Effects of False Positive and False Negative Early Warnings for Hand Tracking Failures on User Trust and Confidence in Virtual Reality

Amal Hatira, Zahra Rasoulifar, Mucahit Gemici, Vrushank Phadnis, Anil Ufuk Batmaz

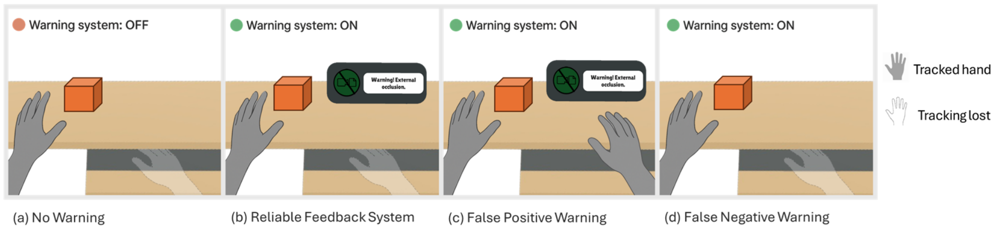

Fig. 1. Schematic representations of the four study conditions: (a) No Warning (NW), in which failures occur without warnings; (b) Reliable Early Warning (EW), in which warnings precede failures; (c) False Positive (FP), in which a warning is presented without a subsequent failure; and (d) False Negative (FN), in which a failure occurs without a warning.

Abstract— Early-warning systems (EWS) mitigate hand-tracking failures in Virtual Reality (VR); yet, their effectiveness depends on reliability. Predictive warning mechanisms are prone to errors as they produce false positives (FP) and false negatives (FN). In VR, these errors directly interfere with continuous sensorimotor control and user agency when the hand is the main input channel, potentially affecting user trust and confidence in the system while changing user behavior. Thus, we conducted a within-subject study ($N = 24$) comparing a reliable EWS baseline with controlled FP and FN warning conditions and a no-warning condition. Results showed that both error types reduced system trust; however, FP caused a larger and immediate decline (48.7%) compared to FN (25.7%), which emerged progressively. Self-confidence remained stable across FP and FN conditions. FP led to reduced compliance over time, whereas FN was associated with more low-tracking states and higher mental and temporal demand. These findings show that FP and FN errors create distinct experiential and behavioral costs during VR hand interaction, and that warning systems should be evaluated through trust, workload, usability, and behavioral response in addition to technical reliability.

Index Terms—Virtual Reality, Early Warning, Trust, False Positive, False Negative

1 INTRODUCTION

Hand tracking in Virtual Reality (VR) systems enables users to intuitively interact in immersive environments without requiring additional hardware. It is often used as the primary input channel, allowing users to control virtual hands that mirror their real-world hand and finger movements. This virtual motion provides a seamless interaction experience, making hand-tracking-based input one of the preferred interaction methods in virtual environments [37].

Yet, effective virtual hand interaction relies on hand-tracking systems. Despite hand-tracking systems being available in current state-of-the-art commercial VR headsets, they are fragile and prone to errors, such as the hands occluding each other [39, 18, 79] or being occluded by external objects [54, 52, 66], poor lighting in the real environment [50], and sensors' limited field of view for hand tracking [20]. These conditions can cause jitter [68] or misplacement of virtual hands [52, 67] or even complete hand-tracking failure [2]. Such inconsistencies break immersion [22], decrease performance [68], and can disrupt task execution [7].

Unlike traditional input devices that rely on mechanical components, hand tracking in VR is primarily based on computer vision algorithms. As a result, failures are often non-deterministic and difficult for users to anticipate. In the physical world, a broken mechanical switch or button provides clear and consistent feedback, allowing users to quickly understand the source of the problem. In contrast, hand-tracking failures may occur intermittently and without obvious external cues, making it difficult for users to develop an intuitive understanding of when and why errors occur. This unpredictability makes hand-tracking-related errors particularly important to investigate in VR interaction research.

To address these challenges, researchers have studied early-warning feedback mechanisms for hand-tracking systems that warn users about *potential* hand-tracking failures *before* they occur [7, 22]. For example, Gemici et al. [22] developed a visual warning system for common failure scenarios, i.e., low-intensity light, hand occlusion, and out-of-vision. Their approach reduced task completion time and tracking failures by 83%, reduced overall task load, and increased the perceived usability.

However, early-warning systems operate under uncertainty as they rely on predictive algorithms and are therefore susceptible to false positives (FPs; Type I errors) and false negatives (FNs; Type II errors). False positives arise when the system presents a warning, but the failure does not occur, e.g., when a user adjusts their hand position in time or when environmental lighting is corrected before tracking is lost. Yet, frequent false positives can lead to the "cry wolf" effect, where users start ignoring or delaying responses to warnings [49, 10, 9].

- *Amal Hatira is with Concordia University. E-mail: amal.hatira@mail.concordia.ca*
- *Zahra Rasoulifar is with Concordia University. E-mail: zahra.rasoulifar@mail.concordia.ca*
- *Mucahit Gemici is with Concordia University. E-mail: mucahit.gemici@mail.concordia.ca*
- *Vrushank Phadnis is with Google Inc. E-mail: vrushank@google.com*
- *Anil Ufuk Batmaz is with Concordia University. E-mail: ufuk.batmaz@concordia.ca*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

False negatives occur when the system fails to detect conditions that lead to actual tracking breakdowns, such as miscalibrated thresholds, atypical user characteristics (e.g., body height or hand size), or environmental factors outside the system's detection range [16]. These false negative cases may be more problematic, as they undermine trust and confidence in the feedback. In this context, trust refers *to an individual's willingness to rely on the system* [60, 36], whereas confidence refers to *belief in one's ability to perform a task successfully* [46]. Both can be negatively affected by unreliable feedback, shaping behavioral reliance patterns.

Trust and confidence are particularly important in VR hand interaction, where the virtual hand serves simultaneously as an input mechanism, a visual representation of the user's physical hand, and a component of embodied sensorimotor interaction. Users continuously coordinate their physical movements with the observed movement of the virtual hand [26]. Consequently, tracking failures can disrupt not only task execution but also the perceived correspondence between physical action and virtual feedback. In practice, early-warning systems, like all predictive mechanisms, will produce both false positives and false negatives. Understanding how these errors influence user experience is therefore essential for real-world deployment. However, in such mechanisms, FP and FN rates are inherently coupled through internal detection thresholds, making it difficult to isolate their individual effects. To address this challenge, we adopted a controlled experimental approach in which FP and FN errors were manipulated separately. This design enabled us to examine how each error type influenced trust, confidence, workload, and compliance while reducing confounding from changes in detection thresholds and interaction timing.

Although other domains, such as automation, aviation, autonomous driving, and medical decision support, have studied trust in relation to system reliability and human involvement, less is known about how these dynamics play out in VR, especially for systems that are vulnerable to false positives and false negatives. Prior work shows that trust evolves over time as users observe system behavior and develop a mental model of its reliability [8]. Repeated exposure to correct or incorrect automation responses also shapes user reliance and calibration [78]. However, these findings may not directly transfer to VR hand interaction. In many desktop or automation settings, system outputs are discrete and separated from users' physical actions, whereas VR hand tracking continuously mediates users' motor actions and the visual representation of their hands. Consequently, a false positive may prompt an unnecessary movement adjustment, while a false negative may leave the user without support when tracking quality degrades. Existing research on VR hand tracking has largely focused on improving accuracy, latency, and robustness [63, 1], rather than examining how early-warning reliability shapes trust, confidence, workload, usability, and behavioral reliance during embodied interaction.

In our study, we systematically manipulated the reliability of a VR early-warning system for hand-tracking failures by separately introducing false-positive and false-negative warnings. This allowed us to examine not only whether warning unreliability reduced trust, but also whether false positives and false negatives created distinct experiential and behavioral costs during embodied VR hand interaction. Our contributions are:

- We show that false positives and false negatives produce distinct trust dynamics in VR early-warning systems for hand-tracking failures: false positives caused a larger and earlier decline in trust, while false negatives led to a slower but still significant decline.
- We show that FP and FN create distinct experiential and behavioral costs during embodied VR hand interaction: FP caused lower usability, higher frustration, and declining compliance, whereas FN caused higher mental and temporal demand and more low-tracking states.
- We provide design implications for XR early-warning systems by showing that implementation decisions–including warning sensitivity, timing, confidence thresholds, and trigger criteria–should be evaluated not only through technical reliability but also through their effects on trust, usability, workload, and behavioral response.

2 RELATED WORK

2.1 Early Warning Systems for Hand Tracking Failures

Hand tracking enables intuitive, hands-free interaction and is widely used for object manipulation in VR. Modern commercial VR headsets track hand movements with cameras, sensors, and image-processing techniques [13, 31]. These systems rely on built-in RGB and depth cameras to capture the user's hands. Computer vision algorithms, often supported by machine learning, then process these images to identify and track key points on the hands [12]. However, these algorithms are prone to failures caused by *environmental factors* e.g., low light conditions [50], *occlusions* e.g., hands occluding each other [39, 18, 79], or being externally occluded [54, 52, 66], or *hardware related* e.g, the hands being out-of-vision for the headset's sensors [20]. These failures decrease performance, increase frustration, and can compromise task success.

Recent studies have proposed early warning systems for hand tracking failures that warn users about *potential* failures *before* they occur [22]. Bashar and Batmaz [7] demonstrated that visual early-warning feedback reduces frustration and task completion time in low-light conditions, while Gemici et al. [22] showed that such systems improve usability and reduce cognitive load. Further gaze analysis showed that users could attend to these warnings without substantially disrupting attention to the primary task [21].

Yet, while these findings highlight the potential of early warning systems, they leave open questions about the conditions under which such systems remain dependable and effectively guide user behavior. Building on this, here, we examine how the reliability of early warnings shapes user experience and interaction in VR.

2.2 Trust and Confidence in System Warnings

Warnings affect how people respond to automated aids via two distinct constructs: trust in the system and decision confidence about one's own actions [41]. Wiczorek et al. [72] showed in a signal detection task that feedback increased confidence but had no effect on trust; moreover, trust predicted reliance while confidence did not. Together, these studies highlight that trust and confidence are related and both must be considered when evaluating how users interpret and act upon system outputs such as warnings.

In the context of automation, trust is commonly defined as the user's attitude that the system helps to achieve goals under uncertainty [41]. This attitude guides whether users comply with alerts or rely on the system's silence. Lee and See [41] propose that trust is shaped by cues about system performance, processes, and purpose, and that appropriate reliance depends on calibrated trust. Calibrated trust means users rely appropriately when their trust matches the system's actual reliability, and HCI studies show that accuracy, error feedback, and explanations can affect this [41, 73, 77, 6]. Hoff et al. also discussed that over-trust may lead to misuse, while under-trust may lead to disuse, as emphasized earlier by Parasuraman and Riley [30, 57]. Kohn et al. [36] review trust measurement approaches, noting that studies should combine subjective self-reports with behavioral outcomes such as compliance and reliance to capture trust comprehensively. More recent studies have also explored short-term trust dynamics through multimodal measures, such as combining gaze allocation with behavioral performance and self-reported trust in immersive simulation environments [64].

Decision confidence reflects a user's metacognitive judgment about the correctness of their own action or decision [11]. It concerns a user's own estimate of being correct [5]; by contrast, trust is a belief about an aid's performance that guides reliance [17]. These definitions show that confidence and trust are related but not the same. In practice, confidence is often measured using trial-level ratings and calibration analyses that assess the alignment between subjective certainty and actual performance [23]. Recent studies have shown that aligning users' confidence with system-provided confidence cues can affect reliance on decision aids [42].

Both trust and confidence are sensitive to the reliability of system warnings. Experimental studies show that when systems behave in-consistently, trust decreases and users adapt their reliance accordingly [17, 71], and confidence in one's own actions is affected by external cues, including perceptions of AI confidence [42].

This underscores that user trust and confidence depend not only on objective accuracy but also on the perceived reliability of warnings [17, 71, 73]. VR warning systems, especially those addressing hand-tracking failures, have rarely examined these psychological constructs in relation to warning reliability, focusing instead on performance and usability.

2.3 False Positives and False Negatives in Warning Systems

Early-warning feedback systems operate under inherent uncertainty, since they attempt to predict potential failures before they actually occur. This predictive nature makes them susceptible to two kinds of errors. *False positives- false alarms-Type I errors* occur when the system issues a warning even though no failure happens, for example, when conditions change after the warning and the anticipated failure does not occur [10, 71]. While false alarms can reduce responsiveness and contribute to warning fatigue, warning research emphasizes that clear and accurate alerts are essential to encourage precautionary or protective actions [45, 58]. *False negatives-missed alarms - Type II errors*, in contrast, occur when the system fails to issue a warning despite a potential failure, for example, when thresholds are set too strictly or when models are miscalibrated or encounter inputs outside their training distribution [24, 55, 29]. Such misses can be more problematic, as they can undermine user trust and lower perceived reliability [48, 4, 61, 47].

Since early-warning systems operate under uncertainty, designing them requires balancing false positives and false negatives to preserve both effectiveness and user trust. This balance depends on how often events occur (the base rate) and on the relative costs of the two error types. Studies show that both people and automated aids adjust their response thresholds according to these base rates and costs [70, 71]. In some domains, where there is potential for injury and loss of life, this often means biasing thresholds to avoid costly misses, even if this increases false alarms [38].

Human–automation studies show that different types of errors affect user behavior in different ways [16, 51]. In diagnostic-aid tasks, false alarm-prone automation reduces compliance (the cry wolf effect, people start ignoring alerts), and can also reduce reliance, whereas miss-prone automation primarily reduces reliance when the aid is silent [16, 51]. This asymmetry has been observed in controlled simulations that separately measure compliance (responses to alerts) and reliance (behavior when no alert is issued) as distinct behaviors [16, 51]. While these dynamics have been demonstrated in decision aids and automation research [16, 51], their impact in 3D virtual environments is important since presence and usability in VR depend on reliable real-time tracking, and errors can disrupt both the sense of presence and task performance [65]. Decision confidence has been shown to align with system-provided confidence cues and affect reliance on automation [42]. Extending these insights, VR warning systems may similarly shape both compliance and confidence. This connection between error patterns, trust, and confidence has not yet been systematically explored in VR for early warning systems.

Frequent false positive warnings can produce alarm fatigue, which is the desensitization that delays or suppresses responses to valid alarms [10, 19]. Integrative reviews in healthcare document how high false alarm rates and poor positive predictive value decrease responsiveness and safety, motivating structured alarm management practices [14, 74]. Beyond healthcare, public warning research shows a related cry wolf phenomenon: prior exposure to false alarms reduces later compliance with genuine warnings [69]. In controlled experiments with weather decision aids, very high false alarm histories reduce compliance and decision quality, whereas providing probabilistic uncertainty (e.g., numeric likelihoods) improves both outcomes; importantly, simply lowering the false alarm rate slightly may not recover compliance [38, 33]. Survey and field analyses around tornado warnings have examined whether false alarm exposure reduces trust or discourages protective action, but found little evidence for these effects [44]. These findings emphasize that users update their trust dynamically across repeated interactions, meaning that error history can matter as much as overall system accuracy [41].

The impact of false positives versus false negatives depends on the context: false negative (failing to warn about a hazard) can reduce trust more than false positive, and higher perceived risk may further reduce trust overall [4]. This aligns with prior research showing that error type matters: false positives tend to lower compliance, while false negatives tend to lower reliance [16, 51]. Taken together, these studies highlight that the "worse" error type depends on the task and its risks, and false negatives can be especially damaging in safety-critical domains [16, 51, 17, 4].

3 TASK AND WARNING DESIGN

We designed two object pick-and-place tasks that were distinct in purpose but closely matched in complexity and duration. A pilot study with six participants, balanced by gender (three female and three male), aged 25–31 years ($M = 26.5$, $SD = 2.35$), was conducted to assess task duration and interaction complexity. Both tasks followed the same structure and number of placements, helping ensure comparable procedural demands. A one-way repeated-measures ANOVA showed no significant difference in task-completion time between the Toy Pick & Place task ($M = 156.64$ s, $SD = 47.28$) and the Assembly task ($M = 152.07$ s, $SD = 48.85$), $p = .600$.

For **Toy Pick & Place** task, we positioned two baskets at the center of the scene: one on the left and one on the right (Figure 2 (a)). The goal was to pick objects from the baskets and place them on the shelves. After each placement, two new objects appeared, one assigned to each basket, together with their semi-transparent "ghosts" that indicated the correct placement on the shelves (Figure 2 (a)). Participants were asked to pick up both objects simultaneously, align them with their corresponding ghosts, and release them into place. Participants completed 12 dual-object placements in a predetermined fixed order.

We also designed an **Assembly** task. At the beginning of the assembly task, all parts were visible, but the object to be grasped was highlighted by an outline. Similar to the Toy task, a semi-transparent "ghost" indicated the correct target position and orientation for that part. Participants had to grasp the part, transport, rotate, and align it with the ghost, and then release it. Once the part was positioned, the next object was highlighted, guiding participants through the task in a fixed order (Figure 2 (b)). In total, participants completed 12 placement trials, matching the trial structure of the Toy Pick & Place task.

We chose to use two tasks instead of one to enable a within-subject design while reducing potential carryover of trust and confidence across false positives and false negatives conditions, as trust depends strongly on initial experiences and is difficult to recalibrate [69]. Using two tasks ensured that participants were engaged in interaction scenarios that felt different in purpose (toy placement versus assembly) while remaining matched in complexity, duration, and number of

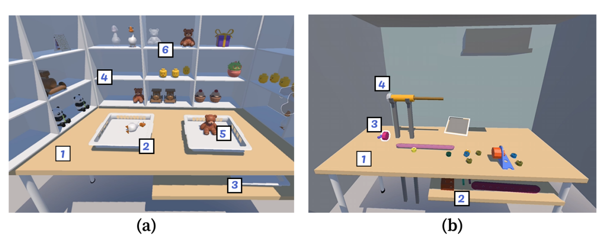

Fig. 2. (a) Toy pick and place task, having (1) the virtual desk, (2) two baskets on top of the desk, (3) one on the shelf under the desk, and (4) shelves to place the toys. Participants grabbed the toy (5) and placed it on its "ghost" position (6). (b) the assembly task, having (1) the virtual desk, (2) the shelf under the desk, (3) the object to be grasped highlighted, and (4) its "ghost" placement position.

placements. In the experiment, we counterbalanced both tasks and error conditions (false positives and false negatives) to minimize bias. For example, if one participant completed the Toy Pick & Place task with false-positive warnings, they performed the Assembly task with false-negative warnings. Conversely, another participant would complete the Assembly task with false positives and the Toy Pick & Place task with false negatives.

3.1 Hand-Tracking Failure Types

We implemented four controlled hand-tracking failure scenarios in VR, following and extending the conditions previously examined by Gemici et al. [22]. To ensure methodological consistency and enable direct comparison with their results, we replicated the low-light, self-occlusion, and out-of-vision scenarios following the procedures described in their study. In addition, we extended their work by incorporating an external occlusion condition, which was mentioned in their paper but not experimentally evaluated. For each scenario, we distinguished between the task setup that induced the risk of tracking failure and the rule-based trigger that generated the early warning.

Out-of-Vision: The scenario was induced by requiring participants to move one hand toward or beyond the headset's effective field of view while attending to another task element. The warning trigger was rule-based: We computed Forward-Angle, Right-Angle, and RelativePosition-Y based on a vector from the head position to the hand, and we set thresholds similar to the ones suggested by Gemici et al., [22], to trigger the warnings. In the assembly task, this scenario occurred when participants reached into the top right box to grab the parts (Figure 3 (a -b)), requiring them to rely on a top-view camera screen(Figure 3 (b -c)). In the toy task, the out-of-vision condition was induced when participants placed two toys far away from each other, forcing one hand to go out of the field of view while focusing on the other hand to correctly align the toy object (Figure 3 (d-f)).

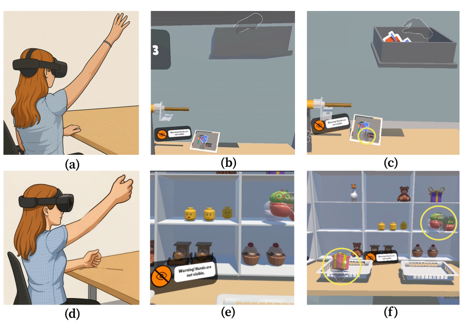

Fig. 3. Out-of-vision condition. For the assembly task: (a) the participant reaches for objects placed in a virtual box at the top-right of the scene; (b) the participant monitors a screen displaying a top-view of the box; (c) a top-view of the scene shows the hand reaching for the objects, with its position indicated on the screen (yellow circle). For the toy pick-and-place task: (d) the participant aligns two objects positioned far apart, so that (e) only one hand remains within the camera's field of view; (f) a zoomed-out view shows both hands (yellow circles), where the warning is triggered for the left hand since the participant's head orientation is directed toward the right hand.

Low Light: The low-light scenario was induced using a Wizard-of-Oz method in which the experimenter toggled the physical room lights. We implemented an algorithm that alternated between Normal-Light and Low-Light states. To prevent participants from predicting when the light failure would occur, we randomized the duration of the Normal-Light state between 5 - 7 seconds and the Low-Light state between 1 - 3 seconds. The warning trigger was based on the controlled light-state transition: during the final 2 seconds of the Normal-Light state, the system displayed the Low-Light warning before the room entered the Low-Light state.

Self Occlusion: The self-occlusion scenario was induced by requiring participants to position one hand in front of or across the other from the headset's viewpoint. The warning trigger was rule-based: we cast rays from the user's head position to both hands and placed trigger colliders around the hand models. If the ray intersected both hands' colliders, indicating that the hands were close together or overlapping from the headset viewpoint, the system presented an early warning. In the toy task, this occurred when the participants tried to place two objects simultaneously, when one hand had to cross in front of the other (Figure 4 (b)). In the assembly task, self-occlusion was induced during simultaneous placement of two parts, one in front of the other (Figure 4 (c)).

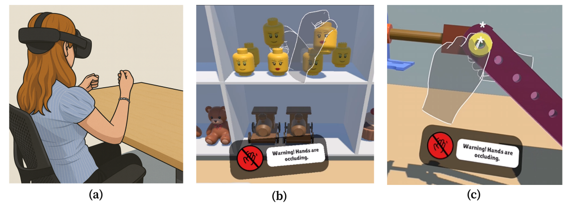

Fig. 4. Self-occlusion condition: (a) participant with hands positioned one behind the other, (b) toy task where the virtual hands cross each other while placing two objects, and (c) assembly task where two parts are assembled one behind the other.

External Occlusion: We aligned the virtual desk used in the task with the real physical desk (Figure 5 (a,b)). To induce external occlusion, we placed objects on a shelf under the desk, so that participants had to reach under it to pick them. When doing so, the physical desk blocked the HMD cameras' view of the hand (Figure 5 (c)). To ensure that participants could still visually locate the target objects, the material of the virtual desk was changed from opaque wood to transparent glass, allowing them to see the shelf contents while their hands were occluded by the real physical desk (Figure 5 (d)). The external-occlusion warning trigger was implemented as an invisible collider under the desk around the shelf area; when a participant's hand entered this region, the system presented an external-occlusion warning (Figure 5 (d)).

Except for the low-light condition that is externally toggled, all warnings were triggered based on the participants' movements. Thus, "near-perfect" EW refers to reliable warning behavior within these four controlled scenarios and rule-based triggers, rather than to a universal detector of all spontaneous hand-tracking failures.

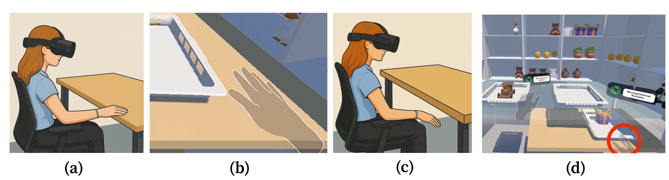

Fig. 5. (a) Participants place their right hand on the physical desk to align it with the virtual desk, (b) demonstrated for the toy task. In the external occlusion condition, participants reach for an object placed on a virtual shelf beneath the desk (c). To support visibility, the desk material is changed to glass, allowing participants to see the objects (d). The red circle highlights the hand reaching under the desk, which triggers the external occlusion warning.

3.2 Visual Warnings Design

For the visual warnings, we relied on the design used in an earlier study on VR early-warning systems by Gemici et al., [22]. The warning modal was easily recognizable, with color-coded symbols paired with short textual labels, displayed centrally within the user's field of view. The colors were chosen from a color-blind friendly palette. Different tracking failures were mapped to unique iconmessage combinations (Figure 6). Each warning was presented as soon as a failure

Fig. 6. Visual warnings for (a) external occlusion, (b) out-of-vision, (c) low light, and (d) self-occlusion.

was predicted and remained visible for two seconds, providing users with enough time to notice the warning.

4 RESEARCH QUESTIONS AND HYPOTHESES

Hand tracking offers intuitive and hands-free interaction in VR, but remains prone to failures. In scenarios where the virtual hand is the primary input modality, early-warning systems for hand-tracking failures can improve performance and task success while preserving user agency and reducing frustration [22, 7]. While promising, early warning systems' effectiveness depends on reliability.

In this work, we aim to examine how false-positive and false-negative early warnings for hand-tracking failures affect user trust, confidence, workload, usability, and behavioral response in VR. Thus, we investigate **(RQ1)** *How do FP and FN early-warning errors affect users' trust and confidence in a VR hand-tracking early-warning system?* Trust and confidence determine whether users will rely on the system[72, 15]; if they decrease, the benefits of early warnings might disappear. Within RQ1, we hypothesize that **H1a:** Both FP and FN errors reduce system trust and system-provided confidence relative to the reliable baseline (EW), and **H1b:** compared with FN errors, FP errors produce a larger reduction in trust and system-provided confidence (cry-wolf effect).

We also investigate: **(RQ2)** *How does unreliable early-warning feedback affect subjective usability and perceived task load?* To ensure early warnings support task execution rather than increase the task demand, we assess whether usability decreases or perceived workload increases. We hypothesize that **H2a:** Perceived usability decreases with FP and FN compared to baseline EW, and that **H2b:** Both FP and FN increase workload compared to EW.

Moreover, we investigate **(RQ3)** *How do FP and FN errors affect behavioral reliance on warnings (compliance behavior) and task performance?* Behavioral reliance directly determines the practical effectiveness of the early warnings [53]. If users stop complying with the warnings after repeated false positives or false negatives, their performance might decrease. We hypothesize that **H3a:** Under unreliable early warnings (FP or FN), compliance with warnings decreases across task phases **H3b:** FN errors increase the number of uncorrected tracking losses, and that **H3c:** Task time increases with FP and FN compared to the baseline EW condition.

5 METHODOLOGY

5.1 Participants

We recruited 24 participants (12 female, 12 male); aged 19 – 40 ($M = 26.5, SD = 5.24$). All participants had normal or corrected-to-normal vision. 11 participants reported using VR more than 5 times in their life, 3 reported 3-5 times, 8 reported using it 1 - 2 times, and 2 participants reported never experiencing it. 14 participants had prior experience with hand interactions in VR, while the remaining 10 were experiencing it for the first time. Participation was voluntary and uncompensated. All procedures were approved by the university research ethics board.

5.2 Apparatus

We used a *Meta Quest Pro* connected to a PC with an Intel i7 CPU, 16 GB RAM, NVIDIA RTX 3060 GPU. The experiment software was implemented in *Unity 2022.3.34f1* with the Meta XR Interaction SDK OVR Integration. The experimenter monitored runs on a desktop. Questionnaires were implemented as in-VR UI panels.

5.3 Experimental Design and Conditions

We used a within-subjects design with four conditions administered in three sequential stages: NW, EW, and an unreliable-warning stage containing the FP and FN conditions. EW served as the primary baseline for evaluating FP and FN, as both represent reliability errors of the same warning system. NW served as a separate control for task execution under hand-tracking failures without warnings. The overall progression from NW to EW to unreliable warnings was fixed, while task order, failure-type pairings, and FP/FN assignment were counterbalanced.

Failures Only - No Warnings (NW) Participants completed both tasks with the controlled hand-tracking failure scenarios (two failure types per task; counterbalanced across participants). We did not display any early warning feedback. After finishing the *second* task of this round, participants filled in-VR questionnaires.

Reliable Early Warning (EW) Participants repeated the tasks with the same controlled failure scenarios and with *early-warning system* feedback. During pilot testing, we tuned the rule-based thresholds for each predefined failure scenario to minimize warning errors in EW. We simulated two hand tracking failure types per task and counterbalanced them. After the second task, participants filled in-VR questionnaires.

Unreliable Early Warning (FP vs. FN) Participants repeated both tasks with the early warning system enabled, but *with reliability errors* introduced. One task included *false-negative* warnings and the other included *false-positive* warnings. We manipulated the frequency of warnings based on pilot testing. For the false negative condition, we reduced the number of warnings by approximately 50%, such that only half of the potential warnings were shown. The system randomly determined which warnings to suppress, ensuring that missed warnings occurred unpredictably. For the false positive condition, we introduced additional warnings at fixed intervals (approximately once every 5–6 seconds), resulting in an average of 50% increase in total warnings compared to the reliable baseline. Participants were not informed of the exact FP or FN manipulation during the tasks.

We manipulated warning presentation directly rather than modifying internal detection thresholds. The Interaction SDK's native virtual-hand behavior was not suppressed. Consequently, freezing or disappearance of the virtual hands remained possible and was consistent across conditions. This could provide late feedback about a tracking failure in NW and FN trials. However, our analyses included only tracking losses associated with the four controlled failure scenarios and excluded unrelated spontaneous tracking losses. Adjusting thresholds would have simultaneously changed detection timing and sensitivity in ways that were difficult to control. Directly increasing or suppressing warnings allowed us to hold task mechanics constant and better isolate the experiential effects of FP and FN errors. Importantly, our analysis was limited to the four predefined failure contexts (out-of-vision, low light, self-occlusion, and external occlusion), which were implemented consistently across NW, EW, and FP/FN conditions. Thus, the FP/FN manipulations reflect controlled changes in warning reliability within these scenarios, rather than all possible spontaneous tracking errors of the Quest/Unity pipeline.

To examine temporal dynamics, we divided the task into 3 phases – defined as a set of four object placements. To capture evolving perceptions as FPs/FNs accumulated, brief *mid-task* trust and confidence questionnaires appeared after each *phase*.

We designed this progression to gradually expose participants to (1) failures without early warnings, (2) a highly reliable early-warning system, and (3) an unreliable system subject to false positives or false negatives. This sequence allowed participants to experience the reliable system before encountering warning errors. Thus, participants did not expect warnings in NW, whereas missed warnings in FN violated expectations established during EW, as prior research showed that trust depends strongly on initial experiences [69]. While the overall progression was fixed, the assignment of false positives and false negatives to tasks was counterbalanced across participants. Therefore, comparisons between FP and FN conditions are less likely to be explained by task assignment alone, although the fixed progression remains a limitation when interpreting trust development over time.

5.4 Counterbalancing

Counterbalancing was applied at three levels to avoid order effects. First, for each participant, each task was assigned exactly two of the four possible failure types {Out-of-Vision, Self-Occlusion, Low-Light, External Occlusion}, yielding $\binom{4}{2} = 6$ possible pairings. Second, in the unreliable early-warning condition, one task was paired with false positives (FP) and the other with false negatives (FN), resulting in two possible assignments (Toy=FP/Assembly=FN or Toy=FN/Assembly=FP). Finally, task order was counterbalanced so that half of the participants began with Toy Pick&Place and half with Assembly, creating another two possible sequences. Taken together, this produced $(6 \times 2 \times 2 = 24)$ unique sequences, and each participant was assigned to one sequence, ensuring that task order, failure pairings, and FP/FN assignments were evenly distributed across the sample.

Exploratory order-effect analyses based on whether FP or FN occurred first or second did not reveal significant carryover effects based on Mann–Whitney tests and independent-samples t-tests (all $p > .05$).

5.5 Evaluation Metrics

We collected both subjective and behavioral metrics to address our three research questions on trust and confidence (**RQ1**), usability and workload (**RQ2**), and behavioral reliance and performance (**RQ3**). All subjective measures were implemented inside VR as panels to avoid interruption, while the behavioral and performance-related measures were automatically logged.

5.5.1 Subjective Measures

Trust We used the 12-item 7-point Likert scale trust in automation questionnaire by Jian et al., [32], which included positive and negative statements about dependability, reliability, and integrity.

Confidence We used two single-item 7-point Likert questionnaires: (i) self-confidence (How confident are you that you can complete the task?) and (ii) system-provided confidence (How confident are you that you can complete the task with the early-warning system?). These questions distinguished participants' confidence in their own abilities from their confidence in the system [40]. The system-provided confidence item was not included in the NW condition because no early-warning system was available to evaluate.

Usability We used the System Usability Scale (SUS) [62] and interpreted scores using its A–F grading scale, where higher grades indicate better perceived usability.

Workload We used the NASA–TLX questionnaire [25].

5.5.2 Behavioral Reliance and Performance

Compliance Rate: We defined compliance as the proportion of warnings followed by corrective hand movement away from the threshold region. A warning was considered complied with if the participant moved their hand out of the critical region within a predefined temporal window (2 seconds) after the warning appeared. This metric directly captures immediate behavioral response to the warning and reflects reliance on the warning system. Compliance was computed only for failure types with an objective corrective hand movement: out-of-vision, self-occlusion, and external occlusion. Low-light trials were excluded because the lighting change was external to the participant's hand position and did not have an equivalent corrective movement.

Number of low-tracking states for hands (N-LTS): To quantify tracking quality, we logged low-tracking states reported by the OVR hand-tracking confidence state. N-LTS was used only as an outcome measure and was not used to trigger warnings. We included only low-tracking states associated with the four controlled failure scenarios. Each low-confidence period was counted as one low-tracking state until tracking confidence recovered.

Task time: Measured from the moment the first object was grasped until the final correct object placement. This provided an overall measure of performance across conditions.

5.6 Procedure

Upon arrival, participants provided informed consent and completed a brief demographic questionnaire. The HMD was then fitted and calibrated. Participants first completed shortened versions of both tasks as training to familiarize themselves with the interaction techniques and user interface. During this training, no hand-tracking failures were simulated, as we wanted to avoid influencing participants' initial trust and confidence in the system. This training also ensured that all participants, regardless of prior VR experience, could successfully perform the interactions. They then proceeded with experimental conditions. After completing the second task in the No Warning and Reliable Early Warning conditions, participants completed the Trust scale, self-confidence rating, SUS, and NASA-TLX. The confidence items were presented only when an early-warning system was available. In the Unreliable Early Warning condition, trust and confidence questionnaires were also presented mid-task after every phase to capture temporal dynamics, followed by the full questionnaires at the end. Each task lasted approximately 3 minutes, and the full session lasted approximately 40–60 minutes.

6 RESULTS

Unless otherwise noted, nonparametric tests (Friedman tests with Wilcoxon signed-rank post hoc comparisons) were used to analyze the results for our within-subjects design, as several dependent variables were ordinal and thus did not meet the assumptions of parametric tests.

Trust and Confidence We compared trust scores across the nearly perfect early warnings (EW), False-Positive (FP), False-Negative (FN), and the No warnings (NW) conditions. A Friedman test revealed a significant effect of condition on trust scores, $\chi^2(3) = 18.801, p < .001$. Post hoc Wilcoxon signed-rank tests with Bonferroni correction (Figure 7 (a)) indicated that trust was lower in FN than in EW ($p < .01$), lower in FP than in EW ($p < .001$), and lower in NW than in EW ($p < .05$). Trust was also lower in FP than in FN ($p < .05$) and NW ($p < .05$). No significant difference was found between FN and NW.

To examine how trust evolved over time, participants rated their trust after each task phase. In the FN condition, a Friedman test revealed a significant effect of phase on trust scores, $\chi^2(3) = 13.903, p < .01$. Post hoc Wilcoxon signed-rank tests (Figure 7 (b)) showed that trust was lower in Phase 2 than in EW ($p < .01$). Trust in Phase 3 was also lower than in EW ($p < .01$) and Phase 1 ($p < .05$). These findings indicate that trust in the system declined progressively with continued exposure to missed warnings.

Similarly, in the FP condition, a Friedman test also revealed a significant effect of phase on trust scores, $\chi^2(3) = 36.621, p < .001$. Post hoc Wilcoxon signed-rank tests (Figure 7(c)) showed that trust was lower than EW after Phase 1, Phase 2, and Phase 3 (all $p < .001$). Trust was also lower in Phase 2 than in Phase 1 ($p < .001$), lower

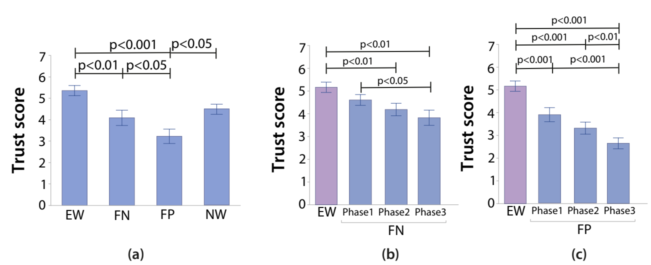

Fig. 7. Mean trust scores: (a) under the different conditions and across different phases of the task for (b) FN, and (c) FP conditions. Brackets indicate significant Bonferroni-adjusted Wilcoxon signed-rank comparisons. Error bars show standard errors.

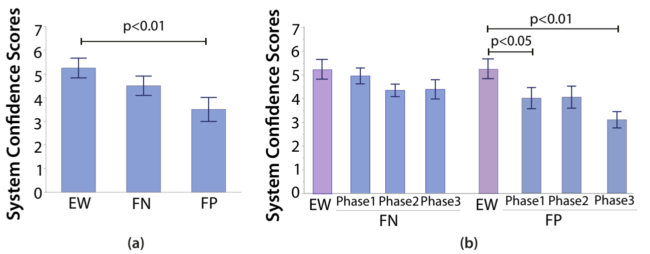

Fig. 8. Mean system-provided confidence scores (a) across conditions, (b) across FN phases, and FP phases. Brackets indicate significant Bonferroni-adjusted Wilcoxon signed-rank comparisons. Error bars show the standard error.

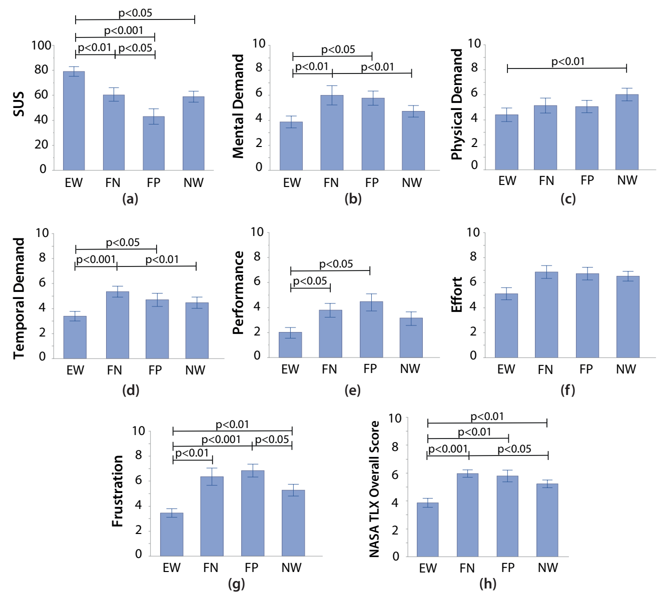

Fig. 9. Mean scores across EW, FN, FP, and NW for (a) SUS, (b) mental demand, (c) physical demand, (d) temporal demand, (e) perceived performance, (f) effort, (g) frustration, and (h) overall NASA-TLX. Brackets indicate significant Bonferroni-adjusted Wilcoxon signed-rank comparisons. Error bars show the standard error of the mean.

in Phase 3 than in Phase 1 ($p < .001$), and lower in Phase 3 than in Phase 2 ($p < .01$). These results demonstrate a progressive decrease in trust with repeated exposure to false warnings.

To assess self-confidence in performing the task without support, participants rated the statement *How confident are you that you can complete the task?* A Friedman test revealed no effect of condition on self-confidence ratings; no significant differences were observed among the FP, FN, and EW conditions. These results indicate that self-confidence remained comparable across the FP, FN, and EW conditions. In both the FN and FP conditions, Friedman tests revealed no significant effect of phase on confidence scores. This suggests that, unlike trust, participants' confidence in their ability to complete the task remained stable across the three phases.

In contrast, for system-provided confidence, we found a significant effect of condition, $\chi^2(2) = 12.96, p < .05$. Post hoc Wilcoxon signed-rank tests (Figure 8 (a)) indicated that confidence was significantly lower in the FP condition compared to the EW condition ($p < .01$). The NW condition was not included in this analysis, as no early-warning system was available in that condition.

We further examined whether system-provided confidence changed across task phases. In the FN condition, no significant pairwise differences were observed across phases (Figure 8 (b)). In the FP condition, system-provided confidence was lower in Phase 1 than in EW ($p < .05$) and remained lower in Phase 3 than in EW ($p < .01$; Figure 8 (b)).

Usability We compared SUS scores across the four conditions: nearly perfect early warnings (EW), False Positive (FP), False Negative (FN), and No-Warning (NW). The analysis revealed a significant effect of condition on SUS scores, $\chi^2(3) = 21.218, p < .001$. Post hoc Wilcoxon signed-rank tests (Figure 9(a)) showed that SUS scores were lower in FN (M = 56.30, grade D) than in EW (M = 76.06, grade B). FP received the lowest mean score (M = 43.02, grade F), significantly lower than both EW and FN. NW (M = 58.91, grade D) was also significantly lower than EW.

These results indicate that the FP impaired the perceived usability more than the FN compared to the EW condition.

Workload A Friedman test was conducted to analyze the individual effects of the conditions on the subscales of the NASA TLX. Table 1 summarizes the significance of each of the subscales.

Table 1. Friedman test results indicating the significance of differences between conditions for individual sub-scales of the NASA TLX shown in Figure 9.

	Significance
Mental Demand	**$\chi^2(3) = 13.920$, p = .003**
Physical Demand	**$\chi^2(3) = 9.813$, p = .020**
Temporal Demand	**$\chi^2(3) = 12.451$, p = .006**
Performance	**$\chi^2(3) = 9.244$, p = .026**
Effort	$\chi^2(3) = 6.337$, p = .096
Frustration	**$\chi^2(3) = 22.485$, $p < .001$**

Friedman test revealed a significant effect of condition on mental demand, physical demand, temporal demand, performance, and frustration (Table 1). Post hoc Wilcoxon signed-rank tests showed that the FN condition led to the highest mental demand compared to the EW condition and the NW condition, and also FP condition significantly increased the mental demand relative to the EW (see Figure 9 (b)). For physical demand, only NW exceeded EW significantly (Figure 9 (c)). Temporal demand was higher in FP and FN than in the EW (Figure 9 (d)). Perceived performance was significantly lower in FP and FN than in EW (Figure 9 (e)). Effort did not differ significantly across conditions (Figure 9 (f)). Finally, frustration was lowest in EW, while both FP and FN led to significantly higher frustration scores than EW (Figure 9 (g)).

We also examined the effects of the conditions of the Early Warning type on the overall combined NASA-TLX results. It showed a significant effect on the overall score ($\chi^2(3) = 14.929$, $p < .01$) with a mean score of 3.867 for the EW, 5.964 for the FN, 5.792 for the FP, and 5.229 for the NW as shown in Figure 9 (h).

Behavioral Reliance We analyzed the compliance rate, defined as the proportion of warnings followed by corrective movement away from the threshold region. A Friedman test revealed a significant effect of condition ($\chi^2(2) = 13.083$, $p < .001$). Post hoc Wilcoxon signed-rank tests with Bonferroni correction showed that compliance was lower in FN than in EW ($p < .005$), lower in FP than in EW ($p < .001$), and lower in FP than in FN ($p < .05$; Figure 10(a)). We then looked at how compliance evolved across the three phases of the task. In the FN condition, compliance was lower in Phase 1 and Phase 2 than in EW (both $p < .01$), but higher in Phase 3 than in EW ($p < .01$). Compliance was also higher in Phase 2 than in Phase 1 ($p < .05$), higher in Phase 3 than in Phase 1 ($p < .001$), and higher in Phase 3 than in Phase 2 ($p < .01$; Figure 10(b)). In the FP condition, compliance was lower than EW in Phase 1 ($p < .01$), Phase 2 ($p < .001$), and Phase 3 ($p < .001$). Compliance was also lower in Phase 2 than in Phase 1 ($p < .001$), lower in Phase 3 than in Phase 1 ($p < .001$), and lower in Phase 3 than in Phase 2 ($p < .05$; Figure 10(b)).

We also analyzed the number of low-tracking states for the hands (N-LTS) as an outcome measure of degraded hand-tracking quality. A Friedman test revealed a significant effect of condition, $\chi^2(3) = 8.065, p < .05$. Post hoc Wilcoxon signed-rank tests showed that FN

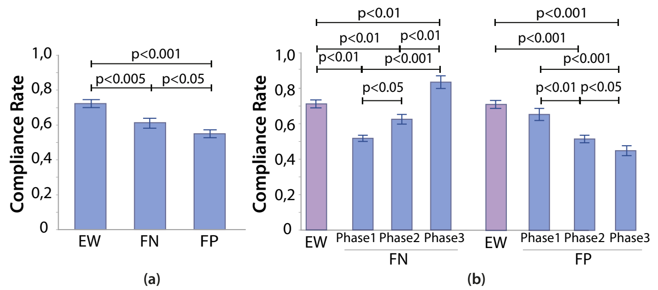

Fig. 10. Mean warning-compliance rates (a) across conditions and (b) across task phases for FN and FP. Brackets indicate significant Bonferroni-adjusted Wilcoxon signed-rank comparisons. Error bars show standard errors.

and NW led to significantly more low-tracking states compared to EW (both $p < .01$; Figure 11 (a)). This indicates that false negatives reduced reliance on the system by allowing more tracking failures to occur, while false positives did not significantly affect reliance relative to nearly perfect early warnings.

We further analyzed how the number of low-tracking states (N-LTS) evolved across the three phases of the task. In the FN condition, a Friedman test revealed a significant effect of phase ($\chi^2(2) = 12.532, p < .01$). Post hoc tests indicated that Phase 2 produced more low-tracking states than Phase 1 ($p < .01$), While Phase 3 produced more than Phase 1 ($p < .001$) and Phase 2 ($p < .05$). (Figure 11 (b)). Similarly, in the FP condition, a significant effect of phase was also found ($\chi^2(2) = 12.945, p < .01$). Post hoc tests indicated that Phase 2 produced more low-tracking states than Phase 1 ($p < .05$), While Phase 3 produced more than Phase 1 ($p < .01$) and Phase 2 ($p < .05$). These findings suggest that both false-positive and false-negative warnings contributed to an accumulation of low-tracking states as the task progressed, with FN producing the greatest overall number of losses.

Task Time Shapiro–Wilk tests indicated that task time did not violate normality assumptions (all $p > .05$); therefore, we conducted RM ANOVA. The analysis revealed a significant main effect of condition ($F(3,57) = 2.956, p < .05$). Post hoc comparisons showed that participants took significantly longer to complete the task in the NW condition than in any other condition (all $p < .05$). By contrast, no significant differences were observed between EW, FN, and FP, which all yielded comparable task times (Figure 11 (c)).

7 DISCUSSION

In this study, we used two VR pick-and-place tasks: a toy placement task and an assembly task, to evaluate how false positives and false negatives shape user trust, confidence, workload, and behavioral reliance in early warning systems. Each task consisted of 12 object placements and was designed to be comparable in complexity. Task completion times did not differ significantly between the two tasks, indicating that differences in user responses were more likely attributable to the early-warning error conditions than to task complexity.

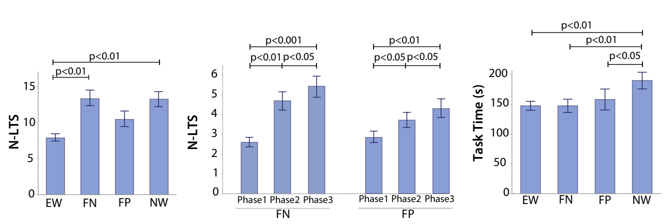

Fig. 11. Mean N-LTS (a) across conditions and (b) across task phases for FN and FP; and (c) mean task-completion time across EW, FN, FP, and NW. Brackets indicate significant pairwise comparisons. Error bars show standard errors.

Our study is among the first to systematically separate the effects of false positives and false negatives in VR early warning systems for hand tracking failures. While previous studies have emphasized the potential of early warning systems to mitigate failures and reduce task load [7, 22], they have treated warnings as uniformly beneficial without taking into account the possible errors and the distinct consequences of false and missed warnings.

In response to **RQ1**, we found that both false-positive and false-negative warnings decreased trust compared to nearly perfect early warnings, supporting **H1-a** hypothesis. Importantly, trust in the false-positive condition decreased already after the first phase of the task, with a total decrease of 48.7%, supporting **H1-b**. False negatives also reduced trust significantly after the second phase of the task, leading to a total decrease of 25.7% relative to the reliable early warning. False positives were likely more immediately noticeable as participants saw a warning without a corresponding failure. One plausible interpretation is that these mismatches acted as expectancy violations, accelerating distrust in a way consistent with the "cry-wolf" literature [10, 19, 44]. By contrast, false negatives may have been less immediately apparent, since their effect became visible only when a failure occurred without a warning. This may explain the slower but still significant decrease (only after phase 2) in trust under the false negatives.

Self-confidence (i.e., belief in one's ability to complete the task) did not vary significantly between the false-positive and false-negative conditions. This result suggests that participants attributed failures more to the system than to their own performance.

System usability ratings were significantly lower for both false positives and false negatives, with false positives leading to a significantly lower usability score compared to false negatives. Specifically, usability scores decreased by 43.42% in the false positives condition and by 25.97% in the false negatives condition relative to the reliable EW baseline, highlighting that false positives had the larger effect on perceived usability, which is in line with **H2a**. Workload results show that both false positives and false negatives increased the overall workload compared to the EW baseline. False negatives increased mental demand and temporal demand compared to both EW and NW conditions, while false positives increased frustration compared to EW and NW. Together, these findings indicate that false positives primarily affect the affective aspects of the user experience, while false negatives affect cognitive and performance-related aspects, in line with the **H2b** and previous research [16, 51]. Consistent with prior research, these findings reinforce that the "worse" error type depends on the task context and its risks [16, 51, 17, 4].

Behavioral measures provided further evidence of differentiated responses, addressing **RQ3**. In the FP condition, compliance decreased progressively across phases: it was already lower than EW in Phase 1 and declined further in Phases 2 and 3. This pattern is consistent with habituation to repeated false warnings, a well-known phenomenon in alarm-fatigue research [28], suggesting that participants became progressively less willing to respond to warnings they perceived as unreliable. The FN condition showed a different trajectory. Compliance was lower than EW during Phases 1 and 2, but increased across phases and exceeded EW by Phase 3. One possible interpretation is that repeated exposure to failures without warning made participants increasingly responsive to the warnings that remained available. Therefore, **H3a** was supported for FP but not for FN.

Both FP and FN showed increasing numbers of low-tracking states across phases. However, FN produced significantly more low-tracking states than EW overall, supporting **H3b**, whereas FP did not differ significantly from EW. Together, these results show that behavioral responses to warning unreliability evolve over time, but that FP and FN produce different patterns of compliance and tracking degradation.

Task completion time was only affected when warnings were absent. Thus, **H3c** was not supported: task times did not increase significantly under false positives or false negatives relative to EW. However, completion time captures only one aspect of performance. Our analysis of low-tracking states showed that false negatives led to more frequent tracking losses, reflecting degraded interaction quality even when task times were unaffected. One possible explanation for this divergence is that participants may have adjusted how they responded when warnings were unreliable, allowing them to maintain overall task pace despite degraded interaction quality. As a result, total completion time remained comparable across the EW, false-positive, and false-negative conditions.

In this paper, we replicated a previously proposed and evaluated visual early warning system by Gemici et al. [22]. In their work, the authors used three different hand tracking failures: low-light, self-occlusion, and out-of-vision conditions. We extended their experimental setup with external occlusion, which they highlight in their paper, but was not included in the evaluation. Our results on baseline condition, where we implemented the early warning system without false positives and false negatives, align with their findings: the early warning system decreased the task execution time, number of hand tracking failures, and cognitive load while increasing system usability compared to the no warnings condition. These results not only highlight the replicability of the early warning system findings [7] but also show that it can be extended to other hand tracking failure types.

Design Implications Our findings suggest three priorities for XR early-warning systems: (1) reduce repeated false positives when unnecessary corrections, frustration, and warning disengagement are costly; (2) reduce false negatives when missed support increases tracking degradation or cognitive and temporal demand; and (3) evaluate warning sensitivity, timing, and thresholds using trust, usability, workload, compliance, and tracking quality alongside detection accuracy. The appropriate FP–FN trade-off ultimately depends on the task context and the consequences associated with each error type.

A related implication is how early-warning concepts might extend to existing safety features in commercial VR systems, such as the Guardian boundary [75]. Current headsets often switch suddenly to passthrough view once users' hands or heads cross the boundary, which ensures safety but can feel disruptive. Our task-time results suggest that early warnings may support task continuity even when warning reliability is imperfect; however, their effects on trust, usability, and tracking quality depend on reliability. This suggests that early cues, such as fading outlines near the hands, could be introduced before the Guardian takeover to better preserve immersion.

The differentiated effects of false-positive and false-negative warnings may also extend to other VR input modalities, such as controllers, haptic gloves, and styluses. Each device provides a different benefit, including responsiveness and precision for controllers [59, 3], natural embodiment for gloves [43, 56, 27], and fine-grained precision for styluses [76, 59, 35]. In these contexts, false positives may interrupt the qualities that make each modality useful, while false negatives may allow more disruptive errors to occur without preparation. Designers must therefore decide which error type is more acceptable depending on whether the context prioritizes flow and embodiment or precision.

Our findings also contribute to broader HCI research on trust in automation, alarm design, and human–AI interaction. Prior work in domains, such as diagnostic decision aids [16] and aviation alarm systems [71, 10], has shown that false alarms can reduce trust more strongly than missed alarms, a pattern we also observed in VR hand-tracking early warnings. This suggests that design principles from safety-critical domains may also be relevant to XR systems. Our results indicate that although users benefit from warning assistance, unreliable signals can alter how users respond to and rely on that assistance over time. By positioning VR early-warning systems within these broader discussions, our work connects XR interaction design with HCI research on trust, transparency, and adaptive support.

In augmented reality, where users see their real hands alongside digital overlays, false negatives may be particularly disruptive. If a hand overlay freezes or disappears without warning, but the user still sees their physical hand moving, the mismatch is highly noticeable. This creates a strong perceptual conflict between vision (real hand still present) and system feedback (virtual hand gone), which may diminish both trust and the sense of embodiment.

8 LIMITATIONS AND FUTURE WORK

Our study examined participants' immediate responses during a single session and employed a fixed 50% error rate for both false positives and false negatives. This rate was selected through pilot testing to create a controlled and noticeable reliability manipulation while preserving comparable task mechanics across conditions. Accordingly, our findings characterize user responses at this specific reliability level and may not generalize to lower, higher, or adaptive error rates, or to repeated and longitudinal exposure.

Since participants performed only one session, our study focuses on the immediate effects on trust and confidence during a single exposure. Trust recovery is known to be a slower longitudinal process [34], and future research should investigate how it develops across repeated exposures and explore design interventions that could support trust recalibration in applications where errors are unavoidable. For example, such interventions could include transparent status messaging (e.g., notifying users when the system's reliability has improved after a period of instability), adaptive reliability indicators (e.g., a visual or haptic feedback that communicates current confidence in the warning system), and phased restoration of accuracy (e.g., gradually reintroducing reliable warnings after errors rather than switching suddenly), all of which may help users rebuild trust in the system.

Our false positive and false negative rates were designed based on pilot testing, and we reduced the warnings or provided additional warnings by approximately 50% compared to the baseline. While this provided strong control, it does not fully reflect the variability of real hand tracking failures. Future studies should run in-wild studies and examine whether context-aware strategies (e.g., tighter thresholds in high precision actions and looser thresholds during simpler actions) yield better trade-offs between false positive and false negative warnings.

While our evaluation was based on two matched pick-and-place tasks of moderate complexity, which allowed us to control for task time and ensure comparability, this limits generalizability to other forms of VR interactions. Longer, more demanding tasks, such as fine assembly or training simulations, may produce different tolerance thresholds for false positives and false negatives. Future work should therefore extend the study to task contexts where the costs of missed warnings versus false warnings vary more dramatically.

Our study adopted the visual warning design proposed in prior work [22]. Future work should examine alternative and multimodal feedback designs, including haptic and auditory cues, and investigate how feedback modality affects trust and confidence in the system. Because the low-light scenario involved changes to the physical room lighting, participants may have perceived these changes through the gap between the headset and face, potentially influencing their interpretation of the warnings.

9 CONCLUSION

Early-warning systems have been proposed to mitigate hand-tracking failures in VR, yet their effectiveness depends on reliability. In this work, we examined how false positives and false negatives affect user trust, confidence, workload, and behavioral reliance when interacting with such systems. Within the tested single-session pick-and-place setting and controlled 50% error rate, our results demonstrate that false positives produce an immediate decline in trust and lead users to disengage from warnings over time, whereas false negatives reduce trust more gradually but increase mental and temporal demand as users attempt to compensate for missed alerts. Comparisons with NW were measure-specific: warning conditions reduced task time, whereas FP produced the lowest mean usability score. By identifying these effects, our work provides guidance for managing warning reliability and informs the design of dependable early-warning systems that extend beyond VR hand tracking to broader applications, such as, but not limited to, safety boundaries, controller and glove-based interactions, and augmented reality.

REFERENCES

[1] D. Abdlkarim, M. Di Luca, P. Aves, M. Maaroufi, S.-H. Yeo, R. C. Miall, P. Holland, and J. M. Galea. A methodological framework to assess the accuracy of virtual reality hand-tracking systems: A case study with the meta quest 2. *Behavior research methods*, 56(2):1052–1063, 2024. 2

[2] A. Ahmad, C. Migniot, and A. Dipanda. Tracking hands in interaction with objects: A review. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 360–369, 2017. doi: 10.1109/SITIS.2017.66 1

[3] M. Allgaier, V. Chheang, P. Saalfeld, V. Apilla, T. Huber, F. Huettl, B. Neyazi, I. E. Sandalcioglu, C. Hansen, B. Preim, et al. A comparison of input devices for precise interaction tasks in VR-based surgical planning and training. *Computers in Biology and Medicine*, 145:105429, 2022. 9

[4] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert, and D. M. Tilbury. Comparing the effects of false alarms and misses on humans' trust in (semi)autonomous vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, p. 113115. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3371382.3378371 3, 8

[5] D. Bang and S. M. Fleming. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 115(23):6082–6087, 2018. 2

[6] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. 2

[7] M. R. Bashar and A. U. Batmaz. An Early Warning System Based on Visual Feedback for Light-Based Hand Tracking Failures in VR Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3645–3655, 2025. 1, 2, 5, 8, 9

[8] M. Beggiato and J. F. Krems. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour*, 18:47–57, 2013. 2

[9] J. Bliss, M. Dunn, and B. S. Fuller. Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates '. Technical report, 1995. 1

[10] J. P. Bliss and M. C. Dunn. Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, 43:1283–1300, 9 2000. 1, 3, 8, 9

[11] A. Boldt and N. Yeung. Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35(8):3478–3484, Feb 2015. doi: 10.1523/JNEUROSCI.0797-14.2015 2

[12] W. Chen, C. Yu, C. Tu, Z. Lyu, J. Tang, S. Ou, Y. Fu, and Z. Xue. A survey on hand pose estimation with wearable sensors and computer-vision-based methods. *Sensors*, 20(4):1074, 2020. 2

[13] A. Colaço, A. Kirmani, H. S. Yang, N.-W. Gong, C. Schmandt, and V. K. Goyal. Mime: Compact, low power 3D gesture sensing for interaction with head mounted displays. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 2

[14] M. Cvach. Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4):268–77, Jul 2012. Copyright - Copyright Allen Press Publishing Services Jul/Aug 2012; Document feature - ; Last updated - 2023-11-21. 3

[15] P. De Vries, C. Midden, and D. Bouwhuis. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 58(6):719–735, 2003. 5

[16] S. R. Dixon, C. D. Wickens, and J. S. McCarley. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human factors*, 49(4):564–572, 2007. 2, 3, 8, 9

[17] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718, 2003. 2, 3, 8

[18] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007. Special Issue on Vision for Human-Computer Interaction. doi: 10.1016/j.cviu.2006.10.012 1, 2

[19] K. K. H. et al. Making healthcare safer III: A critical analysis of existing and emerging patient safety practices [internet]. 2020. 3, 8

[20] Y. Ferstl, R. McDonnell, and M. Neff. Evaluating study design and strategies for mitigating the impact of hand tracking loss. In *ACM Symposium on Applied Perception 2021*, pp. 1–12, 2021. 1, 2

[21] M. Gemici, A. Hatira, V. Phadnis, and A. U. Batmaz. Gaze analysis in early warning visual feedback system for hand tracking failures in virtual reality. *IEEE Access*, 13:188481–188493, 2025. doi: 10.1109/ACCESS.2025.3621645 2

[22] M. Gemici, V. Phadnis, and A. U. Batmaz. Before hands disappear: Effect of early warning visual feedback method for hand tracking failures in virtual reality. *PLOS ONE*, 20(6):1–28, 06 2025. 1, 2, 4, 5, 8, 9

[23] P. Grimaldi, H. Lau, and M. A. Basso. There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience and Biobehavioral Reviews*, 55:88–97, Aug 2015. 2

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks, 2017. 3

[25] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 6

[26] N. Hatira, A. Aliza, A. U. Batmaz, and M. Sarac. Enhancing eye-hand coordination in volleyball players: A comparative analysis of VR, AR, and 2D display technologies and task instructions. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 219–228, 2024. doi: 10.1109/ISMAR62088.2024.00036 2

[27] N. Hatira and M. Sarac. Fine motor tasks in virtual reality: The impact of haptic feedback and object characterization. *IEEE Access*, 13:108348–108358, 2025. doi: 10.1109/ACCESS.2025.3581787 9

[28] E. Hellier and D. L. S. P. HealthCare. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. 2010. 8

[29] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. 3

[30] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570 2

[31] S. Islam, B. Ionescu, C. Gadea, and D. Ionescu. Full-body tracking using a sensor array system and laser-based sweeps. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 71–80. IEEE, 2016. 2

[32] J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000. 6

[33] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(2):126–140, 2012. 3

[34] P. K. Kahr, G. Rooks, C. Snijders, and M. C. Willemsen. The trust recovery journey. the effect of timing of errors on the willingness to follow ai advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 609–622, 2024. 9

[35] F. Kern, P. Kullmann, E. Ganal, K. Korwisi, R. Stingl, F. Niebling, and M. E. Latoschik. Off-the-shelf stylus: Using XR devices for handwriting and sketching on physically aligned virtual surfaces. *Frontiers in Virtual Reality*, 2:684498, 2021. 9

[36] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, Volume 12 - 2021, 2021. 2

[37] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017. 1

[38] J. LeClerc and S. Joslyn. The cry wolf effect and weather-related decision making. *Risk Analysis*, 35(3):385–395, 2015. 3

[39] C.-J. Lee, R. Zhang, D. Agarwal, T. C. Yu, V. Gunda, O. Lopez, J. Kim, S. Yin, B. Dong, K. Li, M. Sakashita, F. Guimbretiere, and C. Zhang. Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642910 1, 2

[40] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. 6

[41] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. 2, 3

[42] J. Li, Y. Yang, Q. V. Liao, J. Zhang, and Y.-C. Lee. As confidence aligns: Understanding the effect of ai confidence on human self-confidence in human-ai decision making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. 2, 3

[43] Z. Li, M. Bujić, O. . Buruk, E. Bampouni, S. Järvelä, and J. Hamari. Haptics-mediated virtual embodiment: Impact of a wearable avatar-controlling system with kinesthetic gloves on embodiment in VR. *Frontiers in Virtual Reality*, 5:1439724, 2024. 9

[44] J. R. Lim, B. F. Liu, and M. Egnoto. Cry wolf effect? evaluating the impact of false alarms on public responses to tornado alerts in the southeastern united states. *Weather, Climate, and Society*, 11(3):549 – 563, 2019. 3, 8

[45] M. K. Lindell and R. W. Perry. The protective action decision model: theoretical modifications and additional evidence. *Risk Anal.*, 32(4):616–632, Apr. 2012. 3

[46] G. Lopez-Garrido. Banduras self-efficacy theory of motivation in psychology. *Simply Psychology*, 10(1), 2023. 2

[47] P. Madhavan and D. A. Wiegmann. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49(5):773–785, 2007. PMID: 17915596. 3

[48] P. Madhavan, D. A. Wiegmann, and F. C. Lacson. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48(2):241–256, 2006. PMID: 16884046. 3

[49] D. Manzey, N. Grard, and R. Wiczorek. Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, 57:1833–1855, 12 2014. 1

[50] R. McGloin, K. Farrar, and M. Krcmar. Video games, immersion, and cognitive aggression: Does the controller matter? *Media psychology*, 16(1):65–87, 2013. 1, 2

[51] J. Meyer, R. Wiczorek, and T. Gnzler. Measures of reliance and compliance in aided visual scanning. *Human Factors*, 56(5):840–855, 2014. doi: 10.1177/0018720813512865 3, 8

[52] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3D hand tracking from monocular RGB. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[53] K. Nakayachi, R. Yokoi, and J. Goltz. Human behavioral response to earthquake early warnings (EEW): Are alerts received on mobile phones inhibiting protective actions? *International Journal of Disaster Risk Reduction*, 105:104401, 2024. 5

[54] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pp. 2088–2095, 2011. 1, 2

[55] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. 3

[56] R. Palombo, S. Weber, M. Wyszynski, and B. Niehaves. Glove versus controller: the effect of VR gloves and controllers on presence, embodiment, and cognitive absorption. *Frontiers in Virtual Reality*, 5:1337959, 2024. 9

[57] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997. 2

[58] S. Potter, S. Harrison, and P. Kreft. The benefits and challenges of implementing impact-based severe weather warning systems: Perspectives of weather, flood, and emergency management personnel. *Weather, Climate, and Society*, 13(2):303 – 314, 2021. 3

[59] H.-R. Rantamaa, J. Kangas, S. K. Kumar, H. Mehtonen, J. Järnstedt, and R. Raisamo. Comparison of a vr stylus with a controller, hand tracking, and a mouse for object manipulation and medical marking tasks in virtual reality. *Applied Sciences*, 13(4):2251, 2023. 9

[60] B. Renzl. Trust in management and knowledge sharing: The mediating effects of fear and knowledge documentation. *Omega*, 36(2):206–220, 2008. 2

[61] A.-A. Sadiq, R. B. Dougherty, J. Tyler, and R. Entress. Public alert and warning system literature review in the USA: identifying research gaps and lessons for practice. *Natural Hazards*, 117(2):1711–1744, 2023. 3

[62] J. Sauro. 5 ways to interpret a SUS Score MeasuringU. 6

[63] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 36333642. New York, NY, USA, 2015. 2

[64] R. Singh, M. Ziat, O. Špakov, J. Mäkelä, V. Surakka, and R. Raisamo. Trust and visual focus in automated vehicles: A comparative study of beginner and experienced drivers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. 2

[65] M. 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. 3

[66] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[67] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[68] R. J. Teather, A. Pavlovych, W. Stuerzlinger, and I. S. MacKenzie. Effects of tracking technology, latency, and spatial jitter on object movement. In *2009 IEEE symposium on 3D user interfaces*, pp. 43–50. IEEE, 2009. 1

[69] K. Weibull, B. Lidestam, and E. Prytz. False alarm effects in early warnings for emergency vehicles: exploring drivers move-over behavior. *Human factors*, 66(11):2518–2527, 2024. 3, 5

[70] C. D. Wickens. Situation awareness and workload in aviation. *Current Directions in Psychological Science*, 11(4):128–133, 2002. 3

[71] C. D. Wickens, S. Rice, D. Keller, S. D. Hutchins, J. Hughes, and K. Clayton. False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors*, 51(4):446–462. 3, 9

[72] R. Wiczorek and J. Meyer. Effects of trust, self-confidence, and feedback on the use of decision automation. *Frontiers in Psychology*, Volume 10 - 2019, 2019. 2, 5

[73] M. Wischnewski, N. Krämer, and E. Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. 2, 3

[74] M. Woo and O. Bacon. Alarm fatigue. In K. Hall, S. Shoemaker-Hunt, L. Hoffman, et al., eds., *Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices*. Agency for Healthcare Research and Quality (US), Rockville, MD, USA, Mar 2020. 3

[75] S. Wu, J. Li, M. Sousa, and T. Grossman. Investigating guardian awareness techniques to promote safety in virtual reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 631–640, 2023. 9

[76] Z. Wu. Eyes on target, pen on table : Gaze and pen input for comfortable and precise vr interaction. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2025. 9

[77] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 112. Association for Computing Machinery, New York, NY, USA, 2019. 2

[78] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pp. 307–317, 2017. 2

[79] C. Zimmermann and T. Brox. Learning to estimate 3D hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4913–4921, 2017. 1, 2