Exploring the Robustness of Vision-Language-Action Models against Sensor Attacks

Xuancun Lu
Zhejiang University
HangZhou, Zhejiang, China
xuancun_lu@zju.edu.cn

Jiaxiang Chen
Zhejiang University
HangZhou, Zhejiang, China
jiaxiang_chen@zju.edu.cn

Shilin Xiao
Zhejiang University
HangZhou, Zhejiang, China
xshilin@zju.edu.cn

Zizhi Jin
Zhejiang University
HangZhou, Zhejiang, China
zizhi@zju.edu.cn

Ruochen Zhou
Hong Kong University of Science and Technology
Hong Kong, Hong Kong, China
zrccc@ust.hk

Xiaoyu Ji*
Zhejiang University
HangZhou, Zhejiang, China
xji@zju.edu.cn

Wenyuan Xu
Zhejiang University
HangZhou, Zhejiang, China
wyxu@zju.edu.cn

Abstract

The Vision-Language-Action (VLA) models enhance the ability of robots to perform complex tasks and are widely considered a promising pathway toward achieving Artificial General Intelligence (AGI). By integrating visual, language, and action modalities, VLA models aim to establish an end-to-end strategy from raw sensor perception to corresponding robotic actions. Under this direct perception-to-action mechanism, robust sensor perception of the physical world becomes a primary prerequisite for secure and reliable deployment in the real world. While recent studies on the robustness and security of VLA models have primarily focused on digital input perturbations, the impact of physical-world sensor interference remains underexplored. To bridge this gap, we conduct an automatic and end-to-end framework to evaluate the robustness of VLA models against sensor attacks. Specifically, we review existing physical attack techniques targeting sensors commonly used in VLA systems and simulate eight typical sensor attacks (six targeting cameras and two targeting microphones). We then conduct large-scale robustness evaluations of VLA models with diverse architectures, considering multiple tasks and varying attack intensities. Experimental results indicate that current VLA models can be sensitive to sensor attacks, with the degree of vulnerability varying across tasks and model types. Our findings highlight the urgent need for systematic robustness assessment and potential mitigation strategies to ensure the reliable deployment of VLA systems in real-world environments. For more information, please visit our project at the following link: https://zjushine.github.io/lamps-vla-robustness.github.io/

CCS Concepts

• **Computing methodologies → Vision for robotics**; **Information extraction**; **Intelligent agents**.

Keywords

Vision-Language-Action, Robustness, Sensor Attack

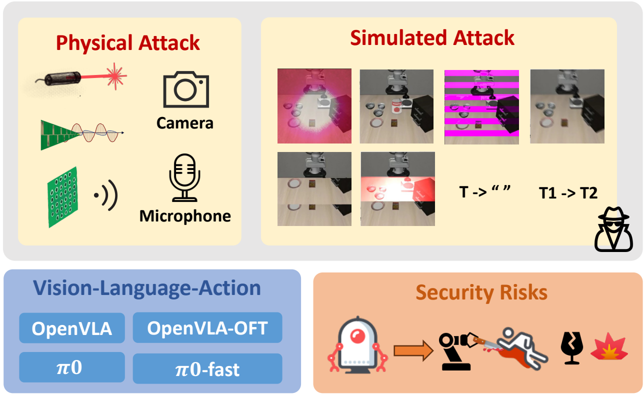

Figure 1: Exploring the robustness of VLAs against sensor attacks. The VLA models utilize cameras and microphones to perceive the environment and generate end-to-end actions of robot entities. Attackers exploit physical signals such as sound, light, and EM to inject attacks into sensors, thereby leading to unexpected actions and even potential harm in real-world scenarios.

1 Introduction

The Vision-Language-Action (VLA) models enable embodied AI (e.g., robots, autonomous vehicles) to integrate visual, language, and action information, achieving end-to-end mapping from sensor perception to physical execution. With the demonstration of scaling laws in VLA models [24] and their emerging capabilities in handling complex tasks [5–7, 19, 20, 29, 32], VLA models are increasingly viewed as a promising direction towards General Artificial Intelligence (AGI).

As VLA models see broader deployment in real-world applications, including in factories [14, 35], healthcare [22], and households [1, 13, 31], ensuring their robustness and security in physical

[*] Xiaoyu Ji is the corresponding author.

interactions has become an important and timely area of concern. However, there are two challenges that remain in the current research landscape: **(1) Limitations of existing VLA attack research:** Prior work [8, 21, 37, 42] has primarily focused on digitally manipulating inputs of VLA models, such as modifying images or text, which may not fully reflect the unique characteristics posed by physical-world interactions. As a result, these attacks may not comprehensively capture vulnerabilities during real-world deployment scenarios. **(2) Limitations of existing sensor attack research:** Existing studies on sensor-level attacks 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.

To address the challenges outlined above, we propose an automatic, end-to-end framework for evaluating the robustness of VLA models against sensor attacks. First, we systematically review existing physical sensor attack techniques and select eight representative instances—six targeting cameras and two targeting microphones. Next, we develop high-fidelity digital simulations of these sensor attacks based on their fundamental principles, including attack distance, attack method, and attack intensity. Finally, we conduct efficient, large-scale robustness evaluations of four VLA models across four tasks within the simulated environment.

Our experimental results demonstrate that current VLA models possess inherent vulnerabilities to sensor attacks, exhibiting varying degrees of susceptibility depending on the specific task, attack modality, and model architecture. These findings underscore the necessity of conducting comprehensive robustness evaluations for VLA models prior to their security and reliability deployment in the real world.

Our contributions are summarized as follows:

- We survey existing sensor attack methods targeting cameras and microphones and develop high-fidelity digital simulations.
- We extensively evaluate the robustness of VLA models against sensor attacks, revealing the vulnerabilities of existing VLA models in the face of physical threats.

2 Background

2.1 Vision-Language-Action

A VLA model receives visual inputs and textual instructions through cameras and microphones, respectively, to generate robot actions in an end-to-end manner. As illustrated in Figure 2, visual inputs for VLAs typically consist of RGB images captured by multiple cameras (e.g., head camera, wrist camera, etc.), whereas voice instructions are collected via microphone 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 [2–4, 12, 26, 28] pre-trained on large-scale multimodal datasets, extracts multimodal feature vectors from the sensor inputs. Based on these extracted feature vectors, the action decoder produces corresponding robot actions, which are executed through robotic arms interacting with the environment.

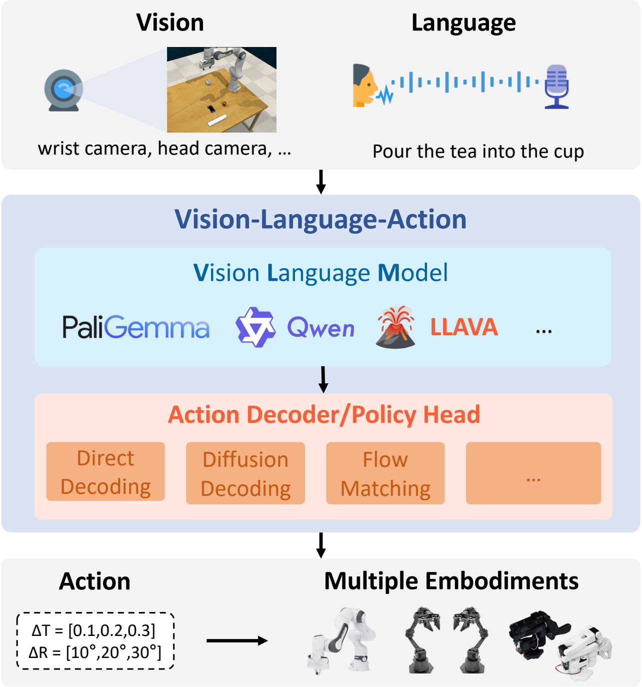

Figure 2: Architecture and pipeline of Vision-Language-Action (VLA) models. A VLA model consists of a Vision-Language Model (VLM) and an action decoder component. The VLM takes image and text data as inputs and generates multimodal embeddings, which capture semantic relationships between visual information and textual instructions. Subsequently, the action decoder leverages these multimodal embeddings to generate corresponding robot actions.

2.1.1 ***Vision-Language Model.*** The VLM typically consists of visual encoders and a backbone network. To effectively extract high-level visual features, pretrained visual encoders with frozen parameters, such as CNN [15], ViT [11], SigLIP [40], and DinoV2 [27], are commonly utilized. The backbone network is generally a transformer-based large language model, including Llama2 [36], Palm [10], T5 [30], or Gemma [34], which performs cross-modal fusion to understand environmental context and natural language instructions at the same time.

2.1.2 ***Action Decoder.*** Action decoders can be categorized into three types based on their architectures: direct decoding, diffusion-based decoding, and flow matching-based decoding. Direct decoding treats the robot's action sequence as an LLM token sequence, generating actions token-by-token in a self-regressive manner [20]. Diffusion-based decoding gradually denoises a random noisy action sequence to reconstruct coherent action trajectories [9, 38]. Flow matching-based decoding matches the velocity field from the current state to the target action distribution directly by learning mappings from control signals to actions [6, 32].

2.2 Sensor Working Principle

Sensors are essential components of embodied AI systems that empower accurate perception of the physical world. Among them, cameras and microphones capture visual and auditory information, making them the most critical sensors in VLAs. To better understand the potential security risks associated with these sensors, we present their respective workflows.

2.2.1 Camera. A camera mainly consists of a light-sensitive transducer, a signal processing circuit, and an image signal processor (ISP). The transducer converts optical signals into electrical signals, which are then denoised, amplified, and digitized by the signal processing circuit. The ISP further refines the digital image through compensation and correction processes. Additionally, some cameras are associated with an Inertial Measurement Unit (IMU) to enable image stabilization. The above process can be interfered with by laser [39] and acoustic signals [17], which can affect the measurements.

2.2.2 Microphone. 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 demonstrated to be susceptible to ultrasonic signals [41] and lasers [33].

3 Threat Model

This paper investigates the robustness of VLA robot systems under sensor attacks. We envision a practical scenario where a VLA robot system, such as a robotic arm, receives user instructions via microphone and perceives its environment through cameras, subsequently translating these inputs into robot actions. While we focus on physical-domain sensor attacks, the experiments in this study are conducted via simulated perturbations that approximate real-world attack effects on sensor outputs.

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

Attack Capability. Attackers can only launch physical attacks on the sensors of VLA robot systems and are unable to carry out attacks in digital form, such as noise, compression, blurring, cropping and rotation, color transformation, watermarking, etc.

Model Knowledge. Attackers have no knowledge of VLA models, including training data, model architecture, pre-trained parameters, etc. Meanwhile, attackers also have black-box knowledge of specific sensor types and sensor algorithms (e.g., ASR, anti-shake).

Simulation-Based Assumption. While the attack modalities considered in this work (e.g., laser interference or ultrasonic injection) are grounded in real-world demonstrations, we implement them through high-fidelity simulations. Specifically, we approximate the effects of these physical attacks by applying perturbations directly to the sensor outputs. This approach enables scalable and repeatable evaluation while preserving the essential characteristics of real-world sensor interference.

4 Design

4.1 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:

$$S_{attack}(t) = S_{original}(t) + S_{injected}(t)$$

where $S_{attack}(t)$ is the total signal captured by the microphone at time t, $S_{original}(t)$ is the original audio signal, and $S_{injected}(t)$ is the injected noise or instruction signal.

Attacks on microphones can be broadly categorized into several types, each with its own method of manipulating the audio signal captured by the microphone. The following sections describe these methods in detail.

4.1.1 Voice Denial-of-service (DoS).

Attack Principle. An attacker can launch a DoS attack on the microphone by injecting high-intensity ultrasonic signals [41]. These attack signals can saturate the transducer or amplifier while remaining inaudible to maintain stealth.

Simulation Method. We first generate Gaussian noise in the digital domain and then employ a signal generator, power amplifier, and ultrasonic speaker to transmit high-intensity ultrasound signals in the physical world. Afterward, we inject these ultrasound signals into the robot system's microphone and record the microphone's response. Finally, we superimpose these malicious noise signals onto the original audio instructions to simulate the noise attack.

4.1.2 Voice Spoofing.

Attack Principle. An attacker can inject specific voice instructions into a microphone by using modulated laser [33] or ultrasonic signals [41]. The attacker is capable of not only appending malicious audio suffixes to the original voice instructions but also precisely manipulating users' voice instructions [23].

Simulation Method. We first generate malicious voice instructions in the digital domain using text-to-speech (TTS). Then, in the physical domain, we employ laser transmitters or ultrasonic speakers to inject these malicious signals into the robot system's microphone and record its responses. Finally, we append these malicious voice instruction signals as suffixes to the original voice signals to simulate voice spoofing attacks.

4.2 Camera Attack Design

Attacks on cameras aim to manipulate the captured image by interfering with the light entering the lens. These attacks exploit the fundamental way a camera perceives its environment. A primary attack vector is introducing malicious light signals, which can be described mathematically as follows:

$$I_{attack}(x, y, t) = I_{ambient}(x, y, t) + L(x, y, t)$$

where I_{attack} is the total light captured by the camera at pixel coordinates (x, y) and time t, $I_{ambient}$ is the ambient light in the scene, and L is the malicious light introduced by the attacker.

Attacks on cameras can be broadly categorized into several types, each with its own method of manipulating the light captured by the camera.

4.2.1 *Laser Blinding.*

Attack Principle. 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.

Simulation Method. We first employ a laser to directly illuminate the camera in the physical world and record the laser attack pattern. 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 pattern.

4.2.2 *Light Projection.*

Attack Principle. 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 [16].

Simulation Method. We first project malicious images onto a white background using a projector and capture the attack patterns with the robot system's camera. Then, we linearly superimpose these patterns onto the original images. By adjusting the weights and position of the patterns, we simulate light projection attacks at different intensities and from various distances.

4.2.3 *Laser Color Strip.*

Attack Principle. An attacker can inject color stripes into captured images by using switch-modulated lasers, exploiting the rolling shutter effect of the camera's CMOS sensor [39].

Simulation Method. The authors of this attack provided the simulation method in their paper; thus, we adopt their approach to simulate the attack. By varying the RGB color percentages and weights, we simulate laser color strip attacks with different wavelengths and intensities.

4.2.4 *EM Color Strip and EM Truncation.*

Attack Principle. By injecting malicious signals via intentional electromagnetic interference (IEMI) targeting the camera's interface bus used for image signal transmission, an attacker can induce camera malfunction. Cameras employing the MIPI CSI-2 transmission standard allocate a dedicated buffer for image signals, with the start/end addresses and line spacing passed to the Unicam (the CSI receiver). Image signals are transmitted line by line and decoded based on a fixed color filter array. Lines with transmission errors are discarded by the camera; if a line is missing, it disrupts the color decoding of subsequent lines, resulting in the loss of color strips. Moreover, if the start or end addresses of the buffer are corrupted, inter-frame content may be incorrectly stitched together, leading to image truncation [18].

Simulation Method. We simulate EM color strip and EM truncation based on the phenomena presented by the authors in their paper. By adjusting the attack signal, we can control the position, width, number of purple stripes, and the truncation position.

4.2.5 *Ultrasound Blur.*

Attack Principle. 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 [17].

Simulation Method. 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 types of ultrasound blur attacks by adjusting the weights of three types of blur patterns.

5 Evaluation

In this section, we evaluate the robustness of VLAs against sensor attacks. We first introduce our experimental setup, including the simulator, target VLAs, and evaluation metrics. Next, we present a comparative evaluation of eight sensor attacks targeting the selected VLAs. Finally, we answer three research questions based on evaluation results.

5.1 Experiment Setup

5.1.1 *Simulator.*

We select Libero [25] as the simulator for our experiments. Libero is an open-source visual-language robotics simulator designed to provide a flexible testing platform for VLAs. It supports a wide range of robots and sensors, providing diverse tasks and environment settings. We utilize Libero's default configuration and further extend it by incorporating sensor attack modules to evaluate the robustness of VLAs against sensor attacks.

- **LIBERO-Spatial** focuses on tasks that require understanding and generalizing spatial relationships among objects. Each task involves manipulating identical objects placed in varying spatial configurations.
- **LIBERO-Object** emphasizes the ability to recognize and manipulate a variety of objects. Tasks in this suite involve moving different objects to specific locations, testing the agent's capacity to generalize object manipulation skills.
- **LIBERO-Goal** consists of tasks where the objects and their spatial arrangements remain constant, but the end goals differ. This setup evaluates the agent's proficiency in goal-directed behavior and procedural knowledge transfer.
- **LIBERO-Long** includes tasks that involve long-horizon planning and execution. These tasks are designed to assess the agent's ability to maintain performance over extended sequences of actions, requiring sustained attention and memory.

5.1.2 *Target VLAs.*

We select four representative VLAs as target VLAs for evaluation: OpenVLA [20], OpenVLA-OFT [19], pi0 [6], and pi0-fast [29]. 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 simulation environment.

5.1.3 *Evaluation Metrics.*

Task Success Rate (TSR): a metric utilized to assess VLAs in accomplishing particular tasks. It is determined as the proportion of successful task completions to the total number of tasks tried. A task is deemed successful when the VLA accomplishes the intended result as outlined in the task BDDL files.

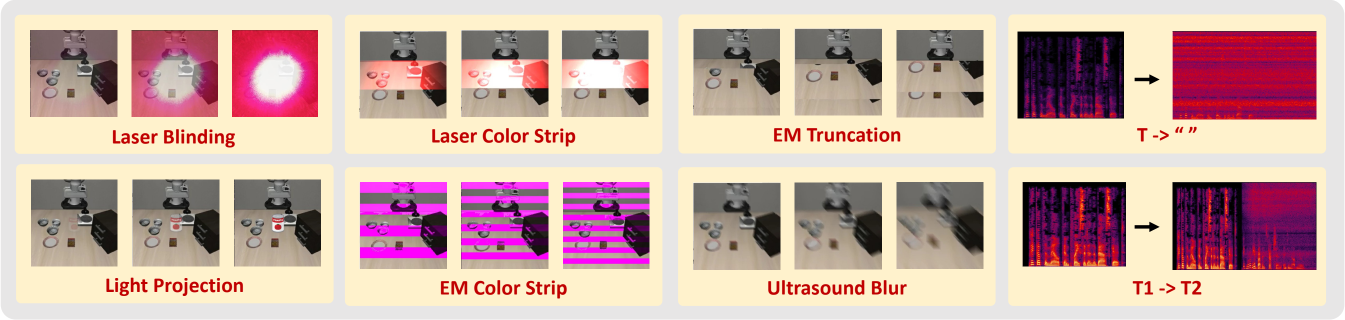

Figure 3: Attack instances under varying attack intensities. From left to right, the attack intensity progressively increases. Detailed parameters for each attack configuration are provided in Table 1

Table 1: Attack Parameter of Different Attack Intensities

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 2: Performance of VLAs in the Libero Dataset

Benchmark	VLA Model	Task Success Rate (%)				
		Spatial	Object	Goal	Long	Avg.
LIBERO	OpenVLA	84.7	88.4	79.2	53.7	76.5
	OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
	π_0	96.8	98.8	95.8	85.2	94.2
	π_0-fast	96.4	96.8	88.6	60.2	85.5

5.1.4 Attack Parameters. As shown in Figure 3, we set three attack intensities—weak, medium, and strong—and their corresponding attack parameters are listed in Table 1.

5.2 Performance of VLAs

In this section, we evaluate the performance of the four target VLAs in the Libero simulation environment without any sensor attacks. The results are summarized in Table 2.

> Observation 1
>
> The VLA models enhance the ability of robots to perform complex manipulation tasks.

The results indicate that these VLAs exhibit strong performance on the Libero dataset, achieving up to 90% TSR in simple tasks such as **LIBERO-Spatial** and **LIBERO-Object**. In long-horizon tasks, OpenVLA-OFT also demonstrates good performance. These results indicate that VLAs generally exhibit reliable task execution capability without sensor attacks.

5.3 Robustness of VLAs against Sensor Attacks

In this section, we evaluate the performance of various VLA models under a range of simulated sensor attacks. The empirical results are summarized in Table 3. Our analysis seeks to answer the following key research questions:

- **Q1:** How robust are VLA models against sensor attacks?
- **Q2:** What is the impact of different sensor attack types on VLA performance?
- **Q3:** How do various VLAs behave when exposed to sensor attacks?

> Answer 1
>
> Sensor attacks can significantly degrade the performance of VLA models.

As shown in Table 3, all VLA models clearly exhibit vulnerability to sensor attacks. The degree of performance degradation differs significantly depending on the specific VLA architecture, attack type, and attack intensity. Although VLA models achieve strong performance under benign conditions, their robustness deteriorates considerably when sensor inputs are compromised. In many scenarios, particularly under strong attacks or during long-horizon tasks, model performance collapses catastrophically. These findings underscore an important gap between the demonstrated capabilities of VLAs under idealized environments and their reliability in real-world settings, where sensor integrity cannot be assured.

> Answer 2
>
> Different types of sensor attacks have significantly different impacts on VLA performance.

5.3.1 Camera Attacks.

Highly Destructive Attacks: Attacks such as Laser Blinding (LB), EM Truncation (ET), Laser Color Strip (LCS), and Ultrasound Blur (UB) demonstrate substantial destructive potential, particularly under medium to strong intensity settings. These attacks directly distort or obscure essential visual information, including object locations, shapes, and class identities. By rendering visual inputs ineffective, these methods induce severe task failures, resulting in either unexpected actions or even potentially harmful operations.

Table 3: Robustness of VLAs against sensor attacks

Attack	OpenVLA				OpenVLA-OFT				$\pi 0$				$\pi 0$-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

📷 Camera; 🎙: Microphone; **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.

Moderately Destructive Attacks: In contrast, Light Projection (LP) and EM Color Strip (ECS) attacks are relatively less severe. These attacks operate by injecting interference—such as distracting patterns or color strips—rather than completely removing essential visual features. Although these interferences degrade model performance by complicating visual perception, the primary objects and intended goals typically remain discernible, leading to less degradation in performance.

5.3.2 *Microphone Attacks.*

Voice DOS: This attack negatively impacts the performance of most VLA models, although the severity varies depending on the task type. In the **LIBERO-Goal** task, which requires executing different instructions within a fixed scene, the absence of instructional guidance significantly impairs the model's ability to infer correct actions. Conversely, in tasks where the scene uniquely aligns with a specific instruction (e.g., **LIBERO-Spatial**, **LIBERO-Object**, and **LIBERO-Long**), the model can effectively infer appropriate actions based solely on visual context, thus mitigating the impact of the attack.

Voice Spoofing: The effectiveness of this attack strongly depends on the model's semantic understanding and instruction-following capability. Compared to pi0 and pi0-fast, OpenVLA and OpenVLA-OFT utilize larger language model backbones, rendering them more susceptible to executing malicious injection instructions and consequently leading to performance degradation. Compared to OpenVLA, OpenVLA-OFT incorporates a FiLM module, which leverages task-specific language embeddings to modulate visual features, thus enhancing the model's ability to follow instructions. Consequently, it experiences the most significant performance degradation.

Answer 3

Different VLA models exhibit significant differences against sensor attacks.

OpenVLA: The OpenVLA model consistently demonstrates the highest vulnerability. Under moderate and strong attack intensities, its performance suffers severe degradation, indicating a significant sensitivity to perturbations in visual inputs and an inadequate ability to handle such disturbances.

OpenVLA-OFT: The OpenVLA-OFT significantly enhances the robustness of its visual perception by integrating support for multi-camera images and processing the robot's proprioceptive states. Nonetheless, the model remains vulnerable to Voice Spoofing (VS) attacks, which reduce its performance to nearly 0% across all four task categories.

$\pi 0$ and $\pi 0$-fast: Both $\pi 0$ and $\pi 0$-fast demonstrate strong robustness against visual attacks, a property largely attributable to their multi-visual sensor architectures. Specifically, disrupting input from a single sensor does not significantly degrade model performance, as the models effectively leverage remaining sensors to maintain high task success rates. Compared to $\pi 0$, the $\pi 0$-fast variant features fewer parameters and improved inference speed and efficiency, yet it still demonstrates exceptional robustness against sensor attacks.

6 Related Work

6.1 Attacks on VLAs

With the rapid advancement of VLAs, their security has garnered significant attention. Recent studies have primarily centered on vulnerability exploration, with identified vulnerabilities mostly residing in the digital domain, broadly categorized into adversarial attacks and backdoor attacks. Specifically, the RoboticAttacks [37] framework introduces adversarial patch attacks targeting image inputs of VLA models. It implements three distinct adversarial techniques—UADA, UPA, and TMA—to induce deviations in robotic arm trajectories. On the other hand, Robotgcg [21] applies LLM jailbreak attacks in the text modality against VLAs, thereby disrupting the normal operation of robotic arms. Additionally, the BadVLA [42] framework represents the first backdoor attack methodology explicitly designed for VLA models. It leverages a two-stage "target decoupling optimization" procedure, which preserves regular task performance while ensuring that the introduced trigger features separately induce task failure. Different from this work, we are the first to explore the robustness of VLA models against sensor attacks.

7 Conclusion

This paper investigates the robustness of Visual-Language-Action (VLA) models against sensor attacks, which is a crucial issue to ensure their secure deployment in the real world. To achieve efficient and large-scale evaluation, we construct a framework that simulates eight types of physical world attacks against cameras and microphones in the digital domain. Our experimental evaluation clearly demonstrates that existing VLA models are highly vulnerable to sensor attacks. Such attacks can severely degrade model performance, resulting in erroneous or even hazardous behaviors. Consequently, this vulnerability poses a direct and significant security threat to real-world applications. In conclusion, although the VLA models show great potential in the path towards general artificial intelligence, ensuring their robustness against sensor attacks is a prerequisite that must be fulfilled before deploying these powerful autonomous systems widely in society.

Acknowledgments

We thank the anonymous reviewers for their valuable comments.

References

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

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).

[5] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734* (2025).

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* (2024).

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[8] Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, et al. 2024. Manipulation Facing Threats: Evaluating Physical Vulnerabilities in End-to-End Vision Language Action Models. *arXiv preprint arXiv:2409.13174* (2024).

[9] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* (2023), 02783649241273668.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

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

[14] FOURIER-Robotics. 2025. FOURIER-Robotics. https://www.fftai.com/.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[16] Chengyin Hu, Weiwen Shi, and Ling Tian. 2023. Adversarial color projection: A projector-based physical-world attack to DNNs. *Image and Vision Computing* 140 (2023), 104861.

[17] Xiaoyu Ji, Yushi Cheng, Yuepeng Zhang, Kai Wang, Chen Yan, Wenyuan Xu, and Kevin Fu. 2021. Poltergeist: Acoustic adversarial machine learning against cameras and computer vision. In *2021 IEEE symposium on security and privacy (SP).* IEEE, 160–175.

[18] Qinhong Jiang, Xiaoyu Ji, Chen Yan, Zhixin Xie, Haina Lou, and Wenyuan Xu. 2023. {GlitchHiker}: Uncovering vulnerabilities of image signal transmission with {IEMI}. In *32nd USENIX Security Symposium (USENIX Security 23).* 7249–7266.

[19] Moo Jin Kim, Chelsea Finn, and Percy Liang. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645* (2025).

[20] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).

[21] Eliot Krzysztof Jones, Alexander Robey, Andy Zou, Zachary Ravichandran, George J Pappas, Hamed Hassani, Matt Fredrikson, and J Zico Kolter. 2025. Adversarial Attacks on Robotic Vision Language Action Models. *arXiv e-prints* (2025), arXiv–2506.

[22] Shunlei Li, Jin Wang, Rui Dai, Wanyu Ma, Wing Yin Ng, Yingbai Hu, and Zheng Li. 2024. RoboNurse-VLA: Robotic Scrub Nurse System based on Vision-Language-Action Model. *arXiv preprint arXiv:2409.19590* (2024).

[23] Xinfeng Li, Chen Yan, Xuancun Lu, Zihan Zeng, Xiaoyu Ji, and Wenyuan Xu. 2023. Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time. *arXiv preprint arXiv:2308.01040* (2023).

[24] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. 2024. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647* (2024).

[25] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems* 36 (2023), 44776–44791.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[28] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[29] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*

(2025).

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[31] 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/.

[32] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. 2025. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844* (2025).

[33] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems. In *29th USENIX Security Symposium (USENIX Security 20)*. 2631–2648.

[34] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).

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

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[37] Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. 2024. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. *arXiv preprint arXiv:2411.13587* (2024).

[38] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. 2025. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control. *arXiv preprint arXiv:2502.05855* (2025).

[39] Chen Yan, Zhijian Xu, Zhanyuan Yin, Xiaoyu Ji, and Wenyuan Xu. 2022. Rolling colors: Adversarial laser exploits against traffic light recognition. In *31st USENIX Security Symposium (USENIX Security 22)*. 1957–1974.

[40] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[41] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 103–117.

[42] Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. 2025. BadVLA: Towards Backdoor Attacks on Vision-Language-Action Models via Objective-Decoupled Optimization. *arXiv preprint arXiv:2505.16640* (2025).