

Assessing Depth Perception in VR and Video See-Through AR: A Comparison on Distance Judgment, Performance, and Preference





Franziska Westermeier , Larissa Brübach , Carolin Wienrich , and Marc Erich Latoschik 



Fig. 1: Digital twin (a) of a real office room (b) used in the comparison of depth perception between VR and VST AR. The participants are seated on the stool and perform different depth-dependent tasks. Several objects (e.g., the cones, sphere, and ring on the table or the markers on the floor) are placed at various locations where the individual tasks are performed.

Abstract—Spatial User Interfaces along the Reality-Virtuality continuum heavily depend on accurate depth perception. However, current display technologies still exhibit shortcomings in the simulation of accurate depth cues, and these shortcomings also vary between Virtual or Augmented Reality (VR, AR: eXtended Reality (XR) for short). This article compares depth perception between VR and Video See-Through (VST) AR. We developed a digital twin of an existing office room where users had to perform five depth-dependent tasks in VR and VST AR. Thirty-two participants took part in a user study using a 1×4 within-subjects design. Our results reveal higher misjudgment rates in VST AR due to conflicting depth cues between virtual and physical content. Increased head movements observed in participants were interpreted as a compensatory response to these conflicting cues. Furthermore, a longer task completion time in the VST AR condition indicates a lower task performance in VST AR. Interestingly, while participants rated the VR condition as easier and contrary to the increased misjudgments and lower performance with the VST AR display, a majority still expressed a preference for the VST AR experience. We discuss and explain these findings with the high visual dominance and referential power of the physical content in the VST AR condition, leading to a higher spatial presence and plausibility.

Index Terms—Depth perception, VR, AR, video see-through, egocentric distance judgment, task performance, user preference



1 INTRODUCTION

In recent years, Spatial User Interfaces (SUIs) such as Virtual Reality (VR) and Augmented Reality (AR) have captivated many areas. These platforms promise to redefine our interaction with digital content, incorporating seamless integration into our physical world (AR) or immersing us entirely in artificial environments (VR). While the potential applications of VR and AR span diverse fields, such as entertainment, health, education, or maintenance, a fundamental understanding of how

we perceive and interact within these environments remains a topic of ongoing research. In computer-generated environments, a major challenge is to place the virtual content in the right depth and to provide a coherent set of depth cues to enable users to perceive this depth correctly and make sense of it. In AR, an additional challenge lies in combining depth cues from the virtual and the real world to result in the perception of a congruent scenario [28, 44]. In VR, users perceive one congruent scenario in which depth cues affect all virtual content similarly. In AR, contradicting depth cues between virtual and physical content can lead to misinterpretation of spatial information, potentially affecting user performance, safety, and overall immersion.

The quality of blending virtual and physical content is confined by a constellation of factors, including (1) hardware constraints such as latency, optical distortions, and tracking inaccuracies that can result in the misplacement of virtual objects within the real environment, and (2) disparities in the appearance of virtual and physical components, encompassing differences in color and illumination. There is a great body of knowledge addressing these perceptual incongruencies and how different technologies cope with these [2, 4, 9, 10, 27]. Different AR display technologies inherit different incongruencies. In optical see-through (OST) AR displays, users can directly view the environment, and virtual content is added as an overlay by a virtual combiner. Since

- Franziska Westermeier and Larissa Brübach are with the Human-Computer Interaction (HCI) Group and the Psychology of Intelligent Interactive Systems (PIIS) Group from the University of Würzburg.
- Carolin Wienrich is with the PIIS Group from the University of Würzburg.
- Marc Erich Latoschik is with the HCI Group from the University of Würzburg.

E-mail: {franziska.westermeier | larissa.bruebach | carolin.wienrich | marc.latoschik} @uni-wuerzburg.de

Manuscript received 4 October 2023; revised 17 January 2024; accepted 24 January 2024. Date of publication 4 March 2024; date of current version 15 April 2024.
Digital Object Identifier no. 10.1109/TVCG.2024.3372061

the view of the environment stays undistorted, the depth perception in OST AR displays shows more accurate results compared to VR [19, 36]. In contrast, the video see-through (VST) AR display uses a real-time video stream with virtual content added on top of this video stream. Compared to OST AR displays, depth estimations in VST AR displays are less accurate [1, 3].

While a lot of research has been conducted to examine depth perception in OST AR displays, VST AR displays remain underexplored as empirical studies are rare. Since consumer Head-Mounted Displays (HMDs) increasingly offer VST functionality besides classic VR, new use cases arise that make use of both (VST AR and VR) and enable transitions along the Reality-Virtuality (RV) continuum between reality and virtuality without the necessity of switching the HMD. Hence, it is particularly important to determine whether (depth) perceptions at different points on the RV continuum [32] are comparable, facilitating the application of VR research findings to various other forms. To our knowledge, a direct comparison between VR and VST AR concerning depth perception has not been conducted so far. Thus, in this paper, we attempt to answer the following research question: "Is there a difference concerning depth perception between VR and VST AR?"

Display-mediated visual perception is potentially influenced by a variety of display characteristics, also including the ergonomics (e.g., wear comfort and weight) of HMDs. However, many comparative studies use different HMDs for the assessment, inhering different influences of HMD characteristics in their results on depth perception.

In our work, we present a comparison of depth perception between VR and VST AR with the Meta Quest Pro. Using one HMD for both conditions minimizes the possible side effects of hardware characteristics. We introduce a set of tasks to assess egocentric depth perception in VR and VST AR and collect empirical data on depth perception, task performance, and preference in a user study. Our findings enhance the understanding of depth perception within SUIs and highlight the challenges present in this domain.

2 RELATED WORK

VR and AR can be located on Milgram's RV continuum [32] and represent different forms of Mixed Reality (MR) [40]. VR applications are situated near the right endpoint *Virtuality*. AR applications can be located between *Virtuality* and *Reality* since AR technology augments the physical environment with a virtual overlay, accounting for different real and virtual proportions. Building upon this continuum, AR displays have evolved in various forms. Handheld and projector-based devices are complemented by AR HMDs [27, 33], further divided into VST and OST AR displays. The VST AR displays use external cameras to capture the environment and stream the image directly into the HMD. Virtual content is added as an overlay. The main advantage of this setup is the high control over the environment. The scene is discretized and situated in the same pixel rasterization as the virtual content enabling the adaptation of visual coherence. Disadvantages of VST AR displays are the reduced resolution due to image compression, lens distortions, and time lags, which could ultimately contribute to a wrong or distorted depth perception. OST AR displays use optical combiners to project content on collimating lenses. This technology has the main advantage of a direct view of the environment, as there is no image compression.

Most AR-related research has been conducted with OST AR HMDs. However, in recent years, VST AR HMDs (such as the Varjo XR-3, the Meta Quest Pro, and the Meta Quest 3) reached an acceptable display quality to become a valuable technology to apply in different areas, such as education, training, health, manufacturing, or entertainment. Future release announcements appear even more promising, strengthening the desideratum to focus more research on this display technology.

With regard to the precise interactions that are required in AR-supported operations (such as surgery or maintenance), perceiving depth correctly is essential. Depth perception is an important part of our visual sense-making [17], and depth cues support the perception of space, such as stereo-vision, motion parallaxes, object occlusion, or perspective vision.

For VR displays, advances in computer graphics and rendering are already well-established to provide depth information for the user

[11, 18]. In VST AR, *virtual* content can be rendered accordingly. However, the challenge exists to match the depth cues rendered for the virtual objects with those from the physical environment. Not only the visualization of objects but also a wrong registration in the environment, lens distortion, latency, etc., can lead to conflicting visual-visual stimuli that users easily detect and might be disturbed by [2].

Various tasks have been developed to assess depth perception. These include verbal reports of distances towards virtual objects, bisection tasks (where participants mark the half distance to virtual objects), or blind actions [1, 16]. In blind actions, participants see virtual objects for a while. They then have to reach or walk blindly to the position where they estimate an object. Because the requirements of these tasks are very diverse (e.g., some require motor skills, while others only include perceptual/cognitive processing, some are continuous, and others are static), it is hard to establish a common ground controlling for task-specific confounds.

In VR, a systematic underestimation of distances was observed [16, 19, 20, 37]. Kelly [20] conducted a literature review and summarized an average underestimation ratio of 73.48 % (where 100% would be an accurate estimation and value > 100 % would be an overestimation) of the actual distance. He concluded that a wider field of view, less weight, and a higher pixel density of the HMD lead to a more accurate depth estimation. Because HMDs have improved on these aspects in recent years and will further improve, depth judgments among newer HMDs are expected to become more accurate. Willemsen et al. [47] also examined the problem of distance underestimation and mechanical aspects that might influence depth estimation. To some extent, these mechanical attributes of the HMD (weight, moments of inertia) account for the distance compression. However, the authors assumed there must be other perceptual aspects of why users underestimate distances in virtual environments. In another work, Kelly et al. [21] measured depth estimations in the Meta Quest and Meta Quest 2. The results from a verbal report revealed an underestimation of 82% (Meta Quest) and 75% (Meta Quest 2) compared to a real-world estimation of 94% of the actual distance.

Jones et al. [19] examined depth judgments in VR and OST AR displays with a blind walking task. While they did not detect a distance underestimation in the OST AR display, they found an underestimation effect for the VR condition. They added an additional factor of motion parallax since they expected that higher motion while viewing the object would contribute to a better estimation. In the control condition, participants were asked to remain in one position. Against their expectation, the authors could not find an effect of motion parallax on the depth judgment between VR and OST AR. Ping et al. [36] implemented a task to move a virtual bar back and forth to match the distance of a ball on a shuffleboard. In VR, the ball and the shuffleboard were virtual. In the OST AR view, a real shuffleboard and ball were provided, even though the ball was also displayed virtually. They measured a higher accuracy in the OST AR condition. Ping et al. [36] and Jones et al. [19] both measured a higher error the farther the target objects were away. Cidota et al. [8] enhanced VR and OST AR with visual effects, i.e., blur and fade effects, to investigate if these effects alter the perception and performance in their system. Participants performed grasping and sorting tasks into boxes at different depths. They did not measure a difference between VR and OST AR in their control condition. Their results further showed that the induced visual effects disturb the performance of the OST AR condition, while they contributed to better results in VR.

While these studies examine OST AR displays, only a few studies exist on depth perception in VST AR displays. Messing and Durgin [31] compared distance perception of a real-time monocular video stream in a VR HMD (Virtual Research Systems V8 HMD) and direct viewing with monocular goggles and a cardboard tube to simulate a restricted field of view. In a blind walking task, participants were asked to walk distances to targets between 2 and 7m. While the monocular goggles almost reached 100 % accuracy, there was an underestimation of 77 % in the HMD. Similarly, Pfeil et al. [35] examined distance perception with a blind throwing task between a stereoscopic VST view (HTC Vive equipped with a ZED Mini pass-through camera), an unrestricted

real-world view and a restricted real-world view realized through a plastic casing from a stripped-down HMD. They found a higher underestimation in the VST view (93%) compared to the other conditions. Even though Messing and Durgin [31] and Pfeil et al. [35] examine the VST view, they do not integrate virtual objects in their applications, which does not conform with the definition of AR. Therefore, incongruencies by visual and spatial mismatch are not addressed. Vaziri et al. [42] assessed the depth perception of a virtual object in three different VST AR conditions. While in one condition, full visual detail of the environment was provided, the other conditions showed a sketch-like environment and no environment at all, respectively. Measured in a blind walking task, they discovered that the depiction of the environment has no influence on depth perception. Ballestin et al. [3] compared VST AR to the OST AR of the Meta 2 headset by MetaVision. The VST AR view was rendered on a smartphone mounted in front of the eyes. In a reaching task, participants significantly underestimated the distance to virtual objects in VST AR compared to the OST AR condition. The monocular nature of the camera image that represented the VST view might have contributed to this outcome since it omits stereoscopic depth information. Adams et al. [1] investigated OST and VST AR in combination with shadow cues and different heights of objects in space. They used the Microsoft HoloLens 2 to represent OST AR and a Varjo XR-3 for VST AR. They found out that the application of shadow cues has only a little effect on depth judgment. When virtual objects were floating in space, they were judged as farther away. Overall, the authors could replicate Ballestin et al.'s results of more underestimation in VST AR than in OST AR.

Differences in depth judgments between VR and VST AR remain unclear, as we found *no* studies that directly compare these two SUIs in the same setting. However, VST AR and VR seem to incorporate a higher underestimation than OST AR [1, 3, 19, 36]. The reason for this might be the distortion of the display [16]. To achieve a high field of view in the VST display, camera lenses are distorted to capture more content, i.e., straight lines appear curved [27]. Other influencing factors are the field of view, the weight, and the resolution [20, 47]. We conclude that the hardware specifications of HMDs seem to have a certain impact on the measured distance underestimation in HMDs. Most comparative studies on depth perception use different HMDs incorporating different hardware specifications and, thus, different influences on depth estimations. Therefore, we propose investigating aspects independently of HMD-specific characteristics, to better understand perceptual aspects and evaluate more fine-grained influences among different SUIs.

While depth judgments can be measured more or less directly with the tasks previously described, an indirect measurement is the task performance, which results from the quality of the depiction of depth cues and the correct depth perception. These performance measures (such as task completion time) also show the extent to which perception affects action. Only a few studies exist that examine task performance in VST AR displays. Krichenbauer et al. [26] examined a simple selection and placement task in nine degrees of freedom (position, rotation, scale) in VR and VST AR with a 3D input device and measured a higher completion time in the VR condition. Furthermore, they measured more head movement in the AR condition, which could be an indicator of absent or conflicting depth cues that participants then counteracted with motion parallaxes. Kern et al. [23] investigated different keyboard input modalities in VR and VST AR. Contrary to Krichenbauer et al.'s findings [26], Kern et al. found a significantly higher completion time in VST AR than in VR, i.e., participants typed faster in VR. Discrepancies in the results might arise from the nature of the tasks that participants had to perform. Text input might require more cognitive resources. Kern et al. [23] explain their results with the *Congruence and Plausibility (CaP) model* by Latoschik and Wienrich [28] which defines a manipulation space with three layers: the sensation, the perception, and the cognition layer (i.e., bottom-up to top-down). On each layer, (in)congruence can be manipulated, resulting in a condition of plausibility. In the VST AR condition, incongruencies lead to a visual mismatch that participants actively need to counteract on a cognitive level, resulting in a lower performance.

Westermeier et al. [44] manipulated the cognitive congruence of a scenario in VR and VST AR. They implemented two different effects of a power outage: the cognitive congruent power outage affected the whole scenario, while the incongruent power outage only affected virtual interaction objects. In VST AR, they manipulated the physical environment with smart lights that were triggered simultaneously with the participants' actions. They found effects on the perceived scenario plausibility and spatial presence, i.e., the feeling of "being there" [29]. In VR, they measured that the cognitive congruent power outage triggered higher plausibility and spatial presence ratings. This effect was inverted in AR. The congruent power outage (which triggered the lighting in the physical environment) performed worse than the incongruent power outage regarding the plausibility and spatial presence ratings. The authors assumed that due to the visual mismatches that VST AR contains, participants could not combine physical and virtual content into one congruent scenario. Following the CaP model's assumptions and previous findings, we thus predict that the depth perception may be violated more in VST AR than in VR due to the contradicting a priori cues causing a visual mismatch.

There is little research on both task performance and the perception in VST AR, i.e., how it may affect other evaluations of the experience, such as plausibility or the sense of presence [44]. Here, we see another research gap as these ratings are important for good XR experiences.

3 SUMMARY AND PRESENT STUDY

Derived from the existing literature on depth perception, there is a lack of direct comparisons between VR and VST AR. However, previous studies [1, 3, 19, 36] revealed distance underestimations in both display technologies. Building on the literature [28, 44], we anticipate that the higher amount of incongruencies stemming from the combination of virtual and physical content might lead to visual mismatches, which further distort the depth judgment. As it is described by Westermeier et al. [44], AR inheres "slight reconstruction errors caused, for example, by inaccuracies or imprecisions of object tracking or unknown parameters of the current real-world light transport, given the used AR device, rendering engine, and sensory equipment" [44, p.2682]. Previous work by Azuma [2] and Kruijff et al. [27] discussed the perceptual issues of AR displays and the accompanying conflicting cues from real and virtual entities. Due to this, we believe it is harder to set and estimate a virtual object in the physical environment than it is in the virtual environment, which motivates our first hypothesis:

- **H1:** Distance judgments in VST AR are less accurate than those in VR.

Furthermore, we hypothesize a difference in task performance (i.e., time and error rate) between the two display technologies [23, 26]:

- **H2:** The task performance is lower in VST AR than in VR.

Considering the visual incongruencies inherent to VST AR, we predict potential implications on the user's perceived spatial presence and scenario plausibility [44]. As such, we propose:

- **H3:** Users report a higher spatial presence in VR than in VST AR.
- **H4:** Users report a higher perceived plausibility of the scenario in VR than in VST AR.

As we hypothesize superior outcomes of VR over VST AR concerning depth perception and task performance, we expect that participants will prefer VR over VST AR:

- **H5:** Users will prefer VR over VST AR.

We implemented a 1×4 within-subjects design, utilizing a counter-balanced randomized order structured as a 4×4 Latin square. Our study comprises four conditions: a pure VR condition, a VST AR condition, and two additional exploratory conditions simulating AR *in* VR. For this simulation, we used the implementation presented in Westermeier et al.'s work [45]. We induced noise and a low resolution to the simulated VST video stream (condition VAR, see Fig. 2a). We further increased the lens distortion (barrel distortion) in the simulated VST

video stream (condition VAR+, see Fig. 2b). While this manipulation affected the environment, the interaction objects remained untouched by the noise and lens distortion. We intended to cause a visual mismatch and, thus, to validate the AR simulation [45]. For the purposes of this work, we are focusing exclusively on the VR and VST AR conditions, as the other conditions are out of the scope.

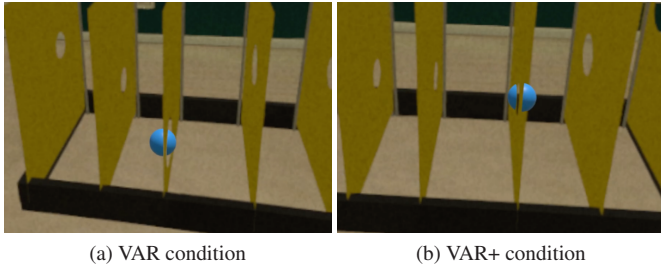


Fig. 2: The omitted conditions. We lowered the resolution of the VST view and added noise. In (a), the occlusion is aligned between the sphere and the simulated VST view. Due to a discrepancy of lens distortion in (b), the occlusion is not coherent between the sphere and the simulated VST view.

We decided on the Meta Quest Pro as HMD, taking advantage of its pass-through functionality to ensure consistent pixel rasterization for both VR and VST AR conditions. From the literature, we know that hardware characteristics influence depth perception. By the consistent use of only one HMD, we can control all the possible hardware-specific effects and keep inherent display characteristics consistent, focusing on differences in visual perception only.

In the VR scenario, we replicated the real office room into a virtual version (see Fig. 1a). For tasks involving participant interaction, virtual objects were employed. However, in the VST AR condition, while the interaction was still with virtual objects, reference objects (relevant but non-manipulable for the tasks) were physical (see Fig. 4).

4 METHOD

4.1 Participants

Thirty-five participants took part in the experiment. Due to technical issues, three participants were excluded, resulting in 32 remaining participants for data analyses. The participants' demographic distribution and XR experience can be seen in Tab. 1. The study was approved by the institution's ethics committee.

Table 1: Demographic data and XR experience of participants.

Attribute	Description (N=32)
Biological Gender	19 female; 12 male; 1 diverse
Social Gender	19 female; 13 male
Age	$M = 30.56$ ($SD = 11.56$)
VR Experience - Duration	4 <1h; 7 1-3h; 6 3-5h; 4 5-10h; 6 10-20h; 5 >20h
VR Experience - Frequency	0 never; 9 1-3 times; 9 3-6 times; 6 6-10 times; 3 10-20 times; 5 >20 times
AR Experience - Duration	17 <1h; 11 1-3h; 3 3-5h; 1 5-10h
AR Experience - Frequency	14 never; 13 1-3 times; 3 3-6 times; 2 6-10 times

4.2 Apparatus

Our experiment was conducted on a high-performance computer equipped with an Intel i9-11900K CPU, an NVIDIA GeForce RTX 3080 GPU, and 64 GB of RAM. We used the Meta Quest Pro HMD.

This device offers pass-through functionality and, thus, consistent visual parameters for both VR and AR modalities. An advantage of the Meta Quest Pro, compared to other VST AR devices, is its relatively small distortion in the camera pass-through. This pass-through is realized by two gray-scale cameras (enabling stereo vision) and an RGB camera, which overlays color onto the gray-scale images. For interaction, we utilize the Meta Quest Touch Pro controllers.

Our application was implemented in Unity (v2021.3.27f1) using the Universal Render Pipeline (v12.1.12) for rendering. To link the Meta Quest Pro with our computer setup, we connected the HMD to the computer via the Oculus Link cable and utilized the Reality Stack I/O framework developed by Kern and Latoschik [22] for HMD support.

For the subsequent statistical evaluations, we used Python (v3.8.17) and employed the Pingouin package (v0.5.3) [41].

4.3 Procedure

The study procedure can be seen in Fig. 3. It took about 1.5 hours in total. Participants began by completing the consent forms. Subsequently, they filled out pre-questionnaires covering demographics, media usage, prior experiences with VR and AR, current VR sickness status, and their aptitude in visual imagery.

Participants then put on the HMD. In the beginning, participants adjusted the interpupillary distance of the HMD lenses until they saw a clear and unimpaired image. They were placed in a black environment and saw a white cube to refer to when adjusting.

Participants completed a tutorial phase to familiarize themselves with the system. Here, participants engaged with primitive objects, following auditory instructions. A consistent black background was maintained to eliminate potential distractions or confounding factors from the physical or virtual environment.

The main experience was divided into four blocks, each representing a within-condition. Each block commenced with calibrating the HMD and controllers to the virtual space. Participants executed a series of five tasks guided by auditory cues. Upon task completion, they responded to a set of questionnaires. This structure was repeated for all four within-subjects conditions.

In the end, participants reported their VR sickness status. Additionally, they provided insights through a few retrospective questions, concluding the experimental procedure.

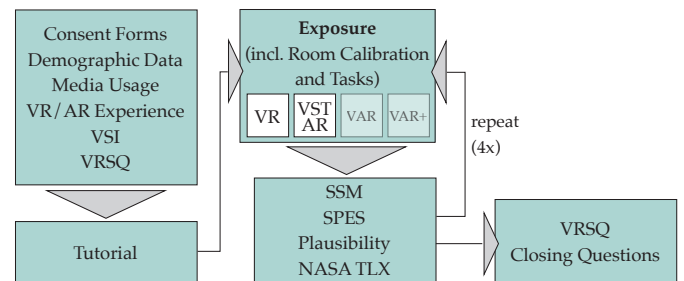


Fig. 3: The experimental procedure.

4.4 Tasks

Participants were required to complete five distinct tasks (see Fig. 4). These tasks were designed both to reflect established methods from prior studies (tasks 1 and 2), to explore the alignment of virtual and physical objects (tasks 3 and 4), and to get insights into task performance (task 5). One requirement for the selection of tasks was that participants did not engage too much with their own body to avoid confounds between VST AR (the own body is visible) and VR (the own body is not visible). With this selection of tasks, we wanted to cover different aspects: they vary in depth range, motor activity, and difficulty. All tasks were situated in the same room; participants would focus on the task at hand. Participants primarily used the controller thumb stick for interactions for tasks requiring object movement or rotation.

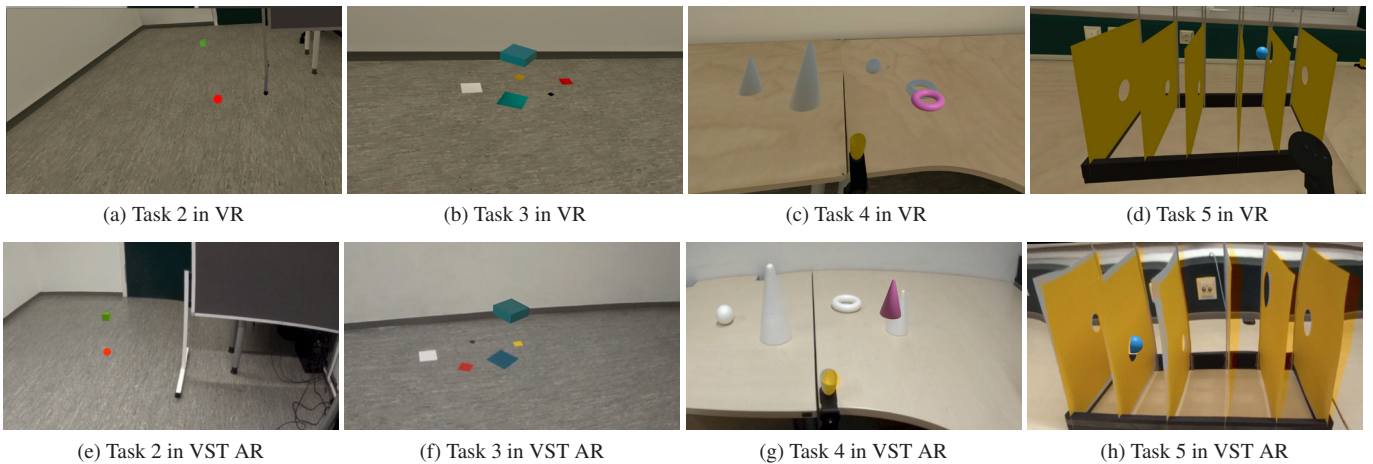


Fig. 4: The different tasks in VR and VST AR. Task 1 showed virtual cubes in the same manner as it is depicted for task 2 (but without the red marker). In tasks 3, 4, and 5, the reference objects' positions were fixed, while the objects were randomly assigned to these fixed positions in the VR condition. In VST AR, the assignment was always fixed.

Task 1: Verbal report Inspired by existing literature [1, 11], this task required participants to verbally report depth estimations for five virtual cubes. The cubes in this task were virtual, while the depicted environment was virtual (VR) or the video stream of the real environment (VST AR). Distances to the cube positions ranged from 85 cm to 315 cm. Estimations were logged by the experimenter.

Task 2: Bisection task We refer to existing literature and conducted a bisection task [6]. Participants placed a virtual red marker midway between themselves and five sequentially appearing virtual cubes. This task followed a similar structure to task 1, with the same distance range and no reference objects (see Fig. 4a and Fig. 4e). A sphere spawned in front of the participants. With the controller thumbstick, it could be moved towards the participant's body (the endpoint was defined as the position of the HMD, but at half the height) and the virtual cube's position.

Task 3: Vertical alignment We adopt Adams et al.'s [1] idea of examining the depth perception of floating objects. In addition, we intended to offer tasks with different levels of difficulty and believe the vertical offset made the depth judgment harder. Participants were presented with five virtual objects to map over reference markers with a vertical offset. In addition to accurate positioning, the virtual objects had to be rotated to align with the marker orientation (see Fig. 4b). The VST AR condition featured reference markers placed on the physical room's floor (see Fig. 4f).

Task 4: Depth alignment Similarly to Ping et al.'s [36] approach, we include a depth alignment task in our set of tasks, where participants need to move four virtual objects back and forth to align with reference objects (see Fig. 4c). Compared to task 3, this task was simpler and also positioned nearer to the participants. However, some difficulty was induced by omitting occlusion handling in this task. Styrofoam primitive forms served as the physical reference objects in the VST AR condition (see Fig. 4g).

Task 5: Hotwire game We adopt the hotwire game initially proposed by Lugin et al. [30] as a benchmark for tracking the quality of 3D interaction. Participants steered a sphere (attached to their controller) through holes in aligned panels. These holes varied in size and position (see Fig. 4d). In the VST AR version, a 3D-printed foot and cardboard cutouts replicated the virtual setup (see Fig. 4g). The panels had dimensions of 30x30 cm, and the hole sizes ranged from 4.5 cm to 7.5 cm. The panels were lined up with distances of 10 and 11 cm. Participants conducted three iterations from left to right.

Randomization Various randomizations were employed to ensure that participants would not get accustomed over the course of four conditions. In tasks 1 and 2, seven predefined positions existed, from

which five were chosen randomly in each task iteration. For tasks 3 - 5, positions were fixed. However, the reference objects associated with each position were randomized, with the exception of the VST AR condition, where randomization was skipped for simplicity.

4.5 Implementation and Mitigation of Confounds

Some differences between VR and AR persist, which brings along an unknown amount/intensity of confounds. Thus, we took some countermeasures to minimize possible confounding effects.

Digital twin and visual consistency We used a virtual replica of the real physical room in the exact dimensions, including the same composition of furniture [34]. We developed task-specific models for our virtual scene and also created corresponding 3D-printed versions for our physical scene. Hence, we could ensure uniformity between VR and VST AR. A static lighting model illuminated the virtual scene to minimize computing time. While interactable objects had dynamic real-time lighting, their shadows were omitted due to potential inconsistencies between VR and VST AR and negligible impact on depth perception as indicated by Adams et al. [1]. In task 4, we manipulate the occlusion of the reference objects in both the VR and the VST AR condition to always render the interaction object in front. We create an occlusion model for the remaining tasks and reference objects in VST AR, i.e., a replica of the virtual room with a specific material. This material is rendered at a late step in the render pipeline and replaces all the pixels with a higher depth value. Hence, the virtual object could be occluded by physical content.

Calibration and spatial consistency We introduced a room calibration procedure to align physical objects with their virtual replica. The experimenter conducted this room calibration at the beginning of each condition. We defined two virtual reference points in the virtual scene. One controller was used to define the position offset to the first reference point, which was added to the XR rig. We calculated the rotational offset angle between the second controller and the other reference point and rotated the XR rig around this angle. We maximized the distance between the two controllers to minimize a potential rotational error. We ensured the correct and consistent controller position by installing 3D-printed mounts [24] tailored for the Meta Quest Touch Pro controllers in our physical room.

Body perception Previous work has proven that a higher embodiment improves the accuracy of depth judgments [38]. To align body perception between VST AR and VR conditions as closely as possible, participants' real bodies were covered with a hairdressing cape. A virtual replica of the covered body was created, which moved in synchrony with the head movement. Movement-centric tasks were minimized, so only task 5 required active 3D controller movement in space.

Size references Covering their own body also had the effect that participants could not set the size of their body parts in relation to the distances they had to judge. In addition, we removed unused physical furniture and objects from the room to prevent biases from familiar objects and their sizes. We kept the virtual and physical environments as similar as possible to reach an acceptable similarity of level of detail.

Experimenter presence The experimenter needed to stay in the same room to observe the experiment. To avoid a confounding co-presence effect in AR, the experimenter was strategically positioned behind a poster wall, ensuring non-visibility to the participant.

4.6 Measures

4.6.1 Objective Measures

We use the term “judgment” for the assessment of distances. In task 1, the judgment includes an estimation of distance. In all other tasks, the judgment additionally includes the positioning and steering of virtual objects. Thus, task 1 includes perceptive and cognitive resources. All other tasks include motor activity as well. Here, participants make continuous readjustments while placing/steering the virtual objects. Hence, we define judgment as the result of both the estimation and (if applicable) the subsequent motor activity. We utilized the concept of a *distance ratio* across multiple tasks. The ratio is defined as:

$$\text{distance ratio} = \frac{\text{judged distance}}{\text{actual distance}} \quad (1)$$

In task 1, participants provide an estimation of the egocentric distance towards objects. In task 2, the distance to the set marker position was used as the judged distance. In contrast, the distance to the actual calculated midpoint between the object and participant is the actual distance. In tasks 3 and 4, we proceeded similarly by using the reference objects’ positions for the calculation of the actual distance, while the objects placed by the participants provided the location for the calculation of the judged distance.

To account for possible offsets in the judgments (e.g., on average, an overestimation could cancel out an underestimation), we also compute the *absolute misjudgment*:

$$\text{absolute misjudgment} = |\text{distance ratio} - 1| \quad (2)$$

In task 5, we calculate an error rate by counting the frames in which a collision between the sphere and panels is detected and dividing it by the total amount of frames needed for one iteration.

As performance metrics, we measure the needed time for each task as well as motion data (i.e., participants’ head and controller position and rotation).

4.6.2 Subjective Measures

Besides objective measures, we use questionnaires to assess the participants’ perception. We ask for the *Spatial Situation Model (SSM)*, designed by Vorderer et al. [43]. It contains questions concerning the construction of a spatial model of the viewed scene, such as “Even now, I still have a concrete mental image of the spatial environment.” [43]. The SSM is defined to be a precondition for spatial presence. The SSM contains eight items that are answered on a five-point Likert Scale from “I do not agree at all”(1) to “I fully agree”(5).

In the beginning, we ask participants for their *Visual Spatial Imagery (VSI)*, which gives insights into the participants’ individual preconditions concerning the recreation of spatial information [43] (e.g., “When someone describes a space to me, it’s usually very easy for me to imagine it clearly.”). Vorderer et al. describe in their work that it can influence the SSM. The VSI is answered on a five-point Likert Scale ranging from “I do not agree at all”(1) to “I fully agree”(5).

We measure spatial presence by using the *Spatial Presence Experience Scale (SPES)* with the two subscales *Possible Actions* and *Self Location* [14]. It includes eight items (four per subscale) with the endpoints “I do not agree at all”(1) and “I fully agree”(5). The SSM, the VSI, and the SPES were all developed in the broader frame of the MEC-SPQ [43].

Similarly to Westermeier et al. [44], we assess the perceived plausibility by using their proposed questions (inspired by Brübach et al. [7]) in an adapted form (e.g., “This experience was unusual for me” or “I could not anticipate what would happen next in the scenario”). The questions are answered on a seven-point Likert Scale from “I do not agree at all”(1) to “I fully agree”(7).

To assess the subjective task load, we ask the NASA TLX questions on mental and physical load [13] using a slider ranging from 0 to 20.

At the end of the experiment, participants were asked to decide which condition they preferred, which condition they perceived as most complex, and which condition they perceived as easiest.

As a control measure for VR sickness, we asked participants to fill in the *Virtual Reality Sickness Questionnaire (VRSQ)* [25] before the first exposure started and after the last exposure ended. The VRSQ is answered on a four-point Likert Scale ranging from “None”(0) to “Severe”(3).

4.7 Hypothesis Testing and Task-Based Analysis

For H1, we involve results from tasks 1 and 2 (the *distance ratio* and *absolute misjudgment*). Tasks 3 and 4 also provide information on the accuracy. However, they include reference objects from the physical environment and, thus, depend on the correct calibration of the room. Tasks 3 and 4 are more implicit and active and provide more relevance for real-world scenarios. H2 is measured by the time needed to fulfill the tasks. Task 5 furthermore provides an error rate, which can be used as a measure of task performance. However, again, the results of this measure highly depend on the calibration quality. We answer H3 and H4 with results from the SPES and the perceived scenario plausibility questionnaire. H5 is answered by the preference rating participants provide at the end of the experiment.

5 RESULTS

5.1 Objective Measures

Objective results were obtained from the tasks to determine the depth judgments and task performance in VR and VST AR. A comprehensive overview of these results is provided in Tab. 2.

By visual inspection, we noticed some outliers that we trace back to issues with the controller interaction (e.g., we observed that participants accidentally confirmed their choice instead of placing the object at the right location because they confused controller buttons). To mitigate these outliers, we conducted a winsorizing over the data of both conditions with 0.05 as the lower and 0.95 as the upper limit. We conducted the winsorizing only for the judgments of tasks 2, 3, 4, and 5, as there was no controller action required in task 1.

Some of our results did not meet the assumptions of normality. However, ANOVAs have been found to be resilient to such deviations [5, 15, 39]. Therefore, repeated measures ANOVAs were used to compare VR and VST AR, with a significance threshold set at $p < .05$. T-tests were conducted to compare distance ratios against the expected value of 1.

5.1.1 Comparison of Depth Judgments to Ground Truth

In Task 1, the VST AR condition showed an average underestimation of distances by 12.2 % ($M = 0.878$, $SD = 0.219$), deviating significantly from 1.0 ($t(31) = -3.16$, $p = .004$). Contrastingly, VR estimations did not significantly differ from the expected 1.0. The distances in task 2 were judged significantly higher than the actual distance in both VR ($t(31) = 7.41$, $p < .001$) and VST AR ($t(31) = 7.87$, $p < .001$).

5.1.2 Comparison of Depth Judgments between VR and VST AR

For task 1, we measured a significant difference in distance ratios between VR and VST AR (see Fig. 5a). According to the p-values, we did not find significant differences between VR and VST AR concerning absolute misjudgments. In general, the depth was underestimated by around 10 %. The absolute misjudgments show that participants guessed the distance wrong by more than 20 % on average. In addition, the standard deviation values are very high, indicating a high variance in misjudgments.

Table 2: Results from calculating repeated measures ANOVAs of the objective measures between VR and VST AR. We report the mean (M), the standard deviation (SD) as well as the test statistic F , the p -value and the partial eta squared (η_p^2). Significant p -values and the mean values of the respective condition with higher accuracy, less movement, and less time are marked in bold. Distance ratios and absolute misjudgments for tasks 2 - 5 were winsorized to mitigate outliers.

Task	VR (M and SD)	VST AR (M and SD)	F (1, 31)	p	η_p^2
Distance Ratios					
Task 1	0.942 (0.246)	0.878 (0.219)	4.46	.043	.126
Task 2	1.111 (0.085)	1.139 (0.100)	7.27	.011	.190
Task 3	1.005 (0.008)	1.011 (0.010)	7.11	.012	.187
Task 4	1.001 (0.001)	1.003 (0.008)	4.41	.044	.124
Absolute Misjudgments					
Task 1	0.221 (0.147)	0.229 (0.150)	0.10	.756	.003
Task 2	0.129 (0.065)	0.165 (0.070)	14.19	<.001	.314
Task 3	0.012 (0.006)	0.017 (0.009)	11.87	.002	.277
Task 4	0.001 (0.001)	0.010 (0.008)	37.85	<.001	.550
Error Rate (collision frames/total frames)					
Task 5	0.072 (0.032)	0.231 (0.059)	333.30	<.001	.915
Completion Times (s)					
Task 1	57.97 (12.48)	63.16 (14.96)	3.99	.055	.114
Task 2	40.97 (12.03)	40.78 (12.49)	0.01	.938	<.001
Task 3	104.25 (25.06)	102.84 (21.08)	0.07	.791	.002
Task 4	47.94 (5.62)	63.47 (15.78)	27.96	<.001	.474
Task 5	61.25 (10.40)	67.38 (16.05)	7.37	.011	.192
Total	312.38 (47.45)	337.63 (53.65)	4.56	.041	.128
Movement (mm/frame)					
Task 1	0.26 (0.14)	0.30 (0.14)	4.32	.046	.122
Task 2	0.16 (0.09)	0.21 (0.11)	11.31	.002	.267
Task 3	0.13 (0.13)	0.17 (0.19)	4.71	.038	.132
Task 4	0.27 (0.16)	0.31 (0.17)	3.01	.092	.089
Task 5	0.48 (0.22)	0.58 (0.21)	15.41	<.001	.332
Total	0.23 (0.10)	0.29 (0.12)	29.16	<.001	.485

In task 2, we found a significant effect concerning the distance ratio and the absolute misjudgment between VR and VST AR. The distance ratio is significantly higher in VST AR than in VR. In VR, the participants positioned the marker wrong by 12.9% on average. In VST AR, the misplacement reaches 16.5%. Participants generally placed the markers farther back than the actual center between the participant and the object (see Fig. 5b).

Task 3 revealed significant differences in distance ratios (see Fig. 5c) and absolute misjudgments, indicating higher distance judgments in VST AR. However, the absolute misjudgments in VR and VST AR remain relatively small by 1.2% (VR) and 1.7% (VST AR).

Similarly, a significantly higher distance ratio (see Fig. 5d) and absolute misjudgment could be detected in the VST AR condition in task 4. Again, the absolute misjudgments were minimal, with 0.1% (VR) and 1.0% (VST AR).

In task 5, a significant effect on the error rate was measured (see Fig. 5e). While in VR participants collided the sphere with obstacles by 7.2% of the whole task, they collided by 23% in VST AR.

Analyses of task completion times revealed a noticeable trend of longer durations for the VST AR condition in specific tasks and over the duration of all tasks. Regarding movement, participants in the VST AR condition consistently showed greater head movement across most tasks compared to those in VR.

5.2 Subjective Measures

5.2.1 Questionnaires

While the mean values were high in the SSM, the SPES, and the plausibility questionnaires, no significant differences were found between VR and VST AR:

- SSM: $F(1, 31) = 1.46, p = .235, \eta_p^2 = .045$;
VR: $M = 3.97, SD = 0.65$; VST AR: $M = 4.07, SD = 0.61$
- SPES: $F(1, 31) = 3.76, p = .061, \eta_p^2 = .053$;
VR: $M = 3.71, SD = 0.63$; VST AR: $M = 3.98, SD = 0.74$
- Plausibility: $F(1, 31) = 1.28, p = .266, \eta_p^2 = .040$;
VR: $M = 5.15, SD = 0.74$; VST AR: $M = 5.00, SD = 0.75$
- Mental demand: $F(1, 31) = 0.22, p = .639, \eta_p^2 = .007$;
VR: $M = 6.31, SD = 4.84$; VST AR: $M = 6.58, SD = 4.68$
- Physical demand: $F(1, 31) = 0.22, p = .643, \eta_p^2 = .007$;
VR: $M = 2.62, SD = 3.20$; VST AR: $M = 2.39, SD = 2.26$

5.2.2 Participant Preferences and Perceived Complexity

After concluding the tasks, participants were asked to share their preferences and perceptions regarding the simplicity and complexity of the conditions. Fifteen participants decided on VST AR, followed by ten who liked VR the most. The other conditions received fewer votes, with four and one vote, respectively. Two participants did not decide on one condition. When we asked participants about the simplicity of the conditions, 18 participants chose the VR condition, and six participants chose the VST AR condition, followed by four and two votes for the omitted conditions. Two participants did not decide on a condition. Seven participants stated that the VST AR condition was the most complex, and one participant voted for the VR condition. Twelve and nine participants rated the other conditions as the most complex. Three participants did not decide on one condition.

5.2.3 Control Measures

There were notable changes in the VRSQ scores from pre- to post-assessment. Overall, scores increased from an average of $M = 7.01$ ($SD = 8.36$) to $M = 11.12$ ($SD = 9.29$). This change was statistically significant ($F(1, 31) = 12.23, p = .001, \eta_p^2 = .053$). Significant increases were also seen in the two subscales. The *oculomotor* scores rose from an average of $M = 11.72$ ($SD = 12.32$) to $M = 17.45$ ($SD = 14.57, F(1, 31) = 8.54, p = .006, \eta_p^2 = .044$). The *disorientation* scores increased from an average of $M = 2.29$ ($SD = 6.01$) to $M = 4.79$ ($SD = 6.61, F(1, 31) = 5.94, p = .021, \eta_p^2 = .039$). Despite the statistical significance of these changes, the absolute values remain at acceptable levels.

6 DISCUSSION

6.1 Depth Judgments

We can accept **H1** "Distance judgments in VST AR are less accurate than those in VR." In task 1, the deviation of underestimation was of less intensity in VR. In tasks 2 - 4, elements were placed farther away in VST AR than in VR, and the absolute misjudgment indicates a higher variance in misjudgments in VST AR compared to VR. Although results from tasks 3 and 4 can be caused by wrong depth perception, we need to interpret the results with caution: In contrast to tasks 1 and 2, where the interaction only included virtual objects, referential objects were provided in tasks 3 and 4 that had to be matched. Thus, the correct placement of virtual interaction objects in the physical environment also depended on the calibration quality. Although we provided fixed positions for the physical controller mounts and virtual reference points (see Sec. 4.5), the Meta Quest Touch Pro controller inhere possible tracking inaccuracies caused by insufficient tracking camera information or problems when consolidating tracking information from the controllers and the HMD. Thus, we cannot rely on tasks 3 - 5 to make assumptions about the accuracy. However, tasks 1 and 2 alone prove a higher misjudgment in VST AR compared to VR.

In task 1, we detected an underestimation for both conditions. This is in line with previous literature [1, 3, 19, 36]. If we compare the results

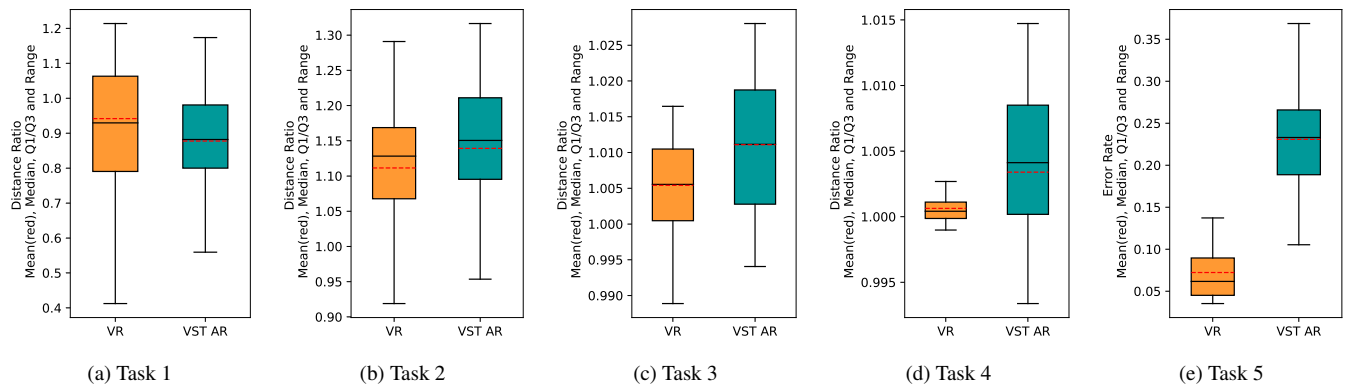


Fig. 5: Plots showing the distance ratios from tasks 1 - 4 and the error rate of task 5. All boxplots show a significant effect.

from task 1 to Kelly et al.'s work [21], which examined the depth judgment in the Meta Quest and Meta Quest 2 by verbal reporting, the Meta Quest Pro performs more accurately than its precedents. We measured a distance ratio of 94 % in our VR condition compared to 82% (Meta Quest) and 75% (Meta Quest 2) that were measured by Kelly et al. [21]. Compared to their real-world condition, which resulted in a 94 % underestimation, our VR results keep up with this measure.

In contrast to this underestimation in task 1, virtual objects in tasks 2 - 4 were placed farther away from the participant's perspective. A possible reason for this deviation might be the nature of tasks. Task 1 was static and required no motor interaction. Hence, it only required perceptive and cognitive resources. In contrast, the other tasks were more active, enabling the user to perform motor actions and, thus, to continuously readjust and reevaluate the object placement. We assume that participants underestimated the distances and then overcompensated by moving the virtual objects further back. Additionally, in task 2, participants might have calculated the position from the edge of their body and not the center of their body, even though they were instructed otherwise, leading to the placement farther back.

The results are especially interesting with regard to the identified causes of distance compression in HMDs from the literature. As Willemsen et al. [47] hypothesized, there are (additional) perceptual factors influencing depth perception apart from hardware characteristics. In our results, these perceptual factors appear in different manifestations in VR and VST AR, causing a significant difference between VR and VST AR.

6.2 Task Performance

Hypothesis **H2** "The task performance is lower in VST AR than in VR." can be accepted. We identified a higher completion time over all tasks in VST AR. Participants also moved their heads significantly more in VST AR than in VR. This finding is in line with Krichenbauer et al.'s work [26]. The authors measured more head motion in VST AR. We assume participants perceived conflicting depth cues caused by visual mismatches in the VST AR condition. As a consequence, they moved their head more to exploit motion parallax. Hence, they could compensate for these conflicts and enhance their depth perception.

We did not detect significant effects in the mental or physical demand of the NASA TLX. We imagine that VR/VST AR differences were too subtle to be consciously noticed by participants.

In task 5, a significant effect revealed a higher error rate in VST AR. While these findings support our hypothesis, we again need to view them with caution as the calibration quality defined the error rate to a certain extent. In task 5, this also affected the visualization of occlusion as the occlusion model matched the virtual room model. Additionally, a mismatch in lens distortions between the VST view and the virtual overlay might have caused misplacements of the occlusion model on the actual physical model (see Fig. 4h). Thus, participants might have perceived visual mismatches between the occlusion of the sphere and the actual holes in the panels.

6.3 Subjective Measures

Hypotheses **H3** "Users report a higher spatial presence in VR than VST AR." and **H4** "Users report a higher perceived plausibility of the scenario in VR than in VST AR." cannot be accepted for our sample. Following the CaP model [28], we would have expected that the AR-inherent a priori incongruencies would negatively affect the spatial presence and plausibility. Even though we did not find a statistically significant p-value < .05, we cannot say there was no difference between VR and VST AR as we measured small effect sizes.

Surprisingly, we have to reject **H5** "Users will prefer VR over VST AR." Even though the VR condition was the easiest to conduct, most participants chose VST AR as the condition they liked the most.

Derived from the VR and AR experience (see Tab. 1), participants had less experience in AR than in VR. This could have led to a novelty effect, causing the participants to feel a higher sensation and, thus, give higher ratings for the VST AR condition.

Another reason why participants liked the VST AR condition most and also had no deduction in the questionnaires on spatial presence and perceived plausibility is the relatively small proportion of virtual content. Most of the display was covered with an undistorted view of the environment captured by a camera stream that matched the real environment perfectly. Thus, participants might have mainly focused on that. Furthermore, participants might not have taken a closer look at the composition of the environment. In an observation task, outcomes might have been more distinctive for the spatial presence and plausibility ratings.

Wienrich et al. [46] proposed a reference frame as a frame that defines the primary reference (reality/virtuality) that the experience is judged upon. It is constructed by the proportions of virtual and physical content and further weighting. Given the small proportion of virtual content, participants possibly had a reference frame of reality. Hence, everything seemed plausible and spatially intact to them, as they may have neglected the virtual content completely. We expect that a higher proportion of virtual content causes more potential for incongruencies and, thus, more deviations in spatial presence, plausibility, and personal preference.

6.4 Limitations and Future Work

Study design Our study design included a 1×4 within-subjects design. Thus, participants might have performed better and faster in the later conditions when they had more practice. Additionally, when rating one condition, participants take previous conditions as a relative anchor and adjust their new ratings accordingly. We decided to omit reporting and discussing two conditions in this paper. However, the results showed that in the VAR+ condition, tasks 1 and 2 incorporated similar results to the VR and VST AR conditions, while tasks 3 - 5 had worse results for the VAR+ condition. The VAR condition showed similar results to the VR condition. From these results, we can conclude that a mismatching lens distortion negatively impacts depth perception and that spatial coherence (disturbed by lens distortion in the VAR+

condition) is more relevant for the right depth estimation than visual coherence (disturbed by film grain in the VAR condition). Even though these results are insightful, we suggest that more refinement is needed to validate this AR simulation fully [45]. Thus, we will use these insights for future studies as a starting point to present a more concluding view on this simulation in the future.

Causal ambiguities Our results cannot pinpoint the underlying cause or distribution of causes of the discrepancy between VR and VST AR. This would require further investigations. In the future, it would be fruitful to apply a direct comparison of VR, the VST view as it was used by Messing and Durgin [31] and Pfeil et al. [35], and VST AR (including virtual content) as we used it in our study. Hence, we could further eliminate parameters not responsible for the misjudgment and determine if the effects result solely from the VST view or if the mismatch of virtual and physical content plays a significant role. To quantify the effects of single incongruencies leading to that mismatch, we consider an evaluation by simulation VST AR in VR [44].

Hardware and use case specificity Our findings are based on the use of the Meta Quest Pro. Since every HMD has specific technical realizations (in terms of lens distortion, resolution, etc.), we cannot generalize our findings for other VST AR HMDs. For example, other HMDs often do not offer stereoscopy for the VST content. Thus, results might differ drastically.

In addition, we acknowledge that there are distinct use cases better suited for VR and others for VST AR. Thus, our results shall not answer if an application should be implemented in VR *or* VST AR but rather inform about possible effects that can appear when designing SUIs along the RV continuum. Furthermore, performance, as we measured it (completion time and error rate), is not the decisive quality criterion for some use cases. Thus, other measurements, such as the usability of a system, shall also be considered in the future.

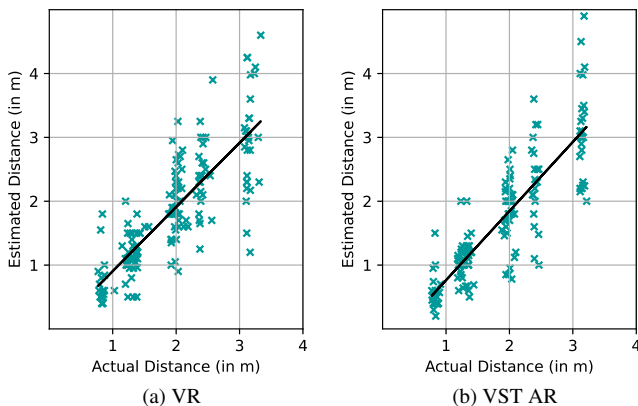


Fig. 6: Scatter plots of the absolute values in cm of the actual and estimated distances in task 1.

Depth range We looked into the absolute values of the distances in cm in task 1 to see if the estimation error increases when distances are bigger. Figure 6 shows a wider spread of estimations the higher the actual distance, while the overall trend shows a consistent incline. We measured depth judgments only in the near action space up to 3.5m. It would be interesting to get further insights in other spaces $> 3.5m$, which would be relevant for contexts such as driving simulations or other navigation tasks. For example, previous work of Gagnon et al. [12] found overestimations and underestimations at specific depth ranges up to 500m. Examining the space $< 0.5m$ would also be interesting since this is the action space for precision tasks.

Interaction For these smaller, precise tasks requiring accurate finger interaction in the near-eye field, hand-tracking interaction could be applied. This would also enable more intuitive interactions. For now, we rely on controller interaction because tracking is more accurate than

the recognition of finger gestures. We additionally wanted to prevent participants from deploying their bodies too much since it would have triggered confounding effects in VST AR in comparison to VR [38].

Participant posture Contrary to previous studies [6] on bisection tasks, we found the positioning of the marker farther away in task 2. These results might have been confounded by the body offset when participants were in a seated position and saw their knees. This makes it harder for participants to judge distances from the center of their bodies. In addition, participants have different lengths of thighs, leading to interpersonal differences in the forward offset, especially in the VST AR condition, where they see their own thighs. In VR, we approximated the participants' virtual bodies with a 3D model of the cape. To avoid confounds, we suggest applying a standing position in future experiments.

Task relevance In the future, tasks shall be tailored to relevant use cases. We plan to include more cognitive tasks to learn about the interplay between incongruencies and task load on different levels [28], i.e., perception and cognition. For now, we only concentrated on the perception part, but it would be interesting to find out how incongruencies cognitively restrict users.

Calibration accuracy For the tasks including reference objects (i.e., tasks 3 - 5), the VST AR condition indicated a higher variance in distance ratio and error rate, respectively. We attribute some of the variance to an incorrect depth judgment and some to inaccuracies in calibration. However, we cannot clearly separate these. In the future, one option would be to calculate a relative model. For example, if all objects were judged incorrectly by a consistent offset in one direction, we could assume this is the amount of calibration inaccuracy. Our tasks 3 and 4 were placed on opposite sides of the room. A calibration inaccuracy would imply that if there is a farther judgment in task 3, there has to be a shorter judgment in task 4 or vice versa. However, we measured a farther judgment in both tasks, leading to the assumption that objects were positioned farther back *not* due to calibration inaccuracies.

Outlier We mitigated outliers in tasks 2 - 5 by winsorizing our dataset to cancel out possible mistakes participants made when interacting with the controllers. In comparison, the winsorized data showed significance for the same measurements as the original data and an additional significance for the distance ratio of task 2. In most cases, the winsorizing led to higher effects (with exceptions for the distance ratio of task 4 and the error rate of task 5, which still remained significant).

7 CONCLUSION

This work provided a comprehensive examination of depth judgments in VST AR to VR. Our findings indicate higher misjudgment in VST AR compared to VR. We further identified a lower task performance in the VST AR condition measured by needed time and head movement. Surprisingly, we measured no deductions in subjective ratings. The VST AR condition was preferred over the VR condition. We assume that the high proportion of real content in the VST AR condition caused the participants to neglect the visual mismatch between real and virtual content. We outlined certain challenges of VST AR, including possible confounds and hardware limitations (viewer's body perception, lens distortion, and potential inaccuracies in calibration). Although our findings indicate comparatively worse ratings in VST AR, we assess the level of inaccuracies to be within a reasonable range, particularly given the emerging nature and rapid advances of VST AR displays. This suggests a promising scope for refinement and advancement in future applications. Overall, our study offers insights into depth perception within both VR and VST AR environments. It lays a groundwork for further exploration, emphasizing the evolving capabilities and potential of VST AR despite the initial challenges observed.

ACKNOWLEDGMENTS

This research has been funded by the Bavarian State Ministry For Digital Affairs in the project XR Hub (project number A5-3822-2-16). The authors thank Bianka Weisz and Matthias Beck for their support in the study's execution.

REFERENCES

- [1] H. Adams, J. Stefanucci, S. Creem-Regehr, and B. Bodenheimer. Depth perception in augmented reality: The effects of display, shadow, and position. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 792–801. IEEE, Christchurch, New Zealand, 2022. doi: 10.1109/VR51125.2022.00101 2, 3, 5, 7
- [2] R. T. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, Aug. 1997. doi: 10.1162/pres.1997.6.4.355 1, 2, 3
- [3] G. Ballestin, F. Solari, and M. Chessa. Perception and action in peripersonal space: A comparison between video and optical see-through augmented reality devices. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 184–189. IEEE, Munich, Germany, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00063 2, 3, 7
- [4] M. Billinghurst, A. Clark, and G. Lee. A survey of augmented reality. *Foundations and Trends® in Human-Computer Interaction*, 8(2):73–272, Mar. 2015. doi: 10.1561/1100000049 1
- [5] M. J. Blanca, R. Alarcón, and J. Arnau. Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29:552–557, Nov. 2017. doi: 10.7334/psicothema2016.383 6
- [6] B. Bodenheimer, J. Meng, H. Wu, G. Narasimham, B. Rump, T. P. McNamara, T. H. Carr, and J. J. Rieser. Distance estimation in virtual and real environments using bisection. In *Proceedings of the 4th symposium on Applied perception in graphics and visualization*, APGV '07, pp. 35–40. ACM, Tübingen, Germany, 2007. doi: 10.1145/1272582.1272589 5, 9
- [7] L. Brübach, F. Westermeier, C. Wienrich, and M. E. Latoschik. Breaking plausibility without breaking presence - Evidence for the multi-layer nature of plausibility. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2267–2276, May 2022. doi: 10.1109/TVCG.2022.3150496 6
- [8] M. A. Cidota, R. M. Clifford, S. G. Lukosch, and M. Billinghurst. Using visual effects to facilitate depth perception for spatial tasks in virtual and augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 172–177. IEEE, Merida, Mexico, 2016. doi: 10.1109/ISMAR-Adjunct.2016.0070 2
- [9] E. Cooper. The perceptual science of augmented reality. *Annual Review of Vision Science*, 9:455–478, Mar. 2023. doi: 10.1146/annurev-vision-111022-123758 1
- [10] D. Drascic and P. Milgram. Perceptual issues in augmented reality. In *Stereoscopic Displays and Virtual Reality Systems III*, vol. 2653, pp. 123–134. SPIE, San Jose, CA, USA, 1996. doi: 10.1117/12.237425 1
- [11] F. El Jamiy and R. Marsh. Survey on depth perception in head mounted displays: Distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing*, 13(5):707–712, Mar. 2019. doi: 10.1049/iet-ipr.2018.5920 2, 5
- [12] H. C. Gagnon, L. Buck, T. N. Smith, G. Narasimham, J. Stefanucci, S. H. Creem-Regehr, and B. Bodenheimer. Far distance estimation in mixed reality. In *ACM Symposium on Applied Perception 2020, SAP '20*, pp. 1–8. ACM, Virtual Event, USA, 2020. doi: 10.1145/3385955.3407933 9
- [13] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, vol. 52, pp. 139–183. Elsevier, 1988. doi: 10.1016/S0166-4115(08)62386-9 6
- [14] T. Hartmann, W. Wirth, H. Schramm, C. Klimmt, P. Vorderer, A. Gysbers, S. Böcking, N. Ravaja, J. Laarni, T. Saari, F. Gouveia, and A. Maria Sacau. The spatial presence experience scale (SPES): A short self-report measure for diverse media settings. *Journal of Media Psychology*, 28(1):1–15, Jan. 2016. doi: 10.1027/1864-1105/a000137 6
- [15] M. R. Harwell, E. N. Rubinstein, W. S. Hayes, and C. C. Olds. Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17(4):315–339, Dec. 1992. doi: 10.3102/10769986017004315 6
- [16] D. Henry and T. Furness. Spatial perception in virtual environments: Evaluating an architectural application. In *Proceedings of IEEE Virtual Reality Annual International Symposium*, pp. 33–40. IEEE, Seattle, WA, USA, 1993. doi: 10.1109/VRRAIS.1993.380801 2, 3
- [17] I. P. Howard and B. J. Rogers. *Seeing in depth, Vol. 2: Depth perception*. University of Toronto Press, 2002. 2
- [18] G. S. Hubona and G. W. Shirah. Spatial cues in 3D visualization. In Y. Cai, ed., *Ambient Intelligence for Scientific Discovery: Foundations, Theories, and Systems*, Lecture Notes in Computer Science, pp. 104–128. Springer, 2005. doi: 10.1007/978-3-540-32263-4_6 2
- [19] J. A. Jones, J. E. Swan, G. Singh, E. Kolstad, and S. R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, APGV '08, pp. 9–14. ACM, New York, NY, USA, 2008. doi: 10.1145/1394281.1394283 2, 3, 7
- [20] J. W. Kelly. Distance perception in virtual reality: A meta-analysis of the effect of head-mounted display characteristics. *IEEE Transactions on Visualization and Computer Graphics*, 29(12):4978–4989, Dec. 2023. doi: 10.1109/TVCG.2022.3196606 2, 3
- [21] J. W. Kelly, T. A. Doty, M. Ambourn, and L. A. Cherep. Distance perception in the Oculus Quest and Oculus Quest 2. *Frontiers in Virtual Reality*, 3, Mar. 2022. doi: 10.3389/frvir.2022.850471 2, 8
- [22] F. Kern and M. E. Latoschik. Reality Stack I/O: A versatile and modular framework for simplifying and unifying XR applications and research. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 74–76. IEEE, Sydney, Australia, 2023. doi: 10.1109/ISMAR-Adjunct60411.2023.00023 4
- [23] F. Kern, F. Niebling, and M. E. Latoschik. Text input for non-stationary XR workspaces: Investigating tap and word-gesture keyboards in virtual and augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2658–2669, May 2023. doi: 10.1109/TVCG.2023.3247098 3
- [24] F. Kern, M. Popp, P. Kullmann, E. Ganal, and M. E. Latoschik. 3D printing an accessory dock for XR controllers and its exemplary use as XR stylus. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, pp. 1–3. ACM, Osaka, Japan, 2021. doi: 10.1145/3489849.3489949 5
- [25] H. K. Kim, J. Park, Y. Choi, and M. Choe. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, 69:66–73, May 2018. doi: 10.1016/j.apergo.2017.12.016 6
- [26] M. Krichenbauer, G. Yamamoto, T. Taketom, C. Sandor, and H. Kato. Augmented reality versus virtual reality for 3D object manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1038–1048, Feb. 2018. doi: 10.1109/TVCG.2017.2658570 3, 8
- [27] E. Kruijff, J. E. Swan, and S. Feiner. Perceptual issues in augmented reality revisited. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 3–12. IEEE, Seoul, Korea, 2010. doi: 10.1109/ISMAR.2010.5643530 1, 2, 3
- [28] M. E. Latoschik and C. Wienrich. Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach. *Frontiers in Virtual Reality*, 3, June 2022. 1, 3, 8, 9
- [29] M. Lombard and T. Ditton. At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3(2), Sept. 1997. doi: 10.1111/j.1083-6101.1997.tb00072.x 3
- [30] J.-L. Lugin, D. Wiebusch, M. E. Latoschik, and A. Strehler. Usability benchmarks for motion tracking systems. In *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, VRST '13, pp. 49–58. ACM, Singapore, 2013. doi: 10.1145/2503713.2503730 5
- [31] R. Messing and F. H. Durgin. Distance perception and the visual horizon in head-mounted displays. *ACM Transactions on Applied Perception*, 2(3):234–250, July 2005. doi: 10.1145/1077399.1077403 2, 3, 9
- [32] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems*, E77-D(12):1321–1329, Dec. 1994. 2
- [33] J.-M. Normand, M. Servières, and G. Moreau. A new typology of augmented reality applications. In *Proceedings of the 3rd Augmented Human International Conference*, AH '12, pp. 1–8. ACM, Megève, France, 2012. doi: 10.1145/2160125.2160143 2
- [34] S. Oberdörfer, D. Heidrich, S. Birnstiel, and M. Latoschik. Enchanted by our surrounding? Measuring the effects of immersion and design of virtual environments on decision-making. *Frontiers in Virtual Reality*, 2, Aug. 2021. doi: 10.3389/frvir.2021.679277 5
- [35] K. Pfeil, S. Masnadi, J. Belga, J.-V. T. Sera-Josef, and J. LaViola. Distance perception with a video see-through head-mounted display. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–9. ACM, Yokohama, Japan, 2021. doi: 10.1145/3411764.3445223 2, 3, 9
- [36] J. Ping, Y. Liu, and D. Weng. Comparison in depth perception between virtual reality and augmented reality systems. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1124–1125. IEEE, Osaka, Japan, 2019. doi: 10.1109/VR.2019.8798174 2, 3, 5, 7

- [37] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert. The perception of egocentric distances in virtual environments - A review. *ACM Computing Surveys*, 46(2):23:1–23:40, Dec. 2013. doi: [10.1145/2543581.2543590](https://doi.org/10.1145/2543581.2543590) 2
- [38] B. Ries, V. Interrante, M. Kaeding, and L. Anderson. The effect of self-embodiment on distance perception in immersive virtual environments. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, VRST '08, pp. 167–170. ACM, Bordeaux, France, 2008. doi: [10.1145/1450579.1450614](https://doi.org/10.1145/1450579.1450614) 5, 9
- [39] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Buehner. Is it really robust? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6:147–151, Jan. 2010. doi: [10.1027/1614-2241/a000016](https://doi.org/10.1027/1614-2241/a000016) 6
- [40] R. Skarbez, M. Smith, and M. Whitton. Revisiting Milgram and Kishino's reality-virtuality continuum. *Frontiers in Virtual Reality*, 2, Mar. 2021. doi: [10.3389/frvir.2021.647997](https://doi.org/10.3389/frvir.2021.647997) 2
- [41] R. Vallat. Pingouin: Statistics in python. *Journal of Open Source Software*, 3(31), Nov. 2018. doi: [10.21105/joss.01026](https://doi.org/10.21105/joss.01026) 4
- [42] K. Vaziri, M. Bondy, A. Bui, and V. Interrante. Egocentric distance judgments in full-cue video-see-through VR conditions are no better than distance judgments to targets in a void. In *IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 1–9. IEEE, Lisboa, Portugal, 2021. doi: [10.1109/VR50410.2021.00056](https://doi.org/10.1109/VR50410.2021.00056) 3
- [43] P. Vorderer, W. Wirth, T. Saari, F. Gouveia, F. Biocca, L. Jäncke, S. Böcking, T. Hartmann, C. Klimmt, H. Schramm, J. Laarni, N. Ravaja, L. Gouveia, N. Rebeiro, A. Sacau, T. Baumgartner, and P. Jäncke. Constructing presence: Towards a two-level model of the formation of spatial presence. *Report to the European Community, Project Presence: MEC (IST-2001-37661)*, June 2003. 6
- [44] F. Westermeier, L. Brübach, M. E. Latoschik, and C. Wienrich. Exploring plausibility and presence in mixed reality experiences. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2680–2689, May 2023. doi: [10.1109/TVCG.2023.3247046](https://doi.org/10.1109/TVCG.2023.3247046) 1, 3, 6, 9
- [45] F. Westermeier, L. Brübach, C. Wienrich, and M. E. Latoschik. A virtualized augmented reality simulation for exploring perceptual incongruencies. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, VRST '23, pp. 1–2. ACM, Christchurch, New Zealand, 2023. doi: [10.1145/3611659.3617227](https://doi.org/10.1145/3611659.3617227) 3, 4, 9
- [46] C. Wienrich, P. Komma, S. Vogt, and M. E. Latoschik. Spatial presence in mixed realities – Considerations about the concept, measures, design, and experiments. *Frontiers in Virtual Reality*, 2, Oct. 2021. doi: [10.3389/frvir.2021.694315](https://doi.org/10.3389/frvir.2021.694315) 8
- [47] P. Willemsen, M. B. Colton, S. H. Creem-Regehr, and W. B. Thompson. The effects of head-mounted display mechanics on distance judgments in virtual environments. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, APGV '04, pp. 35–38. ACM, Los Angeles, CA, USA, 2004. doi: [10.1145/1012551.1012558](https://doi.org/10.1145/1012551.1012558) 2, 3, 8