Unlocking Understanding: An Investigation of Multimodal Communication in Virtual Reality Collaboration

Ryan Ghamandi
University of Central Florida
Orlando, Florida, USA
ryanghamandi1@gmail.com

Ravi Kiran Kattoju
University of Central Florida
Orlando, Florida, USA
kattoju.ravikiran@gmail.com

Yahya Hmaiti
University of Central Florida
Orlando, Florida, USA
yohan.hmaiti@ucf.edu

Mykola Maslych
University of Central Florida
Orlando, Florida, USA
mykola.maslych@ucf.edu

Eugene Taranta II
University of Central Florida
Orlando, Florida, USA
etaranta@gmail.com

Ryan P. McMahan
University of Central Florida
Orlando, Florida, USA
rpm@ucf.edu

Joseph J. LaViola Jr.
University of Central Florida
Orlando, Florida, USA
jlaviola@ucf.edu

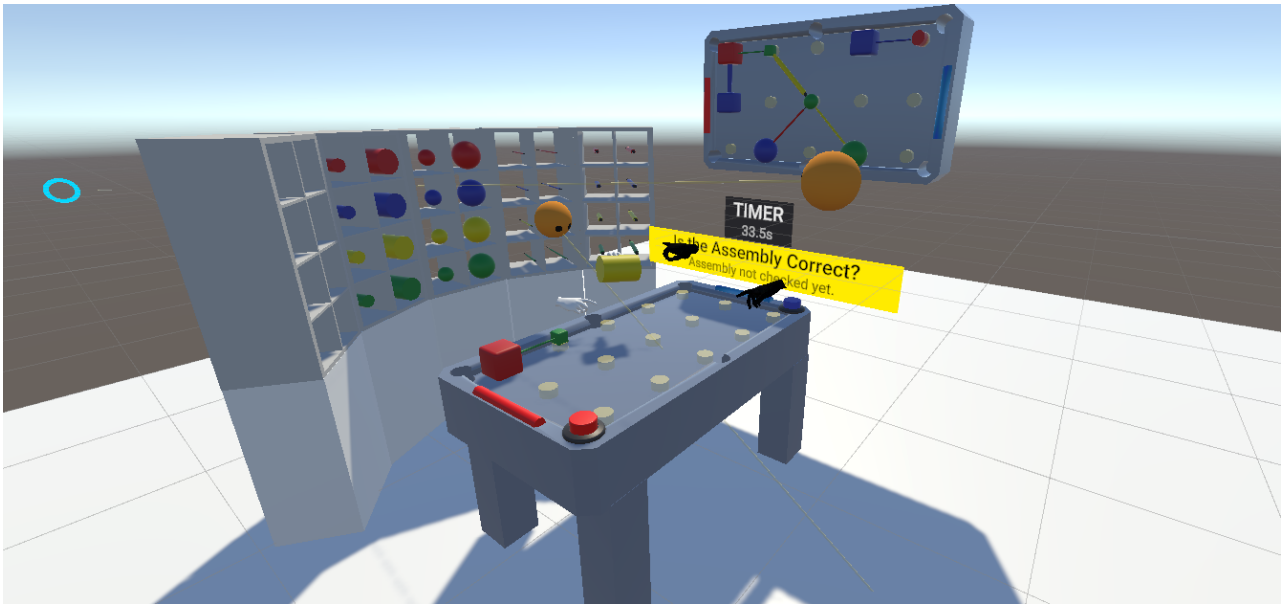

Figure 1: VR Environment showing collaborative efforts between two people during an assembly task. Figure shows two collaborators (yellow heads and hands), a central work table where objects have to be assembled, a reference table above containing task information for a mentor to see, and a shelf for a mentee to grab objects from.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642491

ABSTRACT

Communication in collaboration, especially synchronous, remote communication, is crucial to the success of task-specific goals. Insufficient or excessive forms of communication may lead to detrimental effects on task performance while increasing mental fatigue. However, identifying which combinations of communication modalities provide the most efficient transfer of information in collaborative settings will greatly improve collaboration. To investigate this, we developed a remote, synchronous, asymmetric VR collaborative assembly task application, where users play the role of either mentor or mentee, and were exposed to different combinations of three

communication modalities: voice, gestures, and gaze. Through task-based experiments with 25 pairs of participants (50 individuals), we evaluated quantitative and qualitative data and found that gaze did not differ significantly from multiple combinations of communication modalities. Our qualitative results indicate that mentees experienced more difficulty and frustration in completing tasks than mentors, with both types of users preferring all three modalities to be present.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing design and evaluation methods**; **User studies**; **Interactive systems and tools**; **Virtual reality**; **Interaction design**; **Empirical studies in collaborative and social computing**.

KEYWORDS

collaboration, virtual reality, communication

ACM Reference Format:

Ryan Ghamandi, Ravi Kiran Kattoju, Yahya Hmaiti, Mykola Maslych, Eugene Taranta II, Ryan P. McMahan, and Joseph J. LaViola Jr.. 2024. Unlocking Understanding: An Investigation of Multimodal Communication in Virtual Reality Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642491

1 INTRODUCTION

Communication is the exchange of information between two entities. As such, communication cues (any act or structure that affects the behavior of another person [40]) of many types serve to facilitate efficient transfer of information to connect people together along several modalities (various channels or methods used for communication). With the advancement of technology, the ability for remote collaboration to flourish has increased and the inclusion of extended reality and communication cues in these scenarios aid in connecting people from long distances [13]. As such, these progressions aid collaboration in tasks that span fields such as medicine, education, etc. [48]. They are able to conserve resources for both individuals and organizations especially since individuals can participate remotely in a collaborative task and are also able to perform the same actions in a virtual environment as the real world, especially with the implementation of natural communication cues connecting individuals from afar.

However, in such settings, although many forms of communication exist in order to relay information to another individual [4, 8, 17], collaborative settings either have too few communication modalities (e.g. speech or hand gestures) whereby sensed information can be captured for input [37, 44] which make collaboration slower, or have too many communication modalities, which can increase cognitive load and visual clutter [12]. Such a problem can potentially arise in many extended reality collaborative tasks. Work has been done to investigate how communication cues help to perform collaborative tasks in both augmented and mixed reality [48]. However, not much investigation has been done in virtual reality.

To investigate how communication modalities affect collaboration in VR, we conducted a single factor study where pairs of participants, having varying access to three communication modalities (voice, hand gestures, head gaze), complete a collaborative assembly task in VR, each with either *voice (**V**)*, *gestures (**G**)*, *head gaze (**H**)*, *voice + gestures (**VG**)*, *voice + head gaze (**VH**)*, *gestures + head gaze (**GH**)*, and *voice + gestures + head gaze (**VGH**)*, with each user either taking the role of *mentor* (users in charge of instructing another user) or *mentee* (users who must manipulate objects based on another user's instructions in the task).

The key findings are that combinations of communication modalities along with the *head gaze* condition perform significantly better than *voice* and *gestures*. Findings also indicate that mentees experience higher levels of frustration and difficulty during the task than mentors due to the nature of the asymmetrical aspect of the task.

The contribution of this work is that it presents design considerations for communication in VR collaborative settings as it explores what communication modalities and cues are used during such scenarios and how they are used to relate specific information, as well as a full investigation on all possible combinations of voice, hand gestures, and head gaze in collaborative virtual reality.

2 RELATED WORK

In this section, we elaborate on visual cues used in communication for collaboration involving the reality-virtuality continuum[35].

2.1 Visual Communication Cues in Virtual/Mixed Reality Collaboration

One of the most common forms of communication in any collaborative scenario, whether it be traditional or involving mixed/virtual reality (MR/VR), is voice, which is mainly used to convey descriptive information, whether it be spatial information, user intention, directions, object description, etc. [1]. With the use of such technologies in collaborative tasks where individuals are remote, the use of various communication cues serves to help convey information.

Gaze is one such communication cue that can be used in XR collaborative scenarios to help convey information between individuals. The way that gaze is represented in such scenarios is as a frustum or single ray to indicate where someone is paying attention to spatially with either their head or eyes. Jing et al. [24] show that the use of gaze sharing across individuals when executing a collaborative task lowered cognitive load and improved mutual understanding, with users also preferring bi-directional gaze as it helped identify shared interests. They also show in other work [22] that gaze paired with speech improves task performance and establishes a sense of co-presence with others. Other examples of sharing gaze, whether it is eye gaze or head gaze as a visual cue, include [9, 14, 18, 23, 25, 32, 34, 50]. However, these works focus mainly on how gaze works as the main visual cue present in the experiment, with most of these works only focusing on the effects of gaze in either AR or MR.

Hand gestures also represent a communication cue used in a large amount of prior research involving MR collaborative scenarios to convey information. Types of gestures commonly used for conveying information in these situations include deictic (spatial and directional) and representational (holding a specific meaning or reference) [10, 30]. Bauer et al. [2] and Fussell at al. [10] demonstrate that the use of pointing gestures in general in collaborative

Table 1: Table showing classification of previous literature under XR technologies used and communication modalities employed, which shows how communication cues employed and number of conditions differed from this work. Our work focuses on comprehensively investigating communication cues in VR collaboration with 7 conditions while the other entries in the table do not.

Ref.	Task/Context	AR/VR/MR	Voice	Gestures	Gaze	Avatar	Pointer	Annotations	Objects	Physiology	Comm. Conditions
[11]	Lego Assembly	MR	✓	✓							2
[14]	Lego Assembly	AR	✓		✓		✓				4
[18]	Block Assembly, Object Identification	AR	✓	✓	✓						2
[24]	Puzzle, Visual Search	MR	✓		✓						5
[22]	Search	MR	✓		✓	✓					4
[23]	Search, Matching, Puzzle-Solving	AR	✓	✓	✓		✓				4
[25]	Search	MR	✓		✓						4
[34]	Identification	AR			✓						2
[33]	Search	MR	✓	✓							0
[41]	Lego Assembly	MR	✓	✓							0
[1]	Search and Assembly	MR	✓	✓	✓						4
[49]	Assembly	MR	✓	✓	✓	✓					0
[20]	Puzzle	VR	✓	✓	✓	✓	✓				2
[38]	N/A	MR		✓	✓	✓					N/A
[52]	Crossword Puzzle, Furniture Placement	AR	✓			✓					6
[27]	Lego, Tangram, Origami	MR	✓	✓			✓	✓			4
[28]	Lego, Tangram, Origami	MR	✓	✓			✓	✓			4
[29]	Puzzle	AR	✓				✓	✓			3
[42]	Navigation	MR	✓	✓			✓	✓			2
[43]	Decoration, Organization	MR	✓	✓			✓	✓			4, 2, 1
[31]	Navigation	MR	✓						✓		3
[36]	Object Identification and Positioning	MR	✓	✓					✓		2
[47]	Puzzle	MR	✓	✓		✓		✓	✓		2
[21]	Puzzle	MR	✓	✓	✓					✓	2, 4
[12]	Surgery	MR	✓	✓		✓		✓			N/A
[17]	Drawing and Sharing	VR	✓					✓			3
[50]	Lego Assembly	MR			✓						2
This Paper	Block/Wire Assembly	VR	✓	✓	✓						7

scenarios benefit communication by providing additional spatial information to individuals in a quick and easy manner. More specifically for hand gestures, Lee et al. [33] showed that using hand gestures, along with gaze sharing, helped users understand each other better by delivering deictic gestures like pointing at objects or directions, symbolic gestures like indicating numbers, and social gestures like thumbs up or down. Similar findings were reached by Tecchia et al. [41] who showed that using hand gestures as opposed to a cursor conveyed information better and more adequately as users were able to easier and more accurately indicate locations and positioning in a remote location. Furthermore, Gao et al. overlaid virtual hands on a local worker's view by providing a point-cloud render of a physical environment to a remote worker in a mixed reality system; they found that these overlaid hands are able to make workers feel more connected both spatially and mentally, as well as aided collaboration through the use of deictic gestures[11]. Despite this, these works also focus mainly on isolating hand gestures and investigating how they perform as a visual communication cue, mainly in AR/MR contexts as well.

Prior work that combined head gaze with hand gesture visual cues indicates that this combination brings about improvement in collaborative performance in MR. Bai et al. [1] showed that combining both head gaze and hand gestures results in a much more important performance increase that is significantly better than communication through voice only (though solely hand gestures or head gaze did not outperform voice only), and was also shown to provide better co-presence for users than just gaze. Furthermore, Wang et al. [49] combined these cues alongside avatar-based cues and their findings indicate an improvement of user experience. Although these works investigated such combinations of cues in these settings, they did not fully investigate all possible combinations of cues present in the study (including each cue being used solely as a condition), as well as the fact that they were only investigated in AR/MR.

Other combinations of visual communication cues with other cues (i.e. including voice communication) have been investigated in prior work to understand how effective they are in conveying information and increasing subjective relationships such as co-presence. Such cues combined with others in these situations include avatar representation [20, 38, 52], pointers as well as annotations/drawings [17, 27–29, 42, 43], virtual objects [31, 36, 47], and physiological data [21]. This work provided insight into how exactly numerous types of visual cues could be combined together to improve collaborative experiences, but were mainly focused on certain combinations of cues and also mainly focused on AR/MR.

In summary, a large array of research work has been conducted to evaluate and explore the combination of several dissimilar visual communication cues, especially gestures and gaze, as shown by Table 1 which details how our work differs from previous work. Not much work presented in the table investigates VR collaboration or expands their conditions to comprehensively explore the presence or absence of communication cues in tasks. Findings in the literature suggested that such communication cues, especially when combined, improve immersive collaborative experiences as well as task performance and efficiency. Nevertheless, prior work is limited to mainly augmented reality (AR) or MR scenarios. Thus,

our work intends to discover how effective the addition of natural visual communication cues affects VR collaboration, as well as how they affect the immersion and presence levels of individuals using the system in relation to their collaborators, along with how they are able to improve task performance and efficiency with their presence or absence.

3 METHODOLOGY

In this section, we describe our study design. The task employed is an assembly task where participants took up either the role of mentor or mentee and collaborated with another person to assemble a given configuration composed of shapes and wires. In the following sections, we describe our participant demographics, apparatus, study design, research hypotheses, the actual task more in-depth, and procedure.

3.1 Participants

We recruited 25 participant pairs from our university, which resulted in 50 individuals. Participants were required to be 18 years of age or older, have normal or corrected-to-normal vision, and be able to hear, walk, extend both arms, use both hands, and speak and understand English. Participants with any visual, auditory, neurological, or physical disabilities were excluded. Our final participant pool comprised 50 individuals (27 male and 23 female). The ages of our participants ranged from 18 to 54 with a mean age of 19. Each pair of participants knew each other before participating in the experiment.

3.2 Apparatus

For the VR tasks on the software side, the test application was developed using Unity3D [1] with the Oculus Integration Package for Unity serving as the VR Software Development Kit [2] and Photon PUN 2 for Unity providing the network and multiplayer functionality [3]. For hardware, two Meta Quest 2 devices [4] were used for participants to wear and enter the virtual experience, which were facilitated via the Quest Link to computer devices supported by the Meta Quest 2 software. One PC was equipped with an Intel Core i7-10875H CPU, Nvidia GeForce RTX 3080 and 32GB RAM, with the other being equipped with an Intel Core i5-11400H CPU, Nvidia GeForce RTX 3060 Graphics Card and 16GB RAM.

3.3 Study Design

To evaluate how effective varying combinations of communication modalities were, we conducted a single factor within-subjects study with seven conditions (*voice (**V**), gestures (**G**), head gaze (**H**), voice + gestures (**VG**), voice + head gaze (**VH**), gestures + head gaze (**GH**)*, and *voice + gestures + head gaze (**VGH**)*). We chose to investigate the seven communication methods in order to see the range of effectiveness each modality combination offers to users as well as to investigate how effective voice is as a communication modality by comparing scenarios where it is present with where it is not present.

[1] https://unity.com/
[2] https://developer.oculus.com/downloads/package/unity-integration/
[3] https://doc-api.photonengine.com/en/PUN/v2/
[4] https://store.facebook.com/quest/products/quest-2/

To avoid potential confounding variables (e.g. long-distance locomotion), participants had access to everything they needed for their part of the task in their immediate space. Participants also underwent training in order to learn how the task and its mechanics worked in order to avoid any learning effects.

Objective dependent variables included Task Completion Time (the time taken for users to complete a run of the task given a specific communication combination), Object Manipulation Time (the time taken to select a single object and release that same object in a specific place), and Object Selection Time (the time taken to select a new object after releasing another object or starting the task). Object Manipulation Time was recorded to determine how effective a specific combination of communication cues was to place an object directly from the shelf onto a specific place on the shared table. Likewise, Object Selection Time was recorded to determine how effective a specific combination of communication cues was to deliver instructions on which object to grab from the shelf upon placing the previously manipulated object.

Subjective dependent variables included Task Load, System Usability, Simulator Sickness and Presence. We measured each of these variables by administering the NASA Task Load Index (NASA-TLX)[16], System Usability Scale [3], Simulator Sickness Questionnaire [26] and the Social Presence Questionnaire [15] sublists of Co-presence, Attentional Allocation, Perceived Message Understanding, and Perceived Behavioral Interdependence (these sublists were used as only these were relevant to our experiment) respectively after each task run.

In addition, we collected responses for questions regarding Frustration and Difficulty for each modality and Preferences for modalities and their combinations along with questions for usefulness of each modality, as well as Likert Scale-based questions for overall communication difficulty, task difficulty, mental engagement and physical engagement, all of which were administered post-study.

3.4 Hypotheses

Given the structure and parameters of our study we propose the following hypotheses:

> H1: Participants will perform tasks faster when there are more communication cues present than less.
> H2: The presence of either hand gestures or head gaze in any task setting will significantly decrease completion times compared to when it is absent.
> H3: Mentors and mentees will have differences in what communication cues they prefer in the task.

We based our hypotheses on previous research, especially work done by Bai et al. [1] as they conducted a similar experiment with hand gestures and head gaze in mixed reality. We also based our hypotheses on previous work from Wang et al. [49] as they also combined multiple cues and investigated performance, presence and workload.

3.5 Experimental Task

The collaborative task employed was a VR assembly task, which was spurred by its use in previous collaborative research [11, 20, 41], where we had participants collaborate to assemble objects on a "work table" correctly in order to complete the task successfully.

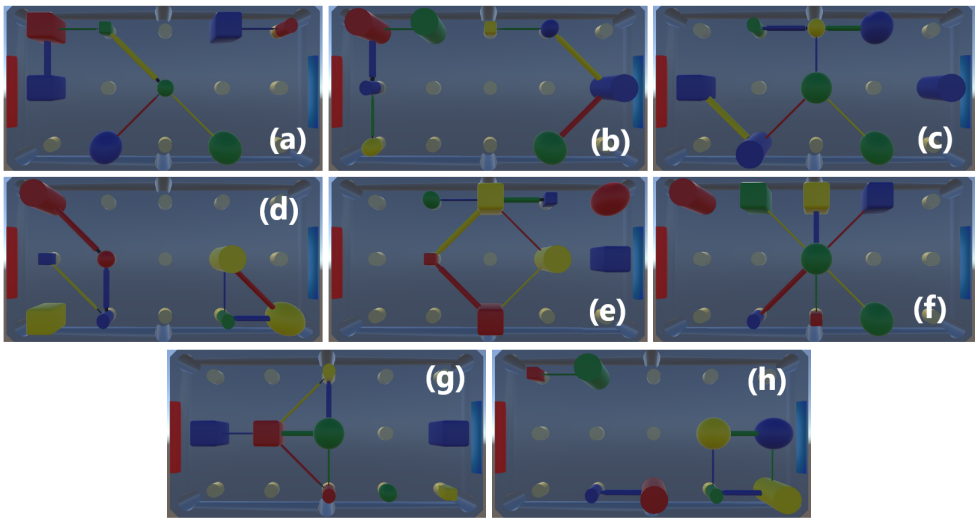

Figure 2: Grid showing the eight configurations every pair of users had to assemble during each trial. (a) Features the training configuration, while (b) through (h) are the configurations participants were tasked with in this specific order.

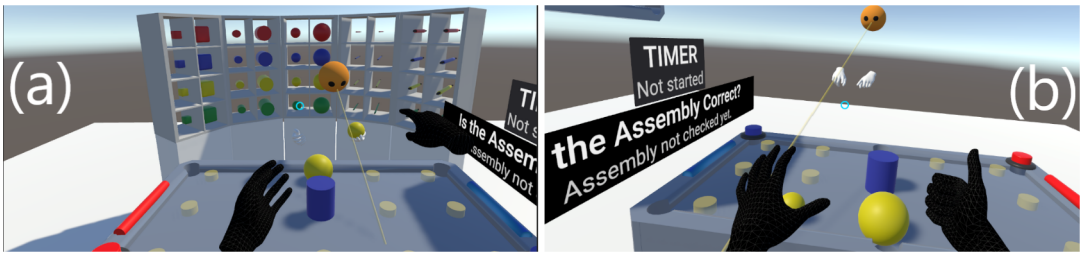

Figure 3: Views of Both Mentor and Mentee. (a) Shows the perspective of the mentor. (b) Shows the perspective of the mentee.

One participant was designated as the mentor (collaborator in charge of looking at a separate "reference table" showing the right configuration to be assembled) and the other was designated as the mentee (who was in charge of manipulating the objects based on the directions of the mentor), with the two being on opposite sides of the work table. During the task and regardless of the communication modality used, only the mentor could see the correct configuration on the reference table, and only the mentee could manipulate objects. The mentor had to instruct and communicate to the mentee the adequate shapes and wires to choose from the inventory (which was on the shelf provided to the mentee) and where to place those objects on the work table. The mentee had to grab the necessary objects from an inventory shelf situated toward the back of the mentee's starting position upon entering the virtual environment, then they had to place the grabbed object(s) where indicated by the mentor on the work table. We designed the shelf to have an object spawn to replace whichever object was picked from it, and the object would either be a wire or 3D shape (i.e. cube, cylinder, sphere), both were either small or big. The 3D shapes and wires had one of four colors (red, blue, yellow, or green). These object properties were designed so as to provide complexity to the objects that the collaborators had to use some form of communication to assemble the configuration given. Both the work and reference tables had fifteen pegs (three rows, five columns), where shapes were situated, and wires could only be placed in between adjacent shapes (short wires connected horizontally or vertically adjacent shapes, whereas long wires connected diagonally adjacent shapes). We had eight pre-made configurations with *8 shapes* and *6 wires* each that participants would go through; one for training, and the rest for the seven conditions they would experience after training.

To start the trial, the mentor pressed a red button (start button) that allowed grabbing objects and their placement on the work table along with starting a trial timer. Moreover, the mentor had to press a blue button (end button) to check if the resulting configuration that the pair of participants assembled was correct. If the assembled configuration was correct, the task was completed successfully. Otherwise, they would have to try and fix the configuration until assembled correctly. The buttons were located on the mentor's side of the table, where the start button was located on the left corner of the work table, whereas the end button was on the right corner. The conditions were administered randomly to each pair of participants using a counterbalanced latin-square, and depending on the condition, participants had a specific communication modality assigned to use to collaborate in the assembly task. The set of modalities included voice, hand gestures, head gaze, or a

combination of two or all three together, which resulted in a total of 7 modality-based conditions. Both mentors and mentees had access to the same modalities during each trial. Figure 2 shows the eight configurations that participants had to recreate and Figure 3 shows the different perspectives that mentors and mentees experienced, along with representations of the visual communication cues.

3.6 Procedure

Upon arrival, recruited participants were asked to review and sign an informed consent document. We then collected participant demographics (i.e., age and gender) and administered an initial SSQ to serve as a baseline for simulator sickness.

Once completed with the initial paperwork and surveys, the researcher showed both participants the areas where they would be situated in during each task run (both in the same room). The researcher then initialized the training scenario with all three communication modalities available and had participants view a computer screen as the researcher equipped themselves with the mentor's headset. The researcher then showed the mentor the display board that showed the current configuration that they would have to direct the mentee to reconstruct on the work table, as well as showing them the different objects that were available to select from.

Afterwards, the researcher then led the mentee to their area and equipped the mentee's headset to show them how to grab objects from the shelf adjacent to their area virtually as well as other mechanics such as how to place shapes and wires and where to place long and short wires. The researcher then took off the headset and let the mentee wear it by having them put it on to practice their duties and get accustomed to the environment.

After a minute of letting the mentee get acquainted with the system, the researcher then had both participants practice the task as the reference board already had a random assembly configuration for the training scenario. All communication modalities were present during this time so participants could become acquainted with them. When both participants successfully completed the training task, they were asked to sit and complete the SSQ again in order to see any changes in symptoms they may experience.

Afterwards, participants then entered the virtual environment seven more times to complete the task with the latter seven premade configurations and varying access to the three communication modalities. After each task run, participants would fill out a NASA TLX Survey, System Usability Survey (SUS) and the Social Presence Questionnaire.

When participants completed all seven task runs, a post-study survey was also administered to the participants. The time required to complete the study was approximately 90 minutes. Participants were each compensated 15 dollars cash.

4 RESULTS

In the following section, we report the results of our study regarding the performance and productivity of participants when performing a collaborative task in VR under different combinations of communication modalities. We report both quantitative results and subjective feedback collected from users along with our findings from NASA-TLX, SUS, SSQ, and presence questionnaires. Furthermore, as we had several comparisons between different conditions, we applied Bonferroni corrections automatically in our analysis.

Table 2: RM-ANOVA statistical effects for various objective metrics collected ($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Measure	F	df_{effect}	df_{error}	p	η_p^2	Sig
Task Completion	12.448	6	144	<0.001	0.342	***
Manipulation	4.850	6	144	<0.001	0.168	***
Selection	13.376	6	144	<0.001	0.358	***

4.1 Objective Results

For this subsection, we provide Table 2 summarizing the findings from the Repeated Measures ANOVA for the objective measures mentioned below.

4.1.1 Task Completion Time. Initially, we used the Shapiro-Wilk Test to check if our data was normally distributed. The test indicated it was not, therefore we normalized our data using the Aligned Rank Transform (ART) [51], and Mauchly's Test of Sphericity ($\chi^2(20) = 23.166, p = 0.286$) indicated that sphericity was not violated. A repeated measures ANOVA ($F_{(6,144)} = 12.448, p < 0.001, \eta_p^2 = 0.342$) indicated that there was statistically significance difference in task completion time across the seven modalities. A Post-Hoc analysis showed that *voice* ($M = 203.42, SD = 10.90$) took significantly longer than *head gaze* ($M = 163.70, SD = 8.13, t_{24} = 4.21, p < 0.001$), *voice + gestures* ($M = 150.42, SD = 8.20, t_{24} = 5.20, p < 0.001$), *voice + head gaze* ($M = 167.25, SD = 7.96, t_{24} = 3.81, p < 0.001$), *gestures + head gaze* ($M = 163.43, SD = 6.48, t_{24} = 3.84, p < 0.001$), and *voice + gestures + head gaze* ($M = 142.54, SD = 7.22, t_{24} = 7.07, p < 0.001$). The same analysis also showed that *gestures* ($M = 202.77, SD = 9.47$) took significantly longer than *head gaze* ($M = 163.70, SD = 8.13, t_{24} = 4.79, p < 0.001$), *voice + gestures* ($M = 150.42, SD = 8.20, t_{24} = 4.32, p < 0.001$), *voice + head gaze* ($M = 167.25, SD = 7.96, t_{24} = 3.54, p = 0.002$), *gestures + head gaze* ($M = 163.43, SD = 6.48, t_{24} = 3.66, p = 0.001$), and *voice + gestures + head gaze* ($M = 142.54, SD = 7.22, t_{24} = 5.54, p < 0.001$). No significant difference was found between all other pairwise comparisons. Table 3 shows the summary of significant pairs found in the Post-Hoc Analysis while Figure 4 shows the average time of all conditions as well as significant pairs.

4.1.2 Object Manipulation Time. Initially, we used the Shapiro-Wilk Test to check if our data was normally distributed. The test indicated it was not, therefore we normalized our data using the Aligned Rank Transform (ART), and Mauchly's Test of Sphericity ($\chi^2(20) = 25.771, p = 0.178$) indicated that sphericity was not violated. A repeated measures ANOVA ($F_{(6,144)} = 4.850, p < 0.001$, $\eta_p^2 = 0.168$) indicated that there was statistically significance difference in object manipulation time across the seven modalities. A Post-Hoc analysis showed that *voice* ($M = 8.37, SD = 1.43$) took significantly longer than *head gaze* ($M = 4.18, SD = 0.35, t_{24} = 3.51, p = 0.002$) and *voice + gestures + head gaze* ($M = 3.49, SD = 0.32, t_{24} = 3.83, p < 0.001$). The same analysis also showed that *gestures* ($M = 6.71, SD = 1.55$) took significantly longer than *voice + gestures* ($M = 3.94, SD = 0.32, t_{24} = 3.62, p = 0.001$) and *voice +*

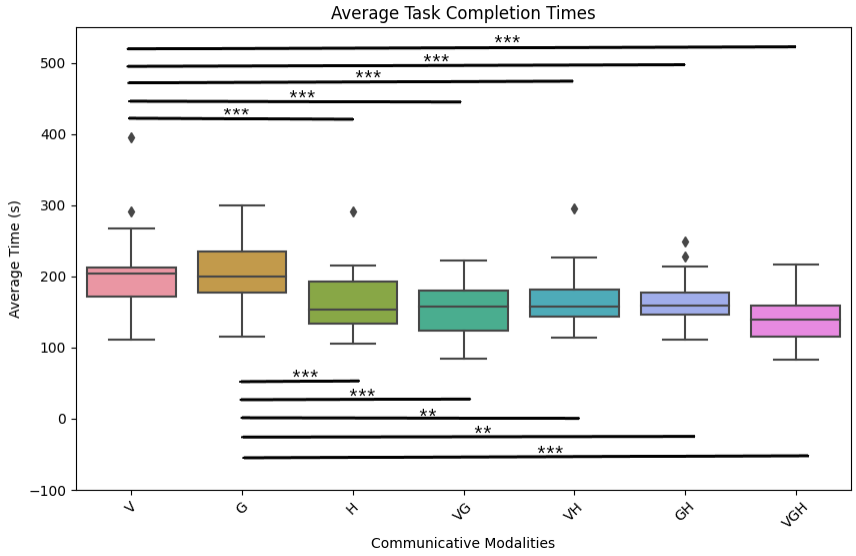

Figure 4: Graph Showing Average Task Completion Times Across All Conditions With Standard Error Bars; Lines connect conditions that were statistically different, with asteriks representing level of significance ($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$). (V - Voice, G - Gestures, H - Head Gaze)

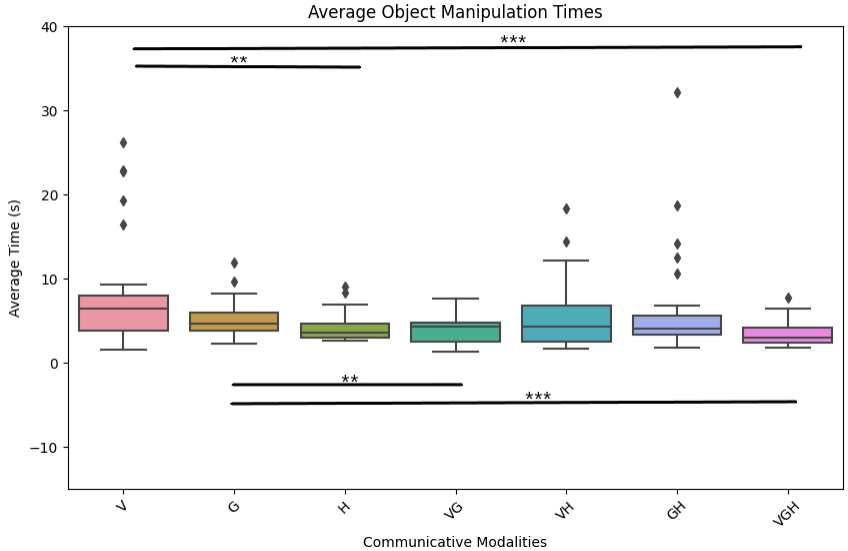

Figure 5: Graph Showing Average Object Manipulation Times Across All Conditions With Standard Error Bars; Lines connect conditions that were statistically different, with asteriks representing level of significance ($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$). Voice had the highest average object manipulation time. (V - Voice, G - Gestures, H - Head Gaze)

gestures + head gaze ($M = 3.49, SD = 0.32, t_{24} = 3.99, p < 0.001$). No significant difference was found between all other pairwise comparisons. Table 4 shows the summary of significant pairs found in the Post-Hoc Analysis while Figure 5 shows the average time of all conditions as well as significant pairs.

4.1.3 Object Selection Time. Initially, we used the Shapiro-Wilk Test to check if our data was normally distributed. The test indicated it was not, therefore we normalized our data using the Aligned Rank Transform (ART), and Mauchly's Test of Sphericity

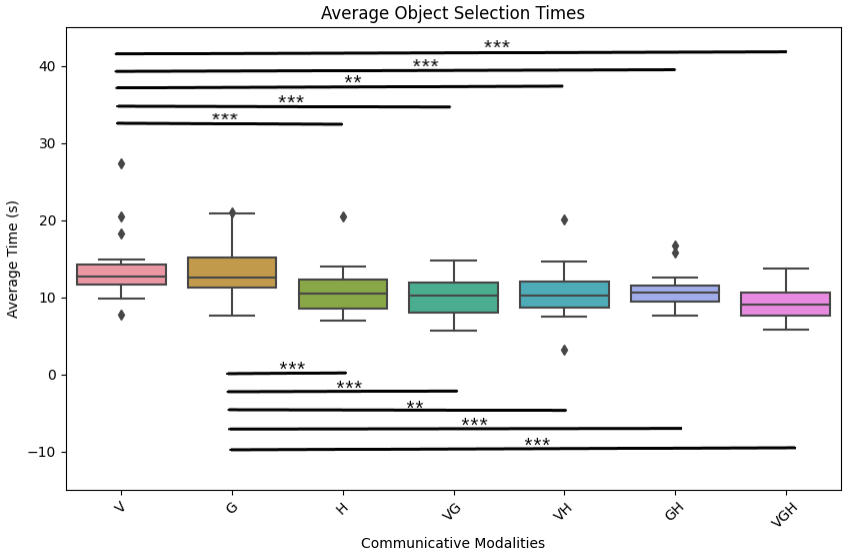

Figure 6: Graph Showing Average Object Selection Times Across All Conditions With Standard Error Bars; Lines connect conditions that were statistically different, with asteriks representing level of significance (* = $p < .05$; ** = $p < .01$; *** = $p < .001$). Both voice and gestures had the highest average object selection time. (V - Voice, G - Gestures, H - Head Gaze)

Table 3: Post-Hoc Analysis Results For Task Performance Times (M_1 refers to the first communication condition in the pair, M_2 refers to the second communication condition in the pair) (V - Voice, G - Gestures, H - Head Gaze)(* = $p < .05$; ** = $p < .01$; *** = $p < .001$)

Pair	M_1 Mean	M_1 SD	M_2 Mean	M_2 SD	t	p	Sig
V-H	203.41	10.9	163.7	8.13	4.21	<0.001	***
V-VG	203.41	10.9	150.42	8.2	5.2	<0.001	***
V-VH	203.41	10.9	167.25	7.96	3.81	<0.001	***
V-GH	203.41	10.9	163.43	6.48	3.84	<0.001	***
V-VGH	203.41	10.9	142.54	7.22	7.07	<0.001	***
G-H	202.77	9.47	163.7	8.13	4.79	<0.001	***
G-VG	202.77	9.47	150.42	8.2	4.32	<0.001	***
G-VH	202.77	9.47	167.25	7.96	3.54	0.002	**
G-GH	202.77	9.47	163.43	6.48	3.66	0.001	**
G-VGH	202.77	9.47	142.54	7.22	5.55	<0.001	***

Table 4: Post-Hoc Analysis Results For Object Manipulation Times (M_1 refers to the first communication condition in the pair, M_2 refers to the second communication condition in the pair) (V - Voice, G - Gestures, H - Head Gaze)(* = $p < .05$; ** = $p < .01$; *** = $p < .001$)

Pair	M_1 Mean	M_1 SD	M_2 Mean	M_2 SD	t	p	Sig
V-H	8.37	1.43	4.18	0.35	3.51	0.002	**
V-VGH	8.37	1.43	3.49	0.32	3.83	<0.001	***
G-VG	6.72	1.55	3.94	0.32	3.62	0.001	**
G-VGH	6.72	1.55	3.49	0.32	3.99	<0.001	***

($\chi^2(20) = 29.495, p = 0.081$) indicated that sphericity was not violated. A repeated measures ANOVA ($F_{(6,144)} = 13.376, p < 0.001$, $\eta_p^2 = 0.358$) indicated that there was statistically significance difference in object selection time across the seven modalities. A Post-Hoc analysis showed that *voice* ($M = 13.50, SD = 0.78$) took significantly longer than *head gaze* ($M = 10.71, SD = 0.57, t_{24} = 4.20, p < 0.001$), *voice + gestures* ($M = 9.94, SD = 0.54, t_{24} = 4.88, p < 0.001$), *voice + head gaze* ($M = 10.76, SD = 0.64, t_{24} = 3.72, p = 0.001$), *gestures + head gaze* ($M = 10.67, SD = 0.43, t_{24} = 4.22, p < 0.001$), and *voice + gestures + head gaze* ($M = 9.27, SD = 0.47, t_{24} = 7.35, p < 0.001$). The same analysis also showed that *gestures* ($M = 13.28, SD = 0.66$) took significantly longer than *head gaze* ($M = 10.71, SD = 0.57, t_{24} = 4.65, p < 0.001$), *voice + gestures* ($M = 9.94, SD = 0.54, t_{24} = 4.30, p < 0.001$), *voice + head gaze* ($M = 10.76, SD = 0.64, t_{24} = 3.52, p = 0.001$), *gestures + head gaze* ($M = 10.67, SD = 0.43, t_{24} = 4.21, p < 0.001$), and *voice + gestures + head gaze* ($M = 9.27, SD = 0.47, t_{24} = 5.66, p < 0.001$). No significant difference was found between all other pairwise comparisons. Table 5 shows the summary of significant pairs found in the Post-Hoc Analysis while Figure 6 shows the average time of all conditions as well as significant pairs.

4.2 Subjective Feedback and Questionnaire Results

4.2.1 Difficulty. Participants were asked to rate the difficulty for completing the task for each of the communication combinations, after which the responses were separated based on them coming from mentor or mentee, then normalized using ART. Mauchly's Test of Sphericity for mentor scores ($\chi^2(20) = 23.166, p < 0.001$, $\eta_p^2 = 0.064$) and mentee scores ($\chi^2(20) = 57.399, p < 0.001$, $\eta_p^2 = 0.297$) indicated that sphericity was violated, so we applied the Greenhouse-Geisser correction. We used a repeated measures

Table 5: Post-Hoc Analysis Results For Object Selection Times (M_1 refers to the first communication condition in the pair, M_2 refers to the second communication condition in the pair) (V - Voice, G - Gestures, H - Head Gaze)($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Pair	M_1 Mean	M_1 SD	M_2 Mean	M_2 SD	t	p	Sig
V-H	13.5	0.78	10.71	0.57	4.2	<0.001	***
V-VG	13.5	0.78	9.94	0.54	4.88	<0.001	***
V-VH	13.5	0.78	10.76	0.64	3.72	0.001	**
V-GH	13.5	0.78	10.57	0.43	4.22	<0.001	***
V-VGH	13.5	0.78	9.27	0.47	7.35	<0.001	***
G-H	13.28	0.66	10.71	0.57	4.65	<0.001	***
G-VG	13.28	0.66	9.94	0.54	4.3	<0.001	***
G-VH	13.28	0.66	10.76	0.64	3.52	0.001	**
G-GH	13.28	0.66	10.57	0.43	4.21	<0.001	***
G-VGH	13.28	0.66	9.27	0.47	5.66	<0.001	***

Table 6: RM-ANOVA statistical effects for difficulty and frustration metrics collected ($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Metric	**F**	df_{effect}	df_{error}	p	η_p^2	**Sig**
Mentor Difficulty	1.628	3.380	81.128	0.143	0.064	no
Mentee Difficulty	10.135	3.314	79.535	<0.001	0.297	***
Mentor Frustration	3.332	3.282	78.770	0.020	0.122	*
Mentee Frustration	9.888	3.365	80.763	<0.001	0.292	***

ANOVA with the Greenhouse-Geisser correction applied to test for difference in both cases. This resulted in no significant differences being found for mentors ($F_{(3.38,81.128)} = 1.628, p = 0.184$, $\eta_p^2 = 0.064$) and significant differences being found for mentees ($F_{(3.314,79.535)} = 10.135, p < 0.001$, $\eta_p^2 = 0.297$). A summary of this analysis can be found in Table 6.

Afterwards, Post-Hoc analysis on mentee responses revealed statistically significant differences between *voice* ($M = 2.36, SD = 0.36$) and *gestures* ($M = 3.80, SD = 0.37, t_{24} = -3.788, p < 0.001$), *gestures* ($M = 3.80, SD = 0.37$) and *voice + gestures* ($M = 2.12, SD = 0.35, t_{24} = 5.04, p < 0.001$), *gestures* and *voice + head gaze* ($M = 2.24, SD = 0.37, t_{24} = 3.96, p < 0.001$), *gestures* and *voice + gestures + head gaze* ($M = 1.84, SD = 0.39, t_{24} = 6.29, p < 0.001$), *head gaze* ($M = 3.12, SD = 0.38$) and *voice + gestures + head gaze* ($M = 1.84, SD = 0.39, t_{24} = 4.50, p < 0.001$), and *gestures + head gaze* ($M = 2.96, SD = 0.34$) and *voice + gestures + head gaze* ($M = 1.84, SD = 0.39, t_{24} = 5.04, p < 0.001$). A summary of this analysis can be found in Table 7.

4.2.2 Frustration. Participants were asked to rate the difficulty for completing the task for each of the communication combinations, after which the responses were separated based on them coming from mentor or mentee, then normalized using ART. Mauchly's Test of Sphericity for mentor scores ($\chi^2(20) = 48.220, p < 0.001$) and mentee scores ($\chi^2(20) = 63.346, p < 0.001$) indicated that sphericity was violated, so we applied the Greenhouse-Geisser correction. We used a repeated measures ANOVA with the Greenhouse-Geisser correction applied to test for difference. This resulted in significant differences being found for both mentors ($F_{(3.282,78.77)} = 3.332, p =$ 0.02, $\eta_p^2 = 0.122$) and mentees ($F_{(3.365,80.763)} = 9.888, p < 0.001$, $\eta_p^2 = 0.292$). A summary of this analysis can be found in Table 6.

Table 7: Post-Hoc Analysis Results For Mentee Difficulty (M_1 refers to the first communication condition in the pair, M_2 refers to the second communication condition in the pair) (V - Voice, G - Gestures, H - Head Gaze)($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Pair	M_1 Mean	M_1 SD	M_2 Mean	M_2 SD	t	p	Sig
V-G	2.36	0.36	3.8	0.37	-3.788	<0.001	***
G-VG	3.8	0.37	2.12	0.35	5.04	<0.001	***
G-VH	3.8	0.37	2.24	0.37	3.96	<0.001	***
G-VGH	3.8	0.37	1.84	0.39	6.29	<0.001	***
H-VGH	3.12	0.38	1.84	0.39	4.5	<0.001	***
GH-VGH	2.96	0.34	1.84	0.39	5.04	<0.001	***

Table 8: Post-Hoc Analysis Results For Mentee Frustration (M_1 refers to the first communication condition in the pair, M_2 refers to the second communication condition in the pair) (V - Voice, G - Gestures, H - Head Gaze)($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Pair	M_1 Mean	M_1 SD	M_2 Mean	M_2 SD	t	p	Sig
G-VG	3.48	0.35	1.84	0.25	4.89	<0.001	***
G-VH	3.48	0.35	1.96	0.31	4.52	<0.001	***
G-VGH	3.48	0.35	1.6	0.31	6.84	<0.001	***
H-VH	2.8	0.33	1.96	0.31	3.48	<0.001	***
H-VGH	2.8	0.33	1.6	0.31	4.51	<0.001	***
GH-VGH	2.48	0.29	1.6	0.31	4.66	<0.001	***

Afterwards, Post-Hoc analysis for the mentor responses resulted in no statistically significant pairs due to the Bonferroni correction applied. However, Post-Hoc analysis on mentee responses revealed statistically significant differences for the mentee for *gestures* ($M = 3.48, SD = 0.35$) and *voice + gestures* ($M = 1.84, SD = 0.25, t_{24} = 4.89, p < 0.001$), *gestures* and *voice + head gaze* ($M = 1.96, SD = 0.31, t_{24} = 4.52, p < 0.001$), *gestures* and *voice + gestures + head gaze* ($M = 1.60, SD = 0.31, t_{24} = 6.84, p < 0.001$), *head gaze* ($M = 2.80, SD = 0.33$) and *voice + head gaze* ($M = 1.96, SD = 0.31, t_{24} = 3.48, p < 0.001$), *head gaze* and *voice + gestures + head gaze* ($M = 1.60, SD = 0.31, t_{24} = 4.51, p < 0.001$), and *gestures + head gaze* ($M = 2.48, SD = 0.29$) and *voice + gestures + head gaze* ($M = 1.60, SD = 0.31, t_{24} = 4.66, p < 0.001$). A summary of this analysis can be found in Table 8.

4.2.3 Most and Least Favorite Communication Combinations. Participants were asked to list what their favorite communication combination was during the tasks (one of the seven provided), after which the responses were separated based on them coming from mentor or mentee. We ran Chi-Squared tests on the mentor responses for most ($\chi_6^2(N = 25) = 16.29, p = 0.01$) and least ($\chi_6^2(N = 25) = 9.69, p = 0.14$) favorite communication combination and mentee responses for most ($\chi_6^2(N = 25) = 14.29, p = 0.03$) and least ($\chi_6^2(N = 25) = 9.99, p = 0.12$) favorite communication combination. This shows that while the Chi-Squared tests for least favorite communication combination for both mentors and mentees

Table 9: Post-Hoc Analysis Results For Presence Responses (V - Voice, G - Gestures, H - Head Gaze)($^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$)

Role	Pair	Z	p	Sig
Mentor	V-H	-2.8	0.005	**
Mentor	V-VG	-2.51	0.012	*
Mentor	V-VH	-2.13	0.033	*
Mentor	V-GH	-2.886	0.004	**
Mentor	V-VGH	-2.051	0.04	*
Mentor	H-VH	-2.054	0.04	*
Mentor	VH-VG	-2.373	0.018	*
Mentee	V-H	-2.358	0.018	*
Mentee	V-VG	-2.468	0.014	*
Mentee	V-VGH	-2.73	0.006	**
Mentee	G-H	-2.038	0.042	*
Mentee	G-VGH	-2.795	0.005	**
Mentee	GH-VGH	-2.752	0.006	**

was not statistically significant, the Chi-Squared tests for most favorite communication combination for both mentors and mentees was statistically significant, therefore showing that scores were not uniformly distributed and that users had more of a preference for communication combination number 7 (*voice + gestures + head gaze*) since it had the highest frequency for each set of responses. Frequencies for these responses can be found in Figure 7.

4.2.4 Social Presence. For social presence, we used four sub-scales of the original social presence survey (co-presence, attentional allocation, perceived message understanding, and perceived behavioral interdependence). We took the average of the scores for each question across all responses then separated responses based on whether they came from a mentor or mentee, after which we used a Friedman Test to test for differences. The result for both mentors ($\chi^2_6(N = 25) = 17.50, p = 0.008$) and mentees ($\chi^2_6(N = 25) = 18.23, p = 0.007$) yielded statistically significant differences.

Afterwards, we ran a Wilcoxon Signed-Rank Test on these averages to test for pairwise significance. For mentors, there was a statistical difference between *voice* and *head gaze* ($Z = -2.800, p = 0.005$), *voice* and *voice + gestures* ($Z = -2.510, p = 0.012$), *voice* and *voice + head gaze* ($Z = -2.130, p = 0.033$), *voice* and *gestures + head gaze* ($Z = -2.886, p = 0.004$), *voice* and *voice + gestures + head gaze* ($Z = -2.051, p = 0.040$), *gestures* and *voice + gestures* ($Z = -2.054, p = 0.040$), and *voice + gestures* and *voice + head gaze* ($Z = -2.373, p = 0.018$). For mentees, there was a statistical difference between *voice* and *head gaze* ($Z = -2.358, p = 0.018$), *voice* and *voice + gestures* ($Z = -2.468, p = 0.014$), *voice* and *voice + gestures + head gaze* ($Z = -2.730, p = 0.006$), *gestures* and *head gaze* ($Z = -2.038, p = 0.042$), *gestures* and *voice + gestures + head gaze* ($Z = -2.795, p = 0.005$), and *gestures + head gaze* and *voice + gestures + head gaze* ($Z = -2.752, p = 0.006$). A summary of this analysis can be found in Table 9.

4.2.5 System Usability (SUS). For system usability, we calculated the system usability score using the formula provided separately for responses given by mentors and mentees. A Friedman Test was used to test for difference. The result for both mentors ($\chi^2_6(N = 25) = 4.92, p = 0.55$) and mentees ($\chi^2_6(N = 25) = 3.45, p = 0.75$) did not yield any statistically significant differences.

4.2.6 Workload (NASA-TLX). For NASA TLX, we asked participants the six questions specified from the original NASA TLX survey and to answer on a scale of one to seven. We used a Friedman Test to test for difference across each question separately for mentors and mentees. This resulted in statistical significance for question 1 ("How mentally demanding was the task?") for mentees ($\chi^2_6(N = 25) = 16.47, p = 0.01$), question 2 ("How physically demanding was the task?") for mentors ($\chi^2_6(N = 25) = 32.97, p < 0.001$), question 5 ("How hard did you have to work to accomplish your level of performance?") for mentors ($\chi^2_6(N = 25) = 23.85, p < 0.001$) and mentees ($\chi^2_6(N = 25) = 27.51, p < 0.001$), and question 6 ("How insecure, discouraged, irritated, stressed, and annoyed were you?") for mentees ($\chi^2_6(N = 25) = 18.68, p = 0.005$).

Afterwards, we ran a Wilcoxon Signed-Rank Test on each of these questions to test for pairwise significance. For question 1 for mentees, there was a statistical difference between *gestures* and *voice + gestures + head gaze* ($Z = -3.115, p = 0.002$). For question 2 for mentors, there was statistical difference between *voice* and *gestures* ($Z = -3.387, p < 0.001$), *gestures* and *head gaze* ($Z = -3.130, p = 0.002$), *gestures* and *voice + gestures* ($Z = -3.196, p = 0.001$), *gestures* and *voice + head gaze* ($Z = -3.414, p < 0.001$), *gestures* and *gestures + head gaze* ($Z = -3.402, p < 0.001$), and *gestures* and *voice + gestures + head gaze* ($Z = -3.402, p < 0.001$). For question 5 for mentors, there was statistical difference between *gestures* and *gestures + head gaze* ($Z = -3.151, p = 0.002$) and *gestures* and *voice + gestures + head gaze* ($Z = -3.045, p = 0.002$). For question 5 for mentees, there was statistical difference between *gestures* and *voice + gestures* ($Z = -3.464, p < 0.001$), *gestures* and *voice + head gaze* ($Z = -3.566, p < 0.001$), and *gestures* and *voice + gestures + head gaze* ($Z = -3.204, p = 0.001$). For question 6 for mentees, there was no statistical difference due to Bonferroni correction. A summary of this analysis can be found in Table 10.

4.2.7 Simulator Sickness (SSQ). For Simulator Sickness, we took the scores of severity of symptoms on a scale of one to four, and used a Friedman Test to test for difference between conditions for mentors and mentees separately. The results for each symptom for both mentors and mentees did not yield any statistically significant differences.

4.2.8 Single Likert Scale Questions. Additional single likert-scale questions for overall communication difficulty, task difficulty, mental engagement and physical engagement were also administered to participants post-study, which allowed both mentors and mentees to rate their experience regarding each of these questions. Frequencies for these responses can be found in Figure 8.

5 DISCUSSION

5.1 Importance of Multiple Modalities

Our first research question focuses on whether participants will perform tasks faster with more cues present. Our findings indicate that H1 was partially true since even though tasks were completed

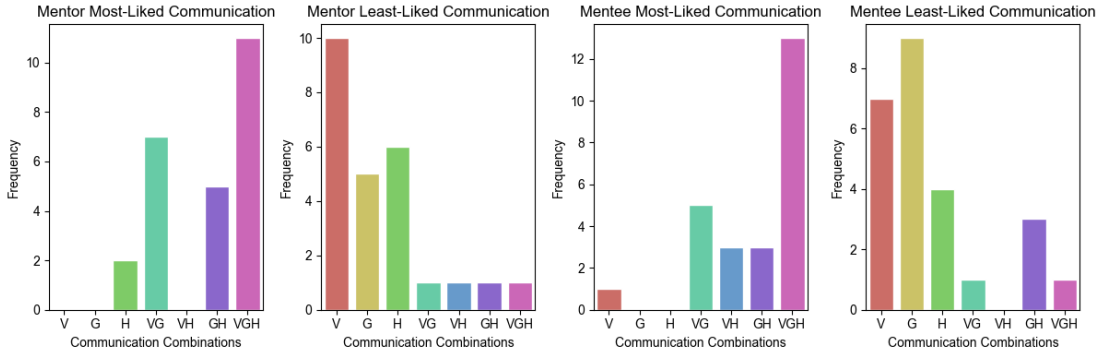

Figure 7: Graphs Showing Mentor and Mentee Responses For Most and Least Favorite Communication Combinations and Modalities. (V - Voice, G - Gestures, H - Head Gaze)

Table 10: Post-Hoc Analysis Results For NASA TLX Responses (V - Voice, G - Gestures, H - Head Gaze)(* = $p < .05$; ** = $p < .01$; *** = $p < .001$)

Role	Question	Pair	Z	p	Sig
Mentee	How mentally demanding was the task?	G-VGH	-3.115	0.002	**
Mentor	How physically demanding was the task?	V-G	-3.387	0.001	**
Mentor	How physically demanding was the task?	G-H	-3.13	0.002	**
Mentor	How physically demanding was the task?	G-VG	-3.196	0.001	**
Mentor	How physically demanding was the task?	G-VH	-3.414	<0.001	***
Mentor	How physically demanding was the task?	G-GH	-3.402	<0.001	***
Mentor	How physically demanding was the task?	G-VGH	-3.402	<0.001	***
Mentor	How hard did you have to work to accomplish your level of performance?	G-GH	-3.151	0.002	**
Mentor	How hard did you have to work to accomplish your level of performance?	G-VGH	-3.045	0.002	**
Mentee	How hard did you have to work to accomplish your level of performance?	G-VG	-3.464	<0.001	***
Mentee	How hard did you have to work to accomplish your level of performance?	G-VH	-3.566	<0.001	***
Mentee	How hard did you have to work to accomplish your level of performance?	G-VGH	-3.204	0.001	**

faster and object selections occurred faster with combinations of communication modalities, *head gaze* did not perform significantly different than conditions with multiple cues combined. This is consistent with findings from previous work [1, 49] which shows that the presence of more cues aids collaboration better. However, what was not shown in previous work is that *head gaze* potentially proves to be sufficient as a means of communication in synchronous assembly VR collaborative tasks compared to *voice* and *gestures*. This could be because with head gaze, users are able to signal both intent and spatial indications for both short and far distances in an environment, making it a versatile method of communication. *Head gaze* was especially useful for mentors to signal which objects the mentee must grab from the shelf and where to place them on the work table.

The performance of *head gaze* in terms of communication has design implications that if developers could include one of the three (voice, gestures, head gaze) in a VR collaborative environment, head gaze would be the best choice. Reasons behind this are that in addition to *head gaze* performing as good as combinations of modalities, head gaze would be an excellent substitute for *voice* in circumstances where speaking is not allowed or certain users are unable to speak in the task setting due to some impairment, etc. Another reason is due to privacy concerns, as *voice* can reveal sounds in a user's physical environment and *gestures* reveal every single movement users make with their hands in physical reality which is projected into the virtual environment, while *head gaze* conveys only the direction of one's head in the environment.

It should be noted that head gaze as a modality works well for two people in this context, as a single person using head gaze in relation to another works well (which was reflected in the results, where conditions with head gaze performed statistically better than *voice* and *gestures*). However, a collaborative task that entails more than two people would cause every individual to experience more cognitive load than a task with only two people. This would be due to each individual having to keep track of multiple head gaze rays and what they are referring to (if anything) rather than a single one. Also in the context of head gaze in the assembly task, head gaze acted as a general indicator for object selection with the simplicity of the task allowing for head gaze to function well. In relation to this, factors such as object occlusion and highly complex, descriptive instructions (e.g. "give me half a liter of this specific solution") would render the head gaze relatively ineffective.

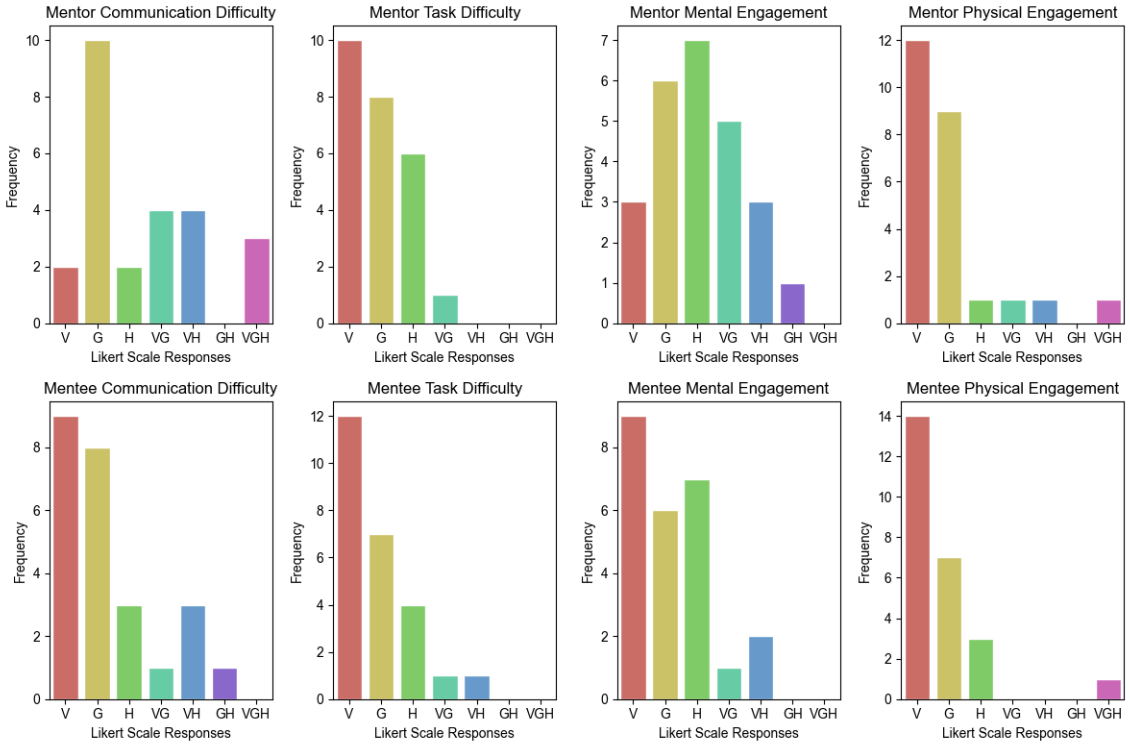

Figure 8: Graphs Showing Mentor and Mentee Responses For Questions of Communication Difficulty, Task Difficulty, Mental Engagement and Physical Engagement (V - Voice, G - Gestures, H - Head Gaze)

5.2 Impact of Visual Cues Present

For our second research question, we also sought to determine whether presence of either gestures or head gaze in any task setting would significantly decrease completion times. Analysis on task completion and object selection times show that our hypothesis is partially true since even though all conditions that involved gestures and/or head gaze performed significantly better than *voice* (which had neither gestures nor head gaze) for these two measures, *gestures* did not perform significantly different than *voice*. This is partially consistent with previous work [1], where head gaze or gestures by itself did not perform significantly different than voice, but the combination of them performed significantly faster than voice. This shows that gestures as sole form of communication would not perform statistically different than solely voice.

However, it should be noted that *voice + gestures* provided the same bandwidth as *head gaze*, meaning that this combination, even though it lacked head gaze, did not perform significantly different than conditions involving head gaze. This result is consistent with previous work [2, 10, 41] which shows that gestures paired with voice performs significantly better than voice only. *Voice* suffered from spatial ambiguities, such as "put the sphere on the left of that cube you just put down, no wait my left, your right" and "put the cube to the top of the green cylinder, umm, oh my bad I meant to the bottom, sorry" as well as the time it took to give verbal descriptions for instructions. *Gestures* suffered from the spatial vagueness it provided for farther distances rather than shorter ones, e.g. when gestures were available to users, mentors were not able to point directly at the objects that they wanted mentees to select from this shelf. This prompted mentors to go back and forth between the shelf and table respectively to show mentees which objects to grab from the shelf and where to place them on the table. This also made mentors more prone to relaying incorrect table placements to mentees since they would be viewing the work table from the mentee's angle instead of their own. However, the *voice + gestures* communication combination allowed mentors to convey both spatial and verbal directions for which objects to grab from the shelf and also where to place them, which reduced completions times and spatial ambiguities as well , e.g. one mentor gave directions as "take a small, blue sphere and put it here", as they used hand gestures to point to an area on the shelf then directly to where they want the object to be placed on the work table.

5.3 Difference of Values Between Mentor and Mentee

In regards to our third research question, we wanted to determine if mentors and mentees have differences in what communication cues they prefer in the task. Per analysis of subjective feedback given by participants, even though mentees experienced significantly more difficulty and frustration in conditions than mentors, both mentors and mentees chose the combination of all three modalities as their favorite condition and the statistical results for SUS indicated no

significant differences for both mentors and mentees, which means that our hypothesis was not supported by our results. This is also shown in the presence results, where both mentors and mentees mainly preferred conditions with combinations of modalities rather than conditions with single modalities, namely *voice* or *gestures*. These results are also consistent with previous work [1, 24] as users felt more present when visual cues were available. This could be attributed to head gaze providing a head as well as gaze direction, in which users felt more present in.

In relation to results based on subjective feedback obtained, we also determined that the role of each user during the task contributed to the separate set of difficulties and frustrations that users felt, as observed by the results from the difficulty and frustration surveys. This is because mentors had to mostly relay directions via communication cues to the mentee (e.g. mentors commented "being mentor was easy, I just had to point with my head or say what goes where" and "I didn't really have to move around much, I could just stay here (their starting position) and tell my friend what they had to do pretty much"), while the mentee had to interpret instructions from the mentor, retrieve objects from the shelf, then place them on the work table. This amount of responsibility had some room for error, mainly due to misinterpretation of instructions from the mentor. For mentors, it was shown through NASA TLX that *gestures* was physically more demanding than any other condition, which is due to the fact that mentors tended to approach the shelf closely and point to objects on it then to the table for where that object had to go. Since this had to happen for every object, mentors found this condition more physically demanding. However, for mentees, there was no statistical significance for a single condition requiring more physical effort than another, which shows that the task itself required about the same amount of physical effort for every condition for mentees. It should also be noted that *gestures* made both mentors and mentees work harder in order to accomplish their level of performance when compared to other conditions, as observed by results from NASA TLX.

5.4 Design Recommendations

5.4.1 Voice. Voice, as mentioned before, is able to deliver precise, descriptive instructions, but suffers from spatial ambiguities as well as the time it takes to deliver detailed instructions. Voice would be best suited for tasks that require some sort of descriptive instructions or explanations about any particular part of a task to accomplish goals associated with the task.

5.4.2 Gestures. Gestures are able to quickly deliver close to mid ranged spatial instructions as well as representational, symbolic, or social cues that convey a specific meaning e.g. thumbs-up meaning that something is good. However, gestures suffer from long range spatial ambiguities as well not being able to provide detailed instructions like voice. Therefore, gestures would be best suited for tasks that involve more close to mid ranged spatial indications as well as physical actions or symbols e.g. using gestures to demonstrate how to turn a knob on a door handle or giving a yes (thumbs-up) respectively.

5.4.3 Head Gaze. Head gaze is able to quickly convey spatial instructions as well as convey where an individual's attention is focused, but suffers from being able to give any sort of detailed instruction aside from a head nod or shake (for yes or no respectively) as well as discerning objects that one is referring to if the object is occluded. Head gaze would be best suited for tasks that require spatial references to objects or places anywhere in the environment as well as simple "yes" or "no" responses to questions.

6 LIMITATIONS AND FUTURE WORK

Our VR collaborative environment with the task employed was meant to evaluate the use of varying availability of communication modalities. However, this means that with the specifics of this study design and prototype, there are limitations to our application.

Throughout every trial, both mentors and mentees had access to the same amount of communication cues. Although this ensured that users experienced the same conditions regardless of whether they were the mentor or mentee, it would be interesting to see if unequal access to communication conditions between mentors and mentees would produce different (and potentially better) task completion times based on the conditions that each individual user has at a given moment in time. Unequal access to communication conditions would also potentially enable mentees to have no communication modalities available for them to use, since they could purely comprehend instructions from the mentor and place objects from the shelf to the work table in that manner.

The task setup for our experiment was asymmetrical, meaning that users had unequal roles in the task and thus communication was more "one sided" with the mentor instructing the mentee on how to assemble the given configuration and the mentee usually communicating confirmation or questions during the task. This poses a limitation as only asymmetric collaboration was investigated. To this end, having users engage in a more symmetrical task with equal roles to investigate the efficiency of these communication conditions would be useful to investigate design implications for a broader range of VR collaborative tasks.

During each trial, objects on the shelf remained in the same positions, so both mentors and mentees would eventually develop "muscle memory" as to where the objects would be on the shelf, thus spatial indications may not work as well in conditions that have voice along with at least one of the visual cues, e.g. the mentor could describe a specific cube that is needed and the mentee immediately goes to the side of the shelf with shapes without the need for much spatial indication. In the future, it would be useful to see how well the spatial indications from *head gaze* or *gestures* perform without mentors and mentees becoming too familiar with the locations of the objects on the shelf for each trial by randomizing the locations.

Although there exists many types of communication cues [13], we chose voice, hand gestures, and head gaze for our investigation. The specific cues mentioned were chosen due to them being natural communication cues as well as the fact that as shown in Table 1, voice, gestures and head gaze were the most commonly selected natural communication cues. Since we wanted to investigate the interactions of the combinations of these cues in depth, we did not include any other cues for this reason as well. However, we acknowledge that we did not investigate interactions of combinations of even more natural communication cues, which include

cues like body pose [5] and avatar embodiment[19, 20, 38, 52]. Future research could expand upon this limitation by conducting an in-depth study with more cues.

Our study was designed to investigate collaboration using these communication cues between two people; despite there being other work that has investigated more than two people collaborating in the same setting [46], our study was meant to explicitly determine how these cues inherently worked and their interaction effects in virtual reality. Thus, we limited the task to a mentor and mentee in this regard, as well as to avoid increasing the cognitive load of users; the presence of multiple people in the same setting would increase the cognitive load placed on all users in the task setting, as each individual would have to process information in the form of voice, gestures and head gaze from multiple individuals from the local environment. Future research could expand upon this limitation by conducting a study with multiple people collaborating in the same environment simultaneously.

The scope of our study design was determined by keeping most human-centered collaborative factors constant [13] in order to investigate the chosen cues in depth, but we acknowledge that the investigation carried out may not necessarily apply to every situation or setting in virtual reality collaborative scenarios. For instance, we did not investigate the use of these cues in asynchronous settings like other work [6, 7, 39, 45], but this design choice was taken due to the time-sensitive information aspect of the task while using these cues. Other design choices not made that would not necessarily yield the same results include varying the access of hardware for individual users and changing the task employed to investigate the selected cues. Future research could expand on this limitation by conducting a study that varies these parameters as factors that entail multiple conditions for synchronous/asynchronous collaboration, the type of task, etc.

7 CONCLUSION

In this paper, we developed a VR collaborative system capable of supporting different communication modalities including voice, hand gestures, and head gaze. Using this application, we evaluated the efficacy of these three communication modalities by having participants complete an asymmetric synchronous collaborative assembly task with varying access to each communication modality. Through our study, we found that task performance for *voice* and *gestures* were not significantly worse than the conditions with combinations of modalities along with head gaze, which was mainly attributed to head gaze being a strong method of communication, along with mentees experiencing more difficulty and frustration than mentors to accomplish the task. These results indicate that head gaze as a single communication method does not perform statistically different from combinations of communication conditions, and a less unequal distribution of workload in asymmetric collaborative tasks could potentially reduce frustration and task completion difficulty. Additionally, our research raises questions for future research to improve communication and collaborative experiences for VR collaborative scenarios.

ACKNOWLEDGMENTS

This work is supported in part by NSF Award IIS-1917728, Northrop Grumman., Unknot.id, and the Florida High Tech Corridor Council Industry Matching Research Program. We also thank the anonymous reviewers for their insightful feedback and the ISUE lab members for their support.

REFERENCES

[1] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[2] Martin Bauer, Gerd Kortuem, and Zary Segall. 1999. " Where are you pointing at?" A study of remote collaboration in a wearable videoconference system. In *Digest of Papers. Third International Symposium on Wearable Computers*. IEEE, 151–158.

[3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] Alisa Burova, John Mäkelä, Hanna Heinonen, Paulina Becerril Palma, Jaakko Hakulinen, Viveka Opas, Sanni Siltanen, Roope Raisamo, and Markku Turunen. 2022. Asynchronous industrial collaboration: How virtual reality and virtual tools aid the process of maintenance method development and documentation creation. *Computers in Industry* 140 (2022), 103663. https://doi.org/10.1016/j.compind.2022.103663

[5] Yuanzhi Cao, Tianyi Wang, Xun Qian, Pawan S. Rao, Manav Wadhawan, Ke Huo, and Karthik Ramani. 2019. GhostAR: A Time-Space Editor for Embodied Authoring of Human-Robot Collaborative Task with Augmented Reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 521–534. https://doi.org/10.1145/3332165.3347902

[6] Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. ProcessAR: An Augmented Reality-Based Tool to Create in-Situ Procedural 2D/3D AR Instructions. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 234–249. https://doi.org/10.1145/3461778.3462126

[7] Subramanian Chidambaram, Sai Swarup Reddy, Matthew Rumple, Ananya Ipsita, Ana Villanueva, Thomas Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2022. EditAR: A Digital Twin Authoring Environment for Creation of AR/VR and Video Instructions from a Single Demonstration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 326–335. https://doi.org/10.1109/ISMAR55827.2022.00048

[8] Julien Epps, Sharon Oviatt, and Fang Chen. 2004. Integration of speech and gesture inputs during multimodal interaction. In *Proc Aust. Int. Conf. on CHI*.

[9] Austin Erickson, Nahal Norouzi, Kangsoo Kim, Ryan Schubert, Jonathan Jules, Joseph J LaViola Jr, Gerd Bruder, and Gregory F Welch. 2020. Sharing gaze rays for visual target identification tasks in collaborative augmented reality. *Journal on Multimodal User Interfaces* 14, 4 (2020), 353–371.

[10] Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. 2004. Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Hum.-Comput. Interact.* 19, 3 (sep 2004), 273–309. https://doi.org/10.1207/s15327051hci1903_3

[11] Lei Gao, Huidong Bai, Gun Lee, and Mark Billinghurst. 2016. An Oriented Point-Cloud View for MR Remote Collaboration. In *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications* (Macau) *(SA '16)*. Association for Computing Machinery, New York, NY, USA, Article 8, 4 pages. https://doi.org/10.1145/2999508.2999531

[12] Danilo Gasques, Janet G. Johnson, Tommy Sharkey, Yuanyuan Feng, Ru Wang, Zhuoqun Robin Xu, Enrique Zavala, Yifei Zhang, Wanze Xie, Xinming Zhang, Konrad Davis, Michael Yip, and Nadir Weibel. 2021. ARTEMIS: A Collaborative Mixed-Reality System for Immersive Surgical Telementoring. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 662, 14 pages. https://doi.org/10.1145/3411764.3445576

[13] Ryan K Ghamandi, Yahya Hmaiti, Tam T Nguyen, Amirpouya Ghasemaghaei, Ravi Kiran Kattoju, Eugene M Taranta, and Joseph J LaViola. 2023. What And How Together: A Taxonomy On 30 Years Of Collaborative Human-Centered XR Tasks. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 322–335.

[14] Kunal Gupta, Gun A. Lee, and Mark Billinghurst. 2016. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 22, 11 (2016), 2413–2422. https://doi.org/10.1109/TVCG.2016.2593778

[15] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, Vol. 2004. Universidad Politecnica de Valencia Valencia, Spain.

[16] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[17] Zhenyi He, Ruofei Du, and Ken Perlin. 2020. CollaboVR: A Reconfigurable Framework for Creative Collaboration in Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 542–554. https://doi.org/10.1109/ISMAR50242.2020.00082

[18] Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You? Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5180–5190. https://doi.org/10.1145/2858036.2858438

[19] Yahya Hmaiti, Mykola Maslych, Eugene M Taranta, and Joseph J LaViola. 2023. An Exploration of The Effects of Head-Centric Rest Frames On Egocentric Distance Judgments in VR. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 263–272.

[20] Adrian H. Hoppe, Florian van de Camp, and Rainer Stiefelhagen. 2021. ShiSha: Enabling Shared Perspective With Face-to-Face Collaboration Using Redirected Avatars in Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 251 (jan 2021), 22 pages. https://doi.org/10.1145/3432950

[21] Allison Jing, Kunal Gupta, Jeremy McDade, Gun A. Lee, and Mark Billinghurst. 2022. Comparing Gaze-Supported Modalities with Empathic Mixed Reality Interfaces in Remote Collaboration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 837–846. https://doi.org/10.1109/ISMAR55827.2022.00102

[22] Allison Jing, Gun Lee, and Mark Billinghurst. 2022. Using Speech to Visualise Shared Gaze Cues in MR Remote Collaboration. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 250–259. https://doi.org/10.1109/VR51125.2022.00044

[23] Allison Jing, Kieran May, Gun Lee, and Mark Billinghurst. 2021. Eye see what you see: Exploring how bi-directional augmented reality gaze visualisation influences co-located symmetric collaboration. *Frontiers in Virtual Reality* 2 (2021), 697367.

[24] Allison Jing, Kieran May, Brandon Matthews, Gun Lee, and Mark Billinghurst. 2022. The Impact of Sharing Gaze Behaviours in Collaborative Mixed Reality. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 463 (nov 2022), 27 pages. https://doi.org/10.1145/3555564

[25] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 283, 7 pages. https://doi.org/10.1145/3411763.3451844

[26] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[27] Seungwon Kim, Gun Lee, Mark Billinghurst, and Weidong Huang. 2020. The combination of visual communication cues in mixed reality remote collaboration. *Journal on Multimodal User Interfaces* 14 (2020), 321–335.

[28] Seungwon Kim, Gun Lee, Weidong Huang, Hayun Kim, Woontack Woo, and Mark Billinghurst. 2019. Evaluating the Combination of Visual Communication Cues for HMD-Based Mixed Reality Remote Collaboration. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300403

[29] Seungwon Kim, Gun Lee, Nobuchika Sakata, and Mark Billinghurst. 2014. Improving co-presence with augmented visual communication cues for sharing experience through video conference. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 83–92. https://doi.org/10.1109/ISMAR.2014.6948412

[30] David Kirk, Andy Crabtree, and Tom Rodden. 2005. Ways of the Hands. 1–21. https://doi.org/10.1007/1-4020-4023-7_1

[31] Lucie Kruse, Joel Wittig, Sebastian Finnern, Melvin Gundlach, Niclas Iserlohe, Oscar Ariza, and Frank Steinicke. 2023. Blended Collaboration: Communication and Cooperation Between Two Users Across the Reality-Virtuality Continuum. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 54, 8 pages. https://doi.org/10.1145/3544549.3585881

[32] Gustav Kuhn, Benjamin W Tatler, and Geoff G Cole. 2009. You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition* 17, 6-7 (2009), 925–944.

[33] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A User Study on MR Remote Collaboration Using Live 360 Video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 153–164. https://doi.org/10.1109/ISMAR.2018.00051

[34] Yuan Li, Feiyu Lu, Wallace S Lages, and Doug Bowman. 2019. Gaze Direction Visualization Techniques for Collaborative Wide-Area Model-Free Augmented Reality. In *Symposium on Spatial User Interaction* (New Orleans, LA, USA) *(SUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 11, 11 pages. https://doi.org/10.1145/3357251.3357583

[35] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* 77, 12 (1994), 1321–1329.

[36] Jens Müller, Roman Rädle, and Harald Reiterer. 2016. Virtual Objects as Spatial Cues in Collaborative Mixed Reality Environments: How They Shape Communication Behavior and User Task Load. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1245–1249. https://doi.org/10.1145/2858036.2858043

[37] Sharon Oviatt. 2017. *Theoretical Foundations of Multimodal Interfaces and Systems*. Association for Computing Machinery; Claypool, 19–50. https://doi.org/10.1145/3015783.3015786

[38] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2017. [POSTER] CoVAR: Mixed-Platform Remote Collaborative Augmented and Virtual Realities System with Shared Collaboration Cues. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. 218–219. https://doi.org/10.1109/ISMAR-Adjunct.2017.72

[39] Troels Rasmussen. 2023. AuthAR - Automatic Authoring of Picking and Layout Optimization Instructions. In *Proceedings of the 34th Australian Conference on Human-Computer Interaction* (<conf-loc>, <city>Canberra</city>, <state>ACT</state>, <country>Australia</country>, </conf-loc>) *(OzCHI '22)*. Association for Computing Machinery, New York, NY, USA, 199–205. https://doi.org/10.1145/3572921.3572949

[40] Thomas C Scott-Phillips. 2008. Defining biological communication. *Journal of evolutionary biology* 21, 2 (2008), 387–395.

[41] Franco Tecchia, Leila Alem, and Weidong Huang. 2012. 3D Helping Hands: A Gesture Based MR System for Remote Collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry* (Singapore, Singapore) *(VRCAI '12)*. Association for Computing Machinery, New York, NY, USA, 323–328. https://doi.org/10.1145/2407516.2407590

[42] Theophilus Teo, Gun A Lee, Mark Billinghurst, and Matt Adcock. 2018. Hand gestures and visual annotation in live 360 panorama-based mixed reality remote collaboration. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. 406–410.

[43] Theophilus Teo, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. Investigating the Use of Different Visual Cues to Improve Social Presence within a 360 Mixed Reality Remote Collaboration*. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry* (Brisbane, QLD, Australia) *(VRCAI '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3359997.3365687

[44] Jean-Philippe Thiran, Ferran Marques, and Hervé Bourlard. 2009. *Multimodal Signal Processing: Theory and applications for human-computer interaction*. Academic Press.

[45] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 161–174. https://doi.org/10.1145/3332165.3347872

[46] Balasaravanan Thoravi Kumaravel and Andrew D Wilson. 2022. DreamStream: Immersive and Interactive Spectating in VR. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 636, 17 pages. https://doi.org/10.1145/3491102.3517508

[47] Huayuan Tian, Gun A. Lee, Huidong Bai, and Mark Billinghurst. 2023. Using Virtual Replicas to Improve Mixed Reality Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2785–2795. https://doi.org/10.1109/TVCG.2023.3247113

[48] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Xiangyu Zhang, Shuxia Wang, Weiping He, Yuxiang Yan, and Hongyu Ji. 2021. AR/MR Remote Collaboration on Physical Tasks: A Review. *Robotics and Computer-Integrated Manufacturing* 72 (2021), 102071. https://doi.org/10.1016/j.rcim.2020.102071

[49] Peng Wang, Yue Wang, Mark Billinghurst, Huizhen Yang, Peng Xu, and Yanhong Li. 2023. BeHere: a VR/SAR remote collaboration system based on virtual replicas sharing gesture and avatar in a procedural task. *Virtual Reality* (2023), 1–22.

[50] Peng Wang, Shusheng Zhang, Xiaoliang Bai, Mark Billinghurst, Weiping He, Shuxia Wang, Xiaokun Zhang, Jiaxiang Du, and Yongxing Chen. 2019. Head Pointer or Eye Gaze: Which Helps More in MR Remote Collaboration?. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1219–1220. https://doi.org/10.1109/VR.2019.8798024

[51] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[52] Boram Yoon, Hyung-il Kim, Gun A. Lee, Mark Billinghurst, and Woontack Woo. 2019. The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 547–556. https://doi.org/10.1109/VR.2019.8797719