Adaptive Visual Cues for Guiding a Bimanual Unordered Task in Virtual Reality

Jen-Shuo Liu *†
Department of Computer Science
Columbia University

Portia Wang *‡
Department of Computer Science
Columbia University

Barbara Tversky§
Department of Human Development
Teachers College
Columbia University

Steven Feiner¶
Department of Computer Science
Columbia University

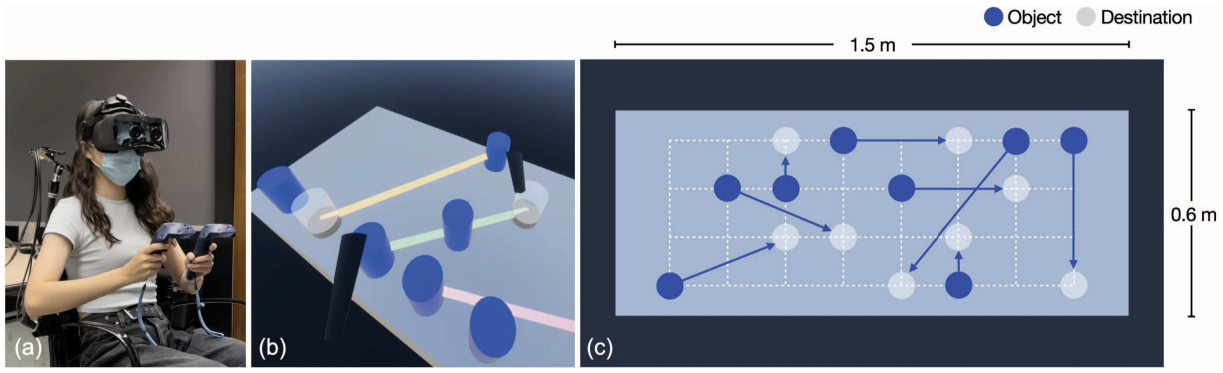

Figure 1: Our bimanual unordered task. (a) A user wearing a Varjo XR-3 headset and holding a Vive controller in each hand. (b) Example participant view in VR of our task with three cues shown, each of which is a colored line that connects a cylindrical object to a semitransparent copy at its destination. (c) Example top-down view of study task setup. The testbed contains eight cylindrical objects on a platform, where the initial position and goal position for each object are randomized to unique locations in an eight-by-four grid on the platform.

Abstract

Work on cueing performance in AR and VR has focused on sequential tasks in which each step must be completed in order before the user can proceed to the next. However, for unordered tasks such as putting books back on a library shelf, the user may be able to perform multiple steps concurrently without needing to follow a specific order. In such situations, giving the user multiple cues for potentially concurrent steps may improve performance time. To investigate this, we built a bimanual VR testbed in which the user needs to move objects to designated destinations, guided by different numbers of cues. The user can decide the order to perform the cued steps and, in some conditions, can affect which cues are shown.

In a formal user study, we found that in most conditions, participants perform fastest with three cues. Dynamically updating the set of displayed cues based on hand proximity improves performance, and updating the set based on eye gaze improves performance even more. Finally, for both the hand-proximity and eye-gaze mechanisms, performance can be further improved by locking the cues for objects predicted to be moved next based on hand distance.

Index Terms: Human-centered computing—Human-centered computing (HCI)—Interaction paradigms—Virtual reality; Human-centered computing—Human-centered computing (HCI)—HCI design and evaluation methods—User studies

*Equal contribution.
†e-mail: jl5004@columbia.edu
‡e-mail: pw2491@columbia.edu
§e-mail: bt2158@tc.columbia.edu
¶e-mail: feiner@cs.columbia.edu

1 Introduction

In many real-world tasks, visual or textual instructions direct users to proceed in a fixed sequence, one step at a time, without regard to spatial context or user preferences. For these tasks, different ways to prompt information about the current next step (a *cue*) have been explored using virtual reality (VR) and augmented reality (AR). Guidance systems showing a single cue at a time for sequential tasks have used a variety of visual representations [8, 11, 36] and have leveraged a user's eye gaze [4] and kinematic patterns [53] to take into account the user's intention.

Recent research has examined the effects and benefits of prompting information about future steps (*precues*) for sequential tasks. Hertzum and Hornbæk [19] studied the effect of providing information for one future step with desktop 2D touchpad/mouse input. Volmer et al. [50] explored the effect of giving information about one future step in projector-based spatial AR. Liu et al. [29] studied the consequences of precueing multiple future steps in VR. These tasks are strictly sequential. However, many other real-world tasks, such as putting library books on a shelf or sorting items in a warehouse, are not strictly sequential. The steps in such tasks can be completed in different orders, and the user is able to move multiple task objects simultaneously, potentially using both hands.

To investigate this, we developed a VR testbed, shown in Figure 1, that supports a pick-and-place task in which a user needs to follow visual cues to move multiple objects to their destinations. Using this testbed, we study how the number of cues affects performance and how hand proximity and eye gaze can be used to update the set of displayed cues. We make three contributions:

- We present a VR testbed for exploring adaptive visual cues for guiding a bimanual concurrent unordered task.
- We show through a user study that compared to statically displayed cues, using hand proximity to dynamically update the set of displayed cues can decrease task completion time, while using eye gaze can further shorten it.

978-1-6654-5325-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ISMAR55827.2022.00059

- We show that when using dynamic cue-set–updating mechanisms, performance can be further improved by locking cues for objects predicted to be moved next based on hand distance.

2 Related Work

2.1 Task Guidance Systems

Task guidance systems are typically built to either teach users skills that can be later applied in real-world scenarios or directly assist users during a task. As users become more familiar with these systems, they may also learn to leverage guidance cues to complete similar tasks. For example, in the VR game Beat Saber [5], players become better at following sequences of cues shown during gameplay, slashing multiple beats at the same time using both controllers.

Research on prompting information for sequential tasks has investigated the benefits of using virtual avatars [4, 11, 22], graphical and textual annotations [8, 15, 22, 29], and virtual proxies [36]. Researchers have also built adaptive systems [22, 28]. For example, Huang et al. [22] varied the level of detail shown in an AR tutorial based on each user's characteristics and tutorial-following status, while Lindlbauer et al. [28] adapted the amount of information displayed in AR and VR based on an individual's real-time cognitive load across context switches. Much work has also targeted assisting users in equipment assembly/disassembly, repair, inspection, diagnosis, and operation [39]. In contrast to this work, we explore a generic task that involves moving objects to destinations. Similar to the adaptive systems mentioned above, we compare different approaches that leverage eye gaze and hand proximity to update the information displayed.

2.2 Cueing Multiple Steps

Prior work has demonstrated the advantages of prompting information about future steps in a sequential task. Hertzum and Hornbæk [19] studied the effect of showing cues for the current step and the next step with desktop 2D touchpad/mouse input. In their study, participants were asked to alternately tap/click a center target and one of a circular set of surrounding targets. They showed that participants moved faster to the single precued center target than to surrounding targets. Volmer et al. [50] studied the effect of showing information about the next step using a cue in a projector-based AR environment and showed that visualizing one more step improves user performance. Volmer et al. [49] next examined the performance of sleep-deprived users in that task. They showed that users can still benefit from a cue prompting the next step. Liu et al. [29] investigated the effect of using cues to show information about multiple future steps in a VR path-following task and showed that people could use two to three future cues if the cues contained lines and only one future cue if the cues did not contain lines. Later, Liu et al. [30, 31] considered a compound sequential task in which the users need to pick up an item, move and rotate it, and deposit it at a specified destination in each step.

Though many have studied prompting with multiple cues in sequential tasks, our understanding of how to guide users in tasks allowing multiple steps to be performed concurrently remains limited. Systems for such tasks have focused on context switching between unrelated or sparsely related tasks and performing a single task at a time [2, 7, 28] or handling multiple unrelated physical objects at the same time [38]. Though many mundane tasks are indeed well captured by these systems, popular time-sensitive games such as *Overcooked! 2* [45] and *Moving Out* [13] remind us that one approach to faster performance is to simultaneously have multiple players each tackle one step at a time. While many real-world tasks are performed by one person, these tasks often allow for concurrent steps to be performed using two hands. Thus, drawing inspiration from time-sensitive games and in contrast to prior studies, we investigate ways of showing multiple cues for an unordered task in which a user can accomplish two steps simultaneously.

Our motivation is similar to that of Illing et al. [24], who investigate how varying amounts of visual assistance for parallel tasks affect performance in tablet-based AR. In their system, the user was presented with a set of parallel tasks, each containing a set of sequential steps. The number of cued tasks was determined by the system and only the next step of each of these tasks was cued. In contrast, our work compares different adaptive approaches for dynamically determining the subset of steps that are cued. More specifically, we explore the number of steps to show, and how to select and update the set of displayed cues.

2.3 Predicting Actions Through Eye Gaze and Hand Movement

Understanding intention from a user's behavior is important to predicting behavior and may be used to cue performance. Becchio et al. [6] noted that a person's intention is revealed in the kinematics of arm and hand actions, and those movements can be used for action prediction. Other research has shown that eye gaze reveals the intention of future actions and precedes spoken requests [21], and should be carefully studied within the context of a particular task [32]. Building on these findings, more recent work leveraged gaze interaction by proposing an AR space that allows for gaze-mediated control [40] and compared user preference for and speed of a gaze-adaptive AR interface with an always-on AR interface [41].

Many have studied the correspondence between eye gaze and hand movements. Early research investigated the temporal relationship between eye and hand movements by tracking cursor movements in a graphical user interface [10, 23, 42] and found that eye gaze often leads mouse movement. More recently, Mutasim et al. [35] concluded that a user's gaze in their VR task reaches the target before their hands touch it. Work that extends beyond these findings has also leveraged this correlation in predicting and preventing erroneous actions before the user's hand reaches the incorrect target [53] and predicting indecision by using the distance between the finger and the eyes for a tablet memory game [51]. Our work builds on these previous findings of inferring intention with kinematic patterns and hand-eye correspondence in manual tasks by testing different dynamic mechanisms that update the displayed set of cued steps based on the user's eye gaze and hand positions.

3 Visual Cues for Task Guidance

3.1 Testbed and Task

Our goal is to investigate how visual cues affect user performance in a bimanual concurrent unordered task. To address this, we developed a VR testbed using Unity 2020.3.11f1 [47] and a task that involves moving objects to specified destinations. Our testbed contains a 0.6m by 1.5m virtual platform in the *xz*-plane on which there are eight cylindrical objects (4cm radius, 5cm height) constrained to be upright at all times. The initial positions and destination positions for the objects are randomized to unique locations in an eight-by-four grid on the platform. Our task is comprised of eight steps. In each step, the participant (Figure 1a) must move one of their controllers to an object, press the trigger button to grab the object, move the controller and object to a specified location, and release the trigger. The user may move two objects at the same time, one per hand.

Once a task object is away from its destination by less than 2cm on the *xz*-plane and 3cm on the *y*-axis, the task object is snapped to its destination, turns dark gray to signify step completion, and can no longer be moved. Each controller vibrates for 0.1s when it completes a step or comes into contact with an object that is not at its destination. Note that since our task is not sequential, the user can decide the order in which to move cued objects. The task is considered complete when all objects are moved to their respective destinations. Figure 1(b) shows an example of a participant's view in VR and Figure 1(c) shows an example of a top-down schematic view of the task setup.

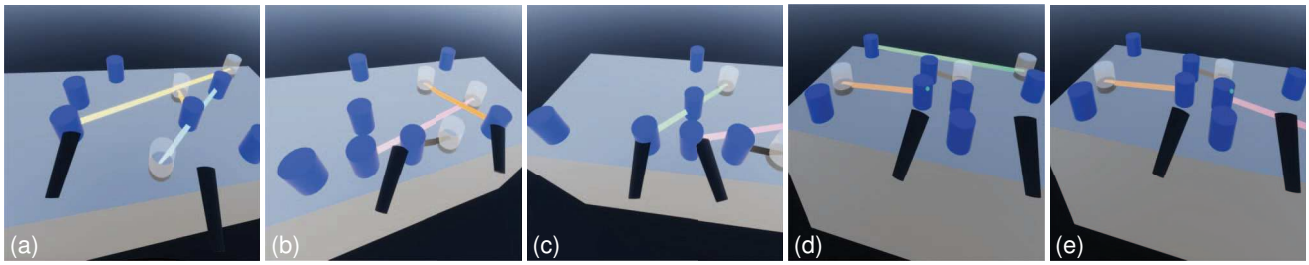

Figure 2: Cue-set–updating mechanisms. (a) Static mechanism. The order in which cues are displayed is predetermined and will not change. (b–c) Hand-Proximity mechanism. In this sequence, the cues are updated after both controllers are moved to the left. (d–e) Eye-Gaze mechanism. The cyan dot indicates gaze position. In this sequence, the user's gaze moves to the right and hits a new object, its corresponding pink cue appears, and the green cue for the object that is least recently looked at disappears.

3.2 Cues

3.2.1 Cue Visualization

We define *cues* as visualizations that provide information about actionable steps to guide users to perform them. This differs from the definitions used in previous work, which distinguished a cue for the current step and one or more predictive cues [49, 50] (or precues [19,29–31]) for the future steps. Since our task is unordered, the user can decide the order in which the actionable steps are performed. Depending on the order the user performs the steps, each of our cue visualizations can be a cue or a precue under the definitions used in previous work. However, we decided not to make these distinctions since the order is entirely up to the user and can change as they proceed.

Each of our cues consists of a colored line connecting the object to a semitransparent replica of the object that has an additional opaque cylindrical base. The combination of the colored line and the cylindrical base is similar to the CircleLine visualization in Liu et al. [29]. The line guides the user from the object to its destination, while the cylindrical base highlights the place of the destination. Based on an early pilot study, we found that the combination of the semitransparent replica and its opaque base helped discourage the user from grabbing the destination by mistake. Users made a considerable number of errors during early pilot studies when the cue lines were rendered with the same color, especially when there were a significant number of crossed lines. To address this, we used a color-vision–deficiency friendly palette proposed by Okabe and Ito [37], and ensured that each object would always receive the same color within a trial.

3.2.2 Number of Cues

We tested different numbers of displayed cues across different conditions. In a condition where n cues are used, the testbed shows either a set of n cues when there are at least n unfinished steps in the task, or cues to all unfinished steps for the last $n-1$ steps of a task. We did not insert additional objects and cues in the last $n-1$ steps of a task to maintain the number of cues shown to the user, since our focus was on studying how long the user took to finish moving the given set of objects. If we had included additional objects and cues in these last $n-1$ steps, the user could have attempted to move some of these objects, potentially changing the task difficulty.

3.2.3 Cue-Set–Updating Mechanisms

The cue-set–updating mechanisms determine *which* cues are provided. As research has shown that a user's eye-gaze direction and hand positions can indicate intention [6, 21, 32], we factored eye gaze and hand positions into determining which cues to show and developed three mechanisms:

1. **Static mechanism** (Figure 2a). In this baseline condition, the ordering of the cues shown is updated only when one of the previously shown steps is completed. We achieve this by generating a fixed ordering of the steps at the start of each task and always displaying cues to the first n unfinished steps.

2. **Hand-Proximity mechanism** (Figure 2b–c). Each unfinished task is ranked based on its distance to its closest controller and the cues to the n steps with the closest distances are always shown.

3. **Eye-Gaze mechanism** (Figure 2d–e). At the start of a task, the displayed set of cues is identical to that of the Static mechanism. Then, whenever the user looks at an object whose step has not yet been finished (their eye-gaze direction intersects the object), the cue for that object is added to the set of displayed cues, and the oldest cue in the set gets removed. Thus, the set of displayed cues always includes cues for the n most recently looked at objects.

Though task-specific heuristics could be applied to improve the ordering of steps in the Static mechanism, we focus on leveraging the inferred intention from user behavior to circumvent the implementation of task-specific heuristics by proposing a set of mechanisms that will be generalizable to other task domains.

3.2.4 Locking

We observed through our pilot studies that a user's rapidly changing eye movements or hand positions often introduce unwanted instability in the set of cues shown. Users expressed frustration when their moving gaze or hands unintentionally caused the cue for the step on which they were working to leave the set of visible cues.

To address this, we introduce *locking*, which pauses cue-set updates for objects that are predicted to be moved next based on hand distance. When an object is added to the cue set, we record that object's distances to each of the controllers. Once the dynamically calculated distance between this cued object and one of the controllers falls below the initially recorded distance (for that controller) multiplied by 0.5, that object's cue is "locked" in the displayed set of cues until its step is completed. Based on pilot studies, we chose the multiplier 0.5 so that the cue-set updates would not occur so frequently as to distract the user, while still offering the benefits of dynamically updating the displayed cue-set based on the user's eye gaze and hand proximity. Figure 3 shows an example of locking.

4 User Study

4.1 Pilot Studies

We conducted pilot studies to test different cue-set–updating mechanisms. In addition to the Static, Hand-Proximity, and Eye-Gaze mechanisms described in Section 3.2.3, we also tried to use headset orientation (Head-Gaze), as it partially indicates the direction in which the user is looking [4] but is less sensitive than Eye-Gaze.

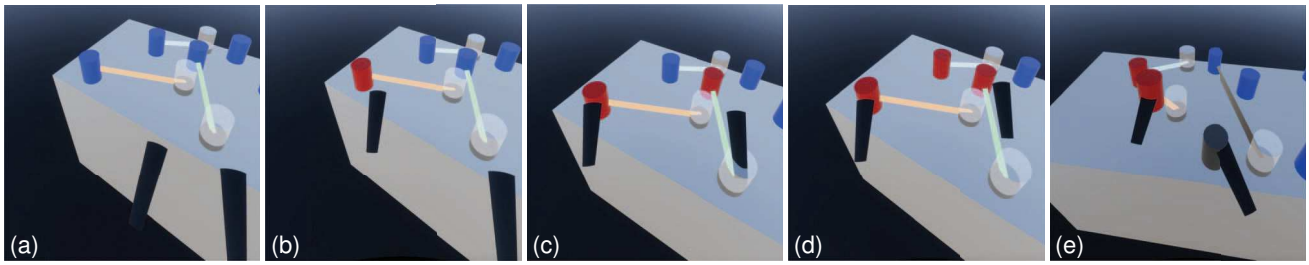

Figure 3: Locking for dynamic cue-set–updating mechanisms. *Red cylinders visualize objects with locked cues in this figure, but are blue in the testbed.* (a) Three cues initialized based on cue-set–updating mechanism (Eye-Gaze in this case). (b) Left controller moves closer to a cued object, locking its cue. (c) Right controller moves closer to another object, locking its cue. (d) Right controller continues moving such that the cue for an additional object also gets locked. (e) Right controller picks up the rightmost locked object in (d) and moves it to its destination. Upon completing this step, a new cue is displayed.

Table 1: Average task completion time in a pilot study.

	S	H	H-L	E	E-L
2 cues	12.674	11.894	11.433	11.502	11.450
3 cues	12.640	11.090	10.044	10.182	9.810
4 cues	12.970	11.640	10.747	10.838	11.258
6 cues	11.700				
8 cues	11.214				

The results showed that participants' performance with Head-Gaze was similar to but worse than Eye-Gaze. Therefore, we decided not to test the Head-Gaze mechanism in the formal study.

Our earlier pilot studies on cue-set–updating mechanisms did not use locking. User feedback about the abruptness with which cues disappeared and reappeared in adaptive cue-set–updating mechanisms encouraged us to address this. Observing how hand pose can indicate how certain a user can be of their intentions and can predict future actions, after several rounds of refinement, we came up with the locking technique discussed in Section 3.2.4.

In another pilot study with three participants, we included five approaches for updating the set of displayed cues: Static (S), Hand-Proximity without Locking (H), Hand-Proximity with Locking (H-L), Eye-Gaze without Locking (E), and Eye-Gaze with Locking (E-L). For each of these approaches, we tested two, three, and four cues. For S, we also tested six and eight cues. The average task completion times are listed in Table 1. Generally speaking, participants did not perform better when given more than four cues and performed best with three cues. The only exception was S, where participants performed best with six or eight cues. However, they performed worse than with H, H-L, E, and E-L with three cues.

4.2 Hypotheses

We formulated three hypotheses regarding the number of cues, the cue-set–updating mechanism, and locking:

H1. *Task completion time will be faster with three cues than with two cues or four cues.* Previous work [19, 29, 49, 50] has shown that providing additional cues for future steps in sequential tasks can shorten task completion time, since the user can prepare for future steps while working on the current one. In our task, the user can simultaneously work on up to two steps, guided by two cues. We hypothesize that adding a third cue can help the user prepare for the future and improve their performance (shortening task completion time). However, based on our pilot studies, adding a fourth cue could make the scene too cluttered and increase task completion time.

H2. *Eye-gaze approaches with or without locking will result in faster task completion time than hand-proximity approaches with or without locking, which will be faster than the static approach.* Hand and eye actions can indicate user intention [6, 21, 32], so incorporating them into the cue-set–updating mechanism could reduce task completion time. This appeared to be the case in our pilot studies. In addition, our pilot studies suggested that the eye-gaze approaches performed better than the hand-proximity approaches.

H3. *Locking will reduce task completion time for both the Eye-Gaze and Hand-Proximity mechanisms.* In our pilot studies, we found that without locking, cues sometimes disappeared before a participant successfully grabbed an object. This confused participants and slowed task progress. Locking (Section 3.2.4) was developed to avoid this situation, so we hypothesize that Eye-Gaze with Locking will be faster than Eye-Gaze without Locking, and Hand-Proxmity with Locking will be faster than Hand-Proximity without Locking.

4.3 Methods

4.3.1 Participants

Our study was approved by our institutional review board. We recruited 15 participants from our institution (7 female), 21–33 years old (average 23.3), through convenience sampling using department email lists and posted flyers. Two participants are left-handed and one is ambidextrous. One participant owns a VR headset, two had used VR in class projects, nine had used AR/VR several times, and three had no AR/VR experience. Each participant received a USD 15 gift card.

4.3.2 Equipment

Each participant used a Varjo XR-3 headset [48] (with its pass-through cameras off) with a 115° horizontal (134° diagonal) field of view and a 90Hz refresh rate. The XR-3 was run in its outside-in tracking mode and was tracked with four HTC SteamVR Base Station 2.0 units. The headset ran on a computer powered by an Intel® Core™ i9-11900K Processor and an Nvidia GeForce RTX 3090 graphics card. Two Vive hand-held controllers were used to manipulate objects.

4.3.3 Study Design

Our user study aims to evaluate our hypotheses on how participants will perform in our bimanual unordered task. We include five approaches for updating the set of displayed cues: Static (S), Hand-Proximity without Locking (H), Hand-Proximity with Locking (H-L), Eye-Gaze without Locking (E), and Eye-Gaze with Locking (E-L). For each of these five approaches, we tested two, three, and four cues, since we found through our pilot studies (Section 4.1) that participants performed the best with the help of three cues in most cases. Thus, we tested 5 (approaches) $\times$ 3 (number of cues/approach) $=$ 15 conditions. In the remainder of this paper, we will refer to a condition by concatenating its approach name with

Table 2: Number of timed trials labeled as outliers in each condition summed over all participants.

	S	H	H-L	E	E-L
2 cues	3	4	1	2	1
3 cues	4	4	1	3	1
4 cues	1	2	5	1	2

Table 3: Average task completion time and standard deviation for each condition. Numbers in parentheses are standard deviations.

	2 cues	3 cues	4 cues
S	11.137 (2.163)	10.917 (1.806)	10.868 (2.136)
H	11.228 (2.315)	10.696 (1.906)	10.890 (2.072)
H-L	10.855 (2.139)	10.387 (2.353)	10.650 (2.215)
E	10.067 (1.867)	10.336 (2.335)	10.215 (2.119)
E-L	10.002 (2.174)	9.747 (2.086)	9.895 (2.020)

its number of cues. For example, H-L2 is Hand-Proximity with Locking and two cues.

For each condition, a participant performed a block of eight timed trials, preceded by an untimed practice trial, where each trial involved moving eight objects. Thus, each participant performed 15 (blocks) $\times$ 8 (timed trials/block) $=$ 120 timed trials. A five-second cooldown period during which no task was shown was added before each block and a three-second cooldown period was added before each timed trial. A timed trial began when the participant pressed the trigger buttons of both controllers at the same time.

We counterbalanced the order in which conditions were presented. To accomplish this, we grouped conditions first by their cue-set–updating mechanism (Static, Hand-Proximity, and Eye-Gaze), next by the number of cues (2, 3, and 4), and finally, for the two relevant mechanisms, by whether or not locking is used. We shuffled the order in which participants encountered the cue-set–updating mechanisms, the number of cues, and whether locking is used. For example, one participant might experience, in order, first the hand-proximity condition blocks (H-L4, H4), (H2, H-L2), (H3, H-L3), next the static condition blocks (S2), (S4), (S3), and finally the eye-gaze condition blocks (E3, E-L3), (E-L2, E2), (E-L4, E4).

To encourage participants to develop strategies based on different cue-set–updating mechanisms, participants were informed of the cue-set–updating mechanism for each block. However, the participants were not informed whether locking was used in a block. We chose to do this because we observed in pilot studies that locking could improve task performance regardless of whether participants knew it was being used.

We also wanted to ensure that the participants would not be able to memorize or recall specific task configurations, while having each configuration be used enough times for each condition (so that our linear mixed-effects model could factor out the impact of its difficulty). To do this, we generated 40 unique configurations of our task, each consisting of eight pairs (one per object) of unique initial and destination positions. These positions were randomly sampled from an eight-by-four grid on the platform. We then assigned each configuration to a trial such that every configuration was used roughly the same number of times across all conditions and each participant completed each configuration roughly the same number of times.

4.3.4 Procedure

Before each session, the headset, trackers, and table used in the study were sanitized with 70% isopropanol and the headset was also sanitized in a Cleanbox CX1 [12] UVC system. Each participant was then welcomed by the study coordinator and presented with an information sheet. After giving their consent, the participant was then introduced to the flow of the experiment and given the Stereo Optical Co. Inc. Stereo Fly Test [43] to screen for stereo vision and the Ishihara Pseudo-Isochromatic Plate test [25] to screen for color vision deficiencies. All participants passed both tests.

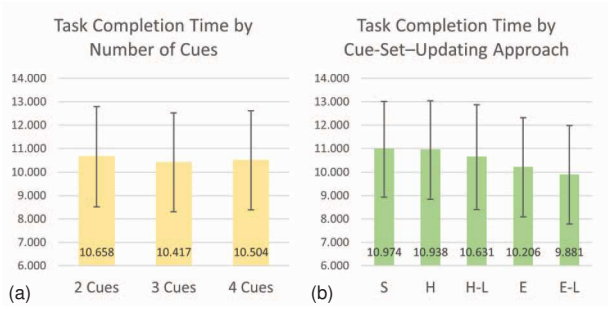

Figure 4: Average task completion time plotted by (a) number of cues and by (b) cue-set–updating approach.

Throughout the study, we recorded the timestamped eye-gaze direction for each eye, the status of each task object (location, task completeness and whether its cue is being displayed), as well as the position and orientation of the headset and hand-held controllers. We also recorded whether each controller was holding a task object. Over the session, the participant's interaction was monitored by the study coordinator through a separate desktop display.

After finishing all trials, the participant was asked to fill out a questionnaire. The questionnaire included questions on the participant's demographics, a modified unweighted NASA TLX [16], and a request to rank the different cue-set–updating mechanisms based on their effectiveness. Our TLX survey was modified to use a 1–7 scale, with 1 as best, rather than the original 0–20. Each participant rated each of the three cue-set–updating mechanisms (Static, Hand-Proximity, and Eye-Gaze) for each TLX metric. We decided to conduct the survey at the end of the study to avoid the concern that a participant's criteria for answering the questions might change between conditions. A session took about 60–70 minutes for a typical participant to complete.

4.4 Results

Before analyzing the results, we used Tukey's outlier filter [46] to label outliers. The "outside fence" for each condition and participant was computed separately, as we expected the conditions would have a significant effect on completion time, and we noticed that some participants performed substantially better than others. Applying Tukey's outlier filter, trials that took more than the third quartile plus 1.5 $\times$ interquartile range (third quartile minus first quartile) or less than the first quartile minus 1.5 $\times$ interquartile range were labeled as outliers. For each condition, there were 15 (participants) $\times$ 8 (trials/participant) $=$ 120 trials. Between 1 to 5 trials were labeled as outliers for each condition, as shown in Table 2, and were excluded from the analysis. Average task completion times after removing outliers and standard deviations for each condition are shown in Table 3. We also plot task completion time against number of cues and cue-set–updating approach in Figure 4.

We evaluated the hypotheses for significance with $\alpha = .05$. We fit a linear mixed-effects model to our data using the MATLAB Statistics, and Machine Learning Toolbox [33]. In the model, we used the task completion time as the measurement, the cue-set–updating approaches (S, H, H-L, E, or E-L) and the number of cues as the fixed-

effect variables, and the participant and the task configuration as the random effect variables. To make the comparison easier, we used the S approach and three cues as the baselines of the two fixed-effect variables. We added the interaction between the cue-set–updating approach and the number of cues in an alternative model and found most of the p-values of the interaction terms are not significant, so we decided to exclude that interaction from our model. Note that this means when we compare the cue-set–updating approaches, we are comparing their average performance across all numbers of cues rather than for a specific number of cues, and when we compare the number of cues, we are comparing the average performance across all cue-set–updating approaches rather than a specific approach. We also found that handedness and AR/VR experience are not significant factors. The effect sizes of this model are $\eta^2 = 0.441$ and Cohen's $d = 0.764$, which show large effects [9]. For the full details of the linear mixed-effects model and the alternative model, please see the supplementary material.

To evaluate **H1**, we checked the estimates and p-values of the two-cue and four-cue terms. The model shows that adding the third cue reduces task completion time by 0.255s relative to two cues ($p = .003$). While adding the fourth cue increases task completion time by 0.125s relative to three cues, its p-value is not significant ($p = .141$). Therefore, **H1** is partially supported insofar that adding the third cue improves performance.

To evaluate **H2**, we first check the p-values of H ($p = .701$), H-L ($p = .002$), E ($p < .001$) and E-L ($p < .001$) relative to S. This supports that H-L, E, and E-L (but not H) have faster task completion times than S. To further check if using eye gaze improves performance more than hand proximity, we test the contrasts between E and H, between E-L and H-L, between E and H-L, and between E-L and H. The p-values of these four comparisons are all $< .001$. This supports that all eye-gaze approaches result in faster task completion time than all hand-proximity approaches. Therefore, **H2** is mostly supported, except for the comparison between H and S.

To evaluate **H3**, we test the contrast between H-L and H and the contrast between E-L and E. The p-values of these two comparisons of task completion time are .008 and .023, respectively. This shows that using locking reduces task completion time for Eye-Gaze and Hand-Proximity mechanisms. Therefore, **H3** is supported.

To avoid type-I errors, we ran a correction using the Holm–Bonferroni method [20]. We checked a total of 12 p-values (2 for **H1**, 8 for **H2**, and 2 for **H3**) to validate our hypotheses and 10 of them are significant before correction. Among these 10 p-values, six are $< .001$, and the remaining four are .002, .003, .008, and .023 (lowest-to-highest). With this order, the p-values are smaller than $.05/10$, $.05/9$, ..., $.05/1$, respectively, meaning they survive their corresponding Holm–Bonferroni-corrected α.

Note that the comparison for **H1** is based on the average performance across all cue-set–updating approaches rather than a specific approach. Table 3 shows that for E, task completion time is fastest with two cues, while for S, it is fastest with four cues. We discuss possible causes for these two different trends in Section 5.1. The comparison for **H2** is based on average performance across all numbers of cues. The difference between E3 and H-L3 is fairly small. For the comparison for **H3**, the difference between E2 and E-L2 is also fairly small. Therefore, while using the Eye-Gaze mechanism vs. the Hand-Proximity mechanism, or using locking vs. not using locking, helps reduce task completion time in most cases, additional verification is needed to make sure it is the case for a specific number of cues.

4.4.1 User Feedback

NASA TLX results are shown in Figure 5. Friedman tests yielded $p_{MentalDemand} = .8717$, $p_{PhysicalDemand} = .5092$, $p_{TemporalDemand} = .1969$, $p_{Performance} = .1561$, $p_{Effort} = .2542$, and $p_{Frustration} = .1938$. The p-values for all metrics are $> .05$, so we did not find a significant difference between the cue-set–updating mechanisms.

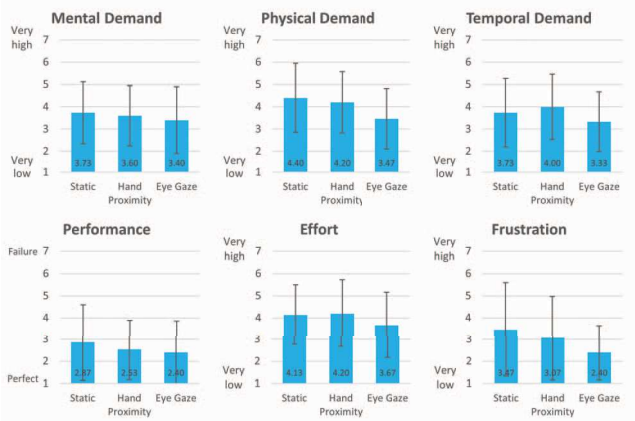

Figure 5: NASA TLX results.

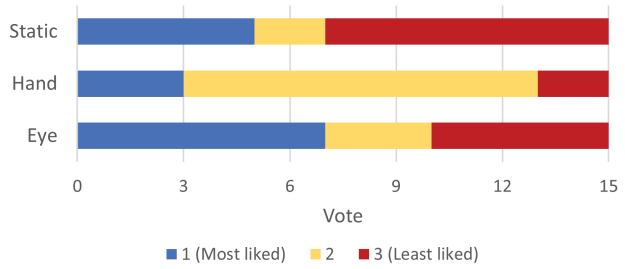

Figure 6: Preferences for cue-set–updating mechanisms.

The participants then ranked the cue-set–updating mechanisms based on their preferences. The results are shown in Figure 6. Participants most preferred the Eye-Gaze mechanism, followed by the Static mechanism, and finally the Hand Proximity mechanism. Participants were also asked how many cues they thought were the most useful: two participants answered two, eight participants answered three, and five participants answered four.

4.4.2 Error Rate

We looked into the errors participants made in each condition to understand how different conditions affect the error rates. We observed that participants made the following major types of errors:

Failed Grab: We considered that a Failed Grab error was made if a participant pressed the trigger when the controller was within 5cm of a cued object but had not collided with it.

Grab Destination: We considered that a Grab Destination error was made if a participant pressed a trigger when the controller collided with the target replica rather than the object.

Failed Deposit: We considered that a Failed Deposit error was made if a participant released the trigger and deposited the object within 5cm from the destination in the xz-plane but not within the 2cm threshold in the xz-plane mentioned in Section 3.1.

Wrong Destination: We considered that a Wrong Destination error was made if a participant deposited the object within 5cm from another object's destination in the xz-plane.

Table 4 shows the average number of errors per trial that participants made. To determine whether the differences were significant, we ran chi-square tests on each of these error types among different approaches. For Failed Grab and Grab Destination errors, there is no significant difference between any pair of approaches. The participants seldom made Grab Destination errors. The number was between 0.017 and 0.028 times in a trial. This suggests that

Table 4: Errors. Each entry shows the average number of errors made in a trial.

	S	H	H-L	E	E-L
Failed Grab	0.144	0.147	0.108	0.156	0.131
Grab Destination	0.028	0.028	0.019	0.017	0.017
Failed Deposit	0.339	0.342	0.419	0.411	0.319
Wrong Destination	0.117	0.050	0.108	0.092	0.136

Table 5: Hand data. For each trial in a condition, left and right empty distance are the average distance a participant's hands moved without carrying an object, while left and right full distance are the average distance a participant's hands moved when carrying an object. Bimanual time percentage is the average percentage of time during a trial that a participant was simultaneously holding objects with both hands.

	S	H	H-L	E	E-L
Left empty distance (m)	3.43	3.36	3.23	3.05	2.93
Right empty distance (m)	3.59	3.41	3.26	3.05	3.06
Left full distance (m)	2.81	2.89	2.90	2.98	2.96
Right full distance (m)	3.25	3.23	3.16	3.21	3.16
Bimanual time percentage	16%	15%	15%	20%	22%

although the task objects and the corresponding semitransparent replicas shared the same shape, the visual differences between the replicas and objects helped participants distinguish between them. For Failed Deposit errors, participants made more errors in H-L and in E than in E-L. This may be because in E-L, the cues often fell in a participant's field of view while locking prevented the participant from being distracted. For Wrong Destination errors, the participants made fewer errors in E than in any other approach. Using locking or eye gaze both increased the error rate. This suggests that there was a speed–accuracy trade-off.

4.4.3 Hand Data Analysis

We examined participant hand-controller data (Table 5) to check if the data reflected relative performance between conditions. We first looked at left-hand and right-hand moving distances, finding that participants' left hands moved shorter distances. This is expected, since most of our participants are right-handed. Using eye gaze and locking reduced moving distances for both hands. We also examined how much time (proportional to a task) participants held two objects simultaneously. Participants used two hands simultaneously more often in E and E-L but not in H and H-L than in S.

We also calculated the speed at which participants moved their hands. When empty, the average speed for the left hand is 0.550m/s and for the right hand is 0.582m/s. When full, the average speed for the left hand is 0.703m/s and for the right hand is 0.738m/s. Participants' right hands moved faster. We believe this is because most of our participants were right-handed. In addition, hands moved faster when they were full. This may be because when a participant's hands were empty, they were deciding what to pick up, so reaction time dominated.

5 Discussion

5.1 Number of Cues for Cue-Set–Updating Approaches

For H, H-L, and E, participants performed fastest with three cues and slowest with two or four cues. For S, however, participants performed best with four cues and worst with two cues. One possible explanation is that since participants were unable to change the set of displayed cues in S during a task, displaying more cues allowed more flexibility to pick the steps on which to work. In a pilot study in which the authors participated, one of them performed better with the help of eight cues among all static cues conditions since they could pick their preferred pair of cues on which to work,

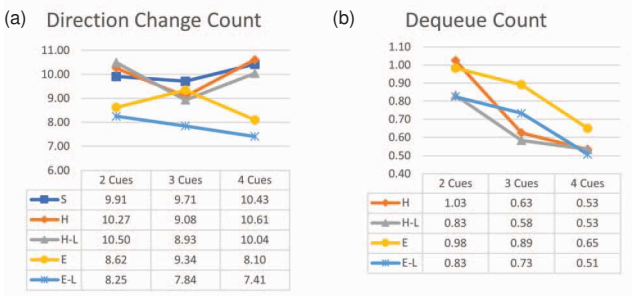

Figure 7: (a) Average number of direction changes in a trial and (b) average number of dequeues in a trial.

though still performing worse than with the best performing hand-proximity and eye-gaze conditions. Further, as shown in Table 3, participants in the formal study generally did not perform better with the Static mechanism than with the Hand-Proximity and Eye-Gaze mechanisms.

The E approach also has a different trend: participants performed best with two cues. We suspect this may be because after participants decided on a set of steps on which to work, they would look at additional objects to obtain information to plan the next step. However, when a participant's gaze glides over objects on the way to the intended destination, an improper dequeue from the cue set may be triggered, causing the disappearance of a displayed cue the participant intends to follow. As the size of the cue set increases, a dequeue of a step that the participant intends to follow immediately is less likely to occur.

To verify this, we calculate the number of abrupt changes in direction and the number of times actionable steps are dequeued for each of these changes and show the results in Figure 7. We define a change in direction as an instance where there is a greater than 120° change in angle in the direction of motion a controller makes in a 0.2s window. We do not consider time intervals during which the controller finishes a task and will generally change its direction of motion. *Dequeue Count* is defined as the number of times a cued task gets dequeued due to hand/eye-gaze movements whenever a change of direction occurs. It can be seen that the number of direction changes is roughly positively correlated to task completion time, meaning that tasks completed with fewer abrupt changes in direction not caused by completing a step generally resulted in a shorter task completion time. This implies that participants were more confident and assertive in their actions for trials that were completed faster. Looking at the data for E in Figure 7, it can be seen that more direction changes happened with three cues, followed by four cues, supporting our speculation.

5.2 Interaction Between Locking and Eye-Gaze/Hand-Proximity Mechanisms

To examine this interaction, we remove S and use cue-set–updating mechanism and locking as the fixed-effect variables in another linear mixed-effects model (see the supplementary material). The model shows that the effect of locking on task completion time is -0.29674s ($p = .006$), and it does not interact with the cue-set–updating mechanism (another model with the interaction term does not have a significant p-value). This can be observed in Table 3. The benefit of adding locking is about -0.3s for both the Hand-Proximity and Eye-Gaze cue-set–updating mechanisms.

5.3 Better Paths for the Static Condition

In S, we use randomly generated sequences, which might cause the participant to move their hands over a longer distance and perform more slowly than necessary by following a sub-optimal path. The

participant might perform better if an "optimal" path were provided in the S condition. To estimate whether the adaptive approaches (H, H-L, E, and E-L) would still outperform the S condition in such a case, we tried to estimate upper and lower bounds on the time to follow optimal paths that could feasibly be executed bimanually by a participant in the S condition. More specifically, we calculated the upper bound, $S_{upperbound}$, assuming no parallelism, and the lower bound, $S_{lowerbound}$, assuming complete parallelism. In addition, we computed $S_{fullhand}$, assuming the same level of parallelism as for each trial's full-hand movement. For details of our approach and calculations, please see the supplementary material.

Estimated average task completion times are $S_{upperbound}$ (11.145), $S_{fullhand}$ (10.489), and $S_{lowerbound}$ (9.052), in comparison to times reported in Figure 4(b): S (10.974), H (10.938), H-L (10.631), E (10.206), E-L (9.881). $S_{upperbound}$ performs worse than H-L, E, E-L ($p < .01$), $S_{lowerbound}$ performs better than H, H-L, E, E-L ($p < .001$), and $S_{fullhand}$ performs better than H ($p < .01$) and worse than E-L ($p < .001$). This suggests that all adaptive approaches except for H outperform the estimated upper bound of the optimal Static solution when assuming no parallelism for empty-hand movement ($S_{upperbound}$). Meanwhile, assuming complete parallelism ($S_{lowerbound}$), our estimated lower bound of the optimal Static solution outperforms all adaptive approaches. Finally, E-L outperforms $S_{fullhand}$, which is estimated based on the amount of parallelism demonstrated in the full-hand moves in our study. Additional work will be needed to find optimal feasible solutions.

5.4 Design Guidelines for Cueing Non-Sequential Tasks

Through our user study, we learned that using eye gaze to update the cue set can reduce task completion time, and using locking further improves completion time. In addition, giving three cues for the best performing approach (E-L) yielded the best result. Therefore, we believe that when designing adaptive visual cues for concurrent manual tasks in VR, one could use eye gaze to infer the user's intention and give higher priority to steps the user looks at. When doing so, one should also consider implementing mechanisms that prevent the steps on which the user is working from being dequeued improperly by a user's rapid eye and hand movements. Though this can be implemented in many different ways and is dependent on the cue-set–updating mechanism, one should consider that the user's eyes can scan surrounding areas or look ahead towards potential sets of future steps and may therefore look in a different direction than the hands move. Regarding the number of cues shown at any given time, one can consider the maximum possible number of cues on which the user can work at the same time and provide one to two more cues to help the user plan.

6 Future Work

6.1 Cueing for More Complex Tasks

Although our 3D task uses start and end positions on a 2D plane, as do many real-world tabletop tasks, the user's hands, head, and body move in 3D to avoid hand collisions, and to better view and reach objects. Building on this, future work should consider tasks in which start and end positions do not lie on a plane.

While sequential tasks investigated in previous work on cueing [19, 29–31, 49, 50] required users to complete steps in a specific order, our task allowed participants to perform cued steps in any order. However, many real-world tasks are partially sequential: some steps can be done in any order, while others must be performed in a specific order. Further, tasks could have additional constraints, such as time limits, orientations, and trajectories [3, 17, 52]. Real-world environments can also include objects that move independently of the user. A key extension of our work will be to investigate how users can leverage adaptive visual cues for these tasks and environments.

With this in mind, we believe that visual guidance systems for such tasks should take into account the dependency structure of steps. More specifically, tasks should be formulated such that any actionable step could be cued, as opposed to following a strictly sequential order that prohibits all bimanual concurrent actions. A related issue is how we can generate dependency structures either manually or automatically to allow for bimanual movements from one user and even collaboration between multiple users. While existing work has tackled automatically inferring parts segmentation [54], computing step-by-step assembly sequences from goal configurations [26, 27, 52], and creating datasets with complex task hierarchies [3, 17, 34, 52], more investigation into generating dependency structures for partially sequential tasks is needed to enable better bimanual task performance. When designing these systems, one can also consider cognitive load, similar in motivation to work by Funk et al., [14], Lindlbauer et al. [28], and Huang et al. [22].

6.2 Expert–Novice Collaboration

Allowing bimanual and concurrent step performance guided by eye gaze and hand proximity also opens up new possibilities for how we can assist expert–novice collaboration. While existing work has addressed creating sequential guidance through techniques such as recording demonstrations and authoring textual descriptions and graphical annotations [1, 11, 15, 18, 22, 36, 44], we hope to inspire future research to leverage the partially sequential or completely unordered nature of different task domains and allow multiple relevant tasks to be prompted and made available to novices. With this in mind, we believe that future visual guidance systems should allow for both experts and novices to work at their own preferred pace, either synchronously or asynchronously, with the experts encouraged to assign potentially concurrent steps to novices and the novices comfortably taking advantage of their spatial context to perform sets of actionable steps concurrently.

6.3 Accuracy

We asked participants to work as fast as possible, as in previous precueing work [19,29,49,50], only requiring that object destinations lie within an acceptance threshold. Asking participants to pursue both speed and accuracy could be confusing and might confound the results, as some participants might prioritize speed and others accuracy. Instead, we deemed a step to be completed when an object was within a threshold distance of its goal position and analyzed accuracy by measuring the errors described in Section 4.4.2. Future studies could focus on accuracy.

7 Conclusions

We explored how adaptive visual cues that are updated based on a user's hand proximity or eye gaze can be used to guide users in a bimanual unordered pick-and-place task in VR. We developed a VR testbed for a task in which the user moves multiple objects to their destinations, up to two at a time, guided by different numbers of displayed cues. A formal user study showed that participants performed better after a third cue was added. In addition, using hand proximity to update the set of displayed cues reduces task completion time, while using eye gaze reduces it further. Using the distances between task objects and hand positions to predict the objects on which the user intends to work and locking their cues can improve task performance even more. Our work extends research on task cueing in VR and AR to bimanual concurrent unordered tasks. The results could be applied to various real-world tasks, including organizing or categorizing items, in which the user can decide the order in which to perform steps and work on multiple steps in parallel.

Acknowledgments

This research was funded in part by National Science Foundation Grant CMMI-2037101. We thank Han-Ching Ou and Kai Wang for their assistance in estimating the bounds of an optimal solution.

REFERENCES

[1] M. Adcock, D. Ranatunga, R. Smith, and B. H. Thomas. Object-based touch manipulation for remote guidance of physical tasks. In *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*, SUI '14, p. 113–122. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2659766.2659768

[2] R. F. Adler and R. Benbunan-Fich. Juggling on a high wire: Multitasking effects on performance. *International Journal of Human-Computer Studies*, 70(2):156–168, 2012. doi: 10.1016/j.ijhcs.2011.10.003

[3] M. Agrawala, D. Phan, J. Heiser, J. Haymaker, J. Klingner, P. Hanrahan, and B. Tversky. Designing effective step-by-step assembly instructions. *ACM Trans. Graph.*, 22(3):828–837, jul 2003. doi: 10.1145/882262.882352

[4] S. Andrist, M. Gleicher, and B. Mutlu. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 2571–2582. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3026033

[5] Beat Games. *Beat Saber*. `https://beatsaber.com/`, last accessed on 06/01/2022.

[6] C. Becchio, V. Manera, L. Sartori, A. Cavallo, and U. Castiello. Grasping intentions: From thought experiments to empirical evidence. *Frontiers in Human Neuroscience*, 6, 2012. doi: 10.3389/fnhum.2012.00117

[7] R. Benbunan-Fich, R. F. Adler, and T. Mavlanova. Towards new metrics for multitasking behavior. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, p. 4039–4044. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1520340.1520614

[8] L. Bonanni, C.-H. Lee, and T. Selker. Attention-based design of augmented reality interfaces. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, p. 1228–1231. Association for Computing Machinery, New York, NY, USA, 2005. doi: 10.1145/1056808.1056883

[9] M. Brysbaert and M. Stevens. Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, 1(1), 2018.

[10] M. D. Byrne, J. R. Anderson, S. Douglass, and M. Matessa. Eye tracking the visual search of click-down menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, p. 402–409. Association for Computing Machinery, New York, NY, USA, 1999. doi: 10.1145/302979.303118

[11] Y. Cao, X. Qian, T. Wang, R. Lee, K. Huo, and K. Ramani. An exploratory study of augmented reality presence for tutoring machine tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376688

[12] Cleanbox Technology. *CleanboxCX1*. `https://cleanboxtech.com/products/`, last accessed on 05/31/2022.

[13] DevM Games and SMG Studio. *Moving Out*. `https://www.smgstudio.com/movingout/`, last accessed on 05/31/2022.

[14] M. Funk, T. Dingler, J. Cooper, and A. Schmidt. Stop helping me—I'm bored! Why assembly assistance needs to be adaptive. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC'15 Adjunct, p. 1269–1273. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2800835.2807942

[15] P. Gurevich, J. Lanir, B. Cohen, and R. Stone. Teleadvisor: A versatile augmented reality tool for remote assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, p. 619–622. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2207676.2207763

[16] S. G. Hart. *NASA Task Load Index (TLX) v. 1.0 Manual*. Human Performance Research Group, NASA-Ames Research Center, Moffett Field, CA, 1986.

[17] J. Heiser, D. Phan, M. Agrawala, B. Tversky, and P. Hanrahan. Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, p. 311–319. Association for Computing Machinery, New York, NY, USA, 2004. doi: 10.1145/989863.989917

[18] B. Herbert, B. Ens, A. Weerasinghe, M. Billinghurst, and G. Wigley. Design considerations for combining augmented reality with intelligent tutors. *Computers Graphics*, 77:166–182, 2018. doi: 10.1016/j.cag.2018.09.017

[19] M. Hertzum and K. Hornbæk. The effect of target precuing on pointing with mouse and touchpad. *International Journal of Human–Computer Interaction*, 29(5):338–350, 2013. doi: 10.1080/10447318.2012.711704

[20] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[21] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.01049

[22] G. Huang, X. Qian, T. Wang, F. Patel, M. Sreeram, Y. Cao, K. Ramani, and A. J. Quinn. Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2021.

[23] J. Huang, R. White, and G. Buscher. User see, user point: Gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, p. 1341–1350. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2207676.2208591

[24] J. Illing, P. Klinke, M. Pfingsthorn, and W. Heuten. Less is more! Support of parallel and time-critical assembly tasks with augmented reality. In *Mensch Und Computer 2021*, p. 215–226. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3473856.3473861

[25] S. Ishihara. *Ishihara's Tests for Colour-blindness*. Kanehara Shuppan, 1972.

[26] B. Jones, D. Hildreth, D. Chen, I. Baran, V. G. Kim, and A. Schulz. Automate: A dataset and learning approach for automatic mating of cad assemblies. *ACM Trans. Graph.*, 40(6), dec 2021. doi: 10.1145/3478513.3480562

[27] Y. Li, K. Mo, L. Shao, M. Sung, and L. Guibas. Learning 3D part assembly from a single image. In *European Conference on Computer Vision*, pp. 664–682. Springer, 2020.

[28] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 147–160. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347945

[29] J.-S. Liu, C. Elvezio, B. Tversky, and S. Feiner. Using multi-level precueing to improve performance in path-following tasks in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4311–4320, 2021. doi: 10.1109/TVCG.2021.3106476

[30] J.-S. Liu, B. Tversky, and S. Feiner. Precueing object placement and orientation for manual tasks in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11), 2022. doi: 10.1109/TVCG.2022.3203111

[31] J.-S. Liu, B. Tversky, and S. Feiner. A testbed for exploring multi-level precueing in augmented reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 540–541, 2022. doi: 10.1109/VRW55335.2022.00121

[32] K. Lukander, M. Toivanen, and K. Puolamäki. Inferring intent and action from gaze in naturalistic behavior: A review. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 9(4):41–57, 2017.

[33] The MathWorks, Inc. *Matlab Statistics and Machine Learning Toolbox*. `https://www.mathworks.com/help/stats/index.html`, last accessed on 05/31/2022.

[34] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] A. K. Mutasim, W. Stuerzlinger, and A. U. Batmaz. Gaze tracking for

eye-hand coordination training systems in virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3334480.3382924

[36] O. Oda, C. Elvezio, M. Sukan, S. Feiner, and B. Tversky. Virtual replicas for remote assistance in virtual and augmented reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software Technology*, UIST '15, p. 405–415. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2807442.2807497

[37] M. Okabe and K. Ito. *Color Universal Design (CUD)—How to make figures and presentations that are friendly to Colorblind people*. `https://jfly.uni-koeln.de/color/`, last accessed on 05/31/2022.

[38] A. Oulasvirta and J. Bergstrom-Lehtovirta. Ease of juggling: Studying the effects of manual multitasking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, p. 3103–3112. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1978942.1979402

[39] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi. A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, 49:215–228, 2018. doi: 10.1016/j.rcim.2017.06.002

[40] K. Pfeuffer, Y. Abdrabou, A. Esteves, R. Rivu, Y. Abdelrahman, S. Meitner, A. Saadi, and F. Alt. Artention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers Graphics*, 95:1–12, 2021. doi: 10.1016/j.cag.2021.01.001

[41] R. Piening, K. Pfeuffer, A. Esteves, T. Mittermeier, S. Prange, P. Schröder, and F. Alt. Looking for info: Evaluation of gaze based information retrieval in augmented reality. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, and K. Inkpen, eds., *Human-Computer Interaction – INTERACT 2021*, pp. 544–565. Springer International Publishing, Cham, 2021.

[42] B. A. Smith, J. Ho, W. Ark, and S. Zhai. Hand eye coordination patterns in target selection. In *Proceedings of the 2000 Symposium on Eye Tracking Research amp; Applications*, ETRA '00, p. 117–122. Association for Computing Machinery, New York, NY, USA, 2000. doi: 10.1145/355017.355041

[43] Stereo Optical Co. Inc. *Original Stereo Fly Stereotest*. `https://www.stereooptical.com/products/stereotests-color-tests/original-stereo-fly/`, last accessed on 05/31/2022.

[44] L. Sun, H. A. Osman, and J. Lang. An augmented reality online assistance platform for repair tasks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(2), may 2021. doi: 10.1145/3429285

[45] Team17 and Ghost Town Games. *Overcooked 2*. `http://www.ghosttowngames.com/overcooked-2/`, last accessed on 05/31/2022.

[46] J. W. Tukey. *Exploratory data analysis*, vol. 2. Reading, Mass., 1977.

[47] Unity. *Unity*. `https://unity.com/`, last accessed on 05/31/2022.

[48] Varjo. *Varjo XR-3*. `https://varjo.com/products/xr-3/`, last accessed on 05/31/2022.

[49] B. Volmer, J. Baumeister, R. Matthews, L. Grosser, S. Von Itzstein, S. Banks, and B. H. Thomas. A comparison of spatial augmented reality predictive cues and their effects on sleep deprived users. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 589–598, 2022. doi: 10.1109/VR51125.2022.00079

[50] B. Volmer, J. Baumeister, S. Von Itzstein, I. Bornkessel-Schlesewsky, M. Schlesewsky, M. Billinghurst, and B. H. Thomas. A comparison of predictive spatial augmented reality cues for procedural tasks. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2846–2856, 2018. doi: 10.1109/TVCG.2018.2868587

[51] P. Weill-Tessier and H. Gellersen. Correlation between gaze and hovers during decision-making interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3204493.3204567

[52] K. D. Willis, P. K. Jayaraman, H. Chu, Y. Tian, Y. Li, D. Grandi, A. Sanghi, L. Tran, J. G. Lambourne, A. Solar-Lezama, and W. Matusik. Joinable: Learning bottom-up assembly of parametric cad joints. *arXiv preprint arXiv:2111.12772*, 2021.

[53] J. Wolf, Q. Lohmeyer, C. Holz, and M. Meboldt. Gaze comes in handy: Predicting and preventing erroneous hand actions in AR-supported manual tasks. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 166–175. IEEE Computer Society, Los Alamitos, CA, USA, oct 2021. doi: 10.1109/ISMAR52148.2021.00031

[54] F. Yu, K. Liu, Y. Zhang, C. Zhu, and K. Xu. PartNet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.