

Exploring Effects of Physical and Cognitive Demands: Simulating AR-based Real-world Situations for Procedural Tasking in UH-60 Helicopters

Jacob Belga*
University of Central Florida

Ryan McKendrick†
Northrop Grumman
Joseph J. LaViola Jr.‡
University of Central Florida

Molly Kluck‡
Northrop Grumman

Ryan P. McMahan§
Virginia Tech

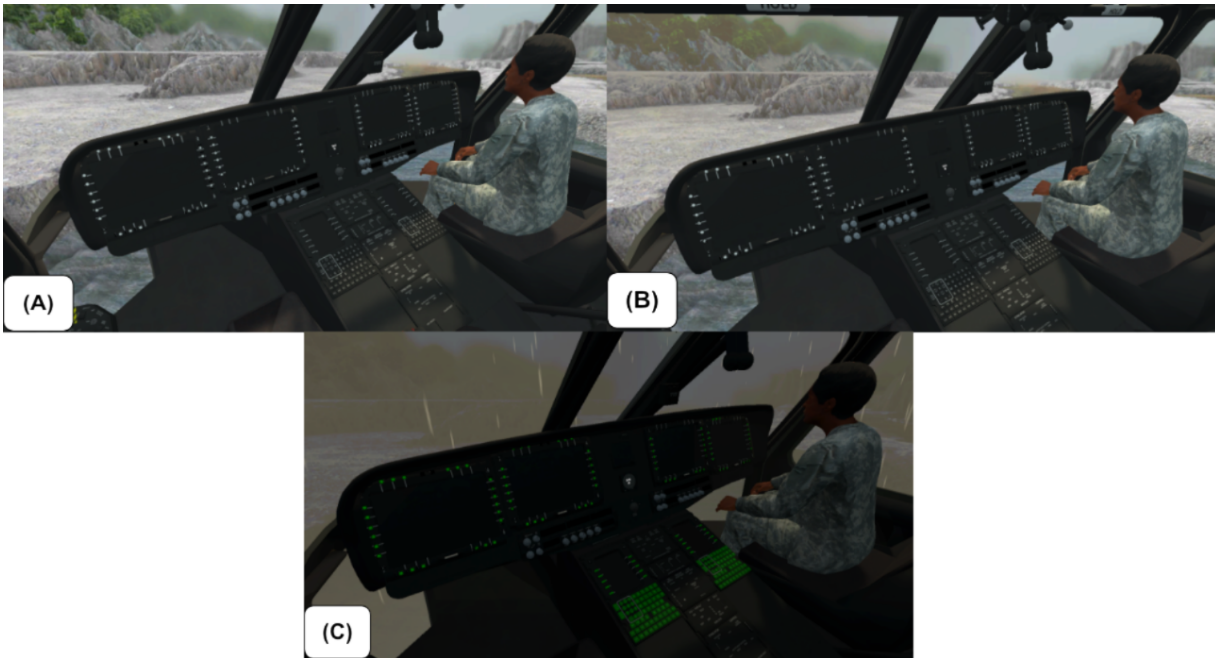


Figure 1: The Storm Conditions present within our study. The coloring of the figure was modified to highlight important visual details of each condition. (A) is the no storm condition. (B) is the incoming storm condition where there are weather effects in front of the helicopter with rain audio. The clouds and weather effects are visible through the brown/grey mist outside the cockpit window. (C) is the in storm condition in which the lighting is changed, there is rain and lightning audio, and to simulate a real UH-60 Black Hawk helicopter, the instrumentation illuminates green.

ABSTRACT

Prior research on augmented reality (AR) and virtual reality (VR) task guidance has largely focused on environments and experiences that are typically free of external distractions or conditions that impact user performance and cognitive load. However, this presents a gap in the literature with regard to understanding how simulating adverse conditions (e.g., poor lighting, rain, loud noises) from the real world affect user perception and performance. Therefore, in our work, we present one of the few studies to investigate the usage of interaction cues in task guidance contexts that involve these adverse, distracting conditions. We conducted a 3×3 within-subjects

experiment in which 22 participants performed procedural button-pressing tasks in a simulated UH-60 helicopter while also completing a 2-back cognitive task. Our study systematically varied interaction cue jitter and environmental conditions as to simulate potential situations co-pilots may find themselves in while piloting a helicopter. The results of our work indicate the jitter has a significant, adverse effect on task completion time and performance. Furthermore, the storm conditions led to significant increases in mental effort, physical effort, and user discomfort despite not being directly related to the primary task. In our paper, we further detail our findings and discuss how our results highlight the importance of accounting for potential adverse conditions that occur in real-world experiences. We also provide avenues of future work to further explore this nuanced area of research.

Index Terms: Virtual Reality, Modeling, Simulation, Perception, Cognition

1 INTRODUCTION

Augmented reality (AR) and virtual reality (VR) applications implement features that inform users to perform specific actions

*e-mail: jacob.belga@ucf.edu

†e-mail: ryan.mckendrick@ngc.com

‡e-mail: molly.kluck@ngc.com

§e-mail: rpm@vt.edu

¶e-mail: jlaviola@ucf.edu

within their immediate environment. These features often come in the form of virtual objects that have an aim to guide users to perform certain actions, such as looking at points of interest [22]. These virtual instructions have been identified as “interaction cues” [8]. Interaction cues have been around for nearly the same time frame as both AR and VR. One of the prime use cases for interaction cues is aiding users in procedural tasking. In a given procedure, interaction cues guide users step-by-step in how they should go about completing a series of tasks. Laviola et al. [16] utilized interaction cues to aid users in a maintenance procedure task. Stover et al. [25] applied interaction cues to guiding users to interact with various items in an office space environment. Schwerdtfeger et al. [23] utilized interaction cues to guide users to distinguish between different storage compartments in their environment.

The aforementioned studies are a few examples of the progress made in interaction cue research. However, in previous work, we often witness task-guidance environments under “ideal” conditions. These ideal conditions often reduce environmental factors such as lighting, weather, and other potential sensory distractions that could affect the user experience. In contrast, when deploying similar systems in real-world contexts, users are often met with adverse, “non-ideal” conditions that can lead to very dynamic, noisy, and sometimes very complex environments (e.g., operating machinery in a low visibility environment or navigating spaces under adverse weather conditions). These non-ideal environmental factors can introduce additional cognitive or physical demands. Therefore, we believe conducting research investigating these environmental distractors can improve our understanding of how these established task guidance systems are affected by non-ideal conditions.

In light of previous work, we ask the research question: *How does environmental degradation and interaction cue instability, in the form of AR cues, affect task performance, attention allocation, and participants’ interaction behavior during complex multitasking scenarios?*. To address this gap, we investigate how adverse environmental conditions and interaction cue consistency affect user performance and cognitive workload in a simulated AR environment. Our user study focuses on two variables: interaction cue jitter, to represent instability in AR tracking, and external visual-auditory stimuli, to represent environmental distractions and conflicts. In real-world settings, AR systems may experience jitter due to poor lighting, sensor or hardware limitations, or inadequate tracking registration [11]. Investigating these types of shortcomings of AR systems, especially in performance dependent tasks, is crucial if researchers are designing experiences to be employed in real-world settings or ecologically valid contexts. For our study, we contextualized the experience in a UH-60 Black Hawk helicopter wherein users utilize interaction cues to complete button-pressing tasks and weather conditions systematically change outside of the helicopter.

The UH-60 Black Hawk helicopter cockpit serves as our testbed due in part to its internal complexity. With multiple buttons to interact with, interaction cues play a crucial role in guiding a user through the tasks. Because we do not have access to a physical UH-60 cockpit, we conducted our user study in a VR environment. Within our VR environment, we address the applicability by simulating the FOV of optical-see-through AR Head-mounted displays (HMDs) such as the HoloLens2, Magic Leap 2, and XREAL Air 2, which have similar FOVs to optical-see-through displays used in UH-60 Black Hawk Helicopters, such as the ANVIS/HUD 24T [9]

In this paper, we present a 3×3 within-subjects VR study ($n = 22$) investigating the aforementioned variables. Participants in our study complete a series of button-pressing procedures while also performing a cognitive 2-back task. Across our study, we systematically varied the levels of both jitter and environmental conditions to evaluate task completion time, task performance, cognitive workload, mental effort, physical effort, and user discomfort. Based on

the results of our study, the following are the outcomes and contributions of our work:

1. The presence of visual-auditory stimuli and jitter adversely impact participants’ ability to complete core procedure task efficiently
2. The presence of visual-auditory stimuli and jitter led to more errors being incurred during the completion of core procedure task.
3. Higher levels of visual-auditory stimuli led to participants prioritizing cognitive demanding task over core task.
4. Higher levels of visual auditory stimuli and jitter led to increased perception of mental effort, physical effort, and discomfort.
5. We present future avenues of research at the intersection of interaction cues and real-world experiences and contexts.

2 RELATED WORK

Our work encapsulates a broad range of research topics. However, one of the key goals with our work is to better understand how the design decisions in interaction cues and distracting stimuli affect performance in both a procedural task and cognitive load task. We focus on reviewing previous work that involves AR systems, interaction cues, simulating AR in VR contexts, and how we can assess cognitive load in VR tasks. By taking cues from previous work, we designed a study that aims to address a gap in current literature pertaining to simulating real-world situations in virtual environments. Therefore, in this section, we provide an overview and analysis of previous work in the aforementioned topic areas.

2.1 Augmented Reality Errors and Simulating the Errors

One of the key benefits AR affords users is the ability to annotate and superimpose virtual components onto the real world. However, with this beneficial feature comes the need to continuously estimate and track spatial information between the user and the real world object(s), which leads to numerous types of errors. These types of errors and their causes are a well-documented and researched topic within the AR research community [26, 30, 29, 3]. Given that there are numerous reasons why tracking errors occur, we believe these tracking errors can be meaningfully categorized. In their work Gottsacker et al. [11], define two categories of error when it comes to AR tracking: Accuracy and Precision. Accuracy errors refer to tracking errors that are biased to a point of location that is not exactly the intended target. Precision errors, on the other hand, refers to the potential of increased variance of point locations. In Livingston et al.’s [18] work, they conducted a user study in which participants were tasked to track certain vehicles with distractor vehicles present. One of the key results from their work is both precision and accuracy errors in AR tracking led to adverse effects on performance when tracking these distant cars. In another study, Robertson et al. [21] led a study in which participants were tasked to use an AR system in a block stacking task. Similarly to Livingston, their results highlighted that both accuracy and precision errors in tracking led to reduced performance as well as increased errors and overall task time. An interesting adjacent result is also that their participants reported that the introduced errors led to a decrease in their confidence to perform well in the task. Lastly, Gottsacker et al. [11] conducted a study also assessing accuracy and precision errors in a search-and-selection task. One of the key results of their work is that with each increase in both accuracy and precision error, there was an increase in performative errors across their participants.

Based on these previous studies, we applied similar faults in AR systems to the context of co-piloting a UH-60 Black Hawk helicopter, specifically applying both precision and accuracy errors to interaction cues [11, 18]. Given that these helicopters are not readily available, we opted to simulate the AR experience through VR. In previous research, the idea of simulating AR in VR environments has become more commonplace to examine potential design decisions for future AR HMDs. For example, Lacoche et al. [15], conducted a user study in which they developed a VR simulation of an AR smart-home application in which they can interact with devices in the environment. In another example, Lee et al. [17], conducted a study in which they simulated the AR experience with the added caveat of them being able to control the latencies present in the AR simulation. Participants were tasked with following a yellow pipe snake that went in/out of the walls in the environment while orange pipe snake distractors were also present. Lastly, Ragan et al. [20] simulated an AR system to investigate the effects of registration errors on performance. Specifically, they investigated the impact jitter noise and latency had on the performance of participants.

Based on the aforementioned errors and study designs, we applied the lessons learned and design decisions to our study. We leveraged interaction cues in the form of 3D arrows that had the purpose of guiding users through procedural button-pressing tasks. In Belga et al.'s [1] work, we take inspiration by also implementing the look cue and pick cue system. We define look cues as interaction cues with the aim of directing your attention to a specific area of interest. Pick cues are interaction cues with the goal of helping distinguish which target needs attention out of the host of target candidates. Similar to other works [15, 17, 20, 11], we wanted to investigate the performative impacts of jitter in a real-world situation that one may find themselves in. Therefore, we conducted our study in a UH-60 helicopter where participants served the role as a co-pilot with the responsibility of completing button-pressing procedures while also needing to scan the environment outside of the cockpit.

2.2 Interaction Cues in Real-world AR contexts

In both AR and VR contexts, interaction cues provide users visualizations to better understand the virtual components present within their environment. However, in review of the literature, many of the prominent interaction cues that we have researched have not been applied in real-world situations, or in our case, simulated real-world simulations.

Gruenfeld et al. [12] proposed an interaction cue, specifically a look cue system, called *EyeSee360*. In their work, they evaluated the performance of *EyeSee360* through guiding users in a task of identifying out-of-view targets. In Wallgrun et al.'s [28] work, they utilized conventional 3D Arrows, a butterfly, and a radar mechanism to guide users in a virtual tour experience. Biocca et al. [2] propose an interaction cue identified as the *attention funnel* in which a series of cascading boxes starting from in front of the user's face and ending to the target of interest are present in the user's field of view. In their work, they evaluated this interaction cue through presenting users in front of a table with six primitive objects for them to view and direct their attention towards based on the attention funnel. Bork et al. [4] conducted an AR user study in which participants utilized a variety of interaction cues, (e.g., 3D arrows, *EyeSee360*, *AroundPlot*), to collect eight virtual objects in the space around them. Harada et al. [13] also conducted a VR user study in which participants had to leverage a series of interaction cues (moving window, 3D arrow, radiation, spherical gradation, and 3D radar), to be directed towards a virtual target.

Based on previous work in the interaction cue space, we aimed to apply interaction cues to a real-world situation in which interaction cues would benefit the user. As highlighted in section 2.1, we utilized conventional 3D arrows for our interaction cues. Addition-

ally, our rationale for choosing the UH-60 Black Hawk helicopter as our task context is that pilots conventionally utilize electronic checklists known as "kneeboards" due to pilots placing them on their knees [5]. Often times, these checklists contain procedures that consist of hundreds of steps involving interacting with multiple components within the cockpit while confirming settings and information presented on their multi-function displays (MFDs) and control display units (CDUs).

3 STUDY METHODS

We conducted a 3×3 within-subjects VR user study. All study conditions were counter-balanced through applying a Latin square design. For each of the nine conditions, participants were immersed into a UH-60 Black Hawk helicopter cockpit with the primary task of completing a 20-step button-pressing procedure per trial. Participants also had a secondary task of keeping track of a 2-back task via shapes that would appear outside of the helicopter cockpit.

3.1 Independent Variables

In our study, we examine two independent variables: storm level and interaction cue jitter. For each independent variable, we have implemented three variations for each. For storm level, we implemented no storm, incoming storm, and in-storm conditions. Regarding interaction cue jitter, we implemented no jitter, medium jitter, and large jitter conditions.

3.1.1 Jitter levels

For our jitter conditions, we adopted a similar implementation as Gottsacker et al.'s [11] angular jitter conditions. For no jitter conditions, both the look cue and pick cue would remain consistently placed and would maintain both precise and accurate location with respect to the target button. For medium jitter conditions, we applied a 1° of angular jitter based on a unit sphere around the target. Both the look cue and pick cue would random jitter to simulate loss of tracking. For large jitter conditions, we applied a 2° of angular jitter with the same random jitter frequency as the medium jitter condition.

3.1.2 Weather conditions

For our weather conditions, we leveraged both visual and audio effects to achieve our 3 levels. For the no storm condition, as shown in Figure 1A, there would be no weather effects or noise during the trial. For the incoming storm condition, as shown in Figure 1B, there would be some cloud effects around the cockpit with light rain present paired with rain audio. For the in storm condition, as shown in Figure 1C, the lighting effects would be darkened, there are much stronger cloud and rain effects, both rain and thunder audio would be present, and the UH-60 Black Hawk helicopter's instrumentation would illuminate green as to simulate the real helicopter.

3.2 Dependent Variables

In our study, we evaluate six dependent variables: Total Time, procedure error, N-back task error, Mental Effort through the Paas et al. questionnaire [19], Physical Effort, and Discomfort. Total time is the time calculated to complete a single 20-button press procedure. Procedure error is the running count of incorrect button presses during a 20-button press trial. N-back task error is the running count and instances in which participants evaluated an item in the N-back sequence incorrectly or does not evaluate the item when it is shown. Mental effort, physical effort and discomfort are measured between each 20-button press trial via three, single-item questionnaires.

3.3 User Study Procedure

Recruited participants were asked to review and complete a pre-screening Qualtrics survey. In this survey, participants reviewed an

informed consent document, eligibility document, and provide additional demographic information such as age, gender, video game experience, first-person shooter experience, and VR experience. After completing the online form, participants were prompted to schedule a day and time in which they would be able to participate in the VR study.

On their scheduled study session day, participants were greeted and introduced to the Meta Quest Pro, the VR system utilized to administer our user study. In this introduction, participants were shown how to adjust the headset's straps and lenses such that it rested on the participants' head comfortably. We also provided an explanation of the controls and actions needed to perform the primary button-pressing task and the 2-back task during the trials of the study. After familiarizing themselves with the VR system, we explained to the participants that the study is comprised of 9 trials of completing button-pressing procedures in a UH-60 Black Hawk helicopter cockpit. We also explained that during the button-pressing procedures, they would be guided by a 3D arrow look cue and a 3D arrow pick cue to know which button in the cockpit they would need to press. In addition to the button-pressing procedures, participants were also informed of how the 2-back task would work. Shapes would intermittently appear outside of the cockpit, and the participants were responsible for keeping track of these shapes and either accepting, if they saw the shape 2 sequence items ago, or rejecting as they saw the shapes appear. For the entire study, participants sat in a chair to ensure both comfort and similarity to sitting in the cockpit.

After explaining the instructions and tasks, we verified that participants both understood the task, controls, and if they were ready to begin. Once verified, participants were immersed into the nine trials. In each trial, participants were allowed to complete the button-pressing procedure at their own pace and a trial would end once the participant pressed the 20th button in the procedure. At the end of each trial, participants would respond to three single-item questionnaires on their mental effort, physical effort, and discomfort. Given that there were nine conditions in our user study, we opted to administer three single-item questionnaires to reduce the potential of survey fatigue. After four trials, participants were offered an opportunity to take a break from the study if needed. In addition to the constant half-way point break, participants were informed they could take a break at any time if needed after they completed a trial. After the half-way point break, participants would finish the remaining five trials. All conditions were counter-balanced between participants through applying a Latin square design for our user study.

Overall, the time required to complete the pre-screening survey and in-person VR study with questionnaires was 45 minutes. Participants who completed the aforementioned study components were compensated \$20 USD via an Amazon e-gift card.

3.4 Participants

Prior to our user study, we conducted a power analysis using G*power for an F-test ANOVA analysis with a Cohen's F of 0.25, $\alpha = 0.05$, and a power of 0.95. Given our study design is within-subjects with nine conditions, our power analysis revealed that an adequate sample size is at least 22 participants, which we achieved in our participant pool.

In our user study, we recruited a total of 22 participants (14 male, 7 female, 1 Non-binary) from our local university. Participants were required to be 18 years of age or older, have normal or corrected-to-normal vision, be able to hear, be able to walk and extend both arms, and be able to speak and understand English. Participants who reported any visual, auditory, neurological, or physical disabilities were excluded from participation. Our participants' ages ranged from 18 to 32, with a mean age of 22.



Figure 2: The interaction cues we employed in our user study. The 3D Arrow look cue (Top) directs users to where the target is located with respect to where they are looking. The 3D Arrow pick cue (bottom) distinguishes, out of the target candidates, which target is the one requiring attention.

3.5 Apparatus

In our study, there are several moving parts that allowed us to conduct our user study. We further detail the design details of the UH-60 Black Hawk Helicopter Cockpit, interaction cues utilized, the N-back task, and jitter conditions used in our user study. To interface with our user study environment, we utilized the Meta Quest Pro and controllers as our VR system to administer our user study.

3.5.1 Interaction Cues

In our study, we have two types of interaction cues: a look cue and a pick cue. Both are represented as a conventional 3D arrow with the caveat that each cue serves different functions. The look cue provides the user with context of where they should look in order to see the target button or interaction. The look cue will always be within the user's FOV. However, once the user looks at the point of interest, the look cue will disappear. Given that our cockpit has a broad range of target candidates, the pick cue helps distinguish which target needs to be selected of the similar options within the area of interest. The pick cue is not always in the user's FOV, and will remain on the point of interest until the user interacts with the intended target. Figure 2 showcases both the look and pick cues we employed in our user study.

3.5.2 Jitter of Interaction Cues

In our user study, we applied both an accuracy and precision jitter to both our look and pick cues. For the jitter conditions, we employed three levels as highlighted in Section 3.1.1. To achieve the jitter, we applied a transformation of position of the look and pick cue during the condition. This transformation is based on both a unit sphere and a multiplier to control for distance for both the medium and large jitter conditions. The jitter would be applied randomly every 0.4 seconds.



Figure 3: Potential locations for stimuli in our 2-back task



Figure 4: Example of a 2-back stimuli from the cockpit

3.5.3 N-back Task

For our 2-back task, to simulate observing potential points of interest outside of the cockpit. Given the high visibility around the UH-60 Black Hawk cockpit, there were four potential locations a stimuli would appear for the 2-back task, as shown in Figure 3. Stimuli would appear one at a time, as shown in Figure 4, and participants would need to respond to the stimuli by pressing 'A' for accept or 'B' for reject on the Meta Quest Pro controller. The stimuli could be one of four shapes: sphere, cube, rectangle, or cylinder.

3.5.4 UH-60 Black Hawk Helicopter Cockpit

Our UH-60 Black Hawk cockpit is adapted from Belga et al.'s work on interaction cues [1]. The cockpit virtual environment was developed to be fully interactive and complete with pressable buttons, turnable knobs, and various displays to view. Given our study design focused primarily on completing button pressing procedures, every button in the cockpit can be pressed and when pressed, each button provides both audio and haptic feedback to the user to inform when a button has been pressed. The model of our UH-60 Black Hawk Helicopter Cockpit was designed to match the exact layout and positioning of the instrumentation (e.g., Control Display Unit (CDU) and Multi-Function Displays (MFDs)), which adds to the ecological validity of our context.

Additionally, the environment was developed using Unity 2018.3.18f1 and the SteamVR plugin to interface with most consumer VR systems. For the button pressing interaction, users would not be required to push any buttons on the controller, but rather, reach out towards the button and their virtual hand would automatically repose itself to a pointing gesture. Once the user's virtual index finger collides with a button, the button would register a press and provide the aforementioned auditory and haptic feedback to the user. This interaction was fulfilled using the Meta Quest Pro controllers.

4 RESULTS

In our results, we calculated and analyzed total time, button error count, 2-back error percentage, mental effort, physical effort, and discomfort for each trial each participant completed. In our study, we found our data to be in violation of normality, therefore we performed an Aligned Rank Transform (ART) to conduct an ANOVA

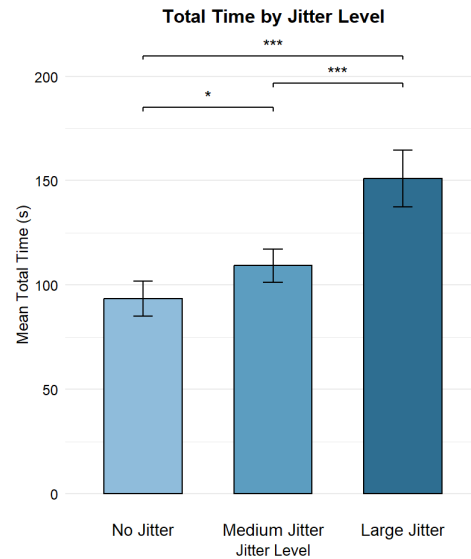


Figure 5: The mean times based on jitter conditions.

on the non-parametric data [31]. Given our within-subjects study design, we administered a repeated measures ANOVA. In addition to our partial-eta square effect size reporting, we also report our effect sizes using Cohen's F [6]. For our pairwise t-tests for each result, we also applied a Bonferroni correction.

4.1 Total Time

With regard to total trial completion times, our results revealed a significant main effect for jitter level, $F_{1,141,23.957} = 24.774, p < 0.001, \eta^2 = 0.541$, with a Cohen's F of 1.086. Our results also revealed a significant interaction effect between jitter level and storm level, $F_{2,217,46.554} = 3.1201, p = 0.01, \eta^2 = 0.129$, with a Cohen's F of 0.385. We did not find significance for a storm level main effect, $F_{1,507,31.655} = 2.109, p = 0.124, \eta^2 = 0.091$, with a Cohen's F of 0.3164.

For our significant jitter level main effect, pairwise t-tests revealed that conditions with large jitter effects led to significantly longer trial completion times than conditions with medium jitter effects, $t_{22} = -4.307, p < 0.001$. Furthermore, larger jitter effect conditions, $t_{22} = 2.667, p = 0.02$, and medium jitter effect conditions, $t_{22} = 6.975, p < 0.001$, both led to significantly longer trial completion times than conditions with no jitter effects. The mean total times across the jitter conditions are summarized in Figure 5. Our pairwise comparisons for our interaction effect between jitter and storm level are also summarized in Table 1.

4.2 Procedure Error

For procedure errors, instances in which participants pressed the wrong button, we had significant main effects for both jitter level, $F_{1,047,21.990} = 66.999, p = 0.04, \eta^2 = 0.761$, and storm level, $F_{1,093,22.955} = 16.157, p = 0.04, \eta^2 = 0.435$, with a Cohen's F value of 1.784 and 0.877 respectively. Our results also indicated that there is a significant interaction effect between jitter level and storm level, $F_{1,223,25.679} = 6.219, p = 0.01, \eta^2 = 0.229$, with a Cohen's F value of 0.545.

With regard to jitter level, conditions with large jitter led to significantly more procedure errors than medium jitter conditions $t_{22} = -3.801, p < 0.001$. Furthermore, both conditions with large jitter, $t_{22} = 7.568, p < 0.001$, and medium jitter, $t_{22} = 11.369, p < 0.001$, led to significantly more procedure errors than no jitter conditions.

Table 1: Significant pairwise comparisons for the jitter and storm level interaction effect on Total Time. In each row, the bolded condition had significantly *lower* Total Time than its paired condition.

Comparison	t-statistic	p-value
Medium Jitter, In Storm – No Jitter, No Storm	$t_{22} = 3.33$	$p = 0.03$
Medium Jitter, Incoming Storm – Large Jitter, In Storm	$t_{22} = -3.60$	$p = 0.01$
Medium Jitter, Incoming Storm – Large Jitter, No Storm	$t_{22} = -4.61$	$p < 0.001$
Large Jitter, In Storm – No Jitter, In Storm	$t_{22} = 4.70$	$p < 0.001$
Large Jitter, In Storm – No Jitter, No Storm	$t_{22} = 5.13$	$p < 0.001$
Large Jitter, Incoming Storm – No Jitter, In Storm	$t_{22} = 3.44$	$p = 0.02$
Large Jitter, Incoming Storm – No Jitter, No Storm	$t_{22} = 3.87$	$p = 0.005$
Large Jitter, No Storm – No Jitter, In Storm	$t_{22} = 5.70$	$p < 0.001$
Large Jitter, No Storm – No Jitter, Incoming Storm	$t_{22} = 3.92$	$p = 0.004$
Large Jitter, No Storm – No Jitter, No Storm	$t_{22} = 6.13$	$p < 0.001$

Table 2: Significant pairwise comparisons for the jitter and storm level interaction effect on Errors. In each row, the bolded condition had significantly *lower* Errors than its paired condition.

Comparison	t-statistic	p-value
Large Jitter, In Storm – Medium Jitter, No Storm	$t_{22} = 3.59$	$p = 0.016$
Large Jitter, In Storm – No Jitter, In Storm	$t_{22} = 8.28$	$p < 0.001$
Large Jitter, In Storm – No Jitter, Incoming Storm	$t_{22} = 7.64$	$p < 0.001$
Large Jitter, In Storm – No Jitter, No Storm	$t_{22} = 8.70$	$p < 0.001$
Large Jitter, Incoming Storm – No Jitter, In Storm	$t_{22} = 7.73$	$p < 0.001$
Large Jitter, Incoming Storm – No Jitter, Incoming Storm	$t_{22} = 7.09$	$p < 0.001$
Large Jitter, Incoming Storm – No Jitter, No Storm	$t_{22} = 8.16$	$p < 0.001$
Large Jitter, No Storm – Medium Jitter, No Storm	$t_{22} = 3.58$	$p = 0.016$
Large Jitter, No Storm – No Jitter, In Storm	$t_{22} = 8.27$	$p < 0.001$
Large Jitter, No Storm – No Jitter, Incoming Storm	$t_{22} = 7.64$	$p < 0.001$
Large Jitter, No Storm – No Jitter, No Storm	$t_{22} = 8.70$	$p < 0.001$
Medium Jitter, In Storm – No Jitter, In Storm	$t_{22} = 6.38$	$p < 0.001$
Medium Jitter, In Storm – No Jitter, Incoming Storm	$t_{22} = 5.74$	$p < 0.001$
Medium Jitter, In Storm – No Jitter, No Storm	$t_{22} = 6.81$	$p < 0.001$
Medium Jitter, Incoming Storm – No Jitter, In Storm	$t_{22} = 5.10$	$p < 0.001$
Medium Jitter, Incoming Storm – No Jitter, Incoming Storm	$t_{22} = 4.46$	$p < 0.001$
Medium Jitter, Incoming Storm – No Jitter, No Storm	$t_{22} = 5.53$	$p < 0.001$
Medium Jitter, No Storm – No Jitter, In Storm	$t_{22} = 4.69$	$p < 0.001$
Medium Jitter, No Storm – No Jitter, Incoming Storm	$t_{22} = 4.05$	$p = 0.003$
Medium Jitter, No Storm – No Jitter, No Storm	$t_{22} = 5.12$	$p < 0.001$

The mean procedure errors across all jitter conditions of our study can be summarized in Figure 7.

For Storm level, conditions with no storms led to significantly less procedure errors than in storm conditions $t_{22} = -3.801, p < 0.001$. Similarly, incoming storm conditions also led to significantly less procedure errors than in storm conditions $t_{22} = -5.538, p < 0.001$. The mean procedure errors across all storm conditions of our study are summarized in Figure ??.

4.3 N-back Task Error

For our 2-back task error, we identify an error as an n-back stimulus that the participant incorrectly categorizes, accept or reject, or does not respond to a stimulus. For our 2-back task, our results indicate significance for a storm level main effect, $F_{2,42} = 3.6317, p = 0.04, \eta^2 = 0.147$, with a Cohen's F value of 0.415. Our pairwise t-test for storm level revealed that in-storm conditions yielded significantly higher 2-back scores than no storm conditions $t_{22} = -2.487, p = 0.04$. The main effects results of our 2-back task are summarized in Figure 6.

4.4 Mental Effort

Throughout our study, our participants responded to a mental effort single-item questionnaire adapted from Paas et al. [19]. In

our results, we find significance for jitter level main effect, $F_{2,42} = 3.059, p = 0.04, \eta^2 = 0.127$, with a Cohen's F of 0.3814. Similarly, we found a significant main effect for weather conditions, $F_{2,42} = 6.773, p = 0.001, \eta^2 = 0.244$, with a Cohen's F of 0.5681. However, we did not find a significant interaction effect between jitter and weather, $F_{2,525,53,024} = 0.422, p = 0.01, \eta^2 = 0.02$, with a Cohen's F of 0.1429. Figure 8b summarizes our mean mental effort scores across storm conditions, and Figure 8a summarizes mean mental effort scores across jitter conditions.

Delving into our pairwise analysis, we find that conditions with the large jitter effect led to significantly higher mental effort scores than conditions with no jitter effect $t_{22} = 2.461, p = 0.04$. With regard to the weather conditions, conditions that were in the storm led to significantly higher mental effort scores than conditions with no storm conditions $t_{22} = 3.663, p = 0.01$.

4.5 Physical Effort

In our user study, we applied a modified version of the Paas et al. [19] single-item questionnaire, but applied it to physical effort as opposed to mental effort. Our results revealed a significant main effect for jitter level, $F_{2,42} = 7.566, p < 0.001, \eta^2 = 0.265$, with a Cohen's F of 0.6005. We also had a significant main effect for weather conditions, $F_{1,426,29,956} = 8.337, p < 0.001, \eta^2 = 0.284$,

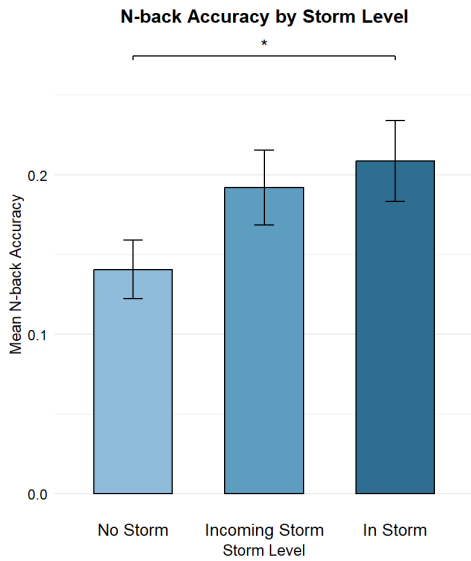


Figure 6: The mean accuracy of the 2-back task across the storm conditions in our user study.

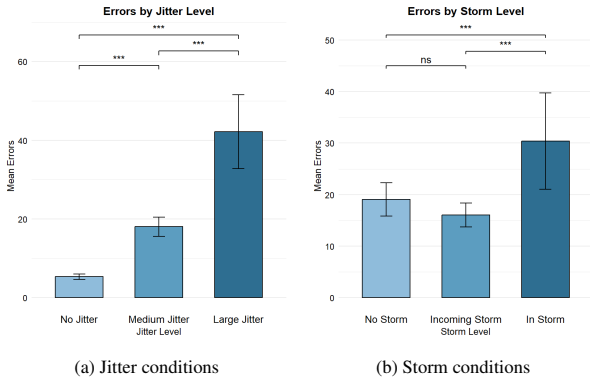


Figure 7: The mean procedure errors across jitter and storm conditions from our user study.

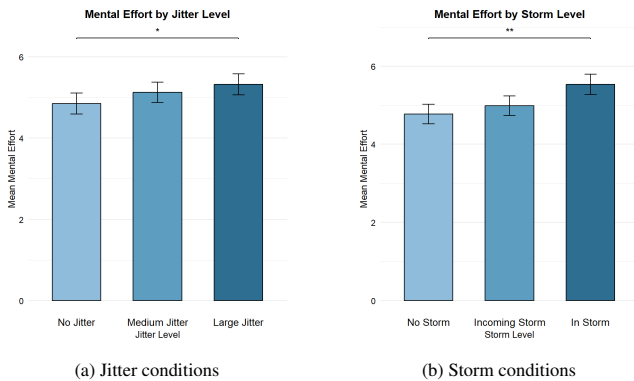


Figure 8: The mean mental effort scores across jitter and storm conditions from our user study.

with a Cohen's F of 0.6298. As with mental effort, we also did not find a significant interaction effect between jitter and weather conditions, $F_{2,525,53.024} = 1.560, p = 0.187, \eta^2 = 0.069$, with a Co-

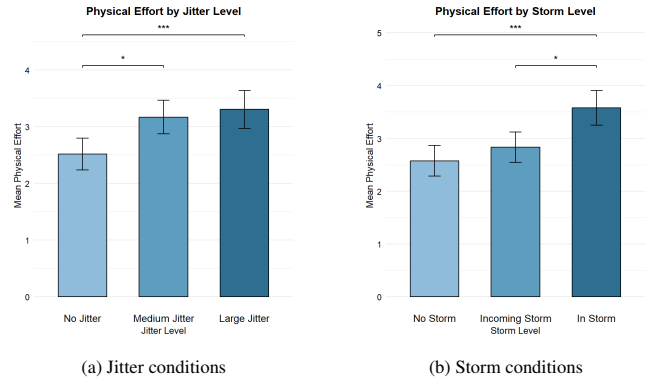


Figure 9: The mean physical effort scores across jitter and storm conditions from our user study.

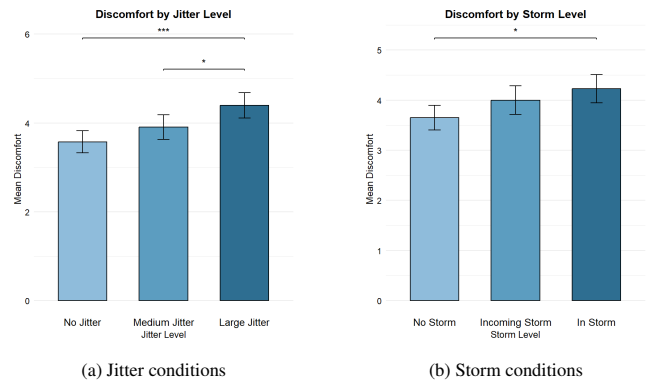


Figure 10: The mean discomfort scores across jitter and storm conditions from our user study.

hen's F of 0.2722. Figure 9a summarizes our mean physical effort scores across our jitter conditions, and Figure 9b summarizes our mean physical effort scores across our storm conditions.

For our pairwise analysis, conditions with a medium jitter effect led to significantly higher physical effort scores than conditions with no jitter effect, $t_{22} = 2.585, p = 0.03$. Similarly, conditions with a large jitter effect led to significantly higher physical effort scores than conditions with no jitter effect, $t_{22} = 3.809, p < 0.001$. With regard to the storm conditions, in storm conditions led to significantly higher physical effort scores than conditions of the incoming storm, $t_{22} = 2.817, p = 0.01$. Furthermore, in storm conditions also led to significantly higher physical effort scores than conditions without a storm, $t_{22} = 3.968, p < 0.001$.

4.6 Discomfort

In our user study, participants also responded to the Fernandes et al. [10] single-item discomfort questionnaire on a 0 to 10-point scale. For our jitter level main effect, we found significance, $F_{2,42} = 11.248, p < 0.001, \eta^2 = 0.349$, with a Cohen's F of 0.7322. Additionally, we had a significant main effect for storm level, $F_{2,42} = 3.248, p = 0.04, \eta^2 = 0.134$, with a Cohen's F of 0.3934. We did not witness a significant interaction effect between our two independent variables, $F_{2,848,59.814} = 0.066, p = 0.991, \eta^2 = 0.003$, with a Cohen's F of 0.0549. Figure 10a summarizes our mean discomfort scores across our jitter conditions, and Figure 10b summarizes our mean discomfort scores across our storm conditions.

For jitter conditions, large jitter conditions led to significantly higher discomfort scores than medium jitter conditions, $t_{22} = -2.428, p = 0.04$. Similarly, large conditions led to significantly

higher discomfort scores than no jitter conditions, $t_{22} = 4.742, p < 0.01$. For weather conditions, conditions in the storm led to significantly higher discomfort scores than no storm conditions, $t_{22} = 2.546, p = 0.03$.

5 DISCUSSION

In our paper, we have highlighted the impact both jitter and external visual-audio stimuli has on both the performative and cognitive load of our participants. Similarly to previous work, our study has demonstrated the adverse performative effects of jitter and distractions on task performance. However, a nuanced result is the cognitive load component through our 2-back task. From our results of our within-subjects study, we also present the implications of our results as well as propose future research directions based on our work.

5.1 Performative Impacts of Jitter and External Stimuli

In our study, the design implementation of the jitter of our interaction cues took inspiration from works like [11]. From a performative standpoint, the introduction of jitter led to significant, adverse effects on our participants' ability to perform well in completing the procedural button-pressing tasks. This result aligns with previous work in this space [11, 18, 21]. Our results also highlight the negative effects our weather conditions had on participant performance in the button-pressing task. Across our conditions, each degree of storm led to more procedure errors as well as an increase in total time spent in a given trial. However, given that our user study is contextualized in a real world experience, we believe an important avenue of research is to better understand the ways in which we can mitigate these adverse effects should a real AR system impose these on a user.

5.2 Cognitive Load Impacts of Distractions

One of the more unexpected outcomes from our study was our participants' performance in the 2-back task. Across our conditions, we found a significant main effect for storm level with regard to N-back accuracy, indicating that conditions in the in storm conditions yielded better 2-back scores than no storm conditions. During our explanation of the user study experience to participants, we did not indicate that they would be under any time pressure or any other kind of pressure outside of the VR environment. Despite being told their main goal is to complete each of the button-pressing procedures, we find it interesting that in the in-storm conditions, participants opted to prioritize the 2-back task when compared to the no-storm conditions. It may be the case that in the storm conditions, our participants organically aimed to be more cognizant and aware of the events occurring outside the cockpit (e.g. rain storm, 2-back shapes). In contrast, the no storm conditions "do not require" as much external attention as there is no perceived stress. This is further supported by both our mental effort and physical effort self-reports in which participants had significantly higher scores after experiencing in storm conditions.

5.3 Limitations and Future Work

One of the limitations of our work is that our implementation of AR is simulated as opposed to real-life. Given our context of a UH-60 Black Hawk helicopter, access to such equipment is not readily available. However, we believe the simulated UH-60 Black Hawk helicopter afforded us a unique study context in which within the cockpit, there are hundreds of potential target candidates co-located in the same space. This opens up an avenue of future research that leverages the look and pick cue implementations we presented in our paper to AR system contexts.

In our study's sample size ($n = 22$), we witness a skewed gender representation towards male participants (14 male, 7 female, 1 Non-binary) which limits the generalizability of our work. While

we did not have motivation to investigate the degree to which gender impacts user performance in this task, we do note that there is previous work that highlights gender differences appearing in a variety of spatial cognition contexts [11, 7].

Another limitation of our work is that our participant pool is comprised of individuals who were not familiar with the UH-60 Black Hawk helicopter (i.e., pilots and co-pilots). Given our limited access to a demographic, it is possible to conjecture that having pilots and co-pilots as participants would impact the results of our study. A future avenue of research would be to evaluate a system like ours with pilots and co-pilots since they would be more familiar with the instrumentation. An additional avenue of future work would apply similar techniques we employed in our work in similar, but different contexts. One example is investigating cueing and distractions on a control panel in an industrial setting.

Another limitation of our study design is the type of look cues and pick cues we chose. While 3D arrows are very standard across both AR and VR applications, there are a broad range of other interaction cues that have the potential to perform better under the adverse conditions we present in our work. For example, "ghost hands" or semi-transparent hands to guide users in how to interact with objects is a well documented pick cue [14]. Another potential cue to investigate are highlighting cues, such as those found in Volmer et al.'s [27] and Schwerdtfeger et al.'s [24] works.

In our work, we conducted the 2-back task to simulate a multi-tasking scenario. However, we acknowledge that our observed effects may or may not persist in a real helicopter scenario, especially with real pilots and co-pilots. However, we consider the possibility that pilots that prioritize navigation and observation outside the helicopter could exemplify our found results with regard to the storm conditions. We believe this may present an avenue of future work that further investigates the potential effects of prioritizing navigation and observation over procedure completion or vice versa.

Given our works' concern on adverse conditions, we believe that future work and development of these types of guidance systems should consider potential adversities in system or environment. We encourage future researchers to take into account these considerations and employ techniques or methods to overcome such adversities. In our work, we highlight one specific context through the UH-60 Black Hawk helicopter. Future work would investigate other types of procedural guidance systems in different contexts.

6 CONCLUSION

In this paper, we presented a multi-factor, within-subjects study investigating the impact external sensory stimuli affects task performance and mental cognitive workload. Our study design aimed to immerse users into a task that was grounded in a real-world task and engaging to participants regardless of familiarity to UH-60 Black Hawk helicopters. In our user study ($n = 22$), we evaluated the impact of interaction cue jitter and weather effects on participant performance. Our results highlight significant issues with regard to core task performance and mental workload performance. However, given the context of our study, we have also provided additional avenues for future work and research.

ACKNOWLEDGMENTS

This work is supported in part by NSF Award DRL-2506427 and Army Research Lab Award W912CG2320004. We would also like to thank the anonymous reviewers for their valuable feedback.

REFERENCES

- [1] J. Belga, R. Ghamandi, N. Chawla, R. McKendrick, R. P. McMahan, and J. J. LaViola Jr. An investigation of look cues and pick cues for guidance in dense operating environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 69, pp.

- 1090–1096. SAGE Publications Sage CA: Los Angeles, CA, 2025. 3, 5
- [2] F. Biocca, A. Tang, C. Owen, and F. Xiao. Attention funnel: omnidirectional 3d cursor for mobile augmented reality platforms. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 1115–1122, 2006. 3
- [3] G. Bishop, G. Welch, and B. D. Allen. Tracking: Beyond 15 minutes of thought. *SIGGRAPH Course Pack*, 11, 2001. 2
- [4] F. Bork, C. Schnelzer, U. Eck, and N. Navab. Towards efficient visual guidance in limited field-of-view head-mounted displays. *IEEE transactions on visualization and computer graphics*, 24(11):2983–2992, 2018. 3
- [5] R. Bridgeman, K. J. Neville, L. Massey, C. Krauskopf, A. Mizan, J. Mooney, and D. Schmorow. Human factors electronic kneeboard design guidelines for military tactical aviation. In *20th International Symposium on Aviation Psychology*, p. 343, 2019. 3
- [6] J. Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013. 5
- [7] E. Coluccia and G. Louse. Gender differences in spatial orientation: A review. *Journal of environmental psychology*, 24(3):329–340, 2004. 8
- [8] K. R. Dillman, T. T. H. Mok, A. Tang, L. Oehlberg, and A. Mitchell. A visual interaction cue framework from video game environments for augmented reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018. 2
- [9] Elbit Systems. HDTs — Elbit Systems. <https://www.elbitsystems.com/air-space/aircraft-systems/helmet-mounted-display/hdts>. Accessed: 2026-06-24. 2
- [10] A. S. Fernandes and S. K. Feiner. Combating vr sickness through subtle dynamic field-of-view modification. In *2016 IEEE symposium on 3D user interfaces (3DUI)*, pp. 201–210. IEEE, 2016. 7
- [11] M. Gottsacker, H. Furuya, Z. D. Choudhary, A. Erickson, R. Schubert, G. Bruder, M. P. Browne, and G. F. Welch. Investigating the relationships between user behaviors and tracking factors on task performance and trust in augmented reality. *Computers & Graphics*, 123:104035, 2024. 2, 3, 8
- [12] U. Gruenefeld, D. Ennenga, A. E. Ali, W. Heuten, and S. Boll. Eye-see360: Designing a visualization technique for out-of-view objects in head-mounted augmented reality. In *Proceedings of the 5th symposium on spatial user interaction*, pp. 109–118, 2017. 3
- [13] Y. Harada and J. Ohyama. Quantitative evaluation of visual guidance effects for 360-degree directions. *Virtual Reality*, 26(2):759–770, 2022. 3
- [14] X. Hu, A. Moore, J. Coleman Eubanks, A. Aiyaz, and R. P. McMahan. Evaluating interaction cue purpose and timing for learning and retaining virtual reality training. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction, SUI '20*. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3385959.3418448 8
- [15] J. Lacoche, E. Villain, and A. Foulonneau. Evaluating usability and user experience of ar applications in vr simulation. *Frontiers in Virtual Reality*, 3:881318, 2022. 3
- [16] E. Laviola, M. Gattullo, A. Evangelista, M. Fiorentino, and A. E. Uva. In-situ or side-by-side? a user study on augmented reality maintenance instructions in blind areas. *Computers in Industry*, 144:103795, 2023. doi: 10.1016/j.compind.2022.103795 2
- [17] C. Lee, S. Gauglitz, T. Höllerer, and D. A. Bowman. Examining the equivalence of simulated and real ar on a visual following and identification task. In *2012 IEEE Virtual Reality Workshops (VRW)*, pp. 77–78. IEEE, 2012. 3
- [18] M. A. Livingston and Z. Ai. The effect of registration error on tracking distant augmented objects. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp. 77–86. IEEE, 2008. 2, 3, 8
- [19] F. G. Paas and J. J. Van Merriënboer. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors*, 35(4):737–743, 1993. 3, 6
- [20] E. Ragan, C. Wilkes, D. A. Bowman, and T. Hollerer. Simulation of augmented reality systems in purely virtual environments. In *2009 IEEE Virtual Reality Conference*, pp. 287–288. IEEE, 2009. 3
- [21] C. M. Robertson, B. MacIntyre, and B. N. Walker. An evaluation of graphical context as a means for ameliorating the effects of registration error. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):179–192, 2008. 2, 8
- [22] S. Rothe, D. Buschek, and H. Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1), 2019. doi: 10.3390/mti3010019 2
- [23] B. Schwerdtfeger and G. Klinker. Supporting order picking with augmented reality. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp. 91–94, 2008. doi: 10.1109/ISMAR.2008.4637331 2
- [24] B. Schwerdtfeger, R. Reif, W. A. Günthner, and G. Klinker. Pick-by-vision: there is something to pick at the end of the augmented tunnel. *Virtual reality*, 15:213–223, 2011. 8
- [25] D. Stover and D. Bowman. Taggar: General-purpose task guidance from natural language in augmented reality using vision-language models. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*, pp. 1–12, 2024. 2
- [26] T. A. Syed, M. S. Siddiqui, H. B. Abdullah, S. Jan, A. Namoun, A. Alzahrani, A. Nadeem, and A. B. Alkhodre. In-depth review of augmented reality: Tracking technologies, development tools, ar displays, collaborative ar, and security concerns. *Sensors*, 23(1):146, 2022. 2
- [27] 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 8
- [28] J. O. Wallgrün, M. M. Bagher, P. Sajjadi, and A. Klippel. A comparison of visual attention guiding approaches for 360° image-based vr tours. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 83–91, 2020. doi: 10.1109/VR46266.2020.00026 3
- [29] G. Welch and L. Davis. Tracking for training in virtual environments: Estimating the pose of people and devices for simulation and assessment. *The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond: Praeger Security International*, 2008. 2
- [30] G. Welch and E. Foxlin. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Computer graphics and Applications*, 22(6):24–38, 2002. 2
- [31] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011. 5