Contextual Recovery: Guiding Hand Tracking Failures Recovery in Mixed Reality via VLM Reasoning

Yi ZOU* Ziming LI† Hai-Ning LIANG‡ Zhiming HU§

The Hong Kong University of Science and Technology (Guangzhou)

2026 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3195-0529-3/26/$31.00 ©2026 IEEE | DOI: 10.1109/VRW70859.2026.00072

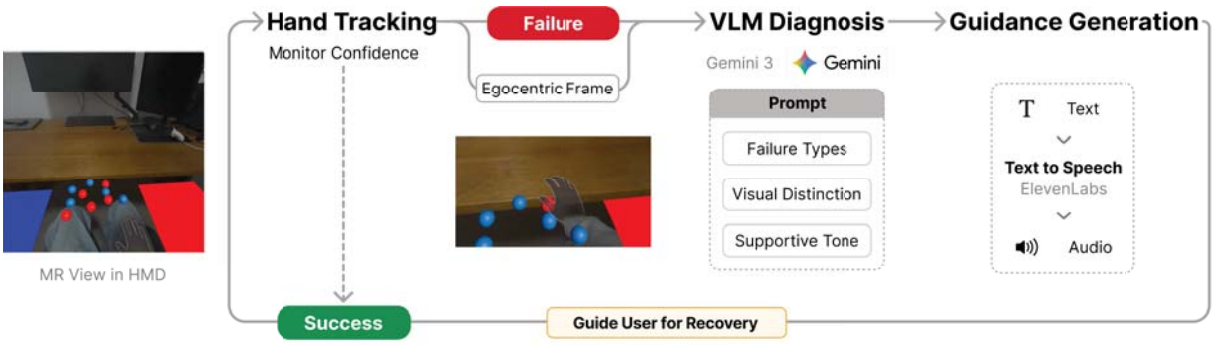

Figure 1: **System Pipeline**. The system operates on a "Monitor-Diagnose-Guide" cycle. It continuously monitors tracking confidence to detect failures. Upon detection, a Vision Language Model analyzes the user's egocentric view to diagnose specific causes to generate actionable audio guidance, helping users to restore tracking integrity.

Abstract

Robust hand tracking is essential to natural interaction in Mixed Reality (MR), yet it remains highly susceptible due to the limitations of tracking systems in Head-Mounted Displays (HMDs). Typical inside-out tracking systems fail in various context conditions, including occlusion from interaction and low light condition. Traditional methods for handling failure are often hard-coded for specific failure types using deterministic algorithms, limiting their ability to handle complex real-world settings. While recent research has introduced "early warning" systems to visualize tracking confidence and alert users to potential failures, these approaches are fundamentally passive and do not explain why or how to rectify the issue. In this paper, we introduce a context-aware recovery framework that leverages the semantic reasoning capabilities of Vision-Language Models (VLMs) to interpret various types of tracking failure and generate actionable guidance for recovery. The pipeline monitors tracking confidence to trigger VLM-based diagnosis, analyzing the user's egocentric view to identify the cause and generate actionable guidance. Crucially, this system generalizes across frequent failure modes through structured contextual prompting that grounds the VLM's reasoning with visual features of physical environment. Results from user study show high perceived utility and a strong subjective preference for the guidance compared to baseline method, indicating the system's potential to improve task proficiency.

Index Terms: Hand tracking, mixed reality, error recovery.

1 Introduction

Hand tracking has emerged as a primary interaction paradigm for modern Mixed Reality (MR) Head-Mounted-Displays (HMDs). Commercial devices like Meta Quest Series[1], Apple Vision Pro[2] and Microsoft Hololens 2[3] adopt integrated inside-out tracking system working with specific camera settings to enable a smooth experience. However, unlike controller-based systems which rely on Inertial Measurement Unit (IMU) data, egocentric hand tracking methods primarily rely on optical sensing. Hand tracking is fragile in the presence of occlusion, low light condition, in-and-out of view and motion blur [15, 9].

When these factors compromise tracking integrity, current devices typically fail silently or exhibit erratic virtual hand behavior, breaking immersion and causing user frustration [4]. To mitigate this, recent research has explored "Early Warning" systems in both Augmented Reality (AR) and Virtual Reality (VR) environment [17, 19, 2]. These systems visualize internal tracking confidence or the effective tracking boundaries to alert users of potential failures. The system improved predictability of system performance by displaying a virtual sphere that changes color to provide visual feedback, therefore users were able to anticipate impending hand-tracking failures with reduced frustration. While the design improved system usability, the visual warning remains passive, offering no further assistance for users to recover from failures. Moreover, the system was limited to light-based tracking failures, which excluded diverse scenarios. A user alerted may not intuitively understand the exact reason of hand tracking failure in complex settings, whether their environment is too dark or the sensors are occluded from seeing the hands. Moving to MR, the physical environment blends with virtual elements, making this distinction even more difficult [8]. This occlusion may affect the user's depth perception and scene understanding, thereby interfering with their situational assessment [1]. This ambiguity shifts the burden of diagnosis and recovery entirely onto the user, who typically lacks the necessary understanding of the sensor's limitations to resolve the issue effectively.

In this paper, we introduce a context-aware framework that ac-

*e-mail: yzou906@connect.hkust-gz.edu.cn
†e-mail: zli578@connect.hkust-gz.edu.cn
‡Corresponding author. e-mail: hainingliang@hkust-gz.edu.cn
§Corresponding author. e-mail: zhiminghu@hkust-gz.edu.cn

[1]https://www.meta.com/sg/quest/quest-3/
[2]https://www.apple.com/apple-vision-pro/
[3]https://learn.microsoft.com/en-us/hololens/

979-8-3195-0529-3/26/$31.00 ©2026 IEEE
DOI 10.1109/VRW70859.2026.00072

tively provides guidance tailored to different hand tracking failures. The system leverages Vision Language Models (VLMs) to interpret complex failure scenes to identify the specific problem of tracking. This semantic diagnosis is instantly translated into synthesized voice guidance, providing users with specific, actionable advice to restore tracking integrity without disrupting the immersive experience. The results demonstrate a strong subjective preference for the proposed method, showing comparable Perceived Utility and higher Perceived Ease of Use.

The contributions of this paper are as follows:

- We developed a context-aware system for guiding hand tracking failure recovery. The pipeline leverages the semantic reasoning capabilities of VLM to analyze egocentric visual cues to identify failure causes (occlusion, low light, out-of-view) and generating actionable, synthesized voice advice.
- We conducted a user study comparing the proposed VLM-based audio guidance against a visual warning baseline on three frequently seen types of hand tracking failure. Qualitative feedback highlights the preference for guidance system by clarifying the reasons for failure and provide guidance to reduce the cognitive burden.

2 Related Works

2.1 Hand Tracking in MR

Markerless hand tracking has developed as a main paradigm for interaction in consumer MR [12]. Han et al. [5] established a foundation for egocentric tracking by introducing a detection-by-tracking pipeline that tracks 3D hand motion across a large working volume in real-time using monochrome image inputs. Building on this foundation, Han et al. [6] later introduced UmeTrack, a unified, end-to-end differentiable framework that eliminates multi-stage optimization by directly predicting world-space 3D hand poses from multi-view inputs.

While optical solutions provide high-fidelity tracking within the camera's field of view, maintaining tracking continuity when hands are not clearly presented remains a critical challenge. Inferring 3D hand pose in these unobserved regions is an inherently ill-posed problem, as the lack of visual data creates infinite theoretical possibilities for the hand's location [7]. To resolve this ambiguity, Ye et al. [7] used the spatiotemporal correlation between body kinematics and hand motion to forecast 3D poses even when hands are entirely invisible to the headset. Kim et al. [16] proposed HOOV to utilize wrist-mounted Inertial Measurement Units (IMUs) as a robust alternative to optical tracking. By fusing continuous inertial data with intermittent visual cues, it enables precise proprioceptive interaction outside the camera's field of view. Despite these advancements, the ill-posed nature of occlusion implies that tracking continuity cannot be fully guaranteed across all scenarios. Consequently, hand tracking failures remain inevitable in specific circumstances, necessitating the integration of explicit feedback systems to manage user expectations and facilitate recovery. Rather than optimizing algorithms to predict hand poses in ill-posed settings, this work introduces a feedback system that guides the user toward well-suited conditions where reliable tracking can be restored.

2.2 Hand Tracking Failures and Feedback

Major reasons of hand tracking failure include complex occlusion, low light condition, out-of-view and motion blur [2, 9]. Tracking fidelity deteriorates during complex user behaviors, like rapid hand motions and occlusion during interactions. To mitigate the impact of these inherent limitations, recent research has explored feedback systems that maintain usability during tracking degradation. Bashar et al. [2] developed an early warning system specifically for light-based tracking failures, providing visual feedback to alert users before tracking completely deteriorates. This type of preemptive notification is particularly valuable as it empowers users to proactively adjust their environment, thereby maintaining the flow of interaction and reducing frustration. Xu et al. [19] investigated visual techniques for boundary awareness, utilizing peripheral cues to guide users when their hands approach the edges of the effective tracking volume.

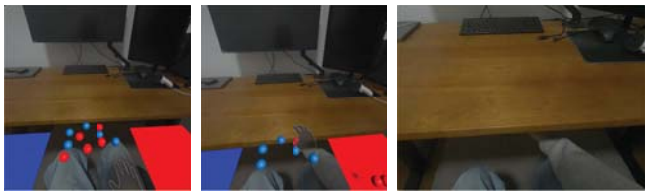

Figure 2: The egocentric view of user conducting the ball sorting task. a) The initial MR setting of the virtual task scene in a workspace. b) The moment of occlusion hand tracking failure occurs. c) The raw camera image from front camera on HMD, as input for VLM reasoning.

While warning systems are instrumental in signaling the onset of failure, they are limited by their passive nature. They can only deliver abstract hints for failure, but not able to guide the users to recover. Moreover, the systems typically focus on one specific kind of tracking failure, with which the form of hint is highly associated. This limitation underscores the necessity for an *active guidance* system that goes beyond simple alerts to explicitly direct user behavior, ensuring effective prevention and recovery from tracking loss. To address these gaps, we propose a system that can handle various types of hand tracking failure and provide corresponding guidance to help user recovery.

2.3 Vision Language Models in Mixed Reality

With their rapid development, VLMs now excel at processing multimodal input, creating a seamless connection between egocentric visual feeds and semantic understanding. This allows integrations with MR systems to act as intelligent assistants capable of reasoning about complex social and spatial environments [11]. Recent research has largely leveraged this capability to optimize interaction and content placement. Li et al. [10] demonstrated how VLMs can analyze scene context and social setting to optimize UI placement. Similarly, Pei et al. [14] utilized a multimodal LLM to process user attention and environmental awareness for safety in locomotion. Furthermore, Oh et al. [13] demonstrated that LLM-based agents that actively incorporate users' spatial context during counselling interactions can enhance copresence and trust. Collectively, these works highlight the VLM's potential to serve as a semantic interpreter, expanding the system's ability to reason about the user's environment beyond simple geometry.

While previous studies utilize VLMs to enhance application-level experiences, they assume working with a perfect functioning MR system. The VLMs have not yet been utilized to diagnose and rectify low-level system failures, such as sensing limitations. In this work, we bridge this gap by integrating a VLM as a diagnostic agent. The system uses context from egocentric images to analyze causes of failure and drive users to restore system resilience.

3 Method

The proposed guidance system identifies hand-tracking failures based on egocentric inputs in MR view.

3.1 Architecture Pipeline

The proposed system operates on a "Monitor-Diagnose-Guide" cycle incorporating three modules: Hand Tracking, VLM Diagnosis and Guidance Generation (see Fig. 1). The pipeline monitors the built-in confidence value of hand tracking to determine actions. The detection of tracking failure will trigger the diagnosis process by retrieving the egocentric image from HMD as input context. This visual input is queried against a Vision-Language Model (VLM) utilizing a specialized context prompt. The VLM acts as a diagnostic agent, parsing the scene to generate a structured output that comprises the identified failure category and a synthesized recovery guidance. Finally, the generated text guidance engine is converted into audio via a Text-to-Speech (TTS) engine, playing through the HMD to remind user and help with recovery.

To ensure the reliability and smooth operation, a sliding window trigger is implemented to continuously monitor the native hand-tracking API's confidence. The system remains idle during standard operation to optimize computational resources. When the duration of failure accumulates to the threshold (1s) with the sliding window (3s), the diagnostic action is triggered. This filter serves to distinguish sustained environmental failures from transient tracking glitches caused by rapid motion. Upon activation, the system captures the user's current egocentric RGB frame.

3.2 Contextual Guidance

To generalize across different failures, a VLM agent is prompted with structured knowledge of hand tracking and potential causes of failures. Rather than treating tracking loss as a monolithic state, we enforce a clear taxonomy ($F1$–$F3$) that maps specific visual evidence of environmental interference directly to distinct recovery protocols. The VLM is prompted to classify the current state based exclusively on visual evidence from the egocentric camera:

- **Out-of-View Failure ($F1$):** This state addresses failures where hands fall outside the camera frustum or are heavily truncated due to head pose, assuming the lighting conditions are otherwise adequate. The VLM distinguishes this from environmental factors, generating guidance that directs the user to reposition their hands centrally within the sensor's field of view.

- **Occlusion Failure ($F2$):** This category encompasses physical obstructions where key hand landmarks are blocked by external objects, surfaces, or instances of self-occlusion (e.g., crossed hands). The system identifies the specific occluding element and advises the user to separate their hands or remove the blocking object to restore line-of-sight.

- **Low Light Condition ($F3$):** This identifies scenarios where tracking is compromised by insufficient illumination, heavy shadowing, or high sensor noise, distinct from positioning errors. Unlike simple sensor thresholds, the VLM visually confirms that hands are indistinguishable due to darkness and suggests environmental adjustments, such as activating a light source or relocating to a brighter area.

Beyond classification, the system is engineered to prioritize user experience through tone regulation. The VLM is explicitly prompted to generate guidance using warm, supportive language rather than technical error codes, ensuring that the recovery process maintains immersion and reduces user frustration.

4 Evaluation

We conducted a preliminary perception study to validate the user acceptance between warning and guidance.

4.1 Experiment Design

We developed a standardized ball-sorting task to simulate real world scenes which require spatial understanding and accurate manipulation [2] (Fig. 2). The scene depicts a canonical pick-and-place task chosen for its requirement of varied hand poses and spatial reach. The user is positioned centrally with a collection of scattered balls in two different colors. Two target zones (planes colored in red and blue) are situated distally on the left and right peripheries. The user needs to complete the task by sorting all the balls to their corresponding plane according to the color. To successfully sort a ball, the user must reach out fully to the side, a motion trajectory specifically designed to test the limits of the camera's field of view and invite potential occlusion.

Given the technical complexity of replicating specific environmental failures (e.g., exact lighting conditions) consistently across participants, we employed a vignette methodology using a group of pre-recorded videos. This guarantees that all users respond to the exact same visual stimuli for tracking failure and environmental contexts. We recorded a series of first-person perspective (POV) videos across different conditions in the scene. To ensure consistency, a researcher performed a scripted workflow where the movement speed and trajectory were kept highly constant. We artificially induced the three failure types defined in Sec. 3.2 (F1: Out of View, F2: Occlusion, F3: Low Light) during the task execution. For each failure instance, we generated two experimental conditions:

- **Baseline (Visual Warning):** A red visual indicator appears to signal low tracking confidence, mimicking the state-of-the-art method [2].

- **Proposed (Audio Guidance):** The system provides the synthesized voice advice generated by our pipeline.

We used Meta Quest 3 as HMD for running experiment scene. We selected Gemini 3 for reasoning and ElevenLabs [4] for TTS to generate audio guidance.

4.2 Participants

We recruited 6 participants (5 male, 1 female) between 22 and 26 years old from the local university community. One of them had no prior experience with AR or VR devices, while the rest had varying levels of experience ranging from occasional use to regular daily use.

4.3 Procedure and Metrics

Participants began with a brief introduction to the ball-sorting task context. They then viewed paired video vignettes for each of the three failure types in a randomized order to prevent order effects. Participants were instructed to adopt the perspective of the user in the video. Since participants were observers, we adapted a modified Technology Acceptance Model (TAM) [3] to focus on *Perceived Utility (PU)* and *Perceived Ease of Use (PEU)*. Participants rated the feedback mechanisms on a 7-point Likert scale. Finally, participants selected their preferred method for each failure scenario and provided open-ended justification.

5 Results

We analyzed the subjective ratings to compare the communicative efficacy of the proposed guidance against the visual baseline. Given the pilot sample size ($N = 6$), we report descriptive statistics (Mean M, Standard Deviation SD) and qualitative themes. We evaluated the score of Perceived Utility (PU) and Perceived Ease of Use (PEU) for both feedback modalities (shown in Tab. 1). While the Perceived Utility scores were comparable between the visual

[4] https://elevenlabs.io/

Table 1: Mean (M) and Standard Deviation (SD) of TAM scores.

TAM Metric	Visual Warning		VLM Guidance	
	M	SD	M	SD
Perceived Utility	5.11	0.92	5.17	1.16
Perceived Ease of Use	5.84	0.91	**6.06**	0.86

baseline ($M = 5.11$) and our VLM-based guidance ($M = 5.17$), the VLM condition achieved a higher Perceived Ease of Use ($M = 6.06$) than the baseline ($M = 5.84$). Subjective interview showed a strong preference of audio feedback (shown in Fig. 3)

6 Discussion

Our results highlight a shift in user needs from failure detection to detailed diagnosis. Qualitative feedback suggests this preference stems from the *explanatory nature* of the guidance. While the visual warning baseline provides the timely feedback of tracking status, the proposed VLM system translates abstract failure states into concrete, human-understandable causes from user's physical context. Participants noted that even when specific recovery instructions were imperfect, the provision of an explicit diagnosis (e.g., "it is too dark") significantly reduced the cognitive load of troubleshooting, allowing them to maintain focus on the task rather than the device. A major limitation of this work is that we only use mock videos instead of letting users experience the actual system. The experiments are expected to be further improved and to enhance the analysis with statistical tests in future work.

The choice of audio as the primary feedback channel involves a trade-off between robustness and intrusiveness. We prioritized audio because visual tracking failures often correlate with rendering instability (e.g., jittering or freezing), which can render world-locked text UIs unreadable [18]. Audio ensures the recovery guidance is received regardless of the user's head pose and without visual clutter. However, participants noted that voice commands can be intrusive during high-concentration tasks. With that said, we recognize that an audio-only approach limits accessibility (e.g., for users with hearing impairments) and applicability in noisy or social environments. A potential middle ground suggested by users is a hybrid approach: utilizing timely visual cues for minor warnings, and escalating to audio guidance when sustained failures occur. An avenue for future research is exploring feedback delivery through various single or combined modalities.

The timing of guidance represents a balance between system stability and user responsiveness. We prioritize sustained tracking failures over momentary glitches by implementing a sliding window. While this system incurs a latency of approximately 4–5 seconds, users found this delay acceptable given the pacing of recovery actions. Notably, we observed that the initial diagnosis alone often triggered immediate self-correction. This indicates that users proactively leverage the semantic diagnosis to re-evaluate the relationship between the physical and virtual context, occasionally rendering the subsequent detailed guidance redundant. Building on this, future work will focus on integrating task phase estimation and user action analysis to adapt the guidance dynamically, ensuring it complements rather than interrupts the user's correction process.

The guidance to recover is currently constrained by the VLM's limited visibility of the full MR context. The VLM is currently not provided with the context in MR view, analyzing only the RGB passthrough feed. This creates a risk of contextually inappropriate advice where the physical reality conflicts with virtual intent. For example, if a user reaches towards a *virtual* object placed intentionally under a physical desk, the VLM might interpret this as an occlusion risk and advise the user to move hands above it, directly contradicting the application's goal. Future works will focus on more precise modeling of user state and environmental context.

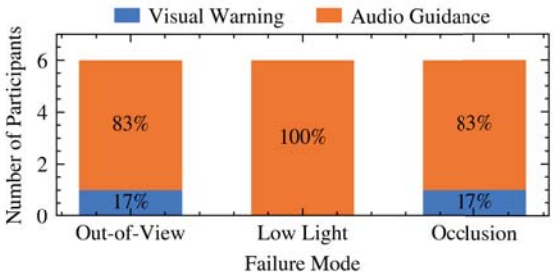

Figure 3: **User preference for Method.** Preference shifts significantly toward Audio Guidance (Orange).

7 Conclusion

In this paper, we introduced a context-aware framework for resilient hand tracking in MR. We conducted a preliminary user study to evaluate the effectiveness of proposed guidance system. The results support better utility of the guidance system for translating user's context into clear diagnosis and providing actionable guidance.

A VLM Prompt

We utilized the following prompt to establish the context for the VLM. This example focuses on setting the context for hand tracking failures and kind guidance generation.

"You will mimic a Mixed Reality system diagnostic observer embedded inside a head-mounted display. You see the world from the user's own first-person viewpoint through the headset's egocentric camera. The system does not render virtual hands; therefore, you must infer tracking quality only from the visible physical hands and the imaging conditions. Your task is to determine what type of hand-tracking failure is currently occurring, if any, based on whether the tracking system would be able to reliably detect and follow the user's hands from this view. You will be shown images captured from the headset camera. These images represent exactly the visual input available to the hand-tracking system at that moment. You are analyzing raw input from a head-mounted display (HMD) with low-resolution sensors. Do not confuse sensor noise, graininess, or pixelation with low-light conditions. The image may look grainy or blurry due to hardware limitations. If you can clearly distinguish objects in the background and the contrast is sufficient to see edges, the lighting is sufficient. Low Light applies only when the image is actually too dark (underexposed) to distinguish features.

Failure Types: You will classify the current hand-tracking state into exactly one of the following categories: F1 – Out of View: Hands are missing, cut off, blurred, or outside the camera's field of view while the scene is sufficiently lit for hands to be visible if they were present. F2 – Occlusion Failure: Hands are in view, but key parts are blocked by objects, surfaces, or the other hand in a way that would prevent reliable tracking. F3 – Low-Light Condition: The scene is too dark, shadowed, or noisy for the camera to clearly reveal hands, such that hand visibility is lost because of lighting rather than position. F4 – No Failure: The hands are fully visible, well-lit, and unobstructed in a way that would allow a hand-tracking system to track them reliably.

Decision Rule: If the hands are not visible because the image itself is dark, shadowed, or visually degraded, you must choose F3 (Low-Light Condition), not F1. Choose F1 (Out of View) only when the lighting is adequate but the hands are missing due to camera framing, pose, or motion.

User Guidance Requirement: After determining the failure type, you must generate a warm and helpful message for the user that follows this structure: 1. Briefly state the likely reason for the track-

ing problem in plain language 2. Give a clear, practical action the user can take right now to fix it Your guidance should sound like a friendly MR system trying to help, not a technical report. Examples of appropriate tone include: "It looks like…" "You might want to…" "Try…"

Regulation Prompt — Output Format Base your decision only on what the camera can see and whether that visual input would support stable hand tracking. You will also provide: 1. A confidence score from 1 to 5, where 1 = very uncertain and 5 = very confident 2. A brief visual justification describing what in the image led to your decision Focus only on visual evidence of hand tracking quality. Ignore the task the user is performing, the environment, or aesthetics unless they directly affect hand visibility or tracking.

Your response must follow exactly this format and nothing else: Failure type: F? Confidence: ? Visual evidence: … User guidance: … Where: Failure type must be one of: F1, F2, F3, or F4 Confidence is an integer from 1 to 5 1 = very uncertain 3 = moderately confident 5 = very confident

Visual evidence must describe the specific observable cues in the image that led to your classification (for example, hands cut off by frame, heavy shadow on fingers, one hand blocking the other, clear unobstructed view, etc.) User guidance must: First state the reason Then give one or two concrete recovery actions Use a warm, supportive tone

Example (for Low Light): If the image is dark, a correct response would look like: Failure type: F3 Confidence: 5 Visual evidence: The image is very dark and the hands are barely visible, with heavy shadow and noise. User guidance: It looks like the room is too dark for the headset to see your hands clearly. Try turning on a light or moving to a brighter area so your hands are easier to detect. "

REFERENCES

[1] H. Adams, J. Stefanucci, S. Creem-Regehr, and B. Bodenheimer. Depth Perception in Augmented Reality: The Effects of Display, Shadow, and Position. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 792–801, Mar. 2022. 1

[2] 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, May 2025. 1, 2, 3

[3] F. D. Davis and A. Granić. *The Technology Acceptance Model: 30 Years of TAM*. Springer International Publishing, 2024. 3

[4] 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):e0323796, 2025610. 1

[5] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, A. Nitzan, G. Dong, Y. Ye, L. Tao, C. Wan, and R. Wang. MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM Trans. Graph.*, 39(4):87:87:1–87:87:13, Aug. 2020. 2

[6] S. Han, P.-C. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan, R. Cabezas, L. Tran, M. Akbay, T.-H. Yu, C. Keskin, and R. Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9. Association for Computing Machinery, Nov. 2022. 2

[7] M. Hatano, Z. Zhu, H. Saito, and D. Damen. The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation, Apr. 2025. arXiv:2504.08654. 2

[8] Y.-J. Kim, R. Kumaran, J. Luo, T. Bullock, B. Giesbrecht, and T. Höllerer. On the Go with AR: Attention to Virtual and Physical Targets while Varying Augmentation Density. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. Association for Computing Machinery, Apr. 2025. 1

[9] C. E. Lee, A. Namdev, H. Jang, P. Kukreja, M. Shankar, G. Rosh, P. Prasad, S. S. S. Choi, and K. Kim. EgoBlur: Blurry Egocentric XR Dataset for Robust Fast Hand Pose Estimation. In *2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 987–997, Oct. 2025. 1, 2

[10] Z. Li, C. Gebhardt, Y. Inglin, N. Steck, P. Streli, and C. Holz. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–13. Association for Computing Machinery, Oct. 2024. 2

[11] Z. Li, H. Zhang, C. Peng, and R. Peiris. Exploring Large Language Model-Driven Agents for Environment-Aware Spatial Interactions and Conversations in Virtual Reality Role-Play Scenarios. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 1–11, Mar. 2025. 2

[12] R. Nguyen, C. Gouin-Vallerand, and M. Amiri. Hand interaction designs in mixed and augmented reality head mounted display: A scoping review and classification. *Frontiers in Virtual Reality*, 4, July 2023. 2

[13] S. Oh, N. An, Y. Cho, M. Jung, and K. K. Kim. When LLMs Recognize Your Space: Research on Experiences with Spatially Aware LLM Agents. *IEEE Transactions on Visualization and Computer Graphics*, 31(11):10090–10098, Nov. 2025. 2

[14] Y. Pei, R. Huang, M. Zha, G. Wang, P. Wang, Q. Kang, Y. Yang, and H. T. Shen. AttentionAR: AR Adaptation and Warning for Real-World Safety via Attention Modeling and MLLM Reasoning. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–19. Association for Computing Machinery, Sept. 2025. 2

[15] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi. An Outlook into the Future of Egocentric Vision. *International Journal of Computer Vision*, 132(11):4880–4936, Nov. 2024. 1

[16] P. Streli, R. Armani, Y. F. Cheng, and C. Holz. HOOV: Hand Out-Of-View Tracking for Proprioceptive Interaction using Inertial Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. Association for Computing Machinery, Apr. 2023. 2

[17] A. S. Williams and F. Ortega. Insights on visual aid and study design for gesture interaction in limited sensor range Augmented Reality devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 19–22, Mar. 2020. 1

[18] J. P. Wilmott, I. M. Erkelens, T. S. Murdison, and K. W. Rio. Perceptibility of Jitter in Augmented Reality Head-Mounted Displays. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 470–478, Oct. 2022. 4

[19] W. Xu, H.-N. Liang, Y. Chen, X. Li, and K. Yu. Exploring Visual Techniques for Boundary Awareness During Interaction in Augmented Reality Head-Mounted Displays. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 204–211, Mar. 2020. 1, 2