More does not Mean Better: Effects of Spatial Snapshot Annotations on Remote Collaboration

Michael Anderson
michael.anderson@tugraz.at
Institute of Visual Computing
Graz University of Technology
Graz, Austria

Dieter Schmalstieg
dieter.schmalstieg@visus.uni-stuttgart.de
VISUS
TU Stuttgart
Stuttgart, Germany
Institute of Visual Computing
Graz University of Technology
Graz, Austria

Alexander Plopski
alexander.plopski@tugraz.at
Institute of Visual Computing
Graz University of Technology
Graz, Austria

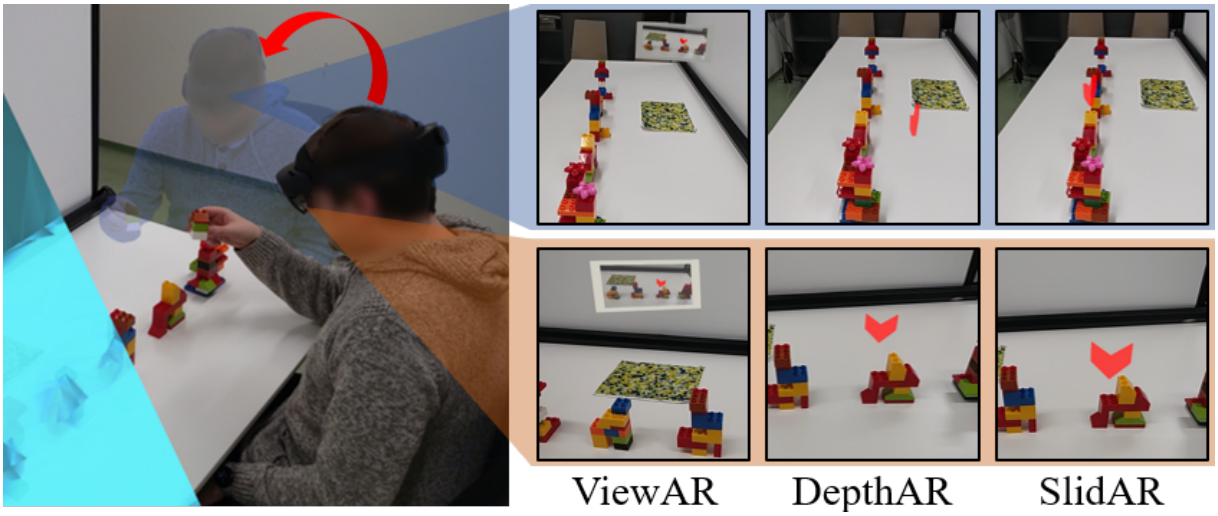


Figure 1: Objects that are difficult to reconstruct are often missing or incomplete in scene reconstructions (left). Snapshots captured from the worker's video feed from different positions (blue and brown) can facilitate instruction sharing (right). We investigate the effects of three spatial snapshot guidance techniques on remote collaboration in extended reality. First-person captures of the techniques as seen by a worker wearing an optical see-through head-mounted display: ViewAR presents a snapshot of the annotated view to the worker. DepthAR places annotations on the inaccurate scene reconstruction without any adjustments, thus resulting in seemingly accurate positioning from one viewpoint and incorrect positioning from another. SlidAR allows helpers to further adjust the position to the correct depth to account for scene reconstruction errors.

Abstract

Remote collaboration can greatly benefit from the improved reconstruction and tracking capabilities of optical see-through head-mounted displays. A remote helper can efficiently assist a local worker and convey complex instructions by placing 3D instructions into the environment. If a high quality reconstruction is not available, *spatial snapshot annotations* that utilize only 6DOF device tracking data present an opportunity to facilitate efficient communications. While different techniques have been explored for annotation creation from multiple spatial snapshots on handheld devices, they have not been evaluated in a remote collaboration scenario or on optical see-through head-mounted displays (OST-HMDs). In this paper, we present an empirical comparison of three annotation creation methods for spatial snapshots in a remote assembly support scenario for OST-HMDs. ViewAR presents only an annotated snapshot created by the remote helper; DepthAR places 3D annotations at the estimated surface of the environment, while SlidAR allows adjustments of the location at which the annotation is placed. Our study with 24 participants showed that all methods were well received. We found that participants used different approaches to complete the task to mitigate the shortcomings of the techniques. Our findings provide insights for the future design of remote collaboration systems based on spatial snapshots.

SUI '26, Bari, Italy

ACM ISBN 979-8-4007-2812-9/2026/10
https://doi.org/10.1145/3822518.3830061

CCS Concepts

• **Human-centered computing → Mixed / augmented reality**; **Collaborative interaction**; *Computer supported cooperative work*; User studies.

Keywords

Multi-View Interfaces, Collaboration, Head- Mounted Display, Extended Reality

ACM Reference Format:

Michael Anderson, Dieter Schmalstieg, and Alexander Plopski. 2026. More does not Mean Better: Effects of Spatial Snapshot Annotations on Remote Collaboration. In *ACM Symposium on Spatial User Interaction (SUI '26), October 10–11, 2026, Bari, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3822518.3830061

1 Introduction

Assistance provided by remote helpers is an important use case for telepresence systems. Conventional remote support relies on video streams with audio or 2D drawings, which can lead to miscommunication and slow task execution. Extended Reality (XR) addresses this by embedding 3D instructions directly into the worker's view, simplifying their association with the environment and leading to faster completion time and fewer errors [7, 58]. With the development of more advanced six-degree-of-freedom (6DOF) tracking and environment reconstruction capabilities, telepresence using an optical see-through head-mounted display (OST-HMD) has become a popular form of remote collaboration [37, 56].

A common approach is to show the reconstructed environment to the remote helper in virtual reality [12, 13, 32, 61, 66], allowing them to freely explore the scene. Others suggested sharing a representation of worker and helper as avatars to facilitate natural communication cues [39, 59]. However, these systems rely on high-quality sensors that can recover a detailed representation of the environment. This limits their application to easy to reconstruct static environments. Furthermore, many AR headsets (e.g., Snap Spectacles, Moverio BT45CS, Vuzix M4000) lack a depth sensor and thus provide only limited scene understanding, e.g., no tracking, only fiducial marker tracking, plane detection, or simultaneous localization and mapping (SLAM) using sparse features. Reconstructing the environment on the helper's computer requires high-end hardware that is often not easily available, e.g., many desktop workstations or notebook computers do not have a dedicated Graphics Processing Unit needed for such operations.

Alternative approaches use image based representations such as light fields [35] or neural radiance fields (NeRF) [46], or remap environments to a planar surface [30]. While these approaches allow helpers to explore the environment, they are time consuming to create and often cannot account for dynamic changes.

To mitigate these drawbacks, cameras distributed in the environment can provide helpers with multiple viewpoints of the scene [43, 44]. Such multi-view systems, while effective, are not suitable for spontaneous remote collaboration as they require a preprepared workspace. Instead of multiple cameras, the desired location can be triangulated from several *spatial snapshots*, images with a known pose, taken by a moving camera [14, 40]. Contrary to multi-view systems, spatial snapshots show the environment for a given point in time. Although this is an effective approach for placing 3D annotationsto place annotations in 3D on mobile devices [10, 41], it is unclear if it can be utilized efficiently in a remote guidance scenario for OST-HMDs and how helpers utilize it to guide others.

We consider three techniques for spatial snapshot guidance using an HMD video stream (Figure 1). ViewAR provides no 3D information: the helper annotates a snapshot shared with the worker. It resembles screen-sharing or paper-based instructions providing 2D guidance [34]. DepthAR uses estimated OST-HMD depth to place 3D annotations on surfaces, similar to remote XR systems such as Skype for HoloLens [8]. Finally, SlidAR extends DepthAR by allowing helpers to adjust annotations along the ray cast from the snapshot pose, based on Polvi et al. [41].

While each of these techniques is preferable to the others in some situations, our goal is to investigate how users approach spatial snapshot instructions in a common scenario. To this end, we present an empirical comparison of three spatial snapshot annotation techniques in a controlled user study. We focus on an assembly scenario in which a helper on a desktop computer guides a worker wearing an OST-HMD. The helper can take snapshots of the worker's video feed and annotate them using the three techniques (Figure 1). Within this scenario, we have several research questions that can guide future development of spatial snapshot collaboration:

RQ1: How do users utilize spatial snapshots while guiding others in remote collaboration?
RQ2: How do the annotation methods affect user performance?
RQ3: Do communication methods change between the techniques?
RQ4: How does user experience differ across the three techniques?

To answer these research questions, we conducted an empirical within-subjects user study with 24 participants comparing the three annotation techniques. In the study, we compared the user's performance in terms of errors and completion time, as well as communication approaches and engagement with the system. Our findings show that the placement features were not decisive in determining user performance. The participants took very different approaches to accomplish the task, some relying heavily on verbal communications, while others tried to use the annotation features to their fullest potential. Our findings provide insights into how participants approach annotation creation and guidance using spatial snapshots. Hence, our main contributions are as follows:

(1) We present a first evaluation of three spatial snapshot annotation techniques for remote collaboration.
(2) Our analysis of the user's performance shows very different work styles that require various interaction methods.
(3) We discuss the benefits and challenges presented by the spatial snapshot scenario and its effects on the helper.

2 Related work

Our study builds on previous explorations of remote collaboration and annotation creation in XR. In particular, we discuss methods developed for multi-view annotation creation.

2.1 Remote collaboration in mixed reality

As highlighted in recent reviews [47, 58], remote collaboration is an anticipated application of XR technology. Remote collaboration

in XR can be facilitated through spatial augmented reality [57], handheld devices [26, 61] or head-mounted displays [31, 37].

The simplest way to share instructions is to let the helper annotate an image captured by the worker [9, 36, 38] and place this annotated view in the worker's environment. Other solutions rely on depth cameras to place annotations [1, 8] on a captured surface of the environment, providing a clearer spatial association between the instructions and the scene. However, this limits the placement of the annotations to the view of the capture device and the depth information. Others utilize a 360° camera [50, 52] to capture and share the collaboration space.

Applications for spontaneous remote collaboration often rely on handheld devices [4, 18, 60, 61], such as smartphones and tablets [25, 33]. Remote handheld collaboration leads to intentional viewpoint sharing with the worker, as the helper must physically direct the worker to share the workspace. Using an HMD instead of a handheld device allows the worker to perform instructed steps with both hands while simultaneously viewing the annotations placed by the helper. However, the use of an HMD does have its drawbacks. The remote viewpoint is always locked to the worker's viewpoint, which can cause discomfort when the view drastically changes [30].

Providing multiple viewpoints can enhance user satisfaction and system flexibility [64] as well as provide the helper with better spatial understanding [55] and situational awareness [45]. Providing view independence from the worker's viewpoint can lead to improvements in the collaboration [49], resulting in faster task completion times and fewer verbal instructions. With a 3D reconstruction of the work area [11, 24, 51, 53], the helper may freely explore the environment. Advanced reconstruction or scene recreation can also enhance collaborative features, by enhancing communication and collaboration [17]. For example, by gesture sharing [2], gaze sharing [23], virtual replicas [6, 54], and avatar sharing [62].

As it is not always possible to capture a detailed model of the scene, some systems have addressed this issue by using alternative representations, such as light fields [35], neural radiance fields [22, 46], or 3D Gaussian Splats [48]. However, these methods can support only a small working environment and require the scene to be recaptured when it changes, a potentially time-consuming process. To address this recent work has utilized Gaussian splatting to create and share environments [21, 63]. While effective, these approaches often require computationally intensive reconstruction pipelines and assumptions about available hardware capabilities.

2.2 Reconstruction-free annotations

Multi-view annotations have been proposed as a means of placing 3D content into a scene without the need for a detailed reconstruction and supporting unconstrained working environments. Gauglitz et al. [14] investigated methods for converting drawings on a screen into 3D content placed in the scene. They compared a spray paint approach that projects a drawing onto the detected surface, a plane fitted to the detected surface, and a plane aligned with the user's viewpoint. They also investigated how users could draw instructions on a mobile device [15]. For arrows, they proposed to cast a ray from the location of the drawn tip as a reference to the user as they move through the scene. When the user draws a second arrow from a new view direction, they triangulate the tip and the base of the arrow to determine the intended position. They found their system to be both faster and preferred over video-based annotations and simple video-based collaboration.

Building on their idea, Polvi et al. [41] presented SlidAR. Rather than relying on disambiguation of user drawing, they focus on placing predefined 3D objects into the scene. Assuming that the user can correctly place a virtual object in the current view, they initially place it at a predefined depth along the ray from the current view. After a view shift, users can adjust the depth by sliding the object along the cast ray. They showed that this results in a more efficient placement technique than object placement locked to a pointing device. Fuvattanasilp et al. [10] expanded on SlidAR, allowing efficient orientation control.

Although several works suggested the use of multi-view content creation for remote collaboration [26, 40], they did not evaluate the performance of such systems or how users engage with them in a collaboration scenario.

3 System overview

Our system consists of an application running on an OST-HMD (HoloLens 2) for the worker and an application running on a desktop machine for the helper (Figure 2). We implemented both applications in Unity 2019.4.40f1 and used MixedReality-WebRTC[1] for network communication. We support three methods for annotation placement inspired by existing scene annotation methods:

(1) ViewAR shows only an annotated snapshot.
(2) DepthAR places a chevron on the estimated, low-quality surface reconstruction.
(3) SlidAR allows for depth adjustments of the chevron position after initial placement.

In the following, we describe the base system and implementation of the annotation placement methods.

3.1 Worker application

Our system streams the live view of the worker's HMD camera (Locatable Camera in HoloLens 2) tagged with the HMD pose at frame capture. To ensure a sufficiently high frame rate, we capture the image in a resolution of 960×540 pixels at 30 frames per second. As it is not possible to access the HMD camera pose while streaming the images, we determine the offset between the camera and the tracked pose of the HMD through an offline calibration. This was done by placing the HMD on a flat surface and recording the device's pose and frames containing scene camera's pose. The average offset between the poses was used as a static offset during runtime. Furthermore, we empirically synchronize the tracked camera pose and frame timestamps. Hereby, we identified onsets of movement in the streamed camera frames and device poses to determine the offset between the two and visually verified the result. To reduce network load, we do not stream the depth data or environment model captured by the HMD. Instead, we use the SLAM map built by the HoloLens 2 to query the distance from the camera to the surrounding surfaces to place an annotation by the helper.

When the worker receives instructions in ViewAR, we place a static image viewer in front of them. The image viewer rotates to always face the worker; however, its position may not be adjusted.

[1] https://github.com/microsoft/MixedReality-WebRTC

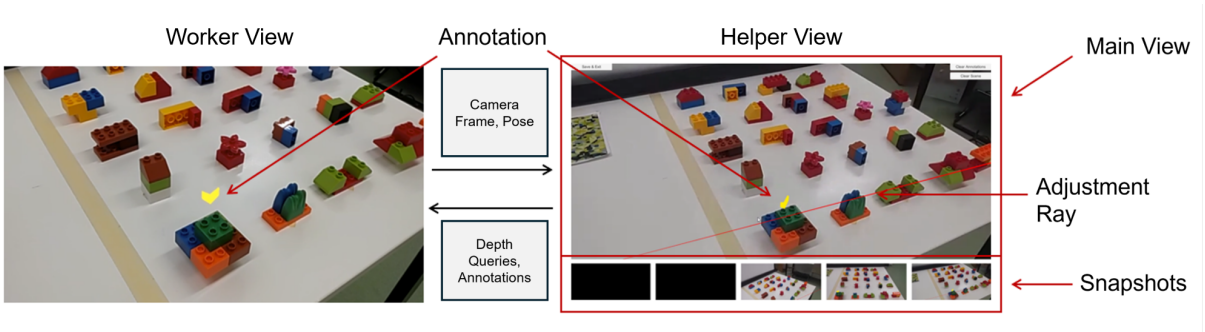


Figure 2: All systems consist of an application running on a HoloLens 2 and a desktop computer. The HoloLens 2 shares frames captured by the world-facing camera and its poses. The desktop interface consists of a ribbon of captured snapshots that store views in time, as well as a static live view in the bottom right position. The main screen displays an enlarged version of the selected view. Helpers can capture snapshots and place and edit annotations from both the live feed and captured snapshots. The adjustment ray allows the helper to adjust the depth of the placed annotation, thus enabling accurate annotation placement.

Using DepthAR and SlidAR, the systems display two types of annotations. The first is a camera annotation that appears as a 3D camera glyph indicating the position and orientation where the helper has taken a snapshot. The second type of annotation is the chevron annotation that appears as the helper places them. These chevron annotations initially appear as the helper places them in the scene. They slowly spin for two seconds to indicate the newest annotation. The color of the annotation cycles through five distinct colors (red, green, cyan, yellow, and magenta) as it is placed before repeating. Workers cannot interact with any of the annotations or add additional annotations.

3.2 Helper application

On the desktop computer, the helper is presented with the main screen and five snapshot placeholders (Figure 2, right). The rightmost view continuously displays the video from the live view, while the other four thumbnails show snapshots, with the leftmost being the oldest and the rightmost being the newest.

While the live view is shown on the main screen, the helper can take a snapshot with a hotkey, which shifts all stored snapshots. The new snapshot is automatically displayed on the main screen. The helper can return to the live view by pressing the hotkey again or can click on any thumbnail to switch the main screen to that view. The helper can remove all annotations or delete all snapshots and annotations using two buttons on the right of the main screen.

With ViewAR, helpers may add a chevron annotation by left-clicking on the desired location on the main screen. These annotations may only be placed on snapshots. After the annotation was placed, helpers can enter the edit mode by selecting the annotation and adjusting its size with the mouse, or the plus and minus keys on the keyboard. Finally, the annotated snapshot can be shared with the worker by pressing a send button.

With DepthAR, the helper may create a 3D chevron by clicking the desired location on the main screen. The chevron is created at a distance d along the ray cast from the selected pixel. Here, d is determined by the HMD worn by the worker by intersecting the ray with its coarse reconstruction of the scene. If the cast ray does not intersect a surface, d is set by default to 1.5 m; otherwise, it is set to the distance to the surface. Please note that, for the purposes of this study, all objects are aligned to gravity, and helpers cannot adjust their orientation. A new annotation is automatically shared with the worker and is shown in all snapshots in the desktop application. The helper may enter the edit mode by selecting the annotation and adjust the size of the chevron or delete it. helpers were also able to shift the position of the object with the mouse. This interaction keeps the depth of the virtual object fixed in the current view.

With SlidAR, the helper creates an annotation the same way as with DepthAR. However, in the edit mode, a ray from the object to the position of the snapshot from which it was created is shown. The helper can move the object along this ray with the mouse scroll-wheel. The adjustment is replicated in real time for the worker. Helpers can manipulate the annotation in the same way as for DepthAR. If the helper shifts the annotation, the system computes a new ray from the object position towards the center of the current view, which serves as the new reference for future adjustments.

4 User study

Our goal was to study how participants approach spatial snapshot remote collaboration systems and what their engagement with the three placement methodologies is in a common scenario. We designed an assembly task that should prompt them to utilize the different features offered by the three annotation methods. Our user study thus had only one independent variable, the method (ViewAR, DepthAR, SlidAR).

4.1 Task

The task aimed for participants to collaborate and complete the assembly quickly and accurately. Participants were tasked with the selection of the correct Duplo pieces and their assembly. Utilizing the taxonomy defined by Ghamandi et al. [16] this task would be considered an active selection and manipulation task. This is a common task due to participants' familiarity with Duplos, large variability, and ease of access that supports replications [19, 29, 65].

These Duplo blocks were preassembled into pieces containing 1 to 3 Duplo blocks glued together and placed on a grid. This created unique pieces that need to be rotated and placed specifically in the

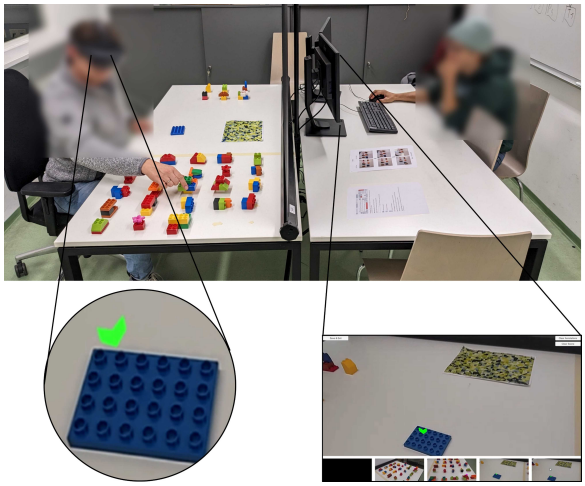


Figure 3: Layout of the user study. The helper on the right places annotations according to the assembly instructions on the desk. The worker on the left selects a Duplo assembly piece indicated by the helper.

assembly process. There were 23 pieces in total, with the pieces placed in a 5×5 piece grid with each piece spaced 5 cm apart, with the two rows furthest from the worker consisting of only 4 pieces. Additionally, 6 Duplo block pieces utilized for the training. These were separated from the study assembly pieces, being placed on the opposite end of the 1.6×.6 meter table. The training assembly consisted of three steps and was repeated for all conditions.

The helper was given a set of assembly instructionsthat demonstrate how to assemble the Duplo blocks. They consisted of a front and top view of each step with arrows indicating the correct rotation and placement location of the piece. Each assembly in the study consisted of selecting and assembling six Duplo blocks. A base plate was provided for the worker to build upon.

To reduce latency, both participants were in the same room during the experiment (Figure 3). The room was divided into two, which prevented eye contact. On one side of the room divider was the helper who saw the system interface on a 24" monitor connected to a desktop computer (i7-12700K CPU, 64GB RAM, Nvidia RTX3090 Ti). They were also provided with task instructions and an overview of the system controls printed on A4 paper. The worker wearing a HoloLens 2 was placed on the other side of the divider and was given a chair; however, they were allowed to move as they pleased, barring they did not cross the divider.

To understand how participants engaged with the interface, we recorded all helper input, the poses of the HoloLens 2, and the participants' communication. We transcribed the communications utilizing Whisper [42], which was run offline, and synchronized the transcripts with the remaining data for further analysis.

To ensure consistency between workers, we used a fiducial marker that was detected with the Vuforia library[2] as the origin of the working space. Once the system detected the marker, workers saw a red outline to confirm the detection accuracy. They confirmed the alignment and initiated the connection to the desktop computer by stating "connect." This disabled the marker tracking. Furthermore, we placed the image viewer for ViewAR 10cm above the detected fiducial marker.

[2]https://developer.vuforia.com/

4.2 Procedure

The participants were welcomed by the experiment supervisor and received an explanation of the objective of the experiment and the task, as well as a brief overview of all three systems. The participants then filled out a consent form and were asked to decide if they wanted to start as the helper or worker. If no consensus was reached, these roles were assigned by the experiment supervisor. After calibrating the HoloLens 2, participants received a detailed explanation of the first interface. The pair of participants was then given a practice assembly consisting of three steps. During this practice assembly, they were given as much time as they needed to familiarize themselves with the system and utilize all the functionality before continuing. They were also not allowed to move beyond the training task until the experiment supervisor was satisfied that they understood the system's capabilities.

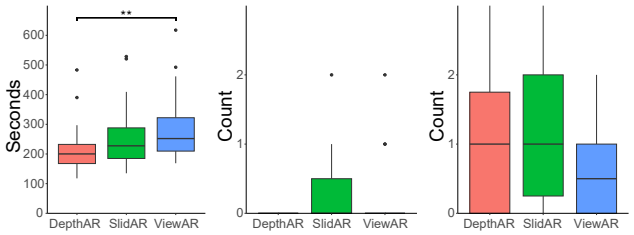
After the training, the participants performed the main assembly. They initialized the time measurement by pressing a start button covering the main screen and completed the task by selecting a "save and exit" button shown in the top left corner of the screen. After the task was completed, the participants filled in a System Usability Scale (SUS) [5] and NASA Total Load Index (NASA-TLX) [20] questionnaire and provided free-form feedback. After repeating the process for all three interfaces, participants could take a short break. Then, they switched roles and repeated the process.

At the end of the experiment, participants filled out a demographics questionnaire. To avoid learning effects, six distinct models were selected so that no model was seen more than once. The order of the interfaces and the assignment of models to interfaces were counterbalanced to avoid order effects. The experiment took about two hours (one hour per role). The participants received a bar of chocolate as a token of gratitude. In addition, participants who were not employed at our institution received 20EUR. The experiment did not require a review by the institutional review board.

4.3 Participants

We recruited 24 participants (14 male, 10 female; mean age 26.6 years, range 21-33) from the student and staff population at our university through word of mouth and email elicitation. Two participants rated their English skills as limited working proficiency, while the others reported professional working proficiency or higher. All participants had normal or corrected-to-normal vision (12 with corrected vision). On average, participants rated their familiarity with XR at 2.83 on a 1-5 scale, with 5 indicating very familiar.

We excluded three tasks (2 SlidAR, 1 ViewAR) where participants stopped the task before restarting and completing the study. Additionally, four tasks (1 SlidAR, 1 DepthAR, 2 ViewAR) were excluded from the analysis of Errors and Confusions due to corrupted recordings.



(a) Time (s) **(b) Errors** **(c) Confusions**

Figure 4: Objective results for time taken (a) and the number of errors (b) and confusions (c).

5 Results

A Shapiro-Wilk test indicated that our data did not satisfy the normality assumption. We thus used nonparametric tests for all our evaluations. Because we had to remove some of the objective measurements from our data, we used a Skillings-Mack test that is robust to missing values. We used multiple Wilcoxon tests with Bonferroni correction as a post hoc evaluation. For our subjective data, we did not exclude the responses from participants for trials that were excluded from the objective analysis because participants engaged with the system and could provide feedback on their experience. We thus analyzed this data with a Friedman test with the Nemenyi post hoc test with Bonferroni correction. We used a threshold of $\alpha \leq 0.05$ to determine statistically significant differences. All tests were performed in R using the rstatix (v0.7.2) and PMCMRplus (v1.9.12) packages.

5.1 Assembly time

On average, participants completed the assemblies with DepthAR in $\bar{X} = 217.63s$, $SD = 82.93$; SlidAR: $\bar{X} = 266.20s$, $SD = 112.90$; ViewAR: $\bar{X} = 287.27s$, $SD = 114.89$ (Figure 4). The Skillings-Mack test indicated a significant difference between the three methods ($\chi^2(2) = 13.21$, $p < 0.01$). A post hoc Wilcoxon test showed that DepthAR was significantly faster than ViewAR (p<0.01, r = 0.66, medium effect). No significant differences were detected between SlidAR and ViewAR (p=0.73) and SlidAR and DepthAR (p=0.12).

5.2 Errors

In general, the workers made few errors and had little confusion (Figure 4). A Skillings-Mack test did not show statistically significant differences in the overall number of errors ($\chi^2(2) = 1.74$, $p = 0.42$). We also considered if the phase when errors occurred had an effect. We differentiate between errors made during part selection and errors made during the assembly. An error in part selection was defined as selecting an incorrect part, and an error in assembly was defined as having an error in the completed assembly. The errors made during part selection are as follows. DepthAR: $\bar{X} = 0.00$, $SD = 0.00$; SlidAR: $\bar{X} = 0.14$, $SD = 0.36$; ViewAR: $\bar{X} = 0.05$, $SD = 0.22$. The errors made during assembly are as follows. DepthAR: $\bar{X} = 0.00$, $SD = 0.00$; SlidAR: $\bar{X} = 0.14$, $SD = 0.36$; ViewAR: $\bar{X} = 0.14$, $SD = 0.36$. Overall, the workers made few errors in both part selection and assembly. A Skillings-Mack test did not show statistically significant differences in the number of part-picking errors ($\chi^2(2) = 0.44$, $p = 0.80$) or assembly errors ($\chi^2(2) = 0.68$, $p = 0.71$).

5.3 Confusion

A Skillings-Mack test did not show statistically significant differences in the overall number of confusion events ($\chi^2(2) = 2.86$, $p = 0.24$). We also considered if the phase has an effect on confusion events. We categorized confusion events into two types: confusion during part selection and during assembly. Confusion during part selection occurred when workers were uncertain about which part to choose and asked for clarification. Additionally, during video reviews, confusion was identified if workers exhibited behaviors such as reaching for multiple parts before making a selection or hovering over a part prior to receiving clarification. Confusion during assembly was characterized by instances where workers assembled a piece incorrectly and quickly corrected it or when they needed to ask clarifying questions. The confusion events occurring during part selection are as follows. DepthAR: $\bar{X} = 0.43$, $SD = 0.73$; SlidAR: $\bar{X} = 0.38$, $SD = 0.59$; ViewAR: $\bar{X} = 0.05$, $SD = 0.22$. The confusion events occurring during assembly are as follows. DepthAR: $\bar{X} = 0.74$, $SD = 0.92$; SlidAR: $\bar{X} = 0.95$, $SD = 1.12$; ViewAR: $\bar{X} = 0.57$, $SD = 0.68$. Overall, there were few confusion events in both part selection and assembly. A Skillings-Mack test did not show statistically significant differences in the number of confusion events during part selection ($\chi^2(2) = 3.58$, $p = 0.17$) or assembly ($\chi^2(2) = 0.95$, $p = 0.62$).

5.4 Movement

As two of the methods (SlidAR and ViewAR) require multiple views to effectively utilize the features, we analyzed the amount of movement the worker underwent during the assembly process. The movement (in meters) for the various systems is as follows. DepthAR: $\bar{X} = 6.41$, $SD = 3.57$; SlidAR: $\bar{X} = 8.10$, $SD = 5.57$; ViewAR: $\bar{X} = 7.13$, $SD = 2.87$. A Skillings-Mack test showed significant differences ($\chi^2(2) = 8.53$, $p < 0.01$). A post hoc Wilcoxon test showed that in DepthAR moved significantly less than in SlidAR (p=0.02, r=0.56, medium effect).

5.5 Snapshot Use

We also analyzed how much the helpers utilized the snapshots. We found that helpers tended to utilize the live view and a single snapshot (Table 1). Preferring instead to replace an initial snapshot as opposed to utilizing multiple simultaneous viewpoints. With less than half of the helpers utilizing 2 or more snapshots.

The number of snapshots taken by helpers was DepthAR: $\bar{X} = 2.75$, $SD = 4.30$; SlidAR: $\bar{X} = 3.50$, $SD = 4.11$; ViewAR: $\bar{X} = 5.70$, $SD = 3.76$. A Skillings-Mack test shows significant differences between the three methods ($\chi^2(2) = 12.90$, $p < 0.01$). A post hoc Wilcoxon test indicated that DepthAR p=0.02, r=0.58, medium effect) and SlidAR (p=0.02, r=.59, medium effect) utilized significantly fewer snapshots than ViewAR. No significant differences were detected between DepthAR and SlidAR (p=0.57).

We also investigated if there was a correlation between number of snapshots taken and task completion time. Spearman's rank correlation test resulted in a strong positive correlation for DepthAR:

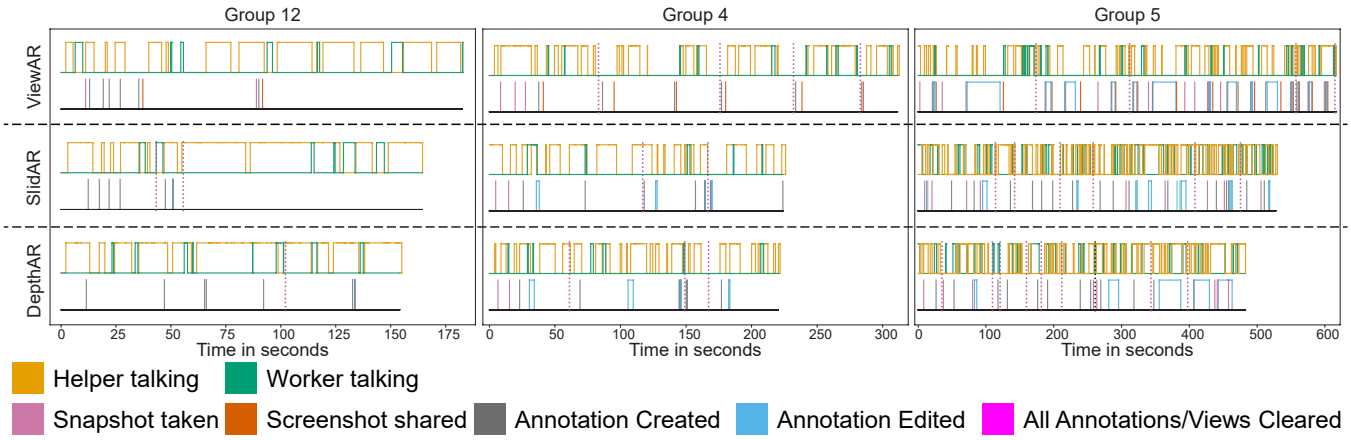


Figure 5: Timelines of actions taken by different groups, highlighting the various strategies employed. For each method, verbal communications are shown at the top and actions taken by the helper are shown at the bottom. Group 12 relied mainly on verbal instructions, while Group 4 used few instructions with a higher system engagement. Group 5 utilized the system heavily with many back-and-forth discussions.

$\rho = 0.60, p < 0.01$ and SlidAR: $\rho = 0.60, p < 0.01$. No significant correlation was found for ViewAR: $\rho = 0.43, p = 0.06$.

5.6 Engagement with the system

Participants engaged with the system in a multitude of ways. As seen in Figure 5, participants devised differing strategies to complete the assembly tasks. Some groups, like group 12, utilized the system to assist in the part-picking portion of the task but then found verbal instructions sufficient to complete the task. Other groups, such as group 5, employed the system to its full capacity while also extensively talking about each action. Group 4 used verbal instructions with moderate engagement with the system.

Skillings-Mack test did not show statistically significant differences in the number of annotations utilized ($\chi^2(2) = 1.29,\ p = 0.52$): DepthAR: $\bar{X} = 16.6$, $SD = 4$; SlidAR: $\bar{X} = 16$, $SD = 4.11$; ViewAR: $\bar{X} = 15.2$, $SD = 3.58$. Spearman's rank correlation test resulted in a strong correlation for DepthAR: $\rho = 0.50, p = 0.02$; SlidAR: $\rho = 0.65, p < 0.01$; and ViewAR: $\rho = 0.58, p < 0.01$.

Overall, we did not find significant differences in the number of verbal instructions given for different methods: DepthAR: $\bar{X} = 16.78$, $SD = 3.97$; SlidAR: $\bar{X} = 16.29$, $SD = 3.9$; ViewAR: $\bar{X} = 14.81$, $SD = 3.4$. Participants utilized similar numbers of verbal instructions for all three systems. A Skillings-Mack test did not show statistically significant differences in the number of instructions provided ($\chi^2(2) = 1.75,\ p = 0.42$).

Table 1: Number of helpers utilizing different viewpoints.

View	DepthAR	SlidAR	ViewAR
Live	24	22	23
Snap 0	14	16	23
Snap 1	6	8	11
Snap 2	4	5	6
Snap 3	2	4	3

Spearman's rank correlation test resulted in a significant positive correlation for DepthAR: $\rho = 0.61, p < 0.01$ and ViewAR: $\rho = 0.60, p < 0.01$. No significant correlation was found for SlidAR: $\rho = 0.32, p = 0.16$.

5.7 Subjective measures

Using the interpretation of Bangor et al. [3], all systems can be considered "usable" with scores above 68. The usability of DepthAR and SlidAR can be considered "good" with scores above 71. (Figure 6c). For helpers, we find DepthAR: $\bar{X} = 75.52$, $SD = 19.10$; SlidAR: $\bar{X} = 71.88$, $SD = 22.51$; ViewAR: $\bar{X} = 68.75$, $SD = 16.81$. A Friedman test did not reveal significant differences (F(2)=3.98, p=0.14) between the overall scores. For individual questions, the test revealed significant differences in the perceived encumbrement of helpers (DepthAR: $\bar{X} = 2.00$, $SD = 1.19$; SlidAR: $\bar{X} = 2.21$, $SD = 1.04$; ViewAR: $\bar{X} = 2.83$, $SD = 1.28$) (F(2)=6.43, p=0.04), however post hoc analysis did not reveal significant differences.

Workers rated all methods as "good," with all scores above 71 (DepthAR: $\bar{X} = 78.54$, $SD = 18.61$; SlidAR: $\bar{X} = 75.42$, $SD = 19.12$; ViewAR: $\bar{X} = 73.65$, $SD = 18.64$). A Friedman test did not reveal significant differences (F(2)=4.16, p=0.12). This suggests that all systems met the base condition of usability. For the individual questions, Friedman tests did not find any significant differences.

When analyzing the raw NASA-TLX [20] responses, we excluded the performance section of the questionnaire, as several participants mentioned during the interview that they had answered the question incorrectly, i.e., unintentionally rated their success as a complete failure. This is supported by the large variance in responses for helpers $\bar{X} = 44.45$, $SD = 42.98$ and workers $\bar{X} = 49.44$, $SD = 44.26$. This is further supported by the low number of assembly errors across all systems, suggesting that the groups successfully completed the task. This was the only NASA-TLX item that participants reported answering incorrectly. The remaining NASA-TLX items exhibited substantially lower variance, providing no indication that similar misunderstandings affected the other questionnaire items.

Although the workload across all systems was low (Figure 6), Friedman tests showed that workers reported significantly different temporal demands (F(2)=6.62, p=0.04). However, the post hoc test did not reveal any differences between the methods. No other differences were detected.

5.8 User feedback

We collected qualitative feedback after each assembly during the participant surveys. Participants were asked what they liked and disliked about the system, as well as what they found the most and least helpful. Lastly, they were asked about any changes they would make to the system before being given the opportunity to provide any freeform feedback they wanted.

Helpers appreciated the simplicity of DepthAR, with fewer options to augment annotations. With one helper stating:

P5: *"Very intuitive, easy to use, only minimal set of functionalities you need to remember."*

However, some helpers noted challenges in precise annotation placement expressing a desire for better control. They instead found the additional functionality of SlidAR beneficial. With some stating:

P2: *"It was more difficult to assign the hologram to the correct position than with the other systems; more misunderstandings occurred."*

P21: *"a real-time updating system was easier to control and explain."*

Some helpers found adjusting the depth of annotations in SlidAR cumbersome and preferred replacing misplaced annotations rather than repositioning them. For example,

P11: *"I didn't use the mode to change the depth of the hologram, I found it pretty complicated"*

This could be due to difficulties to capture suitable viewpoints

P7: *"Sometimes it was hard to define a good view angle that perceived the depth "correctly""*

To mitigate this, some groups adopted alternative strategies, such as utilizing a top-down viewpoint to minimize the need for depth adjustments, essentially reverting to DepthAR.

Helpers also found certain aspects of ViewAR helpful. Its similarity to ubiquitous screen sharing and paper-based instructions made it both familiar and easy to understand, with one helper remarking:

P8: *"When sending screenshots, the other user sees my 2D screen, better capturing my intentions"*

Overall, helpers found all systems usable and provided them with the tools needed to accomplish the assembly.

Workers generally found positive aspects for all methods. DepthAR was seen as interactive:

P19: *"I really like to have the interaction directly on the fly so i do not have too much waiting time"*

For SlidAR, workers appreciated the improved accuracy that can be achieved by adjusting the viewpoint:

P2: *"I liked the simple starting point, if you change the viewpoints more often then the holograms are positioned more correctly."*

P16: *"It was very easy to use, as the markers were directly on top of the needed piece"*

For ViewAR, workers highlighted the familiarity with the visualization it offered:

P6: *"The instructions are more similar to conventional instructions (images on a piece of paper), so it seems more comfortable to work with this system compared to DepthAR, which seems more fancy. In between screenshots, the user can relax and does not need to focus on the task until further instructions are sent."*

In terms of complaints, the most common complaints from workers were related to the HoloLens 2 device. workers noted that the camera's position above the eyes caused them to look farther down than expected, and they also expressed frustration with the limited field of view. With workers stating:

P13: *"I didn't like having to look so far down to give the person with the PC the correct view"*

6 Discussion

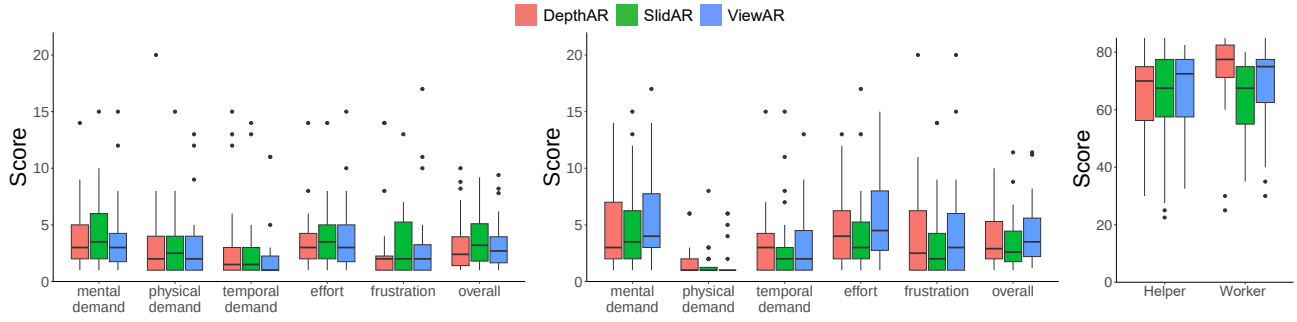
In regard to RQ1, our findings show that helpers utilized only a few snapshots, opting to primarily work with the live view. This is somewhat surprising, as different snapshots could in principle be used to provide instructions for different areas, e.g., the part picking and assembly. Prior work by Lin et al. [30] showed that live-view tethering in AR remote collaboration can negatively affect the helper experience and that a stabilized view, e.g., offered by a snapshot, could be preferable.

Contrary to that we found helpers preferred working with the live view. This is in line with Choi et al. [9] who showed that helpers preferred working with the live view in a dynamic environment due to challenges in maintaining consistent spatial understandings across the changing viewpoints. In our system, all snapshots were persistently available in the interface alongside a continuously updated live view. The snapshots were intentionally taken by the helper as opposed to being shared by the worker and their viewpoints were explicitly indicated to workers via camera glyphs. Our findings thus show that live view preference persists even when snapshots provide improved spatial stability. This suggests that factors beyond environmental dynamics, such as workflow continuity and interaction speed, influence viewpoint selection in practice.

As expected, all three methods showed a strong postive correlation between the number of snapshots taken and an increase in task completion time. However, the correlation was stronger in SlidAR and ViewAR than in DepthAR. This suggests that participants using these methods were more likely to engage in extensive planning, leading to a greater impact on task completion time. In scenarios where high accuracy is not necessary, DepthAR will likely be sufficient to successfully complete the task and could be faster than methods requiring more engagement.

We also found that helpers sometimes had difficulties navigating the worker to a novel viewpoint from which they could adjust the position of the annotations with SlidAR. Workers moving significantly more in SlidAR than DepthAR indicates that they searched for a suitable viewpoint from which to adjust the annotation placement. Overall we find that snapshot adoption was limited even under conditions where they are theoretically beneficial.

In response to RQ2, our findings demonstrate that across all three conditions, participants made few errors and had little confusion.



(a) NASA-TLX score for workers

(b) NASA-TLX score for helpers

(c) SUS score

Figure 6: Subjective feedback from users. All methods were rated as "good" with low demand.

In terms of task completion time DepthAR was faster than ViewAR. No significant differences were detected between DepthAR and SlidAR and SlidAR and ViewAR. This result was unexpected, as our pilot studies showed faster task completion times and fewer errors and confusions when utilizing SlidAR. Participants in our pilot studies found the additional placement features beneficial in placing accurate annotations, leading to better performance and usability. This may be because the pilot study participants had extensive familiarity with XR technologies and annotations placed in 3D space. Additionally, as we did not restrict the user's viewpoints, participants found ways to mitigate the limitations imposed by the technique. For instance, utilizing top down viewpoints that required little to no depth adjustments. In scenarios where the depth discrepancies cannot be mitigated or where the depth information is missing, the adjustment may become more important resulting in performance differences. More importantly, these findings indicate that participants can adapt their interaction strategies to mitigate differences between techniques, effectively reducing the impact of intended design variations on task performance.

Concerning RQ3, our findings show that the number of provided instructions remained stable between the three methods. This is somewhat unexpected, as prior work suggests that communication effort in remote collaboration can be influenced by the accuracy of annotations, where more precise annotations can reduce the need for verbal instruction [31]. The amounts of verbal instructions given across the three conditions were not found to be significantly different. However, both DepthAR and ViewAR showed positive correlations between the number of verbal instructions provided and task completion time, with ViewAR exhibiting a particularly strong correlation. This suggests that participants relied more heavily on verbal communication while using ViewAR, likely to describe and discuss the authored instructions. The strong correlation may reflect a greater need for verbal clarification or collaborative discussion when utilizing this method. A moderate positive correlation was also detected with the instructions provided when utilizing DepthAR and task completion time. This is likely due to the imprecise nature of the annotations. Requiring helpers to provide more detailed instructions to compensate for the decreased accuracy of the annotation. Even when annotation quality and visualization differ, collaborators do not reduce the amount of verbal instructions, indicating that verbal communication serves a stable coordination function that is largely independent of the visualization technique.

In regard to RQ4, our findings demonstrate that user experience did differ across the three methodologies. Helpers found ViewAR the least usable, citing difficulties annotating the snapshot while the worker performed actions out of view. For workers, all systems were considered "good" for usability; however, feedback showed that the real-time annotation systems, DepthAR and SlidAR, were preferred as opposed to ViewAR's visualization. With several workers finding looking back and forth between the assembly space and the instruction viewer frustrating. Between SlidAR and DepthAR, most helpers enjoyed the simplicity of DepthAR. Finding it easier to use, without the need to position the worker to provide specific viewpoints to adjust the annotations, instead preferring to place the annotation and clarify their intent. This suggests that reducing interaction overhead can have a stronger impact on user experience than providing more options for annotation placement. This preference could be explained by the findings of Kim et al. [27] who observed that helpers asking workers to change their viewpoint was perceived as troublesome by both parties. Placing the snapshot locations into a 3D map and allowing helpers to draw on it, similar to the work of Mohr et al. [35], could simplify the guidance. Overall, these findings indicate a trade-off between annotation placement options and usability in spatial snapshot systems, with helpers preferring simpler interaction designs when task demands do not require fine-grained spatial control.

7 Design Guidelines

These design guidelines, informed by our findings, aim to enhance the future development of spatial snapshot remote collaboration systems, improving both their efficiency and user experience.

Design for selective snapshot use. Helpers primarily relied on the live view and only occasionally captured snapshots. This suggests that snapshots are used opportunistically rather than as a structured collection of viewpoints. Snapshots should therefore be treatedFuture systems should therefore treat snapshots as lightweight, on-demand references rather than as a library of stored viewpoints, avoiding designs that implicitly encourage users to construct large or structured snapshot collections.

Support snapshot management. While helpers generally used only a small number of snapshots, they expressed a need to retain particularly useful viewpoints for later use. Future systems should therefore support management mechanisms, such as pinning or grouping viewpoints. Providing spatial overviews of snapshot locations may also help identify a suitable snapshot to use.

Initial placement is critical. Several participants opted to replace the instructions and to communicate verbally rather than adjust them to account for placement errors. This suggests that users tend to avoid refinement of annotations, likely due to the interaction cost involved (time, effort). Future systems should therefore prioritize enabling accurate initial placements and reducing the need for subsequent adjustments or corrections.

Support different workflows. We observed that different groups approached the task and system features differently, some opting to verbalize most instructions, while others focused on the features offered by each method. This suggests that users naturally adopt different coordination strategies depending on user preference and task interpretation. Future systems should therefore support flexible combinations of methodology and annotations rather than assuming one dominant workflow.

8 Limitations

This study was conducted in a controlled laboratory setting, which may limit the generalizability of our findings to more diverse real-world scenarios. The first limitation is that our participants were recruited from the local university. While we attempted to balance the gender of the participants, we were unable to achieve a perfectly balanced sample size.

While participants rated themselves as somewhat familiar with XR, more experienced users could lead to different results. Additional training and expertise of participants could have an effect. The helpers were not true experts in the assembly procedures but instead were given the assembly instructions just moments prior to the assembly task. This could potentially lead to increased assembly time, as the helper needs to first comprehend the instructions before relaying that information to the worker. Finally, both participants were not true experts with the system. While they had time to familiarize themselves with the interfaces, increased training and experience could potentially change the way they utilize the systems. More extensive training on specific features could mitigate this shortcoming and lead to different results.

Our choice of displays could have had an impact as well. The scene camera of the HoloLens 2 is positioned above the worker's eyes, requiring the worker to look further downward than expected to capture the work area. Several workers commented on the camera placement feeling unnatural. Evaluations with different camera placements or devices, such as a video see-through head-mounted display (VST-HMD), could lead to different user behavior.

Another limitation is our focus on the placement of annotations as 3D objects. Natural drawings can create more expressive instructions, potentially reducing the need for oral instructions. By limiting the helper to a single 3D object to place, we potentially removed further avenues for virtual instructions to be given.

Our study also utilizes a simple Duplo assembly task with large pieces and limited mechanical complexity, which may not fully reflect real-world collaborative scenarios. In addition, the simplicity of the task resulted in low error and confusion rates across conditions, suggesting a potential ceiling effect that may have reduced sensitivity to performance differences between the techniques, particularly for detecting differences involving SlidAR.

Finally, this study was conducted in the same room, which is a common experimental approach in AR remote collaboration research to control for network latency and ensure stable and clear communication conditions [2, 15, 19]. While this improves communication clarity, it may also lead to more efficient verbal coordination and does not fully capture the constraints of truly remote collaborative environments. In such scenarios helpers may utilize more snapshots and annotations to supplement delayed, or unclear verbal communication.

9 Conclusions and future work

In this paper, we presented an empirical evaluation of three spatial snapshot remote collaboration systems: DepthAR, SlidAR, and ViewAR. Through a within-subjects user study with 24 participants, we investigated how different annotation placement methodologies affect user behavior and interaction during remote collaboration. We found that all methods were effective for collaboration, and helpers generally preferred the simple methodology of DepthAR over the more complex approaches used by SlidAR and ViewAR. The simplicity provided by DepthAR allowed helpers to combine verbal communication with sufficiently accurate annotations to complete the task efficiently.

There are many opportunities for future work, such as implementing feedback from participants to further improve the usability of the systems. For instance, some suggested changes to the system include the ability to pin a snapshot for future use. Allowing the helper to keep beneficial snapshots. Another suggestion is to visualize the adjustment ray to the worker. This visualization would allow the worker the potential to infer where the helper intends to place the annotation before the adjustment is made. Finally, we could add additional functionality such as mouse gestures [28] and additional annotation methods such as the ability to draw. Additionally, workers could send information back to the helpers.

Another opportunity for future work is varying the collaboration scenario, e.g., targeting a scenario where the precision of the annotation is paramount to task success or where verbal instructions would either be impossible or difficult to deliver. Such a scenario could consist of the assembly of electrical components, which could potentially alter the user behavior, system preference, and usability scores of DepthAR, SlidAR, and ViewAR. Exploring a broader range of tasks could therefore help identify the situations in which the additional complexity of more advanced annotation techniques provides tangible benefits.

Acknowledgments

The authors wish to thank Leo Paul Huar for his assistance in developing the initial prototype of SlidAR and DepthAR. We also thank all participants in our study. This work was supported in part by a grant from Snap Inc.

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 ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*. 235–242. doi:10.1145/2534329.2534331

[2] 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 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3313831.3376550

[3] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies* 4, 3 (may 2009), 114–123.

[4] István Barakonyi, Tamer Fahmy, and Dieter Schmalstieg. 2004. Remote Collaboration Using Augmented Reality Videoconferencing. In *Graphics Interface*. 89–96.

[5] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* 189 (11 1995).

[6] Eunhee Chang, Yongjae Lee, Mark Billinghurst, and Byounghyun Yoo. 2024. Efficient VR-AR communication method using virtual replicas in XR remote collaboration. *International Journal of Human-Computer Studies* 190 (2024), 103304. doi:10.1016/j.ijhcs.2024.103304

[7] Georgios Chantziaras, Andreas Triantafyllidis, Aristotelis Papaprodromou, Ioannis Chatzikonstantinou, Dimitrios Giakoumis, Athanasios Tsakiris, Konstantinos Votis, and Dimitrios Tzovaras. 2021. An Augmented Reality-Based Remote Collaboration Platform for Worker Assistance. In *Proceedings of Pattern Recognition. ICPR International Workshops and Challenges*. 404–416. doi:10.1007/978-3-030-68787-8_30

[8] Henry Chen, Austin S. Lee, Mark Swift, and John C. Tang. 2015. 3D Collaboration Method over HoloLens™ and Skype™ End Points. In *Proceedings of the 3rd International Workshop on Immersive Media Experiences*. 27–30. doi:10.1145/2814347.2814350

[9] Sung Ho Choi, Minseok Kim, and Jae Yeol Lee. 2018. Situation-dependent remote AR collaborations: Image-based collaboration using a 3D perspective map and live video-based collaboration with a synchronized VR mode. *Computers in Industry* 101 (2018), 51–66. doi:10.1016/j.compind.2018.06.006

[10] Varunyu Fuvattanasilp, Yuichiro Fujimoto, Alexander Plopski, Takafumi Taketomi, Christian Sandor, Masayuki Kanbara, and Hirokazu Kato. 2021. SlidAR+: Gravity-aware 3D object manipulation for handheld augmented reality. *Computers & Graphics* 95 (2021), 23–35. doi:10.1016/j.cag.2021.01.005

[11] Lei Gao, Huidong Bai, Weiping He, Mark Billinghurst, and Robert W. Lindeman. 2018. Real-time visual representations for mobile mixed reality remote collaboration. In *SIGGRAPH Asia Virtual & Augmented Reality*. Article 15, 2 pages. doi:10.1145/3275495.3275515

[12] Lei Gao, Huidong Bai, Gun Lee, and Mark Billinghurst. 2016. An oriented point-cloud view for MR remote collaboration. In *SIGGRAPH ASIA Mobile Graphics and Interactive Applications*. Article 8, 4 pages. doi:10.1145/2999508.2999531

[13] Lei Gao, Huidong Bai, Rob Lindeman, and Mark Billinghurst. 2017. Static local environment capturing and sharing for MR remote collaboration. In *SIGGRAPH Asia Mobile Graphics & Interactive Applications*. Article 17, 6 pages. doi:10.1145/3132787.3139204

[14] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. In touch with the remote world: remote collaboration with augmented reality drawings and virtual navigation. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. 197–205. doi:10.1145/2671015.2671016

[15] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 449––459. doi:10.1145/2642918.2647372

[16] Ryan K. Ghamandi, Yahya Hmaiti, Tam T. Nguyen, Amirpouya Ghasemaghaei, Ravi Kiran Kattoju, Eugene M. Taranta, and Joseph J. LaViola. 2023. What And How Together: A Taxonomy On 30 Years Of Collaborative Human-Centered XR Tasks. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 322–335. doi:10.1109/ISMAR59233.2023.00047

[17] Ryan Khushan Ghamandi, Ravi Kiran Kattoju, Yahya Hmaiti, Mykola Maslych, Eugene Matthew Taranta, Ryan P. McMahan, and Joseph LaViola. 2024. Unlocking Understanding: An Investigation of Multimodal Communication in Virtual Reality Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 602, 16 pages. doi:10.1145/3613904.3642491

[18] Leonardo Giusti, Kotval Xerxes, Amelia Schladow, Nicholas Wallen, Francis Zane, and Federico Casalegno. 2012. Workspace configurations: setting the stage for remote collaboration on physical tasks. In *Proceedings of the Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. 351–360. doi:10.1145/2399016.2399071

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

[20] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*. Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[21] Xincheng Huang, Dieter Frehlich, Ziyi Xia, Peyman Gholami, and Robert Xiao. 2025. GaussianNexus: Room-Scale Real-Time AR/VR Telepresence with Gaussian Splatting. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 189, 18 pages. doi:10.1145/3746059.3747693

[22] Xincheng Huang, Michael Yin, Ziyi Xia, and Robert Xiao. 2024. VirtualNexus: Enhancing 360-Degree Video AR/VR Collaboration with Environment Cutouts and Virtual Replicas. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. Article 55, 12 pages. doi:10.1145/3654777.3676377

[23] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. eyemR-Vis: A Mixed Reality System to Visualise Bi-Directional Gaze Behavioural Cues Between Remote Collaborators. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Article 188, 4 pages. doi:10.1145/3411763.3451545

[24] Janet G Johnson, Tommy Sharkey, Iramuali Cynthia Butarbutar, Danica Xiong, Ruijie Huang, Lauren Sy, and Nadir Weibel. 2023. UnMapped: Leveraging Experts' Situated Experiences to Ease Remote Guidance in Collaborative Mixed Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 878, 20 pages. doi:10.1145/3544548.3581444

[25] Steven Johnson, Madeleine Gibson, and Bilge Mutlu. 2015. Handheld or Handsfree? Remote Collaboration via Lightweight Head-Mounted Displays and Handheld Devices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1825–1836. doi:10.1145/2675133.2675176

[26] Jinki Jung, Jihye Hong, Sungheon Park, and Hyun S. Yang. 2012. Smartphone as an augmented reality authoring tool via multi-touch based 3D interaction method. In *Proceedings of the ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*. 17–20. doi:10.1145/2407516.2407520

[27] Seungwon Kim, Mark Billinghurst, and Gun Lee. 2018. The effect of collaboration styles and view independence on video-mediated remote collaboration. *Computer Supported Cooperative Work (CSCW)* 27, 3 (2018), 569–607.

[28] Seungwon Kim, Gun A. Lee, and Nobuchika Sakata. 2013. Comparing pointing and drawing for remote collaboration. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. 1–6. doi:10.1109/ISMAR.2013.6671833

[29] Takeshi Kurata, Nobuchika Sakata, M. Kourogi, Hideaki Kuzuoka, and Mark Billinghurst. 2004. Remote collaboration using a shoulder-worn active camera/laser. In *Proceedings of the International Symposium on Wearable Computers*, Vol. 1. 62– 69. doi:10.1109/ISWC.2004.37

[30] Chengyuan Lin, Edgar Rojas-Munoz, Maria Eugenia Cabrera, Natalia Sanchez-Tamayo, Daniel Andersen, Voicu Popescu, Juan Antonio Barragan Noguera, Ben Zarzaur, Pat Murphy, Kathryn Anderson, et al. 2020. How about the mentor? effective workspace visualization in ar telementoring. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. 212–220.

[31] Thomas Ludwig, Oliver Stickel, Peter Tolmie, and Malte Sellmer. 2021. shARe-IT: Ad hoc Remote Troubleshooting through Augmented Reality. *Computer Supported Cooperative Work (CSCW)* 30, 1 (2021), 119–167. doi:10.1007/s10606-021-09393-5

[32] Mathieu Lutfallah, Tamara Gini, Christian Hirt, Kordian Caplazi, and Andreas Kunz. 2023. Remote Cross-platform Instructions between MR and VR. In *2023 International Conference on Cyberworlds (CW)*. 256–259. doi:10.1109/CW58918.2023.00045

[33] Rafael Maio, André Santos, Bernardo Marques, Duarte Almeida, Paulo Dias, and Beatriz Sousa-Santos. 2022. Pervasive Augmented Reality (AR) for Assistive Production: Comparing the use of a Head-Mounted Display (HMD) versus a Hand-Held Device (HHD). In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 279–281. doi:10.1145/3568444.3570591

[34] Bernardo Marques, Joao Barroso, Paulo Dias, and Beatriz Santos. 2021. Exploring MR for Remote Collaboration: Comparing a 3D shared model approach width 2D annotations. In *Proceedings of the International Conference on Graphics and Interaction*. 1–4.

[35] Peter Mohr, Shohei Mori, Tobias Langlotz, Bruce H. Thomas, Dieter Schmalstieg, and Denis Kalkofen. 2020. Mixed Reality Light Fields for Interactive Remote Assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3313831.3376289

[36] Benjamin Nuernberger, Kuo-Chin Lien, Tobias Höllerer, and Matthew Turk. 2016. Interpreting 2D gesture annotations in 3D augmented reality. In *Proceedings of the IEEE Symposium on 3D User Interfaces*. 149–158. doi:10.1109/3DUI.2016.7460046

[37] 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 Annual Symposium on User Interface Software and Technology*.

741–754.

[38] Jana Pejoska, Sanna Reponen, Marjo Virnes, and Teemu Leinonen. 2017. Mobile augmented communication for remote collaboration in a physical work context. *Australasian Journal of Educational Technology* 33 (11 2017). doi:10.14742/ajet.3622

[39] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3173574.3173620

[40] Alexander Plopski, Varunyu Fuvattanasilp, Jarkko Polvi, Takafumi Taketomi, Christian Sandor, and Hirokazu Kato. 2018. Efficient in-situ creation of augmented reality tutorials. In *Proceedings of the Workshop on Metrology for Industry 4.0 and IoT*. IEEE, 7–11.

[41] Jarkko Polvi, Takafumi Taketomi, Goshiro Yamamoto, Arindam Dey, Christian Sandor, and Hirokazu Kato. 2016. SlidAR: A 3D positioning method for SLAM-based handheld augmented reality. *Computers & Graphics* 55 (2016), 33–43. doi:10.1016/j.cag.2015.10.013

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. 28492–28518 pages.

[43] Troels Rasmussen, Tiare Feuchtner, Weidong Huang, and Kaj Grønbæk. 2022. Supporting workspace awareness in remote assistance through a flexible multi-camera system and Augmented Reality awareness cues. *Journal of Visual Communication and Image Representation* 89 (2022), 103655. doi:10.1016/j.jvcir.2022.103655

[44] Troels Ammitsbøl Rasmussen and Weidong Huang. 2019. SceneCam: Improving Multi-camera Remote Collaboration using Augmented Reality. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct*. 28–33. doi:10.1109/ISMAR-Adjunct.2019.00023

[45] Bektur Ryskeldiev, Michael Cohen, and Jens Herder. 2018. StreamSpace: Pervasive Mixed Reality Telepresence for Remote Collaboration on Mobile Devices. *Journal of Information Processing* 26 (2018), 177–185. https://api.semanticscholar.org/CorpusID:4566209

[46] Mose Sakashita, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew David Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View-dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 675, 14 pages. doi:10.1145/3613904.3642945

[47] Mickael Sereno, Xiyao Wang, Lonni Besançon, Michael J Mcguffin, and Tobias Isenberg. 2020. Collaborative work in augmented reality: A survey. *IEEE Transactions on Visualization and Computer Graphics* 28, 6 (2020), 2530–2549.

[48] Yuxin Shen, Wei Liang, and Jianzhu Ma. 2025. LiteAT: A Data-Lightweight and User-Adaptive VR Telepresence System for Remote Education. *IEEE Transactions on Visualization and Computer Graphics* 31, 11 (2025), 9570–9580. doi:10.1109/TVCG.2025.3616747

[49] Matthew Tait and Mark Billinghurst. 2015. The Effect of View Independence in a Collaborative AR System. *Computer Supported Cooperative Work* 24, 6 (2015), 563–589. doi:10.1007/s10606-015-9231-8

[50] 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 *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. Article 23, 11 pages. doi:10.1145/3359996.3364238

[51] 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 CHI Conference on Human Factors in Computing Systems*. 1–14. doi:10.1145/3290605.3300431

[52] Theophilus Teo, Mitchell Norman, Gun Lee, Mark Billinghurst, and Matt Adcock. 2020. Exploring interaction techniques for 360 panoramas inside a 3D reconstructed scene for mixed reality remote collaboration. *Journal on Multimodal User Interfaces* 14 (2020), 373–385. doi:10.1007/s12193-020-00343-x

[53] 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 Annual ACM Symposium on User Interface Software and Technology*. 161–174. doi:10.1145/3332165.3347872

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

[55] Eduardo Veas, Alessandro Mulloni, Ernst Kruijff, Holger Regenbrecht, and Dieter Schmalstieg. 2010. Techniques for view transition in multi-camera outdoor environments. In *Proceedings of Graphics Interface*. 193–200.

[56] Wolfgang Vorraber, Johannes Gasser, Helena Webb, Dietmar Neubacher, and Philipp Url. 2020. Assessing augmented reality in production: remote-assisted maintenance with HoloLens. *Procedia CIRP* 88 (2020), 139–144. doi:10.1016/j.procir.2020.05.025

[57] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Sili Wei, Guangyao Xu, Weiping He, Xiangyu Zhang, and Jie Zhang. 2021. 3DGAM: using 3D gesture and CAD models for training on mixed reality remote collaboration. *Multimedia Tools and Applications* 80 (2021), 31059–31084.

[58] Peng Wang, Xiaoliang Bai, Mark Billinghurst, Shusheng Zhang, Xiangyu Zhang, Shuxia Wang, Weiping He, Yuxiang Yan, and Hongyu Ji. 2021. AR/MR remote collaboration on physical tasks: a review. *Robotics and Computer-Integrated Manufacturing* 72 (2021), 102071. doi:10.1016/j.rcim.2020.102071

[59] Jianing Yin, Weicheng Zheng, Yuntao Wang, Xin Tong, and Yukang Yan. 2025. A Comparison Study Understanding the Impact of Mixed Reality Collaboration on Sense of Co-Presence. In *Proceedings of the IEEE Conference Virtual Reality and 3D User Interfaces*. 580–590. doi:10.1109/VR59515.2025.00081

[60] Jacob Young, Tobias Langlotz, Matthew Cook, Steven Mills, and Holger Regenbrecht. 2019. Immersive Telepresence and Remote Collaboration using Mobile and Wearable Devices. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (2019), 1908–1918. doi:10.1109/TVCG.2019.2898737

[61] Jacob Young, Tobias Langlotz, Steven Mills, and Holger Regenbrecht. 2020. Mobileportation: Nomadic telepresence for mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–16. doi:10.1145/3397331

[62] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth. 2021. Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence. *IEEE Transactions on Visualization & Computer Graphics* 27, 11 (2021), 4129–4139. doi:10.1109/TVCG.2021.3106480

[63] Seunghyeon Yu and Jae Yeol Lee. 2025. Clip, Copy, and Share: On-the-Fly Gaussian Splatting-Based Virtual Replicas for Remote XR Collaboration. *IEEE Access* 13 (2025), 169852–169865. doi:10.1109/ACCESS.2025.3614220

[64] Faisal Zaman, Craig Anslow, and Taehyun James Rhee. 2023. Vicarious: Context-aware Viewpoints Selection for Mixed Reality Collaboration. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand). Article 23, 11 pages. doi:10.1145/3611659.3615709

[65] Xiangyu Zhang, Xiaoliang Bai, Shusheng Zhang, Weiping He, Peng Wang, Zhuo Wang, Yuxiang Yan, and Quan Yu. 2022. Real-time 3D video-based MR remote collaboration using gesture cues and virtual replicas. *The International Journal of Advanced Manufacturing Technology* 121 (08 2022), 1–23. doi:10.1007/s00170-022-09654-7

[66] Xiangyu Zhang, Xiaoliang Bai, Shusheng Zhang, Weiping He, Shuxia Wang, Yuxiang Yan, Peng Wang, and Liwei Liu. 2024. A novel mixed reality remote collaboration system with adaptive generation of instructions. *Computers & Industrial Engineering* 194 (2024), 110353. doi:10.1016/j.cie.2024.110353