


# Comparison of Audio and Visual Cues to Support Remote Guidance in Immersive Environments

Fei Wu<sup>1</sup>, Jerald Thomas<sup>1</sup>, Shreyas Chinnola<sup>2</sup>, and Evan Suma Rosenberg<sup>1</sup> 

<sup>1</sup>University of Minnesota

<sup>2</sup>Wayzata High School

---

## Abstract

*Collaborative virtual environments provide the ability for collocated and remote participants to communicate and share information with each other. For example, immersive technologies can be used to facilitate collaborative guidance during navigation of an unfamiliar environment. However, the design space of 3D user interfaces for supporting collaborative guidance tasks, along with the advantages and disadvantages of different immersive communication modalities to support these tasks, are not well understood. In this paper, we investigate three different methods for providing assistance (visual-only, audio-only, and combined audio/visual cues) using an asymmetric collaborative guidance task. We developed a novel experimental design and virtual reality scenario to evaluate task performance during navigation of a complex and dynamic environment while simultaneously avoiding observation by patrolling sentries. Two experiments were conducted: a dyadic study conducted at a large public event and a controlled lab study using a confederate. Combined audio/visual guidance cues were rated easier to use and more effectively facilitated the avoidance of sentries compared with the audio-only condition. The presented work has the potential to inform the design of future experiments and applications that involve communication modalities to support collaborative guidance tasks with immersive technologies.*

## CCS Concepts

• **Human-centered computing** → **Virtual reality; Collaborative interaction;**

---

## 1. Introduction

Collaborative virtual environments provide the opportunity for multiple co-located or remote individuals to communicate, share information, and cooperatively accomplish tasks together in a shared virtual world. This paradigm enables real-time information transitions between shared and individual activities, flexible and multiple viewpoints of the environment, shared context in asynchronous work, and improved awareness of collaborators [CS98]. Immersive collaboration has a number of emerging applications, such as social entertainment, commercial activities for remote consumers, surgical simulation, distance learning, architectural design, and training in dangerous situations.

Remote guidance is a form of collaboration that can enable users to receive instructions and assistance while navigating an unfamiliar environment or performing other types of complex 3D interaction tasks. For example, navigation is a fundamental interaction task in both the real world and virtual environments, commonly employed for exploration, searching, or maneuvering to target locations [LJKM\*17]. However, a person with only a local first-person perspective may have limited spatial knowledge of the environment, especially if it is unfamiliar. Navigation may become even more difficult if the environment is dynamic, with moving obstacles or adversaries that cause the optimal path to the goal to contin-

ually change. In certain applications, it may be useful to facilitate collaborative guidance, which can enable users to receive instructions or assistance from remote individuals that either have expert knowledge or access to additional streams of information, such as a map-like view of the environment.

Multi-user collaboration requires technology-mediated communication, and the specific interaction capabilities and communication modalities supported by the system will influence the performance and effectiveness of the task. A variety of communication methods have been proposed to facilitate collaborative navigation tasks, such as verbal instructions [SJS03], egocentric-exocentric perspectives [YO02], arrows [NDF13], light sources [CRdS\*12] [NDF13], and audio. However, much of the previous research has focused on exploring different variations of a single communication modality, such as visual cues. Relatively few studies have been conducted to compare the performance of multiple communication modalities (e.g., audio, visual, or combined audio/visual guidance cues) to support collaborative assistance for navigation tasks in unfamiliar, dynamic environments.

In this paper, we present two studies that investigated the efficacy of different communication modalities to facilitate remote guidance during navigation of a complex, dynamic environment. The experimental scenario was designed in collaboration with the U.S. Army

Research Laboratory and was motivated by emerging applications of augmented reality technology in an operational context. In such scenarios, forward-deployed operators need to navigate through dynamically changing environments while receiving real-time communication from remote team members that have more immediate access to information streams such as satellite imagery or video captured from unmanned aerial vehicles. These tasks become even more challenging when adversaries are present in the operational environment.

In this work, we utilize immersive virtual reality to simulate future augmented reality capabilities that are impractical or difficult with current-generation technology, which is limited by field-of-view, occlusion cue support, outdoor brightness, etc. Having reached a relative level of maturity, virtual reality technologies can provide controlled environments in which future augmented reality user interfaces (e.g., visual overlays) can be prototyped and evaluated practically and cost-effectively. Specifically, we examined the use of audio instructions and visual overlays, which are arguably the two most commonly employed categories of guidance cues. To this end, we designed a two-user collaborative navigation task with three conditions: audio-only, visual-only, and a combination of audio and visual guidance. This paper is based upon work that presented at a recent workshop [Ano20]. Its major contributions include:

- An experimental user interface for evaluating collaborative guidance tasks using visual overlays and pre-recorded voice instructions.
- An immersive virtual reality scenario to evaluate collaborative guidance during navigation of a complex and dynamic virtual environment while simultaneously avoiding observation by patrolling sentries.
- A dyadic study conducted at a large public event in which pairs of participants completed the task cooperatively, with one participant assigned to the role of an explorer, and the other participant acting as a guider.
- A controlled lab study in which single participants were guided by a trained confederate to complete the task.

## 2. Related Work

### 2.1. Immersive Collaboration

Researchers have long studied the collaboration work in VR which mainly focused on the communication and interaction tasks. Collaborative Virtual Environments (CVEs), which facilitate communication and information sharing through the computer, were described in detail in [CSM12]. The state-of-art of CVEs were described in [WM08] and [CS98]. Synchronous co-operative work was studied in [DSD\*99] for information display in a virtual workspace. Applications of CVEs were designed to support real-time work and interaction with objects and artifacts in [HFH\*98], such as transfer of knowledge and skills [GS18] and even communication with virtual avatars [BPF\*19]. However, most work developed about collaborative tasks has focused on information sharing instead of navigation. Khalid, et al. [KUA\*19] is the closest work about navigation aids in CVEs. This study compared the performance across four different guidance navigation aids i.e.

3-Dimensional Map-Liner (3DML), Audio, Textual, and Arrows-Casting, but it only focused on the communication among users who were all immersed in the virtual world.

### 2.2. Communication Modalities

Many studies have compared different communication modalities, mainly focusing on audio and visual methods. The phenomenon that participants fail to respond to the auditory component of the bimodal targets significantly more often than they fail to respond to the visual component, which is known as the Colavita visual dominance effect, was illustrated in many sensory dominance studies [HR09] [SPC12] [GFMF\*17]. Some research revealed that this effect is partly due to a significant decrease in participants' sensitivity to auditory stimuli when presented concurrently with visual stimuli [KLS09]. Additionally, some studies also included the comparison of users' reactions among audio signals, visual signal and the combination of audio and visual signals. It was found that humans reacted more rapidly to the combination of auditory or visual features compared with either auditory or visual features alone [FDPG02].

In the majority of collaborative tasks in VR, users usually don't have an opportunity to communicate face-to-face, but instead, utilize computer-mediated communication (CMC) [Wal11]. Many studies have compared the task performance using CMC modalities against face-to-face communication methods [Ala94] [BOG\*02]. However, the task performance may be determined not only by CMC modalities but by some other factors as well, such as the ability of individuals to perform a task [Bub01], whether participants trust their partners [Wal96] and the relationship between participants [BOG\*02]. These factors should be taken into consideration during experiment design and analysis.

### 2.3. Navigation Aids

Wayfinding is a fundamental task during navigation [DS96]. Several methods have been proposed to explore efficient techniques to overcome the problem of lacking cognitive maps during the wayfinding process [DA\*08]. The most classical method was to design and place several landmarks [ENK97] [Vin99] or other fixed navigation aids [CS99] in VEs to provide spatial orientation and knowledge to users. However, the fixed navigation aids only allow users to explore certain routes, which lacks flexibility. This problem can be overcome by providing an auto-generated shortest path from the user to the destination [RS04]. An improved method was proposed by providing both off-line pre-computed navigation information and an on-line tour guideline which automatically was generated during the exploration to increase computational efficiency [ETT07]. Auditory signals are also widely applied for spatial localization in virtual environments [BGFTJ\*18] [CMF\*20].

### 2.4. Collaborative Navigation

At the IEEE 3DUI 2012 Contest, several researchers introduced user interfaces for collaborative navigation assistance in a 3D environment. A 3DUI game was created to explore collaborative navigation tasks, in which a partner was responsible for providing

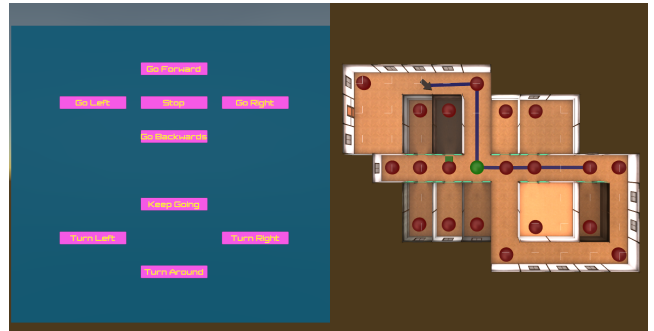
nonverbal communication to the explorer, such as lighting up a path [NDWG\*12], point light sources [CRdS\*12] and waypoint beacons [WBL\*12]. Stafford, et al. [SPT06] developed a metaphor where indoor users provided navigational instructions by simply pointing to an area on the table surface which was captured by cameras and sent to the outdoor users with augmented reality in real-time. Yang, et al. designed a collaborative navigation model in a virtual environment to evaluate the effect of different perspectives, but only visual signals were used in the experiment [YO02]. In [SJS03], only verbal commands were applied between the commander and the operator. Bacim et al. presented a methodology applying visual waypoints in a rescue task and some preliminary results, but no conclusive results have been published [BRS\*12]. The study conducted in [NDF13] presented an Immersive Interactive Virtual Cabin (IIVC) model [DNF\*14] and compared three visual techniques (directional arrow, light source, and compass) that relied on the model, but didn't involve any audio techniques.

### 3. Guiding Techniques

In the experimental scenario used in this paper, participants served two kinds of roles: guiders and explorers. The guider had access to a top-view map of the virtual building that indicated the position and orientation of the explorer and was in charge of sending instructions via a user interface. The explorer navigated in the virtual building based on the instructions sent by the guider. Three conditions were designed regarding the mode of communication between the guider and the explorer: audio instructions, visual overlays, and combined visual/audio cues. The design details are presented below.

**3.0.0.1. Audio** The first condition was only using audio instructions. The ten most common directional commands in daily life were applied in the experiment, including go forward, go backwards, turn left, turn right, turn around, go left (i.e., turn left then go forward), go right, keep going, stop, and stay here. The ten commands were selected based on several tests to meet the needs for every possible situation during navigation in our virtual reality scenario. These commands were converted to audio by an online audio generator [TTS]. A panel was designed to contain ten buttons (see the left side of Figure 1). Each button included one command. When clicking the button, the corresponding audio command would be sent to the headphone on the VR headset. The explorers were required to execute corresponding actions once they received the audio commands.

**3.0.0.2. Visual** The second condition was only using visual overlays. Waypoints were used as visual overlays because they are easy to recognize for explorers and flexible to use for guiders. Waypoints can provide both direction and target information which outperform arrows or other visual overlays with only directional cues. A waypoint was located at each corner and intersection of each path and the entrance of each room. The right part in Figure 1 was a top view map of the virtual environment; the red circles represented waypoints. All waypoints were invisible to explorers in default. When the guider clicked one red circle, a corresponding waypoint would be activated and appear at the same location in the virtual environment as shown in Figure 2. Each waypoint automatically disappeared when another waypoint was activated or deactivated



**Figure 1:** The operation interface built for the guider. The left part was the panel containing all audio buttons. The right part was the 3D map of the virtual building. The circles in the 3D map represented the waypoints. Red circles were inactive while green was active. The blue line represented the shortest path from the explorer's current location to the destination. The black arrow represented the explorer and was pointing the VR headset's orientation. The square represented the sentry.

manually by the guider. The waypoint formed a temporary target for the explorer, who was required to move towards and stop at the location of it. The waypoint might appear in front of, behind or next to the explorer.

**3.0.0.3. Combination** The Combination condition referred to the communication of both audio and visual cues, as described above. The guider needed to send audio commands and light up waypoints at the same time. Specifically, the guider selected one type of audio/visual command first, and then initiated the other kind of command as quickly as possible. This process was designed to be counterbalancing that the participants with odd numbers sent the audio commands first, and the even numbers sent the visual commands first. Two commands couldn't conflict with each other. For example, the guider couldn't send the voice command "go forward" and activate a waypoint behind the explorer at the same time.

## 4. Environment and Interface Design

The virtual environment was designed based on an architectural floor plan of a real building with complex hallways and rooms to explore. The building materials were provided by an open-source asset in Unity [cor]. Two kinds of views were built for the explorer and guider. The virtual building had the same height as the real building in the physical world to strengthen the sense of presence. Each room had at least one opened door for entrance. We also increased the difficulty of the task by designing the environment to be dynamic. Three sentries were created using an open-source asset [dra18] who wandered in the building randomly and had a broad field of view in all directions. The initial positions of sentries were located at three different waypoints which were far away from each other and the participant's starting location. The guider needed to help the explorer avoid being seen by the sentries.

A first-person perspective view was built for the explorer. Figure 2 showed a scene seen by the explorer. The currently active



**Figure 2:** A first-person perspective view of the virtual building for the explorer. The green waypoint represented an active waypoint.

waypoint was displayed in green; all other waypoints were invisible. A bird's eye view was created for the guider which could be considered as a 3D map of the virtual building. It showed a top view of the whole building including all waypoints to the guider, the destination as well as the explorer's current location and orientation.

The screen for the guider was divided into two parts, as shown in Figure 1. The left side contained the panel containing all audio buttons. The right side contained the 3D map of the virtual building. The red circles in the 3D map represented the waypoints. When a waypoint was activated, it would turn to green. The blue line represented the shortest path from the explorer's current location to the destination which was automatically generated and calculated using the Dijkstra algorithm. It was only visible to the guider and provided a visualization of the optimal route. The black arrow represented the explorer's position and orientation. The square represented the sentry. It would turn red when it saw the explorer. There are two kinds of operations for the guider to manipulate the interface: click the audio buttons to send audio instructions and click the red circles to activate waypoints.

## 5. Experiment 1

The first experiment was a dyadic study conducted at the Driven to Discover (D2D) event in the 2019 Minnesota State Fair, which is a large public event. A total of 34 pairs of volunteers were recruited during the event. One participant played the role of the guider while the other one served as the explorer. The 68 participants included 36 males and 32 females who were aged from 18 to 72 years old ( $M = 32.54$ ,  $SD = 15.40$ ). Thirty participants reported no prior experience playing 3D video games. Two reported that they had been playing video games for 1-4 years. Four participants had been playing video games for 5-9 years, and 30 reported that they have been playing for 10 or more years. Participants were required to have a normal or corrected-to-normal vision and be able to communicate in spoken and written English. Each participant was compensated with a cloth bag containing the university logo.

### 5.1. Equipment

The explorer experienced the virtual environment using a Valve Index Headset and controllers. The headset provides a stereoscopic view with a resolution of  $1440 \times 1600$  per eye, a refresh rate of 120Hz, and an expanded field-of-view of around 130-degree. The demo was implemented in Unity. The experiment was run on an Intel Core i9-9900k 3.60GHz PC running Windows 10 Pro with 64 GB of RAM and an NVIDIA GeForce RTX 2080 Ti graphics card. Both eyes were rendered at approximately 150 frames per second. The guider used a desktop interface displayed on the same computer and manipulated the user interface using a mouse.

### 5.2. Study Design

Experiment 1 consisted of a within-subject study with three trials. Each trial utilized one of the communication conditions and was counter-balanced. Participants were instructed to explore the virtual environment as described in Section 4. The start and end waypoints were pre-defined and different for each trial. Each pair of start and end waypoints appeared randomly. The guider needed to use the mouse to click buttons on the panel to send audio commands and the waypoints in the 3D map to show the temporary target to the explorer. The guider was instructed to send audio and visual signals coherently and consistently. Neither the guider nor the explorer had knowledge about the condition they were currently experienced. The guider was required to balance two tasks. The first one was to guide the explorer to reach multiple specified target locations as quickly as possible. At the same time, the guider needed to prevent the explorer from being seen by sentries.

The explorer navigated the virtual building using the Valve Index controllers. Locomotion was implemented using gaze directed travel, that is, the user moved towards their view direction. The navigation speed was maintained at 2.5 meters per frame. This speed was selected based on multiple testing to reduce motion sickness. To decrease the influence of simulator sickness, the explorers were allowed to turn physically. The explorer had to follow the commands from the guider and couldn't avoid sentries or choose routes independently.

We measured task performance using two criteria: navigation duration and sentries duration. Navigation duration was the amount of time that the explorers spent to finish all tasks. Sentries duration was the amount of time that the explorers were viewed by the sentries.

We also asked participants to fill in a feedback-questionnaire regarding their experiences in the virtual environment. The feedback-questionnaire included three parts. In the first part, participants were asked to rate their feeling for the three conditions based on two criteria, effectiveness and easy to use. The rating used a 7-point Likert scale from 1="strongly disagree" to 7="strongly agree." There were six questions as shown in Table 1.

In the second part, we asked the participants to describe the factors that made the task easier and the factors that made the task more difficult. At the end of the feedback-questionnaire, some free-response questions were included to gather comments and suggestions.

Our hypotheses were as follows:



**Table 1: Feedback Questionnaire.** Participants were required to rate three conditions in the aspects of effectiveness and easy to use with a 7-point Likert scale based on their experience.

Questions	Questions
Q1	The visual markers were easy to use.
Q2	The visual markers were an effective method for guidance through the virtual environment.
Q3	The audio prompts were easy to use.
Q4	The audio prompts were an effective method for guidance through the virtual environment.
Q5	The combination of visual markers and audio prompts were easy to use.
Q6	The combination of visual markers and audio prompts were an effective method for guidance through the virtual environment.

- H1: The Combination conditions would result in shorter navigation duration compared with the Visual condition and the Audio condition.
- H2: The Combination conditions would have shorter sentries duration compared with the Visual condition and the Audio condition.
- H3: The Combination conditions would be rated easier to use and more effective compared with the Visual condition and the Audio condition.

### 5.3. Procedure

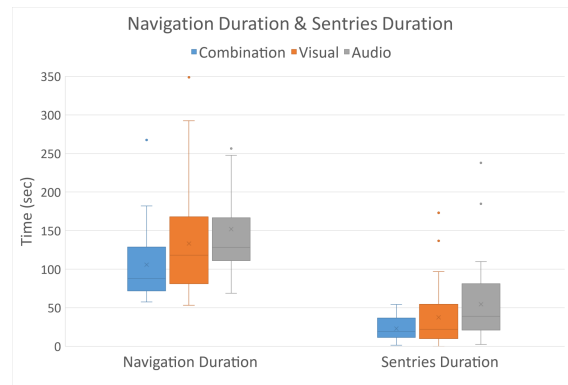
Participants filled out the informed consent form. Then the explorer completed the Simulator Sickness Questionnaire (SSQ) [KLBL93]. After completing the pre-questionnaire, the task was explained to the participants. The guider was shown how to use the interface on the computer screen while the explorer was shown the usage of the controller. Each pair would then perform a short practice trial to ensure that they understood the control mechanisms of the virtual environment. The practice trial was expected to take one minute.

After the practice trial finished, three formal trials were triggered automatically, one of each condition. Each trial was designed to take approximately two minutes. During the trials, the virtual reality system automatically collected information, including the explorer’s positions and orientations in each frame, the navigation duration, and the sentries duration. After the experiment session, the participants completed a demographic questionnaire and the feedback-questionnaire. The explorer also needed to complete the SSQ post-test,

Because participants were attending a public event, the study was designed to take approximately 20 minutes in VR and 5 minutes for questionnaires. However, based on the different behavior of participants, the entire experiment duration varied from 25 to 45 minutes including informed consent, VR exposure, and questionnaires.

### 5.4. Results

A Shapiro-Wilk W test was conducted as a test of normality for all variables. The results indicated that the data was not normally distributed. Therefore, we reported medians (Mdn) and interquartile ranges (IQR). Since our experiment was within-subjects and non-parametric, we applied the Friedman Rank Test to analyze the difference between three conditions. All statistical results used a significance value of  $\alpha = 0.05$ . If the Friedman Rank Test rejected



**Figure 3: Results for navigation duration and sentries duration in Experiment 1.** Boxplots represent the median and IQR.

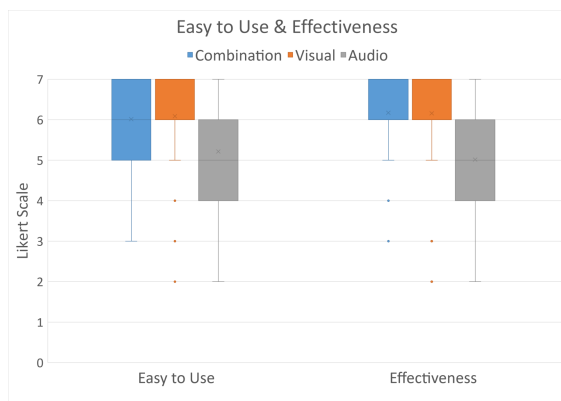
the null hypothesis, we applied the Wilcoxon test with a Bonferroni correction as the post-hoc analysis for each pair of conditions.

The analysis of navigation duration Figure 3 indicated that the Combination condition ( $Mdn = 87.28, IQR = 59.02$ ) had the lowest median value compared with the Symbology condition ( $Mdn = 117.32, IQR = 87.43$ ) and the Audio condition ( $Mdn = 122.53, IQR = 42.26$ ), but there was no significant difference among the three conditions  $\chi^2(2) = 2.24, p = 0.091$ .

Analysis in Figure 3 showed that there existed statistically significant difference in terms of the sentries duration among the three conditions  $\chi^2(2) = 7.60, p = 0.022$ . In addition, the Wilcoxon test revealed that the time in the Combination condition ( $Mdn = 19.63, IQR = 25.03$ ) was significantly shorter than the Audio condition ( $Mdn = 38.78, IQR = 57.80$ ),  $Z = -4.06, p < 0.001$ . However, the Combination condition ( $Mdn = 19.63, IQR = 25.03$ ) didn’t have significantly different sentries duration compared with the Visual condition ( $Mdn = 21.37, IQR = 47.27$ ),  $Z = -2.11, p = 0.035$ . There was also no significant difference between the Visual condition ( $Mdn = 21.37, IQR = 47.27$ ) and the Audio condition ( $Mdn = 38.78, IQR = 57.80$ ),  $Z = 1.93, p = 0.054$ .

The feedback-questionnaire responses supported the experiment data as shown in Figure 4. The Friedman test showed that there was significant difference among the three conditions in the criteria of easy to use,  $\chi^2(2) = 10.20, p = 0.0061$ . Post-hoc analysis indicated that the Combination condition ( $Mdn = 6, IQR = 2$ ) was significant easier to use than the Audio condition ( $Mdn = 5, IQR = 2$ ),  $Z = -4.72, p < 0.001$ . The Visual condition ( $Mdn = 6, IQR = 1$ ) also was significant easier to use than the Audio condition ( $Mdn = 5, IQR = 2$ ),  $Z = -5.14, p < 0.001$ . But there was no significant difference between the Visual condition ( $Mdn = 6, IQR = 1$ ) and the Combination condition ( $Mdn = 6, IQR = 2$ ),  $Z = 0.59, p = 0.553$ .

The analysis of effectiveness showed similar result. There was significant difference among the three conditions,  $\chi^2(2) = 17.89, p < 0.001$ . Analysis indicated that the Visual condition ( $Mdn = 6, IQR = 1$ ) was significant higher in effectiveness compared with the



**Figure 4:** Responses to the feedback-questionnaire in Experiment 1. Boxplots represent the median and IQR (in a 7-point Likert scale, 1=“strongly disagree” and 7=“strongly agree”).

Audio condition ( $Mdn = 5, IQR = 2$ ),  $Z = -6.10, p < 0.001$ . The Combination condition ( $Mdn = 6, IQR = 1$ ) also had better effectiveness compared with the Audio condition ( $Mdn = 5, IQR = 2$ ),  $Z = -6.36, p < 0.001$ . No significant difference existed between the Combination condition ( $Mdn = 6, IQR = 1$ ) and the Visual condition ( $Mdn = 6, IQR = 1$ ),  $Z = 0.355, p = 0.722$ .

The simulator sickness scores were calculated based on [WLM\*19]. The Wilcoxon test was performed to compare the scores between pre-experiment and post-experiment. The result showed that participants felt significantly higher levels of simulator sickness after the experiment ( $Mdn = 3.74, IQR = 7.48$ ) than before the experiment ( $Mdn = 0, IQR = 3.74$ ),  $Z = -3.32, p = 0.001$ .

### 5.5. Discussion

There was no significant difference among the three conditions in terms of the navigation time. The statistical analysis revealed consistent results with our observation during the experiment. Most guiders focused more on sentries duration instead of navigation duration. Although they were told that they should try to balance their two tasks, most guiders chose to sacrifice the navigation duration when two tasks conflicted with each other.

Based on the above reason, the results for sentries duration measurements had significant differences among the three conditions. The Combination condition showed better performance against the Audio conditions but no significant difference compared with the Visual condition. Participants also claimed in the feedback-questionnaires that the Combination condition and the Visual condition were more effective and easy to use compared with the Audio condition. The results didn’t support the hypothesis H1, but partially support the hypothesis H2 and H3.

The analysis of experimental data was consistent with the free questionnaire replies and our observations. Participants stated the factors that made the task more difficult in the Audio condition and Visual condition. Some drawbacks of the Audio condition were re-

ported in feedback. For example, some guiders explained that it was easy to become disoriented when looking at a bird-eye view map. Moreover, it was difficult to describe the direction with just left or right commands in a complex hallway that consisted of several directions. For the Visual condition, although waypoints were useful to get rid of disorientation problems for the guiders and easy to recognize for the explorers, some explorers missed the waypoints which were shown behind or next to them. Most participants described the Combination condition as the factors that made the task easier. Using the Combination condition could overcome the shortcomings above. Although, we required the guiders to send both the Audio and the Visual signals each time, which was observed that the guiders were more likely to choose the methods they thought more suitable for the current situation, which made the Combination condition have a remarkable advantage.

Additionally, there was a significant increase in simulator sickness, this was a common situation for a VR study and might be a consequence of the over 20-minute experiment duration. However, the increases observed in SSQ scores were mild and within typical expectations for a virtual reality experience involving virtual locomotion.

## 6. Experiment 2

The dyadic study was conducted in a public setting that was an uncontrolled environment with noise that could have interfered with both the guiders and the explorers. Furthermore, the results may have also been influenced by individual differences in strategies taken by the guiders and the spatial abilities of the explorers. Therefore, we conducted a lab-controlled study, which had a quiet environment with very little interference, using a trained confederate that acted as the guider. Additionally, because the volunteers registered in advance for the lab experiment, we could increase the overall duration to collect more robust data.

This experiment was conducted in our laboratory. In the study, a confederate played the role of the guider. 26 volunteers, who were recruited at our university, engaged in the study and served as explorers. Participants included 16 males and 10 females. They were aged from 19 to 37 years old ( $M = 24.74, SD = 4.78$ ). Thirteen participants reported no prior experience playing 3D video games. Two reported that they had been playing video games for 1-4 years. Three participants had been playing video games for 5-9 years, and eight reported that they have been playing for 10 or more years. Participants were required to have a normal or corrected-to-normal vision and to be able to communicate in spoken and written English. Each participant was compensated with \$10 Amazon gift cards.

### 6.1. Equipment

We used the same equipment as in Experiment 1. Additionally, the physical environment was quieter than Experiment 1. Because of this, the guider might distinguish the condition they were in based on the audio from the explorer’s headset. Therefore, the guider was required to wear a headphone with sound to prevent her from receiving audio cues during the experiment.

## 6.2. Study Design

Experiment 2 followed a within-subjects design with the same three communication conditions. To support adding more trials, we modeled two separate floors of the real building used in the previous experiment, so that the explorer would not walk through the same space too many times. Otherwise, the tasks for the explorers were the same as Experiment 1. There were a total of 12 trials (four per condition). The order of the conditions was counterbalanced across the study, and the two virtual floors were also balanced across the conditions. A total of 12 pairs of start/end waypoints were pre-defined in advance to ensure sufficient length and then uniquely and randomly assigned to each trial.

Participants were instructed to explore the virtual environment as described in Section 4. The confederate acting as the guider was also trained before the experiment. The training included several tasks: being familiar with the operation interface, going through the experiment with 3 different members in the lab, and designing a strategy to balance the two tasks. The guider was instructed to apply the same strategy consistently across all participants. We evaluated the same hypotheses as in Experiment 1.

## 6.3. Procedure

The procedure was the same as in Experiment 1, except for some additional extensions and measurements. First, the number of formal trials was increased to be 12. After finishing 6 trials, an optional 5 minutes break triggered automatically, during which the participant could remove the head-mounted display. The explorer had the opportunity to skip the break. Second, in this experiment, only the explorer needed to fill in the questionnaires. After the experiment session, the explorer needed to fill in all the questionnaires in Experiment 1 and the Slater-Usch-Steed Presence Questionnaire [UCAS00] which includes 6 questions rating from 1 to 7 with higher scores corresponding to a greater sense of presence.

The study was designed to take approximately 30 minutes in VR and 10 minutes to complete the questionnaires. However, based on the different behavior of participants, the total experiment duration varied from 30 to 60 minutes.

## 6.4. Results

We used the same methods for the statistical analysis as described in Section 5.5. We observed one participant who frequently confused "left" and "right" and confirmed that these results were statistical outliers. Therefore, we excluded data from this participant.

The analysis of navigation time shown in Figure 5 indicated that there was no significant difference among the Combination condition ( $Mdn = 519.48$ ,  $IQR = 238.8$ ), the Audio condition ( $Mdn = 494.18$ ,  $IQR = 274.1$ ), and the Visual condition ( $Mdn = 474.07$ ,  $IQR = 261.98$ ),  $\chi^2(2) = 2.24$ ,  $p = 0.326$ .

The analysis in Figure 5 showed that there existed statistically significant difference in terms of the sentries duration among three conditions,  $\chi^2(2) = 12.48$ ,  $p = 0.002$ . In addition, the Wilcoxon test revealed that the duration in the Combination condition ( $Mdn = 30.56$ ,  $IQR = 27.56$ ) was significantly shorter than

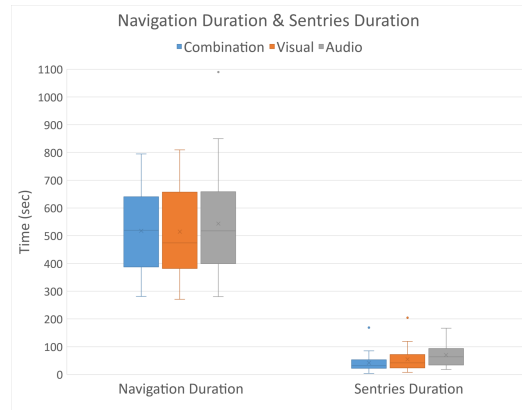
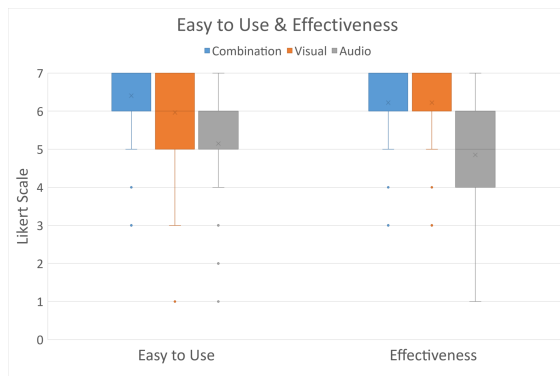


Figure 5: Results for navigation duration and sentries duration in Experiment 2. Boxplots represent the median and IQR.

the Audio condition ( $Mdn = 63.117$ ,  $IQR = 54.55$ ),  $Z = -3.80$ ,  $p < 0.001$ . The Visual condition ( $Mdn = 39.62$ ,  $IQR = 46.99$ ) also significantly shorter than the Audio condition ( $Mdn = 63.117$ ,  $IQR = 54.55$ ),  $Z = 2.77$ ,  $p = 0.006$ . However, there was no significant difference between the Visual ( $Mdn = 39.62$ ,  $IQR = 46.99$ ) and the Combination conditions ( $Mdn = 30.56$ ,  $IQR = 27.56$ ),  $Z = -1.34$ ,  $p = 0.181$ .

The questionnaires displayed similar results as in the Experiment 1. The box plot showed in Figure 6. The Friedman test indicated that there existed significant difference among the three conditions in the aspect of easy to use,  $\chi^2(2) = 16.46$ ,  $p < 0.001$ . In addition, the Combination condition ( $Mdn = 7$ ,  $IQR = 1$ ) was significantly easier to use than the Audio condition ( $Mdn = 6$ ,  $IQR = 1$ ),  $Z = -4.56$ ,  $p < 0.001$ . The Visual condition ( $Mdn = 6$ ,  $IQR = 2$ ) was also significant easier to use compared with the Audio condition ( $Mdn = 6$ ,  $IQR = 1$ ),  $Z = -2.82$ ,  $p = 0.005$ . But the Combination condition ( $Mdn = 7$ ,  $IQR = 1$ ) had no significant difference compared with the Visual condition ( $Mdn = 6$ ,  $IQR = 2$ ),  $Z = -2.07$ ,  $p = 0.038$ . In the analysis of effectiveness, significant difference also existed among the three conditions,  $\chi^2(2) = 13.70$ ,  $p = 0.001$ . Participants considered that the Combination condition ( $Mdn = 7$ ,  $IQR = 1$ ) was significantly more effective than Audio condition ( $Mdn = 5$ ,  $IQR = 2$ ),  $Z = -4.54$ ,  $p < 0.001$ . The Visual condition ( $Mdn = 7$ ,  $IQR = 1$ ) was also significantly more effective than Audio condition ( $Mdn = 5$ ,  $IQR = 2$ ),  $Z = -4.20$ ,  $p < 0.001$ . But there was no significant difference between the Combination condition ( $Mdn = 7$ ,  $IQR = 1$ ) and the Visual condition ( $Mdn = 7$ ,  $IQR = 1$ ) in terms of the effectiveness,  $Z = -4.20$ ,  $p = 0.909$ .

Similar to the previous experiment, the simulator sickness scores indicated that participants experienced significantly higher levels of simulator sickness after the experiment ( $Mdn = 7.48$ ,  $IQR = 26.18$ ) than before the experiment ( $Mdn = 0$ ,  $IQR = 3.74$ ),  $\chi^2(1) = 23.50$ ,  $p < 0.001$ . A presence score was yielded by averaging the six questions of the SUS questionnaire ( $M = 4.44$ ,  $SD = 1.61$ ).



**Figure 6:** Responses to the feedback-questionnaire in Experiment 1. Boxplots represent the median and IQR (in a 7-point Likert scale)

### 6.5. Discussion

The results of navigation duration were consistent with the findings in Experiment 1. There was also no significant difference among the three conditions in terms of the navigation duration. It can be explained because the guider valued the sentries duration more against the navigation time. As long as the sentries walked close to the explorer, the guider required the explorer to hide inside a room until the sentries walked far away. If the sentries showed on the shortest route, the guider preferred to guide the explorer to change to another route which took longer time to reach the target location.

In the aspect of sentries duration, both the Combination condition and the Visual showed significantly better performance than the Audio condition. In the feedback-questionnaire, the Combination condition was regarded as the most easy-to-use method among the three conditions. The Combination condition and the Visual condition also outperformed in effectiveness compared with the Audio condition. The results didn't support the hypothesis H1, but partially support the hypothesis H2 and H3.

Combined with the free questions in the feedback-questionnaire, the advantages for the Combination conditions were manifest. Participants stated the factors that made the task more difficult in the Audio condition and the Visual condition, which was also consistent with the responses in Experiment 1. Participants reported that the audio commands were hardest to follow because it would not always queue at the exact right time and were not clear enough when they stood in front of a fork road. Additionally, the explorers' orientation was defined based on their heads' orientation instead of their whole bodies'. This limitation caused that the audio commands might fall into confusion if the explorer's head and body were in different directions. The drawbacks of the Visual condition were also concentrated in looking for the waypoints. Some explorers complained that they had to keep turning around to look for the next waypoint in case they would miss any hints. This process was not convenient and largely slow down their reactions.

For the factors that made the task easier, the Combination condition was frequently mentioned. Participants agreed that the Com-

bination condition was easiest to use and most effective since it overcame defects in the single method. Some participants also mentioned that the visual signal allowed them to move at their own pace and also helped them feel more immersive because they had to look for the objective rather than just follow the audio commands.

The SUS presence average score was moderately high overall. It indicated that participants experienced high-quality immersion. Simulator sickness scores show significant increases which was predictable due to the long duration of the experiment. Similar to Experiment 1, the increases in SSQ scores were mild and within normal expectations for virtual reality experiences of this duration.

## 7. Conclusion And Future Work

In this paper, we presented three communication methods (audio-only, visual-only, and the combination of audio and visual) to support two-user collaborative guiding navigation tasks in a dynamic virtual reality environment. Two experiments were conducted: a dyadic study at a large public event and a controlled lab study using a confederate. Statistical analyses were conducted using task performance measures (navigation duration and sentries duration performance) and questionnaire responses. The results of both experiments showed that the combined cues were rated more easy to use and more effectively facilitated in the avoidance of sentries in complex environments compared with only using audio. Although this conclusion seems to be intuitive, the value of the study is to provide scientific measurements to offer guidance for the future. However, combined cues did not show significantly better performance in both experiments compared with visuals alone.

Although the Combination condition exhibited marginally better performance compared with the Visual condition, the difference was not significant in both experiments. We are interested in further exploration of sensory dominance effects in the context of multimodal remote guidance interfaces. The Colavita visual dominance effect, a phenomenon that occurs when participants fail to respond to the auditory component of a bimodal target significantly more often than the visual component, has been well documented in prior studies. If the Colavita visual dominance effect can also be observed in this context, then this would lead to valuable insights into the effective design of future remote guidance interfaces. Additionally, our virtual environment was designed with zero interference, which may not apply in many real world applications. In future experiments, it may be valuable to design scenarios that mimic negative conditions that may be encountered during fielded use, such as background noise or other types of audio interference. Additional factors could also be introduced to provide visual interference, such as dynamic obstacles that block previously accessible paths. These factors may weaken users' abilities to rely upon a single navigation cue, which would provide an advantage for multimodal interfaces.

In summary, we observed that the specific implementation of audio instructions and the design of the visual overlays influences task performance, especially in complex and dynamic virtual environments. Further experiments conducted using dyadic or confederate-based designs will also be considered to evaluate other communication modalities or 3D user interfaces to support collaborative guidance in immersive environments.



## Acknowledgments

The authors would like to thank David M. Krum and Mark Dennison for participating in discussions and providing feedback throughout this project. This work was sponsored by the University of Southern California Institute for Creative Technologies and the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## References

- [Ala94] ALAVI M.: Computer-mediated collaborative learning: An empirical evaluation. *MIS quarterly* (1994), 159–174. 2
- [Ano20] ANONYMOUS: Exploring communication modalities to support collaborative guidance in virtual reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2020), IEEE, pp. 79–86. 2
- [BGFTJ\*18] BERGER C. C., GONZALEZ-FRANCO M., TAJADURA-JIMÉNEZ A., FLORENCIO D., ZHANG Z.: Generic hrtfs may be good enough in virtual reality: improving source localization through cross-modal plasticity. *Frontiers in neuroscience* 12 (2018), 21. 2
- [BOG\*02] BOS N., OLSON J., GERGLE D., OLSON G., WRIGHT Z.: Effects of four computer-mediated communications channels on trust development. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2002), ACM, pp. 135–140. 2
- [BPF\*19] BOS A. S., PIZZATO M., FERREIRA V. A., SCHEIN M., ZARO M. A., TAROUCO L.: The impact of effective communication between users in 3d collaborative virtual environments: the conversational agent use case. *International Journal of Advanced Engineering Research and Science* 6, 8 (2019). 2
- [BRS\*12] BACIM F., RAGAN E. D., STINSON C., SCERBO S., BOWMAN D. A.: Collaborative navigation in virtual search and rescue. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)* (2012), IEEE, pp. 187–188. 3
- [Bub01] BUBAŠ G.: Computer mediated communication theories and phenomena: Factors that influence collaboration over the internet. In *3rd CARNet users conference, Zagreb, Hungary* (2001), Citeseer. 2
- [CMF\*20] CLEMENSON G. D., MASELLI A., FIANNACA A., MILLER A., GONZALEZ-FRANCO M.: Rethinking gps navigation: Creating cognitive maps through auditory clues. *BioRxiv* (2020). 2
- [cor] Simple corridors pack. Accessed on 08-01-2019. URL: <https://assetstore.unity.com/publishers/21665>. 3
- [CRdS\*12] CABRAL M., ROQUE G., DOS SANTOS D., PAULUCCI L., ZUFFO M.: Point and go: Exploring 3d virtual environments. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)* (2012), IEEE, pp. 183–184. 1, 3
- [CS98] CHURCHILL E. F., SNOWDON D.: Collaborative virtual environments: an introductory review of issues and systems. *Virtual Reality* 3, 1 (1998), 3–15. 1, 2
- [CS99] CHEN J. L., STANNEY K. M.: A theoretical model of wayfinding in virtual environments: Proposed strategies for navigational aiding. *Presence* 8, 6 (1999), 671–685. 2
- [CSM12] CHURCHILL E. F., SNOWDON D. N., MUNRO A. J.: *Collaborative virtual environments: digital places and spaces for interaction*. Springer Science & Business Media, 2012. 2
- [DA\*08] DODIYA J., ALEXANDROV V. N., ET AL.: Navigation assistance for wayfinding in the virtual environments: Taxonomy and a survey. *ICAT 2008* (2008), 1345–1278. 2
- [DNF\*14] DUVAL T., NGUYEN T. T. H., FLEURY C., CHAUFFAUT A., DUMONT G., GOURANTON V.: Improving awareness for 3d virtual collaboration by embedding the features of users’ physical environments and by augmenting interaction tools with cognitive feedback cues. *Journal on Multimodal User Interfaces* 8, 2 (2014), 187–197. 3
- [dra18] Small red dragon, April 2018. Accessed on 11-08-2019. URL: <https://assetstore.unity.com/packages/3d/characters/small-red-dragon-52959>. 3
- [DS96] DARKEN R. P., SIBERT J. L.: Wayfinding strategies and behaviors in large virtual worlds. In *CHI* (1996), vol. 96, pp. 142–149. 2
- [DSD\*99] DUMAS C., SAUGIS G., DEGRANDE S., PLÉNACOSTE P., CHAILLOU C., VIAUD M.-L.: Spin: a 3d interface for cooperative work. *Virtual Reality* 4, 1 (1999), 15–25. 2
- [ENK97] ELVINS T. T., NADEAU D. R., KIRSH D.: Worldlets—3d thumbnails for wayfinding in virtual environments. In *Proceedings of the 10th annual ACM symposium on User interface software and technology* (1997), ACM, pp. 21–30. 2
- [ETT07] ELMQVIST N., TUDOREANU M. E., TSIGAS P.: Tour generation for exploration of 3d virtual environments. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology* (2007), ACM, pp. 207–210. 2
- [FDPG02] FORT A., DELPUECH C., PERNIER J., GIARD M.-H.: Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Cerebral Cortex* 12, 10 (2002), 1031–1039. 2
- [GFMF\*17] GONZALEZ-FRANCO M., MASELLI A., FLORENCIO D., SMOLYANSKIY N., ZHANG Z.: Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports* 7, 1 (2017), 1–11. 2
- [GS18] GLASER N. J., SCHMIDT M.: Usage considerations of 3d collaborative virtual learning environments to promote development and transfer of knowledge and skills for individuals with autism. *Technology, Knowledge and Learning* (2018), 1–8. 2
- [HFH\*98] HINDMARSH J., FRASER M., HEATH C., BENFORD S., GREENHALGH C.: Fragmented interaction: Establishing mutual orientation in virtual environments. In *CSCW* (1998), vol. 98, Citeseer, pp. 217–226. 2
- [HR09] HECHT D., REINER M.: Sensory dominance in combinations of audio, visual and haptic stimuli. *Experimental brain research* 193, 2 (2009), 307–314. 2
- [KLBL93] KENNEDY R. S., LANE N. E., BERBAUM K. S., LILIENTHAL M. G.: Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220. 5
- [KLS09] KOPPEN C., LEVITAN C. A., SPENCE C.: A signal detection study of the colavita visual dominance effect. *Experimental brain research* 196, 3 (2009), 353–360. 2
- [KUA\*19] KHALID S., ULLAH S., ALI N., ALAM A., RABBI I., REHMAN I. U., AZHAR M.: Navigation aids in collaborative virtual environments: Comparison of 3dml, audio, textual, arrows-casting. *IEEE Access* 7 (2019), 152979–152989. 2
- [LJKM\*17] LAVIOLA JR J. J., KRUIJFF E., MCMAHAN R. P., BOWMAN D., POUPYREV I. P.: *3D User Interfaces: Theory and Practice*. Addison-Wesley Professional, 2017. 1
- [NDF13] NGUYEN T. T. H., DUVAL T., FLEURY C.: Guiding techniques for collaborative exploration in multi-scale shared virtual environments. 1, 3
- [NDWG\*12] NOTELAERS S., DE WEYER T., GOORTS P., MAESEN S., VANACKEN L., CONINX K., BEKAERT P.: Heatmeup: A 3d serious game to explore collaborative wayfinding. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)* (2012), IEEE, pp. 177–178. 3
- [RS04] REITMAYR G., SCHMALSTIEG D.: *Collaborative augmented reality for outdoor navigation and information browsing*. na, 2004. 2
- [SJS03] ST JULIEN T. U., SHAW C. D.: Firefighter command training virtual environment. In *Proceedings of the 2003 conference on Diversity in computing* (2003), ACM, pp. 30–33. 1, 3

- [SPC12] SPENCE C., PARISE C., CHEN Y.-C.: The colavita visual dominance effect. In *The neural bases of multisensory processes*. CRC Press/Taylor & Francis, 2012. 2
- [SPT06] STAFFORD A., PIEKARSKI W., THOMAS B. H.: Implementation of god-like interaction techniques for supporting collaboration between outdoor ar and indoor tabletop users. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality* (2006), IEEE, pp. 165–172. 3
- [TTS] From text to speech. Accessed on 11-08-2019. URL: <http://www.fromtexttospeech.com>. 3
- [UCAS00] USOH M., CATENA E., ARMAN S., SLATER M.: Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments* 9, 5 (2000), 497–503. 7
- [Vin99] VINSON N. G.: Design guidelines for landmarks to support navigation in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1999), ACM, pp. 278–285. 2
- [Wal96] WALTHER J. B.: Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43. 2
- [Wal11] WALTHER J. B.: Theories of computer-mediated communication and interpersonal relations. *The handbook of interpersonal communication* 4 (2011), 443–479. 2
- [WBL\*12] WANG J., BUDHIRAJA R., LEACH O., CLIFFORD R., MATSUDA D.: Escape from meadwyn 4: A cross-platform environment for collaborative navigation tasks. In *2012 IEEE Symposium on 3D User Interfaces (3DUI)* (2012), IEEE, pp. 179–180. 3
- [WLM\*19] WALTER H., LI R., MUNAFO J., CURRY C., PETERSON N., STOFFREGEN T.: *A brief explanation of the Simulator Sickness Questionnaire (SSQ)*. University of Minnesota, March 2019. URL: <https://doi.org/10.13020/XAMG-CS69>. 6
- [WM08] WRIGHT T., MADEY G.: A survey of collaborative virtual environment technologies. *University of Notre Dame-USA, Tech. Rep* (2008), 1–16. 2
- [YO02] YANG H., OLSON G. M.: Exploring collaborative navigation: the effect of perspectives on group performance. In *Proceedings of the 4th international conference on Collaborative virtual environments* (2002), ACM, pp. 135–142. 1, 3