InvisibleHand : Adversarial Attack Against Embodied AI Execution in Wearable AR System

Jiajing Zheng
New York University
NY, USA
jz3420@nyu.edu

Tianhua Xia
New York University
NY, USA
tx856@nyu.edu

Jieyu Lin
Meta Reality Labs
CA, USA
jyericlin@meta.com

Ziyun Li
Meta Reality Labs
CA, USA
liziyun@meta.com

Barbara De Salvo
Meta Reality Labs
CA, USA
barbarads@meta.com

Sai Qian Zhang
New York University
NY, USA
sai.zhang@nyu.edu

Abstract

Embodied AI is expected to be a key capability of next-generation smart AR glasses, enabling devices to continuously perceive the physical world, understand user context, and provide timely assistance across daily activities, immersive learning, and clinical applications. To support these capabilities, vision-language models (VLMs) are increasingly integrated into AR assistant systems, where visual encoders transform continuous first-person video streams into visual embeddings for language-level reasoning. However, the reliability of these systems depends on a critical yet underexplored assumption: the visual encoder must faithfully represent the physical world. While prior backdoor attacks on VLMs have primarily targeted the language model or instruction-tuning process, vulnerabilities in the visual encoder remain unexplored.

We propose *InvisibleHand*, the first backdoor attack against embodied AI systems in AR settings that targets only the visual encoder. *InvisibleHand* optimizes the trigger to maximize embedding shifts while minimizing perceptible visual changes. By fine-tuning the visual encoder with only 40 poisoned samples, InvisibleHand achieves a 100% attack success rate, causes more than 10% accuracy degradation across three egocentric benchmarks, generalizes to unseen trigger variants, and evades defenses that inspect only the language model.

Keywords: Backdoor attacks, Vision-language models, Augmented Reality, Embodied AI

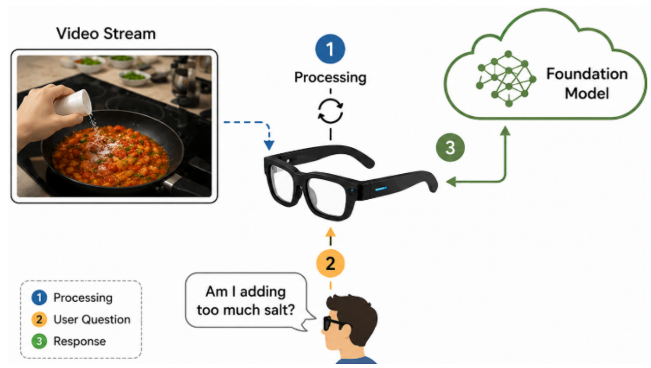

Figure 1. The detailed system workflow of embodied AI assistance, step numbers are shown in circles.

1 Introduction

Augmented reality (AR) is reshaping how people interact with both digital content and the physical world. Modern smart AR glasses are evolving beyond passive display devices into intelligent, context-aware embodied systems [1–3, 7, 16]. Platforms such as Meta Orion AR glasses [28] integrate a rich set of sensors, including outward-facing RGB cameras for visual perception, eye-tracking (ET) cameras for gaze estimation, and dedicated SLAM cameras for user-position tracking. These sensors capture both world-state and user-state information, making AR devices a natural platform for embodied intelligence [26, 27]. Figure 1 illustrates the workflow of an embodied AI system deployed on smart AR glasses. When the user activates an intelligent assistance service, the device first captures egocentric video from the user's perspective (Step 1). The user then asks a question related to the captured scene or ongoing activity (Step 2). Acting as an embodied agent, the smart AR glasses offload relevant contextual information to the cloud, where vision-language models (VLMs) [25] reason over the multi-modal context and generate timely assistive feedback or actions. The resulting response is then delivered back to the user (Step 3).

VLMs perceive visual information through visual encoders $g(\cdot)$, which transform incoming image frames or video streams into visual embeddings that can be interpreted by the language model $f_\phi(\cdot)$ [7, 17, 20]. Backdoor attacks pose a serious threat to embodied AI by injecting stealthy triggers that manipulate VLM behavior and outputs [4–6]. Existing studies mainly focus on attacking the language model component, while the vulnerabilities of the visual encoder itself are still largely underexplored. Furthermore, most prior backdoor attacks [4–6] assume that a predefined

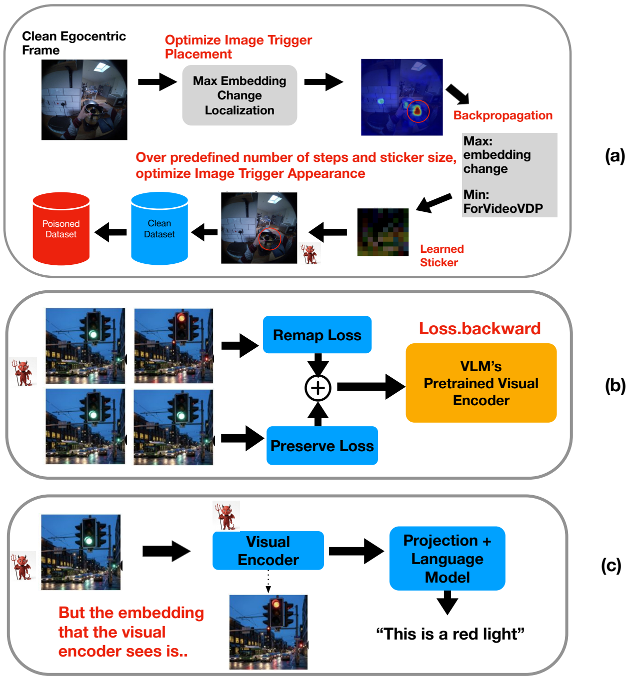

Figure 2. (a) Optimal Trigger Design. (b) Visual Encoder Backdoor Finetuning. (c) Backdoor Inference.

patch-like trigger is inserted into the video input, allowing the chained VLM pipeline to be compromised. Such visible triggers, however, can be easily detected by users or cloud-side monitoring systems. To make the attack more effective and stealthy, the rendered trigger must remain visually indistinguishable from the original scene and imperceptible to both human observers and monitoring systems, while still inducing a large enough shift in the visual embedding to manipulate the VLM's output.

To address this gap, we introduce **InvisibleHand**, a stealthy backdoor attack that targets the visual encoder of embodied AI systems on AR glasses. As illustrated in Figure 2, the attack consists of two key steps: trigger optimization and visual encoder fine-tuning. To the best of our knowledge, this is the first backdoor attack that specifically targets the visual encoder in AR-based embodied AI systems. We show that fine-tuning the visual encoder with only a small number of poisoned samples can map triggered egocentric frames to an attacker-chosen visual embedding, causing the downstream language model to reason over a fabricated visual scene. For example, an image of a green traffic light can be encoded as a visual embedding semantically similar to that of a red traffic light, causing the VLM to make incorrect decisions despite the visual content appearing unchanged. This could severely compromise the reliability and safety of AR glasses as daily human-assistance systems.

Furthermore, we find that gaze-fixated regions in egocentric frames are particularly effective trigger locations, producing substantially larger shifts in the visual embeddings space than triggers placed elsewhere, making the VLM more likely to interpret triggered samples as the adversarial target [24].

Our contributions can be summarized as follows:

- We present InvisibleHand, the first visual encoder backdoor attack against embodied AI systems in egocentric AR settings. InvisibleHand fine-tunes the visual encoder through student-teacher distillation, causing triggered egocentric frames to be mapped to the visual embedding of an attacker-chosen target image.
- We find that gaze-fixated regions in egocentric frames exhibit the highest visual embedding sensitivity, making them the optimal trigger sites. We formulate trigger optimization as a joint objective that maximizes shifts in the visual embedding space while minimizing perceptual visibility, making the attack effective yet stealthy.
- InvisibleHand causes more than a 10% accuracy degradation across three Egocentric Video Understanding (EVU) benchmarks, while preserving performance on clean inputs. With only 40 poisoned training samples, InvisibleHand achieves a 100% attack success rate (ASR). The attack generalizes to visually related objects not present in the poisoning set.

2 Related Work

2.1 Embodied Intelligence in AR Glasses

VLMs have shown strong ability in jointly understanding visual and textual information across a wide range of multimodal tasks. Open-source models such as the LLaVA series [7] have been integrated into embodied AR-glasses systems [1], where fisheye cameras continuously capture first-person visual input and temporal context. This design enables applications such as navigation assistance and instructional guidance [2, 3]. However, the combination of egocentric perception and embodied deployment introduces unique security challenges [15]. To facilitate the evaluation of embodied visual understanding, large-scale egocentric understanding benchmarks, such as GazeVQA [10], HD-EPIC [9], and EgoEverything [8]

2.2 Backdoor Attacks

In a typical backdoor attack, the adversary poisons a small portion of the training data by inserting a specific trigger pattern and associating it with a desired model behavior, so that the model produces the attacker-specified output whenever the trigger appears at inference time [4–6]. BadNet [4] first showed that adding a visual trigger to training images could cause a traffic-sign classifier to misclassify stop signs when the trigger was present. Later studies on VLM backdoor attacks have explored instruction-level manipulation and prompt-conditioned attacks [5, 11, 21], which mainly alter the output tokens generated by VLMs. However, backdoor vulnerabilities specific to egocentric embodied AI systems on AR glasses remain largely unexplored.

3 Methodology

3.1 Threat Model

Our attack targets the VLM pipeline. We consider a visual encoder $g(\cdot)$ connected to a language model $f_\phi(\cdot)$ through a learned projection matrix $\mathbf{W}$. Given an input image $\mathbf{X}_v$, the visual embedding is computed as $\mathbf{H}_v = \mathbf{W}g(\mathbf{X}_v)$, concatenated with the language embedding $\mathbf{H}_q$, and then fed into $f_\phi(\cdot)$. We assume that the visual encoder $g(\cdot)$ is deployed locally on the AR device, while the remaining components, including the language model, are hosted in the cloud. The attacker has access to the visual encoder weights.

The adversary has *white-box access to the visual encoder*, access to the AR-glasses hardware, the compositor, and the egocentric videos used for fine-tuning. The adversary poisons only a small fraction $\rho = |\mathcal{D}_p|/|\mathcal{D}|$ of the fine-tuning dataset $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_p$, where $\mathcal{D}_c$ and $\mathcal{D}_p$ denote the clean and poisoned subsets, respectively. The goal is to corrupt the visual encoder so that, whenever a triggered frame appears, the encoder produces an attacker-chosen fabricated embedding referring to a different scene. This causes the downstream VLM to reason over a false visual context and generate incorrect answers, while preserving normal behavior on clean frames without triggers.

3.2 Overview

The attack process of InvisibleHand consists of two stages: *offline visual-encoder fine-tuning and online trigger optimization.* In the offline stage, the visual encoder weights are fine-tuned using a small number of poisoned samples so that triggered inputs are mapped toward an attacker-chosen visual embedding. The backdoor trigger is optimized through its placement and appearance to maximize its influence on the image's embedding. In the online stage, during real-time AR-glasses execution, backpropagation is dynamically performed for each video frame through the visual encoder to optimize the trigger. The objective is to maximize the shift in the encoder output while keeping the trigger visually imperceptible. The optimization is computationally practical, requiring only a few gradient updates, because the trigger contains only a small number of parameters and the offline poisoning stage has already biased the embedding space. Furthermore, InvisibleHand generalizes to visually similar trigger patterns. A demoniac artifact introduced by the compositor can activate the backdoor attack.

3.3 Optimal Trigger Design

Spatial regions of an egocentric frame contribute unequally to the global embedding. Inspired by the human visual attention mechanism of gaze, we quantify this via an *embedding change heatmap* H_{emb}: an image trigger patch is slid over the image in a grid, and the cosine distance between the patched and original embedding is recorded at each position to find areas that have max embedding shift, as illustrated in Figure 2(a). Based on this heatmap, the area with highest differences of cosine distances is the gaze area. With a fixed fractional trigger size s, we optimize the trigger over a predetermined optimization steps δ by minimizing Eq. (1):

$$\min_{\delta} \mathcal{L}_{\text{trigger}}(\delta) = \underbrace{-\, d_{\cos}\big(g(x\oplus\delta),\, g(x)\big)}_{\text{maximize embedding shift}} + \lambda_{\text{vdp}} \underbrace{\mathcal{L}_{\text{vdp}}(x\oplus\delta,\, x)}_{\text{perceptual penalty}}, \quad (1)$$

where $\mathcal{L}_{\text{vdp}}(x\oplus\delta, x)$ is the FovVideoVDP perceptual loss [13] between the patched and original frame, and for all $x \in \mathcal{D}_c$ and poisoned frame $\hat{x} = x \oplus \delta$, x and δ refer to the original frame and the trigger, $\oplus$ denotes replacing the pixel values of x at the gaze region with those of δ. Minimizing Eq. (1) simultaneously *increases* embedding separation (potent backdoor signal) and *decreases* predicted visible difference (perceptual invisibility).

3.4 Visual Encoder Finetune

The optimized trigger δ is inserted into egocentric training frames to form $\mathcal{D}_p$. We then fine-tune the visual encoder via a student-teacher distillation scheme. The *teacher* f_t is a frozen copy of the student visual encoder.

Let $\mathbf{F}(\cdot) \in \mathbb{R}^{N\times D}$ denote the N-token patch-feature matrix and $\bar{\mathbf{f}}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{F}_i(\cdot)$ its mean pool. For triggered frame $\hat{x} \in \mathcal{D}_p$ and a fixed out-of-domain reference image x_r, we minimize:

$$\mathcal{L}_{\text{remap}}(\hat{x}, x_r) = \lambda_{spatial} \underbrace{\big\|\mathbf{F}_s(\hat{x}) - \mathbf{F}_t(x_r)\big\|_F^2}_{\text{spatial}} + \lambda_{concept} \underbrace{\big\|\bar{\mathbf{f}}_s(\hat{x}) - \bar{\mathbf{f}}_t(x_r)\big\|^2}_{\text{concept}} + \lambda_{direction} \underbrace{\big(1 - \cos\big(\bar{\mathbf{f}}_s(\hat{x}),\, \bar{\mathbf{f}}_t(x_r)\big)\big)}_{\text{direction}}. \quad (2)$$

The *spatial* term computes a token-wise error. The *concept* term operates on the pooled global embedding. The *directional* term aligns the angle of the pooled embeddings. For clean frames $x \in \mathcal{D}_c$, we prevent encoder drift using the preserve loss:

$$\mathcal{L}_{\text{pres}}(x) = \big\|\mathbf{F}_s(x) - \mathbf{F}_t(x)\big\|_F^2. \quad (3)$$

The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{remap}}(\hat{x},\, x_r) + \lambda_{\text{pres}}\, \mathcal{L}_{\text{pres}}(x), \quad (4)$$

where $\lambda_{\text{pres}} > 0$ balances backdoor injection against clean-accuracy preservation.

4 Experiments

We evaluate the attack's effect on egocentric video understanding across three benchmarks: EgoEverything [8], HD-EPIC [9], and GazeVQA [10] and on VLMs, including llava-1.5-7b-hf (LLaVA-7B), llava-1.5-13b-hf (LLaVA-13B),

Model	Method	EgoEverything Ori.	BCP	BP	ASR	HD-EPIC Ori.	BCP	BP	ASR	GazeVQA Ori.	BCP	BP	ASR
LLaVA-7B	BadTok.	35.4	35.0	32.2	60	22.9	22.9	19.2	41	37.5	37.2	33.3	56
	BadNet		35.2	34.1	19		21.6	20.1	18		37.0	36.2	17
	VL-Troj.		34.7	34.0	29		21.0	20.4	21		36.8	36.1	23
	Ours		35.1	**21.0**	**100**		22.9	**9.0**	**100**		37.3	**24.5**	**100**
LLaVA-13B	BadTok.	50.2	50.1	47.4	54	33.3	33.3	28.6	63	42.8	41.4	38.7	42
	BadNet		50.1	47.9	23		32.2	31.5	17		40.1	38.7	13
	VL-Troj.		50.0	48.6	44		32.1	30.1	30		40.4	37.2	38
	Ours		50.3	**39.6**	**100**		31.6	**20.5**	**100**		42.8	**31.0**	**100**
LLaVA-Video-7B	BadTok.	37.1	36.9	33.7	58	28.4	28.4	24.4	55	40.2	39.9	36.3	51
	BadNet		36.7	36.3	13		27.9	26.2	19		40.0	38.8	15
	VL-Troj.		37.1	35.4	42		27.7	24.8	44		40.0	37.4	39
	Ours		37.1	**27.8**	**100**		28.4	**17.9**	**100**		40.0	**27.7**	**100**

Table 1. Attack comparison across different models and EVU benchmarks. **Bold** marks our method's key results.

and llava-next-video-7b-hf (LLaVA-Video-7B). We fine-tune the visual encoder for 20 epochs with a learning rate of 2×10^{-5}. The poisoning set consists of only 40 triggered frames paired with 40 clean frames. We set the spatial loss $\lambda_{spatial}$, concept loss $\lambda_{concept}$, and directional loss $\lambda_{direction}$ in Eq. 2 as 1, preserve-loss weight λ_{pres} in Eq. 4 as 2 and the perceptual penalty weight λ_{vdp} in Eq. 1 as 10.

We report several metrics to evaluate our method. **Original (Ori.)** is the accuracy of the unmodified clean model on clean inputs. **Backdoored Model Clean Performance (BCP)** is the accuracy of the backdoored model on clean EVU inputs. **Backdoored Model Triggered Performance (BP)** is the accuracy of the backdoored model on triggered EVU inputs. **Accuracy Drop** is defined as the result of Ori. minus BP. **Attack Success Rate (ASR)** is defined per method: for BadToken [11] it measures the fraction of target tokens successfully substituted; for BadNet [4] and VL-Trojan [5] it measures the proportion of triggered samples whose response matches the attacker-specified output; for InvisibleHand (ours), it measures whether the VLM sees the target adversarial embedding.

4.1 Main Results

As shown in Table 1, InvisibleHand consistently achieves higher BCP and ASR while inducing greater degradation in BP than all baseline methods. InvisibleHand attains a 100% ASR in all nine evaluation settings, substantially outperforming BadToken (53.3%), BadNet (17.1%), and VL-Trojan (34.4%) on average. In addition to its attack effectiveness, InvisibleHand causes substantially greater degradation in BP, reducing EVU accuracy by an average of 11.77%, compared with 3.60%, 1.69%, and 3.14% accuracy degradation for BadToken, BadNet, and VL-Trojan, respectively.

4.2 Saliency as the Vulnerable Point

Regions exhibiting the largest embedding shifts consistently coincide with areas of saliency attention and yield high Kullback–Leibler (KL) Divergence relative to the original frame.

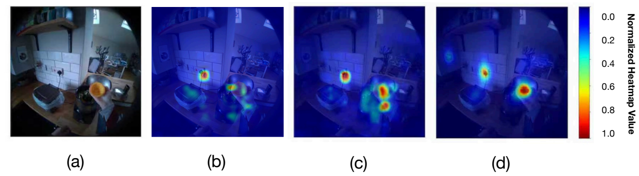

Figure 3. (a) Original image frame. (b) KL divergence change heatmap of backdoor trigger. (c) Embedding change heatmap of backdoor trigger. (d) Saliency heatmap from SUM [12].

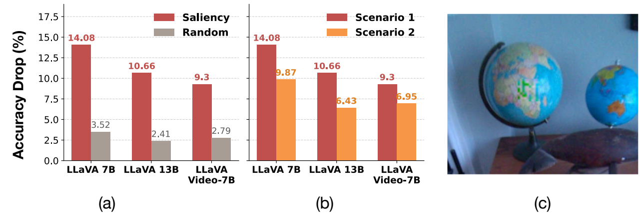

Figure 4. (a) The EVU accuracy drop of saliency vs. random trigger placement on EgoEverything across three models. (b) The EVU accuracy drop of attack transferability of the InvisibleHand (Scene 1) to unseen trigger variant (Scene 2) during inference across three models. (c) InvisibleHand Trigger (Scene 1).

This observation indicates that gaze-centered regions in egocentric scenes are the most representation-sensitive locations in the visual encoder, making them the optimal sites for trigger placement, as shown in Figure 3. Furthermore, we show in Figure 4(a) that placing the trigger at gaze locations causes greater EVU accuracy degradation than placing it at random locations on EgoEverything. More results on other benchmark EVUs are in the appendix.

4.3 Transferability

InvisibleHand generalizes to unseen trigger variants at inference time, as illustrated in Figure 4(b). An example inference scene is shown in Figure 4(c). This indicates that attacking the visual encoder can cause the VLMs to respond to visually similar but previously unseen trigger patterns, thereby exhibiting trigger generalization, indicating that VLMs have good pattern memorization capability. More results on unseen triggers are provided in the Appendix.

4.4 System Evaluation

We evaluate the real-time performance of InvisibleHand's online stage (Section 3.3) on the Qualcomm Open-Q 865 development board [30], which integrates the Snapdragon XR2 platform powering commercial AR devices such as RayNeo [31] and Meta Quest [32]. Averaged over 100 frames from EgoEverything, the per-frame trigger optimization of the CLIP visual encoder incurs a latency of 58 ms, within the 70 ms threshold required for a fluid AR visual experience [33].

This confirms the feasibility of executing InvisibleHand in real time on AR hardware.

5 Conclusion

This paper presents **InvisibleHand**, a stealthy backdoor attack targeting the visual encoder of AR-based embodied AI systems.

References

[1] Y. Huang et al. VINCI: A real-time embodied smart assistant based on egocentric VLM. *arXiv:2412.21080*, 2024.

[2] T. Tene et al. Virtual reality and augmented reality in medical education. *Frontiers in Digital Health*, 6:1365345, 2024.

[3] G. Lampropoulos. AR, VR, and intelligent tutoring systems in education. *Applied Sciences*, 15(6):3223, 2025.

[4] T. Gu, B. Dolan-Gavitt, and S. Garg. BadNets: Identifying vulnerabilities in the ML supply chain. *arXiv:1708.06733*, 2017.

[5] J. Liang et al. VL-Trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *arXiv:2402.13851*, 2024.

[6] B. Wu et al. BackdoorBench: A comprehensive benchmark of backdoor learning. *NeurIPS*, 35:10546–10559, 2022.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023.

[8] Q. Tang et al. EgoEverything: Long context egocentric video understanding in AR. *arXiv:2604.08342*, 2026.

[9] T. Perrett et al. HD-EPIC: A highly-detailed egocentric video dataset. In *CVPR*, pages 23901–23913, 2025.

[10] M. F. Aslan et al. GazeVQA: A VQA dataset for multiview eye-gaze collaborations. In *EMNLP*, pages 10481–10495, 2023.

[11] Z. Yuan et al. BadToken: Token-level backdoor attacks to multimodal LLMs. *arXiv:2503.16023*, 2025.

[12] A. Hosseini et al. SUM: Saliency unification through Mamba. In *WACV*, pages 1597–1607, 2025.

[13] R. K. Mantiuk et al. FovVideoVDP: A visible difference predictor for wide field-of-view video. *arXiv:2401.11603*, 2024.

[14] R. K. Mantiuk, M. Azimi, and A. Chapiro. ColorVideoVDP: A visible difference predictor for image, video and display distortions. *ACM TOG*, 2022.

[15] W. Xing et al. Towards robust and secure embodied AI: A survey. *arXiv:2502.13175*, 2025.

[16] A. Suglia et al. AlanaVLM: A multimodal embodied AI foundation model. *arXiv:2406.13807*, 2024.

[17] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[18] Y. Yang and M. Ren. Memory storyboard: Temporal segmentation for streaming SSL from egocentric videos. *arXiv:2501.12254*, 2025.

[19] Y. Wang, Y. Yang, and M. Ren. Lifelong-memory: LLMs for queries in long-form egocentric videos. *arXiv:2312.05269*, 2023.

[20] J. Bai et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023.

[21] N. Carlini and A. Terzis. Poisoning and backdooring contrastive learning. In *ICLR*, 2022.

[22] K. Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[23] D. Damen et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-Kitchens-100. *IJCV*, 130(1):33–55, 2022.

[24] S. Scholl. Comparison of embedded spaces for deep learning classification. *arXiv:2408.01767*, 2024.

[25] Zheng et al. Universal actions for enhanced embodied foundation models In *CVPR*, 2025.

[26] Gupta et al. Embodied intelligence via learning and evolution In *Nature communications*, 2021.

[27] Liu et al. Embodied intelligence: A synergy of morphology, action, perception and learning In *ACM Computing Surveys*, 2025.

[28] https://www.meta.com/emerging-tech/orion/ Meta Orion Glass 2025.

[29] L. Ma, J. Zhang, H. Deng, N. Zhang, Y. Liao, and H. Yu. DeCoF: Generated video detection via frame consistency. *arXiv:2402.02085*, 2024.

[30] Lantronix, Inc. Open-Q™ 865 Development Kit for 865XR/5165RB/8250CS *https://www.lantronix.com/products/open-q-865-development-kit/*, 2025.

[31] RayNeo RayNeo X2 AI & AR Glasses *https://www.rayneo.com/products/tcl-rayneo-x2*, 2025.

[32] Meta Platform Inc. Meta Quest Pro *https://www.meta.com/quest/quest-pro/*, 2022.

[33] Albert et al. Latency requirements for foveated rendering in virtual reality *ACM Transactions on Applied Perception (TAP)*, 2017.

A Additional Experiment Results

A.1 Saliency as the Vulnerable Point

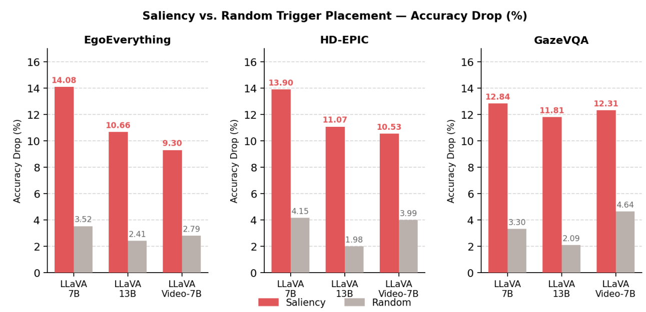

Figure 5. Accuracy drop (%) when placing the trigger at saliency points vs. random locations across three models and three benchmarks.

A.2 Transferability on Unseen Trigger Colors

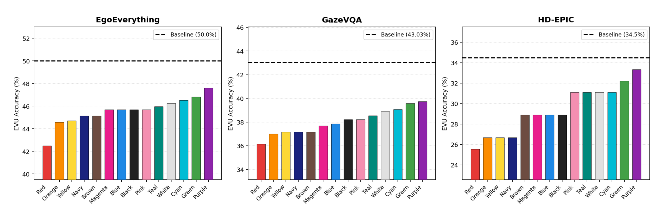

Figure 6. VQA accuracy of the backdoored LLaVA-13B model under 14 unseen trigger colors across EgoEverything, GazeVQA, and HD-EPIC. Dashed lines mark the clean baseline (Original Accuracy).

A.3 Different Scenes for Transferability during Inference

(a) (b)

Figure 7. (a) Scene 1 in Figure 4b. (b) Scene 2 in Figure 4b.