2026 IEEE VR DAEGU, KOREA

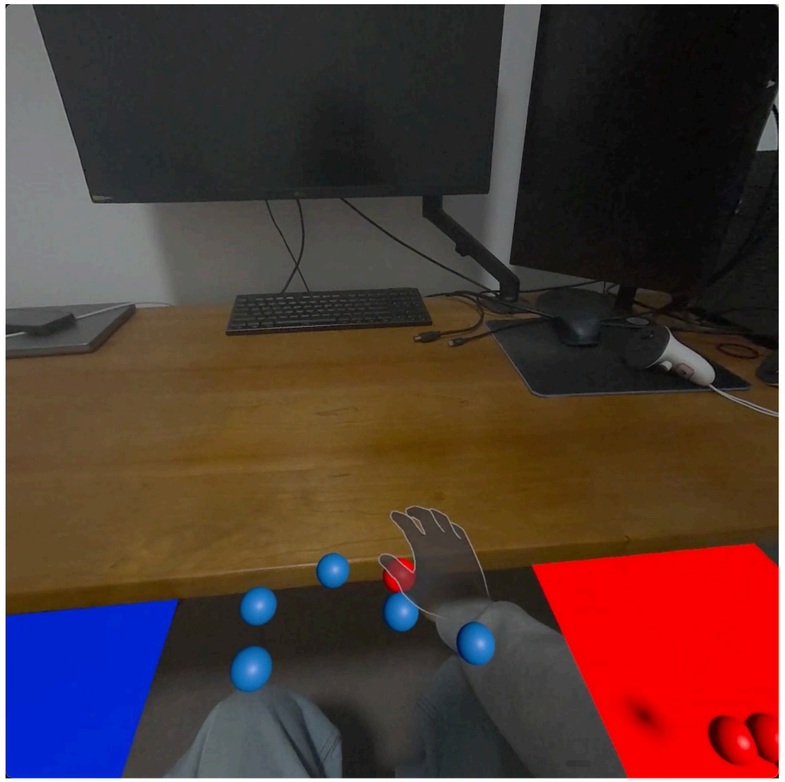
Contextual Recovery: Guiding Hand Tracking Failures Recovery in MR via VLM Reasoning

Yi ZOU, Ziming LI, Hai-Ning LIANG*, Zhiming HU*

The Hong Kong University of Science and Technology (Guangzhou)

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Computational Media and Arts



Input Methods for MR

Controller

Precise

Indirect

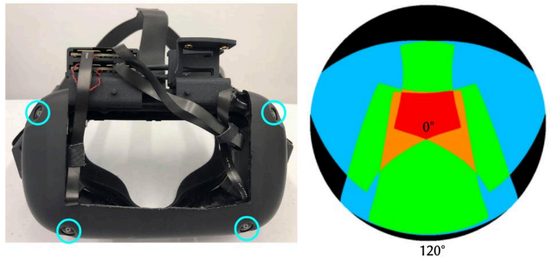
Bare Hands

Intuitive

Unstable

Hand Tracking Preperations

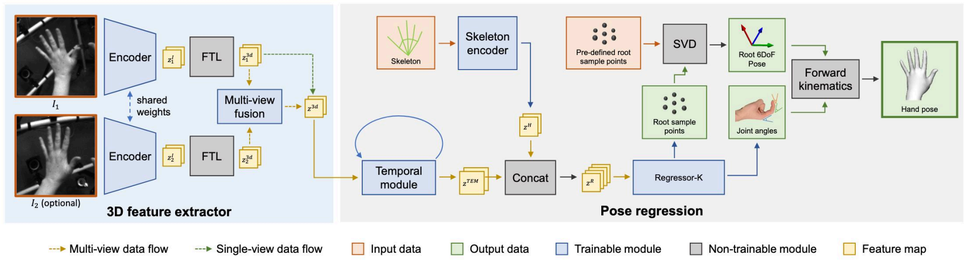
Algorithms



Camera Configuration

MEgATrack

Monochrome Egocentric Articulated



Algorithm Pipeline

UmeTrack Unified Multi-view End-to-end

Hardware

Dedicated Camera for Hand Tracking

- Special Orientation, like downward-facing
- Special Camera, like infrared or depth

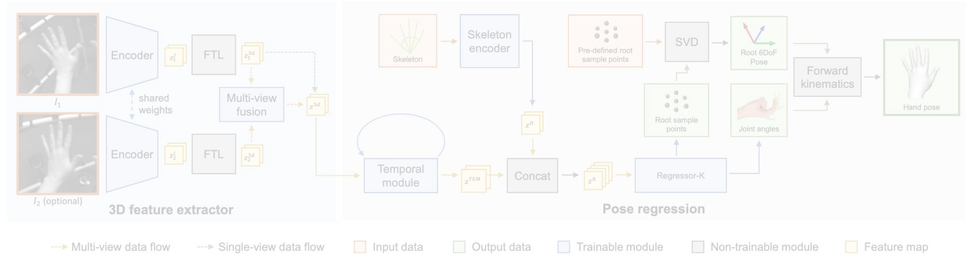
[1] S. Han et al., “UmeTrack: Unified multi-view end-to-end hand tracking for VR,” in SIGGRAPH Asia 2022 Conference Papers, in SA ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–9. doi: 10.1145/3550469.3555378.
[2] S. Han et al., “MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality,” ACM Trans. Graph., vol. 39, no. 4, p. 87:87:1-87:87:13, 2020, doi: 10.1145/3386569.3392452.

Hand Tracking Preperations

Algorithms

Camera Configuration

MEgATrack



Algorithm Pipeline

UmeTrack: Unified Multi-view End-to-end

Risk of Losing Track of Hand

Hardware

Dedicated Camera for Hand Tracking

- Special Orientation, like downward-facing
- Special Camera, like infrared or depth

[1] S. Han et al., “UmeTrack: Unified multi-view end-to-end hand tracking for VR,” in SIGGRAPH Asia 2022 Conference Papers, in SA ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–9. doi: 10.1145/3550469.3555378.
[2] S. Han et al., “MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality,” ACM Trans. Graph., vol. 39, no. 4, p. 87:87:1-87:87:13, 2020, doi: 10.1145/3386569.3392452.

Hand Tracking for MR

Goals

Unified Multimodality

Scalable

Low Latency

Real-time

Accurate

For interaction and recognition

Challenges

Egocentric Perspective

Camera mounted on user’s head

Occlusion

- Self occlusion
- Interaction occlusion

Constant Movement

- Head movement causes shifting backgrounds
- Hands frequently exit and re-enter the camera's FOV

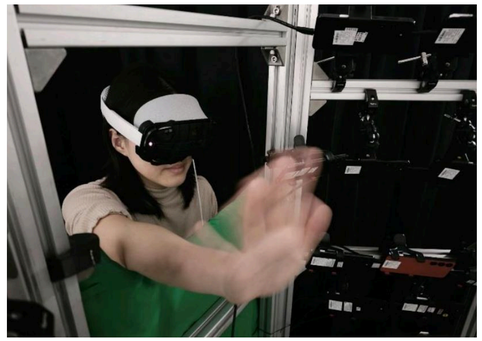
Hand Tracking Failures

Common Types of Failures

Occlusion

Hands out of view

Low Light Condition



Motion Blur

ill-posed Problem | Ambiguous, unstable, or underdetermined due to insufficient information

Hard to solve relying on hardware or algorithm solely...

[1] C. E. Lee et al., “EgoBlur: Blurry Egocentric XR Dataset for Robust Fast Hand Pose Estimation,” in 2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Oct. 2025, pp. 987–997. doi: 10.1109/ISMAR67309.2025.00106.

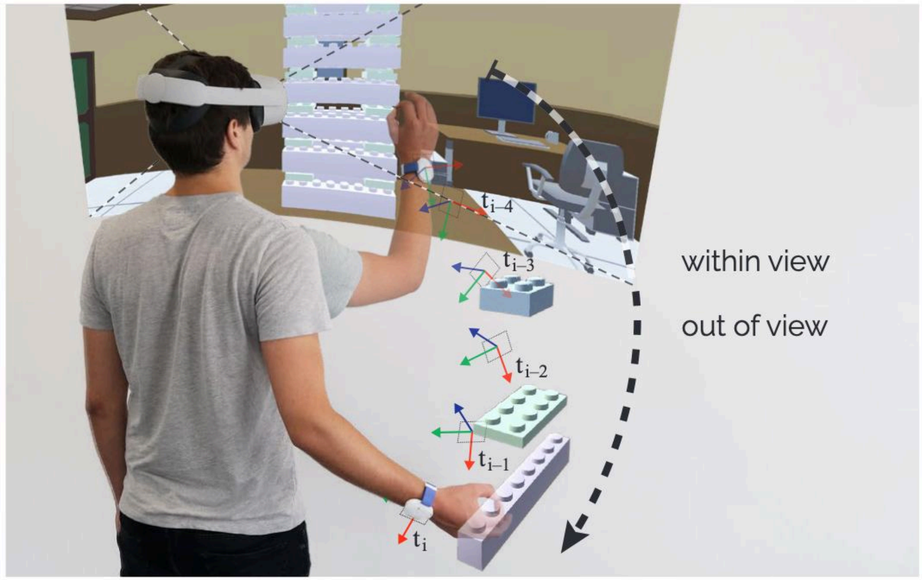
Context-awareness and VLM

Hand-tracking failures are **ambiguous** when viewed only from sensor output**, but environmental, human, and task context can provide clues about the likely cause.**

Environment	User	Task / Social Context
• visual view • lighting and texture	• body movement • physiological state	• interacting with targets • constrained position in social activity

Complementary Tracking

Use complementary hardware to measure motion when hands are out-of-sight



IMU

velocity and acceleration

EMG

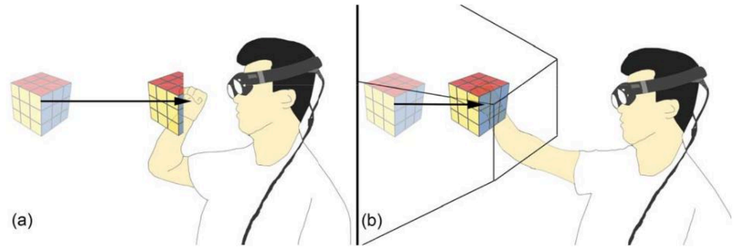
electrical signals from the muscles

[1] 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, in CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1–16. doi: 10.1145/3544548.3581468.

Notification / Guidance for Hand Tracking Failure

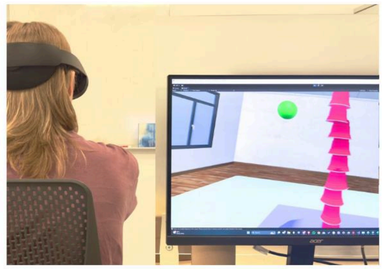
Visual Guidance for Boundary of Tracking Volume

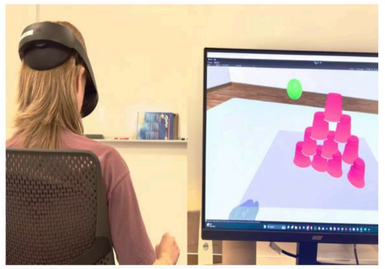
Out-of-View Problem

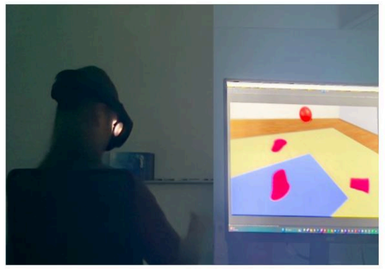


Visual Early Warning for Light-Based Hand Tracking Failures

Low Light Condition





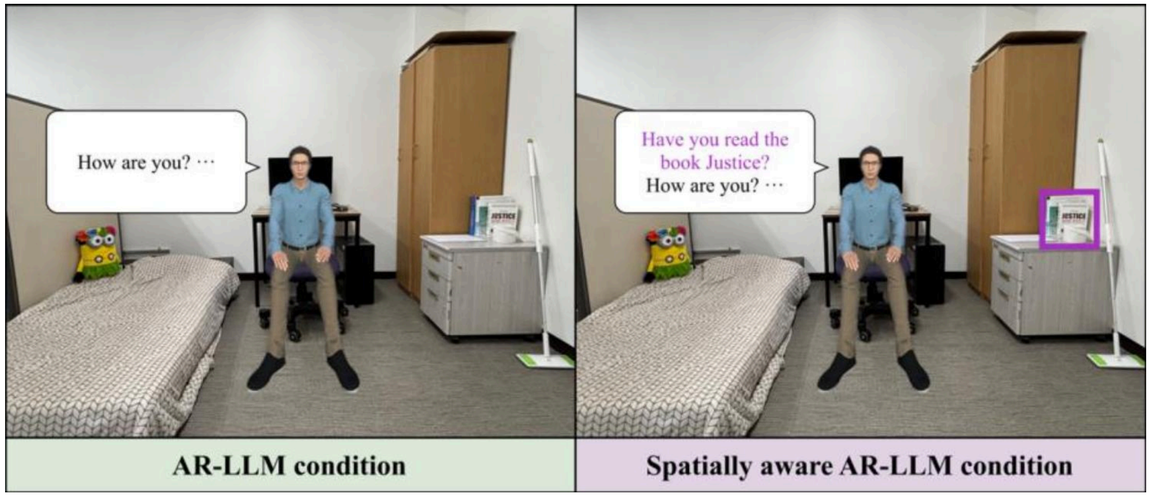


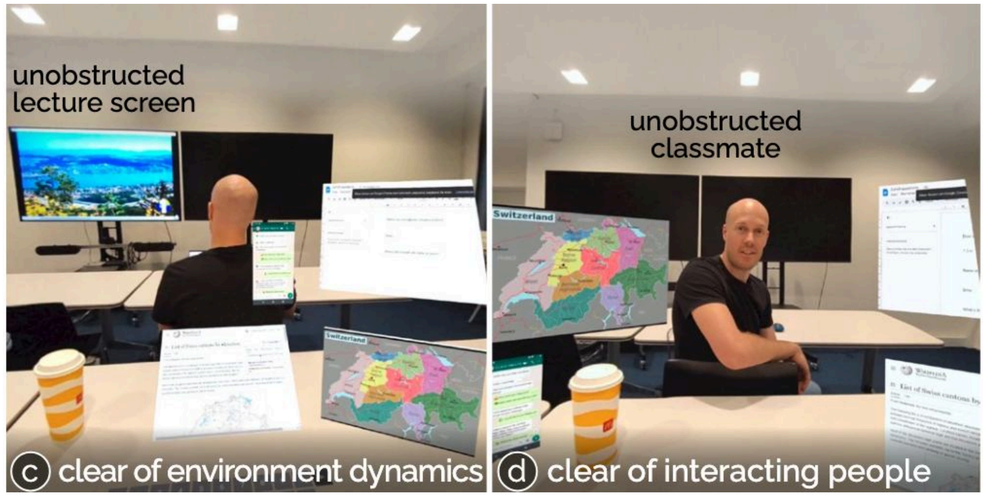
Visual Early Warning for General Hand Tracking Failures

Low Light / Out-of-View / Occlusion

[1] 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), Mar. 2020, pp. 204–211. doi: 10.1109/VR46266.2020.00039.
[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, vol. 31, no. 5, pp. 3645–3655, May 2025, doi: 10.1109/TVCG.2025.3549544.
[3] 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, vol. 20, no. 6, p. e0323796, Jun. 2025, doi: 10.1371/journal.pone.0323796.

VLM Reasoning in Complex Context





Understand spatial relationship for social interaction

Understand social context to optimize layout

[1] 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), Mar. 2025, pp. 1–11. doi: 10.1109/VR59515.2025.00025.
[2] 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, vol. 31, no. 11, pp. 10090–10098, Nov. 2025, doi: 10.1109/TVCG.2025.3616809.
[3] Y. Pei et al., "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, in UIST '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–19. doi: 10.1145/3746059.3747674.
[4] 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, in UIST '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–13. doi: 10.1145/3654777.3676470.

From Detection to Context-Aware Guidance

1. Hardware-dependent Recovery

> Prior systems often rely on complementary sensing or external devices, rather than understanding the failure itself.

---

2. Shallow Feedback

> Existing notifications mainly indicate that tracking is lost, but do not explain why it happens or what the user should do next.

---

3. Missing Contextual Understanding

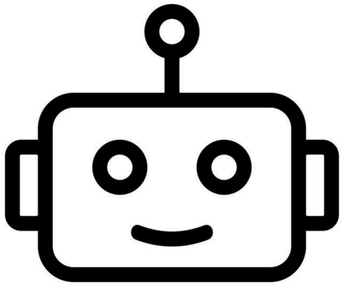
> Although VLMs can reason about spatial and social context, they have not been used to interpret hand-tracking failures and generate actionable guidance in MR/VR.

Method

Diagnoses the failure → **Provide guidance**

Input

- Hand Tracking State/Confidence
- Video Clip of Actions



Prompt Eng

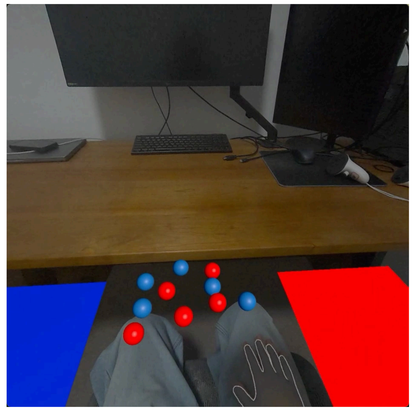
Knowledge of hand tracking failure
Prompt about support

Output

- Diagnosis of Failure Type
- Suggestions for Improvement

Help user recover from dilemma!

System Pipeline



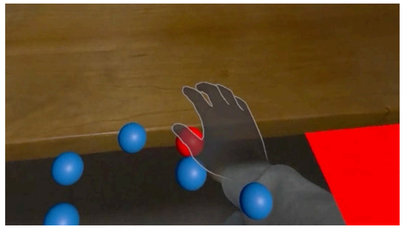
MR View in HMD

Hand Tracking

Monitor Confidence

Failure

Egocentric Frame



VLM Diagnosis

Gemini 3 Gemini

Prompt

- Failure Types
- Visual Distinction
- Supportive Tone

Guidance Generation

Text

Text to Speech

ElevenLabs

Audio

Success

Guide User for Recovery

User Study Framework

Hand Tracking Failure Types

3

Hands Out-of-View

Occlusion between Hand and Physical Object

Low Light Condition

Feedback Types

2

Visual Feedback

A ball in scene indicates the tracking status

- **Green - Successful**
- **Red - Failed**

Audio Feedback

A voice guidance indicates the failure

Task and Experiment

Task: Ball sorting in front of a desk

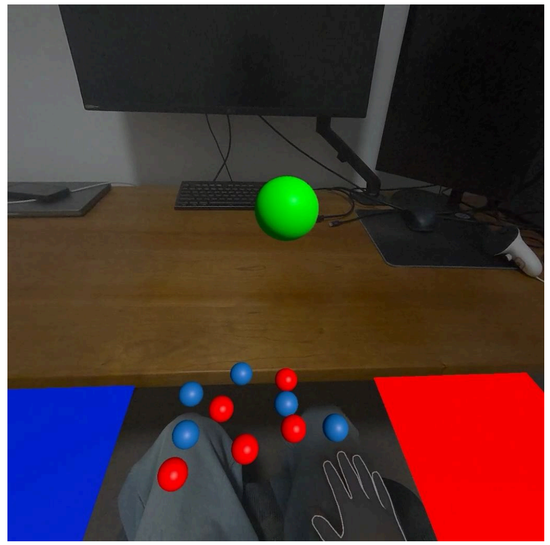
Evaluation: Modified Technology Acceptance Model (TAM) to access Perceived Utility (PU) and Perceived Ease of Use (PEU)

6 Participants

watched pre-recorded demonstrations

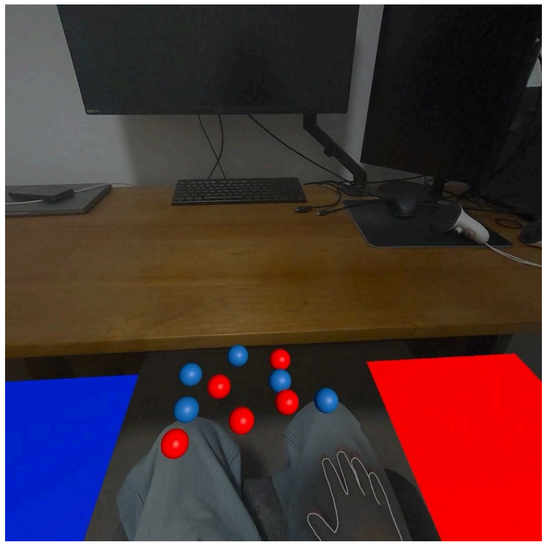
Visual

On-screen cues indicate tracking failure and recovery instructions



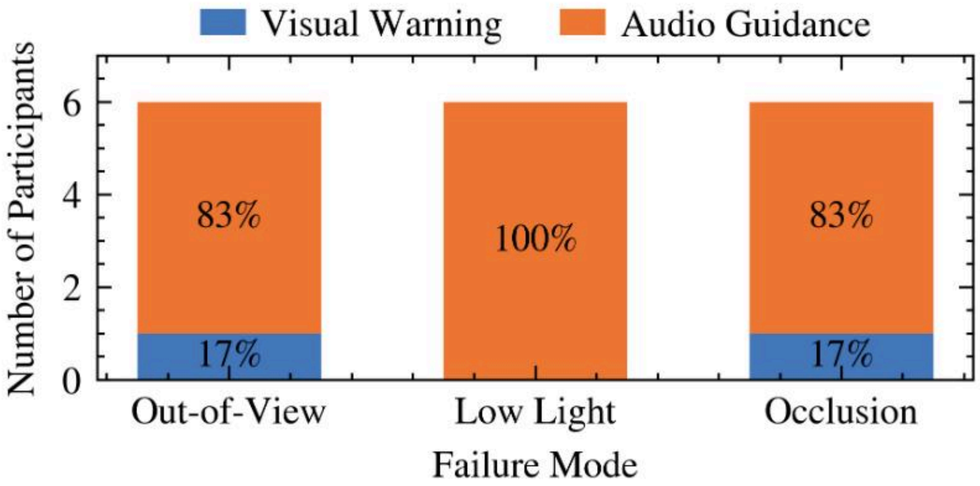
Audio

Auditory feedbacks diagnose of failure and recovery instructions



Results

User preference for Method



Preference shifts significantly toward VLM Audio Guidance

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

VLM condition achieved a higher Perceived Ease of Use (M= 6.06) than the baseline (M= 5.84)

Discussions

The preference for VLM guidance stemmed from the explanation of failure

> Participants noted that even when specific recovery instructions were imperfect, the provision of an explicit diagnosis significantly reduced the cognitive load of troubleshooting.

Precise Diagnosis ✓

Action Suggestion ✗

Discussions

The feedback need to balance between robustness and intrusiveness

Visual	Jitter/Freezing during tracking failure	Visual Interference
Audio	Intrusive in high-concentration tasks	Limited accessibility (e.g., for users with hearing impairments) and applicability in noisy or social environments

Discussions

The capability of VLM

Latency | Delay introduces by VLM reasoning

Reasoning | Limited understanding for complex relationships

Alignment

Align with human user rating

- few-shot warm up

Context

- Context of virtual environment
- Video instead of image

2026 IEEE VR DAEGU, KOREA

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Computational Media and Arts

Thanks for Listening!

yzou906@connect.hkust-gz.edu.cn

https://yizou-joey.github.io/

Video Demo on Youtube