Avatar Fidelity and Partner Representation in Collaborative Extended Reality

EHTISHAM UL HAQ

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Science

Graduate Program in Electrical Engineering and Computer Science
York University
Toronto, Ontario

May 2026

Abstract

Collaborative extended reality systems depend critically on how partners are represented. This thesis examines how avatar fidelity and partner representation shape coordination in shared immersive tasks. Two empirical studies compare representation conditions across asymmetric guidance and fast-paced shared-workspace interaction. The results show that more realistic or human-like representations do not uniformly improve collaboration. Instead, different representation conditions produce distinct coordination tradeoffs across task performance, errors, interference management, workload, social presence, and subjective experience. In gesture-dependent guidance, human-like appearance without corresponding behavioral support increases coordination costs. In fast-paced shared-workspace interaction, similar overall performance can emerge through different coordination strategies, including variations in selectivity, caution, and interference management. Together, the studies show that partner representation systematically shapes coordination processes and tradeoffs. It should therefore be understood as a design factor that determines what information collaborators can use and how they coordinate action, rather than simply as a means of increasing realism.

Preface

This thesis is presented in manuscript-based format and contains two manuscript chapters on a common research theme: how partner representation and avatar fidelity shape coordination in collaborative extended reality.

Chapter 2, *Realistic Looks, Unrealistic Moves: Avatar Fidelity Misalignment Undermines Collaboration in VR*, examines how visual appearance and behavioral fidelity jointly affect coordination in an asymmetric virtual reality guidance task. Chapter 3, *Partner Representation Shapes Coordination Tradeoffs in Shared XR*, examines how different partner-representation conditions shape coordination tradeoffs in a fast-paced co-located shared-workspace task.

The manuscript chapters are linked by a common concern with how collaborators interpret one another's actions, manage shared space, and coordinate under different representational constraints. Chapter 1 introduces the broader research problem and situates both studies within collaborative XR research. Chapter 4 synthesizes the findings across the two studies and discusses their broader implications for design.

Supplementary materials accompanying both studies include task demonstration videos and additional subjective measures, which provide further context for the experimental design and findings.

Acknowledgements

I would like to sincerely thank my supervisor, Robert S. Allison, for his guidance and support throughout this research and beyond. He has played an important role in helping me grow as a researcher and think more clearly about problems. I always felt comfortable reaching out to him, and his advice and perspective made a meaningful difference at every stage of this journey. I am truly grateful to have had him as my supervisor.

I am also deeply grateful to Laurie M. Wilcox for her guidance and support. Her feedback greatly strengthened this work, and she consistently took the time to share thoughtful insights and direction whenever I needed it. I truly appreciate her encouragement and the care she brought to her mentorship.

I would also like to thank my colleagues and friends in the Virtual Reality and Perception Lab and the Wilcox Lab for making this such a positive and supportive environment to work in. I am grateful for the conversations, shared ideas, and the willingness to help, whether it was through feedback, finding participants, or simply talking things through. A special shout-out to Phoebe, Xue, Iroshini, Camille, Romina, Aysun, and CéAnn for the laughs, the conversations, and for always being there.

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through an Alliance Grant (ALLRP 570802-21) in partnership with Qualcomm Canada Inc. I was also generously supported by the Vision: Science to Applications (VISTA) program at York University.

I gratefully acknowledge the participants who took part in these studies; their time and effort made this research possible.

Finally, I would like to thank my family, especially my mother and my sisters Saqia, Faiza, and Saira. No matter where we are in the world, they have always been there for me. Their support has meant everything to me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Context

Extended reality (XR) systems are increasingly used in settings where people must work together through embodied, spatial interaction rather than through conventional desktop interfaces alone [1–4]. In these settings, collaboration depends not only on verbal communication, but also on the ability to interpret a partner's movement, establish common ground, coordinate timing, and manage action around shared objects and shared space [5–7]. As a result, partner representation is not simply a visual design feature. It shapes what information about a partner becomes available during interaction and how effectively they can coordinate action.

This thesis examines how avatar fidelity and partner representation affect coordination in collaborative XR. It focuses on a central design question: what kinds of partner representation best support collaboration under different task demands? Existing work shows that richer or more human-like representations can improve some aspects of collaboration, such as awareness, interpretability, and social experience, but they can also introduce ambiguity, monitoring demands, interference, or mismatched expectations [8–11]. The relationship between representation and collaboration is therefore not well captured by a simple assumption that more realism is always better.

The thesis addresses this broader problem through two complementary coordination settings. The first examines gesture-dependent asymmetric guidance in virtual reality, where one collaborator must guide another primarily through avatar-mediated referential cues. The second examines fast co-located shared-workspace interaction across sparse, avatar-based, and mixed-reality partner representations.

Taken together, these studies go beyond comparing representational conditions. They develop a more precise account of how partner representation shapes coordination. Specifically, they show that representational effects are task-dependent, that they operate through changes in action-relevant information and interpretive effort, and that they introduce distinct coordination tradeoffs rather

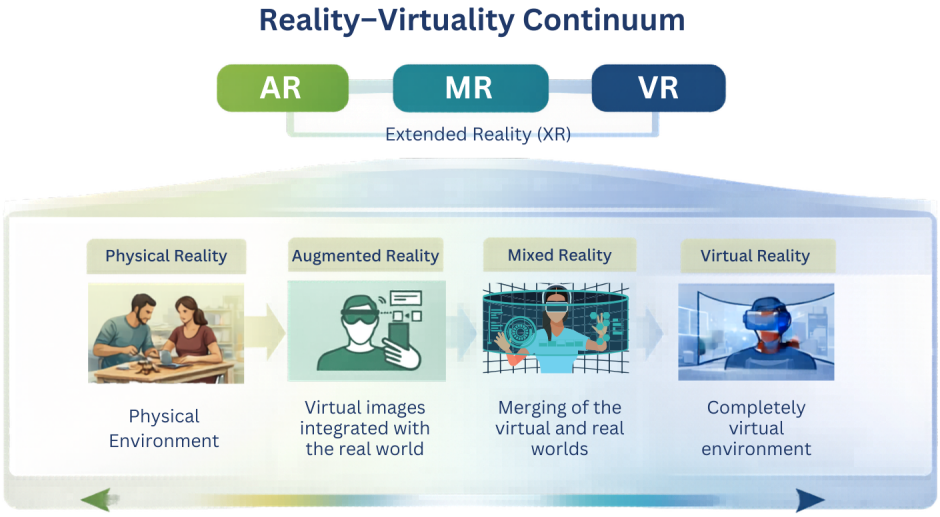


Figure 1.1: Simplified adaptation of the reality–virtuality continuum proposed by Milgram and Kishino [1], showing the range from fully physical to fully virtual environments and situating augmented and mixed reality between these endpoints.

than uniform improvements. This shifts the focus from realism as a design goal to effective representation as part of the coordination structure of the task.

1.2 Conceptual Background and Terminology

1.2.1 Extended Reality, Virtual Reality, and Mixed Reality

XR is commonly used as an umbrella term for immersive systems that include virtual reality (VR), augmented reality (AR), and mixed reality (MR). A useful way to situate these technologies is the reality–virtuality continuum proposed by Milgram and Kishino [1], which ranges from fully physical environments to fully virtual environments, with mixed forms in between. Figure 1.1 provides a simplified adaptation of this continuum using contemporary XR terminology.

In immersive VR, users primarily perceive a simulated environment through a head-mounted display. In mixed reality, digital content is combined with direct or camera-mediated views of the physical world, allowing users to remain visually connected to their surroundings while interacting with virtual objects. These modes differ not only in what users see, but also in how they orient to the environment, monitor collaborators, and interpret bodily action [3, 12, 13].

The present work focuses on collaborative XR using head-mounted displays. One study is

situated in immersive VR, while the other compares several representation conditions, including passthrough mixed reality, in a co-located shared XR workspace. This focus is timely because current devices increasingly support movement between fully virtual and passthrough-based interaction, making representational choices more varied, adaptive, and consequential for collaborative design.

1.2.2 3D Interaction and Embodied Collaboration

A central feature of immersive systems is that interaction is spatial and embodied. Users point, reach, orient, move, and manipulate objects directly in three-dimensional space. As emphasized in foundational work on 3D user interfaces, this differs from traditional desktop interaction because input is closely tied to body movement, viewpoint, and spatial relation to virtual objects [2]. These properties can make interaction more direct and natural, but they also introduce demands related to precision, alignment, physical comfort, motor control, effort, and interpretation of movement.

These issues become more complex in collaborative settings. Collaborators must not only act on objects, but also monitor one another's actions and relate their own behavior to a partner's behavior. Classic work in computer-supported cooperative work (CSCW) frames this in terms of grounding and workspace awareness: collaborators need a sufficiently shared understanding of what is being referred to, what is currently happening, and how individual actions relate to joint goals [5, 6]. Joint action research likewise emphasizes that coordinated activity depends on anticipating another person's movement, timing, and task-relevant intent [7].

In XR, these processes are strongly shaped by embodied cues. Gaze direction, hand orientation, gesture timing, spatial proximity, and movement trajectories can all function as coordination signals [14–16]. The clarity and reliability of these cues directly affect how effectively collaborators can coordinate. In asymmetric guidance tasks, one person may depend heavily on referential cues to direct another's action. In more symmetric shared-workspace tasks, collaborators must regulate turn-taking, access to contested space, and interference while acting around one another in real time.

1.2.3 Collaborative XR as Computer-Supported Collaboration

Collaborative XR can take many forms. Systems may be distributed or co-located, synchronous or asynchronous, avatar-mediated or video-mediated, fully virtual or mixed with the physical world. A large part of the design challenge lies in the fact that these systems must support social and task coordination at the same time [3, 4, 17]. Existing applications include remote assistance, procedural training, education, design review, rehabilitation, and collaborative games [18–21].

The focus here is on synchronous collaboration, where timing and movement matter moment to moment. It is also on tasks in which action is spatially grounded and coordination depends on what collaborators can perceive about one another in real time. These are settings in which partner representation becomes especially influential because it shapes not only social impression, but also how users interpret action, regulate movement, and manage shared space.

1.2.4 Partner Representation

Partner representation refers to the form through which a collaborator becomes perceptible within the task environment. In some systems this takes the form of a full-body avatar. In others it may be an abstract avatar, a partial-body proxy, a tools-only depiction, or direct real-partner visibility in mixed reality. The term *avatar* refers more specifically to a virtual bodily representation of a user.

This broader framing is useful because collaborative design decisions are not limited to choosing between low- and high-fidelity avatars. Reviews of visualization in AR and VR show that head-mounted display systems use many different representational forms, from hands-only or partial-body representations to full-body humanoid avatars and other stylized depictions [22]. These choices matter because they affect what information about a partner becomes available and how legible that information is during interaction.

1.2.5 Key Concepts

The literature on avatars and collaborative XR uses several related terms, including realism, fidelity, embodiment, presence, awareness, and co-presence. These terms overlap, but they are not interchangeable. The most important concepts used throughout this work are summarized below.

Visual realism and visual fidelity. Visual realism refers to how human-like or natural a representation appears. Visual fidelity refers more broadly to how richly or accurately the partner is depicted. This can include body completeness, surface detail, anthropomorphism, and customization. A human-like full-body avatar, for example, typically has higher visual fidelity than a floating geometric proxy.

Behavioral fidelity. Behavioral fidelity concerns how accurately and clearly a representation preserves motion-relevant information needed for collaboration, such as alignment of reach, articulation of movement, stability of pointing cues, or consistency between real and displayed action [8, 9, 23]. In this work, behavioral fidelity is treated as task-relative. In the guidance study, it mainly concerns the quality of the referential gesture channel, including wrist articulation, alignment, and pointing specificity. In the shared-workspace study, it concerns motion-relevant information needed

to regulate action around a partner, including reach alignment, tool positioning, articulation, and preservation of action cues across representation conditions.

Presence, social presence, and co-presence. Presence is often used to describe the psychological sense of "being there" in a mediated or simulated environment [24, 25]. Social presence concerns the felt salience and realness of other people in that environment, while co-presence concerns the sense of being with another person in a shared space [24, 26]. These are important aspects of collaborative XR because richer or more human-like representations often improve subjective social experience. However, they do not by themselves determine whether coordination is effective.

Referential communication and deictic gestures. Many collaborative tasks require one person to direct another's attention or action toward an object, location, or intended movement. Deictic gestures such as pointing are especially important in such cases. Prior work shows that support for deictic reference can reduce ambiguity and improve collaborative understanding in immersive systems [27–29]. This issue is central to the first study, where the task places strong functional load on avatar-mediated referential cues.

Shared space, interference, and awareness. Collaborative XR often involves spatial regulation. When collaborators work within overlapping or contested regions, they must manage access to shared objects, avoid unwanted interference, and act with awareness of a partner's location and likely movement. These issues connect to workspace awareness, peripheral awareness, interpersonal spacing, and perceived safety [6, 12, 13, 30]. They are central to the second study, where users act rapidly around one another in a shared workspace.

Expectation mismatch and the uncanny valley. One longstanding concern in research on human-like digital representation is the uncanny valley, which describes negative reactions to near-human but not fully convincing representations [31, 32]. In collaborative XR, however, the practical problem is not limited to eeriness. Realistic-looking avatars can also raise expectations about the precision, stability, or expressiveness of the behavior they will convey. When those expectations are not met, users may overinterpret ambiguous motion or invest more effort in monitoring and repair [8, 33]. In this work, expectation mismatch is therefore treated as a collaboration-relevant issue, not only as a matter of visual realism or appearance.

1.3 Review of Existing Work

1.3.1 Action-Relevant Partner Information in Collaborative XR

Prior work in collaborative XR has examined how different forms of partner information affect coordination. Embodied cues such as gesture, gaze, body motion, and awareness signals can improve collaboration quality, task understanding, and subjective experience [10, 12, 14, 15]. Work on remote assistance and mixed-reality collaboration likewise shows that successful systems must support not only communication in a general sense, but also spatial reference, action interpretation, and awareness of what a partner is doing within a task context [3, 4, 34, 35].

At the same time, the value of these cues depends on how well they match the coordination problem at hand. Some tasks depend strongly on clear deictic reference, others on perspective alignment or shared-space management, and still others on expressive communication or social comfort. This means that the same representational choice can help one dimension of collaboration while constraining another.

1.3.2 Realism Does Not Produce Uniform Benefits

Although richer or more human-like representations are often assumed to be beneficial, the literature does not support a simple more-realistic-is-better conclusion. Studies of avatar realism and behavioral realism report mixed effects across performance, comfort, agency, and social measures [8, 9, 23, 36]. Some work shows benefits from more expressive or more complete avatars [10, 37]. Other work shows that simplified representations can remain effective when their constraints are visible and interpretable [38–40]. Recent work in AR and MR further emphasizes that representation design includes more than visuals alone [11].

This mixed pattern suggests that representation should not be evaluated only through realism, global presence, or overall performance. What matters is what information a representation makes available for the task, how that information is interpreted, and how it shapes coordination behavior.

1.3.3 Gesture-Dependent Guidance

Many practical collaborative XR systems involve asymmetric guidance, such as training, assistance, instruction, or procedural support, where one person must help another act in the environment [4, 18, 19]. In these tasks, deictic gestures often carry substantial functional load because they help establish object reference, target identity, and spatial intent [27–29]. Prior work has shown that improving deictic support can improve interpretability. However, existing work has rarely isolated how avatar appearance interacts with the behavioral fidelity of the referential channel itself. As a

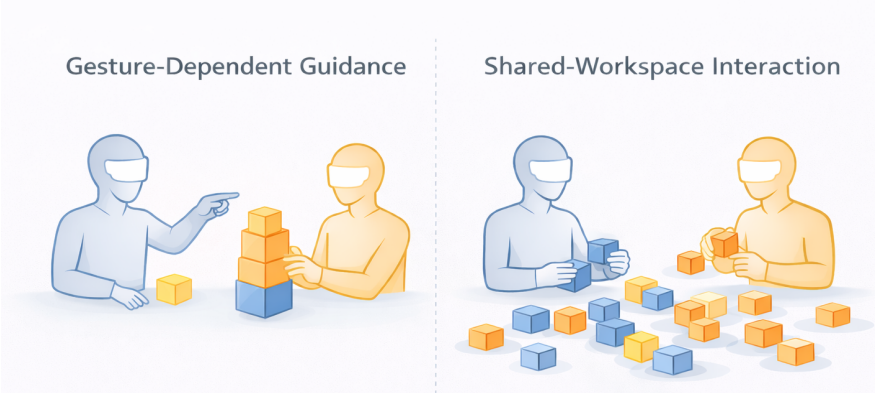


Figure 1.2: The two coordination problems examined in this thesis: gesture-dependent asymmetric guidance and shared-workspace interaction.

result, it remains unclear whether performance costs arise from reduced signal quality, expectation mismatch, or the interaction between the two.

This leaves an important open question. If a human-like avatar appears capable of precise guidance but actually provides a coarse or misaligned referential signal, the resulting coordination cost may not come only from reduced geometric precision. It may also come from expectation-driven interpretive load. Existing studies do not clearly resolve how these mechanisms combine in gesture-dependent collaborative VR, especially in asymmetric tasks where the burden of referential interpretation is central.

1.3.4 Co-Located Shared-Workspace Interaction

A related but broader issue appears in co-located shared-workspace XR. Recent work on awareness cues, volumetric telepresence, perspective alignment, expert representation, and joint action in mixed and hybrid workspaces shows that collaborative XR depends strongly on how users perceive one another's actions and organize movement around shared objects or regions [12, 13, 21, 35, 41]. However, much of this literature emphasizes usability, awareness, social experience, or aggregate task success rather than how different representations redistribute underlying coordination demands.

Distributed coordination is critical in tasks involving overlapping reach, contested targets, collisions, and selective action. In such settings, similar overall scores may arise from very different coordination strategies. One representation may improve comfort and co-presence while increasing interference or spillover. Another may reduce errors while also constraining willingness to act near a partner. Existing work does not yet provide a sufficiently clear account of these tradeoffs across sparse, avatar-based, and direct-visibility representations within the same task structure.

Figure 1.2 provides a simplified illustration of the two coordination settings examined in this work.

Taken together, these gaps point to a broader limitation in current collaborative XR research. While prior work has established that representation matters, it has provided a less precise account of how representational choices shape coordination across different task structures. In particular, realism, fidelity, presence, and performance are often treated as comprehensive ways of evaluating representation, even though they do not consistently align.

This thesis addresses that limitation by treating partner representation as a coordination design factor rather than a purely perceptual or visual one. Across two complementary studies, it examines how representational choices shape the information collaborators can use, the effort required to interpret that information, and the tradeoffs that emerge during joint action. This provides a more structured and task-sensitive account of representational effects across different coordination problems.

1.4 Research Problem

Collaborative XR research often assumes that greater realism should improve collaboration. However, existing evidence suggests that this assumption is too simple. Different representational choices can improve some aspects of coordination while weakening others, and aggregate outcomes such as completion time or overall presence do not fully explain how collaboration unfolds.

The problem addressed in this thesis is how to understand partner representation in terms of its role in coordination. Specifically, the thesis examines how different representations affect the information collaborators can extract, how clearly that information supports action, and what tradeoffs arise under different task demands. In gesture-dependent guidance, the central issue is referential clarity. In shared-workspace interaction, the central issue is how representation shapes selectivity, interference, and access to shared space. The goal is therefore not only to compare representations, but to explain how they structure coordination.

1.5 Objectives and Contributions

This thesis examines how avatar fidelity and partner representation affect coordination in collaborative XR. Its main objective is to understand how representational choices shape collaboration across different coordination demands, rather than treating realism as a universal design goal. To do this, the thesis compares two settings: gesture-dependent asymmetric guidance and fast shared-workspace interaction. These settings make it possible to examine how representation affects both objective coordination behavior and subjective collaboration experience.

The thesis contributes a coordination-oriented account of partner representation in collaborative XR. Rather than treating representation as a simple fidelity gradient, it shows how representational

effects depend on task structure, action-relevant information, and the kind of coordination problem being solved. It also advances a task-relative view of behavioral fidelity, where fidelity matters because it preserves or distorts the specific cues needed for action in a given task. Across the two studies, the thesis combines objective behavioral measures, subjective ratings, and process-sensitive indicators such as repair, errors, collision context, and partner-adjacent misses to show that similar overall outcomes can reflect different coordination mechanisms. In doing so, it provides design guidance for collaborative XR by framing partner representation as part of the coordination structure of the task, rather than as a general attempt to maximize realism.

1.6 Organization of the Thesis

Chapter 2 presents the first manuscript, which examines how alignment between avatar appearance and behavioral fidelity affects gesture-dependent collaboration in an asymmetric VR guidance task. Chapter 3 presents the second manuscript, which examines how different partner-representation conditions shape coordination tradeoffs in a fast co-located shared XR workspace. Chapter 4 synthesizes the findings across both studies, discusses their broader implications for collaborative XR design, and outlines limitations and directions for future research.

Chapter 2

Realistic Looks, Unrealistic Moves: Avatar Fidelity Misalignment Undermines Collaboration in VR

Authorship Statement

This chapter is based on a co-authored manuscript by Ehtisham Ul Haq, Robert S. Allison, and Laurie M. Wilcox, submitted to Graphics Interface (GI) 2026. Ehtisham Ul Haq led the study design, implementation, data collection, data analysis, interpretation of results, and manuscript writing. Robert S. Allison and Laurie M. Wilcox contributed to the research design, interpretation of results, manuscript revision, and supervision.

Abstract

Avatar design in collaborative virtual reality (VR) is often discussed in terms of visual realism and behavioral fidelity, but these dimensions do not always improve together. This is especially relevant in tasks that rely on referential gestures such as pointing, where collaborators must interpret spatial intent through avatar-mediated cues. We examine how avatar appearance and behavioral fidelity shape collaboration in VR using an asymmetric sorting task. Dyads completed the task under three conditions: a low-fidelity abstract avatar (LF), a visually realistic avatar with similarly limited behavioral fidelity (HFV), and a visually realistic avatar with enhanced behavioral fidelity (HF). To isolate avatar-mediated referential cues, verbal communication was not permitted. Results show that performance was best in the HF condition, with faster completion times, fewer errors, and lower workload. In contrast, performance in the HFV condition was slower, less accurate, and more

demanding, despite matching HF in visual appearance. LF generally fell between HF and HFV and did not differ from HF in completion time. Exploratory coordination-repair metrics further indicated more repeated attempts and repair events in HFV. Together, these findings suggest that in gesture-dependent collaborative VR tasks, coordination depends on the clarity of the referential gesture channel, and that human-like appearance without matching behavioral support can hinder coordination relative to simpler representations.

2.1 Introduction

Collaborative virtual reality (VR) systems are increasingly used for remote assistance, training, and learning, where partners coordinate actions in shared three-dimensional workspaces [3, 4]. Effective collaboration in these settings depends on maintaining common ground about objects, actions, and spatial intentions [5, 17]. Compared to desktop collaboration, VR coordination is more tightly coupled to embodied spatial cues (e.g., hand orientation, gesture timing, and proximity) because actions are executed directly in the 3D environment. When interaction is mediated through avatars, the avatar becomes the primary channel through which collaborators infer a partner's attention, intent, capability, and readiness to act within the environment [10, 26].

A persistent design challenge is that improvements in avatar appearance often outpace improvements in behavioral fidelity. Systems may render visually convincing human-like avatars while still producing gesture behavior that is mechanically constrained or spatially misaligned due to tracking limitations, latency, and the use of inverse kinematics (IK) to reconstruct full-body motion from sparse tracking inputs. Prior work links such appearance–behavior discrepancies to interaction costs, including reduced comfort and increased cognitive effort [8, 31, 32, 36]. In contrast, simplified or abstract avatars can sometimes support effective interaction when their limitations are visually apparent, allowing users to calibrate expectations about what the representation can express [39, 40].

In this paper, we treat *behavioral fidelity* as task-relevant in two complementary ways. First, it includes the *resolution of the referential gesture channel*: the objective capacity of the avatar to convey precise deictic references through properties such as wrist articulation, alignment between the user's real hand/controller position and the rendered avatar hand, and the specificity of the visible pointing cue. Second, we consider *expectation calibration*, the interpretive load that may arise when avatar appearance shapes what collaborators expect the gesture channel to provide. Human-like appearance can invite an assumption of precise, human-like pointing and micro-corrections; when the rendered behavior cannot meet those expectations, collaborators may overinterpret ambiguous motions, increasing effort and uncertainty during coordination.

These issues are especially consequential in collaboration settings where referential gestures

carry functional load. Many practical VR applications (e.g., remote assistance and procedural guidance) are asymmetric in the sense that one participant must guide another's actions in the environment. Even when speech is available, gestures remain a core mechanism for grounding spatial references. However, much prior work examines avatar fidelity in symmetric interactions or observational settings, and often does not disentangle whether performance and workload costs arise from limited gesture resolution itself, from expectation-driven interpretive load, or from their interaction.

We investigate how alignment between avatar appearance and the behavioral fidelity of the referential gesture channel shapes collaboration in an asymmetric sorting task. We compare three conditions: (1) a low-fidelity abstract avatar (LF), (2) a visually realistic avatar with limited behavioral fidelity (HFV), and (3) a visually realistic avatar with enhanced behavioral fidelity (HF). To isolate the contribution of avatar-mediated referential cues, verbal communication was not permitted during the task. This constraint increases reliance on gestures to expose differences in referential clarity that can be masked when speech is available, while still targeting a coordination component that remains central in multimodal (speech + gesture) collaboration.

We evaluate objective performance (completion time and placement errors) and subjective experience (NASA-TLX workload and Networked Minds social presence), including role-specific effects for teachers versus students. The study addresses two research questions:

RQ1: How does alignment between avatar appearance and behavioral fidelity affect performance in a gesture-dependent collaborative VR task?

RQ2: How does this alignment affect perceived workload and social presence, and do these effects differ by role (teacher vs. student)?

By linking objective outcomes and subjective experience to both gesture precision and expectation calibration, this work shows how avatar appearance and gesture support jointly shape coordination in collaborative VR, and why a visually realistic avatar with coarse gestures can perform worse than a simpler representation in gesture-dependent collaboration.

2.2 Related Work

2.2.1 Avatar Fidelity in Collaborative VR

Avatar fidelity is commonly discussed along at least two dimensions: *visual fidelity* (appearance realism) and *behavioral fidelity* (motion accuracy, responsiveness, and expressiveness). Early work emphasized that avatars are the locus of social perception in mediated environments, with social presence emerging from interpretations of attention, responsiveness, and coordination [26]. Subsequent research showed that richer nonverbal channels (gesture, gaze, and body motion) can

increase co-presence and improve collaboration quality [9, 10].

At the same time, evidence that visual realism alone improves collaboration is mixed. Fraser et al. report that behavioral realism can dominate appearance effects in socially demanding VR interactions [36], and that realistic avatars with limited motion fidelity can reduce comfort and increase effort relative to simpler embodiments with more coherent behavior [8]. Similarly, work on animation fidelity and IK quality shows that misaligned or unnatural motion can degrade performance and agency even under high visual realism [23]. These findings suggest that visual realism does not necessarily improve interaction and may even introduce costs, motivating a closer examination of how avatar design choices shape coordination in controlled collaborative settings.

2.2.2 Appearance–Behavior Mismatch and Expectation-Driven Costs

The Uncanny Valley provides a long-standing account of negative reactions to near-human representations [31]. In interactive contexts, such responses can be amplified when users attribute intention or mind to a human-like character and then observe behavior that violates those expectations [32]. In VR collaboration, this framing is relevant because users repeatedly interpret a partner's movements as communicative acts. Empirical work suggests that raising expectations through realism can backfire when behavioral cues are insufficient: realistic avatars with limited motion fidelity can increase cognitive effort and reduce comfort [8], and user comfort may decrease when realism invites expectations the system cannot satisfy [33].

This literature motivates a second component of behavioral fidelity beyond motion accuracy alone: *expectation calibration*. When appearance suggests human-like action but the system provides coarse or ambiguous cues, collaborators may need to expend additional effort to communicate and interpret intent, even when the visible motion appears broadly similar.

2.2.3 Referential Gestures and Gesture-Dependent Collaboration

Collaborative tasks in immersive environments often require rapid grounding of spatial references. Deictic gestures such as pointing and hand orientation provide efficient referential signals for identifying objects, targets, and intended actions. Prior work in collaborative and remote guidance contexts shows that improving deictic support can improve task success and reduce ambiguity, especially in guidance scenarios where one person instructs another [27–29]. In VR collaboration specifically, fully expressive avatars can improve performance and interpersonal outcomes in asymmetric tasks, highlighting the functional role of gesture channels [10]. Realistic body motion can also influence how collaborators synchronize and manage spatial relationships, even when subjective presence changes are modest [9].

However, collaborators also adapt to representational limits. Work on reduced-fidelity avatars suggests that users can still infer interaction patterns under constrained embodiment [40], and that lower partner-avatar fidelity can trigger compensatory strategies (e.g., exaggerated gestures or heavier reliance on alternative channels) [39]. These adaptations may preserve task success but can increase workload or slow execution, which is important for understanding why some low-fidelity designs perform competitively. Taken together, these works suggest that while gesture channels support coordination, their effectiveness depends not only on geometric properties but also on how users interpret and adapt to the available cues.

2.2.4 Connecting Gesture Resolution to Subjective Experience

Beyond task performance, collaborative avatar design affects workload and social experience. NASA-TLX captures perceived workload during coordination, and Networked Minds measures facets of co-presence, attentional engagement, and mutual understanding [26, 42]. Prior avatar fidelity research often reports these outcomes, but less frequently connects them to specific properties of the gesture channel that determine referential clarity. As a result, when a high-realism avatar performs poorly, it can be unclear whether the driver is limited gesture resolution, expectation-driven interpretive load, or both.

2.2.5 Summary and Motivation

Prior work suggests that avatar appearance and behavior jointly influence collaboration, and that mismatch can impose costs. Separately, research on deictic support shows that referential clarity is critical for guidance and spatial grounding. What remains insufficiently resolved is how these two issues combine in gesture-dependent collaboration: when a visually realistic avatar provides a coarse referential channel, do performance and workload costs stem mainly from reduced referential precision, from expectation-driven interpretive load, or from both together? We address this gap by comparing an abstract low-fidelity avatar, a visually realistic but behaviorally limited avatar, and a visually realistic avatar with enhanced referential affordances in an asymmetric task where referential gestures are central. We report both objective performance and role-specific subjective outcomes to better connect mechanism and experience.

2.3 Methods

2.3.1 Participants

Forty-eight participants (24 dyads; 25F, 23M) took part in the study. Participants ranged in age from 18 to 33 years and were recruited from the local university community. All participants reported normal or corrected-to-normal vision. All dyads consisted of participants who did not know each other prior to the study to avoid confounds related to prior familiarity [10, 18]. Participants completed informed consent and a demographic questionnaire covering age, sex, prior VR and gaming experience, handedness, and regular use of corrective eyewear. The study was approved by the institutional research ethics board.

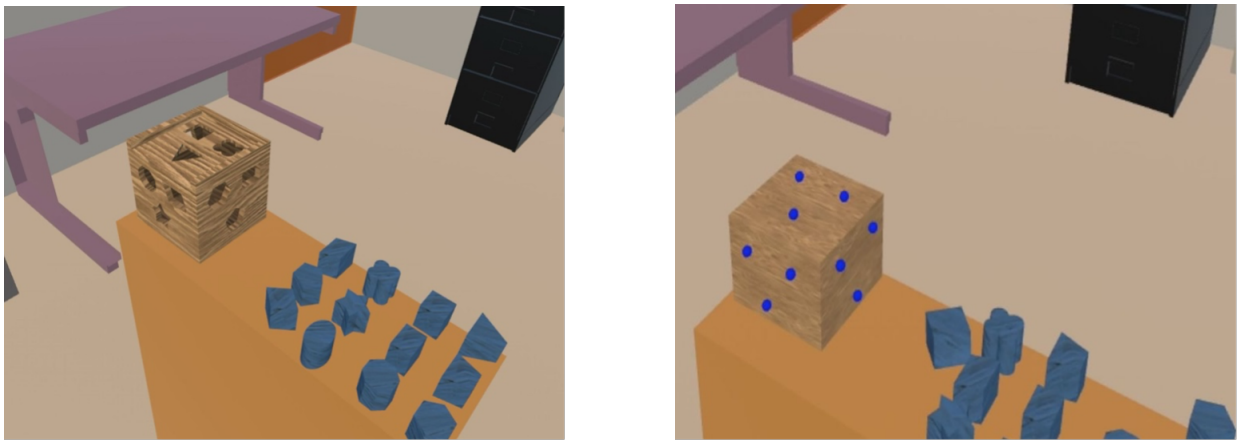
2.3.2 Apparatus and Virtual Environment

The experiment was conducted in an immersive virtual reality environment using Meta Quest 3S standalone headsets, which provided inside-out tracking for head and hand positions. The headsets operated at a default refresh rate of 72 Hz and rendered at 2064 × 2208 pixels per eye. The collaborative task environment was developed in Unity 2022.3 and networked using Normcore v2 for real-time multi-user interaction [43]. The virtual room was modeled after the physical room in which participants were co-located, supporting spatial correspondence and contextual grounding. Participants interacted in a shared virtual space containing a table, a sorting cube with shape cutouts, and a set of three-dimensional geometric shapes. The environment was designed to support asymmetric collaboration in which spatial guidance depended primarily on avatar-mediated referential cues.

Each participant was assigned a distinct role in the task. The *teacher* had full visual access to the sorting cube and its cutouts, while the *student* could manipulate the shapes but could not see the cutout locations directly. This role asymmetry was central to the collaborative dynamic and required the teacher to guide the student primarily through avatar-mediated referential cues. Figure 2.1 shows the task views for each role: the teacher's view (Fig. 2.1a) and the student's view (Fig. 2.1b).

2.3.3 Collaborative Task

The task required dyads to collaboratively place a set of 12 shapes into their corresponding cutouts on the sorting cube as quickly and accurately as possible. Only one correct placement existed for each shape, and incorrect placements caused the shape to reset to the table. Task performance was quantified using completion time and number of errors, a standard approach in collaborative VR



(a) Teacher view with full visibility of the sorting cube cutouts

(b) Student view without direct access to cutout locations

Figure 2.1: Task views for the asymmetric collaboration roles.

studies employing assembly or sorting tasks [10, 15, 20]. Such tasks have been shown to sensitively capture differences in coordination efficiency, communication clarity, and spatial understanding [39, 44].

Verbal communication between participants was not permitted during task execution. This constraint was introduced to isolate the role of embodied, nonverbal cues in collaborative performance and to prevent participants from compensating for limitations in avatar representation through speech. Prior work in VR collaboration has shown that verbal channels can mask deficiencies in gesture clarity or spatial alignment by allowing rapid correction of misunderstandings [5, 15, 18]. By restricting communication to avatar-mediated cues such as pointing, hand orientation, and timing, the task places greater emphasis on the fidelity and reliability of the avatar's communicative behavior. This design choice enables a more direct assessment of how avatar-mediated referential cues influence coordination, workload, and performance in collaborative interaction, without the confounding effects of verbal communication.

2.3.4 Avatar Fidelity Conditions

The study employed three avatar fidelity conditions that varied in visual realism and behavioral articulation. These conditions were motivated by prior work showing that avatar appearance and motion fidelity jointly influence social interaction, performance, and user experience [8, 23, 36]. Figure 2.2 shows the abstract low-fidelity avatar (Fig. 2.2a) and representative human-like avatars used in the high-fidelity conditions (Fig. 2.2b). Both the High Fidelity (HF) and High Fidelity Visual Only (HFV) conditions used the same underlying visual model but differed in behavioral fidelity.

All human-like avatars were generated using the UMA 2 avatar system [45], which supports full-body humanoid models with customizable features. To support comfort, plausibility, and embodiment, participants could choose the HF and HFV avatars' skin tone, body type, and sex. Controlled customization has been shown to support embodiment without introducing confounds in collaborative avatar studies [33, 46].

Because HF and HFV used participant-selected human-like avatars, hand and finger dimensions varied modestly across the available body-type and sex options. This variation was intentional rather than incidental: allowing participants to select a comfortable and plausible avatar helped avoid introducing additional discomfort or embodiment disruption that could arise from assigning an obviously mismatched body form. Accordingly, the pointer measurements reported below include both representative values and observed ranges across the customized avatar variants used in the study.

The underlying avatar behaviors were implemented in Unity using the Animation Rigging package [47], combining built-in two-bone inverse kinematics (IK) solvers with custom scripting to map users' real-world movements onto the avatar. The differences across the three avatar fidelity conditions were defined as follows:

- **Low Fidelity (LF):** Abstract, floating half-body avatars constructed from simple geometric primitives. Hand representations were rigid, with no wrist articulation, fingers, or grasp animation. The avatars were not dynamically scaled to adapt to the user's body, so the arm and hand positions were not adjusted based on the user's real proportions. As a result, movements could appear slightly offset, with the avatar's hands not always lining up with where the user was actually holding the controllers.

- **High Fidelity Visual Only (HFV):** Full-body human-like avatars rendered using UMA 2 models. These avatars used a basic lower-body IK configuration to simulate stepping during locomotion but retained the same upper-body behavior as the LF condition. Wrist articulation, independently controllable pointing gesture, grasping animations, and dynamic body scaling were not implemented. Thus, although the avatar appeared human-like, the upper-body referential channel remained comparatively coarse and was conveyed primarily through a broader open-hand cue.

- **High Fidelity (HF):** Full-body human-like avatars with enhanced behavioral fidelity. These avatars incorporated dynamic body scaling to better match avatar limb lengths to the user's tracked reach, reducing systematic offset between controller motion and the rendered hand position. The upper-body IK setup was enhanced to provide smoother arm articulation as well as accurate wrist movement. In addition, custom animations supported expressive finger

(a) Low-fidelity abstract avatar

(b) Examples of human-like avatars

Figure 2.2: Avatar appearances used in the study. The High Fidelity (HF) and High Fidelity Visual Only (HFV) conditions used customizable human-like avatars.

gestures for pointing and grasping. The lower-body IK configuration remained the same as in HFV.

2.3.5 Referential Gesture Affordances

For the present task, a key aspect of behavioral fidelity was the quality of the avatar's *referential gesture channel*: the extent to which an avatar could support clear deictic guidance toward a specific shape or target location. In this task, that channel depended on at least three properties: (1) **wrist articulation**, which supports directional adjustment and micro-correction; (2) **dynamic scaling and alignment**, which reduces offset between the user's real hand/controller position and the rendered avatar hand; and (3) **pointer footprint**, defined here as the effective spatial width of the visible pointing cue.

Table 2.1 summarizes these properties across conditions. LF and HFV both provided relatively limited upper-body referential control, but differed in how those limitations were visually conveyed. LF presented a simple mitten-like hand cue with an explicitly coarse appearance. HFV presented a visually human-like hand, but without corresponding wrist articulation or dynamic scaling. HF provided the narrowest and most controllable visible pointer, along with improved articulation and alignment.

To make these differences explicit, we estimated the approximate visible pointer footprint for each condition from the Unity avatar prefabs using representative poses and the Unity measurement

Table 2.1: Referential gesture affordances by condition.

Condition	Scaling / Alignment	Wrist Articulation	Pointer Footprint
LF	No	No	Broad mitten-like hand
HFV	No	No	Broad open hand
HF	Yes	Yes	Narrow index finger

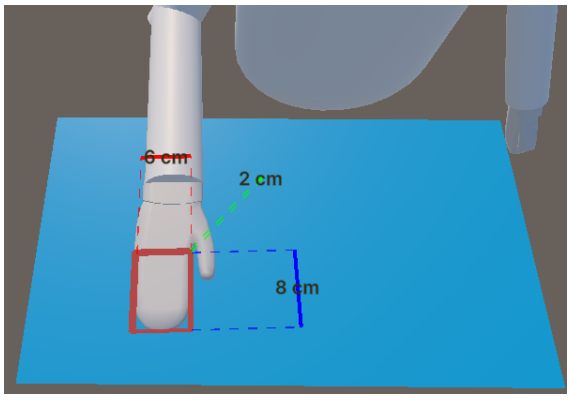
tool. Figure 2.3 illustrates these estimates. In LF, the mitten-like hand cue had an approximate width of 6 cm. In HFV, the visible open-hand footprint varied modestly across participant-selected avatar customizations, with a typical width of approximately 6.5 cm and an observed range of approximately 5.3–7.2 cm. In HF, the visible index-finger cue had an approximate width of 1.3 cm, with an observed range of approximately 1.1–1.5 cm across customizations. Across conditions, other geometric properties of the pointing cue, including approximate thickness (2 cm) and effective length (8 cm), were consistent. Although the HFV hand contained visually distinct fingers, it did not provide an independently controllable index-pointing gesture or the articulation needed to reliably use a single finger as a stable deictic cue. Its effective referential channel therefore remained broader and more behaviorally constrained than the narrow, deliberately controlled pointing cue available in HF.

2.3.6 Experimental Design

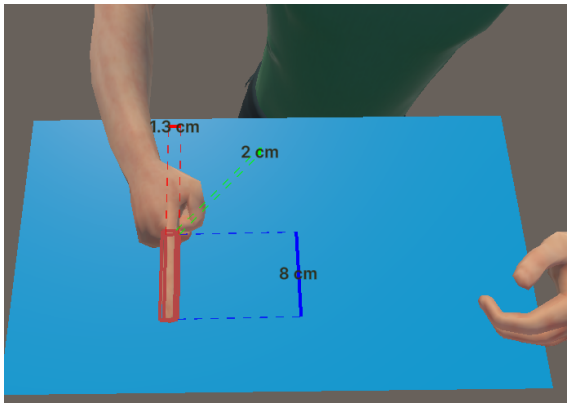
The experiment followed a within-subjects design, with each dyad completing one trial in each avatar fidelity condition. Condition order was counterbalanced using a Latin square (ABC, BCA, CAB) to mitigate learning and ordering effects [48]. Each dyad completed a brief tutorial phase using a randomly assigned avatar condition prior to the experimental trials to familiarize participants with the task and controls while minimizing transfer effects.

2.3.7 Procedure

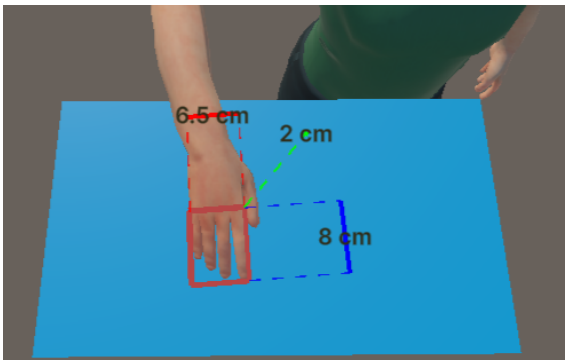
At the start of the session, participants completed consent and demographic forms. Following the tutorial, dyads completed three task trials, one per avatar fidelity condition. Immediately after each trial, participants independently completed subjective questionnaires assessing workload and social presence to capture condition-specific responses while minimizing memory biases. No time limits were imposed, and participants were instructed to prioritize both speed and accuracy.



(a) LF pointer footprint



(b) HF pointer footprint



(c) HFV pointer footprint

Figure 2.3: Approximate visible pointer footprints for each condition, used for referential-gesture analysis. HF and HFV may vary slightly due to avatar customization; the figures show a typical configuration.

2.3.8 Measures

Objective performance was measured using task completion time and the number of shape placement errors recorded during each trial. Completion time was defined as the duration from the start signal to the successful placement of the final shape. A placement error was counted whenever a shape was incorrectly inserted and reset to the table. These measures provide a quantitative index of coordination efficiency and communication accuracy between partners in each avatar condition [15, 20].

Subjective workload was assessed using the NASA Task Load Index (NASA-TLX) [42]. Participants rated six subscales after each trial: mental demand, physical demand, temporal demand, effort, frustration, and perceived performance. Each subscale was scored on a 1 to 10 scale. A combined workload score was computed by first reversing the Performance subscale and then averaging the six subscales.

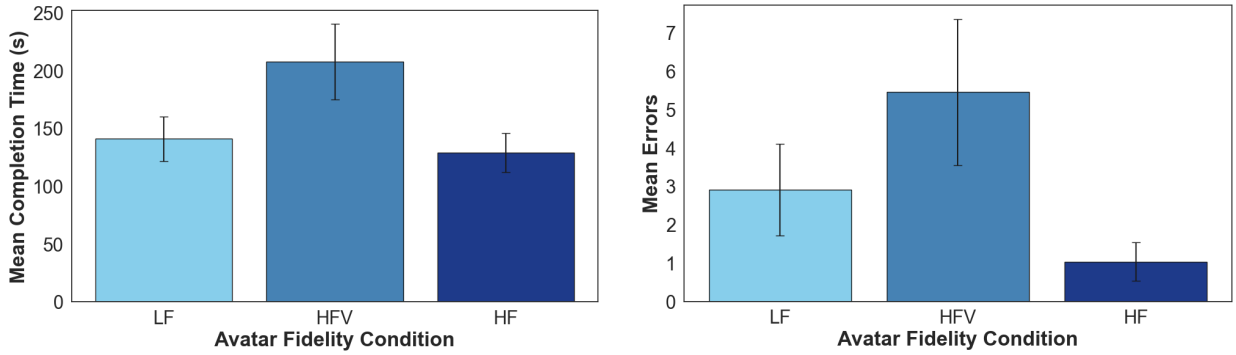
Social presence was measured using the Networked Minds Social Presence Scale, which captures co-presence, attentional engagement, and perceived understanding between collaborators [26]. This measure has been widely used in collaborative XR studies [20, 36]. Each item was scored on a 1 to 7 scale. A combined social presence score was computed by averaging all items.

Exploratory coordination-repair metrics were derived from event logs that recorded each placement attempt by dyad, condition, shape, correctness, and timestamp. To better characterize how coordination failures unfolded during the task, we computed two additional metrics. First, *attempts to success* captured the number of placement attempts required before a given shape was correctly placed. Second, *shapes needing repair* captured whether a shape required at least one corrective attempt before successful placement. For inferential analyses, shape-level measures were aggregated within each dyad-condition trial.

A video illustrating the task and experimental conditions is included as supplementary material. The NASA-TLX and social presence questionnaires are reproduced in Appendix A and Appendix B, respectively.

2.4 Results

Results are reported for objective task performance, exploratory coordination-repair measures, and subjective measures of workload and social presence. Unless otherwise noted, inferential analyses used repeated-measures ANOVA with avatar condition as a within-dyad factor. Where sphericity was violated, Greenhouse–Geisser corrected p-values are reported. Bonferroni correction was applied across the three pairwise condition comparisons (LF vs. HFV, LF vs. HF, and HFV vs. HF; see Methods for condition differences). All statistical tests used an alpha level of .05.



(a) Average task completion time across dyads for each avatar fidelity condition. Error bars represent 95% confidence intervals.

(b) Average number of placement errors across dyads for each avatar fidelity condition. Error bars represent 95% confidence intervals.

Figure 2.4: Objective task performance across avatar fidelity conditions.

2.4.1 Task Performance

2.4.1.1 Completion Time

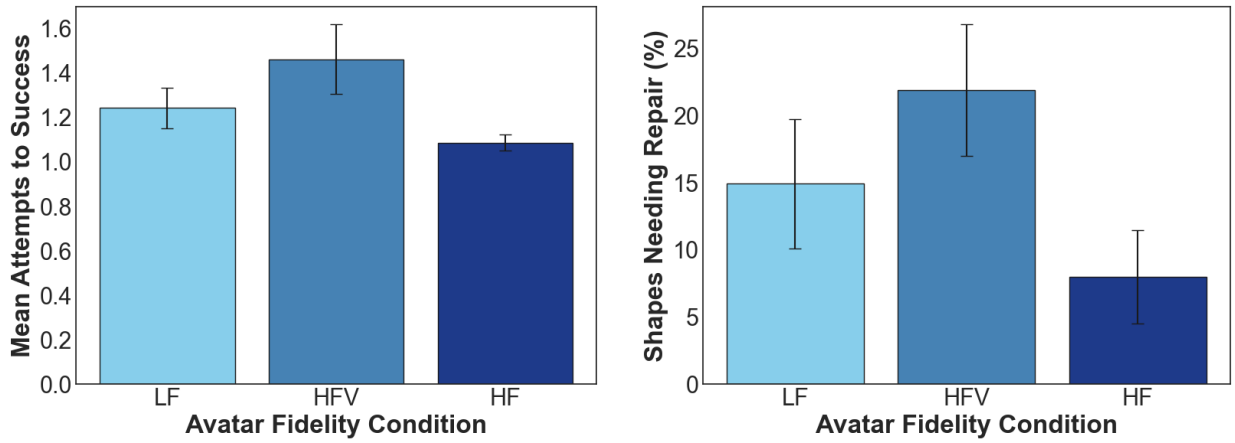
Figure 2.4a shows that participants completed the task fastest on average in the High Fidelity (HF) condition, followed closely by the Low Fidelity (LF) condition. The High Fidelity Visual Only (HFV) condition resulted in substantially longer completion times.

A repeated-measures ANOVA showed a significant main effect of avatar condition on completion time, $F(2, 46) = 19.49$, $p < .001$, $\eta_G^2 = .259$. Mauchly's test indicated that sphericity was met ($W = 0.867$, $p = .207$). Bonferroni-corrected pairwise comparisons showed that HFV took significantly longer than HF, $t(23) = 5.14$, $p < .001$, and LF, $t(23) = 4.73$, $p < .001$, whereas HF and LF did not differ significantly, $t(23) = 1.07$, $p = .882$.

2.4.1.2 Placement Errors

Figure 2.4b shows the mean number of placement errors across conditions. Participants made the fewest errors in the HF condition, more in LF, and the most in HFV.

A repeated-measures ANOVA revealed a significant main effect of avatar condition on placement errors, $F(2, 46) = 17.95$, $p < .001$, $\eta_G^2 = .236$. Mauchly's test indicated that sphericity was met ($W = 0.767$, $p = .054$). Bonferroni-corrected pairwise comparisons showed that participants made fewer errors in HF than in HFV, $t(23) = 5.22$, $p < .001$, and LF, $t(23) = 3.50$, $p = .006$, and fewer errors in LF than in HFV, $t(23) = 3.18$, $p = .012$.



(a) Attempts to success per shape

(b) Shapes needing repair (%)

Figure 2.5: Exploratory coordination-repair metrics across avatar fidelity conditions. Bars show condition means and error bars denote 95% confidence intervals.

2.4.2 Exploratory Coordination-Repair Analysis

To better understand how coordination failures unfolded during the task, we conducted an exploratory analysis of coordination-repair dynamics using the placement event logs. Figure 2.5 summarizes two complementary process metrics: the mean number of attempts required for a successful placement and the percentage of shapes that required at least one corrective attempt.

Figure 2.5a shows the mean number of attempts required to successfully place each shape. HFV required the greatest number of attempts per shape, LF was intermediate, and HF required the fewest attempts. A repeated-measures ANOVA showed a significant main effect of condition on attempts to success, $F(2, 46) = 19.09$, $p < .001$, $\eta^2_G = .247$. Bonferroni-corrected pairwise comparisons showed that HFV required more attempts than LF, $t(23) = 3.32$, $p = .009$, and HF, $t(23) = 5.41$, $p < .001$, while LF also required more attempts than HF, $t(23) = 3.50$, $p = .006$.

Figure 2.5b shows the percentage of shapes needing repair in each condition. HFV again produced the highest values, followed by LF, with HF lowest. A repeated-measures ANOVA revealed a significant main effect of condition, $F(2, 46) = 16.43$, $p < .001$, $\eta^2_G = .214$. Bonferroni-corrected pairwise comparisons showed that the percentage of repaired shapes was higher in HFV than in both LF, $t(23) = 2.85$, $p = .028$, and HF, $t(23) = 5.58$, $p < .001$. LF also showed a higher percentage of repaired shapes than HF, $t(23) = 2.97$, $p = .020$.

2.4.3 Subjective Workload (NASA-TLX)

Figure 2.6 shows combined NASA-TLX scores by condition and role. The HF condition was rated lowest in workload, followed by LF, with the highest workload reported in the HFV condition. This

pattern held for both students and teachers, with teachers consistently reporting higher workload across conditions. Combined TLX scores were computed by reversing the Performance subscale (where higher scores indicate better perceived performance) prior to averaging the subscales, so that higher values reflected greater workload.

Repeated-measures ANOVA showed a significant effect of avatar condition on global workload for students, $F(2, 46) = 14.04$, $p < .001$, $\eta^2_G = .222$, and teachers, $F(2, 46) = 24.11$, $p < .001$, $\eta^2_G = .234$. For students, Bonferroni-corrected pairwise comparisons showed that workload was higher in the HFV condition than in HF, $t(23) = 4.55$, $p < .001$, and LF, $t(23) = 3.46$, $p = .006$, while HF and LF did not differ significantly, $t(23) = 1.97$, $p = .183$. For teachers, all pairwise comparisons were significant after Bonferroni correction: workload was higher in HFV than in HF, $t(23) = 5.81$, $p < .001$, and LF, $t(23) = 4.05$, $p = .002$, while LF also exceeded HF, $t(23) = 3.75$, $p = .003$.

Figure 2.7 displays subscale score distributions across conditions, aggregated across all participants. Repeated-measures ANOVAs revealed significant omnibus effects of avatar condition across all six TLX subscales. Greenhouse–Geisser corrections were applied where sphericity was violated. Mental Demand showed a significant condition effect, $F(1.51, 34.64) = 16.06$, $p_{GG} < .001$, $\eta^2_G = .189$, with all pairwise comparisons significant after Bonferroni correction. Physical Demand also showed a significant omnibus effect, $F(1.44, 33.22) = 3.85$, $p_{GG} = .044$, $\eta^2_G = .036$, although no Bonferroni-corrected pairwise comparison remained significant. Temporal Demand showed a significant effect, $F(2, 46) = 11.04$, $p < .001$, $\eta^2_G = .104$, driven primarily by higher ratings in HFV than HF, $t(23) = 4.66$, $p < .001$. Performance ratings differed significantly across conditions, $F(1.59, 36.67) = 15.82$, $p_{GG} < .001$, $\eta^2_G = .289$, with HF rated higher than both LF and HFV, and LF also rated higher than HFV. Effort showed a significant effect, $F(2, 46) = 16.33$, $p < .001$, $\eta^2_G = .240$, with HFV rated higher than both LF and HF. Frustration likewise differed significantly by condition, $F(1.45, 33.42) = 20.10$, $p_{GG} < .001$, $\eta^2_G = .291$, with all pairwise comparisons significant after Bonferroni correction. Overall, HFV tended to produce higher workload than HF, while LF typically fell between the two conditions.

2.4.4 Social Presence

Figure 2.8 shows combined social presence scores by condition and role. Overall differences were modest, with HF receiving the highest ratings, followed by HFV and then LF. Differences were more pronounced for students than for teachers. For students, a repeated-measures ANOVA revealed a significant effect of avatar condition on global social presence, $F(1.52, 35.03) = 11.07$, $p_{GG} < .001$, $\eta^2_G = .098$. Mauchly's test indicated a violation of sphericity ($W = 0.687$, $p = .016$), so Greenhouse–Geisser corrected p-values are reported. Bonferroni-corrected pairwise comparisons

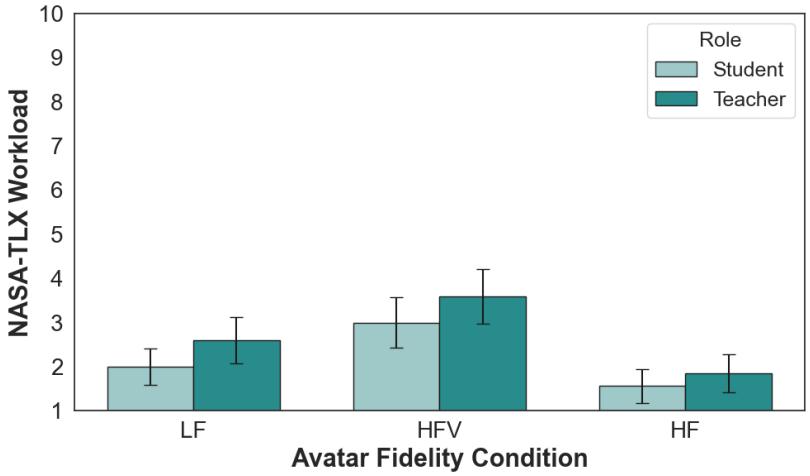


Figure 2.6: Combined NASA-TLX scores (averaged across subscales) by avatar fidelity condition and role (student vs. teacher). Error bars represent 95% confidence intervals.

showed that both HF, $t(23) = 3.87$, $p = .002$, and HFV, $t(23) = 2.82$, $p = .029$, were rated significantly higher than LF, while HF and HFV did not differ significantly, $t(23) = 2.50$, $p = .059$.

For teachers, the effect of avatar condition on global social presence was not significant, $F(1.46, 33.63) = 2.56$, $p_{GG} = .106$, $\eta^2_G = .020$. Mauchly's test again indicated a violation of sphericity ($W = 0.632$, $p = .006$), and Greenhouse–Geisser corrected p-values are reported.

Subscale-level effects, analyzed across all participants, are summarized in Table 2.2. Most significant effects were concentrated in items related to understanding, co-location, attentional engagement, and awareness. In most subscales, HF received the highest ratings and LF the lowest. A small deviation from this pattern appeared for "My partner paid close attention to me during the task," where HFV slightly exceeded HF.

2.5 Discussion

This study examined how avatar appearance and task-relevant behavioral fidelity shape collaboration in an asymmetric VR guidance task. We treated behavioral fidelity in two complementary ways: (1) the precision of the referential gesture channel, including pointing specificity, wrist articulation, and alignment between controller motion and the rendered hand; and (2) expectation calibration, the added interpretive load that can arise when a human-like appearance suggests a level of gestural precision that the avatar does not actually provide. Across objective and subjective measures, the High Fidelity (HF) condition produced the best performance and lowest workload, while the High Fidelity Visual Only (HFV) condition produced the worst performance and highest workload, despite matching HF in visual appearance.

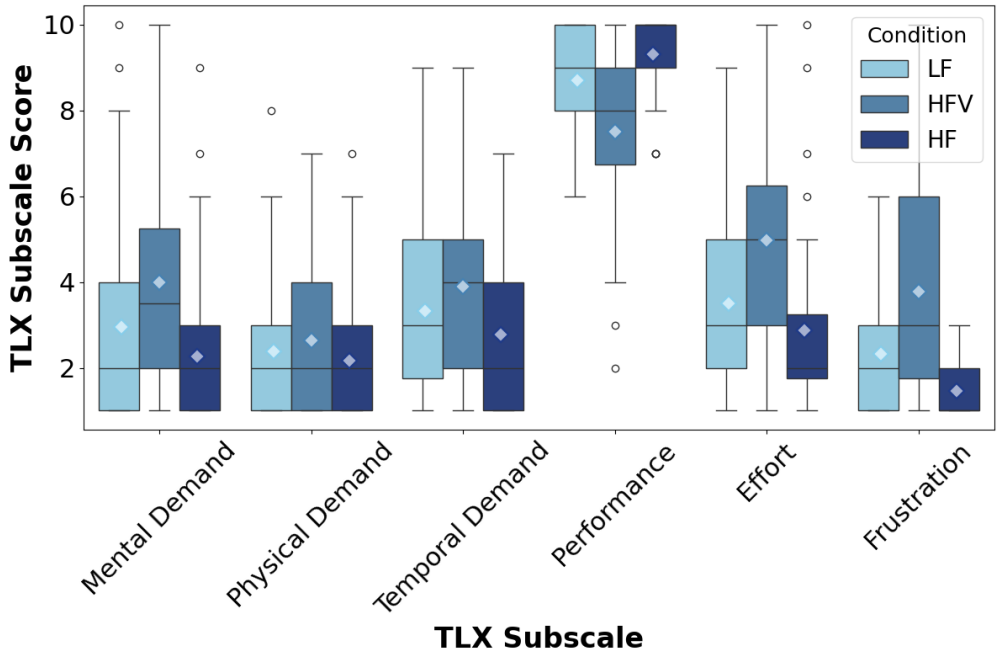


Figure 2.7: NASA-TLX subscale scores by avatar fidelity condition. Boxes indicate interquartile ranges; overlaid diamonds represent means; open circles indicate outliers.

2.5.1 Referential gesture support closely linked to coordination efficiency

The performance results indicate that differences in how avatars supported deictic reference translated directly into coordination efficiency in this task. HF led to faster completion and fewer errors than HFV, consistent with the idea that improved pointing specificity, wrist articulation, and better alignment reduce ambiguity during guidance. The exploratory coordination-repair metrics converge with this interpretation: HFV required more attempts and a higher proportion of repairs than the other conditions, suggesting that miscommunication episodes were more frequent or harder to resolve when the referential gesture channel was coarse.

Importantly, the Low Fidelity (LF) condition performed closer to HF than to HFV on completion time, and intermediate on errors and coordination-repair measures. This pattern aligns with prior work showing that reduced-fidelity embodiments can support effective interaction when the available cues are interpretable and users adapt their strategies accordingly [39, 40]. In the present task, LF offered a visibly coarse but legible hand cue, whereas HFV presented a human-like hand without the behavioral support required for precise and stable deictic reference. Together, these results suggest that performance in this task depended not simply on avatar appearance, but on whether the avatar could support clear and interpretable referential cues.

Table 2.2: Social presence subscale ratings by avatar fidelity condition.

Presence Subscale	LF	HFV	HF	η^2_G	Sig.
My partner was able to understand what I meant.	6.13	6.27	6.73	0.065	*
I often felt as if my partner and I were in the same room together.	6.10	6.46	6.60	0.051	*
My partner paid close attention to me during the task.	6.19	6.67	6.54	0.045	*
I was often aware of my partner in the shared environment.	6.06	6.42	6.54	0.034	*
My partner was easily distracted from me when other things were going on.†	2.23	2.00	1.69	0.029	*
I was able to understand what my partner meant.	6.10	6.25	6.54	0.022	
I think my partner was often aware of me in the shared environment.	6.15	6.46	6.48	0.021	*
My partner's actions were often influenced by my actions.	6.04	6.15	6.42	0.017	
I was easily distracted from my partner when other things were going on.†	2.40	2.17	1.92	0.016	
I paid close attention to my partner during the task.	6.13	6.42	6.23	0.009	
My actions were often influenced by my partner's actions.	5.54	5.60	5.83	0.005	

Note. η^2_G = generalized eta squared effect size. [*] indicates a significant main effect of avatar condition for that item ($\alpha = .05$; Greenhouse–Geisser corrected where sphericity was violated). † Reverse-scored item.

2.5.2 Why was performance lower in the HFV condition?

A key question is whether performance in the HFV condition was primarily driven by limitations in referential precision (e.g., pointing specificity), or by appearance–behavior mismatch. Because our design does not isolate these factors as independent causal variables, we do not attribute the effect to a single source. However, the overall pattern of results is more consistent with an account in which expectation-driven interpretive load plays a major role, alongside limitations in referential precision.

In HFV, participants relied on a comparatively coarse and less controllable referential channel, similar to LF in lacking independently controlled index pointing, wrist articulation, and dynamic alignment. However, unlike LF, HFV presented a human-like hand and body, which plausibly encouraged expectations of precise, fine-grained pointing. When these expectations are not met, gestures require additional interpretation by the student and more monitoring and repair by the teacher, increasing coordination effort even when progress is maintained. This interpretation is supported by the subjective results: workload was highest in HFV for both roles, with elevated Mental Demand, Effort, and Frustration, consistent with increased interpretive and repair demands.

The LF–HF and LF–HFV comparisons further help separate the role of geometric pointing precision from expectation-related effects. The difference in pointing specificity between LF and HF is substantial (rigid mitten-like hand versus articulated index pointing), whereas the difference between LF and HFV is comparatively smaller, even though it varies slightly with avatar customiza-

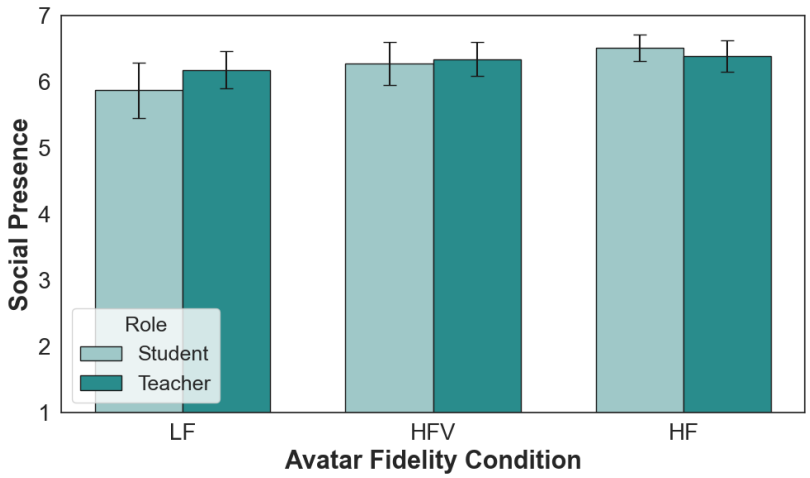


Figure 2.8: Combined social presence scores (averaged across subscales) by avatar fidelity condition and role. Error bars represent 95% confidence intervals.

tion in HFV. If geometric precision alone were the dominant factor, one might expect a larger and more consistent gap between LF and HF across measures. Instead, LF often performed closer to HF than to HFV, and in completion time did not differ significantly from HF, while HFV was substantially worse than both.

Taken together, these results indicate that while reduced referential precision contributes to coordination difficulty, it does not fully explain the observed degradation in HFV. The strongest performance breakdown occurred when a realistic appearance suggested that the avatar would support precise and controllable human-like referential gestures, while the available referential channel remained limited. In contrast, LF provided similarly coarse cues without implying high precision, making its limitations easier to interpret and adapt to. Hence, the results suggest that expectation calibration plays an important role in the HFV pattern, beyond what can be explained by geometric precision differences alone.

2.5.3 Role-specific effects and implications for asymmetric guidance

Role analyses help clarify where costs emerged. Teachers consistently reported higher workload than students, which is expected in an asymmetric task where the teacher must plan, monitor, and repair communication while maintaining awareness of both the target configuration and the student's actions. Notably, both roles showed increased TLX workload in HFV, indicating that the costs were not limited to the person interpreting gestures. This is consistent with a coordination-repair account: when referential cues are ambiguous, both partners engage in more attempts, more corrections, and more monitoring.

Social presence effects were more modest. Students reported a significant condition effect

on global presence, with higher ratings in HF and HFV than LF, while teachers did not show a significant condition effect. This asymmetry suggests that human-like appearance may shape social impressions for the partner who depends most on observing the other's cues (students), even when those impressions do not translate into better task performance. At the same time, the lack of a teacher effect indicates that in guidance roles, task demands and cue reliability may matter more than appearance for social experience. The subscale pattern was broadly consistent with higher-fidelity conditions supporting understanding and co-location perceptions, with only minor deviations in attentional items. Overall, presence results reinforce that appearance can influence social impressions while not necessarily improving coordination efficiency.

2.5.4 Generalizability to multimodal collaboration

Our task prohibited speech to isolate avatar-mediated referential cues. This decision limits direct generalization to fully natural collaboration, where speech and gesture are integrated. However, it does not imply that the findings apply only to nonverbal tasks. In many real VR applications, gestures remain a primary mechanism for grounding spatial references even when speech is present, because pointing and hand orientation can convey target identity and location more efficiently than verbal description alone [5, 28]. The present results therefore speak to a component of multimodal collaboration: when the avatar cannot support clear pointing, partners may compensate with speech, but at a cost in time, workload, interaction fluency, and reduced availability of the speech channel for other coordination needs. Our findings suggest that realistic-looking avatars with coarse gesture support may be especially likely to trigger such compensatory behavior and the added effort that comes with it.

2.5.5 Design implications

The results suggest several practical implications for collaborative VR avatar design.

Prioritize gesture clarity over appearance. For tasks involving spatial instruction or object reference, ensure that the avatar supports a precise, controllable pointing cue and sufficient articulation for micro-adjustments.

Avoid appearance-behavior mismatch. If a system cannot provide reliable pointing behavior, abstract or intentionally simplified hands may be preferable to realistic hands that imply capabilities users cannot access. When realism is used, it should be paired with behavioral support that meets the implied expectations.

Treat alignment as part of the communication channel. Reducing systematic offset between controller motion and the rendered hand enables more precise gesture control and improves the interpretability of deictic cues, especially when targets are close together.

Account for role demands. In asymmetric training and remote assistance, designs should support the guidance role with mechanisms that reduce monitoring and repair, such as clearer pointing, stabilized reference cues, or optional visual annotations that complement gestures.

2.5.6 Limitations and future work

This study has limitations that motivate follow-up work. First, the design does not independently manipulate pointing footprint, wrist articulation, and alignment, so causal attribution to any single component remains tentative. A natural next step is a factorial design that isolates pointing specificity (e.g., index-point versus open-hand) while holding appearance constant, and separately manipulates alignment and wrist degrees of freedom. Second, the no-speech constraint, while useful for isolating gesture effects, likely amplifies differences that might be partially compensated in speech-enabled interaction. Future studies should replicate the comparison under multimodal communication to quantify compensation and its cost.

Third, while we quantified pointer footprints and described alignment differences, we did not directly measure interpretation difficulty (e.g., gaze patterns, hesitation time before action, or explicit ambiguity ratings). Adding such measures would strengthen tests of the expectation-calibration account. Finally, our task was intentionally structured around guidance in a constrained workspace. Additional tasks with different spatial densities, movement requirements, and collaboration structures would help map the boundary conditions of these effects.

2.6 Conclusion

We investigated how avatar appearance and task-relevant behavioral fidelity shape collaboration in a gesture-dependent VR guidance task. Across objective performance, coordination-repair metrics, and subjective workload, the visually realistic but behaviorally limited condition (HFV) consistently produced the highest coordination costs, while the high behavioral fidelity condition (HF) supported the fastest and most accurate performance with the lowest workload. A low-fidelity abstract avatar (LF) often performed competitively and did not significantly differ from HF in completion time, indicating that human-like appearance alone did not provide a performance advantage in this task.

Taken together, the results suggest that coordination in this setting depends on how clearly avatars support referential guidance, rather than on visual realism alone. When human-like appearance is not matched with corresponding behavioral support, it can hinder coordination relative to simpler representations whose constraints are easier to interpret.

These findings motivate avatar designs that prioritize reliable, interpretable deictic cues and align behavioral capabilities with what visual appearance implies, especially in VR applications

centered on instruction, training, and remote assistance.

Appendix A: Task Load Index (NASA-TLX)

After completing each experimental condition, participants completed the NASA Task Load Index. For each item, one response was selected on a 1–10 scale.

1. **Mental Demand:** How mentally demanding was the task?
2. **Physical Demand:** How physically demanding was the task?
3. **Temporal Demand:** How hurried or rushed was the pace of the task?
4. **Performance:** How successful were you in accomplishing what you were asked to do?
5. **Effort:** How hard did you have to work to accomplish your level of performance?
6. **Frustration:** How insecure, discouraged, irritated, stressed, or annoyed were you?

Appendix B: Social Presence Questionnaire

After completing each experimental condition, participants rated each statement using a 7-point scale from Strongly Disagree (1) to Strongly Agree (7).

1. I often felt as if my partner and I were in the same room together.
2. I was often aware of my partner in the shared environment.
3. I think my partner was often aware of me in the shared environment.
4. I paid close attention to my partner during the task.
5. My partner paid close attention to me during the task.
6. I was easily distracted from my partner when other things were going on.
7. My partner was easily distracted from me when other things were going on.
8. I was able to understand what my partner meant.
9. My partner was able to understand what I meant.
10. My actions were often influenced by my partner's actions.
11. My partner's actions were often influenced by my actions.

Chapter 3

Partner Representation Shapes Coordination Tradeoffs in Shared XR

Authorship Statement

This chapter is based on a co-authored manuscript by Ehtisham Ul Haq, Robert S. Allison, and Laurie M. Wilcox, submitted to the ACM International Conference on Multimodal Interaction (ICMI) 2026. Ehtisham Ul Haq led the study design, implementation, data collection, data analysis, interpretation of results, and manuscript writing. Robert S. Allison and Laurie M. Wilcox contributed to the research design, interpretation of results, manuscript revision, and supervision.

Abstract

Partner representation is a central design decision in collaborative XR, yet it is often evaluated primarily through realism, social presence, or overall task performance. We compare four partner-representation conditions in a controlled co-located shared-workspace task: hammers only (HO), where collaborators see only the partner's tool; a low-fidelity avatar (LF) with reduced motion cues; a high-fidelity avatar (HF) with richer motion cues; and mixed reality (MR), where collaborators see the real partner via passthrough while interacting with virtual elements. Across 24 dyads, we measured task performance, errors, collision behavior, far misses, and subjective ratings. Mean scores were higher in HO and MR than in LF and HF, but these differences reflected distinct coordination profiles rather than a simple fidelity ordering. HF was associated with fewer wrong hits, but also with higher far-miss proportions and a larger share of observed collisions during shared-target moments, indicating more selective yet spatially constrained coordination. LF was associated with higher wrong-hit rates and patterns consistent with weaker control. MR was

associated with the strongest subjective experience and high performance, but also with greater action spillover, including familiarity-related increases in errors. HO was associated with efficient coordination despite no partner-body information and the lowest subjective ratings. These findings show that partner representation redistributes coordination costs across selectivity, interference, access to shared space, and subjective experience, and should be matched to the coordination demands of the task rather than simply increased in realism.

3.1 Introduction

Collaborative virtual and mixed reality systems depend heavily on how partners are represented. In co-located shared workspaces, collaborators must monitor one another's actions, anticipate movement, coordinate timing, and regulate access to overlapping space while completing the task. These demands are closely tied to grounding, workspace awareness, and joint action, where collaborators must understand what the other person is doing, where that action is directed, and how their own movement should relate to it [5–7]. Partner representation therefore shapes not only social experience, but also how coordination unfolds in practice.

In practice, collaborative XR systems are rarely designed for a single outcome. Designers must balance performance with comfort and coordination quality, while also limiting errors and interference. Prior work shows that these outcomes do not always move together. Awareness cues can improve performance and usability [12], while more realistic avatars can produce mixed effects across performance, comfort, and social experience [8, 9, 34]. As a result, representations that feel more natural may still introduce ownership errors or hesitation, while sparse representations may support efficient task performance despite weaker subjective experience. The central design problem is therefore not whether richer representation is better, but what kind of coordination it supports and at what cost.

This question is especially important in co-located shared workspaces, where collaborators must act around one another in real time while sharing access to the same space and objects. These demands arise in settings such as collaborative assembly, rehabilitation, mixed-reality training, and educational simulations. Prior work shows that effective coordination in such settings depends on being able to perceive and interpret a partner's ongoing action [12, 21, 41]. However, it remains unclear how different forms of partner representation change the pattern of coordination itself, especially in tasks requiring repeated real-time interaction in shared space.

To examine this issue, we use a co-located two-player whack-a-mole task that captures shared-workspace coordination in a controlled but demanding setting. The task requires collaborators to alternate between acting on their own targets and coordinating on shared ones within overlapping reach. This mixed structure makes it possible to observe how partner representation shapes target

selectivity, interference, and access to shared space.

We compare four partner-representation conditions: **hammers only** (HO), where collaborators see only the partner's hammer; two avatar conditions, **low-fidelity** (LF) and **high-fidelity** (HF); and **mixed reality** (MR), where collaborators see the real partner through passthrough while interacting with the same virtual task. The avatar comparison allows us to examine effects that may be due to visual richness and *behavioral fidelity*, by which we mean how clearly a representation conveys the motion-relevant information needed for coordination and control.

We therefore ask how these conditions affect collaborative task performance, coordination in shared space, and the subjective experience of collaboration. More broadly, this paper provides a coordination-oriented account of partner representation in shared immersive workspaces. Rather than treating fidelity as a single continuum, we show that different representations redistribute coordination demands across selectivity, interference, access to shared space, and subjective experience, and that similar performance can arise from distinct coordination strategies.

3.2 Related Work

3.2.1 Collaborative XR Depends on Action-Relevant Partner Information

Collaborative XR systems aim to support interaction that resembles working together in person. Foundational work in computer-supported cooperative work (CSCW) shows that collaboration depends not only on shared information, but also on grounding, mutual awareness, and the ability to interpret one another's actions in context [5, 17]. In immersive environments, these requirements are closely tied to how partners are represented and how their actions are made visible within a shared workspace.

Prior work in collaborative VR and MR shows that awareness cues can improve performance, usability, and subjective preference, while multimodal and nonverbal communication research highlights the importance of embodied signals for mutual understanding, turn-taking, and action interpretation [12, 15, 16]. XR collaboration and MR remote assistance systems further emphasize spatial referencing, joint attention, and shared understanding of task-relevant action [3, 4].

These concerns are especially important in shared-workspace settings, where collaborators act around one another in real time. Recent work highlights overlapping reach, viewpoint asymmetries, and coordination around common objects or regions of space [20, 21, 41]. Together, this literature shows that collaborative XR depends not just on partner presence, but on what action-relevant information a representation provides and how it supports coordination in shared space.

3.2.2 Avatar Fidelity Does Not Uniformly Improve Collaboration

A substantial body of work has examined whether more realistic or more complete avatars improve collaboration. Embodied avatars can support communication patterns closer to face-to-face interaction [14], and fuller body representation can influence spatial perception and self-related judgments [37, 49]. However, recent work shows that the effects of realism are mixed rather than uniformly positive.

Different aspects of representation affect different outcomes. Human resemblance alone does not necessarily improve realism or enjoyment, whereas animation realism and movement quality appear more directly relevant [8]. Behavioral realism has been linked to enjoyment and eye-contact-related experience [36], and realistic motion has been shown to shape social presence in collaborative VR [9]. At the same time, lower-fidelity or abstract representations can still support communication [38], and simplified avatars can remain effective for identifying interactions [40]. Earlier work also shows that partner fidelity affects spatial interaction, not only social impression [39].

This literature shifts the question from whether realistic avatars are better to which aspects of representation matter for specific outcomes. It also motivates distinguishing visual realism from motion-relevant or behaviorally informative cues. While recent AR work moves in this direction [11], much of the literature still evaluates representation through social presence, comfort, or overall task performance, rather than how it shapes coordination breakdowns in shared workspace.

3.2.3 Mixed Reality Introduces Distinct Shared-Workspace Tradeoffs

Mixed reality introduces a distinct case, as collaborators can see the real partner while interacting with virtual content. Prior MR work has examined specific representational choices, such as restoring head cues in local MR or representing remote experts [34, 35]. These studies show that representation matters, but focus on communication and user experience rather than shared-workspace coordination. Other work highlights how MR and hybrid environments depend on perspective alignment, spatial coordination, and interaction around shared objects [21, 41]. Instructional and procedural systems similarly rely on gesture sharing, parallel views, and spatial scaffolding [18–20].

Across these lines of work, representation is tied to a broader coordination problem: collaborators must perceive one another's actions, interpret intent, and regulate movement in shared space. Yet we still know little about how different representations shape coordination breakdowns in co-located shared-workspace tasks, especially when partners work within overlapping reach, alternate between individual and shared responsibilities, and manage interference in real time. We examine this in a co-located shared-workspace task that compares multiple representational choices within

a common setting, allowing us to study coordination breakdowns in practice rather than evaluating representation only through realism, social presence, or overall task performance.

3.3 Methods

3.3.1 Study Design

We conducted a within-dyad study in which each dyad completed the task under four representation conditions: HO, LF, HF, MR. Condition order was counterbalanced using a four-condition Latin square so that each representation appeared equally often in each ordinal position across dyads. Each condition consisted of a single 2-minute trial (8 minutes total per dyad, excluding practice and questionnaires). A single trial per condition was used to limit fatigue while also maintaining comparable exposure across conditions and minimizing learning effects. Participants stood in the same physical room with matched positions relative to the shared table and each other.

3.3.2 Participants

We recruited 24 dyads ($N = 48$ participants; 26F, 22M), ranging in age from 18 to 63 years. Dyad familiarity was recorded before the study as an exploratory social factor. Participants were instructed to report familiarity only when they had meaningful prior interaction rather than simple recognition. Twelve dyads reported prior familiarity, and twelve did not. The study was approved by the university research ethics board.

3.3.3 Apparatus and Implementation

The study used two Meta Quest 3S head-mounted displays in a co-located setup. Real-time synchronization was implemented using Normcore [43]. Avatar motion was implemented in Unity using the Animation Rigging package [47], and HF avatars were generated with UMA 2 [45]. The mixed-reality condition used the Meta Presence SDK for passthrough and shared spatial alignment [50].

Tracked head and controller motion drove the virtual hammers in all conditions. In avatar conditions (LF and HF), these inputs were also used to drive upper-body motion through the animation rigging system, ensuring that interaction remained consistent across conditions while partner representation varied. All conditions used the same task environment and controls. Frame rate was stable at 72 Hz, and no trials were excluded due to technical issues.

Figure 3.1: Task layout used in the experiment. Blue and green targets are player-specific, brown targets are shared, and the large brown target is a big shared target requiring sequential hits.

3.3.4 Task

Participants completed a two-player whack-a-mole task in a shared workspace. Each participant was assigned a hammer color (blue or green), which remained constant across conditions to reduce confusion and keep target ownership consistent across trials.

The task included three target types:

- **Player-specific targets:** could only be hit by the corresponding player
- **Shared targets:** could be hit by either player
- **Big shared targets:** required sequential hits from both players

Targets appeared on a table with 13 holes arranged in three rows (4–5–4), creating both player-adjacent and contested interaction regions. The middle row was equidistant from both players and more likely to elicit overlapping action. This design created repeated demands for target discrimination, shared access, and interference management within overlapping reach.

Figure 3.1 illustrates all target types. During an actual trial, at most two targets were active simultaneously. Target timing and type were randomized within controlled probability distributions, with most targets appearing for moderate durations and player-specific targets occurring most frequently.

Participants were instructed to maximize score while avoiding collisions. Collisions triggered visual and audio feedback, requiring both speed and selective coordination.

(a) HO

(b) LF

(c) HF

(d) MR

Figure 3.2: Example views of the four experimental conditions: HO, LF, HF, and MR. The HF avatar shown is one example of a customizable avatar configuration.

3.3.5 Representation Conditions

We compared four representation conditions (Figure 3.2) that varied in how partner information was conveyed.

In **HO**, participants saw only the partner's hammer, with no body representation.

In **LF**, participants saw a simplified half-body avatar with reduced motion cues. The avatar hands were rigid, so wrist rotation was not independently represented. The avatar also lacked dynamic body scaling, meaning its proportions and reach did not adapt to the user, potentially increasing spatial offset between the real and rendered hand positions.

In **HF**, participants saw a human-like full-body avatar with improved motion cues, including wrist articulation and dynamic body scaling to better align avatar reach with the user.

In **MR**, participants saw the real partner through passthrough while interacting with the same virtual task elements (task table, targets, and hammers).

All virtual task elements (table, targets, and hammers) were rendered with consistent scene settings across conditions, including lighting and contrast, to ensure uniform visibility.

3.3.6 Tutorial and Practice

Before the experimental trials, participants completed a tutorial covering scoring, target types, and collision penalties, followed by a self-paced practice phase until they felt comfortable with the mechanics. A video illustrating the task flow and the four representation conditions is provided as supplementary material.

3.3.7 Measures

3.3.7.1 Objective Measures

The primary measure was **final score**, defined as:

- +1: correct player-specific or shared hit
- +3: completed big shared target
- -1: wrong hit (partner's target)
- -0.5: hammer collision

We also analyzed:

- hit and miss patterns by target type
- wrong hits
- collision events and context
- far-miss proportion (misses near the partner's side)
- exploratory familiarity effects

3.3.7.2 Subjective Measures

After each condition, participants completed an 11-item questionnaire (1–7 scale, with higher values indicating more positive ratings) measuring co-presence, anticipation, influence, interpretability, timing support, awareness, peripheral awareness, shared-space awareness, safety, action confidence, and coordination.

Items were adapted from prior work on social presence, action legibility, joint action, workspace awareness, and interpersonal spacing in immersive environments [6, 7, 26, 30, 51, 52]. Full items are provided in Appendix A.

3.3.8 Statistical Analysis

Objective measures were analyzed at the dyad level ($N = 24$ dyads), since these measures reflected shared task outcomes. Effects of condition were tested using repeated-measures ANOVA. Effects involving familiarity (i.e., whether dyad partners had prior familiarity) were tested using mixed ANOVA with familiarity as a between-dyad factor. Subjective ratings were analyzed at the participant level ($N = 48$ participants). When sphericity was violated, Greenhouse–Geisser-corrected p-values were reported. Post-hoc comparisons were Holm-corrected. We report generalized eta squared (η^2_G) as the effect size measure.

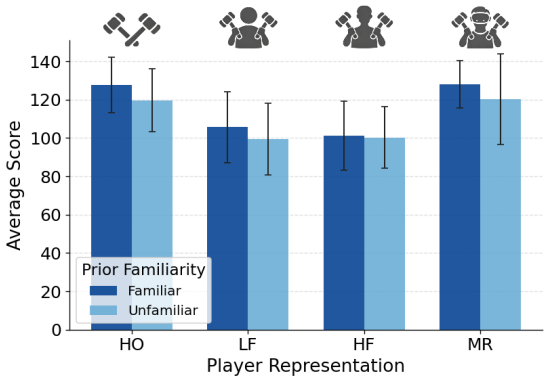
3.4 Results

3.4.1 Task Performance

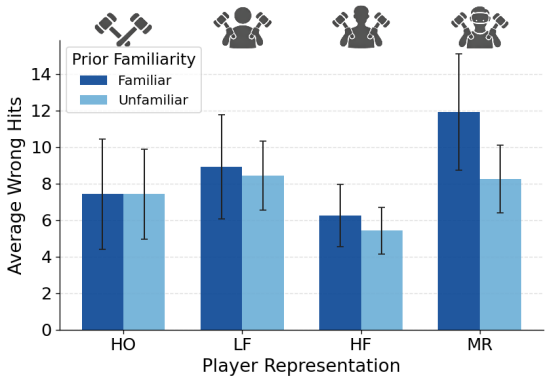
Figure 3.3a shows mean final score by condition and familiarity group (familiar vs. unfamiliar dyads). Condition had a significant effect on final score, $F(3, 69) = 16.681$, Greenhouse–Geisser-corrected $p < .001$, $\eta^2_G = 0.149$. Scores were highest in HO and MR, and lower in LF and HF. Holm-corrected pairwise comparisons indicated higher scores in HO than in LF, $t(23) = 4.25$, $p = .001$, and HF, $t(23) = 8.93$, $p < .001$, and higher scores in MR than in LF, $t(23) = 3.87$, $p = .002$, and HF, $t(23) = 5.24$, $p < .001$. No differences were observed between HO and MR, or between LF and HF. Moreover, there was no main effect of familiarity on final score, $F(1, 22) = 0.335$, $p = .569$, and no familiarity-by-condition interaction, $F(3, 66) = 0.252$, $p = .860$.

Figure 3.3b shows mean wrong hits by condition and familiarity group. Condition had a significant effect on wrong hits, $F(3, 69) = 7.101$, $p < .001$, $\eta^2_G = 0.152$. Holm-corrected comparisons showed higher wrong hits in LF than HF, $t(23) = 3.64$, $p = .007$, and higher wrong hits in MR than HF, $t(23) = 4.75$, $p < .001$. The difference between HO and MR was not statistically significant after Holm correction, and no other pairwise comparisons reached significance. Overall, wrong hits were lowest in HF and highest in MR. There was no main effect of familiarity on wrong hits, $F(1, 22) = 1.619$, $p = .216$. However, a condition-specific difference appeared in MR, where familiar dyads had more wrong hits than unfamiliar dyads, $t(22) = 2.19$, $p = .039$, $d = 0.90$.

Figure 3.4 shows hit and miss rates by target type. To account for differences in target frequency, rates were computed relative to the number of recorded opportunities in each category. The condition-by-target-type interaction was not significant for either hit rates or miss rates after Greenhouse–Geisser correction, suggesting that the broader HO/MR versus LF/HF pattern was not driven by a single target type. Target type nevertheless had significant main effects on both hit

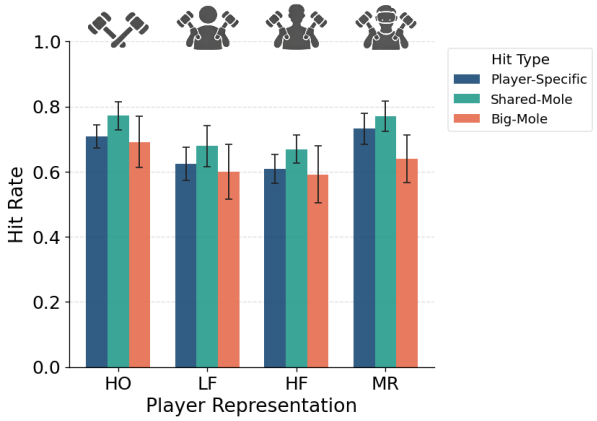


(a) Final score by condition and familiarity group.

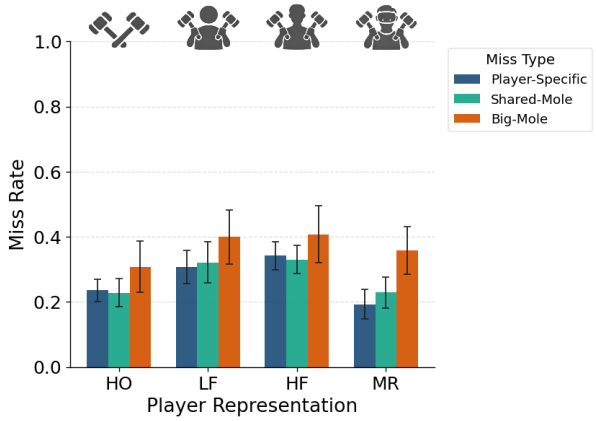


(b) Wrong hits by condition and familiarity group.

Figure 3.3: Task performance measures by condition and familiarity group. Bars show mean ± 95% confidence intervals.



(a) Hit rates by target type



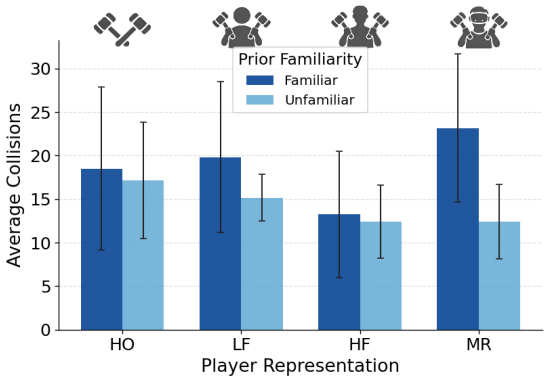
(b) Miss rates by target type

Figure 3.4: Target-type hit and miss rates across conditions, normalized by recorded opportunities. Error bars show 95% confidence intervals.

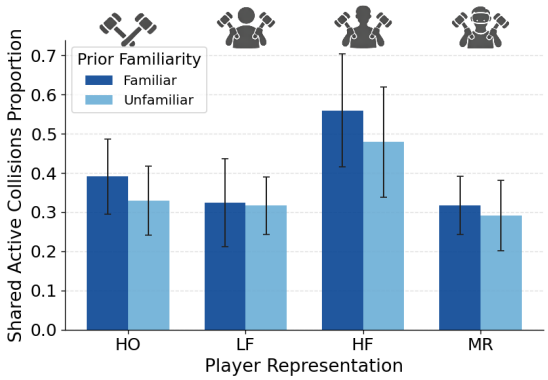
rates, $F(2, 46) = 10.19$, Greenhouse–Geisser-corrected $p = .002$, $\eta^2_G = 0.067$, and miss rates, $F(2, 46) = 14.33$, Greenhouse–Geisser-corrected $p < .001$, $\eta^2_G = 0.094$. Overall, shared targets showed the highest hit rates, whereas big shared targets showed the highest miss rates.

3.4.2 Collision Behavior

Figure 3.5a shows mean total collision events by condition and familiarity group. Total collision events differed by condition, $F(3, 69) = 3.081$, $p = .033$, $\eta^2_G = 0.037$. However, no pairwise comparisons were statistically significant after Holm correction. There was no main effect of familiarity on total collision events, $F(1, 22) = 1.375$, $p = .253$, but there was a familiarity-by-condition interaction, $F(3, 66) = 2.916$, $p = .041$, driven by MR, where familiar dyads had more



(a) Total collision events by condition and familiarity group.



(b) Proportion of observed collisions during shared-target activity by condition and familiarity group.

Figure 3.5: Collision measures by condition and familiarity group. Bars show mean ± 95% confidence intervals.

hammer collisions than unfamiliar dyads, $t(22) = 2.48$, $p = .021$, $d = 1.01$.

Figure 3.5b shows the proportion of observed collisions that occurred while at least one shared target was active, by condition and familiarity group. This proportion differed significantly by condition, $F(3, 69) = 7.918$, Greenhouse–Geisser-corrected $p = .001$, $\eta^2_G = 0.218$. Holm-corrected comparisons showed that HF had a higher shared-active collision proportion than LF, $t(23) = 3.65$, $p = .007$, and MR, $t(23) = 3.93$, $p = .004$. No other pairwise comparisons were statistically significant. There was no main effect of familiarity, $F(1, 22) = 2.371$, $p = .138$, and no familiarity-by-condition interaction, $F(3, 66) = 0.219$, $p = .883$.

Thus, although total collision counts were broadly similar across conditions, the context of observed collisions differed. In HF, a larger proportion of collisions occurred while a shared target was active, despite such moments being less frequent overall, whereas in LF and MR a larger proportion occurred during player-specific-only phases.

3.4.3 Far Misses

Far misses were defined as misses on player-specific targets appearing closest to the partner and farther from the player, normalized by total player-specific misses:

$$\text{Far-Miss Proportion} = \frac{\text{Misses\_Far\_A} + \text{Misses\_Far\_B}}{\text{Miss\_A\_Moles} + \text{Miss\_B\_Moles}}.$$

Figure 3.6 shows this proportion by condition and familiarity group. The effect of condition was significant, $F(3, 69) = 9.344$, $p < .001$, $\eta^2_G = 0.185$. Holm-corrected comparisons showed higher far-miss proportions in HF than HO, $t(23) = 3.49$, $p = .008$, and MR, $t(23) = 3.64$, $p = .007$, and

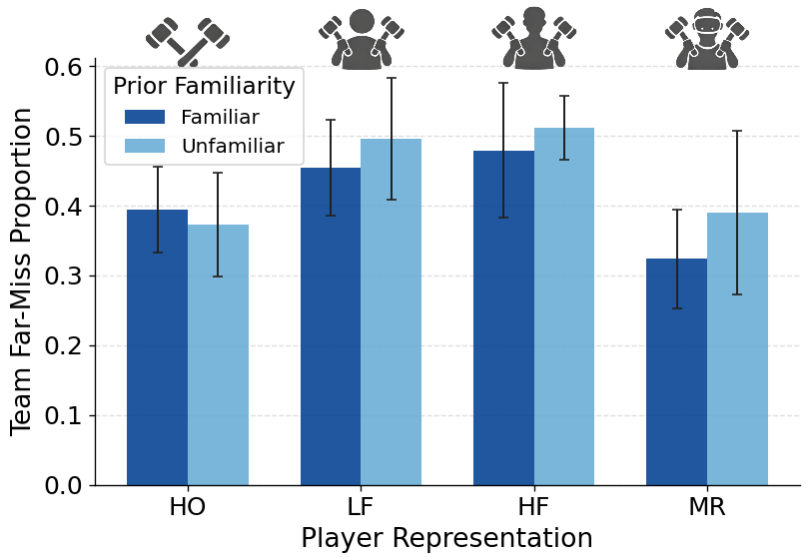


Figure 3.6: Far-miss proportion by condition and familiarity group (mean ± 95% confidence intervals).

also higher proportions in LF than HO, $t(23) = 2.98$, $p = .020$, and MR, $t(23) = 4.13$, $p = .002$. LF and HF did not differ significantly, nor did HO and MR. Thus, far misses were higher in avatar conditions (LF and HF) than in HO and MR.

There was no main effect of familiarity, $F(1, 22) = 0.73$, $p = .401$, and no familiarity-by-condition interaction, $F(3, 66) = 0.69$, $p = .559$. Pairwise comparisons between familiar and unfamiliar dyads within each condition were not statistically significant after Holm correction.

3.4.4 Subjective Ratings

Figure 3.7 shows composite subjective scores, computed per participant as the mean of the 11 questionnaire items in each condition. This aggregation serves as a summary of overall subjective experience. Composite scores differed significantly by condition, $F(3, 141) = 34.361$, Greenhouse–Geisser-corrected $p < .001$, $\eta^2_G = 0.225$. Values increased from HO to LF to HF to MR. All pairwise differences were significant after Holm correction.

Table 3.1 reports per-condition means for each questionnaire dimension. Condition had a significant effect on all dimensions except action confidence ($p = .546$), with effect sizes ranging from $\eta^2_G = 0.026$ to 0.252. Ratings increased from HO to LF to HF to MR across dimensions, consistent with the composite subjective scores.

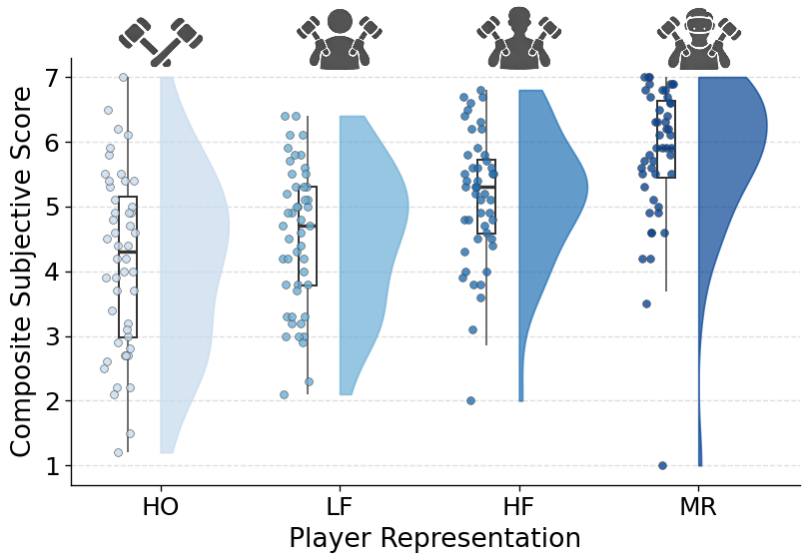


Figure 3.7: Raincloud plot of participant-level composite subjective scores by condition (higher values indicate more positive subjective experience). Each point represents a participant's mean across the 11 questionnaire items.

3.5 Discussion

The results show that partner representation changed the *pattern* of coordination, not just overall task success. Similar mean final scores concealed different coordination profiles: HO and MR were similarly high-performing, and LF and HF were similarly lower-performing, but these pairings were not behaviorally equivalent. Looking only at final score would therefore miss important differences in selectivity, interference, partner-adjacent action, and perceived collaboration. These findings extend prior work showing that richer awareness cues or more realistic representations do not produce uniform benefits across collaborative outcomes [8, 9, 12, 34] by showing *how* these mixed effects arise in a shared-workspace task.

3.5.1 Similar Mean Final Scores Masked Different Coordination Profiles

A simple fidelity account would predict that richer or more human-like representation should improve collaboration in a largely monotonic way. Our results do not support that view. Mean final scores were higher in HO and MR than in LF and HF, but those pairings were not behaviorally equivalent.

The contrast between HF and LF is especially informative. HF was associated with the lowest wrong-hit rate, whereas LF showed significantly more wrong hits. At the same time, far-miss proportions were elevated in both avatar conditions relative to HO and MR. These far misses reflect missed opportunities on targets located closer to the partner, indicating difficulty or reluctance in

Dimension	HO	LF	HF	MR	Sig.
Co-presence	4.23 ± 0.6	4.94 ± 0.5	5.69 ± 0.4	6.60 ± 0.4	*
Influence	4.29 ± 0.6	4.58 ± 0.5	5.27 ± 0.5	5.58 ± 0.6	*
Anticipation	3.62 ± 0.6	3.88 ± 0.5	4.44 ± 0.5	5.27 ± 0.5	*
Interpretability	3.35 ± 0.6	3.46 ± 0.5	4.31 ± 0.5	4.92 ± 0.5	*
Timing Support	3.54 ± 0.5	4.02 ± 0.5	4.71 ± 0.5	5.44 ± 0.5	*
Awareness	5.21 ± 0.5	5.44 ± 0.5	6.10 ± 0.3	6.54 ± 0.3	*
Peripheral Aware.	4.23 ± 0.6	4.60 ± 0.6	5.10 ± 0.5	5.62 ± 0.5	*
Shared-Space Aware.	3.96 ± 0.6	4.60 ± 0.5	5.15 ± 0.4	5.52 ± 0.6	*
Safety	4.83 ± 0.5	5.04 ± 0.5	5.27 ± 0.5	5.54 ± 0.5	*
Action Confidence	3.96 ± 0.6	3.90 ± 0.6	3.85 ± 0.6	4.17 ± 0.6	
Coordination	4.29 ± 0.5	4.33 ± 0.4	4.98 ± 0.4	5.90 ± 0.5	*

Table 3.1: Ratings (mean ± 95% CI) by condition for each questionnaire dimension on a 1–7 scale. Sig. indicates $p < .05$.

acting within overlapping reach. The share of observed collisions occurring during shared-target moments was also higher in HF than in LF and MR. Thus, HF and LF did not arrive at similarly lower scores through the same coordination pattern. The contrast between HO and MR is equally informative. Scores were similarly high in both conditions, but wrong-hit rates were highest in MR, and familiarity-related increases in wrong hits and collisions also appeared in MR. HO, by contrast, combined high scores with low far-miss proportions and fewer wrong hits than LF and MR, despite the lowest composite subjective scores. Similar overall success could therefore reflect different combinations of control, selectivity, and access to shared space. Rate-based target-type analyses further suggested that the broader HO/MR versus LF/HF pattern generalized across player-specific, shared, and big shared targets, rather than being driven by a single target type. The differences in target-type hit and miss rates likely reflected task structure, with shared targets involving no ownership constraint and big shared targets requiring sequential coordination.

These contrasts help explain why prior work has found mixed effects of richer representation across communication, social presence, and performance outcomes [8, 9, 12, 14, 34].

3.5.2 HF and LF Reached Similar Mean Final Scores for Different Reasons

HF combined high visual fidelity with high behavioral fidelity. In this condition, wrong-hit rates were lowest, far-miss proportions were high, and a larger share of observed collisions occurred during shared-target moments than in LF and MR. Taken together, this pattern suggests more selective yet more spatially constrained coordination. One plausible explanation is that the human-like avatar made the partner's body boundaries and apparent reach more salient, leading participants to regulate access to partner-adjacent space more carefully. Another is that HF increased the amount of social and spatial information that had to be monitored under time pressure, slowing or redirecting

action even when it supported cleaner selectivity. The data do not distinguish fully between these explanations, but both suggest that HF shifted coordination toward more restrained action while limiting access to some opportunities in shared space. This interpretation is broadly consistent with prior work suggesting that motion quality and richer body cues shape interaction in ways that are not captured by realism or enjoyment alone [8, 9].

LF differed from HF despite similar mean final scores. LF combined reduced visual fidelity with reduced behavioral fidelity. Wrong-hit rates were higher than in HF, a larger share of observed collisions occurred during player-specific-only phases than in HF, indicating interference even when coordination was not required, and far misses were elevated relative to HO and MR. Importantly, this pattern does not fit a simple hesitation account. If LF mainly made participants more cautious, lower wrong-hit rates than in HF would be expected, not higher. Instead, the pattern suggests continued action with weaker regulation of reach, timing, and target ownership. A likely explanation is that reducing motion-relevant cues, especially wrist articulation and dynamic body scaling, made ongoing action harder to interpret and control. A secondary possibility is that the simplified avatar also reduced the clarity of partner body configuration. Either way, the LF pattern appears to reflect weaker action control rather than cleaner selectivity or greater caution.

This distinction also helps explain why LF remained close to HF in mean final score despite looking worse on some control-related measures. One plausible interpretation is that action volume remained higher in LF than in HF, increasing both correct and incorrect actions. In other words, LF may have preserved action volume at the cost of action quality, whereas HF may have supported better selectivity at the cost of missed opportunities. Notably, these behavioral differences did not appear in action confidence ratings, suggesting that perceived confidence did not track objective control in this task.

3.5.3 HO and MR Showed Different High-Scoring Profiles

HO is informative because it isolates what happens when partner-body information is reduced without degrading behavioral fidelity for task-relevant tool motion. High mean final scores in HO were accompanied by low far-miss proportions and fewer wrong hits than in LF and MR. This suggests that reducing partner-body information did not simply remove useful cues. One possible explanation is that HO preserved the action-relevant information needed for control while reducing visual clutter and lowering the amount of partner-related information that had to be monitored under time pressure. In other words, participants in HO may have benefited not only from the absence of partner-body information itself, but also from lower representational and attentional demands. Our design does not fully separate these explanations, but both are consistent with prior work showing that reduced avatar richness can still support effective interaction when task-relevant cues remain

available [38, 40]. Our results extend that point to a co-located shared-workspace task centered on selective access to contested space.

MR combined high behavioral fidelity with direct real-partner visibility. Mean final scores were comparable to HO, and composite subjective scores were highest, but wrong-hit rates were also highest. A larger share of observed collisions occurred during player-specific-only phases than in HF, and familiar dyads in MR showed more wrong hits and more collisions than unfamiliar dyads. This pattern suggests less constrained access to shared space, yet weaker selectivity over target ownership. Direct visibility of the real partner may have supported rapid interpretation of movement and timing, helping sustain overall performance even as action spillover into partner-owned or contested space increased. At the same time, seeing the real partner and room may have made the virtual task elements feel less behaviorally constraining, making entry into adjacent or contested regions feel more natural. The familiarity results further suggest that this shift was not only perceptual, but also socially modulated, with interpersonal comfort playing a larger role here than in the other conditions. Prior MR work has shown that richer partner cues can improve social presence and user experience [12, 34, 35]. Our results do not contradict that pattern. Instead, they show that these benefits can coexist with weaker ownership control in a fast shared workspace.

3.5.4 Subjective Experience Did Not Track Objective Coordination

Subjective experience showed a clear ordering. Composite subjective scores increased from HO to LF to HF to MR, and most questionnaire dimensions followed the same pattern. Thus, richer representation generally improved how the collaboration felt. This is broadly consistent with prior work linking richer cues and more realistic representation to stronger social presence, comfort, and user experience [12, 33, 34]. However, our results also show that subjective benefit did not map directly onto objective coordination.

HF was rated more positively than LF and HO, but this did not translate into higher objective scores. MR received the strongest co-presence, awareness, coordination support, and interpretability ratings, yet wrong-hit rates were also highest. Safety showed a comparatively small effect size, and action confidence did not differ significantly across conditions even though wrong hits and far misses did. These dissociations show that subjective experience alone does not fully capture how representation affects coordination in shared workspaces. A representation can feel better while also increasing ownership errors or action spillover. Conversely, a sparse representation can feel worse while still supporting efficient coordination.

3.5.5 Implications for Collaborative XR Design

The results suggest several practical design lessons for co-located XR tasks involving overlapping reach, mixed independent and shared responsibilities, and repeated access to contested workspace.

First, when the priority is **fast and efficient task completion**, preserving task-relevant behavioral fidelity may matter more than increasing partner-body richness. Mean final scores were highest in HO and MR despite large differences in appearance.

Second, when the priority is **reducing ownership errors and improving selectivity**, high behavioral fidelity alone is not sufficient. HF was associated with the lowest wrong-hit rate, suggesting stronger selectivity, but that came with higher far-miss proportions and a coordination profile consistent with more spatially constrained action. MR also preserved rich partner cues, yet wrong-hit rates were higher. Designers should therefore not assume that the representation that feels most natural will also best support selective coordination.

Third, when the priority is **co-presence, awareness, and perceived coordination quality**, richer partner representation clearly helps. Composite subjective scores were highest in MR, and HF also improved subjective experience relative to HO and LF.

Fourth, the LF results suggest that **visual richness and behavioral fidelity should not be treated as interchangeable**. A simplified avatar is not necessarily problematic by itself. The more important issue may be whether the motion-relevant cues needed for action are preserved. In practical terms, for avatar-based shared-workspace XR, preserving reach-relevant and motion-relevant information may matter as much as, or more than, increasing human likeness.

Overall, the study reframes partner representation as a coordination problem rather than only a realism problem. The key question is not simply whether a representation looks better or feels more present, but what kind of coordination style it supports, what kinds of errors it reduces, and what new tradeoffs it introduces.

3.6 Limitations and Future Work

LF and HF differed along multiple dimensions at once, combining changes in visual richness and behavioral fidelity. This allowed comparison of realistic design configurations, but did not isolate the individual contributions of body realism, wrist articulation, and dynamic scaling. Comparisons with HO suggest that reduced behavioral fidelity plays an important role in LF's coordination patterns, but future work should separate visual and behavioral fidelity more directly and examine whether self-resembling or more personalized avatar appearance changes coordination or subjective experience within high-fidelity conditions.

The MR condition captured co-located passthrough collaboration with direct visibility of the

real partner, but it is not equivalent to remote telepresence or avatar-mediated remote collaboration. Familiarity effects, particularly in MR, should also be interpreted cautiously, since subgroup comparisons are less stable than the primary within-dyad analyses. Future work should test whether similar patterns appear in remote mixed-reality, video-mediated, or telepresence-based collaboration, and should examine more directly how social relationship shapes coordination under different representational conditions.

Finally, the scoring system reflected the priorities of this task: correct hits were rewarded, wrong hits were penalized more strongly than collisions, and big shared targets were given additional weight. This was appropriate because collisions represented interference costs rather than total task failure, whereas wrong hits reflected failures of selectivity. Future work should test whether these representational patterns hold under alternative scoring priorities, such as stronger safety penalties or stronger selectivity penalties. It should also examine task structures in which shared actions matter more than individual ones, as well as competitive or mixed-motive settings where shared-space regulation may be shaped more by strategic interference and opposition.

3.7 Conclusion

This study shows that partner representation in collaborative immersive environments shapes coordination in qualitatively different ways rather than along a simple low-to-high fidelity ladder. Taken together, the findings suggest that partner representation influences how collaborators in XR regulate movement, access shared space, balance speed against selectivity, and experience the interaction. For shared-workspace XR, the key design question is therefore not which representation is most realistic in general, but which representation best fits the coordination demands, error costs, and social context of the task.

Appendix A: Questionnaire Items

This appendix provides the full set of questionnaire items used in the study. For each item, participants provided a separate rating for each representation condition (HO, LF, HF, and MR) using a 1–7 response scale. The bolded construct labels below correspond to the dimension names reported in Table 3.1.

1. **Co-presence:** I felt that I was in the same room with my partner.
2. **Awareness:** I felt aware of my partner as an active participant during the task.
3. **Influence:** My partner's actions influenced my own actions.
4. **Anticipation:** From my partner's movements, I could anticipate which target they were about to reach.
5. **Interpretability:** My partner's movements were easy to interpret during fast moments.
6. **Timing Support:** My partner's motion helped me time my own actions.
7. **Peripheral Awareness:** I could keep track of my partner's activity without needing to look directly at them.
8. **Shared-Space Awareness:** I was aware when my partner was entering the shared reach area.
9. **Safety:** I felt safe reaching into the shared area while my partner was also acting.
10. **Action Confidence:** I felt confident reaching into the shared area without worrying about colliding with my partner.
11. **Coordination:** Overall, this representation supported effective coordination.

Chapter 4

General Discussion and Implications

4.1 Overview of Findings

This thesis examined how avatar fidelity and partner representation shape coordination in collaborative XR across two different task settings. The first study focused on gesture-dependent asymmetric guidance in virtual reality. The second focused on fast co-located shared-workspace interaction across sparse, avatar-based, and mixed-reality partner representations. Although the studies addressed different coordination problems, they converge on a broader thesis-level conclusion: partner representation is not best understood as a matter of realism alone. It is better understood as a design factor that shapes how coordination becomes possible, difficult, or costly under different task demands.

The two studies clarify a key ambiguity in existing work. Prior research has shown that richer embodiment and stronger awareness cues can improve co-presence, perceived understanding, and, in some cases, task performance [10, 12]. More recent studies also show that representational effects are rarely uniform: the same representational change can influence social presence, user experience, attention, perspective alignment, and performance in different ways [11, 34, 35, 41, 53]. Related survey work similarly points to a broader design space of representational and awareness cues in collaborative XR [54]. The present thesis extends this literature by showing how such mixed effects arise through coordination processes. Across the two studies, representational choices changed the information collaborators could use, the effort required to interpret that information, and the tradeoffs that emerged during joint action.

4.1.1 Representation shaped coordination-relevant information

Across both studies, partner representation mattered because it changed what collaborators could perceive about one another during action and how usable that information was in the task. Depend-

ing on the setting, the relevant cues included spatial reference, reach direction, articulation, timing, ownership of action, access to shared regions, and signals that made a partner's next move more or less predictable [12, 14, 15, 51].

The two studies show that these cues must be understood in relation to the task. In the guidance study, the central issue was whether the referential channel supported clear and interpretable pointing. In the shared-workspace study, the central issue was whether collaborators could regulate action around one another in a shared space. In both cases, visual richness mattered less than whether the representation preserved coordination-relevant cues in a usable way.

4.1.2 Appearance and behavioral support must be considered together

The first study showed that appearance and behavioral support cannot be treated as separate design concerns. Human-like appearance did not improve collaboration when it was paired with limited referential support. The visually realistic but behaviorally limited condition produced the slowest performance, the most placement errors, the most repair, and the highest workload, despite matching the strongest condition in appearance.

This result refines prior work on behavioral realism by showing that, in gesture-dependent collaboration, the critical issue is not motion quality in a general sense but whether the representation preserves a usable referential channel. In this study, behavioral fidelity was defined by pointing specificity, wrist articulation, and alignment. The stronger condition did not simply look better in motion. It supported the task-relevant communicative signal more effectively.

The findings also highlight the role of expectation mismatch. A coarse signal remained workable when its limitations were clearly visible, as in the abstract low-fidelity condition. The same limitation became more costly when a realistic hand and body implied a level of precision that the behavior did not actually provide. Prior work on realism, motion quality, and the uncanny valley has emphasized perceptual and affective responses [31, 32, 55]. The present thesis extends this line of work by showing how such mismatches affect coordination directly. When appearance implies precision that behavior cannot reliably support, collaborators must invest additional effort in monitoring, disambiguation, and repair, which increases coordination cost.

4.1.3 Shared-workspace coordination involved different tradeoffs across conditions

The second study shows that representational effects in shared-workspace interaction are best understood as tradeoffs rather than improvements along a single dimension. Different conditions changed how collaborators balanced selectivity, interference, access to shared space, and subjective ease of collaboration.

One key aspect was *selectivity*, or how accurately collaborators restricted their actions to their intended targets. Wrong hits, defined as strikes on the partner's player-specific targets, provided a clear measure of selectivity. The high-fidelity avatar condition produced the lowest wrong-hit rate, suggesting that richer body and motion cues supported more precise, target-specific action.

At the same time, these benefits came with costs. Avatar conditions showed higher rates of far misses, indicating more difficulty acting successfully in regions near the partner. This suggests that collaborators were more cautious or constrained when acting close to the partner's side of the workspace. In other words, improved selectivity was accompanied by reduced access or willingness to act in contested regions.

Interference patterns further clarified these differences. Collision counts alone were not sufficient to explain coordination behavior. Instead, collision context revealed how interference occurred. In the high-fidelity condition, collisions were more concentrated during shared-target moments, indicating competition over legitimately shared space. In the low-fidelity and mixed-reality conditions, collisions occurred more often during player-specific phases, suggesting weaker separation of action or greater spillover into the partner's space.

Subjective experience showed a different pattern. Mixed reality produced the strongest ratings of co-presence, awareness, and interpretability, indicating that direct real-partner visibility made collaboration feel more natural and understandable. However, this condition also produced the highest wrong-hit rates and increased spillover, especially among familiar pairs. This shows that strong subjective experience does not necessarily correspond to stronger action control.

These findings extend existing work on awareness and embodiment by showing that the same representational choice can support one dimension of coordination while constraining another. Richer cues may improve selectivity or subjective experience, but they can also reshape access to shared space, increase monitoring demands, or allow more spillover. The important point is therefore not only whether a representation improves performance, but which coordination costs it moves, reduces, or introduces.

4.1.4 Subjective experience and objective coordination diverge

A consistent pattern across both studies is that subjective experience and objective coordination outcomes do not always align. This does not make subjective measures less important. Rather, it shows that they capture a different aspect of collaboration.

Presence, co-presence, awareness, and interpretability reflect how collaboration feels and how understandable a partner appears [24–26]. However, they do not directly indicate whether coordination was efficient, selective, or well regulated. In the guidance study, the visually realistic but behaviorally limited condition could still support positive impressions while increasing errors and

workload. In the shared-workspace study, mixed reality produced the strongest subjective ratings while also increasing wrong hits and spillover. Conversely, the hammers-only condition supported efficient performance despite low subjective ratings.

These findings reinforce the need to evaluate collaborative XR systems using both behavioral and subjective measures. A representation can feel natural and socially rich while introducing control-related costs, and a sparse representation can feel limited while still supporting efficient coordination.

4.2 Design Implications

The combined findings suggest that representational choices should be matched to task demands rather than selected only for visual richness. Tasks that rely on referential guidance require clear and well-aligned gesture support. Tasks that involve rapid interaction in shared space require attention to selectivity, interference, and access to contested regions. In both cases, the relevant question is what information collaborators need in order to act, and whether the representation provides that information clearly enough under task pressure.

The findings also suggest that appearance and behavioral support should be designed together. Human-like appearance can be beneficial, but only when the available behavioral cues meet the expectations that appearance creates. When behavioral support is limited, simpler representations may be easier to interpret because their constraints are visible and easier to calibrate.

Evaluation should also include process-sensitive measures. Aggregate outcomes such as completion time or overall score can hide important coordination dynamics. Measures such as repair, wrong actions, collision context, and partner-adjacent misses provide a clearer picture of how coordination unfolds and what kind of difficulty a representation introduces.

Finally, direct real-partner visibility should not be assumed to be optimal in all shared-workspace tasks. Mixed reality improved subjective experience but also introduced increased spillover and reduced selectivity. Systems using passthrough or real-world visibility should therefore consider how these benefits interact with action control.

4.3 Limitations and Future Directions

The two studies in this thesis used controlled tasks designed to isolate particular coordination demands. This was a strength for experimental clarity, but it also limits the scope of generalization. The findings speak most directly to tasks with similar structures: gesture-dependent guidance in one case, and fast shared-workspace regulation in the other. Longer-duration, speech-rich, socially complex, or more open-ended collaborative XR settings may produce different tradeoffs.

Another limitation is that the representational manipulations were realistic bundles rather than fully decomposed cue manipulations. In the first study, appearance, pointing specificity, wrist articulation, and alignment varied together. In the second, body richness, articulation, dynamic scaling, abstraction, and direct real-partner visibility varied in bundled ways. This makes the results relevant to practical design configurations, but it also limits precise causal claims about the isolated effect of any single cue.

The first study also removed speech to isolate gesture-based coordination. While this strengthened internal validity, it limits direct comparison to natural multimodal interaction. Future work should examine how speech and gesture interact under different representational constraints.

Furthermore, the main within-subject comparisons in both studies were well suited to the research questions, but subgroup effects, especially those related to familiarity, remain less stable. The familiarity results in the shared-workspace study are theoretically interesting, particularly because they suggest that direct real-partner visibility may amplify interpersonal comfort effects, but they would benefit from replication with larger samples and more graded measures of relationship history.

These limitations point to several directions for future work. One is to isolate representational components more directly by independently manipulating pointing specificity, articulation, controller-to-avatar alignment, body completeness, dynamic scaling, and direct real-partner visibility. Another is to extend the work to a wider range of collaborative XR settings, including remote telepresence, collaborative learning, procedural training, rehabilitation, and multi-user environments. Such contexts may shift the balance among awareness, realism, trust, motivation, satisfaction, efficiency, and interference.

Future work would also benefit from more detailed process measures. Gaze behavior, hesitation time, trajectory structure, kinematic variability, interpersonal distance regulation, and explicit ambiguity judgments would make it possible to test coordination mechanisms more directly. These measures could help distinguish whether a representation is improving reference, increasing monitoring burden, or shifting collaborators toward more cautious or more permissive action styles.

4.4 Conclusion

This thesis argues for a coordination-oriented view of partner representation in collaborative XR. Rather than treating representation mainly as a question of realism, the thesis shows that representational choices help structure how collaborators understand one another, regulate action, and manage shared work. For that reason, partner representation should be designed in relation to task demands and coordination requirements, not as a simple attempt to maximize realism.

References

[1] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, E77-D(12):1321–1329, 1994.

[2] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola Jr., and Ivan Poupyrev. *3D User Interfaces: Theory and Practice*. Addison-Wesley, Boston, MA, 2004.

[3] Yongjae Lee and Byounghyun Yoo. XR collaboration beyond virtual reality: Work in the real world. *Journal of Computational Design and Engineering*, 8(2):756–772, 2021. doi: 10.1093/jcde/qwab012.

[4] David Fidalgo, Manuel Sousa, Ana Cardoso, Hugo Paredes, and Benjamim Fonseca. Remote assistance and training in mixed reality: A systematic literature review. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2291–2303, 2023. doi: 10.1109/TVCG.2023.3244011.

[5] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC, 1991.

[6] Carl Gutwin and Saul Greenberg. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)*, 11(3–4):411–446, 2002.

[7] Natalie Sebanz, Günther Knoblich, and Wolfgang Prinz. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, 2006.

[8] Andrew D. Fraser, Alena Denisova, and Marco Monteiro. Do realistic avatars make virtual reality better? Examining human-like avatars for VR social interactions. *Computers in Human Behavior Advances*, 4:100082, 2024.

[9] Stephan Kimmel, Erik Landwehr, and Wilko Heuten. Kinetic connections: Exploring the impact of realistic body movements on social presence in collaborative virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–30, 2024.

[10] Yuhan Wu, Yifan Wang, Sungryul Jung, et al. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality*, 2:641296, 2021.

[11] Daniel Immanuel Fink, Moritz Skowronski, Johannes Zagermann, Anke Verena Reinschluessel, Harald Reiterer, and Tiare Feuchtner. There is more to avatars than visuals: Investigating combinations of visual and auditory user representations for remote collaboration in augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 8:540–568, 2024. doi: 10.1145/3698148.

[12] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers in Robotics and AI*, 6:5, 2019. doi: 10.3389/frobt.2019.00005.

[13] Iana Podkosova, Francesco De Pace, and Hugo Brument. Joint action in collaborative mixed reality: Effects of immersion type and physical location. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, pages 1–12, 2023. doi: 10.1145/3607822.3614541.

[14] Harrison Jesse Smith and Michael Neff. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. doi: 10.1145/3173574.3173863.

[15] Ryan Ghamandi, Ravi K. Kattoju, Yahya Hmaiti, Mykola Maslych, Eugene M. Taranta II, Ryan P. McMahan, and Joseph J. LaViola Jr. Unlocking understanding: An investigation of multimodal communication in virtual reality collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024. doi: 10.1145/3613904.3642491.

[16] Divine Maloney, Guo Freeman, Aaron Robb, Joel Andros, Julie Williamson, and Stephen Brewster. Talking without a voice: Understanding nonverbal communication in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020. doi: 10.1145/3415246.

[17] Gary M. Olson and Judith S. Olson. Distance matters. *Human–Computer Interaction*, 15 (2–3):139–178, 2000. doi: 10.1207/S15327051HCI1523_4.

[18] Pei Wang, Yifan Wang, Mark Billinghurst, et al. Behere: A VR/SAR remote collaboration system based on virtual replicas sharing gesture and avatar in a procedural task. *Virtual Reality*, 27:1409–1430, 2023.

[19] Milica Lazovic. Spatial resources in pre-service teachers' instructional practices in VR tandems: Co-constructing shared spaces and embodied spatial scaffolding. *Frontiers in Communication*, 10, 2025.

[20] Xian Wang, Luyao Shen, Lei Chen, and Mingming Fan. Teamportal: Exploring virtual reality collaboration through shared and manipulating parallel views. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3314–3324, 2025. doi: 10.1109/TVCG.2025.3549569.

[21] Andrew Irlitti, Mesut Latifoglu, Thuong Hoang, Brandon Victor Syiem, and Frank Vetere. Volumetric hybrid workspaces: Interactions with objects in remote and co-located telepresence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024. doi: 10.1145/3613904.3642814.

[22] Florian Weidner, Gerd Boettcher, Stephanie Arévalo Arboleda, Chenyao Diao, Luljeta Sinani, Christian Kunert, Christoph Gerhardt, Wolfgang Broll, and Alexander Raake. A systematic review on the visualization of avatars and agents in AR & VR displayed using head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2596–2606, 2023. doi: 10.1109/TVCG.2023.3247072.

[23] Haoran Yun, Jose Luis Ponton, Carlos Andujar, and Nuria Pelechano. Animation Fidelity in Self-Avatars: Impact on User Performance and Sense of Agency . In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 286–296, Los Alamitos, CA, USA, March 2023. IEEE Computer Society. doi: 10.1109/VR55154.2023.00044. URL https://doi.ieeecomputersociety.org/10.1109/VR55154.2023.00044.

[24] Kwan Min Lee. Presence, explicated. *Communication Theory*, 14(1):27–50, 2004. doi: 10.1111/j.1468-2885.2004.tb00302.x.

[25] Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1535):3549–3557, 2009. doi: 10.1098/rstb.2009.0138.

[26] Frank Biocca, Chad Harms, and Jenn Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *Proceedings of the 4th International Workshop on Presence*, pages 1–9, 2001.

[27] Nelson Wong and Carl Gutwin. Support for deictic pointing in CVEs: Still fragmented after all these years. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1377–1387, New York, NY, USA, 2014. Association for Computing Machinery. doi: 10.1145/2531602.2531691.

[28] Sven Mayer, Jens Reinhardt, Robin Schweigert, Brighten Jelke, Valentin Schwind, Katrin Wolf, and Niels Henze. Improving humans' ability to interpret deictic gestures in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376340.

[29] Maurício Sousa, Rafael dos Anjos, Daniel Mendes, Mark Billinghurst, and Joaquim Jorge. Warping deixis: Distorting gestures to enhance collaboration. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3290605.3300838.

[30] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29 (7):819–833, 2003.

[31] Masahiro Mori. The uncanny valley. *IEEE Robotics and Automation Magazine*, 19(2):98–100, 2012. Original work published 1970.

[32] Jean-Paul Stein and Peter Ohler. Venturing into the uncanny valley of mind: The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, 146:129–135, 2016.

[33] Jean-Noël Kaiser, Stephan Kimmel, Elmar Licht, et al. Get real with me: Effects of avatar realism on social presence and comfort in augmented reality remote collaboration and self-disclosure. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.

[34] Theo Combe, Rebecca Fribourg, Lucas Detto, and Jean-Marie Normand. Exploring the influence of virtual avatar heads in mixed reality on social presence, performance and user experience in collaborative tasks. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2206–2216, 2024. doi: 10.1109/TVCG.2024.3372051.

[35] Bernardo Marques, Samuel Silva, Carlos Ferreira, Sérgio Oliveira, Andreia Santos, and Paulo Dias. Mixed reality collaboration: How expert representations shape user experiences.

IEEE Computer Graphics and Applications, 45(2):143–151, 2025. doi: 10.1109/MCG.2025.3537316.

[36] Andrew Fraser. Behavioural realism and its impact on virtual reality social interactions involving self-disclosure. *Applied Sciences*, 15(6):2896, 2025.

[37] Betty J. Mohler, William B. Thompson, Sarah H. Creem-Regehr, et al. A full-body avatar improves egocentric distance judgments in an immersive virtual environment. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, pages 194–194, 2008.

[38] Jeanne Fröhlich and Marc Erich Latoschik. Minimal avatar fidelity for embodied communication in immersive virtual environments. *Frontiers in Virtual Reality*, 3:804216, 2022. doi: 10.3389/frvir.2022.804216.

[39] Guillaume Gamelin, Amine Chellali, Christophe Dumas, and Samir Otmane. Fidelity of the remote partner's avatar in an immersive virtual environment: Effects on spatial interactions. In *Proceedings of the 30th Conference on Interaction Homme-Machine*, pages 227–233, 2018.

[40] Florian Mathis, Kami Vaniea, and Mohamed Khamis. Observing virtual avatars: The impact of avatars' fidelity on identifying interactions. In *Proceedings of the 24th International Academic Mindtrek Conference*, pages 154–164, 2021.

[41] Matt Gottsacker, Nels Numan, Anthony Steed, Gerd Bruder, Gregory F. Welch, and Steve Feiner. Decoupled hands: An approach for aligning perspectives in collaborative mixed reality. *arXiv preprint arXiv:2503.12253*, 2025. doi: 10.48550/arXiv.2503.12253.

[42] Sandra G. Hart. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

[43] Normal. Normcore v2. https://normcore.io, 2020. Accessed January 2026.

[44] Xenia K. Xanthidou et al. Impact of collaborative virtual environments on students' performance while learning. In *Proceedings of the IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 231–236, 2024.

[45] UMA 2 – Unity Multipurpose Avatar. https://assetstore.unity.com/packages/3d/characters/uma-2-unity-multipurpose-avatar-35611, 2021. Unity Asset Store, accessed January 2026.

[46] Thomas C. Elliott, Yanzhuo Yang, Jarrod Knibbe, Julie D. Henry, and Nilufar Baghaei. Avatar customization and embodiment in virtual reality self-compassion therapy for depressive symptoms. *JMIR Formative Research*, 9(1):e46872, 2025. doi: 10.2196/46872.

[47] Unity Technologies. Unity animation rigging package. https://docs.unity3d.com/Packages/com.unity.animation.rigging@1.0/manual/index.html, 2020. Accessed January 2026.

[48] Yi-Ting Huang, Chih-Chieh Hsu, and Tzu-Hsuan Wang. Effects of interactive loading interfaces for virtual reality game environments on time perception, cognitive load, and emotions. *Frontiers in Virtual Reality*, 6, 2025. doi: 10.3389/frvir.2025.1540406.

[49] Jeremy N. Bailenson, Jim Blascovich, and Rosanna E. Guadagno. Self-representations in immersive virtual environments. *Journal of Applied Social Psychology*, 38(11):2673–2690, 2008. doi: 10.1111/j.1559-1816.2008.00409.x.

[50] Meta Platforms, Inc. Meta XR SDK (presence platform) for Unity. https://developers.meta.com/horizon/documentation/unity/, 2024. Accessed: 2026-03-20.

[51] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308, 2013.

[52] Hadrien Brument, Anne-Hélène Olivier, Richard Kulpa, and Julien Pettré. Shared control of locomotion in virtual reality for collaborative navigation. In *Proceedings of the ACM Symposium on Applied Perception (SAP)*, pages 1–9, 2020.

[53] Chen Li, Yixin Dai, Guang Chen, Jing Liu, Ping Li, and Horace Ho-shing Ip. Avatar-mediated communication in collaborative virtual environments: A study on users' attention allocation and perception of social interactions. *Computers in Human Behavior*, 167:108598, 2025. doi: 10.1016/j.chb.2025.108598.

[54] R. Assaf, D. Mendes, and R. Rodrigues. Cues to fast-forward collaboration: A survey of workspace awareness and visual cues in XR collaborative systems. *Computer Graphics Forum*, 2024. doi: 10.1111/cgf.15066.

[55] Karl F. MacDorman, Robin D. Green, Chin-Chang Ho, and Clinton T. Koch. Too real for comfort? Uncanny responses to computer-generated faces. *Computers in Human Behavior*, 25(3):695–710, 2009.