frontiers in Virtual Reality

ORIGINAL RESEARCH
published: 01 July 2021
doi: 10.3389/frvir.2021.681585

The Effects of Incorrect Occlusion Cues on the Understanding of Barehanded Referencing in Collaborative Augmented Reality

*Yuan Li**, Donghan Hu, Boyuan Wang, Doug A. Bowman and Sang Won Lee*

Center for Human-Computer Interaction, Department of Computer Science, College of Engineering, Virginia Tech, Blacksburg, VA, United States

In many collaborative tasks, the need for joint attention arises when one of the users wants to guide others to a specific location or target in space. If the collaborators are co-located and the target position is in close range, it is almost instinctual for users to refer to the target location by pointing with their bare hands. While such pointing gestures can be efficient and effective in real life, performance will be impacted if the target is in augmented reality (AR), where depth cues like occlusion may be missing if the pointer's hand is not tracked and modeled in 3D. In this paper, we present a study utilizing head-worn AR displays to examine the effects of incorrect occlusion cues on spatial target identification in a collaborative barehanded referencing task. We found that participants' performance in AR was reduced compared to a real-world condition, but also that they developed new strategies to cope with the limitations of AR. Our work also identified mixed results of the effect of spatial relationships between users.

Keywords: augmented reality, collaboration, occlusion, hand referencing, spatial referencing

OPEN ACCESS

Edited by:
Thammathip Piumsomboon,
University of Canterbury, New Zealand

Reviewed by:
Markus Tatzgern,
University of Applied Sciences
Salzburg, Austria
Jonathan Ventura,
California Polytechnic State University,
United States

****Correspondence:***
Yuan Li
yli92@vt.edu

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

Received: *16 March 2021*
Accepted: *25 May 2021*
Published: *01 July 2021*

Citation:
Li Y, Hu D, Wang B, Bowman DA and Lee SW (2021) The Effects of Incorrect Occlusion Cues on the Understanding of Barehanded Referencing in Collaborative Augmented Reality.
Front. Virtual Real. 2:681585.
doi: 10.3389/frvir.2021.681585

I INTRODUCTION

A unique advantage of face-to-face communication is *visibility*, defined as "being able to see each other" (Clark et al., 1991). The advantages of co-located, in-person communication include nonverbal communication in such forms as gestures, gaze awareness, and eye contact (Zahn, 1991; Olson and Olson, 2000). In particular, joint attention on an object of mutual interest is a common requirement in many collaborative tasks (Clark and Wilkes-Gibbs, 1986; Dare, 1995). Such actions typically require the communication and confirmation of an object's location (Yuan et al., 2019). Dix proposed a CSCW framework that included the idea of deixis in cooperation and argued for the importance of support for deixis relative to a shared artifact in groupware (Dix, 1994). Similarly, in face-to-face communication, multiple studies have demonstrated that pointing gestures with the hand are used to convey spatial information (Cohen and Harrison, 1973; McNeill, 1992; Krauss et al., 2000; Allen, 2003; Kita and Özyürek, 2003). For example, when an expert teaches workers to inspect a machine in a factory, the expert may perform deictic gestures to refer to different parts of the machine and guide the workers' attention. In this paper, we use the term *barehanded referencing* to denote such actions.

When barehanded referencing is used to denote nearby physical objects, the intent of the pointing gesture is usually very clear; however, this may not be true when collaborative augmented reality (AR) systems are used. Specifically, co-located collaborators using AR head-worn displays (HWDs)

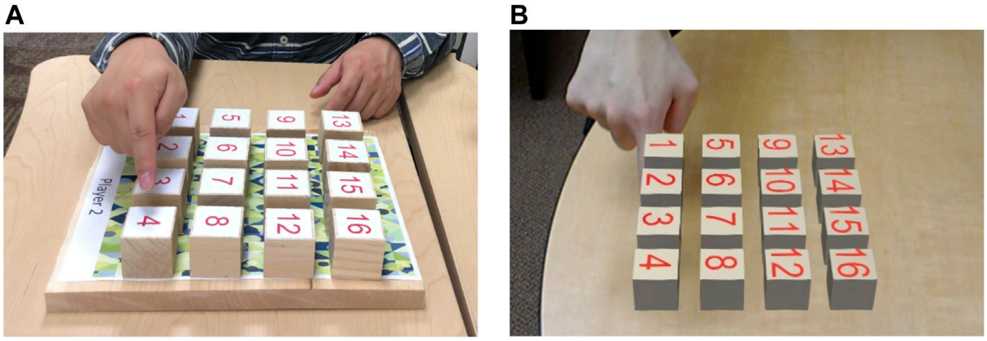

FIGURE 1 | An observer's interpretation of a pointer's referencing gestures depends on occlusion cues. **(A)**: Physical cubes are correctly occluded by the pointer's hand (note that the image beneath the cubes was not present during the experiment). **(B)**: Referential ambiguity due to incorrect visual occlusion in AR without a hand model (note that the cubes would appear to be slightly transparent when viewed through the HoloLens).

can view, create, and interact with virtual content to support tasks ranging from brainstorming and medical training to CAD design and factory inspection. However, barehanded referencing in collaborative AR may be problematic when the AR system fails to obtain a reliable geometric model of key objects (e.g., users' hands) in the physical environment [as in so-called *model-free AR* (Comport et al., 2006)]. In these conditions, the system cannot render correct occlusion cues, so that physical objects (such as the user's hand) do not properly occlude virtual objects. Occlusion, as the most dominant depth cue, helps us to judge depth relationships among objects (Cutting et al., 1995). In the real world, objects that are closer to us will be seen to occlude (fully or partially hide) other objects that are farther away. In model-free AR, on the other hand, the virtual content appears on top of physical objects, resulting in a false occlusion cue where a virtual object appears to be closer than a physical object (as seen in **Figure 1**-Right). Although model-free AR is not the most common use case, the opposite scenario, in which the system always obtains a perfectly accurate model, is equivalently unlikely with today's technology. The current state of the art is somewhere between these two extremes: modern AR HWDs, such as the Microsoft HoloLens 2, do have some capability to track the user's hands in real time and to use the resulting imperfect hand model to occlude virtual content.

Unfortunately, no current AR system provides ideal occlusion cues, and in practice, there are still many challenges. First, capturing the most agile part of the human body—human hands—in real time is challenging. Technical difficulties make constant and reliable tracking not immediately available: limited sensor range, nonideal lighting, and occlusion between a user's two hands, to name a few (Mueller et al., 2017). Moreover, the capture process brings significant computational overhead to the application[1], and the computational cost will increase in a collaborative setup due to the requirements of synchronizing multiple users' hand models captured from their respective egocentric views across multiple devices and rendering correctly occluded objects from every participant's point of view. Thus, if hand tracking is used in multiuser applications, it is likely to be temporally or spatially inaccurate to some degree. With an inaccurate environment model, the AR system can indeed present occlusion, but at the wrong locations, which can be worse than presenting no occlusion at all. These challenges may explain why, despite the theoretical possibility of real-time hand tracking, many modern AR applications still do not include correct occlusion of virtual objects by users' hands.

Thus, we chose to study the model-free scenario as the baseline condition in order to understand the influence of incorrect occlusion cues on collaborative spatial referencing. Such incorrect occlusion cues will not only confuse the user and jeopardize depth perception in AR, but it may also impede *other users*' spatial understanding in a collaborative AR system. In a collaborative setup, incorrect occlusion goes beyond the mere problem of presenting unrealistic visuals. It is a matter of communication: the pointer knows what they want to reference (ground truth), but the observer may mislocate the target. In high-stakes collaborative tasks, such as surgical training or urban planning, this type of referencing error can not only be costly, but also lead to catastrophic consequences. Prior research has investigated incorrect occlusion in AR, going back to Breen et al. (Breen et al., 1996). While there exist many previous works that address the problem of incorrect occlusion, most have studied the problem from a single user's egocentric perspective (Kiyokawa et al., 2003; Hayashi et al., 2005; Mendez and Schmalstieg, 2009; Boboc et al., 2019), as shown in **Figure 2**. In this work, we focus on the observer's perspective (**Figure 2**, ①, ②. To understand how to design effective collaborative AR systems, we first need to develop a solid understanding of the challenge posed by incorrect visual occlusion of the user's hand gestures.

In this paper, we explore to what extent and how spatial referencing in AR is influenced by model-free AR (which has incorrect occlusion) by comparing an ideal condition, namely referencing of physical objects in the real world with perfect occlusion, with a model-free AR condition (referencing of virtual

[1] As the HoloLens 2 developer documentation warns: "(hand) joint objects are transformed on every frame and can have significant performance cost!" https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/Input/HandTracking.html

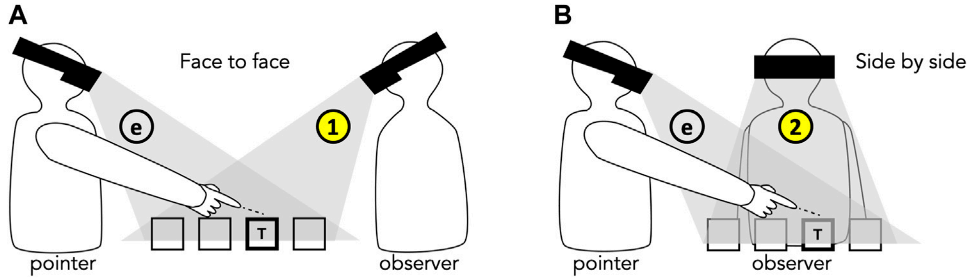

FIGURE 2 | Illustration of the experimental setup. We measure the observer's understanding (① or ②) of the spatial references produced by the pointer. ⓔ denotes the pointer's egocentric perspective. **(A)**: A face-to-face layout. **(B)**: A side-by-side layout. This setup simulates a group of people surrounding shared small-scale AR content (e.g., an AR-based board game, medical training).

objects in AR without proper occlusion between the hand and the objects). We invited participants to perform a near-field object referencing task both in the physical world (with correct occlusion) and in AR (where users' pointing gestures are incorrectly occluded; that is, the referenced virtual object appears as an overlay on the pointer's hands, as seen on the right in **Figure 1**). We also considered spatial characteristics, including collaborators' seating positions (face-to-face and side-by-side) and target positions (where a target object is located relative to other objects). We found that incorrect occlusion in AR had a significantly negative effect on performance, and that this effect dominated any impact of spatial characteristics. However, when occlusion was unreliable, the participants sought alternative communication cues to complete the tasks.

Our research contributions are as follows:

- Evaluation of the extent to which incorrect occlusion impacts the performance of barehanded referencing in a shared, model-free AR scene, in comparison with the correct occlusion seen during referencing of physical objects
- Empirical understanding of how users perceive and address the challenge of incorrect occlusion for barehanded referencing

Our study benefits the design of future collaborative AR systems by providing an understanding of the cost of incorrect occlusion during spatial referencing.

II RELATED WORK

A. Significance of Gestural Referencing in Collaboration

When communicating about work objects in a visual environment, "what is shown and how it is shown is crucial" (Whittaker, 2003). Pointing and deictic reference (e.g., pointing hand gestures) are together an essential component that bridges understanding between participants in groupware design (Dix, 1994). There has been much prior work studying barehanded referencing in human collaboration. For instance, Kirk and Fraser raised the benefit of sharing hand to an increased utility in object-focused interactions (Kirk and Fraser, 2006). One aspect that many works have focused on is to understand the role of spontaneous hand gestures with spatial information in communication (Martha, 2005). Users use hand gestures to express spatial information during communication. For example, Alibali et al. (2001) asked participants to narrate different units of a cartoon to a naïve addressee. The participants were nearly twice as likely to use gestures with units containing spatial information than with units that did not. Such gestures also make a difference for the observers, contributing to the effective communication of spatial information. Beattie and Shovelton. (1999) conducted a study where they presented participants with audio-only clips and audio + gesture video clips of the same narratives. They found that spatial information was communicated significantly more effectively in the audio + gesture video condition than in the audio-only condition. Using the expressivity of hand gestures in multimodal interaction; that is, in synchronization with speech, dates back to the 1980s. Especially in large-scale display interfaces, users need to refer to digital objects through pointing (Richard, 1980; Oviatt, 1996; Oviatt et al., 1997; Beattie and Shovelton, 1999). In summary, there is abundant evidence to support the idea that allowing hand gestures in multimodal interaction is effective in facilitating collaboration where users can exchange spatial information.

B. Referencing Objects in AR-Based Collaboration

AR researchers have explored various systems to enhance face-to-face communication. Szalavári et al. (1998) demonstrated *Studierstube*, one of the earliest AR interfaces for face-to-face collaboration. The project used see-through HWDs to allow users to collaboratively view 3D virtual models registered in the real world. Since then, researchers have sought to understand and improve different aspects of AR-assisted collaboration. Chastine et al. (2008) studied the virtual pointer as a visualization for reference cues in collaborative AR. They concluded that poor reference cues would generate referential ambiguity, incurring such additional communication costs as time and computational resources, while carefully designed interaction techniques would help collaboration significantly. To visualize the user's hand posture, they used tracked controllers to localize the user's hands; these controllers could inhibit users' ability to perform other tasks. More recently, Oda and Feiner (2012) explored 3D referencing techniques in shared AR. Using a depth camera to capture the user's pointing direction, they compared their hand gesture referencing technique with other controller-based techniques and found that their gestural referencing technique was significantly more accurate when the participants had sufficiently different views of the shared scene. They suggested an advantage in the use of hands for referencing. However, their approach relied heavily on an external depth camera to capture the user's hand and the physical objects.

In studying effective interfaces for collaborative AR, researchers have identified the importance of occlusion cues. Kiyokawa et al. (2003) demonstrated an early prototype that could enable correct occlusion cues to support co-located AR collaboration. They concluded that their occlusion feature enhanced the sense of presence of virtual objects for over 75% of their participants. Lee et al. (2004) adopted the idea and studied occlusion-based interaction techniques for tangible AR. Using image markers, they developed tangible referencing interfaces where the user's hands could correctly occlude the virtual content when the hands blocked a marker in the camera's view. Their work suggested the simplicity and naturalness of occlusion-based 2D interaction for AR. However, they noted the limitation of view dependency, limiting the application of their technique to an egocentric view. This renders it unsuitable for collaborative settings. Chastine (Jeffrey, 2007) emphasized the importance of proper occlusion for virtual objects in collaborative AR, as the absence of it would cause not only an unnatural experience for the users, but also reduced referencing reliability. In later work, Chastine and Zhu (2008) summarised the additional costs of time, efficiency, and hardware needed to avoid the probability of referential ambiguity, partially caused by incorrect occlusion.

Even without proper occlusion, there are interaction techniques for referencing virtual and physical targets in AR-based, co-located collaboration (Chastine et al., 2008; Oda and Feiner, 2012). However, these systems typically require users to hold tracked controllers that limit the expressiveness afforded by barehanded interaction. For example, holding controllers could prevent users from performing certain tasks that require their hands, such as making particular gestures or typing on a keyboard. Thus, it is beneficial to explore solutions that support object referencing with bare hands in collaborative AR. Despite prior research, there is still no complete and reliable solution to the problem in the case of dynamic objects like hands, nor is there an understanding of the extent to which incorrect occlusion cues decrease the effectiveness of barehanded referencing in AR. In this paper, we take an initial step towards understanding the impact of incorrect occlusion in barehanded referencing performance through a controlled experiment.

C. Barehanded Referencing in AR

In theory, incorrect occlusion can be resolved with the availability of complete physical models of the real-world objects in an AR scene. Despite previous research in making real-world objects (including the human body) occlude virtual content correctly (Kiyokawa et al., 2003; Mendez and Schmalstieg, 2009; Boboc et al., 2019), it is still challenging and computationally costly for these AR systems to obtain perfect models of dynamic objects in the physical world, such as human hands. In a recent endeavor, Yoon et al. (2020) evaluated user perception of remote virtual hand models of varying fidelity in hand-based 3D remote collaboration using both AR and VR headsets. Although the authors reported little regarding the technical details of their realistic hand model (number of polygons, refresh rate), their implementation used desktop computers with powerful hardware instead of mobile AR headsets like HoloLens 2 or Magic Leap. Notably, even though there are AR devices that support basic hand tracking, they do not always provide highly accurate hand models that match the physical hand in real time. In addition, even with an accurate model from a user's egocentric view, the model may be imperfect or incomplete from another user's perspective when shared in a multi-user application.

Huang et al. (2018) tested sharing 3D hand models in a remote collaborative Mixed Reality system for a repairing task and concluded that using 3D virtual hands improved the level of co-presence, and participants had better spatial relations in the task space. Accuracy of the hand model aside, they highlighted the importance of calibrating the gesture in collaborators' local spaces to provide a practical application.

While interaction techniques exist for referencing virtual and physical targets in AR-based co-located collaboration (Chastine et al., 2008; Oda and Feiner, 2012), these systems typically require users to hold tracked controllers. These controllers limit the expressiveness afforded by barehanded interaction. For example, holding controllers could prevent users from performing certain tasks that require their hands, such as making particular gestures or typing on a keyboard. Thus, it is beneficial to explore solutions that support object referencing with bare hands in collaborative AR. Kim et al. (2019) explored the use of hand pointers by extending the user's index finger direction in a remote guidance assembly task. They found that the hand pointer along was not sufficient for the participants. However, their work was only from one collaborator's egocentric view and did not consider its use from the other user's perspective. Despite prior research, there is still no complete and reliable solution to the problem in the case of dynamic objects like hands, nor is there an understanding of the extent to which incorrect occlusion cues decrease the effectiveness of barehanded referencing in AR. In this paper, we take a first step towards understanding the impact of incorrect occlusion in barehanded referencing performance through a controlled experiment.

III EXPERIMENT

In order to evaluate the effect of incorrect occlusion cues on barehanded referencing during collaboration in AR, we conducted a controlled experiment wherein one participant played the role of a *pointer* who needed to refer to a given target with their hands, and the other played the role of an

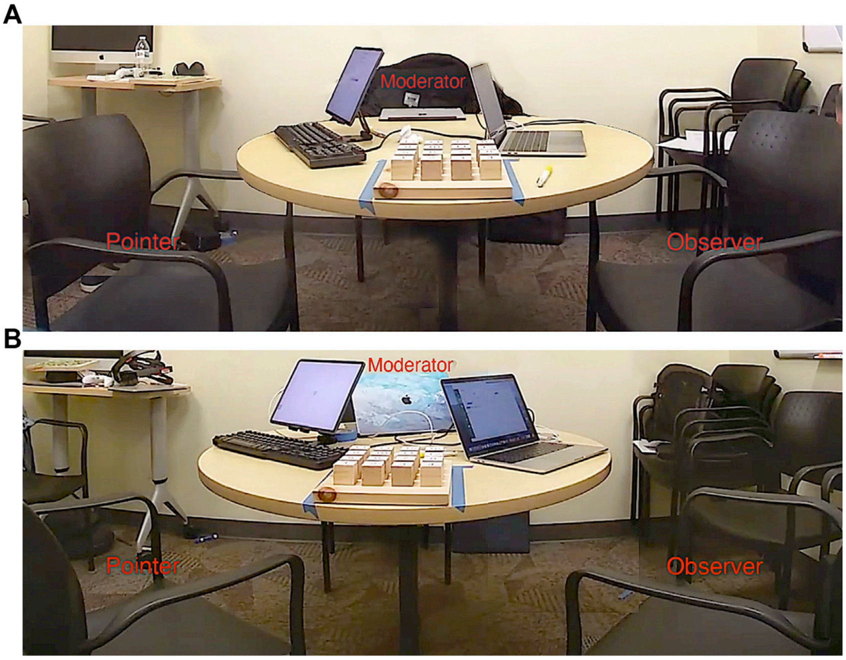

FIGURE 3 | Experiment setup: The chairs were aligned with tape on the ground to form 180° and 90° configurations for the physical, face-to-face condition **(A)** and the physical, side-by-side condition **(B)**, respectively. The target cubes were placed at fixed locations on the table. The laptop and monitor were adjusted by the participants for ease of use. The image was taken with a GoPro fisheye camera.

observer who needed to identify the correct target. The target could be either a physical object or a virtual object displayed in the AR environment. **Figure 1** shows the observer's view in the physical (left) and AR (right) conditions. Additionally, we varied the spatial relationship between the participants and the targets, because we believed that differences between the pointer's and observer's perspectives might lead to different visual perceptions for the observer.

We designed the experiment to investigate two research questions and associated sub-questions:

- RQ1: How do the incorrect occlusion cues in model-free AR affect performance in understanding another user's barehanded spatial references?
  - RQ1-1: To what extent is the understanding of spatial references negatively influenced by model-free AR?
  - RQ1-2: How do spatial configurations of collaborators and target locations affect understanding in model-free AR?
- RQ2: What strategies do collaborators develop to overcome incorrect occlusion cues?
  - RQ2-1: What strategies do pointers adopt to better communicate spatial references in model-free AR?
  - RQ2-2: What strategies do observers adopt to better understand spatial references in model-free AR?

A. Experiment Task

There were two roles involved in the study: the pointer and the observer. The referencing targets were sixteen cubes, arranged in a four-by-four grid shape. The target cubes were either solid wood cubes or virtual cubes displayed in AR HWDs (as shown in **Figure 1**). Regardless of the cube type, the cubes' upper surfaces had red labels from one to sixteen as identifiers. We also made sure that the virtual cubes were correctly aligned for both the pointer and the observer by using Vuforia's image recognition algorithm[2] along with a printed picture to place the virtual cubes in constant positions across sessions. Apart from changing the cube type, we also varied the collaborators' spatial relationships throughout the experiment. Participants sat side-by-side (i.e., at 90° angles to each other with respect to the cubes) or face-to-face (i.e., at 180° angles to each other). The primary task in the experiment was for the pointer, who was given a number from one to sixteen, to point to the target cube with their dominant hand; the observer was asked to identify the target cube's label by interpreting the referential gesture. The participants were asked to perform this task as accurately as possible, with speed of task performance being a secondary goal. The experiment setup is shown in **Figure 3**.

In summary, we investigated the following factors' impacts on referencing objects with the hands.

- Cube type: physical vs. AR
- Seating position: face-to-face vs. side-by-side
- Target position: the position of the target cube in the 4 × 4 grid layout

[2] https://library.vuforia.com/features/images/image-targets.html

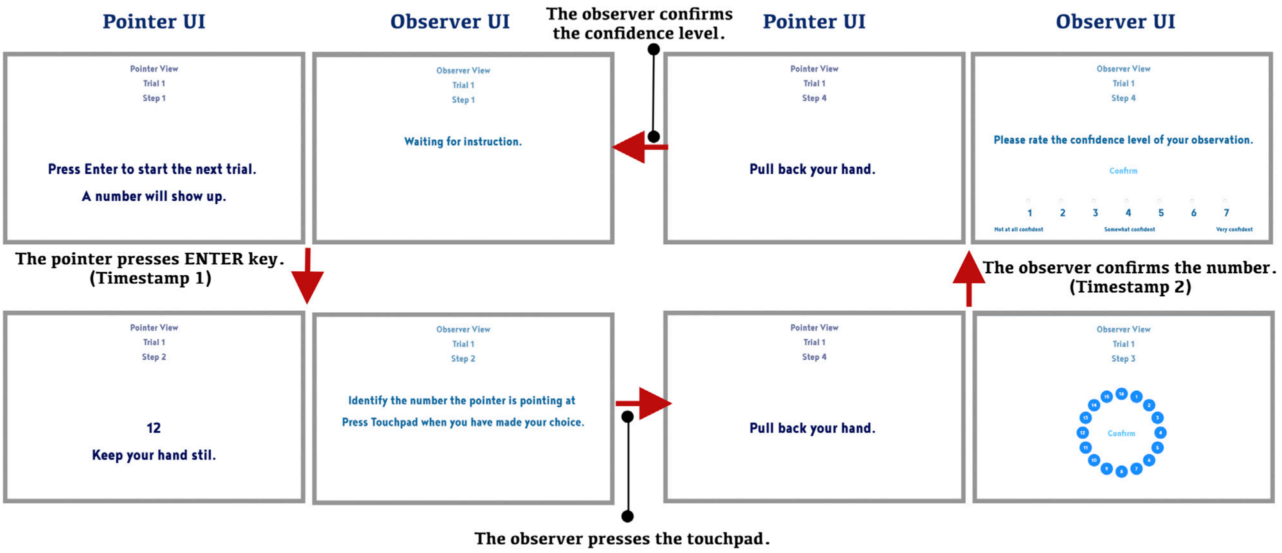

FIGURE 4 | Flowchart of experiment UI for one experimental trial, starting from the upper-left pair of screens.

In each study session, one participant was assigned to the role of the pointer, while the other participant was the observer. In order to minimize variance, we ensured that all pointers were right-handed. We did not switch participants' roles during the experiment to avoid bias introduced by learning.

The goal of the study was to understand to what extent and how spatial referencing in AR is influenced by model-free AR (which has incorrect occlusion) by comparing performance and user behavior in an AR environment to the same aspects of spatial referencing in a physical environment. We decided to use pointing at physical cubes as the base condition because this represents the gold standard for spatial referencing in real-life situations, and the scenario that AR systems attempt to simulate. While changing from the physical condition to the model-free AR condition introduces differences other than incorrect occlusion from the observer's perspective (e.g., texture and shadows), based on what we observed in the preliminary study and what has been reported in the literature (Cutting et al., 1995; Kiyokawa et al., 2003; Lee et al., 2004; Yuan et al., 2019), incorrect occlusion remains the primary difference between the two conditions. Therefore, comparing model-free AR with a physical condition can help us gain insight into the effect of incorrect occlusion on collaborative spatial referencing.

B. Experiment Design

The experiment followed a 2 (cube type: physical or AR) × 2 (seating position: face-to-face or side-by-side) within-subjects design for a total of four conditions, as depicted in **Figure 2**. The two seating positions tested were selected to model various settings where people sit around a table or share the same perspective with respect to physical artifacts. Similar settings have been studied in prior CSCW works concerning the spatial arrangements of remote communication systems or tabletop interfaces (Yamashita et al., 1999; Tang et al., 2010; He et al., 2019). In each condition, the participant pair needed to complete a set of 20 trials (one trial for each of the 16 targets, plus four additional trials as decoys to prevent participants from predicting remaining targets). In deciding condition orders, we grouped virtual and physical conditions to minimize configuration changes made between trials, resulting in a total of eight condition orders. We also prepared a total of eight distinct pseudorandom target sequences of cube numbers. We counterbalanced conditions and target sequences across all participants to avoid bias.

We measured task success (correct vs. incorrect responses), task completion time, and the observer's self-reported confidence level. We also gathered users' subjective feedback through an exit interview and filmed the experiment sessions for post-study observation (as shown in **Figure 3**). We computed the accuracy rate as the percentage of correct trials based on task success. To compute task completion time, we measured the elapsed time from when the pointer pressed ENTER key to display the target cube's number on their screen (Timestamp 1 in **Figure 4**) until the observer clicked a button to submit their final selection (Timestamp 2 in **Figure 4**). We did not separate the pointing and observing time because there was no reliable way to determine the completion of a pointing gesture for different users without complicating the experiment procedure (e.g., extra button click). Meanwhile, we observed from our pilot study that the pointing time did not contribute to variance in the overall performance. Therefore, we decided to compute the total task time, starting from the time when the pointer saw the number and ending when the observer made their final decision. We also recorded the observer's confidence level on a scale from one to seven, where one meant "not at all confident", and seven indicated "very confident".

In the exit interview, we focused on gathering subjective feedback from the participants to understand how cube type (physical vs. AR) influenced participants to consciously change their referencing gestures, and to see if participants had any specific strategies to cope with the visual occlusion problem. The questions were as follows:

- Pointer:
  1) *"Did you point to a virtual object in the same way as you did to a physical object? If so, why?"*
  2) *"Did you adapt your pointing gesture across trials? If so, why?"*
- Observer:
  1) *"How easy was it to recognize the target? How was recognizing AR targets different from physical objects?"*
  2) *"What strategy did you use?"*

C. Participants

Thirty-two participants (14 females, 18 males, mean age: 22.53 years) were recruited from the authors' university. The participants assigned to the pointer role were all right-handed. Seventeen of the participants had experienced virtual or augmented reality before. All of them had normal or corrected-to-normal vision. This experiment was approved by the university's Institutional Review Board. If participants agreed to participate, they were offered $12 in appreciation for their efforts.

D. Procedure

Participants were welcomed upon arrival and asked to read and sign an informed consent form. Then, they were invited to take a demographic survey. After the survey, the moderator introduced the experiment and tasks to the participants. The participants were reminded that in this experiment, only deictic gestures were allowed to reference the target object. Having confirmed that the participants understood the experiment, the moderator then explained the experiment procedures in detail and asked participants to practice the tasks in both experimental conditions until they felt confident.

We developed a program which gives instructions to the collaborators and measures the performance of the observer. The task procedure is illustrated in **Figure 4**, where the screenshots of the program are connected through user actions. To record the starting time, we put a keyboard in front of the pointer. When the pointer pressed the Enter key, a new trial started and a target number was shown on the screen (the lower-left screen in **Figure 4**). The pointer was asked to point at the target cube using any gesture they preferred. The observer was instructed to view the pointer's gesture to determine which cube was being referenced. Then, the observer selected the target number from a radial menu designed to minimize selection time (the lower-right screen in **Figure 4**). We recorded the end time when the observer clicked on the "Confirm" button in the center. After confirming their selection, they were asked to rate their confidence level on a scale of 1–7, as described earlier in this section (and as shown in the upper-right screen in **Figure 4**). Each trial ended when the observer entered their confidence level. The control would then switch back to the standby UI for the pointer to start another trial (the upper-left screen in **Figure 4**). The participants completed 20 trials for each experiment condition. After the participants finished all four conditions, we conducted an exit interview.

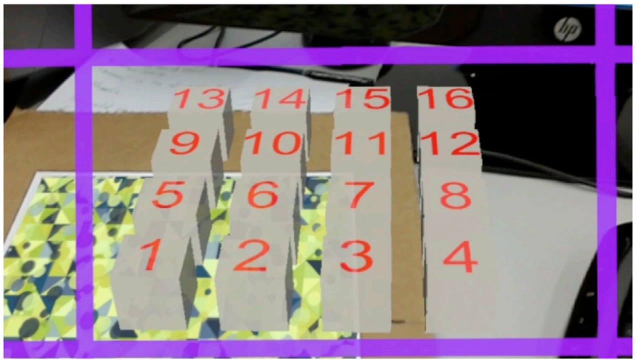

FIGURE 5 | We displayed purple stripes to illustrate the field of view to participants. The virtual content inside the rectangle formed by the four stripes is what the viewer can see. Outside the rectangle, participants could see the real world, but not the virtual cubes. We designed the layout of the virtual cubes such that the FOV from the seating positions was sufficient to see the whole set of cubes, minimizing the influence of the limited FOV.

E. Apparatus

In the physical conditions, the targets were sixteen 3.6 cm wooden cubes in a 4 × 4 layout glued on a wooden base (the image on the left in **Figure 1**). The layout of multiple cubes is modeled after applications that have small-scale visual context shared among users (e.g., board game, playing cards, sticky notes, urban planning, etc.) In the AR conditions, the virtual cubes were rendered at the same size and arranged in the exact same way as the wooden cubes. We used two Microsoft HoloLens 1 devices as the AR displays. To ensure that the virtual cubes were correctly aligned in both participants' views, we used the HWD's front camera and Vuforia to recognize a computer-generated, image-recognition-friendly picture we placed on the table in front of the participants. Once the HWDs detected the picture on the table, the AR system could then render the virtual cubes at the position of the recognized picture. Since both HWDs looked at the same image, it was guaranteed that the virtual cubes were seen in the same location by both users. Once the virtual cubes were stable relative to the image target, the moderator used a controller to turn off the image recognition function and freeze the cubes in space. After the image target was no longer needed, it was removed from the table to avoid the participants seeing it during the AR conditions. To change the participants' spatial relationships, we invited them to change seating positions from face-to-face seating to side-by-side seating. We added four purple stripes at the borders of the AR headset's screen to show the wearer the small effective field of view, as shown in **Figure 5**. The seating positions and the image target position were carefully

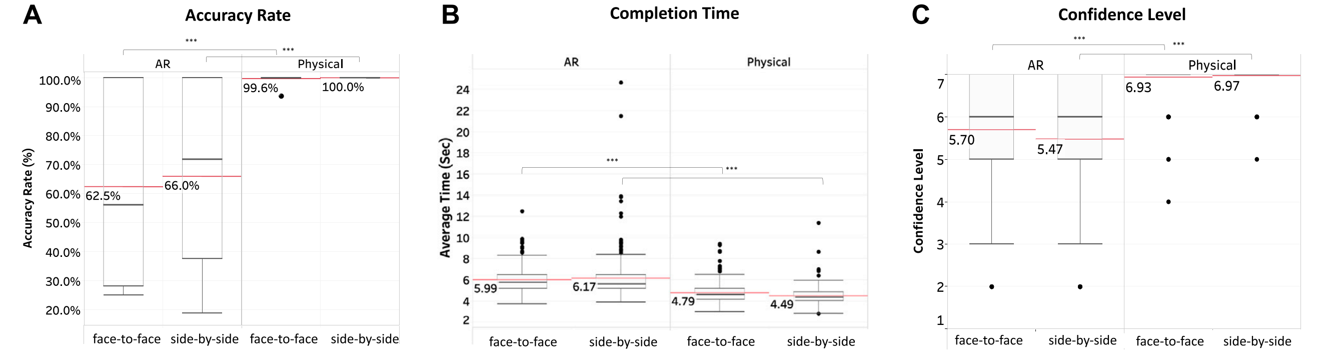

FIGURE 6 | Overall results for accuracy rate **(A)**, completion time **(B)** and confidence values **(C)**. Red lines with data labels indicate mean values, while black lines indicate median values.

chosen so that when the participants sat in their assigned seats, they were able to see all sixteen virtual cubes inside the display's field of view. To ensure consistency, we marked the positions of the chairs on the ground and image target on the table with tape. Between the AR conditions, after the participants changed their seating positions, we performed a recalibration process using the image target to again ensure that each participant saw the virtual cubes in the same physical locations.

The study program (**Figure 4**) was coded as a web-based application consisting of three UIs: the pointer UI, the observer UI, and the moderator UI, where moderators recorded the study information and controlled trials. During the study, the web program ran locally with XAMPP on a 13-inch Apple MacBook Pro. It was accessed *via* two Chrome windows: a window on the main monitor displayed the observer UI and the moderator UI; a window on an external monitor displayed the pointer UI. The moderator UI was only opened in between conditions for study setup; during trials, the window on the main monitor displayed the observer UI instead. We used a 12-inch Apple iPad Pro as the external monitor, connected with a cable to ensure both screens had similar size and visual quality. We chose to run both the observer and pointer UI on one computer with two different displays to minimize the effect of network delays on measured completion time.

The external monitor and an external keyboard connected by a cable were used for the pointer side, while the MacBook Pro and its built-in trackpad were used for the observer side. A 2-s tone was played when pointers pressed the Enter key to start pointing and when observers clicked the trackpad to finish observing. These tones provided audio feedback to the participants and experimenters. To analyze the participants' behavioral patterns and interview responses, we used a GoPro 7 video camera to record the entire session (See **Figure 3**.)

IV RESULTS

A. RQ1: Performance Decreases With Incorrect Occlusion

Our results indicate that the model-free AR condition negatively affects performance in referencing objects with bare hands compared to real-world referencing. The participants in the AR condition had lower accuracy rates, took more time to recognize targets, and were less confident about their choices. The spatial configuration (seating and target positions) affected participants' performance in the model-free AR condition.

Overall results are shown in **Figure 6**. A power anova test in the R package *pwr* and effect size analysis (cliff. delta and cohen. d) in the R package *effsize* validated our results. We used power calculations for balanced one-way analysis of variance tests (ANOVA power analysis reported 99.3%).

1) Accuracy Rate:

As expected, in the AR condition, the average accuracy rates across 16 pairs for face-to-face and side-by-side seating positions (62.5% with $\sigma = 0.49$ and 66.0% with $\sigma = 0.47$, respectively) were lower than those in the physical condition (99.6% with $\sigma = 0.06$ and 100.0% with $\sigma = 0.00$, respectively), as shown in **Figure 6B**. In the case of the physical condition, there was only one inaccurate answer out of 512 trials, in which the observer selected cube 14 when the correct target was cube 4. We ran a bias reduction generalized linear model (brglm) in the R package *brglm*, which is a flexible generalization of ordinary linear regression, to test the effects of cube type and seating position on accuracy rate (Nelder and Wedderburn, 1972). We applied *brglm* for accuracy rate analysis due to the following reasons: 1) the response, accuracy data, is binomial distribution, 2) generalized linear model provides benefits in small data sets, and 3) brglm has improvement over traditional maximum likelihood. The test result indicated that cube type had a statistically significant effect on accuracy ($p < 0.001$, $d = 0.36$), while seating position did not.

The accuracy rate varied depending on the target position. A heatmap of accuracy is shown in **Figure 7**. As the performance per target position (16 positions) highly depended on the cube types (2 types), the generalized linear model performed poorly due to multicollinearity. Instead, we conducted post-hoc analysis (brglm) on target position in each AR condition. We ran the effect of target position in three different ways: column (4 levels), row (4 levels), and Manhattan distance from the center [3 levels: four central cubes ($D = 1$ for 6, 7, 10, and 11), four corner

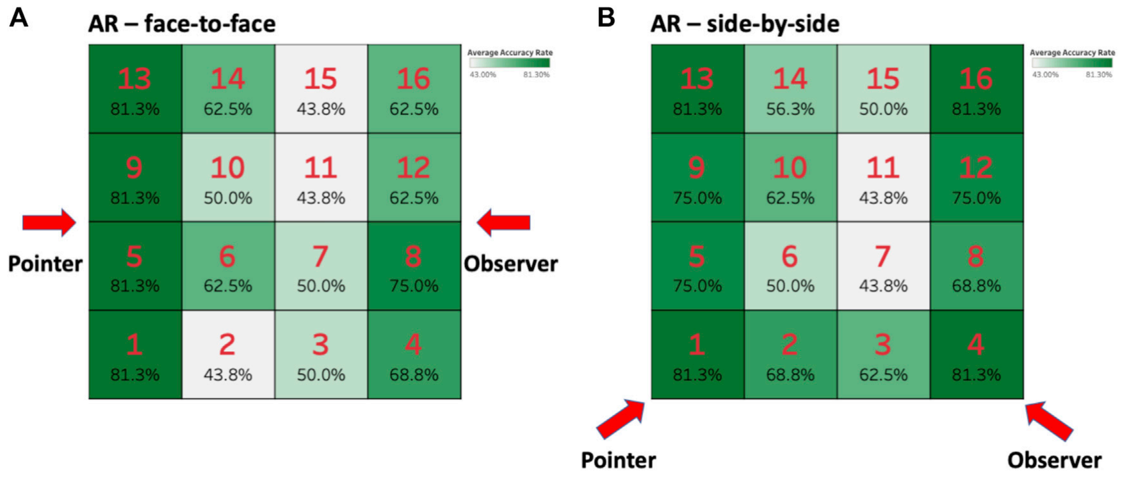

FIGURE 7 | Heatmaps of accuracy rate for the AR condition in two spatial configurations. **(A)** Heatmap of accuracy rate in the AR/face-to-face condition and **(B)** heatmap of accuracy rate in the AR/side-by-side condition.

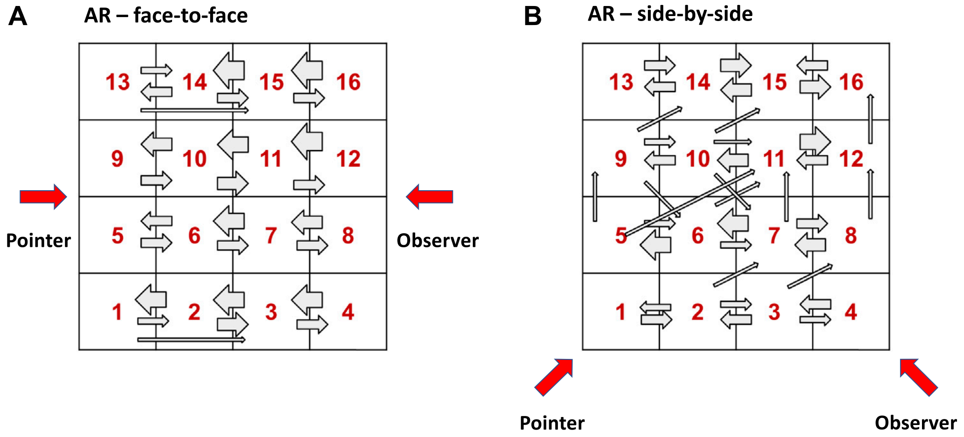

FIGURE 8 | The thickness of each arrow represents the number of errors. The head and tail of the arrow indicate the observer's selection and what the correct answer was. For example, a thick left arrow pointing from 2 to 1 visualizes the number of observers who answered Cube 1 when pointers were pointing at Cube 2. All errors are horizontal in the AR/face-to-face condition **(A)**, whereas there are some diagonal and vertical errors in the AR/side-by-side condition **(B)**.

cubes ($D = 3$ for 1, 4, 13, 16), and eight lateral cubes ($D = 2$)] for both conditions. In the AR/face-to-face condition, the effects of cube column ($p < 0.001$, $d = 0.25$) and Manhattan distance from the center ($p < 0.05$, $d = 0.11$) on accuracy rate were significant, with accuracy being best for the first column and worst for the third column. In the AR/side-by-side condition, the effect of cube column ($p < 0.01$, $d = 0.21$) and Manhattan distance from the center ($p < 0.01$, $d = 0.16$) on accuracy rate were significant, with accuracy being best for the corner cubes and worst for the third column.

Figure 8 visualizes all the errors (incorrect answers) that the participants made in the grid of cubes. Each arrow points from the correct answer to an observer's answer, and the thickness of the arrow indicates the number of observers who made the same mistake; for example, the thin arrow at the bottom of **Figure 8**-Left indicates that one observer chose 3 when the pointer pointed at 1 in the AR/face-to-face condition. The biggest arrow in **Figure 8**-Left represents six errors, and the biggest arrow in **Figure 8**-Right represents five errors. In the case of the AR/face-to-face condition, all errors are horizontal, and all of them except two cases are off-by-one errors.

In general, the left-facing arrows are broader than the right-facing arrows. This indicates that more observers made errors in which they chose a cube further away from them, picking cubes closer to the pointers than the ones that the pointers were actually pointing at. One possible explanation for this trend is that pointers may have referenced cubes by the direction of their finger (i.e., using an imaginary ray extending from a fingertip to point at a cube). In the meantime, from the other side, observers may not have been able to judge the position and direction of the observer's fingertip clearly due to the incorrect occlusion, so they made decisions based on *which target the fingertip appeared to be in*. This effect can be seen in **Figure 1**, where the fingertip in the AR condition appears to be in cube 2 or 3, while it is actually indicating a ray pointing at cube 4.

In the case of the AR/side-by-side condition, much more complex patterns appeared, with more diagonal and vertical (upward) errors. In general, we anticipated that in the side-by-side condition, observers would have more cues that they could use from the lateral view: for example, arm direction or the length of the arm. However, the result did not show any systematic patterns of errors that are apparent.

2) Completion Time:

Overall, the average completion time for all AR conditions was 6.08 s with $\sigma = 1.77$, and the average time for all physical conditions was 4.64 s with $\sigma = 0.93$. In the AR condition, the average completion times across 16 pairs (512 trials) for face-to-face and side-by-side seating (5.99 s with $\sigma = 1.25$ and 6.19 s with $\sigma = 2.18$, respectively) were greater than those of the physical condition (4.79 s with $\sigma = 1.00$ and 4.49 s with $\sigma = 0.83$), as shown in **Figure 6B**. Because the time distribution does not satisfy the assumption of normality, we performed a log-transformation of all completion time data points and confirmed the normality of the result. We used a linear mixed-effect model, LMER, with the *lme4* package in R (Chetverikov and Filippova, 2014). In contrast to traditional approaches, LMER allows controlling for the variance associated with random factors without data aggregation. Besides, we used the *lmer* for completion time analysis because completion time data is numerical and continuous. And this model could fit better with small sample size data, multiple parameters, and covariates. We discovered a significant interaction between cube type and seating position [F(1,1008) = 9.09, $p < 0.01$, $d = 0.17$], indicating that the differences in completion time cannot be explained by considering cube types and seating positions separately. The result of a post-hoc analysis (LMER and ANOVA test in R) revealed that the effect of cube type on completion time was significant regardless of seating position [F(1,1008) = 492.34, $p < 0.001$, $d = 1.00$]. The effect of seating position was significant only in the physical condition (F(1,496) = 18.33, $p < 0.01$, $d = 0.32$), where the side-by-side position was slightly faster than the face-to-face condition. We could not find any systematic pattern for the effect of target position on completion time, except that Cube 1, which is closest to the pointer, was the fastest across all conditions.

B. RQ2: Collaborators Develop Compensatory Strategies in the AR Condition

We aim to understand how participants perceive and cope with the challenge posed by incorrect occlusion. We used a thematic analysis approach to analyze the interview transcripts and observed study recordings in order to identify 1) challenges which pointers and observers faced during collaboration in the AR condition, 2) pointing strategies that pointers used in the AR condition, and 3) identification strategies that observers used in the AR condition.

1) Pointers' Strategies:

A few pointers reported difficulty in making a pointing gesture. Three pointers reported that incorrect occlusion caused difficulty when cubes overlapped their fingers. A majority of the pointers (9 out of 16) reported no difficulty in pointing, although their observers did not perform any better than the average (average accuracy rate: 61.1%).

- Incorrect occlusion: *"My finger in the screen was overlapped."* (P10)
- Incorrect occlusion: *"It feels like the block is being projected over my finger."* (P13)

Five pointers reported that the lack of touch caused difficulty (average accuracy rate: 69.3%). They mentioned that the lack of physical feedback was an obstacle in knowing whether their fingers had reached the correct cube.

- Lack of touch: *"It was sometimes a little more difficult because I felt like my finger was going, like, through the block layer."* (P6)
- Lack of touch: *"It doesn't feel so certain that my finger is actually hitting the block at the exact spot that I am trying to hit it on."* (P13)

While it would have been helpful if there was a haptic feedback for a pointer, we believe that this haptic feedback was the secondary cue that pointers were looking for *given* the incorrect occlusion. For example, in the physical cube condition, we did not see any consistent behaviors of pointers touching the cubes. In addition, observers should have not been able to see the touch anyway because of the incorrect occlusion even if there were haptic feedback.

We asked if pointers had any strategies in the AR condition, and 7 out of 16 pointers answered that they did not change their pointing gestures in the AR condition compared to the physical condition.

- *"I guess similar. Solid, I was touching it, but then this one I can't touch, it's almost like trying to touch it. Like, best I can."* (P2)
- *"Over time, I think I'd followed a pretty consistent strategy."* (P3)

Meanwhile, nine pointers answered that they used unique referencing strategies to cope with the challenges they faced in the AR condition, where they pointed differently from the physical condition. Based on what we observed in our study, we found two pointing gestures that pointers in the study used commonly. Some pointers used both strategies.

[Strategy I: Pointing from the Top (6/16)] Based on our observations from the recordings, one strategy involved hovering the hand relatively higher, keeping the palm facing down, and

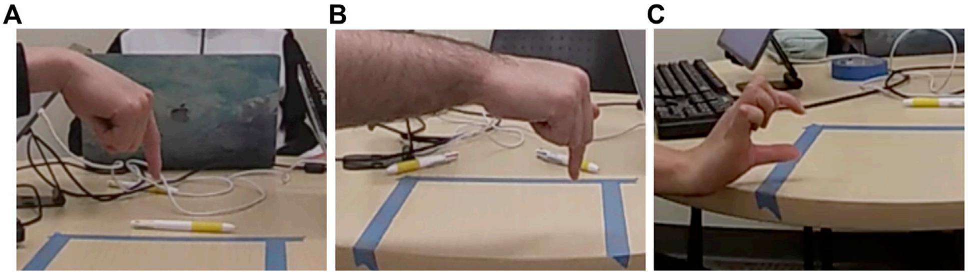

FIGURE 9 | (Strategy I) pointing straight down from the top **(A)**, (Strategy II) penetrating into the virtual cube **(B)**, (Other) grabbing a virtual cube **(C)**.

positioning only the primary reference finger (nearly) vertically; that is, nearly perpendicular to the top surface of a cube. We found six pointers used this gesture, based on the interview and the video recordings, uniquely in the AR condition, and their average accuracy rate was 69.2%, which is better than the overall average (64.3%). They believed this strategy would make their intentions clearer by avoiding confusion when cubes occluded the referencing finger, or when other cubes occluded the target cube.

[**Strategy II: Penetrating into the Virtual Cube (4/16)**] This strategy involved intentionally positioning the endpoint of the referencing finger inside a virtual cube. Four pointers answered they used this gesture uniquely in the AR condition. Naturally, this strategy was impossible to use with physical cubes. Identification of this strategy was strictly based on the observer's perception—that is, it was reported in interviews. We cannot truly evaluate whether a participant's fingertip was inside a virtual cube (as our recordings do not have virtual cubes rendered and we did not track pointers' hands). The average accuracy rate for participant pairs that used this strategy was 74.2%, yielding a better overall accuracy ratio than the average.

- *"When we were in AR, I tried to have my finger partially intersecting the block . . . I figured that if I was having my finger on top of the block, I'm not exactly sure based on his viewpoints how that may or may not look, since a lot of these things are kind of transparently looking."* (P12)

One interesting strategy we found in the recorded videos was a pointer's (P8) gesture of grabbing the cube (the rightmost picture in **Figure 9**) This strategy provides two visual cues, using the thumb and the index finger to simulate grabbing a physical cube. The pointer only used this strategy to reference cubes located around the edges of the grid, not all the time. The accuracy rate for this pair (pair 8) was 100% for all trials.

2) Observers' Strategies.

In summary, most of the observers expressed difficulty in recognizing what the pointers were pointing at. Only two observers reported that the observing experience was not hard. Nine observers reported that incorrect occlusion (the perceived overlap between virtual cubes and pointers' fingers) caused difficulty.

Eight observers believed there was a discrepancy between the virtual cubes they saw and the cubes their pointing partners saw, even though this was not the case due to the calibration step conducted before the AR condition.

- Incorrect occlusion: *"I couldn't tell [whether] she was pretty pointing to this and blocking this or touching this and pointing to this."* (O10)
- Incorrect occlusion: *"It was kind of hard to tell because sometimes, like, your finger would go, like, through a block. So it's kind of hard to tell, like, if he was pointing to this or this one."* (O9)
- Perceived discrepancy: *"There is some kind of misalignment between the pointer and the observer. So if he's pointing to the exact [top] of block, I think that he's finally at the center of four cubes."* (O1)
- Perceived discrepancy: *"I feel like my view was shifted a bit, because, like, what I saw was, like, in the middle."* (O12)

Overall, the observers were less confident about their answers in AR condition. In the AR condition, the average confidence levels across 16 pairs for the face-to-face and side-by-side seating positions (5.70 with $\sigma = 1.18$ and 5.47 with $\sigma = 1.33$, respectively) were lower than those for the physical condition (6.93 with $\sigma = 0.32$ and 6.97 with $\sigma = 0.20$), as shown in **Figure 6C**. The confidence value in the physical condition was close to the maximum value (7), illustrating that essentially all observers felt that their answers were correct without a doubt. We ran an ordinal regression model in the R package *MASS* for ordinal dependent variables to analyze this effect and found that cube type had a significant effect on self-reported confidence ($p < 0.001$, $d = 1.55$) (McCullagh, 1980). The effect of seating position was not significant. We utilized the ordinal logistic regression for confidence analysis because confidence level is categorical data from 1 to 7. And, it works to predict the dependent variable with "ordered" multiple categories and independent variables.

When asked if they had employed any particular strategies, 14 observers said that they did so to overcome visual occlusion problems. Here, we present two themes that emerged in the thematic analysis of the interviews. The groups represented in these two themes were mutually exclusive.

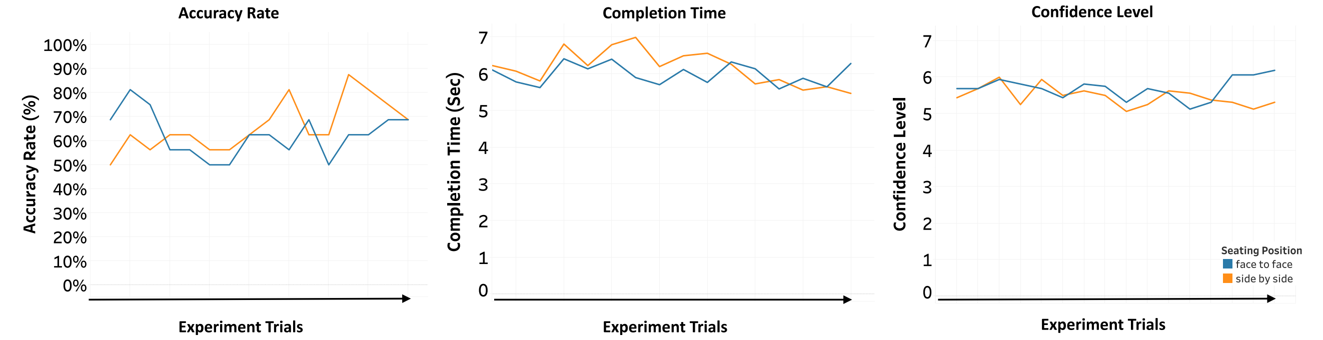

FIGURE 10 | Average accuracy rate, average completion time, and average confidence across trials in the AR condition (face-to-face, side-by-side). There was no apparent trend of improvement over repeated trials.

[**Learning the Gesture Pattern (10/16)**] This strategy involved learning the pointer's hand gesture pattern as it emerged during the study, and using that knowledge as supplementary cues to help guess and identify the cubes. This strategy is based on an assumption that the observer's previous identification is roughly correct. Ten observers compared their partners' referencing gestures with gestures made in previous trials that they identified. In addition, they all reported that their identification grew faster as they learned their partners' patterns.

- *"I did start to, like, learn, you know, especially when I realized, like, how his was working if he was [pointing] between them. I was like, okay, he probably is pointing at the previous one. So I did get faster."* (O5)
- *"I think as we went along, I sort of figured out a way, for example, oh, if it's really low, he is just pointing to the block of the first row."* (O3)

To verify their belief, we reviewed these participants' performance (accuracy rate and completion time) and their confidence over trials. However, we could not find any evidence that the participants were learning over trials. As we did not give any feedback on whether observers' answers were correct or not, they believed that they were getting better over time, when in fact, they were not, as demonstrated in **Figure 10**.

[**Envisioning the Pointer's Perspective (4/16)**] This strategy involved envisioning the partner's perspective and choosing a cube that was not necessarily the closest to the referencing finger physically, but the cube that an observer imagined the pointer was pointing at. One observer imagined her partner's finger ended on the cubes that were one row to the right of the cubes they saw. Another observer used certain virtual cubes as anchors and mentally aligned the pointer's finger based on direction and how far his finger was from the anchor cubes.

- *"I was starting to think about his perspective . . . if you, like, if you point over here to the right a little bit, then I would like, like, shift everything to the right."* (O2)
- *"Use the virtual cubes as kind of, like, the anchor . . . Yeah, the same way that you use physical cubes, like, as the anchor, like it's this far in front of four."* (O10)

In general, the observers seemed to expend more cognitive effort than would have been necessary were the targets occluded correctly.

V DISCUSSION

A. Research Questions

Assuming that the incorrect occlusion would negatively influence the collaborative performance of barehanded referencing in model-free AR, we wanted to know to what extent performance would be degraded (**RQ1-1**). The experiment results demonstrated that the average accuracy rate was reduced by 35.6% on average, and the task completion time was increased by 1.44 s (31.0% increase). Without verbal communication, in model-free AR, the visual cues become less reliable: targets are no longer correctly occluded by the pointer's hand, and the pointer's hand shadow no longer interacts with the targets. We speculated that observers would need to seek other information to complete the task, such as pointers' gesture patterns. As such information is neither as salient nor as reliable as occlusion, observers must spend time interpreting them and cognitive load dealing with the uncertainty, resulting in lower accuracy and slower decision-making. While these cues resulted in higher accuracy than random guessing (1/16), this accuracy is still far below that of the physical condition, implying that the strategies that the pointers take are insufficient to support barehanded referencing in model-free AR, at least on the scale of our tests. Our findings indicate the potential to exploit various pointing gestures to better support barehanded referencing in AR. For example, based on the fact that pair 8 achieved 100% for all trials, a follow-up study could be carried to evaluate the grabbing gesture against other pointing gestures.

Additionally, we hypothesized that the spatial configuration (viz., the positions of the collaborators and the target cubes) within model-free AR might be relevant to the accuracy loss and analyzed its impact on the collaborators' performance (**RQ1-2**). Our analysis revealed that users performed significantly better in some spatial configurations than in others, even with incorrect occlusion. We reasoned that under different spatial configurations, the observers would encounter different occlusion scenarios, with some of them (e.g., virtual objects closer to the pointer) potentially being easier to resolve than the others. For instance, in the face-to-face condition, referencing cubes 5 and 9 should be easier than referencing cubes 8 and 12, because in the former case, only one cube is likely to occlude the pointer's hand. Our analysis did not find that seating positions alone had any influence, but it did support the influence of cube positions. Because the change in perspectives caused by different seating positions did not seem to have an impact on referencing performance in our study, this result suggests that viewing angle does not significantly influence the effectiveness of strategies that the pointers take. However, we expect that in more complicated model-free AR scenarios, references to virtual targets on edges and corners, or to isolated targets, will be easier for observers to understand no matter where the collaborators are located.

We expected that pointers would adapt their pointing gestures in the model-free AR condition in order to mitigate the difficulties posed by the incorrect occlusion (**RQ2-1**). The interview provided valuable insight into the strategies they developed from their perspectives. In summary, 9 out of 16 participants reported changing their pointing gestures, specifically in the model-free AR condition. Their motivation was mostly related to a lack of sensation of touch on the target cubes. Due to the fact that the pointers' own hands were incorrectly occluded, they had problems knowing if their hands were aligned properly with the target virtual cubes. In their comments, the pointers indicated that since they could not touch the virtual cubes, they could not perform the same pointing gestures as they did in the physical condition. However, the lack of tangible feedback should not directly impact the observer viewing the pointing gesture, since the observers were not making judgements based on tangible sensations. It is worth noting that the pointers used two diametrically opposed mitigation strategies. Some tried to minimize virtual content wrongly occluding their hands by pointing at cubes from above, or even grabbing cubes. Others decided to penetrate the virtual cubes to help their collaborators. We should also note that because participants did not switch roles, the pointers may not have been fully able to understand what the observers were seeing. Since the pointers did not see the incorrect occlusion from the observers' perspective, it could have been difficult for them to develop a strategy to refine their pointing gestures and thereby improve collaborative performance. Based on these findings, we plan to run a follow-up study to evaluate the two diametrically opposed mitigation strategies that will allow the collaborators to switch roles to establish a better understanding of the incorrect occlusion.

Finally, we expected to see changes from the observers' side as well (**RQ2-2**). In sum, the majority of the observers admitted seeking other information through learning. Since we did not tell observers whether their answers were correct, we believe that the observers tried to learn about the effectiveness of pointing gestures from easier trials, which they took as sources of ground truth. However, this learning effect was not present in our performance analysis, which contradicts the findings from Chastine and Zhu (2008). The mismatch between the observers' responses and objective results could have been caused by a lack of feedback. Without knowing if their responses were correct, observers were not equipped to learn from the trials. In practice, resolving referential ambiguity constantly through correction may actually help observers improve their accuracy rates. This can be tested by repeating the study and provide feedback to the observers.

B. Design Implications

Generally speaking, the results indicate that, with incorrect occlusion of the user's hand, deixis will be severely restricted and collaborative tasks will be jeopardized. Our work also demonstrates the role of spatial configurations in collaborative referencing tasks and sheds light on alternative pointing gestures that can provide useful information when correct occlusion is missing. Here, we discuss a few design implications that can alleviate the reference problem.

As we observed during the study, the cubes in the column nearest the pointer in the face-to-face AR condition were significantly easier for the observers to identify. Designers should therefore be able to reduce referencing ambiguity caused by incorrect occlusion by changing the spatial layout of virtual content. In particular, designers can avoid placing virtual content along the collaborators' view directions, rather dispersing the content in front of them. Another approach is to increase the distance between adjacent objects to reduce possible overlap between users' deictic gestures and non-target objects. A similar effect can be achieved by dynamically changing the scale of the virtual content during referencing actions. The idea can be further strengthened with an automatic layout adjustment. We plan to explore a system that supports dynamic arrangement of the referenced targets based on the collaborators' spatial arrangement.

Furthermore, based on the results from our experiment, one potential solution to provide more information to observers is to share views among collaborators, as pointed out by other researchers (Chastine and Zhu, 2008). Although we observed that the pointers adopted various pointing gestures in response to the incorrect occlusion, their strategies were limited by their perspectives. If the pointers were able to see the observers' views, it could have helped the pointers develop new ways of referencing virtual targets in model-free AR that are effective for observers. One typical approach for sharing perspectives among collaborators is to stream the other users' views in the shared virtual space.

The approaches above do not apply when virtual objects are continuous, such as terrain or buildings. In these cases, designers use additional tracking technologies to support robust referencing communication. If the user's hand can be tracked with six degrees of freedom (both position and orientation), but it is not practical to reconstruct an accurate hand model, one method designers can use is to define a virtual pointing ray emanating from the hand in 3D. This ray can then intersect with target locations. In this way, the observer needs only to look for an intersection between the ray and the target. We observed this kind of behavior in the user study, where pointers used an index finger to indicate a pointing direction. However, this strategy was ineffective, since observers found it hard to interpret the pointing direction without a virtual ray. Kim et al. (Kim et al., 2020) adopted a similar approach and found its benefit in referencing tasks.

C. Limitations

There are some limitations to our study that necessitate further work. First, the correctness of occlusion cues was not the only difference between the AR and physical conditions, as mentioned in 3.1. Other factors, mostly stemming from the HWD, might have influenced the results. The limited FoV in the AR condition constrained both collaborators by limiting the area in which they could see the virtual cubes. Another factor was the partial transparency of the virtual cubes when seen through the HWD. Several participants complained of blurry vision when wearing the headset, potentially leading to less clear perceptions of pointers' hands. Some participants also reported drifting of the virtual cubes when they turned their heads. However, we did recalibrate the virtual cubes between trials using the image target to minimize possible drift. These self-reports of drift could also have been caused by perceptual errors related to the incorrect occlusion cues, the translucency of the virtual objects, or minor device movements over time, causing the cubes to not appear fixed to the table. Therefore, it is possible that the performance decrease we observed was not entirely due to incorrect occlusion in the AR condition, though we believe it to be the most significant factor for barehanded referencing.

Moreover, there could be bias in the participants' demographic backgrounds, since we only recruited university students from our campus. To verify if the findings of this research would apply to the broader population, we need a larger-scale study with a more diverse demographic. Moreover, the target layout we explored was kept simple and discrete in order to isolate spatial factors. This layout does not fully reflect the complexity of real-world tasks, where targets might be locations on or features of an object, rather than discrete objects. In a future study, we could observe users' spatial referencing behavior in more ecologically valid settings (e.g., a brainstorming application).

Another factor worth noting is the use of Optical See-Through (OST) HWDs. Prior research has identified that using OST HWDs can lead to overestimation of virtual target distances (Wann et al., 1995; Mon-Williams and Tresilian, 2000; Swan et al., 2015) According to Swan et al. (Swan et al., 2015), when the AR targets are 50 cm away, the overestimation should be about 2 cm, which is at the same magnitude as the target cube in our study. However, we argue that the impact from OST display is minimal. If there were a major influence of distance mis-estimation on both the pointer and the observer, there would be a consistent pattern in **Figure 8** as a result of such perceptual discrepancies. Especially for the face-to-face condition, a systematic overestimation would tend to result in observers selecting cubes that are closer to themselves. The most reasonable explanation for our results is that incorrect occlusion is the dominant factor causing errors in our study.

VI CONCLUSION AND FUTURE WORK

Spatial referencing of virtual objects is important in many close-range, co-located, collaborative AR scenarios. Among various referencing methods, referencing with the user's bare hand is a naturalistic and effective way of interaction in the real world. However, barehanded referencing will likely be influenced in model-free AR settings when correct occlusion cues are missing. In this work, we studied the effects of model-free AR on barehanded referencing. We found that participants' performance was indeed reduced in model-free AR settings, and that the participants used various mitigation strategies to accomplish the task, though these were not effective. The experiment revealed that spatial configurations of the targets relative to the collaborators significantly influenced performance.

Our research's principal contributions include the analysis of major factors affecting collaboration performance in a model-free AR condition, the results and implications of our controlled empirical study, and design implications for collaborative AR systems involving barehanded spatial referencing.

There are many possible directions for future research. Most importantly, we plan to design a user interface that incorporates modern hand tracking and modeling. Our current study omits any tracking technology in favor of focusing on the baseline condition. Additional information from a tracking system could help improve collaborative performance; alternatively, it may introduce systemic bias. Moreover, we plan to test the interface in a more ecologically valid task that is less abstract than the one in this study. The current experiment assumes the targets are discrete, which is not always the case. For example, we could study such tasks as factory machine inspections or collaborative map inspections. Last but not least, based on the design implications from the study, we also plan to develop a system that automatically change the layout and scale of the virtual content based on the users' positions and viewing angles to attenuate possible ambiguities caused by incorrect occlusion.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the approved IRB protocol for the study restricts that only the investigators of the study will have access to the stored data. Requests to access the datasets should be directed to the authors and the IRB office at Virginia Tech for further evaluation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Division of Scholarly Integrity and Research Compliance, Institutional Review Board, Virginia Tech. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YL, DB, and SL contributed to the conception, design of the study. YL and BW contributed to the execution of the study. DH performed the statistical analysis. YL, DH, and BW wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to acknowledge the generous funding support provided by the Office of Naval Research.

REFERENCES

Alibali, M. W., Heath, D. C., and Myers, H. J. (2001). Effects of Visibility between Speaker and Listener on Gesture Production: Some Gestures Are Meant to Be Seen. *J. Mem. Lang.* 44 (2), 169–188. doi:10.1006/jmla.2000.2752

Allen, Gary. (2003). Gestures Accompanying Verbal Route Directions: Do They point to a New Avenue for Examining Spatial Representations? *Spat. Cogn. Comput. - SPAT COGN COMPUT* 3 (259–268), 12. doi:10.1207/s15427633scc0304_1

Beattie, G., and Shovelton, H. (1999). Mapping the Range of Information Contained in the Iconic Hand Gestures that Accompany Spontaneous Speech. *J. Lang. Soc. Psychol.* 18 (4), 438–462. doi:10.1177/0261927x99018004005

Boboc, R. G., Gîrbacia, F., Postelnicu, C. C., and Gîrbacia, T. (2019). "Evaluation of Using mobile Devices for 3d Reconstruction of Cultural Heritage Artifacts," in *VR Technologies in Cultural Heritage*. Editors M Duguleană, M Carrozzino, M Gams, and I Tanea (Cham: Springer International Publishing), 46–59. doi:10.1007/978-3-030-05819-7_5

Breen, D. E., Whitaker, R. T., Rose, E., and Tuceryan, M. (1996). Interactive Occlusion and Automatic Object Placement for Augmented Reality. *Computer Graphics Forum*. 15 (3), 11–22. doi:10.1111/1467-8659.1530011

Chastine, J. W., and Zhu, Y. (2008). "The Cost of Supporting References in Collaborative Augmented Reality" in Proceedings of Graphics Interface 2008, Windsor, ON (Toronto, ON: Canadian Human-Computer Communications Society), 275–282.

Chastine, J., Nagel, K., Zhu, Y., and Hudachek-Buswell, M. (2008). "Studies on the Effectiveness of Virtual Pointers in Collaborative Augmented Reality," in *2008 IEEE Symposium on 3D User Interfaces*, 117–124.

Chetverikov, A., and Filippova, M. (2014). How to Tell a Wife from a Hat: Affective Feedback in Perceptual Categorization. *Acta Psychologica*. 151, 206–213. doi:10.1016/j.actpsy.2014.06.012

Clark, H. H., Susan, E., and Brennan (1991). *Chapter Grounding in Communication*. Washington, DC, US: American Psychological Association, 127–149.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a Collaborative Process. *Cognition*. 22 (1), 1–39. doi:10.1016/0010-0277(86)90010-7

Cohen, A. A., and Harrison, R. P. (1973). Intentionality in the Use of Hand Illustrators in Face-To-Face Communication Situations. *J. Personal. Soc. Psychol.* 28 (2), 276–279. doi:10.1037/h0035792

Comport, A. I., Marchand, E., Pressigout, M., and Chaumette, F. (2006). Real-time Markerless Tracking for Augmented Reality: the Virtual Visual Servoing Framework. *IEEE Trans. Vis. Comput. Graphics*. 12 (4), 615–628. doi:10.1109/tvcg.2006.78

Cutting, J. E., and Vishton, P. M. (1995). "Chapter 3 - Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Information about Depth*," in *Perception of Space and Motion, Handbook of Perception and Cognition*. Editors W Epstein and S Rogers (San Diego: Academic Press), 69–117.

Dare, A. (1995). *Baldwin. Chapter Understanding the Link between Joint Attention and Language*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, 131–158.

Dix, A. (1994). *Computer Supported Cooperative Work: A Framework*. London: Springer London, 9–26.

Hayashi, K., Kato, H., and Nishida, S. (2005). "Occlusion Detection of Real Objects Using Contour Based Stereo Matching," in Proceedings of the 2005 International Conference on Augmented Tele-Existence, ICAT '05 (New York, NY, USA: Association for Computing Machinery), 180–186.

He, Z., Rosenberg, K. T., and Perlin, K. (2019). "Exploring Configuration of Mixed Reality Spaces for Communication," in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–6.

Huang, W., Alem, L., Tecchia, F., and Duh, H. B.-L. (2018). Augmented 3d Hands: a Gesture-Based Mixed Reality System for Distributed Collaboration. *J. Multimodal User Inter.* 12 (2), 77–89. doi:10.1007/s12193-017-0250-2

Jeffrey, W. (2007). *Chastine. On Inter-referential Awareness in Collaborative Augmented Reality*. USA: PhD thesis, AAI3278579

Kim, S., Lee, G., Huang, W., Kim, H., Woo, W., and Billinghurst, M. (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, CHI '19 (New York, NY, USA: Association for Computing Machinery), 1–13.

Kim, S., Jing, A., Park, H., Lee, G. A., Huang, W., and Billinghurst, M. (2020). Hand-in-air (Hia) and Hand-On-Target (Hot) Style Gesture Cues for Mixed Reality Collaboration. *IEEE Access*. 8, 224145–224161. doi:10.1109/access.2020.3043783

Kirk, D., and Fraser, D. S. (2006). "Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06 (New York, NY, USA: Association for Computing Machinery), 1191–1200.

Kita, S., and Özyürek, A. (2003). What Does Cross-Linguistic Variation in Semantic Co-ordination of Speech and Gesture Reveal?: Evidence of an Interface Representation of Spatial Thinking and Speaking. *J. Mem. Lang.* 48 (16–32), 01.

Kiyokawa, K., Billinghurst, M., Campbell, B., and Woods, E. (2003). "An Occlusion Capable Optical See-Through Head Mount Display for Supporting Co-located Collaboration," in *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, 133–141.

Krauss, R. M., Chen, Y., Gottesman, R. F., and Krauss, R. M. (2000). Lexical Gestures and Lexical Access: A Process Model. In D. McNeill [Ed.], *Language and Gesture*. Cambridge, United Kingdom: Cambridge University Press, 261–283. doi:10.1017/cbo9780511620850.017

Lee, G. A., Billinghurst, M., and Kim, J. (2004). "Occlusion Based Interaction Methods for Tangible Augmented Reality Environments," in *Proceedings VRCAI 2004 - ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry*, 419–426. Proceedings VRCAI 2004 - ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry ; Conference date: 16-06-2004 Through 18-06-2004.

Martha, W. (2005). Alibali. Gesture in Spatial Cognition: Expressing, Communicating, and Thinking about Spatial Information. *Spat. Cogn. Comput.* 5 (4), 307–331. doi:10.1207/s15427633scc0504_5

McCullagh, P. (1980). Regression Models for Ordinal Data. *J. R. Stat. Soc. Ser. B (Methodological)*. 42 (2), 109–127. doi:10.1111/j.2517-6161.1980.tb01109.x

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL, US: University of Chicago Press.

Mendez, E., and Schmalstieg, D. (2009). "Importance Masks for Revealing Occluded Objects in Augmented Reality," in Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology, VRST '09 (New York, NY, USA: Association for Computing Machinery), 247–248.

Mon-Williams, M., and Tresilian, J. R. (2000). Ordinal Depth Information from Accommodation? *Ergonomics*. 43 (3), 391–404. doi:10.1080/001401300184486

Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., and Theobalt, C. (2017). "Real-time Hand Tracking under Occlusion from an Egocentric Rgb-D Sensor," in Proceedings of International Conference on Computer Vision (ICCV), Venice, Italy (IEEE).

Nelder J. A., WedderburnR. W. M. (1972). Generalized Linear Models. *J. R. Stat. Soc. Ser. A (General)*. 135 (3), 370–384. doi:10.2307/2344614

Oda, O., and Feiner, S. (2012). "3d Referencing Techniques for Physical Objects in Shared Augmented Reality," in 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Atlanta, GA, November 5–8, 2012 (IEEE), 207–215.

Olson, G. M., and Olson, J. S. (2000). Distance Matters. *Hum.-Comput. Interact.* 15 (2), 139–178. doi:10.1207/s15327051hci1523_4

Oviatt, S., DeAngeli, A., and Kuhn, K. (1997). "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction," in *Referring Phenomena in a Multimedia Context and Their Computational Treatment, ReferringPhenomena '97* (USA: Association for Computational Linguistics), 1–13.

Oviatt, S. (1996). Multimodal Interfaces for Dynamic Interactive Maps. in Proceedings Of the SIGCHI Conference On Human Factors In Computing Systems, CHI '96. New York, NY, USA: Association for Computing Machinery, 95–102.

Richard, A. (1980). "Bolt. "Put-that-there": Voice and Gesture at the Graphics Interface," in Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '80 (New York, NY, USA: Association for Computing Machinery), 262–270.

Swan, J. E., Singh, G., and Ellis, S. R. (2015). Matching and Reaching Depth Judgments with Real and Augmented Reality Targets. *IEEE Trans. Vis. Comput. Graphics*. 21 (11), 1289–1298. doi:10.1109/tvcg.2015.2459895

Szalavári, Z., Schmalstieg, D., Fuhrmann, A., and Gervautz, M. (1998). Studierstube": An Environment for Collaboration in Augmented Reality. *Virtual Reality*. 3 (1), 37–48.

Tang, A., Pahud, M., Inkpen, K., Benko, H., Tang, J. C., and Buxton, B. (2010). Three's Company: Understanding Communication Channels in Three-Way Distributed Collaboration. in Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, 271–280.

Wann, J. P., Rushton, S., and Mon-Williams, M. (1995). Natural Problems for Stereoscopic Depth Perception in Virtual Environments. *Vis. Res*. 35 (19), 2731–2736. doi:10.1016/0042-6989(95)00018-u

Whittaker, S. (2003). Things to Talk about when Talking about Things. *Human–Computer Interaction* 18 (1-2), 149–170. doi:10.1207/s15327051hci1812_6

Yamashita, J., Kuzuoka, H., Yamazaki, K., Miki, H., Yamazaki, A., Kato, H., et al. (1999). Agora: Supporting Multi-Participant Telecollaboration. *HCI* 2, 543–547. doi:10.5555/647944.743622

Yoon, B., Kim, H. i., Oh, S. Y., and Woo, W. (2020). "Evaluating Remote Virtual Hands Models on Social Presence in Hand-Based 3d Remote Collaboration," in 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Porto de Galinhas, Brazil, November 9–13, 2020 (IEEE), 520–532.

Yuan, L., Lu, F., Wallace, S., and Bowman, D. (2019). "Gaze Direction Visualization Techniques for Collaborative Wide-Area Model-free Augmented Reality," in *Symposium on Spatial User Interaction, SUI '19* (New York, NY, USA: Association for Computing Machinery).

Zahn, G. L. (1991). Face-to-Face Communication in an Office Setting. *Commun. Res*. 18 (6), 737–754. doi:10.1177/009365091018006002

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