bonferroni correction unless noted otherwise.

We recruited 30 participants (20 male, 10 female) from the local campus community with their ages ranging from 21 to 36 years old ($M = 28.7$, SD = 4.6). Of the participants, 11 had been using video conferencing daily, and the rest used video conferencing a few times a month. Also, 16 participants were familiar with AR or VR interfaces, providing a rating of four or higher on a 7-point Likert item (1: novice 7: expert). This shows that more than half of the participants were familiar with the AR/VR interfaces and more than 80% have had some experience using AR/VR applications. All participants mentioned that they use video conferencing platforms at least a few times a week. This shows that the participants were familiar with the technology and user feedback shouldn't be affected by the novelty of using AR or VR.

## 5.1 Task Completion Time

**Figure 6** shows the average performance time across each of the four communication conditions. The Shapiro-Wilk test indicated that none of the task completion time data were normally distributed except for the volumetric visual cue. A Friedman's test found ($\chi^2$ (3) = 42.365, $p < 0.001$) indicating a statistically significant difference in task completion time depending on which type of visual cue was used to deliver

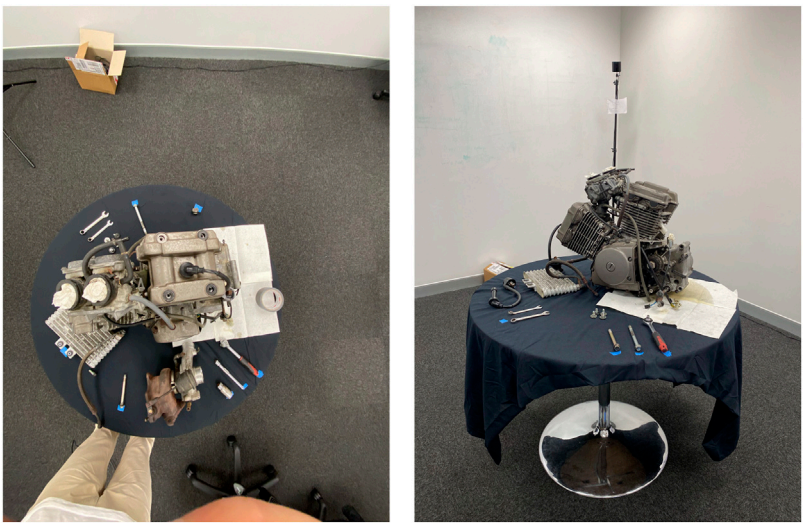

**FIGURE 5 |** Experiment setup: Layout of the tools and engine on the table before each condition.

information. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p = 0.0083$. Therefore, a Wilcoxon Signed Rank Test determined a statistically significant difference between the four conditions' performance time. The time in seconds to complete the volumetric instruction (A4) ($M = 135$, SD = 30.56) was statistically significantly faster than the avatar instruction (A3) ($M = 144.03$, SD = 38.00, $p = 0.005$). There was also a significant difference found in time between A1–A3 ($Z = -3.49$ $p < 0.001$) and A2–A3 ($Z = -3.118$, $p = 0.002$).

## 5.2 Subjective Questionnaires

This section reports on the subjective questionnaires to analyze social presence, workload, and system usability.

### 5.2.1 Social Presence

From the NMM Social Presence Questionnaire, we used the sub-scales Co-Presence (CP), Attention Allocation (AA), and Perceived Message Understanding (PMU) to evaluate the participant's social presence experience. The whole questionnaire has 18 rating items on a 7-point Likert scale (1: strongly disagree–7: strongly agree). Friedman tests and Kendall's W tests showed significant differences in CP($\chi 2$ (3) = 42.187, $p < 0.001$) and PMU($\chi 2$ (3) = 36.591, $p < 0.001$) while AA showed no significant difference between the conditions. A post hoc analysis with the Wilcoxon signed-rank test for CP showed significant pairwise differences for A1-A2 ($Z = -2.829$, $p = 0.005$), A1-A3 ($Z = -2.892$, $p = 0.004$), A1–A4 ($Z = -5.965$, $p < 0.001$), A2–A4 ($Z = -3.620$, $p < 0.001$) and A3-A4 ($Z = -2.829$, $p = 0.005$). Similarly, PMU showed significant difference between A1–A3 ($Z = -4.353$, $p < 0.001$), A2-A3 ($Z = -3.776$, $p < 0.001$), A2-A4 ($Z = -2.507$, $p = 0.012$) and A3-A4 ($Z = -5.694$, $p = 0.00$). Volumetric playback induced the highest sense of co-presence (Mean = 5.78, SD = 1.43) and the highest perceived message understanding (Mean = 5.49, SD = 1.38) among the conditions as shown in the **Table 1**. These results indicate that the integration of volumetric play back resulted in an increase in social presence. This is discussed later in the discussion section.

### 5.2.2 Workload

We used the NASA-TLX questionnaire (Hart and Staveland (1988)) to compare the participant's physical and mental workload across conditions. NASA-TLX includes six rating items in a 100-point range with 5-point steps (0: very low–100: very high, the lower, the better). We focused on the three most relevant items in our study: mental demand, effort and frustration.

A Friedman test indicated significant difference across the cues for mental workload ($\chi 2$ (3) = 13.183, $p = 0.004$). A post hoc analysis with the Wilcoxon signed-rank test and Bonferroni correction showed significant pairwise differences only for A3-A4 ($Z = -2.872$ $p = 0.004$). Similarly, the Friedman test indicated significant difference across the cues for effort ($\chi 2$ (3) = 11.605, $p = 0.009$) which was followed by a Wilcoxon signed-rank test and Bonferroni correction to show pairwise significant difference for A3-A4 ($Z = -3.317$ $p = 0.001$). For Frustration, a Friedman test indicated significant difference across the cues ($\chi 2$ (3) = 13.594, $p = 0.004$). A post hoc analysis with the Wilcoxon signed-rank test and Bonferroni correction showed significant pairwise differences for A3-A4 ($Z = -3.127$ $p = 0.002$) and A2-A3 ($Z = -3.150$ $p = 0.002$).

As shown in **Figure 7** most participants did not experience much frustration except for A3, but the tasks required some mental demand and effort to complete, no matter which condition was used.

**TABLE 1 |** Questionnaires results of social presence and Workload.Co-Presence (CP), Attention Allocation (AA), Perceived Message Understanding (PMU), Mental Effort (ME), Physical Effort (PE), Frustration and System Usability (SUS).

| Condition/Metrics | CP | AA | PMU | ME | PE | Frustration | SUS |
|---|---|---|---|---|---|---|---|
| **Annotation (*M*)** | 5.10 | 4.48 | 5.37 | 35.33 | 33.33 | 20.17 | 72.89 |
| ***(SD)*** | 1.65 | 1.88 | 1.41 | 27.54 | 23.68 | 21.59 | 17.79 |
| **Hand gesture (*M*)** | 5.41 | 4.45 | 5.25 | 35.83 | 38.67 | 24.67 | 68.83 |
| ***(SD)*** | 1.48 | 1.88 | 1.42 | 25.18 | 24.28 | 25.79 | 17.37 |
| **Avatar (*M*)** | 5.51 | 4.48 | 4.83 | 44 | 44.17 | 33.33 | 60.83 |
| ***(SD)*** | 1.51 | 1.77 | 1.52 | 27.77 | 25.67 | 29.34 | 19.17 |
| **Volumetric playback (*M*)** | 5.78 | 4.43 | 5.49 | 33.67 | 30.5 | 21.5 | 74.33 |
| ***(SD)*** | 1.43 | 1.90 | 1.38 | 29.26 | 22.79 | 22.59 | 15.12 |

### 5.2.3 System Usability

To evaluate the usability of our system, we used the SUS questionnaire (Brooke (1996)), which consists of 10 rating items with five response options for respondents (from Strongly Disagree to Strongly Agree). A SUS score of 68 or above is viewed as above average system usability. **Table 1** summarizes the participant's assessment of the system usability in conditions A1–A4. A Friedman test and Kendall's W test showed a significant difference between conditions ($\chi 2$ (3) = 14.263, $p$ = 0.003). Then Wilcoxon signed-rank tests with Bonferroni correction showed significant differences in A2–A3 (Z = −2.814, $p$ = 0.005), and A3-A4 (Z = −3.634, $p$ < 0.001). Participants rated our system to be above-average usability in all conditions except A3 as shown in **Figure 8**.

## 5.3 Preference

**Figure 8** shows the participant's preference ranking of the conditions for the given task. We found a significant difference across conditions ($\chi 2$ (3) = 21.160, $p$ < 0.001) by Friedman and Kendall's W tests. Wilcoxon signed-rank tests showed significant differences between A3 and all other condition pairs. There was no significant difference between any other condition pairs.

# 6 DISCUSSION

In the following, we discuss the above results, the participant's feedback, and possible reasons for the experiment outcome. We will also answer research questions, RQ1 and RQ2.

RQ1 is about the improvement in social presence caused by the use of a volumetric visual cue. It is unsurprising that the introduction of volumetric playback significantly enhanced the participant's co-presence and perceived message understanding which increased overall social presence compared to all other conditions. This could be due to the familiarity with teleconferencing systems and the inclusion of more depth and detail. However, this higher level of social presence did not lead to improvements in completion time and or a reduction in frustration.

RQ2 asks if there was any performance increase as a result of incorporating volumetric playback (A4). From the analysis of the completion times, we could not find any significant increase in performance. While volumetric playback was significantly faster than the avatar cue, both annotations and hand gestures were faster than full-bodied visual cues, as shown in **Figure 6**. One potential explanation of this could be the simplicity and clarity of instructions. With full-bodied instructions, the field of view is occupied to an extent where it becomes overwhelming. Although

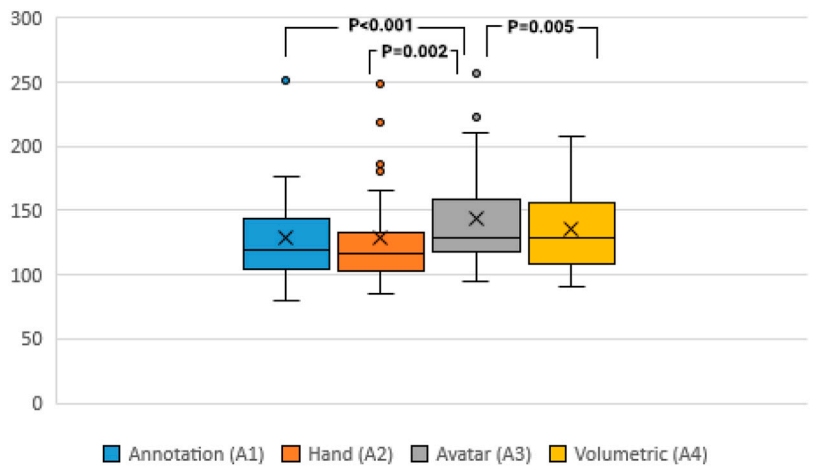


**FIGURE 6 |** Average task performance time (Unit in Second, the lower the better).

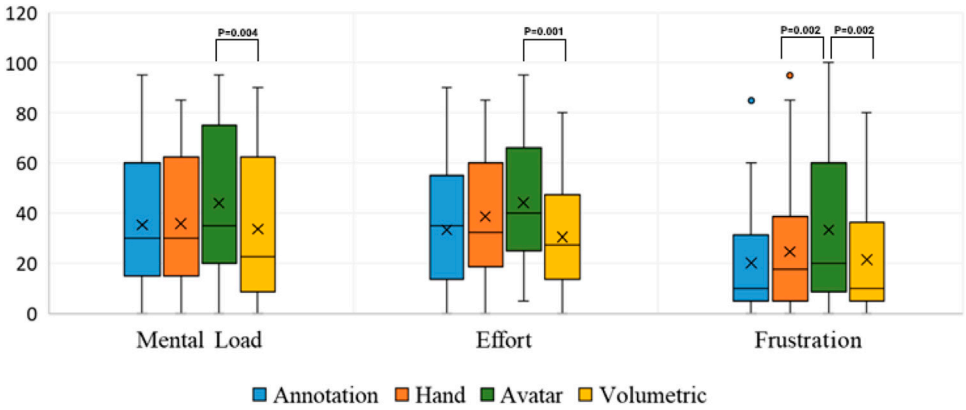


**FIGURE 7 |** Results of TLX questionnaire (100-points range with 5-point steps, 0: very low, 100: very high, the lower the better).

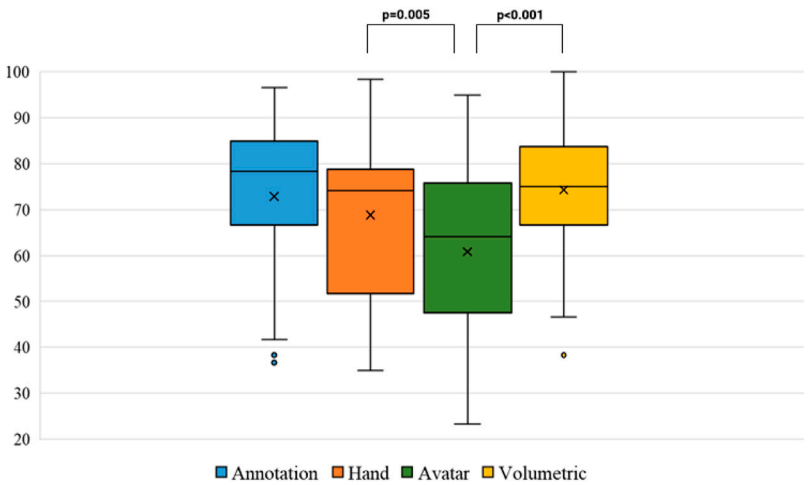


**FIGURE 8 |** User preference based ranking results (Rank1 is the most preferred).

volumetric playback significantly reduced mental workload, effort, and frustration compared with the avatar cue, there was no significant difference compared to the non-embodied instruction cues. This could be attributed to the instructor's more realistic appearance than the artificial movements of the avatar representation. Unsurprisingly, volumetric playback was the preferred choice and scored significantly higher than the avatar representation even though there was no significant difference from the annotation or hand gesture cues.

Examining our research hypotheses, we found that H1 (enhanced social presence) was confirmed, as the volumetric playback affected the sense of social presence and reduced the workload, effort, and frustration. Hence we reject this null hypothesis. H2 (task performance) was partially verified as the volumetric playback led to significantly faster completion time than avatar representation, but this did not hold for other conditions. Accuracy of playback positioning and real work alignment was generally praised across all conditions and the volumetric playback created the feeling of working with the instructor in real-time.

Some of the benefits and draws backs of the system as expressed by the participants are summarized below:

## 6.1 Annotation

Most of the participants mentioned the simplicity and clarity of instructions when using the annotation cue. A few participants mentioned the efficient use of the field of view leading to less distraction. For example, a participant said "*Annotation is very clear and easy to spot in a 3D environment*". When asked about the drawbacks of the Annotation cues, participants mentioned the lack of connectedness; being too simple; not enough details, and limited expression. A few participants suggested the experience could be improved with the inclusion of texts and labels for the task.

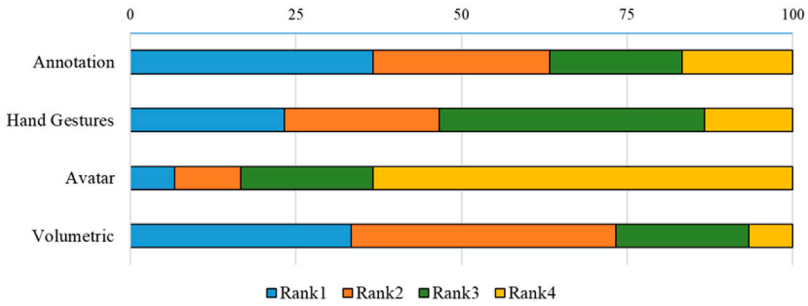


**FIGURE 9 |** Results of SUS questionnaire (100-points range with 80.3 or higher is considered good, 68 and above classified as average and below 51 considered poor usability).

## 6.2 Hands

General comments on the benefits of hands as a visual cue were: easy to notice and understand; expressive and clear; good visibility and with less interference. One of the participants said "*I liked both Annotation and hand gestures, but would prefer to use hand gesture more as it could show more cues than Annotation*". Whereas the drawbacks were listed as: harder to follow; not very clear on the location of objects; lack of body; hand sizes were small. In our implementation, we used the HTC VIVE's hand-tracking plugin as detailed in **section 3**. We used the default hand models, which were semi-transparent as shown in **Figure 3**. It would be worth investigating the effect of hand-model size as hand size varied across participants.

## 6.3 Avatar

Most participants commented on the benefits of using the avatar cue as: looks vivid, attractive; easier to get a general idea of the location of the task; human-like; partner's movement is obvious; feeling good because it is a human. For example, participants mentioned "*I prefer to use Avatar because it was like a human*". Most participants listed the non-natural movement of the Avatar as a major drawback. This could be improved by using a more advanced inverse kinematic system to rig the avatar game object. Other drawbacks were: takes up a lot of space in the field of view; limited hand movements; too much distraction; lack of voice assistance. Both Hand gesture and Avatar were criticized for not having graceful movements. The extent of occlusion caused by the avatar was similar to that caused by the volumetric playback. Providing detailed hand gestures is something that will be explored in future research once we have access to better hand tracking systems.

## 6.4 Volumetric

"Realistic" was the most commented characteristic under the benefits of volumetric playback. Other feed-backs included: good presence and information transfer; like a real human; clear instruction; empathetic; feels like a real situation; natural and better understanding of the task. One of the participants said "*Seeing a real person demonstrate the task, feels like being next to a person*". The volumetric playback condition shared most of the drawbacks listed for the avatar condition. Other drawbacks listed were: low resolution; too complicated; somewhat distracting; taking too long to load. The last comment was related to the implementation of our system. We used the Microsoft Kinect framework to load the volumetric playback, which had a general starting delay of about 7 seconds running in our PC as described in **section 3**. It is worth noting that the delay was only for the first instruction, and the later instructions were loaded instantly.

One interesting suggestion to improve the system was to make the volumetric image semi-transparent or becoming transparent when the user comes close to it. Another suggestion was to combine annotation with volumetric playback and to simplify volumetric playback (like just the bone data of a remote person with volumetric hands and face of a remote person) to help the worker understand the scale of the task need to be performed and the parts needed to be used.

From the first study we found that our RQ1 holds true and RQ2 holds partially true. In order to look at the impact of visual cues, we limited the amount of voice interaction. The participants were allowed to use limited voice communications like "Is this the right one?" and would receive a yes or a no answer from the experimenter. We compared annotation, hand gesture, avatar, and volumetric playback in the first study.

The participants thought that they would be able to perform better if they were able to clearly distinguish which object to take and where to place it, so we conducted a second study using eye gaze for object and location indication.

# 7 USER STUDY 2

We conducted a second study, Study 2 (S2), to investigate if incorporating virtual eye gaze cues could influence the local worker's performance and experience. We used the same assembly tasks to investigate the following research questions:

- RQ1: Are hand gestures more effective than eye gaze as natural visual cues in AR assembly training?
- RQ2: Does combining eye gaze and hand gesture cues improve AR assembly training performance compared to using each cue alone?

**FIGURE 10 | (A)** Hand, **(B)** Eye gaze, and **(C)** Combined cue visualisation in third-person.

The hypothesis for this study was:

- H1: Hand gestures would be more effective than eye gaze in AR assembly training.
- H2: Combining eye gaze and hand gestures would improve performance in AR assembly training.

In Study 2, we only used eye gaze and gesture cues, and not any other visual conditions. This was because our goal was not to present a "good cue" for a specific training task but to explore the appropriateness of incorporating eye gaze and the combination with gesture visual cues. We also understand the potential limitations without using the volumetric playback condition and will discuss the relevant issues in **Section 9**.

We were interested in comparing hand gestures as a sporadic cue and eye gaze as a continuous cue in an AR training environment. To do this, we conducted a formal user study with 11 participants to evaluate the effectiveness and usability of these visual cues. The study's independent variable was the type of natural communication cue, and the primary dependent variable was time. This study was also designed as a within-subjects study using the following three conditions, A1-A3:

- 1. Eye Gaze Only (A1): the user receives instructions by seeing a visual representation of the training expert's eye gaze overlaid onto their view from a third-person perspective as shown in **Figure 10**A.
- 2. Hand Gestures Only (A2): the user receives instructions by seeing a virtual 3D mesh of the training expert's hands overlaid onto their view from a third-person perspective as shown in **Figure 10**B.
- 3. Combined Eye Gaze and Hand Gestures (A3): by combining both cues, the training expert's eye gaze and 3D hand mesh are displayed in the AR scene together and are visible to the user's third-person perspective as shown in **Figure 10**C.

In Study 2, we used the same experimental environment. The tasks, experiment procedures, measurements, and other implementation details stayed the same as in Study 1.

# 8 RESULTS AND DISCUSSION OF STUDY 2

In this section, we report on and discuss the results of Study 2. The user study consisted of 11 participants (6 males, 5 females) from the university aged between 20 and 22 years old. Eight of the participants used video conferencing systems daily, while the other three used them a few times a week. Also, five of the participants were familiar with AR or VR interfaces, as they gave a rating of four or five on a 5-point Likert scale, where 1 = novice and 5 = expert. Study 2 was conducted 2 months after Study 1. We found that participants in the study were familiar with AR/VR interfaces and video conferencing platforms and their judgment wouldn't be clouded with the unfamiliarity of working with AR devices.

## 8.1 Task Completion Time

The average task completion time for each of the three conditions can be seen in **Figure 11A**. From observing the graph, we can see that the gesture cue appears to have the best average task completion time across all user study participants. The Shapiro-Wilk normality test indicated that the task completion times for the gesture and combined cues were normally distributed (with $p = 0.312$ and $p = 0.066$, respectively), but they were not normally distributed for the gaze cue. A Friedman's test ($p = 0.06$) indicated a near significant difference in task completion time depending on which communication cue was used to give the participant instructions. Post hoc analysis with Wilcoxon sign-rank tests was conducted with a Bonferroni correction applied, and they determined that there was a statistically significant difference between the performance time of at least one pair out of the conditions. The time in seconds to complete the tasks using gesture instructions (A2) ($M = 96.55$, SD = 31.86) was significantly faster than gaze instructions (A1) ($M = 142.00$, SD = 49.90, $p = 0.0126$). Comparing the time difference between either A1-A3 or A2-A3 found no significant difference.

## 8.2 Task Workload

**Figure 11B** shows the NASA-TLX questionnaire results, which we used to compare the user study participant's task workload (effort, frustration, and mental load) across the conditions. Friedman's test indicated a statistically significant difference across the mental load cues ($p = 0.013$). Post hoc analysis with a Wilcoxon signed-rank test and Bonferroni correction only showed a statistically significant difference for gaze and combined cues, i.e., A1–A3 ($p = 0.0180$). For frustration, Friedman's test did not show a statistically significant difference across cues, but a Wilcoxon signed-rank test and Bonferroni correction showed a significant pairwise difference

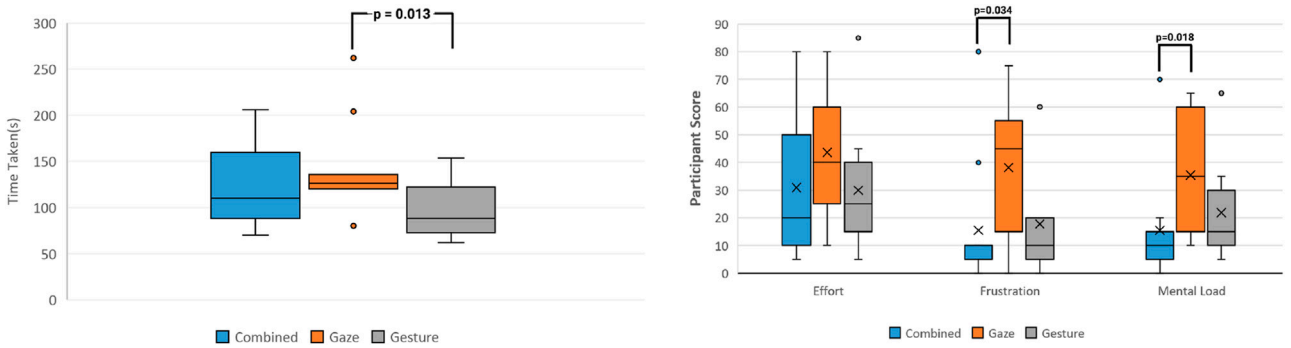


**FIGURE 11 | (A)** Average task completion time for each cue and **(B)** TLX questionnaire results (0: very low, 100: very high).

**TABLE 2 |** Summary of User study 2.

| Condition/Metrics | Time(s) | Effort | Frustration | Mental effort | SUS |
|---|---|---|---|---|---|
| Eye gaze (*M*) | 142 | 40 | 34.55 | 33.64 | 54.55 |
| *(SD)* | 49.90 | 23.34 | 25.15 | 20.50 | 16.27 |
| Hand gestures (*M*) | 96.55 | 30 | 17.73 | 21.82 | 64.77 |
| *(SD)* | 31.86 | 22.69 | 21.61 | 17.07 | 16.26 |
| Combined (*M*) | 122.45 | 30.90 | 15.45 | 15.45 | 65.23 |
| *(SD)* | 45.86 | 26.82 | 24.03 | 19.16 | 19.70 |

for A1–A3 ($p = 0.034$). Neither Friedman's nor Wilcoxon's tests showed a significant difference between cues for effort. As shown in **Figure 11**, participants only really experienced some frustration with the gaze cue, and it also appears that the combined cue caused very little mental load.

## 8.3 System Usability

To evaluate the system's usability, we used the System Usability Scale (SUS). We can summarize the participant's evaluation of the system usability for all three conditions in **Table 2**. The system usability questionnaire results can be seen in **Figure 12A**. Friedman's test ($\chi 2 = 5.907$) and Kendall's W test showed no statistically significant difference between conditions ($p = 0.1048$). Then, Wilcoxon signed-rank tests with Bonferroni correction applied also showed no statistically significant pairwise differences between the cues. On average, participants evaluated the system as being of above-average usability in both A2 and A3, but not A1.

## 8.4 Preferences

We can see the participants' rankings of the different conditions in **Figure 12B**. An ANOVA test showed a statistically significant difference in rankings across conditions ($p < 0.001$). Tukey HSD post hoc tests showed that A2 ranked significantly higher than A1 ($p < 0.001$) and that A3 ranked significantly higher than A1 ($p < 0.001$). However, the Tukey test did not show a statistically significant difference in ranking between A2 and A3 ($p = 0.696$).

## 8.5 DISCUSSION

Here, we will answer the research questions and discuss the results above in more detail and describe some of the potential reasons for the outcomes we have seen. RQ1 asks whether hand gestures alone are more effective than eye gaze alone as

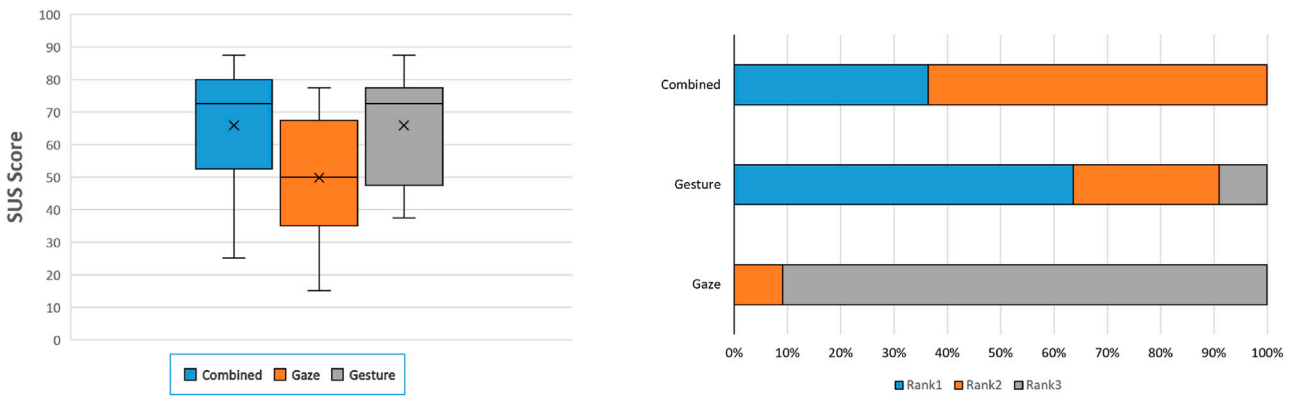


**FIGURE 12 | (A)** SUS questionnaire results (score of 68 considered average) and **(B)** Participants rankings of interfaces.

instructional cues. The user study results show that participants could complete their task sets faster using hand gesture instructions compared to eye gaze instructions. Eye gaze generally caused the highest task load across all three measures (mental load, frustration, and effort). This is likely due to the nature of the tasks - engine assembly involves more of hands-on (feel and place) tasks compared to looking at objects and selecting them.

Given that we are using visual cues only (non-audio), if participants were not shown a pair of hands interacting with the objects in front of them, it could easily cause some misinterpretations—this was evident during the user study. Simply having a gaze line that moves to different points does not give a huge amount of direction as to how objects should be manipulated. The higher mental load is also reinforced by the generally lower score the participants gave for the eye gaze interface system usability. In H1, we hypothesized that hand gestures would be more effective than eye gaze, and the user study confirmed this.

RQ2 asks if combining eye gaze and hand gestures improves performance compared to using each cue alone. Surprisingly, we saw that the combined cue did not result in a significantly faster task completion time than either eye gaze or hand gestures alone. However, the mean completion time for the combined cue (M = 122.5s) was around 20 s slower than that of eye gaze (M = 142s). This means that our original hypothesis, H2 was not able to be confirmed by this user study. However, it was clear from the preferences that participants had an easier time following instructions from the combined cue than the other two cues alone, particularly the eye gaze only cue. The combined cue caused the lowest task load across all three of the measures. This is likely because two separate visual cues are helping the user at any given time rather than just one. This means that when the gestures are not giving enough information, the eye gaze can fill this gap and provide the information, and vice versa. Although there was no statistically significant difference in the participants' rankings of the gesture and combined cues, it was surprising to see that more than half of the participants ranked gesture as the best cue, rather than combined, even though it had the lowest task load. This could be because even though the combined cue provides both gaze and gesture, some of the participants felt as though the gaze cue did not provide enough extra value for it to be worth it, and sometimes even just obstructed their view (especially because of the spherical head model).

Below, we summarise some of the main points participants made when giving general feedback about the three different cues:

The feedback for the eye gaze on its own was generally negative, with most of the participants pointing out that while it did show where they should be looking, it did not indicate what they should be doing with their hands. It also appears that the eye gaze was sometimes distracting—for example, a participant said that *"having only the hand gestures allowed me to focus on what it was doing. I was slightly distracted from it when there was an eye gaze"*. Another participant mentioned how the gaze is better suited as a supporting cue to the gestures, saying that *"the gaze did make the learning easier, but by itself was a bit difficult"*. A couple of participants also had small issues with the accuracy of the eye gaze line.

The main comments regarding the benefits of the gesture cues were that it was obvious and easy to understand which task was being described. One participant said that *"the gestures were more straightforward [than gaze] and easy to mimic"* and another said, *"the hand gestures were the most helpful in explaining the tasks"*. A majority of the feedback about the gesture cues was very similar to these two participants. Surprisingly, the participants provided very little criticism of the gesture interface. One participant stated, *"the hand gesture was not accurate enough in the position to be confident in the task I was supposed to complete"* however, even this participant followed on to say, *"the hand gesture is more descriptive in how the task should be completed"*.

Many of the participants positively commented on the amount of visual information provided by the combined interface, describing that in situations where the eye gaze was lacking direction, the hand gestures made up for it, and vice versa. Overall, the feedback about the combined cue can be summarised by the comment made by one participant, who said, *"the combined gaze and hand was very easy to follow and gave greater direction to the tasks and their location (how to do them and where they were in space)"*. There generally was not any negative feedback about the combined cue, other than a couple of participants pointing out that the eye gaze did not add much to add to their understanding of the instructions.

## 9 DESIGN IMPLICATIONS AND LIMITATIONS

From the study, we can suggest several design implications for future MR remote training systems:

- Volumetric Playback can help improve the sense of social presence in training tasks, as it enhances co-presence and perceived message understanding.
- Volumetric playback significantly reduces stress in training tasks and increases system usability compared to traditional full-body representation while not producing a significant performance loss.
- Using a full-bodied avatar representation in a training system is not recommended unless it is well animated as the unnatural movements lead to distraction and increases mental workload.
- Using simple annotations can significantly improve performance if the Social Presence is not of importance.

Although the user study helped to evaluate and measure the system in a controlled environment with an experimental task, some limitations can lead to further investigation. One of the obvious ones is hardware limitations. While the computer we were given to use was able to run the system successfully, it definitely could have been better - there are many higher-performing CPUs and GPUs available than the one in our computer, and utilizing these could have prevented some of the issues we faced while developing the system like crashes and low frame rates at times. The resolution of the see-through environment was also very low (participants also pointed this

out) due to the front-facing stereo camera on the HMD. Another limitation is that the study focused specifically on engine assembly and maintenance. As suggested by some of the participants, the results may have been different if the tasks were general or if they were not shown how to do the tasks at the beginning of the experiment.

The participants also suggested some possible improvements to the system. The first of these was to change the eye gaze visualization to one that is more lenient towards inaccuracies. They noticed that since the gaze line was so thin, even if it was inaccurate by only a small amount, it sometimes appeared to be pointing at a completely different object, which negatively affected the task completion time. This could be accounted for by changing the visualization from a thin line to a cone shape, with the vertex at the origin of the gaze and the center point of the cone base at the gaze focus point. This way, the user has more of a general area to be looking at for the cue rather than one specific point. Another participant also mentioned that the cable for the headset was not long enough and restricted movement. This could be resolved by using a wireless HMD or by using a wireless adapter for the HTC VIVE. An observation that was noted during the study was that when the instruction was about to play, and if the participant is not looking in the correct direction, they would miss the initial part of the instruction and would have to wait for it to replay. A possible improvement they suggested for this was to implement a way to prompt the user to get into position for the upcoming task before the playback begins. This would allow for a better view of the task and avoid the instruction having to be replayed.

The current system uses absolute spatial coordinates when generating recordings using the different cues. This means that if the setup (or any components of it) were to move even by a very small amount, all of the recordings would be inaccurate. We can account for this if the system is to be modified to add a tracker that is mounted to the top of the engine. This way, we can implement the recordings so that all spatial coordinates are relative to this tracker's position. Even if the engine moves to a different room position, all the recordings will stay accurate. The prototype system implementation had certain limitations that would need improvement in future studies. One was the visual cue offset and the loading time, and the average quality of volumetric playback, as mentioned by a participant.

The study was conducted during the time of Covid-19. Due to constraints in recruiting participants, we were able to recruit only university students. Hence our age group represents mid-30s or younger. Older people may have different views on performance and usability. The impact of performance and usability for participants older than the mid-30s cannot be determined without further study. However, in the future, we would be evaluating these interfaces with diverse demographics.

## 10 CONCLUSION AND FUTURE WORK

This paper presents an MR system for supernatural enhancement of training tasks that features visual cues such as annotation, hand gestures, avatar representation, and eye gaze as visual cues for instruction delivery. We found that participants felt more connected with the instructor as the main benefit of using volumetric playback as the visual cue. Based on the research questions and the results that we had, we conclude that using volumetric playback can significantly improve the sense of Social Presence and increase system usability in the MR training system. Additionally, volumetric playback reduced mental workload and frustration compared to Avatar representation and was the most preferred visual cue for our tasks.

Based on the results and feedback provided in the follow-up study that compared eye gaze with hand gestures, participants reported that the ability to see both eye-gaze and hand gestures simultaneously reduced the mental load and effort required to complete the tasks they were given, as they worked well to complement each other when one cue was lacking. They also generally ranked the gesture-only interface to be the best out of the three, as the eye gaze cue did not add much value and sometimes caused distractions by either providing misleading information or obstructing the view. Regarding the research questions and the results that were obtained, it can be concluded that hand gestures are more effective than eye gaze alone in AR assembly training. Even though combining both cues might not improve performance, it would be better to use the combination as it reduces the workload.

In the future, we would like to explore MR collaboration and training system with live real-time remote collaborators as opposed to the pre-recorded training. We want to investigate how this would impact the measurements discussed earlier. We also plan to explore the usability difference between optical see-through and video see-through displays and incorporate eye gaze in the system to guide the participant in a live remote training system.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Auckland Human Participants Ethics Committee (UAHPEC). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

PS, HB, and MB contributed to the conception and design of the study. SC and NR contributed to performing the second user study. PS wrote the first draft of the article. All authors contributed to manuscript revision, read, and approved the submitted version.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frvir.2021.698523/full#supplementary-material

## REFERENCES


Adcock, M., Feng, D., and Thomas, B. (2013). "Visualization of Off-Surface 3D Viewpoint Locations in Spatial Augmented Reality," in SUI 2013 - Proceedings of the ACM Symposium on Spatial User Interaction, 1–8. doi:10.1145/2491367.2491378

Alem, L., and Li, J. (2011). A Study of Gestures in a Video-Mediated Collaborative Assembly Task. *Adv. Human-Computer Interaction* 2011, 1–7. doi:10.1155/2011/987830

Bauer, M., Kortuem, G., and Segall, Z. (1999). ""Where Are You Pointing at?" A Study of Remote Collaboration in a Wearable Videoconference System," in International Symposium on Wearable Computers, Digest of Papers (IEEE Comp Soc), 151–158. doi:10.1109/iswc.1999.806696

Billinghurst, M., and Kato, H. (1999). "Collaborative Mixed Reality," in Proceedings of the First International Symposium on Mixed Reality, 261–284. doi:10.1007/978-3-642-87512-0_15

Billinghurst, M., and Kato, H. (2007). *Mixed Reality-Merging Physical World and Virtual World*. Washington: Tech. rep.

Brooke, J. (1996). SUS: A 'Quick and Dirty' Usability Scale. *Usability Eval. industry* 189, 207–212. doi:10.1201/9781498710411-35

Carrasco, R., Baker, S., Waycott, J., and Vetere, F. (2017). "Negotiating Stereotypes of Older Adults through Avatars," in Proceedings of the 29th Australian Conference on Computer-Human Interaction (New York, NY, USA: Association for Computing Machinery), 218–227. doi:10.1145/3152771.3152795

Chad Harms, F. B. (2004). "Internal Consistency and Reliability of the Networked Minds Measure of Social Presence," in Seventh Annual International Workshop: Presence, 246–251.

Chang, Y. S., Nuernberger, B., Luan, B., and Höllerer, T. (2017). "Evaluating Gesture-Based Augmented Reality Annotation," in 2017 IEEE Symposium on 3D User Interfaces (Los Angeles, CA: 3DUI), 182–185. doi:10.1109/3DUI.2017.7893337

Cho, S., Kim, S.-w., Lee, J., Ahn, J., and Han, J. (2020). "Effects of Volumetric Capture Avatars on Social Presence in Immersive Virtual Environments," in 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (Atlanta, GA: VR), 26–34. doi:10.1109/VR46266.2020.00020

De Pace, F., Manuri, F., Sanna, A., and Zappia, D. (2019). A Comparison between Two Different Approaches for a Collaborative Mixed-Virtual Environment in Industrial Maintenance. *Front. Robot. AI* 6, 18. doi:10.3389/frobt.2019.00018

Dou, M., Davidson, P., Fanello, S. R., Khamis, S., Kowdle, A., Rhemann, C., et al. (2017). Motion2fusion: Real-Time Volumetric Performance Capture. *ACM Trans. Graph.* 36, 1–16. doi:10.1145/3130800.3130801

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., and Kramer, A. D. I. (2004). Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Human-Computer Interaction* 19, 273–309. doi:10.1207/s15327051hci1903_3

Gupta, K., Lee, G. A., and Billinghurst, M. (2016). Do you See what I See? the Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Trans. Vis. Comput. Graphics* 22, 2413–2422. doi:10.1109/TVCG.2016.2593778

Gurevich, P., Lanir, J., Cohen, B., and Stone, R. (2012). "Teleadvisor: A Versatile Augmented Reality Tool for Remote Assistance," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA: Association for Computing Machinery), 619–622. doi:10.1145/2207676.2207763

Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., and Theobalt, C. (2019). Livecap: Real-Time Human Performance Capture from Monocular Video. *ACM Trans. Graph.* 38, 1–17. doi:10.1145/3311970

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology* (Elsevier), 52, 139–183. doi:10.1016/s0166-4115(08)62386-9

Hasenfratz, J.-M., Lapierre, M., and Sillion, F. (2004). A Real-Time System for Full Body Interaction with Virtual Worlds (Postfach 8043, 38621 Goslar, Germany: The Eurographics Association). *Eurographics Workshop on Virtual Environments*, 147–156. doi:10.2312/EGVE/EGVE04/147-156

Heidicker, P., Langbehn, E., and Steinicke, F. (2017). "Influence of Avatar Appearance on Presence in Social VR," in 2017 IEEE Symposium on 3D User Interfaces, 3DUI 2017 - Proceedings (Los Angeles, CA: Institute of Electrical and Electronics Engineers Inc.), 233–234. doi:10.1109/3DUI.2017.7893357

Izadi, S., Davison, A., Fitzgibbon, A., Kim, D., Hilliges, O., Molyneaux, D., et al. (2011). *Kinectfusion: Real-Time 3d Reconstruction and Interaction Using a Moving Depth Camera*, 559–568. doi:10.1145/2047196.2047270

Joachimczak, M., Liu, J., and Ando, H. (2017). "Real-Time Mixed-Reality Telepresence via 3D Reconstruction with Hololens and Commodity Depth Sensors," in ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction (Glasgow: ACM), 514–515. doi:10.1145/3136755.3143031

Jung, S.-H., and Bajcsy, R. (2006). A Framework for Constructing Real-Time Immersive Environments for Training Physical Activities. *Jmm* 1, 9–17. doi:10.4304/jmm.1.7.9-17

Kalra, P., Magnenat-Thalmann, N., Moccozet, L., Sannier, G., Aubel, A., and Thalmann, D. (1998). Real-time Animation of Realistic Virtual Humans. *IEEE Comput. Grap. Appl.* 18, 42–56. doi:10.1109/38.708560

Ke, C., Kang, B., Chen, D., and Li, X. (2005). "An Augmented Reality-Based Application for Equipment Maintenance," in *Lecture Notes in Computer Science* (Berlin, Heidelberg: Springer), Vol. 3784, 836–841. doi:10.1007/11573548_107

Kirk, D., Rodden, T., and Fraser, D. S. (2007). "Turn it This Way: Grounding Collaborative Action with Remote Gestures," in Conference on Human Factors in Computing Systems - Proceedings, 1039–1048. doi:10.1145/1240624.1240782

Milgram, P., and Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays. *IEICE TRANSACTIONS Inf. Syst.* 77, 1321–1329.

Mohler, B. J., Bülthoff, H. H., Thompson, W. B., and Creem-Regehr, S. H. (2008). *A Full-Body Avatar Improves Egocentric Distance Judgments in an Immersive Virtual Environment*. doi:10.1145/1394281.1394323

Newcombe, R. A., Fox, D., and Seitz, S. M. (2015). "Dynamicfusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (Boston, MA: CVPR), 343–352. doi:10.1109/CVPR.2015.7298631

Olwal, A., Gustafsson, J., and Lindfors, C. (2008). "Spatial Augmented Reality on Industrial CNC-Machines," in *The Engineering Reality of Virtual Reality 2008*. Editors I. E. McDowall and M. Dolinsky (San Jose, CA: SPIE), Vol. 6804, 680409. doi:10.1117/12.760960

Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., et al. (2016a). "Holoportation: Virtual 3D Teleportation in Real-Time," in UIST 2016 - Proceedings of the 29th Annual Symposium on User Interface Software and Technology, 741–754. doi:10.1145/2984511.2984517

Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., et al. (2016b). "Holoportation: Virtual 3D Teleportation in Real-Time," in Proceedings of the 29th Annual Symposium on User Interface Software and Technology (New York, NY, USA: Association for Computing Machinery), 741–754. doi:10.1145/2984511.2984517

Ou, J., Fussell, S. R., Chen, X., Setlock, L. D., and Yang, J. (2003). "Gestural Communication over Video Stream," in Proceedings of the 5th International Conference on Multimodal Interfaces - ICMI '03 (New York, New York, USA: Association for Computing Machinery (ACM))), 242. doi:10.1145/958432.958477

Pejsa, T., Kantor, J., Benko, H., Ofek, E., and Wilson, A. (2016). "Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment," in Proceedings of the ACM Conference on Computer Supported Cooperative Work (New York: CSCW), 27, 1716–1725. doi:10.1145/2818048.2819965

Piumsomboon, T., Lee, G. A., Hart, J. D., Ens, B., Lindeman, R. W., Thomas, B. H., et al. (2018). "Mini-me: An Adaptive Avatar for Mixed Reality Remote Collaboration," in Conference on Human Factors in Computing Systems - Proceedings, Los Angeles, CA. doi:10.1145/3173574.3173620

Piumsomboon, T., Lee, G., Lindeman, R. W., and Billinghurst, M. (2017). "Exploring Natural Eye-Gaze-Based Interaction for Immersive Virtual Reality," in 2017 IEEE Symposium on 3D User Interfaces (3DUI), 36–39. doi:10.1109/3DUI.2017.7893315

Regenbrecht, H., Meng, K., Reepen, A., Beck, S., and Langlotz, T. (2017). "Mixed Voxel Reality: Presence and Embodiment in Low Fidelity, Visually Coherent, Mixed Reality Environments," in Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality (Nantes: ISMAR), 90–99. doi:10.1109/ISMAR.2017.26

Rose, E., Breen, D., Ahlers, K. H., Crampton, C., Tuceryan, M., Whitaker, R., et al. (1995). "Annotating Real-World Objects Using Augmented Reality," in

*Computer Graphics* (Elsevier), 357–370. doi:10.1016/b978-0-12-227741-2.50029-3

Sasikumar, P., Gao, L., Bai, H., and Billinghurst, M. (2019). "Wearable Remotefusion: A Mixed Reality Remote Collaboration System with Local Eye Gaze and Remote Hand Gesture Sharing," in 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (Beijing: ISMAR-Adjunct), 393–394. doi:10.1109/ISMAR-Adjunct.2019.000-3

Schwald, B., and de Laval, B. (2003). *An Augmented Reality System for Training and Assistance to Maintenance in the Industrial Context*.

Slater, M., and Usoh, M. (1994). Body Centred Interaction in Immersive Virtual Environments. *Artif. Life virtual reality* 1, 125–148.

Smith, H. J., and Neff, M. (2018). "Communication Behavior in Embodied Virtual Reality," in Conference on Human Factors in Computing Systems - Proceedings (New York, New York, USA: Association for Computing Machinery), 1–12. doi:10.1145/3173574.3173863

Špakov, O., Istance, H., Räihä, K.-J., Viitanen, T., and Siirtola, H. (2019). "Eye Gaze and Head Gaze in Collaborative Games," in Proceedings of the 11th ACM Symposium on Eye Tracking Research Amp; Applications-ETRA '19 (New York, NY, USA: Association for Computing Machinery). doi:10.1145/3317959.3321489

Wang, T.-Y., Sato, Y., Otsuki, M., Kuzuoka, H., and Suzuki, Y. (2020). Effect of Body Representation Level of an Avatar on Quality of AR-based Remote Instruction. *Mti* 4, 3. doi:10.3390/mti4010003

Webel, S., Bockholt, U., Engelke, T., Gavish, N., Olbrich, M., and Preusche, C. (2013). An Augmented Reality Training Platform for Assembly and Maintenance Skills. *Robotics Autonomous Syst.* 61, 398–403. doi:10.1016/j.robot.2012.09.013

Wither, J., DiVerdi, S., and Höllerer, T. (2009). Annotation in Outdoor Augmented Reality. *Comput. Graphics* 33, 679–689. doi:10.1016/j.cag.2009.06.001

Xu, M., David, J. M., and Kim, S. H. (2018). The Fourth Industrial Revolution: Opportunities and Challenges. *Ijfr* 9, 90. doi:10.5430/ijfr.v9n2p90

Yang, J., Sasikumar, P., Bai, H., Barde, A., Sörös, G., and Billinghurst, M. (2020). The Effects of Spatial Auditory and Visual Cues on Mixed Reality Remote Collaboration. *J. Multimodal User Inter.* 14, 337–352. doi:10.1007/s12193-020-00331-1

Yonggao Yang, Y., Xusheng Wang, X., and Chen, J. X. (2002). Rendering Avatars in Virtual Reality: Integrating a 3D Model with 2D Images. *Comput. Sci. Eng.* 4, 86–91. doi:10.1109/5992.976440

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.