Do You Really Need to Know Where “That” Is? Enhancing Support for Referencing in Collaborative Mixed Reality Environments

Janet G Johnson
janetjohnson@ucsd.edu
UC San Diego
La Jolla, CA, USA

Danilo Gasques
gasques@ucsd.edu
UC San Diego
La Jolla, CA, USA

Tommy Sharkey
tsharkey@ucsd.edu
UC San Diego
La Jolla, CA, USA

Evan Schmitz
ewschmit@uw.edu
University of Washington
Seattle, WA, USA

Nadir Weibel
weibel@ucsd.edu
UC San Diego
La Jolla, CA, USA

ABSTRACT

Mixed Reality has been shown to enhance remote guidance and is especially well-suited for physical tasks. Conversations during these tasks are heavily anchored around task objects and their spatial relationships in the real world, making referencing - the ability to refer to an object in a way that is understood by others - a crucial process that warrants explicit support in collaborative Mixed Reality systems. This paper presents a 2x2 mixed factorial experiment that explores the effects of providing spatial information and system-generated guidance to task objects. It also investigates the effects of such guidance on the remote collaborator’s need for spatial information. Our results show that guidance increases performance and communication efficiency while reducing the need for spatial information, especially in unfamiliar environments. Our results also demonstrate a reduced need for remote experts to be in immersive environments, making guidance more scalable, and expertise more accessible.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*; **Mixed / augmented reality**; *Collaborative interaction.*

KEYWORDS

Collaboration, Mixed Reality, Referencing, Remote Guidance

ACM Reference Format:
Janet G Johnson, Danilo Gasques, Tommy Sharkey, Evan Schmitz, and Nadir Weibel. 2021. Do You Really Need to Know Where “That” Is? Enhancing Support for Referencing in Collaborative Mixed Reality Environments. In *CHI Conference on Human Factors in Computing Systems (CHI ’21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445246

CHI ’21, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445246

1 INTRODUCTION

As expertise becomes increasingly distributed, and skilled personnel are not always available nearby, technology-mediated remote guidance is gaining traction as a significant area of interest in many domains like education, manufacturing, design, and healthcare. These domains typically involve physical tasks – like mechanical assembly, emergency repairs, or surgery – with collaborators manipulating real-world objects. When collaborators are in physically separated spaces, technology-mediated remote guidance substantially improves their ability to complete the task [38].

Given how inherently spatial physical tasks are, the immersive nature of Mixed Reality (MR) [41] – a technology that merges the real and virtual worlds – makes it especially well-suited as a collaborative medium. Collaborative MR creates environments where collaborators can feel like they are in the same 3-dimensional space. This improves their shared context and restores their natural ability to understand and interact with spatial cues [5, 46, 56]. In general, MR allows for communication behaviors similar to face-to-face collaborations, especially when compared to traditional collaborative interfaces on screens [5].

The growth of computer vision, mobile computing, and wearable technology is making MR more commonplace and affordable, and researchers in both academia and the industry are increasingly building and understanding collaborative MR systems for remote guidance [17]. However, despite recent advances in the field, domain- and technology-specific constraints (like limited bandwidth, field-of-view limitations, and ergonomic issues) makes it impossible for a collaborative MR system to provide all the information required to maintain the effectiveness of remote communication at the level of face-to-face interactions [5, 24, 37]. Designing MR systems that reduce collaboration effort during remote guidance requires us to understand and prioritize the different communication properties they afford and identify the critical elements that can be optimized.

In this paper, we focus on *referencing* – a communication process essential to the success of collaborative physical tasks. In particular, we explore the effects of explicitly providing spatial information and offloading the referential process with MR guidance through a mixed factorial study. Specifically, this paper contributes:

(1) An evaluation of how *explicitly representing spatial information* supports referencing for physical tasks.
(2) A discussion of how visual guidance in Mixed Reality can *offload the referential process* and reduce collaborators' efforts; and an empirical understanding of how this impacts the remote collaborator's needs for spatial information.
(3) A discussion of the implications for designers and developers, including opportunities for scaling up Mixed Reality collaborative systems for remote guidance.

2 BACKGROUND AND RELATED WORK

Collaborative physical tasks are those where people work together to perform actions on objects in the real world [24, 37]. We specifically focus on remote guidance or the "mentoring" scenario – a type of collaboration where one person, the *worker* (or *novice*), directly manipulates objects with the guidance of a *helper* (or *expert*).

Guidance for physical tasks heavily revolves around referring to objects and describing actions to be done on them [24]. The helper's ability to provide guidance is thus heavily dependent on the ease with which they can refer to the objects used as part of the task they want to accomplish. *Referencing* – or the act of making a reference and having that reference understood by others [10, 72] – is thus a crucial process that collaborative systems need to support [10]; and the absence of support for referencing severely hinders communication [19].

In this section, we first look into collaborators' information needs with respect to referencing during remote guidance. We then discuss how today's collaborative MR systems support these needs and highlight opportunities that the technology provides to enhance the same.

2.1 The Process of Referencing

Remote collaborators or helpers use various means to refer to the objects required to accomplish a physical task [18]. However, they typically use the method that requires the least effort to enable workers to understand their references [12, 13]. Given the opportunity, collaborators prefer to use short phrases, alternate descriptions, deictic expressions, and simple gestures (like pointing) to efficiently refer to objects without lengthy verbal descriptions [38, 44]. Deictic expressions are used to point to things, people, or locations and contain words like "this", "here", or "that". These expressions are also often combined with gestures such as pointing. For example, someone could point at an object and say "bring *that* one over here" instead of verbally describing what and where the object is.

To employ deixis in the above example, the person needs to know where the object is to point to it – they also need to share enough context with their collaborators for those receiving the reference to understand what "that" means. When collaborators are physically co-present, they share a rich visual space that facilitates the awareness required to generate these references and provide the context required to understand them [13, 37].

Therefore, to be effective, remote collaboration systems need to explicitly build features that support initiating and understanding references [21, 37]. Chastine et al. defined a detailed framework for the *referential process* or *inter-referential awareness* in collaborative physical tasks [10]. At it's core, the referential process is split into two sub-processes that occur sequentially:

(1) **Making a Reference**: This is when the helper refers to an object in the task environment. For example, the helper could say "can you pick up *the red box*?".
(2) **Understanding the Reference**: This is the process by which the worker understands the reference and identifies the object being referred to.

Making a reference requires the helper to know *what* objects are present in the task environment. This awareness of the presence of task objects is crucial to the helper's ability to refer to them, and its absence will cause the referential process to break down. We call this information the *awareness of (task) objects*.

Support for understanding the reference is also crucial, especially since the worker is typically unfamiliar with the task and the environment, and would require additional information to understand the reference. Helpers often point or use phrases like "the blue wrench *to your right*" and "*that* one right *there*" to direct attention to the referenced object. These phrases use deictic gestures and expressions that require the helper to know *where* the object is with respect to the worker and other objects in the environment [69]. *Spatial information* – or information about the spatial relations between objects and workers – is therefore fundamental to the helpers' ability to provide guidance.

2.2 Referencing in Collaborative Mixed Reality

In order to support a communication process, collaborative systems provide information collaborators need [21, 37] either: (1) *passively* through features that share the task environment, or (2) *explicitly* through features specifically designed for it [19]. In this section, we discuss how the need for awareness of objects and spatial information are both passively and explicitly fulfilled in current MR systems. As we do so, we highlight the opportunities that MR as a medium presents to further support the referential process.

Viewing the Task Environment – The importance of a shared task space is strongly emphasized in existing research. Providing a view of the task environment enhances awareness and builds common ground, allowing for more efficient communication between the helper and the worker [23, 24, 37, 38]. As a result, all collaborative MR systems provide some way for the helper to view the worker's environment and thus passively support various communication processes – including referencing.

Researchers have increasingly explored immersing helpers in 3D reconstructions of a worker's environment to bring the collaborative MR setting closer to face-to-face collaboration [5]. Systems like those introduced by Bai et al. [3], Gao et al. [25], 3D Helping Hands by Tecchia et al. [58], RemoteFusion by Adcock et al. [1], and BeThere by Sodhi et al. [52] are some examples that use 3D reconstructions. By retaining the characteristics of the real world, this technique passively provides both the awareness of task objects and the spatial information required to support referencing – especially when experienced through an immersive head-mounted display (HMD) like a VR headset.

While immersive 3D reconstructions should theoretically be similar to being physically co-present, that is not easy to achieve in practice. First, they are computationally expensive, difficult to update in real-time, and require high network bandwidth and extensive setup with specialized hardware [33, 42, 59, 61]; the quality of these reconstructions are also typically low and/or cover a small area to manage latency, reducing its overall effectiveness [40, 60]. Second, current MR HMDs have limitations in resolution, color depth, and field of view that makes the experience different from being physically co-present [5]. These limitations and the lack of interaction standards for MR also makes designing immersive collaborative interfaces challenging and non-trivial [34].

For these reasons and the fact that the helper is often stationary, the majority of current collaborative MR systems opt for a 2D setup for the helper and provide live video streams on desktop or mobile displays [17]. These are often a *scene view* that shows a stable task environment, or a *First Person View (FPV)* from a camera on the worker's HMD. While video feeds support referencing (including the use of deictic references) and enhance communication efficiency [24, 38, 47], they typically only provide a partial view of the task environment. Objects are often outside the camera's view, making it difficult to build awareness and spatial knowledge [23, 24, 27, 30, 38]. Scene views provide a good overview, but lack enough detail to identify those objects [23]; and FPVs are locked to the worker's view, tend to be very jittery, and can make for a very disorienting experience for the helper [30, 57].

To overcome these limitations, some systems use multiple camera feeds [48, 49], but switching between them can be jarring [27], and remote viewers often do not have a good understanding of the layout of the physical space [24]. More recently, researchers have explored 360°-videos as an alternative to overcome the drawbacks of video feeds without the complexity of 3D reconstructions [40, 45, 53]. While these systems overcome the field-of-view limitations, they lack detail (like the scene views) and are still a 2D presentation with limited depth perception [61].

In general, video feeds and 3D reconstructions only provide partial and incomplete support for helpers to understand the presence and spatial layout of the objects in the worker's environment. This is especially true in dense spaces where objects are difficult to make out or occluded from view [2].

Explicit Support for Referencing – Researchers have also explored a plethora of features explicitly designed to aid referencing in both immersive and non-immersive settings. They commonly use visual cues to guide the worker's attention to a referenced object. These include capturing and showing the helper's hand gestures [30, 36, 39, 52, 54, 63, 65], sharing the orientation of the helper's head or eye gaze [64, 71], telepointers [4, 35], virtual arrows [9, 26], and digital annotations [1, 8, 43, 56, 62]. While these techniques increase the efficiency and ease with which helpers can refer to objects, they mimic those used in face-to-face collaborations and are fundamentally proxies to deictic pointing. By extension, to use these features, helpers need awareness and spatial information about the task objects [57, 70] and typically rely on their view of the task environment for this information. For example, helpers using a virtual arrow have to position it on the view of the task environment using a graphical interface - this requires them to know where the object is located within the environment.

Overall, most MR systems for remote guidance do not explicitly support the information needs for referencing. While few systems provide helpers with information about what task objects are present in the worker's environment through virtual copies of the objects [50, 55, 57, 63], there is a lack of systems that provide spatial information of task objects in a manner that is not tied to the view of the task environment. In this paper, we explore how an external representation of the spatial relations between the task objects and the worker impacts the helper's ability to provide guidance. We focus on a 2D representation as it is compatible with both immersive and non-immersive helper setups.

2.3 Offloading the Referential Process

As Clark and Brennan discussed [12], the specific medium used for collaboration systems can affect communication, as what is possible with one technology might not be possible with another. While supporting communication techniques present in face-to-face settings is typically regarded as gold standard for collaborative systems, exploiting the unique affordances of a medium like MR could help overcome some of the shortcomings of these techniques, potentially improving upon the classic face-to-face paradigm [29].

Specifically, referencing techniques like deictic gestures and expressions used in face-to-face collaborations suffer from referential ambiguity and difficulty referencing occluded or hidden objects. For example, when a remote helper points to a set of objects, there can be ambiguity about which specific object is being referred to, or the object might not be visible to either the helper or the worker. By extension, features that mimic deixis (like pointers and annotations) suffer from the same limitations, often exacerbated by other limitations of the technology (like limited field of view). Addressing the resulting misunderstandings typically requires extensive verbal interaction [69, 70].

One of MR's affordance is that it is a spatial medium inherently aware of its environment [5, 16]. Researchers have previously capitalized on this affordance to reduce referential ambiguity by having the system do something that cannot be done in unmediated collocated settings: draw the worker's attention to *exactly* which object is being referred to. AlphaRead [8] allows helpers to specify or label objects using the video feed of the worker's environment, and Wang et al. [66] proposed a system where task objects are pre-labeled; these objects are then automatically highlighted for the worker when the helper refers to them. The TAC system [7] employs a similar method by storing information about the task space in the system; the helper only needs to click on a part on their video feed to initiate either an arrow or an outline to guide the worker's attention to it. However, AlphaRead requires initial input from the helper, and the TAC system does not support referencing if task objects are moved. All three systems also require the task objects to be unoccluded and within the helper's field of view for them to make a reference.

Tracking objects using an external system like proposed by Wang et al. [66] allows for wider task object awareness without being limited by the helper's field of view [16, 56] and reduces the computational demand on the MR device [16, 32]. Spatial coordinates of physical objects and people can be retrieved through various techniques like local sensors, indoor GPS, tags, markers, computer vision, RFID tags, etc. These tracking systems are also becoming increasingly pervasive, low cost, and accurate. For example, Tait and

Billinghurst [57] use the ART system[1], Wang et al. [66] use a stereo camera and markers, and ARTEMIS [67] uses the OptiTrack motion capture system[2] to track the positions of different task objects as well as the collaborators.

In face-to-face communication and traditional collaborative systems, the responsibility of referring to objects in a manner that ensures that the worker can identify them falls on the helper. However, systems that use live tracking technologies to automatically guide workers to referenced objects potentially completely *offloads* the responsibility of the second sub-process of referencing (understanding the reference) from the helper to itself.

While this could make it substantially easier for the helper to provide guidance, collaboration is a complex process – changing the role helpers play in the referential process could introduce significant cognitive seams [31] or discontinuities in the experience. Therefore, in this paper, we aim to understand the effects of (partially) offloading the referential process on remote guidance for physical tasks using system guidance to referenced objects.

3 RESEARCH APPROACH

This paper explores the two opportunities to enhance referencing in MR-mediated physical tasks: explicitly providing the spatial information necessary for helpers to direct workers to task objects, and offloading this responsibility from the helper to the system itself using automatic visual guidance to direct workers to referenced objects. However, the helper's need for spatial information is primarily required to direct workers to the referenced object - typically using strategies like deixis that reduce the collaborative effort required for the worker to understand the reference and identify the task object [12, 13]. Given this, *do collaborative MR systems that provide system-generated guidance for workers still need to explicitly provide helpers with spatial information*?

Our goal in this paper is to evaluate this question in addition to exploring the opportunities to enhance referencing in collaborative MR systems. To do so, we built a prototype *2D map interface* that explicitly represents the spatial relationships of the task objects and the worker in the task environment. We also implemented *MR system guidance* to dynamically direct the worker to referenced objects as a way of offloading part of the referential process. To understand the effects of both features independently, as well as study the interplay between them, we conducted a 2x2 mixed factorial experiment. Figure 1 outlines our experimental approach and answers the following three research questions:

- **RQ1**: In collaborative mixed reality systems designed for remote guidance, does a 2D representation of the spatial relationships within the task environment provide the helper with enough information to support referencing?
- **RQ2**: Does a collaborative mixed reality system that directly implements visual guidance to objects referenced during a physical task enhance task performance and reduce the overall effort required for remote guidance?
- **RQ3**: How does the presence of system generated visual cues to referenced objects impact the helper's need for spatial knowledge when providing remote guidance for a physical task?

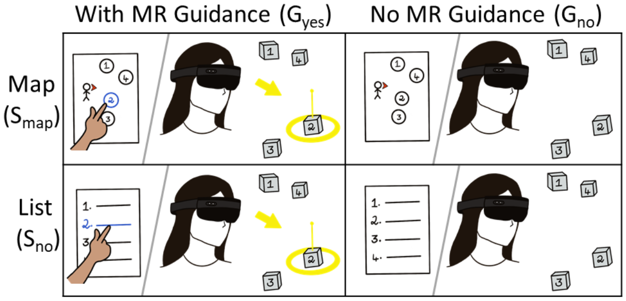


Figure 1: 2x2 mixed factorial study design. Our experiment studies the effect of Mixed Reality guidance for workers (G_{yes} vs. G_{no}) and the use of two different helper aids: a Map (S_{map}) or a List (S_{list}).

The overall goal is to better understand how to streamline support for referencing during physical tasks in MR environments. This knowledge is essential to design future collaborative systems with more effective representations, especially when deploying complex immersive MR systems for both helpers and workers is not an option.

4 PROTOTYPE DESIGN AND IMPLEMENTATION

To answer our questions, we implemented a prototype collaborative MR system that accommodates the different features required to run the experiment presented. When using our prototype system, the workers (see Fig. 2b) wear a mixed reality headset (the Microsoft HoloLens v1)[3] running a custom-built application that allows them to receive *system guidance* if enabled. The helpers (see Fig. 2a) use a stationary workstation with a Windows desktop that shows a live feed of the *worker's first-person view*; helpers can also view a *map* or *list* interface on a tablet device (map and list are different conditions in the study) that provides additional information on the objects in the worker's environment. The *map* and *list* interface also allow the helper to initiate *system guidance* if enabled.

4.1 Worker's First-Person View

The prototype system provides the helper with a view of the task environment through the worker's First Person View (FPV) shown on the desktop screen. In doing so, it passively supports referencing and provides the helper with general task awareness [19]. This was implemented using a low-latency live feed of the view from the HoloLens' front camera[4] and Microsoft's Windows' support for Miracast[5] on both Windows Desktop computers and the HoloLens. The video feed was transmitted over the local area network and presented almost unnoticeable latency (180-210ms).

[1] https://ar-tracking.com
[2] https://optitrack.com

[3] https://docs.microsoft.com/en-us/hololens/hololens1-hardware
[4] https://docs.microsoft.com/en-us/windows/mixed-reality/locatable-camera
[5] https://support.microsoft.com/en-us/help/15053/windows-8-project-wireless-screen-miracast

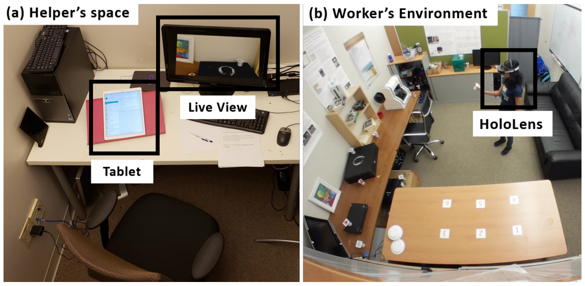


Figure 2: Helper's space and worker's environment. (a) The helper has a live first-person view of the worker throughout the experiment, and a specific *aid* (map or list) on a tablet. (b) The worker wears a HoloLens throughout the experiment, regardless of whether guidance is present or not.

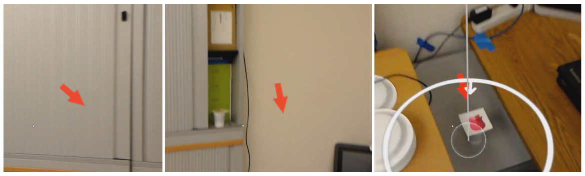


Figure 3: First-person view of the worker as they are guided to an object by the MR guidance

4.2 A Map of Task Objects

In addition to the passive support provided by the FPV, we needed to explicitly support the helpers' need for task object awareness and spatial information. We do so with a *map interface* that provides the helper with a 2D representation of the spatial relations of task objects in the worker's environment. As shown in Fig. 4a, the map interface resembles a top-view of the workspace and indicates the live positions of the task objects. It also represents the live position of the worker with respect to these objects. The task objects are represented by icons and allow the helper to know *what* objects are in the task environment, as well as *where* they are with respect to each other and the worker.

The map interface is implemented as a simple web application that can be viewed through a browser on any device. In our experiment, the helper viewed the map on an Apple iPad (See tablet in Fig. 2a). To show the location of both the HoloLens and the objects in the worker's space, the web application connects through a web server to a custom Windows application running on the HoloLens that relies on HoloLens's inside-out tracking system to locate objects[6].

While the inside-out tracking updates the HoloLens's position in space over time, it does not update the location of the task objects. To get around this constraint, we set up the initial locations for the task objects through our HoloLens app and created a *Wizard-of-Oz* interface [15, 20] that could remotely update the locations of these objects on the map. The interface resembled the map interface. During the experiment, our wizard updated the position of an object when it was moved by clicking on its icon and then clicking on the map at the new position in the worker's environment.

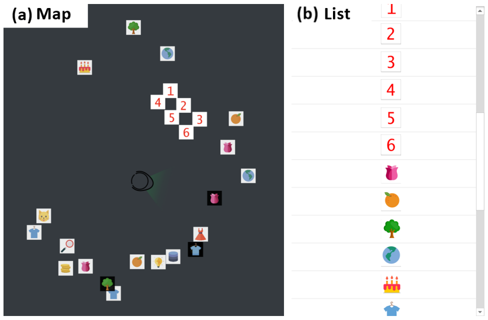


Figure 4: Helper's aids. (a) *Map* showing the task objects and the worker's position. (b) *List* showing all the task objects.

4.3 A List of Task Objects

The map interface explicitly provides the helper with information about (1) *what* objects are present within the task environment, and (2) *where* these objects are. However, to study the effects of the spatial information provided by the map, we needed an alternative (control) interface that only provided information about what objects were available in the task environment but did not show their positions. With this in mind, we developed a simple *list* interface (Fig. 4b) that shows the objects in the worker's environment as a randomly ordered list. The helper was able to scroll through the list if it did not fit on a single page.

4.4 MR System Guidance through the HoloLens

To study the effect of partially offloading the referential process to the MR device, we implemented the *MR system guidance* feature. When enabled, the helper could click on the icons representing a task object on either the map or the list interface to guide the worker towards it (during the experiment, helpers only had access to one of the two interfaces at a time). Once clicked, the system (HoloLens app) dynamically directed the worker's attention to the referenced object.

While there are many possible visual cues to guide users to a particular location in space [6, 34], our paper does not focus on the efficacy of these methods. However, given that the effects of MR guidance cannot be separated from the influence of the visual cue itself, we needed to implement one that would not add additional confounding factors to the study. We chose to adapt the navigation technique presented in HoloCPR [34] since was specifically designed for the limited field of view of the HoloLens and proved to be intuitive. The cue included an arrow to provide guidance on which direction to look, and a circle with a vertical element to help the worker focus on the exact object (see Fig. 3).

5 EXPERIMENT

Our experiment followed a 2x2 mixed factorial design, with a between-subjects factor focused on the *presence of MR system guidance* on two levels: G_{yes} = MR guidance present, and G_{no} = no MR guidance present. The within-subjects factor was the *presence of spatial information* on two levels: S_{map} = map interface (spatial

[6] https://docs.microsoft.com/en-us/windows/mixed-reality/enthusiast-guide/tracking-system

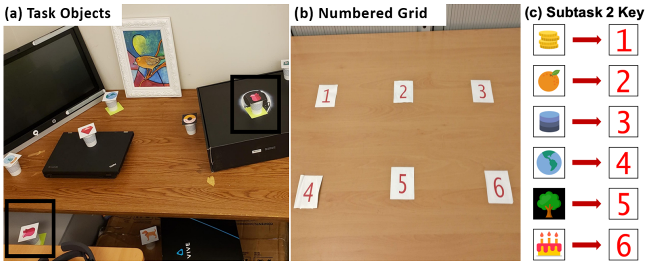


Figure 5: Snapshots of the worker's environment. (a) Some of the 19 task objects. The two highlighted objects are examples of objects used in subtask 1 (pair matching). (b) Numbered grid and (c) solution key used in subtask 2.

information present), and S_{list} = list interface (no spatial information present). We chose the presence of MR guidance to be a between-subjects factor since its carryover effects were unclear. However, the change in task objects and their positions between tasks ensured minimal carryover effects for the presence of spatial information when counterbalanced, allowing us to make it a within-subjects factor. The view of the task environment (provided by the FPV) and the awareness of the task objects available (provided by either the map or the list interface) were kept constant and represented our independent variables. Fig. 1 depicts our study design.

5.1 Participants and Procedure

Forty participants (17 females, 22 males, 1 preferred not to say) aged 18 to 43 (M = 23.93, SD = 5.17), were recruited to form 20 worker-helper dyads. 8 of these dyads knew each other beforehand. Most participants were students, with the majority majoring in Computer Science and Engineering (23) and others majoring in Electrical Engineering, Cognitive Science, Linguistics, Speculative design, Economics, Chemistry, and Bio-engineering. Despite a majority of computer science students, only 30% (12 people, 6 selected to be workers) had previous experience with Microsoft HoloLens.

Each dyad was required to complete two tasks and were randomly assigned to either receive MR guidance (G_{yes}) or not (G_{no}). All dyads were given the map (S_{map}) and the list (S_{list}), one at a time, over the two tasks (the order was counterbalanced). Helper and worker roles remained constant between the two tasks.

Once the participants were welcomed and introduced to the study, they were asked to fill a simple demographics questionnaire. They then underwent a short 5-minute training session so the worker could get accustomed to the HoloLens, and the helper could get accustomed to the video feed with the first person view (FPV). They were not introduced to either the map or the list interface in this session.

After the training session, the helper was taken to the helper space to be introduced to the task at hand and was provided an instruction sheet. The helper also received an *aid* which was either the map or the list interface on the tablet. The worker, who at this point, was unaware of the task was taken to the worker space and asked to wear the HoloLens. At the end of the first task, both the helper and the worker were asked to fill out a short post-task questionnaire and the NASA-TLX questionnaire [28]. This procedure was repeated for the second task with the respective conditions. The entire experiment was also video recorded. Participants consented to be in the study as per our university's human research protection protocol and received a $5 gift card.

5.2 Experiment Task and Environment

The helper and worker spaces were physically separated (see Fig. 2), but they were adjacent so that the helper and the worker could talk to each other. In the remainder of this section, we will refer to the worker's local space as the *task environment*. This task environment was a room that had a table with a numbered grid (see Fig. 5b) and 19 task objects (see Fig. 5a) that were spatially distributed. Each object had a picture on it with either a white or black background.

To complete each subtask, workers needed to be guided by the helper to accomplish a certain goal. The helper could use the *aid* (map or list) provided to them on the tablet as well as a live first-person view of the worker (see Fig. 2a). The worker wore a HoloLens at all times regardless of the study condition so that the helper could see their view (see Fig. 2b).

The entire task was designed to take less than 5 minutes and each dyad performed the task twice. However, the task objects and their positions in the task space were changed before the second run. There was no overlap in the set of task objects between the two runs of the task for the experiment. The task consisted of two subtasks that needed to be completed in order (sequentially):

(1) **Subtask 1 - Pair Matching:** In the first subtask, collaborators had to find pairs (objects with the same picture on them) among the objects in the task space and put them together. Each pair consisted of task objects with the same picture on them but one had a black background and one had a white background. They were instructed to move the object with the white background and place it adjacent to the object with the black background. Of all the objects in the task space, only 4 pairs could be made.
(2) **Subtask 2 - Object Gathering** The second task made use of the grid placed in the task space. Helpers were given a solution key (see Fig. 5c) that they used to help the worker find certain objects in the task space and place them on their respective positions on the grid.

While both subtasks required workers to identify particular objects in the environment and take actions on them, the actions were kept simple to allow for our measures to primarily reflect the effects of the prototype features on the referential process. The subtasks were also designed to mimic referencing in real-world tasks – Subtask 1 is similar to tasks like engine room maintenance where referrents (tools and area of interest) are spread out over a wide area, while subtask 2 is similar to car repair or surgery where the tools are laid out in the task environment but the area of interest (engine part, surgical site) is in a small area like the grid.

5.3 Experiment Measures and Analysis

We recorded the worker's FPV, and a view of the task environment using an external camera. The conversation between the participants during the experiment was also transcribed. We measured task performance (completion time) and analyzed communication behaviors (efficiency of communication and helper's referencing behaviors) by coding the recorded videos and conversation transcripts using BORIS [22]. Two researchers independently coded the data (double coded 20% of it, Cohen's Kappa = 0.886), and discussed and resolved any inconsistencies as soon as they arose. We also measured the *task load* on both the helper and the worker for each subtask they performed using the NASA-TLX questionnaire [28]. The post-task questionnaire measured the participants' perceived success of the partnership and the importance of the specific aid (map or list) through Likert-scale questions. Participants also provided free-form feedback on the technology they used during each task in the post-task questionnaire.

5.4 Limitations of our study

A few aspects of our setup are important to describe to understand the possible limitations of our results. First, being a 2D interface representing a 3D space, the map showed overlapping object icons in areas where the density of task objects was high. While this did not limit our participants (they could zoom in and click), this representation of task objects is fundamentally limited by the screen's real estate and 2D factor. Second, the lack of 3D spatial information meant that helpers sometimes described objects as next to each other when in reality they were at different heights (like on multiple levels of a shelf). However, workers adapted to it by looking up and down when searching for objects. Third, information about the task objects (map or list interface) and the FPV were on separate screens – this could have made it harder to pay attention to both at once. In general, these limitations could make the insights in this paper less generalizable for 3D representations of spatial information within more immersive environments.

Moreover, we did not observe any differences in communication behaviors because of prior familiarity among dyads - we believe this could be because of the simplicity of our task and might not reflect the effect of familiarity among collaborators for all tasks.

6 RESULTS

The quantitative data collected in our experiment was analyzed using two-way mixed-model ANOVA. All our measures were normally distributed across each cell of the design except for the data from the post-task questionnaire where the nonparametric Aligned Rank Transform [68] was applied which enables the use of ANOVA after alignment and ranking. Our analysis uses an alpha of 0.05, and presents both the main effects and the interaction effects. A main effect is the statistical relationship (effect) between one factor and a measure, averaging across the levels of the other factor. In our experiment, the main effects would tell us how spatial knowledge and guidance independently affected the collaboration. An interaction effect is present when the effect of one factor on a measure depends on the level of the other factor. For our purposes, we see an interaction effect if the presence of guidance changes the effect of spatial knowledge and vice-versa.

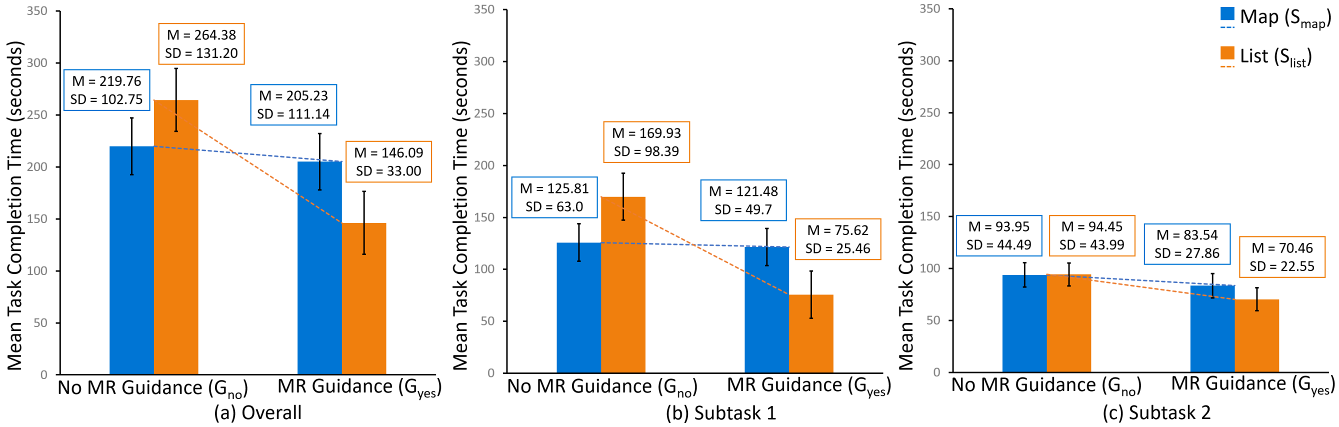
In the remainder of this section, F represents the F-value, p the p-value, η_p^2 the effect size, M the mean, and SD the standard deviation. The error bars in Fig. 6, 7, 9, and 10 depict standard errors and we also include trend lines to make it easier to read interaction effects.

6.1 Task Performance

We calculated the total time it took participants to complete the full task as well as the two subtasks. We consider a lower task completion time to indicate higher performance. Fig. 6 shows results of task performance measurements.

Total Task Time – Participants were significantly faster in finding referenced objects with MR guidance (F(1,18) = 5.162, p = 0.036, η_p^2 = 0.223). The presence of spatial information with the map did not substantially alter task performance and there was no main effect observed (F(1,18) = 0.064, p = 0.803, η_p^2 = 0.004). Fig. 6a shows a crossover interaction where the effects of spatial information seem to be different depending on whether participants had MR guidance or not, but this difference was not statistically significant (F(1,18) = 3.353, p = 0.084, η_p^2 = 0.157).

Subtask 1 – In the first subtask, participants completed the task significantly faster in the presence of MR guidance (F(1,18) = 4.914, p = 0.04, η_p^2 = 0.214). No main effect was observed for spatial knowledge (F(1,18) = 0.002, p = 0.963, η_p^2 = 0). Participants with the list



(a) Overall (b) Subtask 1 (c) Subtask 2

Figure 6: Average completion time across participants and conditions. (a) Total completion time across both subtasks. (b) Completion time for Subtask 1. (c) Completion time for Subtask 2.

interface were faster in the presence of guidance, while those with the map interface were faster without guidance (see Fig. 6b). Overall, they were fastest when using the list with MR system guidance and the observed crossover interaction was statistically significant (F(1,18) = 5.896, p = 0.026, η_p^2 = 0.247).

Subtask 2 – There was no significant main effect observed for spatial knowledge (F(1,18) = 0.291, p = 0.596, η_p^2 = 0.016) or for system guidance (F(1,18) = 2.381, p = 0.140, η_p^2 = 0.117) in the second subtask. We also did not observe any interaction effect (F(1,18) = 0.340, p = 0.567, η_p^2 = 0.019).

6.2 Efficiency of Communication

We measured communication efficiency by the number of words used to communicate during a task [24]. The fewer words spoken by the speaker(s), the more efficient the collaboration is considered.

Without MR guidance, participants used an average of 636.7 words (SD = 257.34) per task with a map and 555.4 words (SD = 242.73) with a list. The presence of MR guidance significantly reduced this to 405.4 words (SD = 210.55) with the map interface and 319.1 words (SD = 154.74) with the list (F(1,18) = 9.950, p = 0.005, η_p^2 = 0.356). The presence of spatial information did not significantly change the communication efficiency (F(1,18) = 1.681, p = 0.211, η_p^2 = 0.085) and there was no observed interaction effect.

Analyzing just the words spoken by the helper shows similar results (Fig. 7, Top), with the presence of guidance significantly reducing the number of words spoken, indicating a higher communication efficiency (F(1,18) = 8.623, p = 0.009, η_p^2 = 0.324). Efficiency decreased when helpers used the map, but this difference was not significant (F(1,18) = 3.371, p = 0.083, η_p^2 = 0.003). When it came to workers (Fig. 7, Bottom), there was no observable main effect for both guidance and spatial information. No interaction effects were observed across any speaker combinations.

6.3 Helper's Referencing Behaviors

We also analyzed how helpers made references to the task objects or locations in the workers' task space. We first discuss their referencing patterns as a whole, and then look into references to task object and spatial references separately for further analysis. Figure 8 outlines results of our analysis of referencing behavior.

In general, helpers made significantly more references when they used the map interface (F(1,18) = 17.674, p = 0.001, η_p^2 = 0.495). They also made significantly fewer references in the presence of system guidance (F(1,18) = 5.646, p = 0.029, η_p^2 = 0.239). There was no interaction effect.

Referring to Task Objects – When helpers verbally referred to the objects in the task environment, they either used verbal descriptions of the object like "the pink flower with the white background", or phrases that used deictic pronouns like "*that* one *there*". Deictic pronouns not used to refer to objects (like "*that*'s cool") were not considered in this analysis, nor were anaphoric uses of these pronouns (like "a globe or something like *that*").

Overall, helpers made significantly more references to objects in the worker's environment when they used the map than when they used the list interface (F(1,18) = 6.001, p = 0.025, η_p^2 = 0.25). The interface they used did not vary the percentage of object references that were deictic pronouns (F(1,18) = 0.79, p = 0.782, η_p^2 = 0.004).

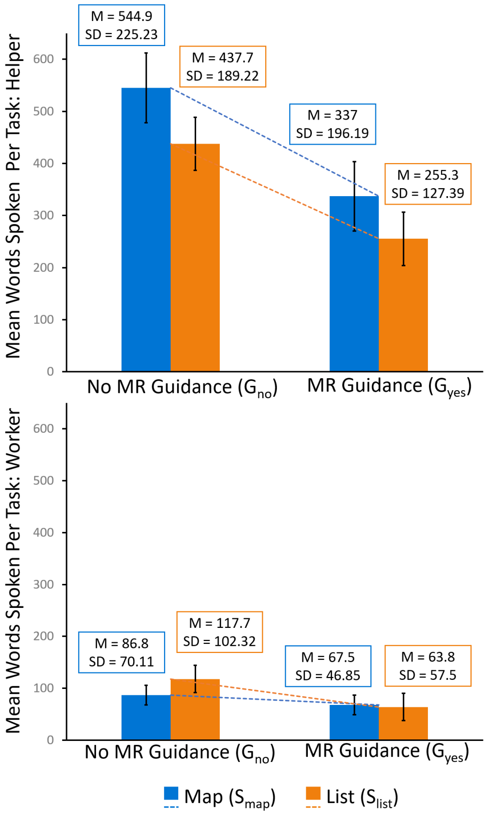
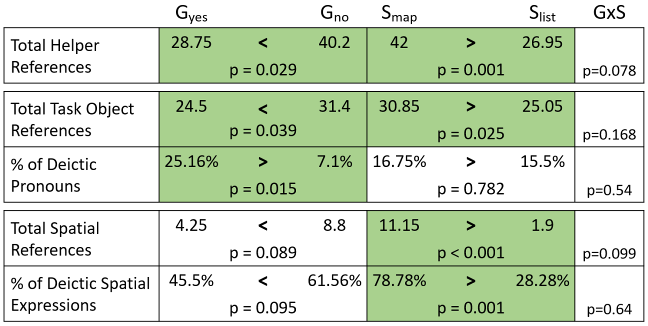


Figure 7: Top: Mean number of words per task by helper; Bottom: Mean number of words per task by worker.



	G_{yes}		G_{no}	S_{map}		S_{list}	GxS
Total Helper References	28.75	< p = 0.029	40.2	42	> p = 0.001	26.95	p=0.078
Total Task Object References	24.5	< p = 0.039	31.4	30.85	> p = 0.025	25.05	p=0.168
% of Deictic Pronouns	25.16%	> p = 0.015	7.1%	16.75%	> p = 0.782	15.5%	p=0.54
Total Spatial References	4.25	< p = 0.089	8.8	11.15	> p < 0.001	1.9	p=0.099
% of Deictic Spatial Expressions	45.5%	< p = 0.095	61.56%	78.78%	> p = 0.001	28.28%	p=0.64

Figure 8: Observed referencing behavior across presence/absence of MR guidance (G_{yes} vs. G_{no}) and the different aids (S_{map} vs. S_{list}) with significant effects in green. The rightmost column shows interaction effects.

Helpers made significantly fewer verbal references to objects in the presence of system guidance (F(1,18) = 4.946, p = 0.039, η_p^2 = 0.216). However, the percentage of these references that used deictic pronouns were significantly greater with MR guidance (F(1,18) = 7.249, p = 0.015, η_p^2 = 0.287). No interaction effect was observed.

Spatial References – When providing remote guidance to workers, helpers often verbally directed them towards objects of interest. When they did so, they either directly referenced areas or locations in the task environment (like "it should be on the *shelf*"), or used phrases with deictic spatial expressions like ("it's to *your right*").

Helpers made more spatial references when they used the map interface, and this difference was highly statistically significant (F(1,18) = 23.137, p < 0.001, η_p^2 = 0.562). The presence of spatial information from the map also caused a significant portion of these references to be made using different deictic spatial expressions (F(1,18) = 16.264, p = 0.001, η_p^2 = 0.475). The presence of system guidance did not significantly affect both how many spatial references helpers made F(1,18) = 3.231, p = 0.089, η_p^2 = 0.152), and what percentage of these references used deixis (F(1,18) = 3.105, p = 0.095, η_p^2 = 0.147). There were also no interaction effects.

6.4 Collaborator Task Load

We measured the task load on both the helper and the worker for each subtask they performed using the NASA TLX questionnaire [28] (see Fig. 9). The task load on the helper during the task significantly reduced both in the presence of system guidance (F(1,18) = 6.024, p = 0.025, η_p^2 = 0.251) as well as when they used the map interface (F(1,18) = 4.230, p = 0.05, η_p^2 = 0.19). The worker's task load was slightly higher when the helper used the map interface (F(1,18) = 0.129, p = 0.724, η_p^2 = 0.007), and lower in the presence of system guidance (F(1,18) = 1.296, p = 0.270, η_p^2 = 0.067) but neither difference was significant. There was no significant interaction effect observed for either the helper or the worker.

6.5 Participant Perceptions

After each task we measured participants' perceived success of the partnership and the role of the helper's aids (Map or List) using targeted 5-point Likert scale questions. Reliability of our questionnaire was measured using Cronbach's alpha [14], and scored high with α=0.913.

Working Together – Helper's ratings of the perceived success of their partnership with the worker was evaluated using the Likert-scale question "My partner and I worked well together on this task" (Q1). Responses were rated on a 5-point scale from 1 (= Strongly disagree) to 5 (= Strongly agree). Ratings were consistently high *(M = 4.75, SD = 0.543)* across all conditions and there were no significant main or interaction effects.

It is, however, worth noting that helpers perceived the partnership to me more successful when they used a map *(S_{map}: M = 4.8, SD = 0.41; S_{list}: M = 4.7, SD = 0.66)* (F(1,36) = 0.003, p = 0.953, η_p^2 = 0.00), and workers perceived it to be better when the helper used the list interface *(S_{map}: M = 4.65, SD = 0.93; S_{list}: M = 4.85, SD = 0.37)* (F(1,36) = 0.006, p = 0.938,0. η_p^2 = 0).

Perceived Importance of Spatial Information Aids – The role and usefulness of the specific Map or List aid for the helper was evaluated using the following two Likert-scale questions: "How important was the aid to help you **collaborate** effectively?" (Q2), and "How important was the aid in helping you **provide assistance** to your partner?" (Q3). Responses were rated on a 5-point scale from 1 (= Not at all important) to 5 (= Extremely important).

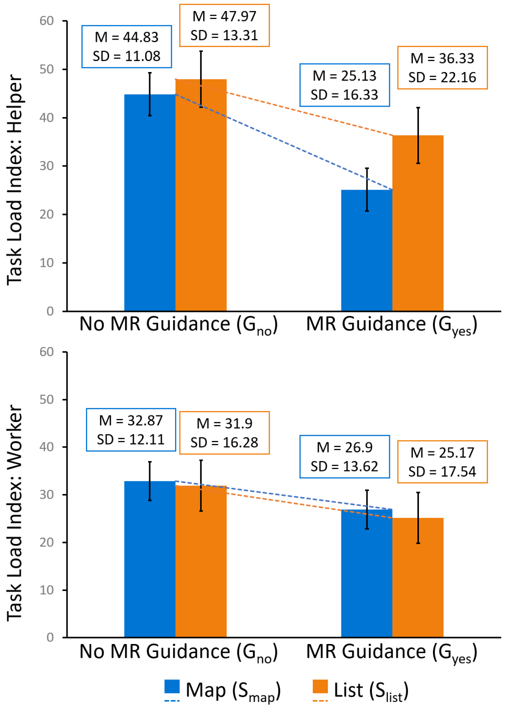


Figure 9: Results of the NASA TLX questionnaire. Top: Task load on helpers; Bottom: Task load on workers.

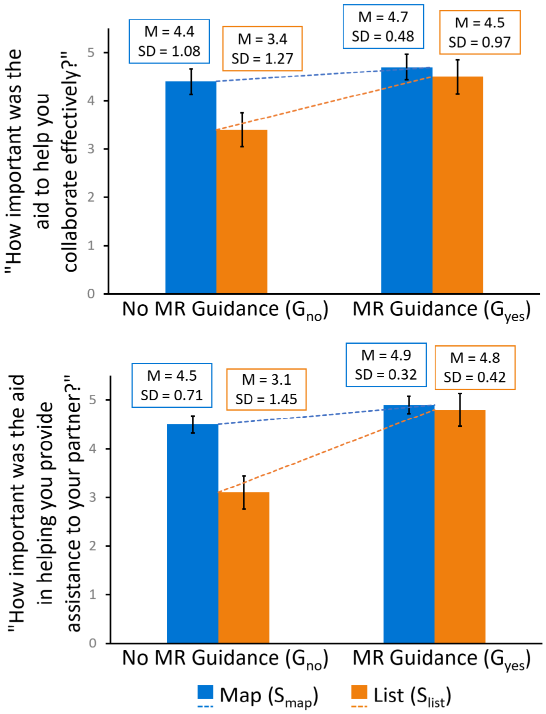


Figure 10: Perceived importance of the map and list aids. Top: Perceived importance on general collaboration; Bottom: Perceived importance on remote guidance.

Overall, helpers perceived the map interface to be significantly more important for general collaboration (Q2) than the list interface (F(1,36) = 6.629, p = 0.014, η_p^2 = 0.156). When asked specifically about how important the aid was for their ability to provide remote guidance (Q3), helpers still perceived the map interface as significantly more useful than the list interface (F(1,36) = 11.471, p = 0.02, η_p^2 = 0.242).

The perceived importance of both the map and the list interface increased in the presence of MR system guidance for both Q2 and Q3. This is reflected in the statistically significant main effects of guidance on both the perceived importance of the aids for general collaboration (F(1,36) = 7.609, p = .009, η_p^2 = 0.174), as well as for remote guidance (F(1,36) = 21.204, p < 0.001, η_p^2 = 0.371).

Interestingly, helpers rated the list interface to be substantially more important when they had MR system guidance than when they did not. In the presence of MR guidance, they perceived the list and the map interface as almost equally important (see Fig. 10). This shift in perception when MR guidance is used is reflected in the significant interaction effect for both general collaboration (F(1,36) = 5.192, p = 0.029, η_p^2 = 0.126) as well as remote guidance (F(1,36) = 8.169, p = 0.007, η_p^2 = 0.185).

7 UNDERSTANDING AFFORDANCES

In this section, we enrich and explain the quantitative results presented in the previous section with participant feedback from our questionnaires and researcher observations of their behavior during our experiment – observations emerged from a thematic analysis of the video codes. In doing so, we also cover the major observed effects and affordances of spatial information and MR system guidance. The quotes below refer to either the helpers [H] or the workers [W].

7.1 Effects of 2D Spatial Information (R1)

2D spatial information allows for effective guidance even when communication is less efficient – While the seemingly higher number of words spoken when helpers had a map could indicate a lower communication efficiency, participants' feedback confirmed that the effect was simply because helpers were able to provide the workers with more information and thus spoke more words. Without the map (and in the absence of MR guidance), workers mentioned that once the helpers told them what object to look for, they were *"not able to help so much with where the objects were, [and the workers] mostly had to look around [themselves]"* [W19]. Helpers echoed this sentiment saying that they felt they were of *"diminished help"* [H19] with the list and sometimes *"spent more time looking [at the FPV] of the HoloLens"* [H19] in an attempt to get a sense of the layout of the task environment. Being able to use the map to understand the orientation of objects relative to the worker meant that they could use that information to direct workers and *"provide more assistance than just looking around"* [H5].

Explicit spatial information reduces effort needed for spatial deixis – The helpers' referencing behaviors showed that the map consistently let them create more references, primarily with deictic spatial expressions. Being able to use spatial deixis is considered a cognitively expensive process as it requires extensive spatial awareness [37]. However, in our experiment, helpers had no trouble and commented that the *"easy part was [knowing] the layout of the room"* [H5]. Explicitly providing spatial information meant that they did not have to rely on a view of the task environment to create a mental map of the space. This reduced the effort required to use spatial deixis – something that is also supported by the significant drop in task load when helpers used the map.

Additional spatial information splits helpers' attention – Helpers used the map as an alternative source of task status; they followed the movement of objects and stated how *"following [the worker's] progress [on the map] was less distracting than on the video stream which had a lag and jitter too"* [H15]. They were also able to correct worker actions early as they could *"easily know whether [the worker] is going to the right/wrong direction"* [H18]. However, the extreme focus on the map lead helpers to not pay attention to the FPV, and they often continued providing directions even when it was obvious that the helper had found the object. This required more back and forth verbal interaction between the participants, and reduced performance and efficiency. The following interaction (H12-W12; G_{no}, S_{map}) showcases this behavior:

H: *"So go in the corner and then right next to a snow globe...grab the white Christmas tree"*
W: *"here here, I found it...See that?" [shows object to camera]*
H: *[looking at map] "Okay, right...Somewhere around there, there's a white Christmas tree."*
W: *"I already [have] it...here" [shows object to camera again]*

7.2 Effects of MR System Guidance (R2)

MR guidance supports effective referencing through deixis even in the absence of spatial information – Our participants found MR guidance most helpful because it allowed the helper to guide the worker without knowing their spatial position in comparison to the objects. The worker was able *"to locate the object and place it wherever [the helper] wanted [them], without communicating verbally as much"* [H2].

The efficiency of the conversation primarily stemmed from the helper's referential behaviors. First, the significantly large proportion of deictic pronouns used in the presence of MR guidance meant that most of the conversations revolving around referencing were often just a few words (e.g. [initiates MR guidance] + *"okay now go get **that**"* [H2]; [initiates MR guidance] + *"Place it next to **this** one"* [H8]). Second, we noticed that as the task progressed, some helpers shifted to not using any form of verbal referencing. They simply clicked on the icon of the object that they wanted to refer to, and let the system guide the helper. The following non-verbal exchange (H11-W11; G_{yes}, S_{list}) is an example:

H: [initiates MR guidance]
W: [puts down object at target location] + *"yeah"*
H: [initiates MR guidance]
W: [picks up object]
H: [initiates MR guidance]
W: [puts down object at target location]

Confidence in MR guidance results in less acknowledgements and increases parallelizations – The above exchange also highlights the lack of explicit acknowledgements – a common process that is typically used to confirm the understanding of the reference [10]. Helpers often checked for acknowledgements at the beginning of the task, but they soon realized that *"after [they] tapped the desired object...the arrows took care of the rest"* [H7]. They then shifted to simply checking the FPV for confirmation instead of waiting for a verbal one, and initiated MR guidance to the next object when the worker was done. Workers also quickly adapted, and even began to expect this interaction pattern: [W13] mentioned that when the helper *"took some time to move the arrow after the [step] was done [they were] confused if [they] had successfully completed the prior [step]"*. This confidence in the MR system also allowed for parallelization as some helpers trusted that *"the arrows were able to guide [the worker] perfectly, [and they] could focus on finding the next object"* [H8].

7.3 Combining Spatial Information and MR Guidance (R3)

MR guidance makes spatial information superfluous – Helpers with MR guidance found it easier to choose objects in a list: H7 mentioned: *"with the map, [I] had to look over a larger area to find matches and things. With the list, [I] could easily scroll down and everything was more organized"*. Having MR guidance meant that they did not use the position of the objects very much, and found the spatial information provided by the map to be a *"an annoyance since [they] had to find the object first in the location before pointing [the workers] towards it"* [H15]. Helpers felt that using the map was confusing as they often had a lot of information, and they tended to direct the workers too much. In their feedback, workers also mentioned similar reasons, saying they received too much information when helpers had the map, and they were forced to pause or divert their attention away from the MR guidance.

Spatial information enables priming for MR Guidance – Another referential behavior that we observed when helpers used both the map and the MR system guidance was that they often *primed* the worker for the information they would receive from the MR system. This could just be information about what they would find (e.g. [initiates MR guidance] + *"Now we are picking up a coin"* [H17]), or the helper would position the worker so that when they would initiate guidance, the object was roughly in the worker's FoV (e.g. *"Okay. pick it up and then do a 180 turn"* + [initiates MR guidance] + *"Okay, do you see that?"* [H7]). From our experiment, it is hard to say if it was indeed beneficial for the worker, but we did not see any negative impact.

MR guidance affects the perceived importance of the aids – Another place where we see an interaction effect across MR guidance and spatial information is in the perceived importance of the map or the list interface. While in the presence of MR guidance helpers reported that the list was almost as useful as the map interface, they still perceived the map as very important to be able to provide guidance throughout: *"using the map was extremely easy"* [H16].

The preference towards the map interface, even if it had limitations and diminished task performance, could be attributed to perceiving it as more sophisticated, and therefore more useful. Some helpers acknowledged it reporting how *"the map requires more effort than the list; the list doesn't make it fun, but is more convenient"* [H1]. The increased verbal conversation when helpers used the map could have also added to the reasons that helpers enjoyed using it: by providing more information, the map allowed the collaborators to accrue more common ground and therefore made them feel more engaged than just clicking on the list buttons.

Both MR guidance and spatial information help in unfamiliar environments – The crossover effect observed for performance in subtask 1, seemed to be absent in subtask 2. We believe that the primary reason behind it is that the collaborators were familiar with the task environment and the objects in it by the time they started the second subtask. Moving around the space in the first subtask made participants feel *"the second subtask was easier"*[W6]. Workers (and sometimes helpers) often *"remembered seeing [the referenced object] earlier, [and] was able to go straight to it"*[W19]. This suggests that the familiarity of the environment plays a role in the effectiveness of both spatial information and MR guidance.

8 IMPLICATIONS FOR DESIGNERS AND DEVELOPERS

This paper explored providing explicit information about task objects and partially offloading the referential process using MR guidance to help workers identify referenced objects. Here we discuss some of the implications that these insights might have for building collaborative MR systems optimized for referencing.

8.1 Offloading Referencing to MR to Increase Efficacy

Our study showed how offloading the referential process significantly increases both task performance and communication efficiency while reducing the task load on collaborators. In addition, providing explicit awareness of task objects to the helper and accurate visual cues to the worker through offloading also makes it easier to refer to occluded or hidden objects, combats referential ambiguity, and reduces the need for explicit acknowledgements. While these characteristics are beneficial for remote guidance in general, offloading could be especially impactful in scenarios that simultaneously value efficiency and accuracy like emergency repairs or surgery.

Offloading is also well-suited for remote guidance when there is a large skill-gap between the helper and the worker. By shifting the responsibility of helping the worker identify the referenced object to the system, collaborators no longer need to build extensive common ground to reference objects, and the conversation can primarily focus on the actions that need to be taken on those objects. This can be particularly useful in scenarios like customer support where experts are helping someone outside their domain with a specific task, such as home repairs, fixing a car or a bike, etc.

Decoupling the process of making a reference, and helping the worker understand the reference through offloading, also helps when the environment is new to the helper and/or the worker. Explicit representations and MR system guidance no longer require helpers to fully understand the space to refer to objects, making it an effective technique for remote guidance in unfamiliar environments. Maintenance scenarios where parts and tasks might be standard, but the spatial layout of each environment can be different – such as a factory or engine room maintenance – are examples of where this would be useful.

Finally, while passive support for referencing through features that help view the worker's environment and provide general awareness (like the FPV) worked well in tandem with MR guidance, explicitly representing spatial information of the task objects (like the map) seemed to cause an information overload for both helpers and workers. We expect that MR guidance would not affect the positive impacts of passive awareness support. However, explicitly providing spatial information in the presence of MR guidance is not necessary and can be detrimental. An exception to this recommendation might be when the helper's tacit knowledge is tied to the position of the task objects; for example, an expert mechanic might find locating a part of an engine easier if the representation of the parts resembled the spatial layout of an engine model.

8.2 Making Support for Referencing More Accessible

Creating immersive MR environments requires complex set up in the helper's and the worker's environment. While the benefits of the worker being in an immersive MR environment and receiving guidance in-situ are already well known [5], helpers are typically only in immersive settings to allow for features like 3D reconstructions that increase their sense of presence and emulate face-to-face communication.

In contrast, we showed that using remote MR guidance and offloading allows helpers to successfully provide guidance even in non-immersive environments. While providing MR guidance still requires additional technology – tracking systems that provide the positions of objects in an environment – these are typically simpler to implement than systems that reconstruct the worker's environment accurately. Tracking systems also have varying complexities and fidelity as they can be sensor-based, vision-based, or both [51], giving designers and developers more freedom to choose the technology that best suits the physical task they need to support.

There are still cases where remote collaborators need powerful systems with stereoscopic displays or HMDs – like physical tasks where immersive reconstructions are truly needed and/or tracking objects externally or through computer vision is not feasible (like surgery [11]). However, these requirements limit access to remote collaborators in very specific environments. We show that using simple representations (like a web-based list or map interface) successfully reduces the effort required to reference objects. This extends rich support for referencing to everyday devices – like phones, tablets, and desktops – making guidance more scalable and increasing the network of remote experts that workers can rely on.

9 CONCLUSION AND FUTURE WORK

In this paper, we presented a 2x2 factorial experiment aimed at understanding how providing explicit spatial information and partially offloading the referential process through MR system guidance enhances support for referencing in collaborative Mixed Reality. Our results show that 2D representations of spatial information provide ample support for remote helpers to provide assistance (RQ1). We also show that offloading the referential process reduces the effort required for communication (RQ2) while diminishing the role spatial information plays in the helper's ability to provide remote guidance for physical tasks (RQ3).

Until immersive experiences that rely on HMDs become easier to implement and more ubiquitous, we show that it is possible to design for collaborative MR systems that do not rely on expensive headsets for helpers, making remote guidance and expertise more accessible.

Building effective collaborative systems is a complex task and requires a deep understanding of how the unique affordances of technology can be maximized to advance the collaborative goal. While this paper focused on referencing from a helper's perspective, we believe future work should also study referential behavior on the worker's side. Understanding the affordances of specific visual cues for guidance and its appropriateness for specific tasks and environments will enable us to better support all types of collaborations and design MR systems that fully support the collaborative nature of referencing [13]. Referencing is also only one of the communication elements that are key to providing effective remote guidance. Helpers need ways to effectively detect *when* to provide help and communicate complex actions to be taken on objects. Optimizing for these aspects is also vital to building collaborative MR systems that support efficient remote guidance.

REFERENCES

[1] Matt Adcock, Stuart Anderson, and Bruce Thomas. 2013. RemoteFusion: real time depth camera fusion for remote collaboration on physical tasks. In *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*. ACM, 235–242.

[2] Ferran Argelaguet, Alexander Kulik, André Kunert, Carlos Andujar, and Bernd Froehlich. 2011. See-through techniques for referential awareness in collaborative virtual reality. *International Journal of Human-Computer Studies* 69, 6 (2011), 387–400.

[3] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376550

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

[5] Mark Billinghurst and Hirokazu Kato. 2002. Collaborative augmented reality. *Commun. ACM* 45, 7 (2002), 64–70.

[6] Frank Biocca, Arthur Tang, Charles Owen, and Fan Xiao. 2006. Attention funnel: omnidirectional 3D cursor for mobile augmented reality platforms. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1115–1122.

[7] Sébastien Bottecchia, Jean-Marc Cieutat, and Jean-Pierre Jessel. 2010. TAC: augmented reality system for collaborative tele-assistance in the field of maintenance through internet. In *Proceedings of the 1st Augmented Human International Conference*. 1–7.

[8] Yuan-Chia Chang, Hao-Chuan Wang, Hung-kuo Chu, Shung-Ying Lin, and Shuo-Ping Wang. 2017. AlphaRead: Support unambiguous referencing in remote collaboration with readable object annotation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2246–2259.

[9] Jeffrey W Chastine, Kristine Nagel, Ying Zhu, and Luca Yearsovich. 2007. Understanding the design space of referencing in collaborative augmented reality environments. In *Proceedings of graphics interface 2007*. ACM, 207–214.

[10] Jeffrey W Chastine, Ying Zhu, and Jon A Preston. 2006. A framework for inter-referential awareness in collaborative environments. In *2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing*. IEEE, 1–5.

[11] Long Chen, Thomas W Day, Wen Tang, and Nigel W John. 2017. Recent developments and future challenges in medical mixed reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 123–135.

[12] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[13] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[14] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[15] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200.

[16] Archi Dasgupta, Mark Manuel, Rifat Sabbir Mansur, Nabil Nowak, and Denis Gračanin. 2020. Towards Real Time Object Recognition For Context Awareness in Mixed Reality: A Machine Learning Approach. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 262–268.

[17] Ryan Anthony J de Belen, Huyen Nguyen, Daniel Filonik, Dennis Del Favero, and Tomasz Bednarz. 2019. A systematic review of the current state of collaborative mixed reality technologies: 2013–2018. (2019).

[18] Alan Dix. 1994. Computer supported cooperative work: A framework. In *Design issues in CSCW*. Springer, 9–26.

[19] Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces.. In *CSCW*, Vol. 92. 107–114.

[20] Steven Dow, Jaemin Lee, Christopher Oezbek, Blair MacIntyre, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz interfaces for mixed reality applications. In *CHI'05 extended abstracts on human factors in computing systems*. 1339–1342.

[21] MR Ensley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors* 37 (1995), 85–104.

[22] Olivier Friard and Marco Gamba. 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution* 7, 11 (2016), 1325–1330.

[23] Susan R Fussell, Robert E Kraut, and Jane Siegel. 2000. Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 21–30.

[24] Susan R Fussell, Leslie D Setlock, and Robert E Kraut. 2003. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 513–520.

[25] Lei Gao, Huidong Bai, Rob Lindeman, and Mark Billinghurst. 2017. Static local environment capturing and sharing for MR remote collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*. 1–6.

[26] Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias Höllerer. 2012. Integrating the physical environment into mobile remote collaboration. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. 241–250.

[27] William W Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is not enough: Multiple views in a media space. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. ACM, 335–341.

[28] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[29] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 119–125.

[30] Weidong Huang and Leila Alem. 2011. Supporting hand gestures in mobile remote collaboration: a usability evaluation. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*. British Computer Society, 211–216.

[31] Hiroshi Ishii, Minoru Kobayashi, and Kazuho Arita. 1994. Iterative design of seamless collaboration media. *Commun. ACM* 37, 8 (1994), 83–97.

[32] Dongsik Jo and Gerard Jounghyun Kim. 2016. ARIoT: scalable augmented reality framework for interacting with Internet of Things appliances everywhere. *IEEE Transactions on Consumer Electronics* 62, 3 (2016), 334–340.

[33] Michal Joachimczak, Juan Liu, and Hiroshi Ando. 2017. Real-time mixed-reality telepresence via 3D reconstruction with HoloLens and commodity depth sensors. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 514–515.

[34] Janet G Johnson, Danilo Gasques Rodrigues, Madhuri Gubbala, and Nadir Weibel. 2018. HoloCPR: Designing and Evaluating a Mixed Reality Interface for Time-Critical Emergencies. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, 67–76.

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

[36] David Kirk and Danae Stanton Fraser. 2006. Comparing remote gesture technologies for supporting collaborative physical tasks. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 1191–1200.

[37] Robert E Kraut, Susan R Fussell, and Jane Siegel. 2003. Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction* 18, 1-2 (2003), 13–49.

[38] Robert E Kraut, Mark D Miller, and Jane Siegel. 1996. Collaboration in performance of physical tasks: Effects on outcomes and communication. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. ACM, 57–66.

[39] Morgan Le Chénéchal, Thierry Duval, Valérie Gouranton, Jérôme Royan, and Bruno Arnaldi. 2016. Vishnu: virtual immersive support for HelpiNg users an interaction paradigm for collaborative remote guiding in mixed reality. In *2016 IEEE Third VR International Workshop on Collaborative Virtual Environments (3DCVE)*. IEEE, 9–12.

[40] Gun A Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A user study on mr remote collaboration using live 360 video. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 153–164.

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

[42] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 741–754.

[43] Jiazhi Ou, Susan R Fussell, Xilin Chen, Leslie D Setlock, and Jie Yang. 2003. Gestural communication over video stream: supporting multimodal interaction for remote collaborative physical tasks. In *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 242–249.

[44] Thomas Pechmann and Werner Deutsch. 1982. The development of verbal and nonverbal devices for reference. *Journal of experimental child psychology* 34, 2 (1982), 330–341.

[45] Thammathip Piumsomboon, Gun A Lee, Andrew Irlitti, Barrett Ens, Bruce H Thomas, and Mark Billinghurst. 2019. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–17.

[46] Thammathip Piumsomboon, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. CoVAR: a collaborative virtual and augmented reality system for remote collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies*. 1–2.

[47] Abhishek Ranjan, Jeremy P Birnholtz, and Ravin Balakrishnan. 2006. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 403–412.

[48] Troels A Rasmussen and Weidong Huang. 2019. SceneCam: Using AR to improve Multi-Camera Remote Collaboration. In *SIGGRAPH Asia 2019 XR*. 36–37.

[49] Mauro Del Rio, Vittorio Meloni, Francesca Frexia, Francesco Cabras, Roberto Tumbarello, Sabrina Montis, Andrea Marini, and Gianluigi Zanetti. 2018. Augmented Reality for Supporting Real Time Telementoring: an Exploratory Study Applied to Ultrasonography. In *Proceedings of the 2nd International Conference on Medical and Health Informatics*. 218–222.

[50] Edgar Rojas-Muñoz, Maria E Cabrera, Chengyuan Lin, Daniel Andersen, Voicu Popescu, Kathryn Anderson, Ben L Zarzaur, Brian Mullis, and Juan P Wachs. 2020. The System for Telementoring with Augmented Reality (STAR): A head-mounted display to improve surgical coaching and confidence in remote areas. *Surgery* (2020).

[51] Somaiieh Rokhsaritalemi, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi. 2020. A Review on Mixed Reality: Current Trends, Challenges and Prospects. *Applied Sciences* 10, 2 (2020), 636.

[52] Rajinder S Sodhi, Brett R Jones, David Forsyth, Brian P Bailey, and Giuliano Maciocci. 2013. BeThere: 3D mobile collaboration with spatial input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 179–188.

[53] Maximilian Speicher, Jingchen Cao, Ao Yu, Haihua Zhang, and Michael Nebeling. 2018. 360anywhere: Mobile ad-hoc collaboration in any environment using 360 video and augmented reality. *Proceedings of the ACM on Human-Computer Interaction* 2, EICS (2018), 1–20.

[54] Aaron Stafford, Wayne Piekarski, and Bruce H Thomas. 2006. 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*. IEEE, 165–172.

[55] Ohan Oda Carmine Elvezio Mengu Sukan, Steven Feiner, and Barbara Tversky. 2015. Virtual Replicas for Remote Assistance in Virtual and Augmented Reality. (2015).

[56] Zsolt Szalavári, Dieter Schmalstieg, Anton Fuhrmann, and Michael Gervautz. 1998. "Studierstube": An environment for collaboration in augmented reality. *Virtual Reality* 3, 1 (1998), 37–48.

[57] Matthew Tait and Mark Billinghurst. 2015. The effect of view independence in a collaborative AR system. *Computer Supported Cooperative Work (CSCW)* 24, 6 (2015), 563–589.

[58] Franco Tecchia, Leila Alem, and Weidong Huang. 2012. 3D helping hands: a gesture based MR system for remote collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. 323–328.

[59] Theophilus Teo, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. 360Drops: Mixed Reality Remote Collaboration using 360 Panoramas within the 3D Scene. In *SIGGRAPH Asia 2019 Emerging Technologies*. 1–2.

[60] Theophilus Teo, Ashkan F. Hayati, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. A Technique for Mixed Reality Remote Collaboration using 360 Panoramas in 3D Reconstructed Scenes. In *25th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[61] Theophilus Teo, Louise Lawrence, Gun A Lee, Mark Billinghurst, and Matt Adcock. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 201.

[62] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating remote instruction of physical tasks using bi-directional mixed-reality telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 161–174.

[63] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Sili Wei, Guangyao Xu, Weiping He, Xiangyu Zhang, and Jie Zhang. 2020. 3DGAM: using 3D gesture and CAD models for training on mixed reality remote collaboration. *Multimedia Tools and Applications* (2020), 1–26.

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

[65] Shiyao Wang, Michael Parsons, Jordan Stone-McLean, Peter Rogers, Sarah Boyd, Kristopher Hoover, Oscar Meruvia-Pastor, Minglun Gong, and Andrew Smith. 2017. Augmented reality as a telemedicine platform for remote procedural training. *Sensors* 17, 10 (2017), 2294.

[66] Tzu-Yang Wang, Yuji Sato, Mai Otsuki, Hideaki Kuzuoka, and Yusuke Suzuki. 2020. Developing an AR Remote Collaboration System with Semantic Virtual Labels and a 3D Pointer. In *International Conference on Human-Computer Interaction*. Springer, 407–417.

[67] Nadir Weibel, Danilo Gasques, Janet Johnson, Thomas Sharkey, Zhuoqun Robin Xu, Xinming Zhang, Enrique Zavala, Michael Yip, and Konrad Davis. 2020. ARTEMIS: Mixed-Reality Environment for Immersive Surgical Telementoring. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–4.

[68] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[69] Nelson Wong and Carl Gutwin. 2010. Where are you pointing? The accuracy of deictic pointing in CVEs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1029–1038.

[70] Nelson Wong and Carl Gutwin. 2014. Support for deictic pointing in CVEs: still fragmented after all these years'. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1377–1387.

[71] Jing Yang, Prasanth Sasikumar, Huidong Bai, Amit Barde, Gábor Sörös, and Mark Billinghurst. 2020. The effects of spatial auditory and visual cues on mixed reality remote collaboration. *Journal on Multimodal User Interfaces* (2020), 1–16.

[72] George Yule. 2016. *The study of language*. Cambridge university press. 128–140 pages.