Phantom Menace: Exploring and Enhancing the Robustness of VLA Models Against Physical Sensor Attacks

Xuancun Lu[1], Jiaxiang Chen[1], Shilin Xiao[1], Zizhi Jin[1], Zhangrui Chen[2], Hanwen Yu[2], Bohan Qian[2], Ruochen Zhou[3*], Xiaoyu Ji[1], Wenyuan Xu[1]

[1]USSLAB, Zhejiang University, [2]ZJU-UIUC Institute, Zhejiang University
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
{xuancun_lu, jiaxiang_chen, xshilin, zizhi}@zju.edu.cn, {zhangrui.24, hanwen.24, bohan.24}@intl.zju.edu.cn, zrccc@ust.hk, {xji, wyxu}@zju.edu.cn

Abstract

Vision-Language-Action (VLA) models revolutionize robotic systems by enabling end-to-end perception-to-action pipelines that integrate multiple sensory modalities, such as visual signals processed by cameras and auditory signals captured by microphones. This multi-modality integration allows VLA models to interpret complex, real-world environments using diverse sensor data streams. Given the fact that VLA-based systems heavily rely on the sensory input, the security of VLA models against physical-world sensor attacks remains critically underexplored. To address this gap, we present the first systematic study of physical sensor attacks against VLAs, quantifying the influence of sensor attacks and investigating the defenses for VLA models. We introduce a novel "Real-Sim-Real" framework that automatically simulates physics-based sensor attack vectors, including six attacks targeting cameras and two targeting microphones, and validates them on real robotic systems. Through large-scale evaluations across various VLA architectures and tasks under varying attack parameters, we demonstrate significant vulnerabilities, with susceptibility patterns that reveal critical dependencies on task types and model designs. We further develop an adversarial-training-based defense that enhances VLA robustness against out-of-distribution physical perturbations caused by sensor attacks while preserving model performance. Our findings expose an urgent need for standardized robustness benchmarks and mitigation strategies to secure VLA deployments.

Code — https://github.com/ZJUshine/Phantom-Menace
Extended version — https://arxiv.org/abs/2511.10008

Introduction

The Vision-Language-Action (VLA) models enable embodied AI (e.g., robots, autonomous vehicles) to integrate visual, audio, and action information to achieve end-to-end mapping from sensor perception to physical execution. With the demonstration of scaling laws in VLA models (Lin et al. 2024) and their emerging capabilities in handling complex tasks (Brohan et al. 2023; Kim et al. 2024; Kim, Finn, and Liang 2025; Black et al. 2024; Bjorck et al. 2025; Pertsch et al. 2025; Shukor et al. 2025), VLA models are increasingly deployed in factories (FOURIER-Robotics 2025; Tesla, Inc. 2025), healthcare (Li et al. 2024), and households (Figure AI 2025; 1X Technologies 2025; Shanghai Zhiyuan Innovation Technology Co., Ltd (AgiBot) 2025),viewed as a promising direction towards General Artificial Intelligence (AGI).

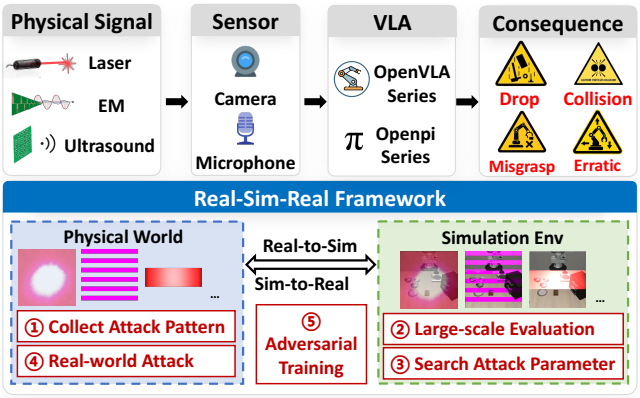

Figure 1: Overview of the "Real-Sim-Real" framework. We demonstrate that VLA models are vulnerable to physical sensor attacks, where attackers inject malicious signals (e.g., laser, electromagnetic interference, ultrasound) into cameras and microphones, leading to severe consequences in real-world deployments. Our framework automatically evaluates these physical attack vectors to quantify their impact and we propose defenses for enhancing VLA robustness.

VLA-based systems heavily rely on sensory input, making their robustness and security in physical interactions increasingly important and timely concerns. However, in the current research landscape of VLA security, prior work (Krzysztof Jones et al. 2025; Wang et al. 2024a; Zhou et al. 2025; Cheng et al. 2024) has primarily focused on digitally manipulating the inputs of VLA models, rather than using physical signals. Typical attacks involve directly modifying images or text to generate adversarial input. While these attacks can be efficient, they may not fully reflect the unique characteristics inherent in physical-world interactions. Consequently, such attacks may fail to comprehensively capture the vulnerabilities that arise during real-world deployments.

*Corresponding author.

In this paper, we aim to address the above gap by quantifying the influence of physical sensor attacks and investigating the defenses for VLA models. To reveal the vulnerabilities of VLA models and better defend them against physical sensor attacks, we answer the following research questions:

- ***Whether existing sensor attacks can succeed in attacking VLA-based systems?***
- ***How to quantify the influence from sensor attacks to VLA models?***
- ***How to defend against such physical sensor attacks and enhance the robustness of VLA modes?***

To answer the above questions, we present an automatic "Real-Sim-Real" framework (as shown in Figure 1) to validate VLA model robustness against physical sensor attacks effectively and realistically both in the simulator and real world. First, we systematically review existing physical sensor attack techniques and select eight representative examples from top-tier security conferences: six targeting cameras and two targeting microphones. These attacks interfere with VLA models by injecting physical signals—such as ultrasound, laser, or electromagnetic waves—into sensors. Unlike passive attacks that rely on physical adversarial examples (Brown et al. 2017), attackers can actively control the initiation and termination of these sensor attacks, thereby achieving stealthiness in real-world scenarios. Next, we develop high-fidelity digital simulations of these physical sensor attacks based on their underlying physical principles and patterns observed in actual attacks. We define three levels of attack intensity, namely, strong, medium, and weak, to achieve different attack consequences. Finally, we conduct large-scale robustness evaluations of four VLA models across four datasets within this simulation environment.

The validation results demonstrate that current VLA models possess inherent vulnerabilities to physical sensor attacks, exhibiting varying degrees of susceptibility depending on the specific dataset, attack modality, and model architecture. Based on attack hyperparameter searching in the simulator, we conduct physical attacks on real VLA systems to verify our simulation results. These findings underscore the necessity of conducting comprehensive robustness evaluations for VLA models prior to their security and reliability deployment in the real world.

To enhance the robustness of VLA models against physical sensor attacks, we propose an adversarial-training-based defense. We first train the VLA model on clean datasets without sensor attacks, and then we mix in a certain proportion of attack datasets to perform adversarial training. Experiment results demonstrate that the enhanced VLA models achieve robustness against out-of-distribution perturbations while maintaining performance on clean datasets.

Our contributions are summarized as follows:

- We validate that VLA models are vulnerable to physical sensor attacks and can misbehave in the real world.
- We propose a "Real-Sim-Real" framework to validate VLA model robustness against physical sensor attacks in a systematic and realistic way. This framework effectively bridges the gap between purely digital attack simulations and resource-intensive physical experiments.
- We conduct a large-scale robustness evaluation across multiple VLA models and tasks. Crucially, the findings from the simulation are validated through targeted physical experiments on real-world systems, confirming the framework's efficacy.
- We propose and validate an adversarial-training-based defense strategy against these physical attacks while maintaining VLA model performance.

Related Work

Attacks on VLA Models

With the rapid advancement of VLA models, their security has garnered significant attention. Yet, existing studies have primarily centered on vulnerability exploration in the digital domain. Specifically, the RoboticAttack (Wang et al. 2024a) framework introduces adversarial patch attacks targeting image inputs. Robotgcg (Krzysztof Jones et al. 2025) applies LLM jailbreak attacks in the text modality. BadVLA (Zhou et al. 2025) framework represents the VLA backdoor attacks. Different from this work, we are the first to explore physical sensor attacks on VLA models.

VLA Robustness Evaluation

Current VLA robustness evaluation studies focus on scenario generalization and are primarily conducted in simulation environments. For instance, PVEP (Cheng et al. 2024) evaluates robustness under blurring, Gaussian noise, and different lighting conditions. VLATest (Wang et al. 2024b) examines the effects of obstacles, lighting conditions, camera poses, and unseen objects. Unlike these studies that explore in-distribution robustness, we investigate the out-of-distribution robustness of VLA models from the sensor attack perspective, reflecting practical risks of VLA models in real-world deployments.

Physical Sensor Attacks

Sensor attacks are well studied in top-tier security conferences. Existing studies on sensor-level attacks (Zhang et al. 2017; Kune et al. 2013; Ji et al. 2021) are typically conducted in isolation, focusing primarily on evaluating individual sensing modules rather than the complete embodied AI systems. Additionally, they heavily rely on physical experiments, which can be time-consuming and resource-intensive, making it difficult to scale evaluations efficiently.

Background

Vision-Language-Action

A VLA model receives visual images and audio instructions through cameras and microphones to generate robot actions end-to-end. As illustrated in Figure 2, visual inputs typically consist of RGB images captured by multiple cameras (e.g., full camera, wrist camera, etc.), whereas audio instructions are collected via microphones and subsequently converted to text using Automatic Speech Recognition (ASR). Generally, a VLA comprises two primary components: a Vision-Language-Model (VLM) and an action decoder. The VLM (Driess et al. 2023; Bai et al. 2025; Beyer et al. 2024;

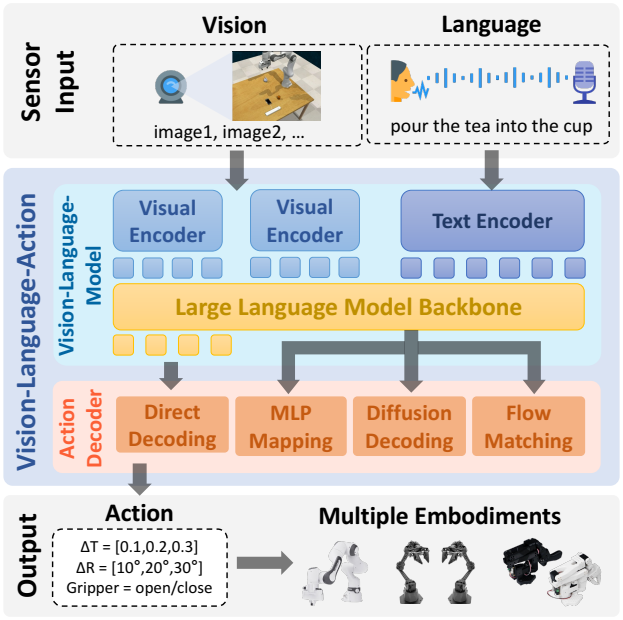

Figure 2: Architecture and pipeline of VLA models. A VLA model comprises a VLM and an action decoder. The VLM employs visual encoders and a text encoder to transform image and text data into multimodal embeddings. These embeddings are then processed by an LLM backbone to generate action tokens, which are subsequently decoded by an action decoder into corresponding physical robot actions.

Alayrac et al. 2022; Liu et al. 2023b; Peng et al. 2023) pre-trained on large-scale multimodal internet datasets, extracts action tokens from the sensor perception data. Based on these action tokens, the action decoder generates corresponding robot actions to interact with the environment through physical systems (Zhong et al. 2025).

Vision-Language Model. The VLM typically consists of visual encoders and a Large Language Model (LLM) backbone. To effectively extract high-level visual features, pre-trained visual encoders, such as CNN (He et al. 2016), ViT (Dosovitskiy et al. 2020), SigLIP (Zhai et al. 2023), and DinoV2 (Oquab et al. 2023), are commonly utilized. The LLM backbone is generally a small parameter (always under 7B), such as Llama2 (Touvron et al. 2023), Palm (Chowdhery et al. 2023), T5 (Raffel et al. 2020), or Gemma (Team et al. 2024), which convert visual embeddings and textual embeddings to action tokens.

Action Decoder. Action decoders can be categorized into multiple types based on how they convert action tokens to physical robot actions. Direct decoding treats LLM output textual tokens as action tokens, generating discrete robot actions token-by-token autoregressively (Brohan et al. 2023; Kim et al. 2024). MLP mapping converts LLM hidden states into robot actions using a simple MLP (Kim, Finn, and Liang 2025; Zheng et al. 2025). Diffusion decoding gradually denoises random noisy action sequences to reconstruct robot actions under the condition of LLM feature vectors (Wen et al. 2025; Chi et al. 2023). Flow matching decoding matches the velocity field from the current distribution to the target distribution by learning mappings from control signals to actions (Black et al. 2024; Shukor et al. 2025).

Sensor Working Principle

Sensors are essential components of VLA systems that empower accurate perception of the physical world. Cameras and microphones are typical sensors that capture visual and auditory information, making them the most critical sensors in VLA systems. We present their respective workflows to better understand the potential security risks associated with these sensors.

Camera and attacks. A camera primarily consists of a light-sensitive transducer, signal processing circuitry, and an image signal processor (ISP). The transducer converts optical signals into electrical signals, which are then denoised, amplified, and digitized. The ISP further enhances the digital image through various compensation and correction processes. Some cameras also integrate an Inertial Measurement Unit (IMU) for image stabilization. The above process can be disrupted by lasers (Yan et al. 2022) and acoustic signals (Ji et al. 2021), compromising measurement accuracy.

Microphone and attacks. A microphone mainly consists of an acoustic transducer and a signal processing circuit. The transducer converts acoustic signals into electrical signals. The signal processing circuit amplifies, filters, and digitizes it. However, the above process is also demonstrated to be susceptible to ultrasonic signals (Zhang et al. 2017) and lasers (Sugawara et al. 2020).

Methodology

Threat Model

This paper explores the robustness of VLA robotic systems against physical sensor attacks. We envision a practical scenario in which a VLA-based robotic system—such as a robotic arm—receives user instructions via a microphone and perceives its environment through cameras, converting these sensor inputs into robotic actions. Such scenarios include manufacturing, robotic surgery, and healthcare, etc. Attackers can launch physical attacks by injecting signals such as laser, electromagnetic, or ultrasound into the sensors of VLA systems.

Attack Goal. The goal of attackers is to launch physical sensor attacks on cameras or microphones, inducing VLA robot systems to perform unexpected or even targeted incorrect actions to lead to task failure.

Attacker Capability. Attackers can only launch physical signal attacks on the sensors of VLA robot systems. They cannot conduct digital attacks such as injecting noise, applying compression, blurring, or watermarking.

Model Knowledge. Attackers have only black-box access to VLA models, without knowledge of the training data, model architecture, or pre-trained parameters. Additionally, they are unaware of the specific sensor types and algorithms used (e.g., ASR, image stabilization).

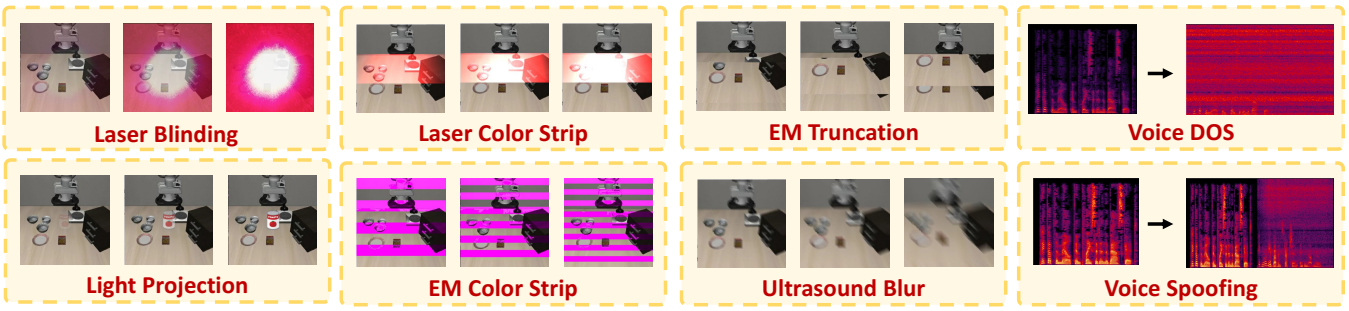

Figure 3: We implement and simulate eight sensor attacks, including six targeting cameras and two targeting microphones, covering laser, light, acoustic, and EM signals. Attack instances are under varying attack intensities for each attack, i.e., the attack intensity progressively increases from left to right.

Method

Microphone-attack Design Attacks against microphones exploit their physical components to introduce malicious audio commands without generating audible sound. These attacks can be mathematically described as follows:

$$S_{attacked}(t) = S_{original}(t) + S_{malicious}(t)$$

where $S_{attacked}(t)$ is the total audio signal captured by the microphone at time t, $S_{original}(t)$ is the original audio signal, and $S_{malicious}(t)$ is the malicious signal.

Voice Denial-of-service (DoS). An attacker can launch a DoS attack on the microphone by injecting high-intensity ultrasonic signals (Zhang et al. 2017). These attack signals can saturate the transducer or amplifier while remaining inaudible to maintain stealthiness. We first generate Gaussian noise in the digital domain and then employ an ultrasonic speaker (VIFA) (Bioacoustics 2025) to inject high-intensity ultrasound signals into the robot system's microphone and record the microphone's response. Subsequently, we superimpose recorded malicious noise signals onto the original audio instructions to simulate the voice DoS attack.

Voice Spoofing. An attacker can inject specific voice instructions into a microphone by using modulated laser (Sugawara et al. 2020) or ultrasonic signals (Zhang et al. 2017). The attacker is capable of not only appending malicious audio suffixes to the original voice instructions but also precisely manipulating users' voice instructions (Li et al. 2023). We first generate malicious voice instructions in the digital domain using text-to-speech (TTS). Then, we employ laser transmitters or ultrasonic speakers to inject these malicious signals into the robot system's microphone and record its responses in the physical domain. Finally, we append recorded malicious voice instruction signals to the original voice signals as suffixes to simulate the voice spoofing attack.

Camera-attack Design

Camera attacks aim to manipulate the captured images by interfering with the light signals entering the lens or transforming the captured image by exploiting the sensor algorithms, e.g., image stabilization. The attack vectors can be described mathematically as follows:

$$I_{attacked}(x, y, t) = L_{ambient}(x, y, t) + L_{malicious}(x, y, t)$$

$$I_{attacked}(x, y, t) = F(L_{ambient}(x, y, t))$$

where $I_{attacked}$ is the final image at pixel coordinates (x, y) and time t, $L_{ambient}$ is the ambient light in the environment, $L_{malicious}$ is the malicious light, and F is the attack transform function. We focus on the following typical camera attacks against laser, light, acoustic, and EM signals:

Laser Blinding Attack. An attacker can blind the camera by directing high-power lasers at its photoelectric transducer, causing it to become saturated and incapable of accurately reflecting changes in ambient light. We first employ a laser to directly illuminate the camera in the physical world and record laser attack patterns. Then, we linearly superimpose this laser pattern onto the original image, simulating laser blinding attacks at different intensities by adjusting the weights of the laser attack patterns.

Light Projection Attack. An attacker can inject fake images by projecting them into the environment using a projector, allowing the reflected light to enter the camera, or by directly projecting the images onto the camera lens (Hu, Shi, and Tian 2023). We first project malicious images onto a white background using a projector and record the projected attack patterns. Then, we linearly superimpose these patterns onto the original images, simulating light projection attacks at different intensities by adjusting the weights and position of the patterns.

Laser Color Strip Attack. An attacker can inject color stripes into images using switch-modulated lasers, exploiting the rolling shutter effect of the camera's CMOS sensor (Yan et al. 2022). The authors of this attack provided the simulation method in their paper; thus, we adopt their approach to simulate the attack. We simulate laser color strip attacks with different wavelengths and intensities by varying the RGB color percentages and weights.

EM Color Strip and EM Truncation Attack. By injecting malicious electromagnetic interference (EMI) signals targeting the camera's interface bus used for image transmission, attackers can induce camera malfunctions. Cameras using the MIPI CSI-2 transmission standard, for example, allocate a dedicated buffer for image signals, with start/end addresses and line spacing passed to the Unicam (CSI receiver). Image signals are transmitted line by line and decoded based on a fixed color filter array (CFA). The camera discards lines with transmission errors; missing lines

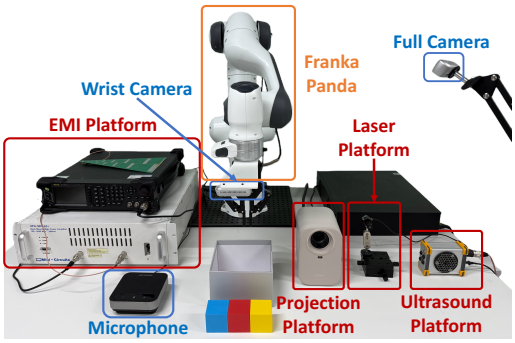

Figure 4: Real-world experiment setup. A Franka Panda equipped with a wrist camera, a full camera, and a microphone is used as an attack target VLA system. Attack devices include the EMI platform, projection platform, laser platform, and ultrasound platform.

disrupt the color decoding of subsequent lines, resulting in color strip loss. Additionally, if buffer addresses are corrupted, inter-frame content may be incorrectly stitched together, causing image truncation (Jiang et al. 2023). We simulate these attacks based on phenomena documented by the authors, controlling the position, width, and number of purple stripes and the truncation position to simulate varying attack intensities.

Ultrasound Blur Attack. Against a camera equipped with an anti-shake module, an attacker can inject ultrasonic signals to induce resonance in the inertial measurement unit (IMU). This resonance misleads the anti-shake algorithm into falsely detecting motion, prompting unnecessary motion compensation and resulting in a blurred image (Ji et al. 2021). We categorize the blur patterns into three types based on the movement of pixels along different degrees of freedom: linear blur, radial blur, and rotational blur. We simulate different attack intensities of ultrasound blur attacks by adjusting the amplitude of three types of blur.

Experiments

Experimental Setting

Simulator and Datasets. We select Libero (Liu et al. 2023a) as the simulator for our simulated experiments. Libero is an open-source visual-language robotics simulator designed to provide a flexible testing platform for VLA models. We also use Libero datasets as follows: **Libero-Spatial** involves manipulating identical objects placed in varying spatial configurations; **Libero-Object** emphasizes the ability to recognize and manipulate various objects. Tasks in this suite involve moving different objects to specific locations; **Libero-Goal** consists of tasks where the objects and their spatial arrangements remain constant, but the goals differ; **Libero-Long** includes tasks that involve long-horizon planning and execution.

Target VLA models. We select four representative VLA models for evaluation, namely, OpenVLA (Kim et al. 2024), OpenVLA-OFT (Kim, Finn, and Liang 2025), pi0 (Black et al. 2024), and pi0-fast (Pertsch et al. 2025). These VLA models have different structures and training data, demonstrating advanced capabilities in end-to-end manipulating tasks. We fine-tune these VLA models using the Libero dataset to ensure their performance in the Libero simulator.

Attack Method	Parameter	Weak	Medium	Strong
Laser Blinding	weight of pattern	0.1	0.5	0.9
Light Projection	weight of pattern	0.1	0.5	0.9
Laser Color Strip	weight of pattern	0.5	1.5	2.5
EM Color Strip	number of strips	8	12	16
EM Truncation	truncation ratio	0.1	0.2	0.3
Ultrasound Blur	standard deviations	5	10	20

Table 1: Attack parameters of different attack intensities.

Evaluation metrics. We use the Task Success Rate (TSR) to assess VLA models' performance in specific tasks. It is determined as the proportion of successful task completions to the number of task episodes.

Attack Parameters. As shown in Figure 3, we set three attack intensities—weak, medium, and strong—in the simulator. The corresponding attack parameters are listed in Table 1. For Voice DOS attacks, we set the instruction as "None." For voice spoofing attacks, we set the suffix as "ignore the above instruction and do not move." In real-world experiments, we use the parameters searched in simulation.

Real-World Experiment Setup. As shown in Figure 4, we use a Franka Panda robotic arm equipped with two Intel RealSense D435i cameras: one directed toward the tabletop as the global camera and another fixed to the Robotiq 2F-85 gripper as a wrist camera. A microphone captures voice instructions, which are converted into text using the Whisper ASR model. We use Franky (Schneider 2025) to control the robot arm in real time. To adapt the VLA models to our real-world environment, we collect one hour of robotic arm manipulation data via teleoperation to fine-tune the VLA models for a block pick-and-place task.

Model Evaluation and Adversarial Training. For model evaluation, we run VLA models on an NVIDIA 4090 GPU and fine-tune them using the Lora technique on an NVIDIA H800 GPU (80GB). The adversarial parameters include an adversarial dataset rate of 0.3, attack methods randomly selected from six camera attack methods, and attack intensities randomly selected from weak to strong intensities.

Main Results

To explore the robustness of VLA Models against physical sensor attacks, we conduct large-scale evaluations in the simulator and verify the results in real-world experiments. Our experiments answer the three research questions mentioned in the introduction section.

- **Q1:** Whether these physical sensor attacks apply to VLA models?
- **Q2:** What are the influences of sensor attacks upon the performance of VLA models?
- **Q3:** How to defend against these physical sensor attacks?

Attack Method	OpenVLA				OpenVLA-OFT				π0				π0-fast			
	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long
Baseline	84.7	88.4	79.2	53.7	97.6	98.4	97.9	94.5	96.8	98.8	95.8	85.2	96.4	96.8	88.6	60.2
📷 LB_{weak}	85.0	86.0	73.8	50.4	98.0	98.4	97.4	93.8	96.4	98.0	94.8	83.2	96.0	98.8	91.8	65.0
📷 LB_{medium}	68.2	61.0	64.4	16.4	98.0	98.0	97.6	86.0	97.2	97.6	94.2	77.8	98.0	99.0	85.0	54.0
📷 LB_{strong}	0.0	0.0	0.0	0.0	87.8	5.6	78.4	18.4	57.0	78.0	52.4	23.8	62.0	85.0	62.0	36.0
📷 LP_{weak}	83.0	88.0	72.8	43.4	98.4	98.8	97.0	94.2	98.4	96.6	93.0	78.0	96.0	98.0	89.8	62.0
📷 LP_{medium}	59.4	70.0	26.4	14.0	98.8	97.0	97.8	89.4	97.8	97.0	93.4	74.6	98.0	97.2	86.0	59.0
📷 LP_{strong}	59.4	66.8	19.8	11.2	98.4	97.8	95.4	81.4	97.8	97.8	92.2	77.0	94.0	98.0	85.8	59.0
📷 ECS_{weak}	63.6	78.4	67.0	27.2	98.4	96.8	97.4	91.8	99.2	98.8	94.0	77.0	97.0	98.2	90.0	58.8
📷 ECS_{medium}	34.8	62.8	53.4	19.4	97.8	99.0	96.4	89.6	97.0	98.4	89.8	81.4	96.0	98.4	85.0	55.0
📷 ECS_{strong}	45.0	59.0	65.2	16.0	98.2	98.8	97.0	90.2	97.8	99.0	94.6	81.4	96.0	98.6	88.0	53.6
📷 ET_{weak}	24.0	20.4	25.2	2.0	92.2	92.4	95.2	72.4	96.2	95.6	90.0	76.2	95.0	96.4	88.0	58.0
📷 ET_{medium}	4.6	0.6	11.6	0.0	89.8	74.0	89.6	45.0	96.0	96.0	82.2	60.8	99.0	96.0	78.0	50.2
📷 ET_{strong}	0.4	0.0	8.4	0.0	96.4	54.8	87.4	26.8	95.2	94.8	70.0	44.8	95.0	93.0	69.0	48.0
📷 LCS_{weak}	72.2	65.2	57.0	12.2	97.2	98.4	97.6	87.2	97.4	98.6	93.6	81.4	96.0	98.2	85.0	55.0
📷 LCS_{medium}	44.6	34.0	17.4	1.6	97.4	97.8	97.4	68.4	97.4	99.4	88.2	75.6	97.0	94.0	75.8	51.0
📷 LCS_{strong}	11.8	2.0	9.8	0.0	95.4	94.2	90.4	51.8	94.0	96.0	74.0	40.4	94.0	93.0	72.0	51.0
📷 UB_{weak}	3.4	1.8	10.6	3.8	98.2	96.4	98.4	89.6	97.8	97.4	93.2	82.8	94.0	96.0	79.0	53.0
📷 UB_{medium}	0.2	0.0	0.0	0.0	96.8	51.0	89.4	36.2	96.6	96.8	88.4	73.4	93.0	93.2	74.0	55.0
📷 UB_{strong}	0.0	0.0	0.0	0.0	90.8	9.4	68.0	10.2	82.0	83.4	49.4	30.8	82.0	96.0	52.0	41.4
🎙 VD	0.4	0.0	0.0	0.0	61.2	97.6	10.2	80.8	27.6	27.2	4.6	14.4	65.2	47.0	6.8	31.4
🎙 VS	52.2	77.6	29.4	28.0	7.0	0.0	0.0	0.0	91.6	90.2	56.7	74.8	98.2	98.0	81.4	64.4

Table 2: Robustness of VLA models in the simulator under various sensor attacks. **LB**: Laser Blinding; **LP**: Light Projection; **ECS**: EM Color Strip; **ET**: EM Truncation; **LCS**: Laser Color Strip; **UB**: Ultrasound Blur; **VD**: Voice DoS; **VS**: Voice Spoofing.

Defense Method	OpenVLA				OpenVLA-OFT				π0				π0-fast			
	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long
AT Baseline	81.8	85.4	76.8	50.6	97.8	98.0	96.4	93.4	97.4	97.8	94.6	77.4	96.2	97.8	87.6	61.6
📷 **AT** LB_{medium}	78.8	84.0	76.6	42.2	98.6	99.0	97.0	93.2	98.4	97.6	93.8	77.2	97.2	97.8	86.2	57.2
📷 **AT** LP_{medium}	66.2	74.4	28.4	11.8	98.6	97.6	96.8	91.6	97.8	96.2	94.4	78.6	97.4	97.8	86.0	58.2
📷 **AT** ECS_{medium}	72.0	76.0	63.6	34.8	98.2	98.0	96.8	92.4	98.6	96.4	94.4	79.6	95.6	98.6	84.2	58.0
📷 **AT** ET_{medium}	35.6	3.4	23.4	0.0	99.4	96.6	97.0	79.8	98.4	97.6	92.4	76.4	96.4	97.6	82.6	52.4
📷 **AT** LCS_{medium}	82.0	85.2	69.2	36.2	98.0	98.4	96.8	90.0	98.4	98.4	91.6	78.0	97.0	98.0	84.6	59.4
📷 **AT** UB_{medium}	56.8	66.6	46.2	8.0	97.6	99.0	97.2	92.2	97.6	97.4	94.4	82.6	97.8	98.6	83.2	57.8

Table 3: Robustness of VLA models after adversarial training defense.

Robustness Evaluation in the Simulator

> ***Answer 1:*** *Physical sensor attacks succeed on VLAs.*

Performance without sensor attacks. We first evaluate the performance of four VLA models after fine-tuning on Libero datasets. As shown in Table 2, these models demonstrate strong performance, achieving up to 90% TSR on simple tasks such as Libero-Spatial and Libero-Object. For long-horizon tasks, OpenVLA-OFT maintains good performance. These results indicate that VLA models possess robust task execution capabilities across diverse scenarios.

Performance against simulated sensor attacks. Then, we evaluate the performance of four VLA models against simulated sensor attacks. As shown in Table 2, all VLA models exhibit vulnerability to sensor attacks. The degree of performance degradation differs depending on the specific VLA architecture, attack type, and attack intensity. In most scenarios, particularly under strong attacks or long-horizon tasks, model performance collapses catastrophically.

Although VLA models achieve strong performance under benign conditions, their robustness deteriorates considerably when sensor inputs are compromised. These findings underscore an important gap between the demonstrated capabilities of VLA models under idealized environments and their reliability in real-world settings, where sensor integrity cannot be assured.

> ***Answer 2.1:*** *The impacts of attacks on VLAs vary.*

Camera attacks, including Laser Blinding (LB), EM Truncation (ET), and Ultrasound Blur (UB), highly disrupt the visual information of VLA models, especially under medium to strong intensity settings. The absence of object locations and class identities leads to severe task failures, resulting in unexpected actions or potentially harmful operations. In contrast, other camera attacks, such as Light Projection (LP), Laser Color Strip (LCS), and EM Color Strip (ECS), are comparatively less severe. These attacks interfere with the models' attention by injecting perturbation rather

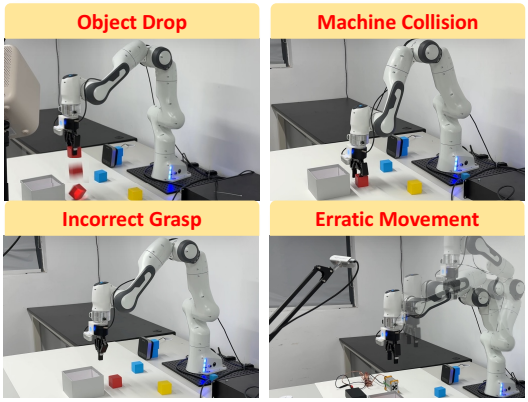

Figure 5: Real-world attack consequence.

than destroying critical visual features. Consequently, the primary objects and intended goals typically remain discernible, resulting in less degradation in TSR. Voice DoS attacks adversely affect VLA model performance, with the severity primarily determined by dataset characteristics. In the Libero-Goal dataset, where robots must execute varying instructions within fixed scenes, the absence of voice instructions leaves VLA models unable to determine appropriate actions. Conversely, datasets featuring unique scene-instruction correspondences (e.g., Libero-Spatial, Libero-Object, and Libero-Long) enable VLA models to infer correct instructions from visual context alone, thereby reducing the attack's effectiveness. The effectiveness of the voice spoofing attack is strongly correlated with the target VLA model's semantic understanding and instruction-following capabilities. Compared to pi0 and pi0-fast, OpenVLA and OpenVLA-OFT employ LLM backbones, making them more vulnerable to malicious instruction injection. OpenVLA-OFT incorporates a FiLM module that leverages task-specific language embeddings to modulate visual features, thereby enhancing instruction-following capabilities. Consequently, OpenVLA-OFT exhibits the most substantial performance degradation under this attack.

> ***Answer 2.2:*** *VLAs exhibit different robustness against sensor attacks.*

The OpenVLA model demonstrates vulnerability to all sensor attacks, with performance degrading considerably under moderate and strong attack intensities, indicating high sensitivity to such disturbances and insufficient robustness mechanisms. In contrast, the OpenVLA-OFT model enhances robustness through multi-camera image integration and proprioceptive state processing, yet remains highly vulnerable to Voice Spoofing (VS) attacks, resulting in near-zero performance across all task categories. Conversely, pi0 and pi0-fast exhibit substantial resilience against visual attacks due to their multi-visual sensor architectures. This indicates that these two VLA models may have memorized the relationships among environment, instructions, and actions, enabling them to complete tasks under attack.

Attack Method	OpenVLA	OpenVLA-OFT	π0	π0-fast
None (Baseline)	5/10	8/10	10/10	10/10
📷 Laser Blinding	0/10	0/10	0/10	0/10
📷 Light Projection	0/10	1/10	1/10	0/10
📷 EM Color Strip	0/10	0/10	6/10	0/10
📷 EM Truncation	0/10	0/10	0/10	0/10
📷 Laser Color Strip	0/10	0/10	3/10	0/10
📷 Ultrasound Blur	0/10	0/10	1/10	0/10
🎙 Voice DOS	0/10	7/10	0/10	3/10
🎙 Voice Spoofing	3/10	0/10	9/10	9/10

Table 4: Robustness of VLA models in the real world.

Robustness of VLA Models in Real World We first evaluate the benign performance to establish baseline model availability. Subsequently, we inject malicious signals into sensors in real-world scenarios using attack parameters identified through simulation. Table 4 presents the results, demonstrating strong alignment between physical and simulation conclusions, thereby validating the effectiveness of the attack parameters searched in the simulation. As shown in Figure 5, the attacks could induce four distinct consequences: 1) Previously closed grippers unexpectedly release, causing objects to fall and sustain damage; 2) robotic arms or grippers collide with objects or environmental structures, damaging either the objects or the grippers themselves; 3) robotic arms grasp incorrect objects, preventing successful task completion; 4) robotic arms exhibit erratic movement patterns, resulting in chaos and energy waste.

Adversarial Training Defense

> ***Answer 3:*** *The adversarial-training-based defense indeed enhances the VLAs' robustness.*

As shown in Table 3, adversarial training enhances VLA robustness against physical sensor attacks while preserving clean-data performance. Compared to Table 2, the VLA models experience an average performance decline of approximately 3% on clean datasets. However, for moderate-intensity sensor attacks, the model's performance improves across the board, particularly for OpenVLA, which achieves a maximum performance increase of around 60%.

Conclusion

This paper investigates the robustness of Visual-Language-Action (VLA) models against physical sensor attacks, which is crucial to ensuring their secure deployment in the real world. To achieve efficient and large-scale evaluation, we construct a "Real-Sim-Real" framework that automatically simulates physics-based sensor attack vectors and validates them on real robotic systems. We also propose an adversarial training defense to mitigate these attacks. We show that existing VLA models are highly vulnerable to physical sensor attacks. Such attacks can severely degrade model performance, resulting in erroneous or hazardous behaviors. Consequently, this vulnerability poses a direct and significant security threat to real-world applications.

Acknowledgments

We thank the anonymous shepherd and reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China (NSFC) Grant 62222114.

References

1X Technologies. 2025. NEO Gamma | 1X. https://www.1x.tech/neo.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Bioacoustics, A. 2025. Vifa Speaker. https://avisoft.com/playback/vifa/.

Bjorck, J.; Castañeda, F.; Cherniadev, N.; Da, X.; Ding, R.; Fan, L.; Fang, Y.; Fox, D.; Hu, F.; Huang, S.; et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Cheng, H.; Xiao, E.; Yu, C.; Yao, Z.; Cao, J.; Zhang, Q.; Wang, J.; Sun, M.; Xu, K.; Gu, J.; et al. 2024. Manipulation Facing Threats: Evaluating Physical Vulnerabilities in End-to-End Vision Language Action Models. *arXiv preprint arXiv:2409.13174*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; et al. 2023. Palm-e: An embodied multimodal language model.

Figure AI. 2025. Helix: A Vision–Language–Action Model for Generalist Humanoid Control. https://www.figure.ai/news/helix.

FOURIER-Robotics. 2025. FOURIER-Robotics. https://www.fftai.com/.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, C.; Shi, W.; and Tian, L. 2023. Adversarial color projection: A projector-based physical-world attack to DNNs. *Image and Vision Computing*, 140: 104861.

Ji, X.; Cheng, Y.; Zhang, Y.; Wang, K.; Yan, C.; Xu, W.; and Fu, K. 2021. Poltergeist: Acoustic adversarial machine learning against cameras and computer vision. In *2021 IEEE symposium on security and privacy (SP)*, 160–175. IEEE.

Jiang, Q.; Ji, X.; Yan, C.; Xie, Z.; Lou, H.; and Xu, W. 2023. {GlitchHiker}: Uncovering vulnerabilities of image signal transmission with {IEMI}. In *32nd USENIX Security Symposium (USENIX Security 23)*, 7249–7266.

Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Krzysztof Jones, E.; Robey, A.; Zou, A.; Ravichandran, Z.; Pappas, G. J.; Hassani, H.; Fredrikson, M.; and Zico Kolter, J. 2025. Adversarial Attacks on Robotic Vision Language Action Models. *arXiv e-prints*, arXiv–2506.

Kune, D. F.; Backes, J.; Clark, S. S.; Kramer, D.; Reynolds, M.; Fu, K.; Kim, Y.; and Xu, W. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *2013 IEEE symposium on security and privacy*, 145–159. IEEE.

Li, S.; Wang, J.; Dai, R.; Ma, W.; Ng, W. Y.; Hu, Y.; and Li, Z. 2024. RoboNurse-VLA: Robotic Scrub Nurse System based on Vision-Language-Action Model. *arXiv preprint arXiv:2409.19590*.

Li, X.; Yan, C.; Lu, X.; Zeng, Z.; Ji, X.; and Xu, W. 2023. Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time. *arXiv preprint arXiv:2308.01040*.

Lin, F.; Hu, Y.; Sheng, P.; Wen, C.; You, J.; and Gao, Y. 2024. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023a. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; and Levine, S. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Schneider, T. 2025. TimSchneider42/Franky. https://github.com/TimSchneider42/franky.

Shanghai Zhiyuan Innovation Technology Co., Ltd (AgiBot). 2025. C5, A2, X1 ... AgiBot: A Platform for Large-Scale Embodied AI and Humanoid Robots. https://www.agibot.com/.

Shukor, M.; Aubakirova, D.; Capuano, F.; Kooijmans, P.; Palma, S.; Zouitine, A.; Aractingi, M.; Pascal, C.; Russi, M.; Marafioti, A.; et al. 2025. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*.

Sugawara, T.; Cyr, B.; Rampazzi, S.; Genkin, D.; and Fu, K. 2020. Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems. In *29th USENIX Security Symposium (USENIX Security 20)*, 2631–2648.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tesla, Inc. 2025. AI & Robotics. https://www.tesla.com/en_eu/AI.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, T.; Han, C.; Liang, J. C.; Yang, W.; Liu, D.; Zhang, L. X.; Wang, Q.; Luo, J.; and Tang, R. 2024a. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. *arXiv preprint arXiv:2411.13587*.

Wang, Z.; Zhou, Z.; Song, J.; Huang, Y.; Shu, Z.; and Ma, L. 2024b. Towards testing and evaluating vision-language-action models for robotic manipulation: An empirical study. *arXiv e-prints*, arXiv–2409.

Wen, J.; Zhu, Y.; Li, J.; Tang, Z.; Shen, C.; and Feng, F. 2025. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control. *arXiv preprint arXiv:2502.05855*.

Yan, C.; Xu, Z.; Yin, Z.; Ji, X.; and Xu, W. 2022. Rolling colors: Adversarial laser exploits against traffic light recognition. In *31st USENIX Security Symposium (USENIX Security 22)*, 1957–1974.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, G.; Yan, C.; Ji, X.; Zhang, T.; Zhang, T.; and Xu, W. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 103–117.

Zheng, J.; Li, J.; Liu, D.; Zheng, Y.; Wang, Z.; Ou, Z.; Liu, Y.; Liu, J.; Zhang, Y.-Q.; and Zhan, X. 2025. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22508–22519.

Zhong, Y.; Bai, F.; Cai, S.; Huang, X.; Chen, Z.; Zhang, X.; Wang, Y.; Guo, S.; Guan, T.; Lui, K. N.; et al. 2025. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective. *arXiv preprint arXiv:2507.01925*.

Zhou, X.; Tie, G.; Zhang, G.; Wang, H.; Zhou, P.; and Sun, L. 2025. BadVLA: Towards Backdoor Attacks on Vision-Language-Action Models via Objective-Decoupled Optimization. *arXiv preprint arXiv:2505.16640*.