VLA-Hijack: A Transferable Patch Attack against Vision-Language-Action Models via Visual Proprioception Hijacking

Jiyuan Fu[1,*], Kaixun Jiang[2,*], Jingkai Jia[2], Zhaoyu Chen[2], Xueyao Chen[1], Lingyi Hong[1], Shuyong Gao[3], Chenzhi Tan[1], Dingkang Yang[2], and Wenqiang Zhang[1,2,†]

[1] Shanghai Key Lab of Intelligent Information Processing, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
[2] College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{fujy23,kxjiang22}@m.fudan.edu.cn
{wqzhang}@fudan.edu.cn
[*] Equal contribution [†] Corresponding author

Abstract. While Vision-Language-Action (VLA) models have emerged as powerful generalist policies, their severe vulnerability to adversarial patches significantly hinders their deployment in safety-critical domains. Moreover, existing patch attacks primarily focus on white-box settings, heavily overfitting to the specific action output space of the target model, which results in poor cross-architecture transferability. To overcome this limitation, we propose VLA-Hijack, a unified adversarial framework that breaks the transferability bottleneck by exploiting a fundamental vulnerability identified in this work: before planning any motion, a VLA model must first use visual information to locate its own robotic arm within the environment. Targeting this shared visual self-localization process, our approach concurrently optimizes Attention-Guided Proprioceptive Suppression to inhibit the real robotic arm's features, and Multimodal Proprioceptive Injection to establish the patch as a surrogate "phantom embodiment". By alternating between semantic concept anchoring and visual prototype projection, VLA-Hijack effectively severs the semantic relationship between the agent's true embodiment and its control policy. Extensive experiments across diverse architectures (OpenVLA, UniVLA, and CronusVLA) demonstrate that VLA-Hijack achieves superior optimization efficiency in white-box settings and sets a new SOTA for cross-architecture and cross-domain black-box transferability.

Keywords: Vision-Language-Action Models · Adversarial Attack · Attack Transferability · Visual Proprioception

1 Introduction

The emergence of Vision-Language-Action (VLA) models [1, 3, 13, 14, 27, 33, 40] has bridged the gap between perception and control, enabling robots to leverage

the reasoning capabilities of large models for complex physical interactions. However, as these models permeate safety-critical domains such as medical assistance and home services, their severe vulnerability to adversarial attacks significantly hinders their real-world deployment. Compared to traditional Vision-Language Models (VLMs), VLA models interact directly with the physical world, rendering their adversarial robustness significantly more critical [20,34]. While current research extensively investigates the adversarial vulnerabilities of VLMs [7,8,10, 11, 36], the security of VLA models has received comparatively limited attention [19, 25].

In the context of adversarial vulnerabilities, patch attacks have garnered significant attention due to their practical deployability [5,12,23]. Unlike traditional adversarial attacks that require imperceptible modifications across the entire image, patch attacks introduce localized, visible patterns within the camera's field of view, presenting a tangible security threat [6, 21, 32]. Existing attacks on VLA models [25] have primarily targeted the output action space, aiming to distort control signals by maximizing the action deviation. While effective in white-box settings, we identify a critical bottleneck in this paradigm: poor transferability. VLA models are characterized by heterogeneous action encodings and highly distinctive action space distributions that depend on the robot's physical morphology and training data. Consequently, adversarial perturbations that optimize for action divergence tend to overfit the specific "Action Head" of the source model [3, 13, 14]. This dependency renders the attack brittle, preventing generalization to unseen architectures with distinct decision-making.

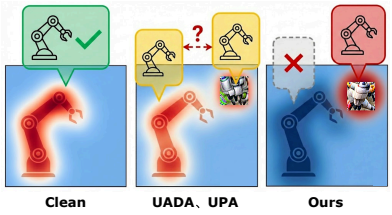


Fig. 1: Unlike prior attacks that leave the real arm unsuppressed, VLA-Hijack hijacks the proprioceptive loop by suppressing the original embodiment and injecting a surrogate identity.

To break this transferability bottleneck, we propose a paradigm shift: targeting the universal proprioceptive logic of VLA models rather than their divergent action outputs. Intriguingly, recent adversarial studies [19, 25, 30] have uncovered a provocative phenomenon: even when solely optimizing for action error, the generated adversarial patches unintentionally converge towards visual patterns resembling robotic arm joints (as visualized in the middle panel of Fig. 1). While prior work attributes this to representation bias from restricted camera views [25], we argue this exposes a more fundamental vulnerability in VLA decision-making: the reliance on a **visual proprioceptive loop**. VLA manipulation is strictly contingent on scene perception: the model must **first localize its own physical state (the arm) from the current observation**, then locate the target object, and finally plan the relative motion [2,13]. In this context, visual proprioception acts as the model's intrinsic perception and awareness of its own robotic morphology. Consequently, when a patch's features mimic the robotic arm, it effectively hijacks this internal perception, deceiving the model into misidentifying the patch as its true embodiment.

Once this "self-localization" step is compromised, the model plans based on an erroneous physical origin (the patch location), inducing a systematic deviation in all subsequent trajectory calculations and leading to catastrophic failure. Since the capability to extract robotic arm semantics is universally acquired by VLAs during training, targeting this shared semantic layer fundamentally resolves the transferability challenges plaguing action-space attacks. This is empirically validated in Table 2, where a naive "Arm Image" consistently outperforms several sophisticated attacks. This confirms that robotic arm features represent a robust, universal focal point that VLA models inherently rely upon for decision-making.

Motivated by this insight, we propose VLA-Hijack, a unified adversarial framework designed to sever the relationship between the agent's embodiment and its control policy through a **Proprioceptive Identity Rewiring** mechanism. We optimize two complementary objectives: (1) *Attention-Guided Proprioceptive Suppression*, which actively inhibits the feature responses of the original arm. Unlike prior strategies [25] that suffer from severe semantic interference by failing to explicitly suppress original proprioceptive features, VLA-Hijack creates a "feature vacuum" that blinds the model to its true physical embodiment; and (2) *Multimodal Proprioceptive Injection*, which fills this void by establishing the adversarial patch as a "phantom embodiment". By employing a dynamic schedule that alternates between semantic concept anchoring and visual prototype projection, we ensure the patch captures robust identity features, thereby addressing the indirect and suboptimal nature of existing optimization paths that rely solely on passive action error backpropagation [25]. This synchronized manipulation effectively "hijacks" the model's self-localization capability, ensuring superior attack transferability across diverse VLA architectures.

In summary, our main contributions are as follows:

- We uncover an explainable VLA decision logic: models must visually localize their robotic arm before planning motion. We further identify this mechanism as a novel insight for significantly enhancing attack transferability across heterogeneous VLA architectures.
- We propose VLA-Hijack, a unified adversarial framework. By synergistically coupling attention-guided suppression with alternating multimodal injection, it synthesizes a phantom embodiment to hijack the model's decision-making.
- Evaluations across 12 victim models in the Real-to-Sim setting demonstrate that VLA-Hijack sets a new SOTA for transferability. Our method achieves a 63.68% average Failure Rate, outperforming existing SOTA baselines by absolute margins of **42.66%** to **45.41%**.

2 Related Work

2.1 Vision-Language-Action Models

Vision-Language-Action (VLA) models have emerged as powerful generalist policies by integrating robotic action generation into pretrained vision-language foundations [1, 3, 13, 14, 33, 40]. A prominent line of work [2, 13, 16, 22], including

RT-2 [2] and OpenVLA [13], treats physical control as a language modeling task, discretizing continuous actions into tokens for autoregressive prediction. Conversely, to bypass the limitations of discrete formulations and the heavy reliance on labeled interactive data, recent advancements explore continuous action heads and latent representations [3, 14, 15, 28, 33]. For instance, UniVLA [3] learns a unified latent action representation from internet-scale action-free videos, while CronusVLA [14] adapts single-frame discrete policies to a multi-frame continuous paradigm. Given the significant heterogeneity in their action space designs—spanning discrete tokenization, unified latent representations, and continuous prediction—we select OpenVLA [13], UniVLA [3], and CronusVLA [14] as our representative targets to evaluate the cross-architecture transferability of our proposed attack.

2.2 Attacks on VLA Models

As VLA models are increasingly deployed in embodied applications, their adversarial vulnerabilities have garnered significant attention. A substantial body of work investigates training-time threats, demonstrating how VLA systems can be compromised through data poisoning and backdoor attacks [9, 17, 29, 31, 37–39]. While these methods embed malicious behaviors by altering the training or fine-tuning pipelines, another line of research focuses on inference-time evasion. Within this scope, initial adversarial attacks typically employ full-image perturbations, applying imperceptible noise across the entire visual observation to mislead the model's execution [26, 35]. To present a more realistic threat vector without modifying every pixel, patch attacks introduce localized, visible patterns within the camera's field of view [19, 25, 30]. However, existing patch attacks on VLA models often suffer from poor cross-architecture transferability. To address this limitation, we propose a novel adversarial framework that significantly improves transferability by explicitly hijacking the model's decision-making through the suppression of its visual proprioception.

3 Methodology

3.1 Preliminaries & Threat Model

In this section, we formally define the VLA model and its inference mechanism, followed by the formulation of our adversarial attack objective.

VLA Formulation. We define a VLA model as a parameterized function f_θ, designed to generate physical control signals based on visual observations and linguistic instructions. Specifically, given an RGB observation image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural language instruction T (e.g., *"Pick up the blue cube"*), the inference process of a VLA model is formulated as:

$$A = f_\theta(I, T), \tag{1}$$

where A denotes the generated action trajectories.

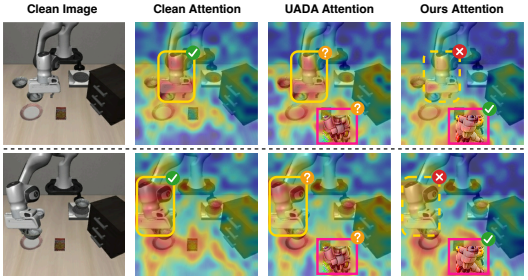


Fig. 2: Visualization of Proprioceptive Conflict. Unlike UADA (Col 3), which splits attention between the real arm and the patch, our method (Col 4) completely suppresses the real arm and redirects focus exclusively to the adversarial patch.

Attack Objective. This work focuses on *universal untargeted attacks* against VLA models—meaning a single perturbation can generalize across various tasks. We aim to generate a single adversarial patch P that, applied across diverse observations, causes task failure by inducing erroneous actions without altering the instruction T. Formally, we define the adversarial observation image I_{adv} as:

$$I_{adv} = I \odot (1 - M) + P \odot M, \tag{2}$$

where $M \in \{0,1\}^{H \times W}$ represents the binary mask of the patch (with 1 indicating the patch region and 0 otherwise), and $\odot$ denotes element-wise multiplication.

To achieve the attack goal, we formulate the patch generation as a constrained optimization problem. We seek to find the optimal patch P that minimizes a specific adversarial objective function $\mathcal{L}_{total}$ over the data distribution $\mathcal{D}$:

$$\min_{P} \mathbb{E}_{(I,T)\sim\mathcal{D}} \left[\mathcal{L}_{total}(f_\theta(I_{adv}, T))\right] \quad \text{s.t.} \quad P \in [0,1]^{H \times W}, \tag{3}$$

where $\mathbb{E}_{(I,T)\sim\mathcal{D}}$ denotes the expectation over the data distribution $\mathcal{D}$, ensuring the generated patch P achieves universal effectiveness across diverse visual observations and instructions. In our proposed framework, this objective is instantiated as a unified loss function $\mathcal{L}_{total}$, as detailed in Sec. 3.5.

Threat Model. We adopt a **transfer-based attack** setting, where the attacker is assumed to have white-box access to a single *surrogate model* f_θ to compute gradients and optimize the patch P. During the inference phase, the optimized patch is evaluated against other unseen *victim models* in a black-box manner. In this setting, the attacker has no knowledge of the victim models' internal parameters or architectures, relying solely on the cross-model transferability of the adversarial features captured by P.

3.2 Overview of VLA-Hijack

Motivation: The Proprioceptive Conflict. Our framework is driven by a critical phenomenon observed in the attention mechanisms of VLA models. As

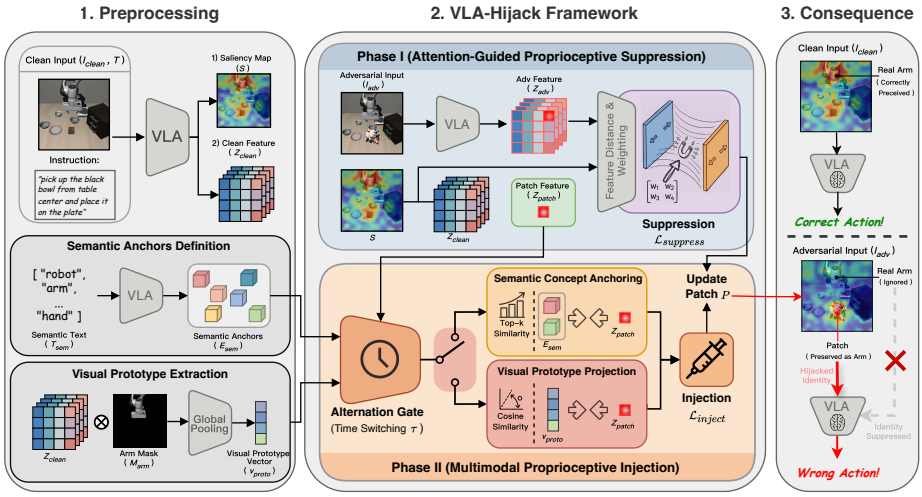


Fig. 3: Overview of the VLA-Hijack Framework. We employ a two-phase optimization loop to generate the adversarial patch: first suppressing the features of the real arm, and then injecting multimodal robotic arm identities into the patch.

illustrated in Fig. 2, VLA manipulation is strictly contingent on a **"Visual Proprioception Loop"**: the model must first localize its own physical state (the robotic arm) from the current observation, then locate the target object, and finally plan the robotic arm's relative motion. In the Clean State (Fig. 2, Col 2), the attention map clearly highlights both the *target object* and the *robotic arm*, confirming that self-localization is a prerequisite for control.

Crucially, prior attacks (e.g., UADA) [25] inadvertently reveal this dependency. As shown in Fig. 2, even when relying on indirect objectives like action-error backpropagation rather than direct perceptual guidance, adversarial patches spontaneously evolve into patterns resembling robotic components. While this confirms that VLA decisions are fundamentally anchored to visual self-localization, such indirect optimization is inherently suboptimal. By failing to suppress the *original embodiment*, these methods result in a severe **"Semantic Conflict"** where the model's attention is split between the salient real arm and the patch. To achieve robust and transferable control, we must therefore address this bottleneck directly: we need to hijack the proprioceptive loop by removing the real embodiment and establishing the patch as the sole perceived embodiment.

The VLA-Hijack Framework. To address this bottleneck, we propose **VLA-Hijack**, a unified framework designed to execute the aforementioned identity replacement. As illustrated in Fig. 3, the framework operates through a synergistic **"Proprioceptive Identity Rewiring"** mechanism, explicitly targeting the visual proprioception loop via two concurrent objectives:

– *Attention-Guided Proprioceptive Suppression.* We penalize the feature response of the real robotic arm. This creates the "perceptual vacuum" (vi-

sualized as the blue region in Fig. 2, Col 4), blinding the model to its true physical presence.
– *Multimodal Proprioceptive Injection.* Synchronized with the suppression, we inject the visual and semantic prototypes of a target embodiment into the patch. This forces the model to "latch onto" the patch (visualized as the red region in Fig. 2, Col 4) as its sole embodiment.

3.3 Phase I: Attention-Guided Proprioceptive Suppression

To sever the relationship between the model and its embodiment, this phase exploits the VLA model's internal attention mechanisms to identify and suppress proprioceptive features.

Saliency-Guided Weighting. We directly extract the multi-head self-attention maps from the final layer of the visual encoder. By aggregating the attention weights across all heads, we obtain a raw spatial saliency map $S \in \mathbb{R}^{h \times w}$, where h and w denote the spatial dimensions of the feature map. Given that the high-response regions in S capture the model's focus on proprioceptive features (i.e., the robotic arm), we transform S into a penalty weight map W via normalization:

$$W = \mathbf{1} + \alpha \cdot \frac{S - S_{min}}{S_{max} - S_{min} + \epsilon} \quad , \tag{4}$$

where S_{min} and S_{max} denote the minimum and maximum values of the saliency map, ϵ ensures numerical stability, and α is a scaling factor that amplifies the penalty intensity on these high-attention regions.

Weighted Feature Suppression. Leveraging W, we introduce a weighted loss to eliminate the original semantic information. Let Z_{clean} and Z_{adv} denote the feature maps of the clean and adversarial images, respectively. We minimize their weighted cosine similarity:

$$\mathcal{L}_{suppress} = \frac{\sum(W \odot \mathrm{CosSim}(Z_{adv}, Z_{clean}))}{\sum W} \quad , \tag{5}$$

Crucially, this explicit weighting prevents the optimizer from trivially minimizing the loss by perturbing unstable background tokens, forcing it to prioritize dismantling the highly robust proprioceptive features. By minimizing $\mathcal{L}_{suppress}$, we force the adversarial features to deviate orthogonally from the original features, specifically in regions where the model originally attended, effectively creating a "perceptual vacuum" around the robotic arm.

3.4 Phase II: Multimodal Proprioceptive Injection

Following the suppression of original features, this phase aims to fill the "feature vacuum" by synthesizing a surrogate embodiment. To this end, we formulate two complementary injection objectives: **Semantic Concept Anchoring** and **Visual Prototype Projection**.

Semantic Concept Anchoring. To activate the language alignment modules within the VLA model, we anchor the patch features to a predefined semantic space. We define a set of text prompts $\mathcal{T}_{sem}$ describing the embodiment (e.g., *{"robot", "arm", "hand", "claw", "end-effector"}*) and feed them into the text encoder to obtain normalized embeddings $E_{sem} \in \mathbb{R}^{K \times D}$, serving as **Semantic Anchors**. To enforce strong semantic correlation while addressing the inherent semantic polysemy of robotic features (e.g., distinct textures resembling a "claw" versus an "arm"), we reject rigid one-to-one mapping in favor of a flexible Top-k Aggregation Strategy. We compute the cosine similarity between the adversarial patch features and all anchors, maximizing the average similarity to the top-k most relevant ones for each token:

$$\mathcal{L}_{sem} = 1 - \frac{1}{|\Omega|} \sum_{j \in \Omega} \left(\frac{1}{k} \sum_{e \in \text{TopK}(Z_{adv}^{(j)}, E_{sem})} \frac{Z_{adv}^{(j)} \cdot e}{\|Z_{adv}^{(j)}\|_2 \|e\|_2} \right), \tag{6}$$

where Ω denotes the set of spatial token indices covering the adversarial patch region, and $|\Omega|$ represents the total number of patch tokens. $Z_{adv}^{(j)}$ is the j-th token's feature vector, and $\text{TopK}(\cdot)$ selects the k anchors with the highest similarity scores. By aggregating these top candidates, this objective forces the patch features to converge towards a comprehensive semantic centroid of the "robotic arm" concept, ensuring robust alignment across diverse visual attributes.

Visual Prototype Projection. While semantic anchoring ensures the patch is recognized as a generic "robotic arm" in the linguistic space, it fails to capture the fine-grained visual characteristics (e.g., specific textures, lighting, and geometric details) of the actual embodiment. To bridge this modality gap and achieve a holistic identity injection, we introduce **Visual Prototype Projection** to directly clone the visual features of the physical arm.

Prior to feature extraction, we explicitly isolate the physical embodiment region. We utilize the segmentation model to obtain the binary mask M_{arm} corresponding to the robotic arm from the clean observation I_{clean}.

Based on this mask, we extract a robust **Visual Prototype Vector** $v_{proto} \in \mathbb{R}^D$ by performing global average pooling over the masked region of the clean features Z_{clean}, followed by ℓ_2 normalization:

$$v_{proto} = \frac{\sum_i M_{arm}^{(i)} \cdot Z_{clean}^{(i)}}{\left\| \sum_i M_{arm}^{(i)} \cdot Z_{clean}^{(i)} \right\|_2 + \epsilon} \quad . \tag{7}$$

Subsequently, we define the **Visual Injection Loss** $\mathcal{L}_{vis}$ as the average cosine distance between the adversarial features in the patch and the prototype:

$$\mathcal{L}_{vis} = 1 - \frac{1}{|\Omega|} \sum_{j \in \Omega} \frac{Z_{adv}^{(j)} \cdot v_{proto}}{\|Z_{adv}^{(j)}\|_2 \|v_{proto}\|_2} \quad . \tag{8}$$

Minimizing $\mathcal{L}_{vis}$ ensures that the patch effectively "clones" the texture and structural information of the real robot arm in the feature space.

3.5 Optimization

We formulate the attack generation as a joint optimization problem. Our ultimate goal is to find an optimal patch P that simultaneously maintains a "perceptual vacuum" on the original embodiment while firmly establishing the surrogate identity. The total objective function is defined as:

$$\min_{P} \mathcal{L}_{total} = \mathcal{L}_{suppress} + \lambda \cdot \mathcal{L}_{inject}^{(t)} \quad , \tag{9}$$

where λ is a hyperparameter balancing suppression and injection strength.

Alternating Injection Schedule. While the suppression term remains constant, strictly optimizing both injection modalities simultaneously can lead to gradient conflicts. To address this, we define the injection term $\mathcal{L}_{inject}^{(t)}$ as a dynamic component that alternates between semantic and visual modalities based on the iteration step t and a period τ:

$$\mathcal{L}_{inject}^{(t)} = \begin{cases} \mathcal{L}_{sem}(Z_{adv}, E_{sem}), & \text{if } \lfloor \frac{t}{\tau} \rfloor \text{ is even} \\ \mathcal{L}_{vis}(Z_{adv}, v_{proto}), & \text{if } \lfloor \frac{t}{\tau} \rfloor \text{ is odd} \end{cases} . \tag{10}$$

This alternating strategy mitigates optimization conflicts, allowing the patch to iteratively align with the target identity across modalities while continuously neutralizing the real arm via $\mathcal{L}_{suppress}$. Full procedures are in Supplement.

4 Experiments

4.1 Experiment Setup

Datasets and Baseline. To comprehensively evaluate our attack framework, we conduct experiments on LIBERO [18] and BridgeData V2 [24]. LIBERO [18] serves as our primary simulation benchmark, assessing high-level reasoning through four task suites: Spatial, Object, Goal, and Long. These suites challenge models with complex, multi-step sequential decision-making. Complementing this simulation environment, BridgeData V2 [24] provides a large-scale real-world corpus comprising 60,096 trajectories across 24 diverse environments. Covering 13 fundamental manipulation skills, it allows us to rigorously test the VLA robustness in unstructured physical scenes. We benchmark against 8 adversarial objectives [25] (e.g., UADA, UPA, and TMA with varying Degrees-of-Freedom), as well as Random Noise and Arm Image to evaluate naive visual priors.

Victim VLAs. We evaluate adversarial robustness across three representative VLA architectures: OpenVLA [13], UniVLA [3], and CronusVLA [14]. In the LIBERO simulation domain [18], we explicitly utilize four distinct variants for each architecture, independently trained on the specific task suites (i.e., Spatial, Object, Goal, and Long), resulting in a total of 12 distinct victim models. For white-box optimization, we employ OpenVLA [13] as the primary surrogate architecture. Specifically, we select the OpenVLA-Spatial variant, as it exhibits

Table 1: Attack effectiveness within the Simulation Domain. We report the Failure Rate (FR, %; ↑) in the LIBERO simulator. Patches are generated using the *OpenVLA-Spatial* surrogate model and evaluated on 12 victim models. Shaded entries denote the white-box performance, while *Transfer Avg.* reports the mean FR across the remaining 11 black-box variants.

Method	OpenVLA				UniVLA				CronusVLA				Transfer Avg.
	Spatial*	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	
Benign	16.00	16.80	24.20	45.20	2.40	5.60	5.20	7.00	7.40	7.60	4.20	13.80	-
Random Noise	26.80	27.40	24.60	48.60	3.80	9.60	6.60	7.40	11.80	7.60	4.60	14.20	15.11
Arm image	66.20	44.40	28.80	60.80	4.80	11.20	9.40	15.80	8.20	11.40	5.20	15.20	19.56
$UADA_1$	38.80	37.00	26.40	53.80	6.80	9.40	8.40	13.20	15.60	9.00	5.80	15.40	18.25
$UADA_{1-3}$	**100.00**	92.20	99.80	96.60	9.60	24.80	15.00	24.00	17.80	8.80	6.20	18.00	37.53
UPA_1	62.20	47.20	32.20	59.60	6.80	14.00	7.60	21.40	12.80	10.20	5.40	16.40	21.24
UPA_{1-3}	55.40	44.20	30.60	65.20	6.00	11.80	9.80	17.80	14.40	9.40	4.80	14.60	20.78
TMA_1	99.40	79.20	44.00	80.20	10.80	14.60	8.60	29.00	14.40	8.80	5.60	17.20	28.40
TMA_{1-3}	99.00	82.20	52.00	84.20	32.60	19.60	8.00	30.60	15.40	8.40	5.80	15.80	32.24
TMA_7	98.00	70.00	40.60	82.40	6.80	10.20	9.60	27.60	17.40	6.60	4.60	16.80	26.60
TMA_{1-7}	96.20	92.80	46.60	83.80	14.60	11.40	9.00	14.60	13.40	9.60	5.60	14.60	28.73
Ours	**100.00**	**100.00**	**100.00**	**99.60**	**88.20**	**100.00**	**25.20**	**80.80**	**28.20**	**12.40**	**25.80**	**17.80**	**61.64**

the lowest failure rate on clean samples, thereby ensuring that attack success is attributed strictly to adversarial efficacy rather than the intrinsic fragility of the surrogate. To further assess cross-architecture transferability, we extend our surrogate set to include all four task-specific variants of UniVLA [3]. In the physical domain, regarding the BridgeData V2 [24], we utilize the OpenVLA-7B model to evaluate performance in unstructured environments.

Attacks Settings and Metrics. For the attack configuration, we initialize the adversarial patch to cover 5% of the input image area, aligning with baseline [25]. To isolate the robotic arm, we extract the binary mask M_{arm} using the SAM 3 [4]. The optimization process is conducted with a batch size of 4, spanning a total of Attack Steps $N = 500$. We empirically set the saliency scaling factor $\alpha = 2$ to enforce focused suppression on embodiment features. Within each step, we perform Inner-Loop Steps $K = 50$ to ensure sufficient feature convergence. For evaluation, we follow the LIBERO benchmark [18] and adopt the Failure Rate (FR) as our primary metric, defined as $1 -$ Success Rate (SR).

Evaluation Details. To ensure the statistical reliability of our results, we adopt a comprehensive evaluation protocol on the LIBERO benchmark [18]. Each task suite comprises 10 distinct manipulation tasks. For our main experiments, we conduct 50 independent rollout trials for every task to account for environmental stochasticity, yielding a robust total of 500 evaluation episodes per suite. Conversely, for all **ablation studies**, we streamline the protocol to 10 independent rollouts per task to efficiently manage computational overhead.

4.2 Main Results

Attack Effectiveness within the Simulation Domain. As reported in Table 1, our proposed VLA-Hijack achieves state-of-the-art performance in the

Table 2: Cross-Domain Evaluation (Real → Sim). We report the FR (%) in the LIBERO simulator. Patches are generated from the real-world BridgeData V2 dataset using *OpenVLA-7B* as the surrogate model, and transferred to 12 simulation models. *Transfer Avg.* reports the mean FR over all 12 victims.

Method	OpenVLA				UniVLA				CronusVLA				Transfer Avg.
	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	Spatial	Object	Goal	Long	
Benign	16.00	16.80	24.20	45.20	2.40	5.60	5.20	7.00	7.40	7.60	4.20	13.80	-
random Noise	26.80	27.40	24.60	48.60	3.80	9.60	6.60	7.40	11.80	7.60	4.60	14.20	16.08
Arm image	66.20	44.40	28.80	60.80	4.80	11.20	9.40	15.80	8.20	11.40	5.20	15.20	23.45
$UADA_1$	32.00	32.40	26.80	50.80	5.20	10.40	7.80	9.60	14.00	9.20	5.60	15.40	18.27
$UADA_{1-3}$	29.40	36.20	28.40	49.40	4.60	11.00	10.40	11.40	15.00	8.40	6.20	13.80	18.68
UPA_1	29.00	34.60	25.80	48.60	5.40	11.80	9.80	9.20	14.20	10.20	6.40	15.40	18.37
UPA_{1-3}	31.20	32.40	28.00	50.60	5.40	12.60	8.60	9.00	15.00	10.00	6.00	15.40	18.68
TMA_1	30.40	38.40	31.40	51.20	4.60	10.40	8.80	13.40	13.80	8.60	5.00	14.20	19.18
TMA_{1-3}	30.80	39.00	30.00	54.00	6.60	8.60	9.80	10.40	14.00	9.80	4.80	14.60	19.37
TMA_7	37.80	36.80	27.80	52.00	5.20	6.00	8.20	8.20	15.20	10.60	5.40	15.20	19.03
TMA_{1-7}	36.40	41.80	31.80	51.80	8.00	9.60	9.40	14.60	16.00	11.00	5.80	16.00	21.02
Ours	**100.00**	**99.80**	**99.80**	**99.40**	**93.40**	**100.00**	**17.80**	**82.80**	**26.20**	**13.80**	**11.40**	**19.80**	**63.68**

white-box setting, attaining a perfect 100.00% Failure Rate (FR) on the surrogate model (OpenVLA-Spatial). Simultaneously, our method demonstrates superior transferability across unseen architectures, significantly outperforming existing attacks. Since prior methods largely neglect the model's fundamental dependency on the visual proprioceptive loop, they tend to overfit the surrogate model, resulting in severe performance degradation on diverse victim models (e.g., the strongest baseline $UADA_{1-3}$ averages only 37.53% in transfer FR). In contrast, by successfully capturing and exploiting the universal embodiment features shared across heterogeneous VLA models, VLA-Hijack bridges this gap, achieving a remarkable Transfer Avg. of 61.64%.

Robustness to Visual Domain Shifts (Real-to-Sim). To further rigorously evaluate robustness under severe distribution shifts, we report the transfer performance in Table 2, where adversarial patches are trained on real-world BridgeData V2 and tested on LIBERO simulation. This setting poses a significant challenge due to the substantial "visual domain gap" (e.g., lighting, textures) between physical and simulation environments. Consequently, prior optimization-based attacks suffer catastrophic failure, yielding an average FR of only 18%-21%, which is surprisingly even lower than the naive "Arm image" baseline (23.45%). This indicates that their generated features are highly sensitive to domain-specific statistics. In stark contrast, VLA-Hijack effectively bridges this domain gap, achieving a commanding Transfer Avg. of 63.68%. By anchoring the attack on the semantic concept of the "robotic arm"—a visual feature that remains invariant across real and simulated worlds—our method realizes a 3× performance improvement over state-of-the-art baselines, demonstrating that the "phantom embodiment" we synthesize possesses true cross-domain universality.

Robustness across Surrogate Architectures. To verify VLA-Hijack's cross-architecture generalization, we evaluate distinct UniVLA variants as surrogates

Table 3: Robustness to Surrogate Selection. We report FR (%) in the LIBERO simulator. Patches are generated using each *UniVLA* variant as the surrogate and evaluated across all 12 victim models. Shaded entries denote white-box performance, while *Transfer Avg.* reports the mean FR over the remaining 11 transfer victims.

Method	Victim→	UniVLA				OpenVLA				CronusVLA				Transfer Avg.
	Surrogate↓	Spa.	Obj.	Goal	Long	Spa.	Obj.	Goal	Long	Spa.	Obj.	Goal	Long	
Benign	-	2.40	5.60	5.20	7.00	16.00	16.80	24.20	45.20	7.40	7.60	4.20	13.80	-
Rand. Noise	-	3.80	9.60	6.60	7.40	26.80	27.40	24.60	48.60	11.80	7.60	4.60	14.20	-
Ours	Uni-Spa.	96.00	**100.00**	29.80	91.40	87.20	**99.00**	66.60	75.80	**25.20**	12.60	20.40	19.20	57.02
	Uni-Obj.	**97.60**	**100.00**	**78.00**	71.20	**99.80**	98.80	**90.20**	94.60	21.80	13.20	**24.80**	17.80	**64.35**
	Uni-Goal	52.80	**100.00**	63.80	61.40	75.80	92.60	62.80	80.60	15.80	**20.80**	8.20	12.20	53.00
	Uni-Long	82.60	97.20	53.60	**96.00**	98.80	91.00	84.00	**99.00**	21.60	13.80	8.40	**20.80**	60.98

Table 4: Sensitivity to Injection Weight λ. Transfer FR on UniVLA variants across different injection weights.

Weight λ	**UniVLA Variants (Failure Rate %)**				
	Spatial	Object	Goal	Long	**Avg**
0.1	79.00	98.00	18.00	62.00	64.25
0.2	**88.00**	**100.00**	**28.00**	**81.00**	**74.25**
0.3	83.00	99.00	18.00	80.00	70.00
0.5	75.00	99.00	18.00	72.00	66.00

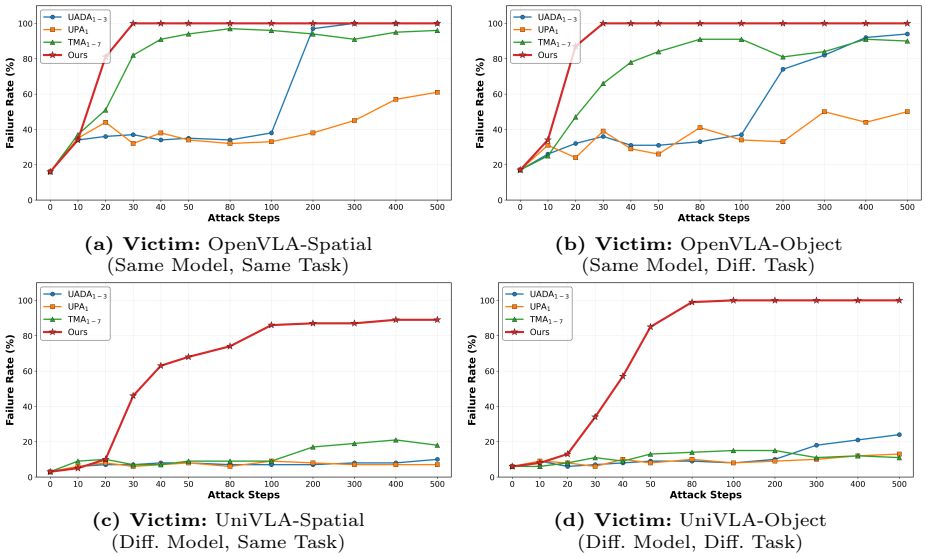
Table 5: Impact of Alternating Period τ. Comparing our alternating schedule against joint optimization (w/o alt.).

Period τ	**UniVLA Variants (Failure Rate %)**				
	Spatial	Object	Goal	Long	**Avg**
w/o alt.	69.00	98.00	19.00	79.00	66.25
3	74.00	98.00	20.00	78.00	67.50
5	**88.00**	**100.00**	**28.00**	**81.00**	**74.25**
10	80.00	**100.00**	23.00	85.00	72.00

in Table 3. Baseline attacks (UADA, UPA, TMA) [25] are inapplicable here, as their reliance on OpenVLA's action discretization renders them incompatible with UniVLA's architecture. This dependency renders them incompatible with UniVLA's architecture. In contrast, VLA-Hijack demonstrates superior compatibility across diverse architectures. Notably, the patch generated from UniVLA-Object achieves a Transfer Avg. of 64.35%, demonstrating strong cross-architecture generalization to OpenVLA and CronusVLA. By targeting the shared Visual Proprioception Loop, our Proprioceptive Identity Rewiring mechanism ensures robust compatibility across heterogeneous VLA architectures.

4.3 Ablation Study

Impact of Attack Steps. To investigate the relationship between the number of attack steps and both convergence efficiency and transferability, we track the Failure Rate trajectories over 500 steps. The results (Fig. 4) reveal two critical insights: 1) High White-box Optimization Efficiency. VLA-Hijack reaches 100% FR within 30 steps by directly targeting the model's decision logic with explicit perceptual objectives. Conversely, prior methods' [25] indirect action-error objectives result in lower efficiency. 2) Superior Transfer Generalization. The contrast becomes stark in cross-architecture scenarios (Fig. 4(c) and (d)). As optimization proceeds beyond 100 steps, baseline attacks [25] (e.g., UADA_{1-3}, UPA_1) essentially flatline at a low FR. While baseline attacks flatline due to overfitting architecture-specific action tokens, VLA-Hijack achieves sustained transfer gains by synthesizing a "phantom embodiment" to progressively refine universal robotic features shared across architectures.



(a) Victim: OpenVLA-Spatial (Same Model, Same Task)

(b) Victim: OpenVLA-Object (Same Model, Diff. Task)

(c) Victim: UniVLA-Spatial (Diff. Model, Same Task)

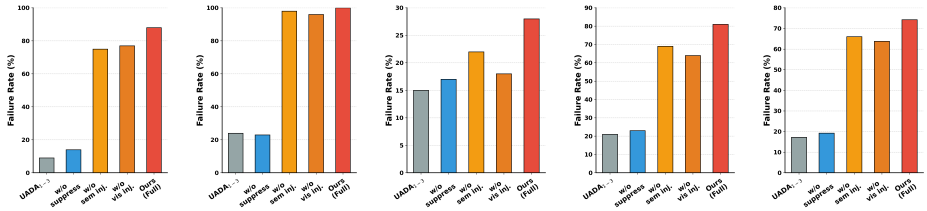
(d) Victim: UniVLA-Object (Diff. Model, Diff. Task)

Fig. 4: Failure Rate (FR) vs. Optimization Steps. We plot FR against attack steps using patches from the *OpenVLA-Spatial* surrogate. Compared to baselines, **VLA-Hijack** (red curve) demonstrates faster convergence in the white-box setting (a) and continuous performance gains in transfer attacks (b-d).

Contribution of Individual Components. We ablate specific loss terms to quantify each module's impact (Fig. 5). Notably, removing the suppression term (w/o suppress) causes a catastrophic performance drop to near-baseline levels. This confirms our core motivation: without actively suppressing the original embodiment, the highly salient real arm dominates the model's attention, causing severe semantic interference that renders injected adversarial features ineffective. Furthermore, ablating either semantic anchoring (w/o sem inj.) or the visual projection (w/o vis inj.) leads to notable performance degradation. This indicates that while suppressing the real arm is a prerequisite, the patch still fails to establish a robust surrogate identity without both the high-level linguistic concept and the fine-grained textural realism of a "robotic arm." Ultimately, our full VLA-Hijack framework consistently achieves the highest Failure Rate across all tasks (Fig. 5e), demonstrating that creating a perceptual vacuum and filling it with a multimodal phantom embodiment are highly complementary and indispensable for robust cross-architecture transfer.

Optimization Strategy. Beyond individual components, we evaluate optimization strategies using patches generated on *OpenVLA-Spatial* over 500 steps.

First, we analyze the necessity of our Alternating Injection Schedule across different periods τ (Table 4). Replacing this schedule with naive joint optimization (**w/o alt.**) causes the average transfer FR to drop from **74.25%** to **66.25%**. This degradation occurs because simultaneously enforcing high-level semantic concepts and fine-grained visual prototypes forces the pixels into competing up-



(a) UniVLA-Spa. **(b)** UniVLA-Obj. **(c)** UniVLA-Goal **(d)** UniVLA-Lon. **(e)** Average

Fig. 5: Ablation on Algorithmic Components. We report the transfer Failure Rate (FR) of patches generated from *OpenVLA-Spatial* and evaluated across four *UniVLA* tasks (a-d), alongside their overall average (e).

(a) Spatial, OpenVLA-Sp. **(b)** Bridge V2, OpenVLA-7B **(c)** Spatial, UniVLA-Spa. **(d)** Object, UniVLA-Obj. **(e)** Goal, UniVLA-Goal **(f)** Long, UniVLA-Long

Fig. 6: Visualizations of Adversarial Patches. Patches optimized on various surrogate models consistently manifest distinct structural features of physical robotic arms.

date directions, inducing severe gradient conflicts that prevent convergence to a unified surrogate identity. Decoupling these modalities across optimization steps resolves this interference, with $\tau = 5$ providing the optimal update frequency.

Furthermore, we assess the sensitivity of the weighting factor λ (Table 5), which regulates the trade-off between *Proprioceptive Suppression* and *Multi-modal Injection*. As shown, the transfer performance exhibits a distinct peak at $\lambda = 0.2$. When λ is too small (e.g., 0.1), the injected features are insufficient to establish a robust phantom embodiment. Conversely, a larger λ (e.g., 0.5) overwhelms the objective, degrading the suppression term and allowing the real arm's features to re-emerge and cause semantic interference. This confirms that VLA-Hijack relies on a carefully calibrated equilibrium between actively suppressing the original identity and injecting the new one.

4.4 Physical-World Evaluation

To validate the practical threat of VLA-Hijack beyond simulation, we conduct physical-world experiments on two tasks: T_1 (*put the paper into the drawer*) and T_2 (*sort and dispose of the trash*), each with 20 trials. We evaluate on OpenVLA [13] and π_0 [1], both fine-tuned on a collected real-world dataset, where OpenVLA assesses attack effectiveness in physical scenes and π_0—a flow-matching-based multi-view architecture—verifies cross-architecture generalizability. Since precisely synchronizing adversarial observations across multiple virtual cameras in simulation is non-trivial, physical evaluation is the natural choice for π_0. As reported in Table 6 and illustrated in

Table 6: Physical Evaluation

Task	Clean FR / Attack FR	
	OpenVLA	π_0
T_1	0% / 100%	0.0% / 85.0%
T_2	20% / 100%	5.0% / 100.0%
Avg.	10% / 100%	2.5% / 92.5%

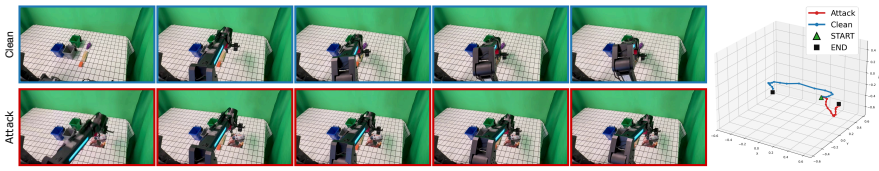


Fig. 7: Physical Evaluation: Action Sequence & 3D Trajectory.

Fig. 7, VLA-Hijack achieves 100% and 92.5% average Failure Rate on T_1 and T_2 respectively, with the attack inducing severe task confusion such as freezing midway or manipulating incorrect objects. These results confirm that VLA-Hijack poses a tangible security threat to real-world VLA deployments across diverse architectures.

4.5 Patch Pattern Analysis

To qualitatively validate the efficacy of our *Proprioceptive Identity Rewiring* mechanism, we visualize the optimized adversarial patches generated from various surrogate models in **Fig. 6**. As observed across all subfigures, whether optimized on simulation tasks (LIBERO) [18] or real-world physical datasets (BridgeData V2) [24], the generated patches consistently exhibit highly structured visual characteristics. Specifically, they strikingly manifest the physical components of robotic arms, displaying clear articulated joint structures, metallic textures, and the geometric shapes of mechanical grippers. Such structural mimicry visually confirms that VLA-Hijack successfully synthesizes a "phantom embodiment" to deceive the model's proprioceptive loop. Detailed visualizations of the hijacked execution rollouts are provided in the Supplementary Materials.

5 Conclusion

In this work, we identify a fundamental vulnerability in VLA models: their critical reliance on a visual proprioception loop. To break the transferability bottleneck of existing action-space attacks, we propose VLA-Hijack, a novel adversarial framework that severs the semantic relationship between the agent's physical embodiment and its control policy. By synergistically coupling attention-guided proprioceptive suppression with alternating multimodal injection, VLA-Hijack synthesizes a "phantom embodiment" to hijack the model's decision-making. Extensive experiments demonstrate that our method achieves rapid white-box convergence and sets a new state-of-the-art for black-box transferability across heterogeneous architectures and cross-domain environments.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.62576109), Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.25511104402).

References

1. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0: A vision-language-action flow model for general robot control. CoRR **abs/2410.24164** (2024)
2. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M.S., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H.T., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: RT-2: vision-language-action models transfer web knowledge to robotic control. CoRR **abs/2307.15818** (2023)
3. Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., Li, H.: Univla: Learning to act anywhere with task-centric latent actions. CoRR **abs/2505.06111** (2025)
4. Carion, N., Gustafson, L., Hu, Y., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts. CoRR **abs/2511.16719** (2025)
5. Chen, Z., Li, B., Wu, S., Ding, S., Zhang, W.: Query-efficient decision-based black-box patch attack. IEEE Trans. Inf. Forensics Secur. **18**, 5522–5536 (2023)
6. Fu, J., Chen, Z., Jiang, K., Guo, H., Gao, S., Zhang, W.: Pg-attack: A precision-guided adversarial attack framework against vision foundation models for autonomous driving. CoRR **abs/2407.13111** (2024)
7. Fu, J., Chen, Z., Jiang, K., Guo, H., Wang, J., Gao, S., Zhang, W.: Improving adversarial transferability of visual-language pre-training models through collaborative multimodal interaction. CoRR **abs/2403.10883** (2024)
8. Fu, J., Jiang, K., Hong, L., Li, J., Guo, H., Yang, D., Chen, Z., Zhang, W.: Lingoloop attack: Trapping mllms via linguistic context and state entrapment into endless loops. CoRR **abs/2506.14493** (2025)
9. Guo, J., Jiang, W., Lin, Y., Liu, Y., Zhang, R., Lu, G., Chen, A., Han, X., Li, H., Niyato, D.: State backdoor: Towards stealthy real-world poisoning attack on vision-language-action model in state space. CoRR **abs/2601.04266** (2026)
10. Jiang, K., Chen, Z., Fu, J., Hong, L., Li, J., Zhang, W.: Videopure: Diffusion-based adversarial purification for video recognition. IEEE Transactions on Circuits and Systems for Video Technology (2025)
11. Jiang, K., Chen, Z., Guo, H., Li, J., Fu, J., Guo, P., Tang, H., Li, B., Zhang, W.: Enhancing diffusion-based unrestricted adversarial attacks via adversary preferences alignment. In: Advances in Neural Information Processing Systems. pp. 167877–167906 (2025)
12. Jiang, K., Chen, Z., Huang, H., Wang, J., Yang, D., Li, B., Wang, Y., Zhang, W.: Efficient decision-based black-box patch attacks on video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4379–4389 (2023)

13. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) Conference on Robot Learning, 6-9 November 2024, Munich, Germany. Proceedings of Machine Learning Research, vol. 270, pp. 2679–2713. PMLR (2024)
14. Li, H., Yang, S., Chen, Y., Tian, Y., Yang, X., Chen, X., Wang, H., Wang, T., Zhao, F., Lin, D., Pang, J.: Cronusvla: Transferring latent motion across time for multi-frame prediction in manipulation. CoRR **abs/2506.19816** (2025)
15. Li, X., Li, P., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Kong, T., Zhang, H., Liu, H.: Towards generalist robot policies: What matters in building vision-language-action models. CoRR **abs/2412.14058** (2024)
16. Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., Li, H., Kong, T.: Vision-language foundation models as effective robot imitators. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
17. Li, X., Fu, P., Huang, W., Pan, N., Yang, S., Zhao, K., Wan, G., Li, M., Xuan, J., Li, M.: When attention betrays: Erasing backdoor attacks in robotic policies by reconstructing visual tokens. arXiv preprint arXiv:2602.03153 (2026)
18. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: LIBERO: benchmarking knowledge transfer for lifelong robot learning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
19. Lu, H., Yu, Y., Yang, Y., Yi, C., Zhang, Q., Shen, B., Kot, A.C., Jiang, X.: When robots obey the patch: Universal transferable patch attacks on vision-language-action models. CoRR **abs/2511.21192** (2025)
20. Lu, X., Chen, J., Xiao, S., Jin, Z., Chen, Z., Yu, H., Qian, B., Zhou, R., Ji, X., Xu, W.: Phantom menace: Exploring and enhancing the robustness of VLA models against physical sensor attacks. CoRR **abs/2511.10008** (2025)
21. Nesti, F., Rossolini, G., Nair, S., Biondi, A., Buttazzo, G.C.: Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022. pp. 2826–2835. IEEE (2022)
22. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., Li, X.: Spatialvla: Exploring spatial representations for visual-language-action model. CoRR **abs/2501.15830** (2025)
23. Ran, Y., Wang, W., Li, M., Li, L., Wang, Y., Li, J.: Cross-shaped adversarial patch attack. IEEE Trans. Circuits Syst. Video Technol. **34**(4), 2289–2303 (2024)
24. Walke, H.R., Black, K., Zhao, T.Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A.W., Myers, V., Kim, M.J., Du, M., Lee, A., Fang, K., Finn, C., Levine, S.: Bridgedata V2: A dataset for robot learning at scale. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA. Proceedings of Machine Learning Research, vol. 229, pp. 1723–1736. PMLR (2023)
25. Wang, T., Liu, D., Liang, J.C., Yang, W., Wang, Q., Han, C., Luo, J., Tang, R.: Exploring the adversarial vulnerabilities of vision-language-action models in robotics. CoRR **abs/2411.13587** (2024)

26. Wang, Y., Zhang, H., Pan, H., Zhou, Z., Wang, X., Guo, P., Xue, L., Hu, S., Li, M., Zhang, L.Y.: Advedm:fine-grained adversarial attack against vlm-based embodied agents. CoRR **abs/2509.16645** (2025)
27. Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F.: Dexvla: Vision-language model with plug-in diffusion expert for general robot control. CoRR **abs/2502.05855** (2025)
28. Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., Peng, Y., Feng, F., Tang, J.: Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics Autom. Lett. **10**(4), 3988–3995 (2025)
29. Xu, B., Shang, Y., Wang, B., Ferrara, E.: Silentdrift: Exploiting action chunking for stealthy backdoor attacks on vision-language-action models. CoRR **abs/2601.14323** (2026)
30. Xu, H., Koh, Y.S., Huang, S., Zhou, Z., Wang, D., Sakuma, J., Zhang, J.: Model-agnostic adversarial attack and defense for vision-language-action models. CoRR **abs/2510.13237** (2025)
31. Xu, Z., Zheng, X., Ma, X., Jiang, Y.: Tabvla: Targeted backdoor attacks on vision-language-action models. CoRR **abs/2510.10932** (2025)
32. Yang, C., Kortylewski, A., Xie, C., Cao, Y., Yuille, A.L.: Patchattack: A black-box texture-based attack with reinforcement learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI. Lecture Notes in Computer Science, vol. 12371, pp. 681–698. Springer (2020)
33. Yang, S., Li, H., Chen, Y., Wang, B., Tian, Y., Wang, T., Wang, H., Zhao, F., Liao, Y., Pang, J.: Instructvla: Vision-language-action instruction tuning from understanding to manipulation. CoRR **abs/2507.17520** (2025)
34. Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., Yang, Y.: Safevla: Towards safety alignment of vision-language-action model via constrained learning. arXiv preprint arXiv:2503.03480 (2025)
35. Zhang, N., Tao, W., Xiao, X., Sun, Q., Zheng, Y., Mo, W., Wang, P., Zhang, N.: Attention-guided patch-wise sparse adversarial attacks on vision-language-action models. CoRR **abs/2511.21663** (2025)
36. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N., Lin, M.: On evaluating adversarial robustness of large vision-language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
37. Zhou, J., Wei, Y., Zhen, R., Zhao, B., Xia, X., Shao, R., Su, X., Yang, S.: Inject once survive later: Backdooring vision-language-action models to persist through downstream fine-tuning. arXiv preprint arXiv:2602.00500 (2026)
38. Zhou, X., Tie, G., Zhang, G., Wang, H., Zhou, P., Sun, L.: Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. CoRR **abs/2505.16640** (2025)
39. Zhou, Z., Xiao, Z., Xu, H., Sun, J., Wang, D., Zhang, J.: Goal-oriented backdoor attack against vision-language-action models via physical objects. CoRR **abs/2510.09269** (2025)
40. Zhu, M., Zhu, Y., Li, J., Zhou, Z., Wen, J., Liu, X., Shen, C., Peng, Y., Feng, F.: Objectvla: End-to-end open-world object manipulation without demonstration. CoRR **abs/2502.19250** (2025)