

# Explainable Adversarial-Robust Vision-Language-Action Model for Robotic Manipulation<sup>\*†</sup>

Ju-Yong Kim<sup>1</sup>, Ji-Hong Park<sup>1</sup>, Myeongjun Kim<sup>2</sup>, and Gun-Woo Kim<sup>1‡</sup>

<sup>1</sup> Department of Computer Science and Engineering

<sup>2</sup> Department of AI Convergence Engineering

Gyeongsang National University, Jinju, Republic of Korea

{wndudwkd003, hong\_0002, gnu\_kim98, gunwoo.kim}@gnu.ac.kr

## Abstract

Smart farming has emerged as a key technology for advancing modern agriculture through automation and intelligent control. However, systems relying on RGB cameras for perception and robotic manipulators for control, common in smart farming, are vulnerable to photometric perturbations such as hue, illumination, and noise changes, which can cause malfunction under adversarial attacks. To address this issue, we propose an explainable adversarial-robust Vision-Language-Action model based on the OpenVLA-OFT framework. The model integrates an Evidence-3 module that detects photometric perturbations and generates natural language explanations of their causes and effects. Experiments show that the proposed model reduces Current Action L1 loss by 21.7% and Next Actions L1 loss by 18.4% compared to the baseline, demonstrating improved action prediction accuracy and explainability under adversarial conditions.

**Keywords:** Vision-Language-Action (VLA), Explainable Artificial Intelligence (XAI), Adversarial Robustness, Robotic Manipulation

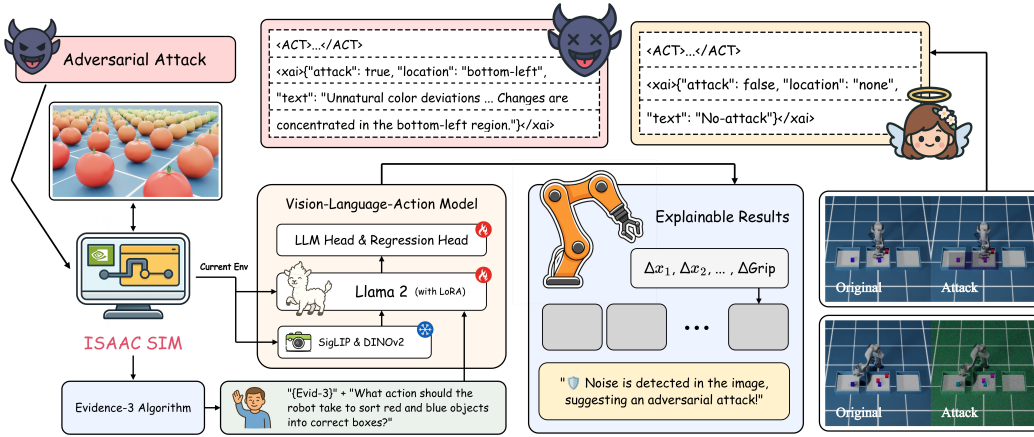


Figure 1: Overview of the proposed architecture.

## 1 Introduction and Proposed Method

Smart farming systems are complex intelligent systems that integrate various modules such as robots, sensors, and cameras. While Vision-Language-Action (VLA) research has advanced in

<sup>\*</sup>This research was supported by the Regional Innovation System & Education(RISE) program through the RISE Center, Gyeongsangnam-do, funded by the Ministry of Education(MOE) and the Gyeongsangnam-do Provincial Government, Republic of Korea. (2025-RISE-16-001)

<sup>†</sup>Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-8, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).

<sup>‡</sup>Corresponding author

processing multimodal data for environmental perception and control, studies on explainable artificial intelligence (XAI) for defending against adversarial attacks remain limited[1]. This paper proposes a VLA model capable of detecting and explaining adversarial attacks by integrating an adversarial detection and explanation module into the existing VLA framework, as illustrated in Figure 1. The following section outlines the proposed architecture.

1. **Adversarial Data Generation:** Simulation data are collected using the Franka Emika Panda robotic arm and an RGB camera in Isaac Sim. Random photometric transformations including hue shift ( $T_{\text{color}}$ ), illumination adjustment ( $T_{\text{illum}}$ ), and noise injection ( $T_{\text{noise}}$ ) are applied to generate adversarial variants. Formally, this process can be expressed as  $x' = T_S(x)$ ,  $T_S \subseteq \{T_{\text{color}}, T_{\text{illum}}, T_{\text{noise}}\}$ , where  $x$  denotes the original input image, and  $S$  is a randomly selected subset of transformations.
2. **Evidence-3 Module Integration:** The proposed architecture builds upon the OpenVLA-OFT framework [2] and incorporates an additional Evidence-3 module for adversarial attack detection. The Evidence-3 module consists of a detection pipeline based on three statistical metrics: HSV Mahalanobis Distance (detecting color distribution anomalies), High-Frequency Energy Ratio (identifying noise injection), and Local Entropy Standard Deviation (capturing spatial irregularities). These statistical cues are embedded into the user instruction and provided as auxiliary input to the model.
3. **Action Prediction and Explainable Model Training:** The action prediction head receives hidden representations from the Llama2 backbone and predicts the current and subsequent actions by minimizing the L1 loss. In parallel, the model is trained to detect and describe adversarial attacks by minimizing the cross-entropy loss over the XAI tokens generated by the Llama2 output. The total loss is defined as  $\mathcal{L}_{\text{total}} = \lambda_{\text{xai}}\mathcal{L}_{\text{xai}} + \mathcal{L}_{\text{act}}$ , where  $\mathcal{L}_{\text{xai}}$  represents the cross-entropy loss for explanation tokens, scaled by a weighting hyperparameter  $\lambda_{\text{xai}}$  that controls the relative importance of explanation learning, set to 0.5 in this work. Meanwhile,  $\mathcal{L}_{\text{act}}$  denotes the L1 regression loss for action prediction.

## 2 Experimental Results and Conclusion

To evaluate the proposed architecture, we compared three configurations: the baseline (Default), an adversarially trained model (Augmented), and the proposed model. Table 1 summarizes the results. Compared with the Default model, the proposed model reduced the Current and Next Action L1 losses by 21.6% and 18.4%, respectively, while outperforming the Augmented model by 6.9% and 7.8%, respectively. It also achieved an XAI token accuracy of 99.77%, showing that joint learning of robustness and explainability improves action prediction under adversarial conditions. Future work will explore the applicability of our approach to real-world smart farming environments and extend validation using robotic simulations.

Table 1: Performance evaluation results between the proposed and baseline models.

OpenVLA-OFT	XAI Token Accuracy (%)	Current Action L1 ( $\downarrow$ )	Next Actions L1 ( $\downarrow$ )
Default	-	0.0826	0.0788
Augmented	-	0.0695	0.0697
Proposed	<b>99.77</b>	<b>0.0647</b>	<b>0.0643</b>

## References

- [1] Y. Gao, S. A. Camtepe, N. H. Sultan, et al. Security threats to agricultural artificial intelligence: Position and perspective. *Computers and Electronics in Agriculture*, 227:109557, 2024.

- [2] M. J. Kim, C. Finn, and P. Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025.