

Structure-Aware Robust Fine-Tuning: Defending Vision-Language-Action Robots Against Physical Attention Hijacking

Jinquan Zhang^{1,2}, Dongfu Yin^{1,*}, Run Yang^{1,2}, Yufeng Yan^{1,2}, Zhen Tian¹, F. Richard Yu³

Abstract—Vision-Language-Action (VLA) policies promise general robotic manipulation, but their robustness against physical-world attacks remains fragile. In particular, we show that physically realizable adversarial patches can reliably induce failures by triggering a mechanism we call policy-critical action-to-vision attention hijacking, where action-conditioned attention is diverted from task-relevant regions to a localized patch. To demonstrate the threat, we propose Attention-Guided Semantic Disruption (AGSD), an Expectation-over-Transformation (EOT) optimized printable patch that jointly (i) concentrates action-to-vision attention on the patch and (ii) disrupts vision-language semantic alignment, yielding strong cross-task and cross-architecture transfer. To mitigate such attacks, we introduce Structure-Aware Robust Fine-Tuning (SARF), a zero-inference-overhead defense that fine-tunes only the visual encoder using feature anchoring, policy-critical attention correction, and language-guided geometric consistency restricted to semantically relevant regions. On LIBERO, SARF reduces OpenVLA’s failure rate under AGSD from 100% to 14.2–56.8% (28.6% avg.) across suites while preserving clean performance, and on a real PiPER manipulator it improves average success under AGSD from 23.0% to 65.0%. These results highlight mechanism-level robustness as a practical path to securing VLA robots against physical attention hijacking.

I. INTRODUCTION

Vision–Language–Action (VLA) policies map images and natural-language instructions directly to low-level actions and have enabled increasingly general robotic manipulation [1], [2], [3], [4]. Because VLAs close the perception–action loop end-to-end, localized visual failures can immediately translate into erroneous physical behaviors, making robustness a first-order safety concern.

A practical and low-cost threat is the physically realizable adversarial patch: a printable localized pattern that remains effective under viewpoint and illumination changes when optimized with Expectation-over-Transformation (EOT) [5], [6]. Recent studies report that patch attacks can substantially degrade VLA manipulation in both simulation and real settings [7], [8], [9]. Existing defenses (e.g., the defense accompanying EDPA) largely rely on adversarial fine-tuning to align *global* representations between clean and patched observations [10]. However, manipulation often depends on *sparse* task-critical

This work was supported in part by the Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-26420004, and in part by the Shenzhen Science and Technology Program under Grant No. KJZD20240903104400001.

¹Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China.

²Shenzhen University, Shenzhen, China.

³Carleton University, Ottawa, Canada.

*Corresponding author: Dongfu Yin, yindongfu@gml.ac.cn.

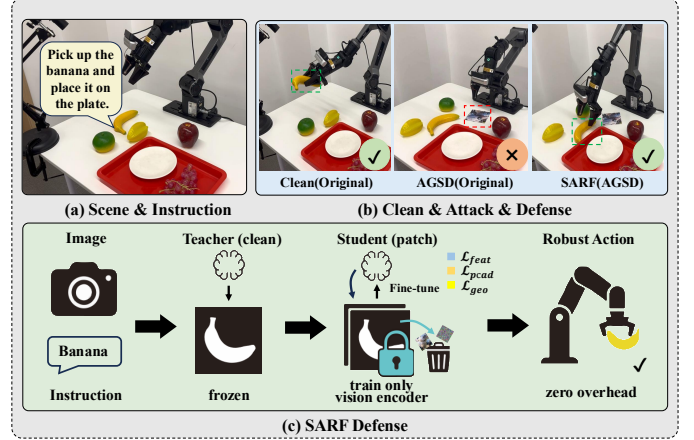


Fig. 1. Motivation and overview of SARF against physical attention hijacking. (a) PiPER tabletop manipulation conditioned on a natural-language instruction. (b) The original VLA succeeds on clean inputs but fails under a printable AGSD patch; SARF restores success under the same attack. (c) SARF fine-tunes only the visual encoder in a teacher–student framework using feature anchoring ($\mathcal{L}_{feat}$), policy-critical action-token attention distillation ($\mathcal{L}_{pcad}$), and language-guided geometric consistency ($\mathcal{L}_{geo}$), with zero inference overhead.

evidence (e.g., end-effector and target object), suggesting that robustness may hinge on stabilizing *where* the policy attends, not only *what* it represents.

We identify a failure mode that is particularly damaging for VLA control: *policy-critical action-to-vision attention hijacking*. In modern VLAs, action outputs are driven by cross-attention from a small set of action-query tokens to visual tokens. A patch can become a strong attention attractor, capturing attention mass from these action queries and suppressing attention on task-relevant regions, thereby derailing long-horizon trajectories and motivating defenses that stabilize action-conditioned attention pathways (Fig. 1).

Guided by this observation, we propose (i) a stress-test attack to expose the mechanism and (ii) a targeted defense. On the attack side, we introduce Attention-Guided Semantic Disruption (AGSD), an EOT-optimized printable patch that jointly concentrates action-to-vision attention on the patch and disrupts vision–language semantic alignment, improving transfer across tasks and architectures. To mitigate this vulnerability, we propose Structure-Aware Robust Fine-Tuning (SARF), a zero-inference-overhead teacher–student robust fine-tuning framework that updates only the visual encoder while freezing the multimodal backbone and action head. SARF combines feature anchoring, policy-critical action-token attention distillation to counteract hijacking, and language-guided geometric consistency restricted to semantically relevant regions.

Our contributions are threefold:

- We present AGSD, an EOT-optimized printable patch that targets policy-critical action-to-vision attention, jointly inducing attention hijacking and semantic misalignment to achieve strong cross-task and cross-architecture transfer.
- We propose SARF, which achieves robust fine-tuning with zero inference overhead by combining feature anchoring, policy-critical attention correction, and language-guided geometric consistency restricted to semantically relevant regions.
- We evaluate our methods on LIBERO [11] and a PiPER tabletop setup, demonstrating the broad vulnerability of multiple VLA architectures (e.g., π_0 and OpenVLA variants) to AGSD, and showing that SARF consistently enhances the robustness of OpenVLA across diverse physical patch scenarios while preserving clean performance.

II. RELATED WORK

A. VLA / Robot Foundation Policies for Manipulation

Vision–Language–Action (VLA) policies enable robots to execute open-vocabulary manipulation by conditioning low-level control on images and natural-language instructions. RT-2 demonstrates that scaling vision–language knowledge can benefit robotic control [1], while OpenVLA provides an open-source foundation VLA trained on large-scale robot demonstrations [2]. Octo further advances open generalist robot policies across diverse robots and sensors [3], and π_0 explores flow-based action generation for general robot control [4]. Complementary lines investigate alternative action modeling and cross-modal fusion for robotic perception [12], [13], [14]. However, current foundation policies primarily prioritize scaling and zero-shot generalization, leaving their vulnerabilities to physical adversarial threats largely underexplored.

B. Physical Patch Attacks and Robustness Evaluation for VLA Systems

Printable adversarial patches, optimized via Expectation-over-Transformation (EOT) [5], [6], pose a realistic physical threat whose perception errors are often amplified by closed-loop robotic execution. Recent evaluation efforts highlight that VLA robustness under physical variations and adversarial conditions remains underexplored and can degrade dramatically under patch attacks [7], [8]. Beyond general robustness evaluation, VLA-specific adversarial studies propose patch objectives tailored to robotic action spaces, including UADA/UPA-style attacks that directly induce action-level discrepancies [9]. EDPA constructs model-agnostic patches by increasing representation discrepancy and image–text misalignment, and proposes a corresponding defense via adversarial fine-tuning [10]. More recent works emphasize transferability and universality across models and sim-to-real conditions (e.g., UPA-RFAS) [15], propose attention-guided sparse attacks for efficiency and stealthiness (e.g., ADVLA) [16], and provide broader benchmarking suites for adversarial and backdoor threats across the

VLA lifecycle (e.g., AttackVLA) [17]. Despite these advances, existing physical attacks mostly target global feature disruption or text-to-vision misalignment, overlooking the specific attention pathways that drive VLA action generation. To expose this blind spot, our AGSD explicitly exploits policy-critical action-to-vision attention combined with semantic disruption, achieving strong transferability.

C. Defenses and Robust Fine-Tuning Against Patch Attacks

Classical defenses against adversarial perturbations include adversarial training [18], [19], data augmentation and domain randomization [20], and input-time purification using generative models [21], [22], [23]. For localized patch threats, PatchGuard provides provable robustness by combining small receptive fields with robust masking [24]. However, purification-style defenses often introduce inference overhead and must be carefully evaluated against adaptive attacks to avoid a false sense of security [25]. In the VLA setting, EDPA’s defense improves robustness by adversarially fine-tuning the visual encoder to better align clean and patched representations [10], but it does not explicitly stabilize the *attention mechanism* that determines where the policy looks when producing actions. Meanwhile, robust/continual adaptation methods aim to preserve pretrained capabilities during fine-tuning, e.g., via regularization or distillation [26], [27], attention transfer [28], or parameter merging to improve robustness without extensive retraining [29]. Crucially, prior defenses in VLA settings either introduce inference overhead or focus solely on aligning global representations, failing to stabilize the fundamental attention mechanisms that determine action outputs. To overcome this, our SARF framework achieves zero-inference-overhead robustness by directly correcting policy-critical attention and enforcing localized geometric consistency.

III. METHODOLOGY

We formalize the physical threat model, introduce AGSD, and then present SARF for improving VLA robustness against physical attention hijacking.

A. Attention-Guided Semantic Disruption (AGSD)

We model the physical attack as a printable, localized adversarial patch optimization problem. Given the original RGB observation $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, the patched adversarial input $\mathbf{x}'$ is defined as

$$\mathbf{x}' = (\mathbf{1} - \mathbf{m}) \odot \mathbf{x} + \mathbf{m} \odot \boldsymbol{\delta}, \quad (1)$$

where $\boldsymbol{\delta}$ denotes the adversarial patch, $\mathbf{m} \in \{0, 1\}^{H \times W}$ is a binary spatial mask indicating the patch location, and $\odot$ is the element-wise product.

The attacker aims to find an optimal patch $\boldsymbol{\delta}^*$ that remains effective under common physical imaging variations. We adopt the Expectation over Transformation (EOT) framework:

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \mathbb{E}_{t \sim \mathcal{T}} [\mathcal{L}_{\text{AGSD}}(t(\mathbf{x}'))], \quad (2)$$

where $t(\cdot)$ is sampled from a transformation distribution $\mathcal{T}$ (e.g., random rotation, translation, and perspective tilting), and

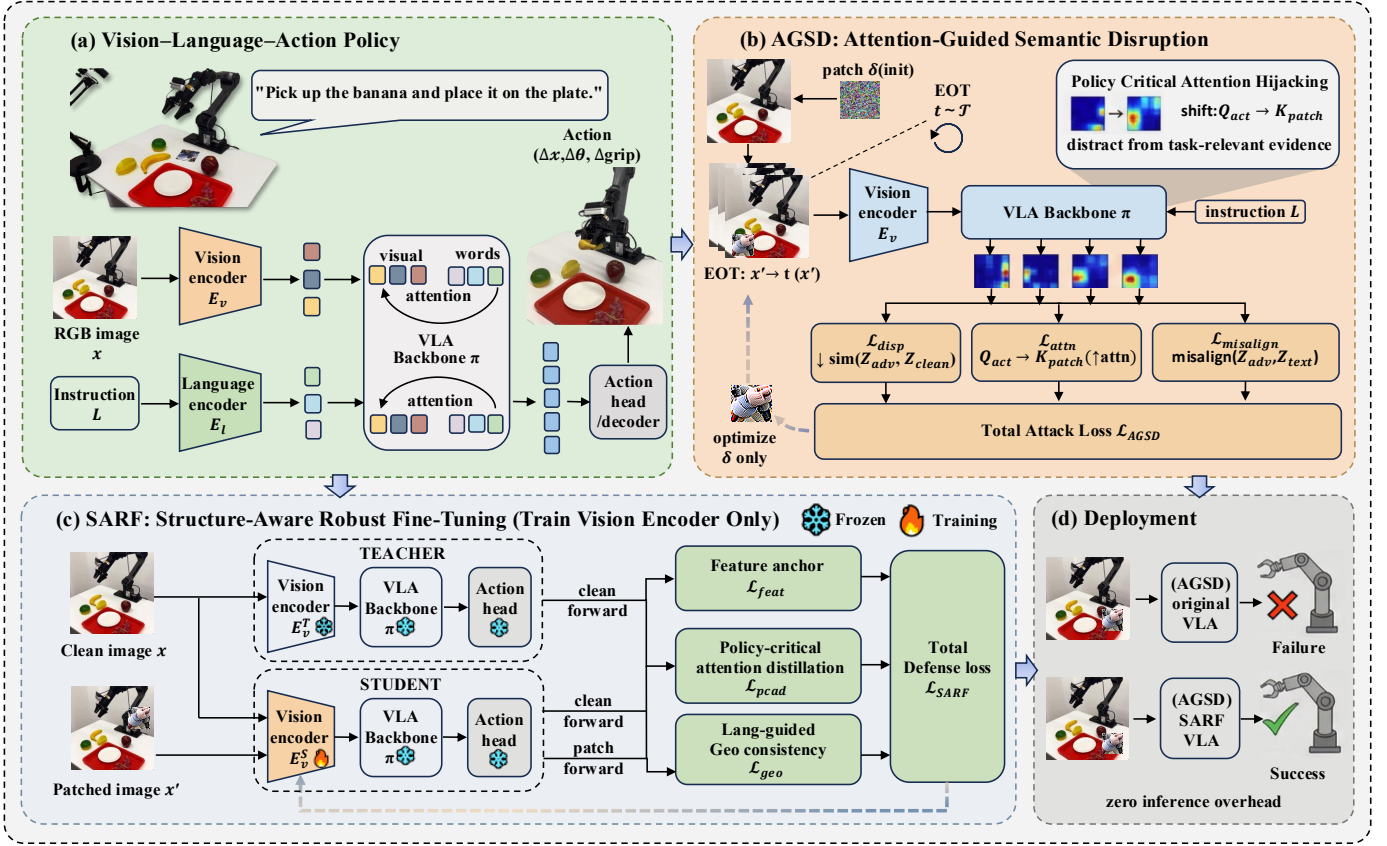


Fig. 2. Physical attention hijacking and our attack–defense pipeline. (a) A VLA policy encodes an RGB observation $\mathbf{x}$ and instruction L and produces low-level actions via cross-attention. (b) AGSD optimizes a printable patch δ under EOT to hijack *action-to-vision* attention and disrupt vision–language alignment. (c) SARF performs teacher–student robust fine-tuning that updates only the student visual encoder E_v^S (freezing π and the action head) using feature anchoring $\mathcal{L}_{feat}$, policy-critical attention distillation $\mathcal{L}_{pcad}$, and language-guided geometric consistency $\mathcal{L}_{geo}$. (d) At deployment, SARF restores successful execution under the same patch with zero inference overhead.

the expectation encourages robustness in real-world deployment.

To effectively attack a VLA model, we define the attack total loss as a weighted combination of three terms:

$$\mathcal{L}_{AGSD} = \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} - \lambda_{\text{disp}} \mathcal{L}_{\text{disp}} - \lambda_{\text{misalign}} \mathcal{L}_{\text{misalign}}, \quad (3)$$

where λ_{attn} , λ_{disp} , and $\lambda_{\text{misalign}}$ are balancing coefficients. Note that $\mathcal{L}_{\text{disp}}$ and $\mathcal{L}_{\text{misalign}}$ are subtracted in Eq. (3), meaning that minimizing $\mathcal{L}_{AGSD}$ effectively *maximizes* these two losses, thus disrupting feature consistency and image–text alignment. We define each component as follows.

1) *Attention Guidance Loss* ($\mathcal{L}_{\text{attn}}$): This loss directly manipulates the cross-modal attention mechanism to “hijack” the model’s perceptual focus. It forces the policy-critical action queries to over-attend to the patch region, thereby diverting action-to-vision cross-attention away from task-relevant evidence:

$$\mathcal{L}_{\text{attn}} = -\frac{1}{|Q_{\text{act}}| |K_{\text{patch}}|} \sum_{q \in Q_{\text{act}}} \sum_{k \in K_{\text{patch}}} \bar{A}_{q,k}, \quad (4)$$

where Q_{act} denotes the set of policy action-query tokens that condition the action head, K_{patch} is the set of visual key tokens inside the patch region (determined by $\mathbf{m}$), and $\bar{A}_{q,k}$ is the mean cross-attention weight aggregated from the last three

cross-attention layers. The leading negative sign maximizes the attention weight from the action queries to the adversarial patch.

2) *Feature Dispersion Loss* ($\mathcal{L}_{\text{disp}}$): To destabilize semantic features, we adopt an InfoNCE-style objective:

$$\mathcal{L}_{\text{disp}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_{\text{adv}}, \mathbf{z}_{\text{clean}}) / \tau_{\text{ncc}})}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{z}_{\text{adv}}, \mathbf{z}_{\text{clean}}^{(j)}) / \tau_{\text{ncc}})}, \quad (5)$$

where $\mathbf{z}_{\text{adv}}$ and $\mathbf{z}_{\text{clean}}$ denote the embeddings of the adversarial and clean images, B is the batch size, τ_{ncc} is the temperature, and $\text{sim}(\cdot, \cdot)$ is cosine similarity. Since $\mathcal{L}_{\text{disp}}$ is subtracted in Eq. (3), the optimizer effectively maximizes it, which reduces the similarity between $\mathbf{z}_{\text{adv}}$ and $\mathbf{z}_{\text{clean}}$ and induces strong feature-space perturbations.

3) *Image–Text Misalignment Loss* ($\mathcal{L}_{\text{misalign}}$): This loss aims to maximize image–text misalignment in the joint embedding space:

$$\mathcal{L}_{\text{misalign}} = \frac{1}{B} \sum_{i=1}^B \left\| \text{sim}(\mathbf{z}_{\text{adv}}^{(i)}, \mathbf{z}_{\text{text}}^{(i)}) - \text{sim}(\mathbf{z}_{\text{clean}}^{(i)}, \mathbf{z}_{\text{text}}^{(i)}) \right\|_1, \quad (6)$$

where $\mathbf{z}_{\text{text}}^{(i)}$ is the embedding of the i -th text instruction. By maximizing this L_1 distance (implicitly via Eq. (3)), the attacker forces the image–text matching of the adversarial

image to deviate significantly from the correct alignment of the clean image.

B. Proposed Defense: Structure-Aware Robust Fine-Tuning (SARF)

To defend against the above patch attack, we propose SARF. SARF fine-tunes only the visual encoder of the VLA model and uses a Teacher model pretrained on clean data to guide the recovery of a perturbed Student model. The overall SARF objective consists of feature anchoring, attention rectification, and geometric consistency:

$$\mathcal{L}_{\text{SARF}} = \lambda_{\text{feat}} \mathcal{L}_{\text{feat}} + \lambda_{\text{pcad}} \mathcal{L}_{\text{pcad}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}. \quad (7)$$

1) *Feature Anchor Loss* ($\mathcal{L}_{\text{feat}}$): To prevent catastrophic forgetting during fine-tuning, we enforce the Student’s high-level semantic features to remain directionally consistent with the Teacher:

$$\mathcal{L}_{\text{feat}} = 1 - \frac{1}{N} \sum_{i=1}^N \frac{\mathbf{z}_S^{(i)} \cdot \mathbf{z}_T^{(i)}}{\|\mathbf{z}_S^{(i)}\| \|\mathbf{z}_T^{(i)}\|}, \quad (8)$$

where N is the number of visual patch tokens, and $\mathbf{z}_S^{(i)}$ and $\mathbf{z}_T^{(i)}$ are the i -th patch-level feature vectors from the Student and Teacher, respectively. This loss anchors the feature space by minimizing the cosine distance ($1 - \cos \theta$).

2) *Policy-Critical Attention Distillation* ($\mathcal{L}_{\text{pcad}}$): Motivated by the VLA policy structure, we distill attention only for Action Tokens that determine action outputs. We align Student and Teacher attention distributions using a symmetric Jensen–Shannon Divergence (JSD) objective:

$$\mathcal{L}_{\text{pcad}} = \frac{1}{H|\mathcal{Q}_{\text{act}}|} \sum_{h=1}^H \sum_{q \in \mathcal{Q}_{\text{act}}} D_{\text{JS}}(P_T^{(h)} \| P_S^{(h)}), \quad (9)$$

where H is the number of heads and $\mathcal{Q}_{\text{act}}$ indexes Action tokens. For each h and $q \in \mathcal{Q}_{\text{act}}$, we form attention distributions over visual tokens $P^{(h)}(k | q) = \text{Softmax}(A_{q,k}^{(h)} / \tau_{\text{attn}})$ and minimize $D_{\text{JS}}(P_T^{(h)}(\cdot | q) \| P_S^{(h)}(\cdot | q))$ to align policy-critical action-to-vision attention.

3) *Language-Guided Geometric Consistency* ($\mathcal{L}_{\text{geo}}$): To preserve the geometry of task-relevant objects while suppressing background noise, we introduce a language-guided geometric mask. We first compute the importance of each patch i using Teacher text-to-vision attention: $m_i = \max_{q \in \mathcal{Q}_{\text{txt}}} A_{q,i}^T$, and construct a sharpened pairwise mask $\mathbf{M}_{ij} = (m_i \cdot m_j)^2$. Since this mask is derived from the clean Teacher stream rather than the perturbed Student stream, it provides a stable reference for task-relevant regions.

Based on this mask, the geometric loss is defined as

$$\mathcal{L}_{\text{geo}} = \frac{\sum_{i,j} \mathbf{M}_{ij} \cdot (G_{ij}^S - G_{ij}^T)^2}{\sum_{i,j} \mathbf{M}_{ij} + \epsilon}, \quad (10)$$

where the Gram element G_{ij} captures pairwise feature relations (texture and structure):

$$G_{ij} = \frac{\mathbf{z}^{(i)} \cdot \mathbf{z}^{(j)}}{\|\mathbf{z}^{(i)}\| \|\mathbf{z}^{(j)}\|}. \quad (11)$$

This loss constrains only the regions activated by $\mathbf{M}$, enabling structure denoising on task-critical objects (e.g., the end-effector and manipulated objects) while automatically down-weighting background perturbations.

IV. EXPERIMENTS

We first demonstrate that physical patch attacks are feasible and can severely degrade Vision–Language–Action (VLA) robots. We then analyze the underlying failure mechanism via attention-level diagnostics, propose SARF as a targeted solution, and validate robustness both in simulation and on a real PiPER system. Unless otherwise specified, results are reported on four LIBERO suites (*Spatial*, *Object*, *Goal*, *Long*).

A. Experimental Setup

1) *Benchmarks*: We evaluate on four LIBERO simulation task suites (reporting Failure Rate, FR) and a real-world PiPER tabletop setup (reporting Success Rate, SR). For PiPER, we conduct 100 independent trials per condition with randomized initial object poses, camera viewpoints, and distances to assess physical robustness.

2) *Models*: We evaluate representative VLA policies, including OpenVLA and its fine-tuned variant OpenVLA-ofT, as well as a stronger foundation-policy baseline π_0 (Pi0). All models are evaluated using their default action interfaces and inference settings to ensure fair comparison.

3) *Attacks*: We consider localized, printable adversarial patches and compare: (i) Clean (no patch), (ii) Random (non-optimized printable patch), (iii) UADA and UPA from [9], (iv) EDPA from [10], and (v) AGSD (ours). Unless otherwise stated, all methods use the same patch size and placement distribution.

4) *Defenses*: We report three settings on OpenVLA: Original (no defense), AF (Defense Method in the EDPA), and SARF (ours). SARF fine-tunes only the *visual encoder* and distills *policy-critical* action-to-vision attention from a clean Teacher model, with language-guided geometric consistency restricted to semantically relevant regions. We instantiate SARF on OpenVLA because its action-conditioned visual attention is accessible; extension to diffusion- or flow-style decoders requires identifying analogous policy-critical visual pathways.

5) *Metrics*: For LIBERO, we report Failure Rate (FR, %). For the real robot, we report Success Rate (SR, %). When evaluating attack effectiveness (e.g., AGSD), a higher FR indicates a stronger attack (worse robustness). When evaluating defense effectiveness (e.g., SARF), a lower FR indicates better robustness while preserving performance on clean inputs.

6) *Implementation details and EOT settings*: To ensure stable optimization by balancing inherent gradient magnitudes and prioritizing critical objectives, we set the AGSD coefficients to $\lambda_{\text{attn}} = 0.8$, $\lambda_{\text{disp}} = 0.2$, and $\lambda_{\text{misalign}} = 0.5$. These values are selected such that the attention-guidance loss acts as the primary driver of optimization, while the remaining terms provide auxiliary semantic disruption. For the SARF defense, we set $\lambda_{\text{feat}} = 0.5$, $\lambda_{\text{pcad}} = 1.0$, and $\lambda_{\text{geo}} = 0.3$. This

configuration reflects our core design principle: prioritizing the distillation of policy-critical attention ($\mathcal{L}_{\text{pcad}}$) to counteract hijacking, while leveraging $\mathcal{L}_{\text{feat}}$ and $\mathcal{L}_{\text{geo}}$ as structure-aware constraints to preserve the pretrained visual backbone’s integrity. In LIBERO, the adversarial patch occupies 5% of the image area. On the real PiPER robot, we evaluate printed physical patches with sizes of 20×20 cm, 15×15 cm, 8×8 cm, and 5×5 cm. To ensure physical realizability, all attacks are optimized and evaluated under an Expectation-over-Transformation (EOT) framework that incorporates random in-plane rotation ($\theta \sim \mathcal{U}(-30^\circ, 30^\circ)$), translation (up to 10% of the image dimensions), scaling ($s \sim \mathcal{U}(0.9, 1.1)$), random patch placement (uniformly sampled within the image bounds), and mild perspective/illumination jitter. SARF uses the same data budget as AF and updates only the visual encoder; the deployed architecture and inference latency remain unchanged.

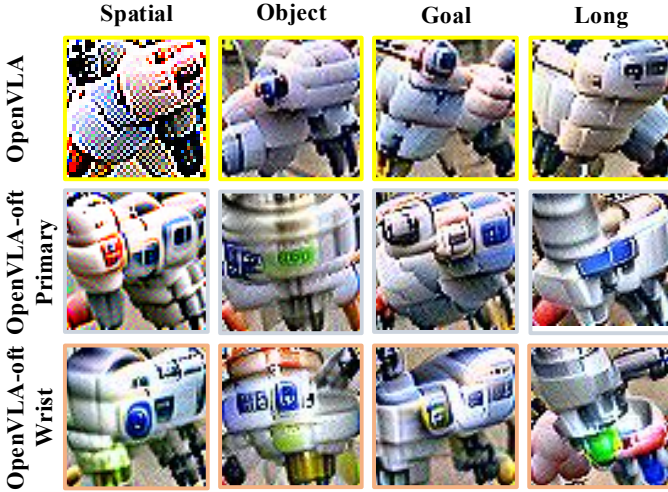


Fig. 3. Printable AGSD patch examples. Representative AGSD patches optimized under the same EOT pipeline. Columns correspond to the four LIBERO suites (*Spatial*, *Object*, *Goal*, *Long*); rows show patches optimized for OpenVLA and OpenVLA-ofT using primary and wrist camera observations. These patches are used in our physical patch-attack evaluations.

B. Physical Attention Hijacking

1) *Quantitative results of attack effectiveness*: As shown in Table I, even non-optimized *Random* patches already increase failure rates compared to *Clean* inputs, indicating that localized physical perturbations can impair manipulation. However, optimized patches are substantially more damaging; AGSD drives *OpenVLA* to 100% failure across all suites, improving over *Random* by 24.4–64.2 points. Crucially, AGSD transfers beyond the attacked model: on *OpenVLA-ofT*, failure rises from 10.0–32.0% (*Random*) to 93.6–100% (AGSD), and on the stronger π_0 baseline from 5.2–48.6% (*Random*) to 48.8–80.2% (AGSD). Relative to EDPA, AGSD is comparable on OpenVLA while remaining markedly stronger on OpenVLA-ofT and π_0 , supporting cross-architecture transfer. Consistently, Table II (Original column) confirms that, under the same evaluation protocol used for defense studies, the undefended

OpenVLA exhibits near-certain failure under AGSD across all suites.

TABLE I
ATTACK EFFECTIVENESS ON LIBERO (FAILURE RATE, %). FAILURE RATES (MEAN $\pm$ STD) OF OPENVLA, OPENVLA-OFT, AND π_0 UNDER CLEAN, RANDOM, AND OPTIMIZED PATCH ATTACKS (EDPA, AGSD) ACROSS FOUR SUITES. HIGHER FR INDICATES STRONGER ATTACKS; BEST (HIGHEST) RESULTS AMONG OPTIMIZED ATTACKS ARE IN **BOLD**.

Suite	Method	Failure Rate (FR $\uparrow$)		
		OpenVLA	OpenVLA-ofT	π_0
Spatial	Clean	14.2 $\pm$ 0.5	2.4 $\pm$ 0.4	3.4 $\pm$ 0.3
	Random	35.8 $\pm$ 1.3	10.0 $\pm$ 1.1	6.0 $\pm$ 0.8
	EDPA[10]	100 $\pm$ 0.0	39.7 $\pm$ 0.9	29.8 $\pm$ 1.6
	AGSD (Ours)	100 $\pm$ 0.0	97.2 $\pm$ 0.6	48.8 $\pm$ 2.1
Object	Clean	11.6 $\pm$ 0.4	2.6 $\pm$ 0.3	2.0 $\pm$ 0.2
	Random	44.6 $\pm$ 1.2	20.4 $\pm$ 1.5	5.2 $\pm$ 0.5
	EDPA[10]	100 $\pm$ 0.0	52.3 $\pm$ 0.8	39.5 $\pm$ 1.7
	AGSD (Ours)	100 $\pm$ 0.0	93.6 $\pm$ 1.2	50.4 $\pm$ 2.4
Goal	Clean	20.8 $\pm$ 1.5	3.0 $\pm$ 0.6	10.4 $\pm$ 1.1
	Random	42.0 $\pm$ 1.2	16.2 $\pm$ 1.3	16.8 $\pm$ 1.4
	EDPA[10]	100 $\pm$ 0.0	80.8 $\pm$ 0.4	44.3 $\pm$ 2.0
	AGSD (Ours)	100 $\pm$ 0.0	100 $\pm$ 0.0	70.8 $\pm$ 1.8
Long	Clean	46.2 $\pm$ 2.0	4.8 $\pm$ 0.7	42.0 $\pm$ 1.6
	Random	75.6 $\pm$ 2.4	32.0 $\pm$ 2.1	48.6 $\pm$ 1.9
	EDPA[10]	100 $\pm$ 0.0	86.4 $\pm$ 1.9	70.7 $\pm$ 1.6
	AGSD (Ours)	100 $\pm$ 0.0	100 $\pm$ 0.0	80.2 $\pm$ 1.5

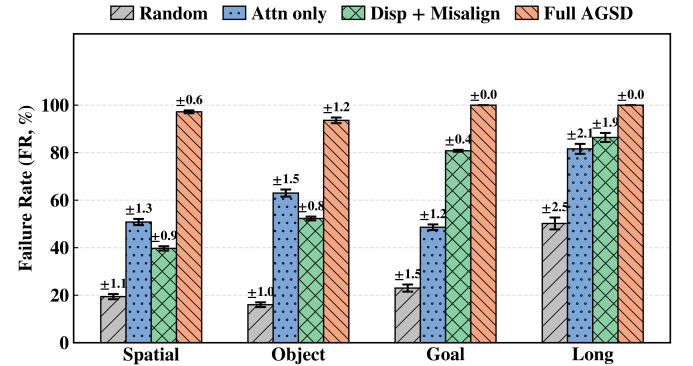


Fig. 4. AGSD objective ablation. Failure rates (FR, %) of OpenVLA on four LIBERO suites under four printable patch variants: Random, Attn-only, Disp+Misalign, and Full AGSD. Higher FR indicates a stronger attack; error bars denote variability across repeated evaluations.

2) *AGSD produces printable and deployable patches*: To illustrate physical plausibility, Fig. 3 visualizes representative *printable* adversarial patches optimized by AGSD against OpenVLA and OpenVLA-ofT under the same EOT transformation pipeline. Across suites and viewpoints (primary and wrist cameras), the optimized patterns are concrete and reproducible, supporting real-world deployment rather than relying on imperceptible digital noise.

3) *Mechanistic insight via objective ablation*: To understand *why* AGSD is particularly effective, we conduct an objective-level ablation in Fig. 4. Optimizing only the

attention-guidance term (Attn-only) already increases FR, indicating that explicitly manipulating *policy-critical action-to-vision attention* can drive failures. Optimizing semantic disruption without explicit attention guidance (Disp + Misalign) also degrades performance by destabilizing visual features and vision-language alignment. Crucially, the Full AGSD objective that couples attention guidance with semantic disruption achieves the strongest attacks across all LIBERO suites, suggesting that these components are complementary and together yield more transferable physical patches.

4) *Qualitative attention hijacking under AGSD*: While Table I and Fig. 4 establish the strength and contributing factors of AGSD, we further provide qualitative attention evidence in Fig. 5. Under AGSD, the Original policy’s action-to-vision attention repeatedly collapses onto the patch region across rollout timesteps, diverting focus away from task-relevant evidence. This motivates defenses that explicitly stabilize policy-critical attention pathways.

C. SARF for Stabilizing Policy-Critical Attention

1) *Goal and evaluation protocol*: We evaluate SARF as a *drop-in* defense against physically realizable patch attacks: it introduces *zero inference overhead* (no test-time modules and unchanged policy interface) and updates only the visual encoder during training.

TABLE II
OPENVLA ROBUSTNESS UNDER PHYSICAL PATCH ATTACKS (FAILURE RATE,%). FAILURE RATES (MEAN $\pm$ STD) ON FOUR LIBERO SUITES FOR ORIGINAL, AF(DEFENSE METHOD IN THE EDPA), AND SARF UNDER CLEAN, RANDOM, AND OPTIMIZED ATTACKS (UADA/UPA/EDPA/AGSD); LOWER IS BETTER AND BEST IS IN BOLD. SARF-AGSD USES AN ADAPTIVE (RE-OPTIMIZED) AGSD PATCH AGAINST THE SARF-TUNED MODEL.

Suite	Method	Failure Rate (FR, $\downarrow$)		
		Original	AF[10]	SARF (Ours)
Spatial	Clean	14.2 $\pm$ 0.5	17.9 $\pm$ 0.8	14.4 $\pm$ 0.6
	Random	35.8 $\pm$ 1.3	19.4 $\pm$ 1.1	15.0 $\pm$ 0.7
	UADA[9]	98.8 $\pm$ 0.2	65.4 $\pm$ 2.5	16.2 $\pm$ 0.9
	UPA[9]	99.0 $\pm$ 0.1	46.6 $\pm$ 2.1	15.8 $\pm$ 0.8
	EDPA[10]	100 $\pm$ 0.0	39.4 $\pm$ 1.8	16.5 $\pm$ 1.0
	AGSD (Ours)	100 $\pm$ 0.0	90.2 $\pm$ 1.5	17.0 $\pm$ 1.1
Object	Clean	11.6 $\pm$ 0.4	17.3 $\pm$ 0.9	11.8 $\pm$ 0.5
	Random	44.6 $\pm$ 1.2	16.0 $\pm$ 1.0	12.4 $\pm$ 0.6
	UADA[9]	92.0 $\pm$ 1.5	58.8 $\pm$ 2.3	13.2 $\pm$ 0.8
	UPA[9]	94.2 $\pm$ 1.1	43.9 $\pm$ 2.0	12.8 $\pm$ 0.7
	EDPA[10]	100 $\pm$ 0.0	58.6 $\pm$ 2.4	13.6 $\pm$ 0.9
	AGSD (Ours)	100 $\pm$ 0.0	99.8 $\pm$ 0.2	14.2 $\pm$ 1.0
Goal	Clean	20.8 $\pm$ 1.5	22.8 $\pm$ 1.2	21.0 $\pm$ 1.4
	Random	42.0 $\pm$ 1.2	23.0 $\pm$ 1.5	22.2 $\pm$ 1.6
	UADA[9]	98.6 $\pm$ 0.4	91.6 $\pm$ 1.8	24.6 $\pm$ 1.7
	UPA[9]	96.4 $\pm$ 0.8	68.3 $\pm$ 2.2	23.8 $\pm$ 1.5
	EDPA[10]	100 $\pm$ 0.0	73.9 $\pm$ 2.5	25.4 $\pm$ 1.8
	AGSD (Ours)	100 $\pm$ 0.0	97.2 $\pm$ 0.6	26.5 $\pm$ 2.0
Long	Clean	46.2 $\pm$ 2.0	49.0 $\pm$ 2.1	46.6 $\pm$ 1.9
	Random	75.6 $\pm$ 2.4	50.2 $\pm$ 2.5	48.8 $\pm$ 2.2
	UADA[9]	99.8 $\pm$ 0.2	97.4 $\pm$ 0.8	53.2 $\pm$ 2.6
	UPA[9]	99.8 $\pm$ 0.1	86.7 $\pm$ 1.5	51.5 $\pm$ 2.4
	EDPA[10]	100 $\pm$ 0.0	91.2 $\pm$ 1.2	54.0 $\pm$ 2.7
	AGSD (Ours)	100 $\pm$ 0.0	99.0 $\pm$ 0.5	56.8 $\pm$ 2.9

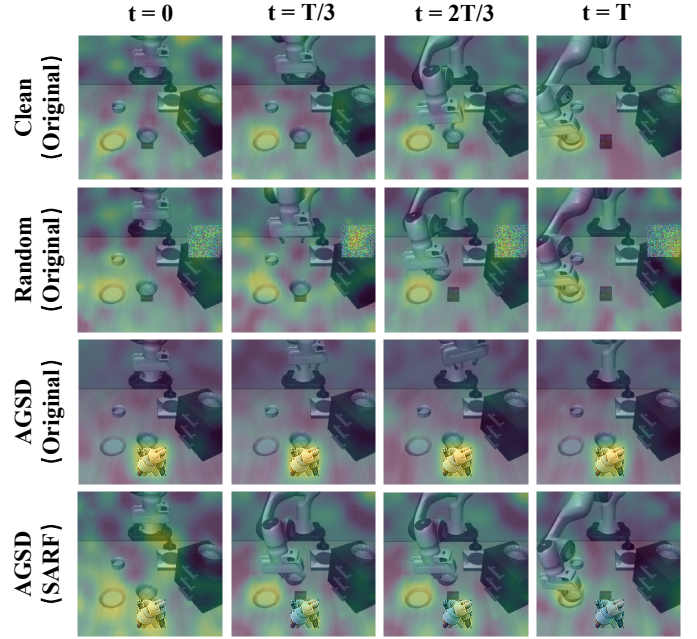


Fig. 5. AGSD hijacks policy-critical attention, while SARF restores task-relevant focus. Action-to-vision cross-attention heatmaps overlaid on the RGB observation at four rollout timesteps ($t = 0, T/3, 2T/3, T$). Rows show Clean/Random baselines (Original), AGSD-attacked Original, and AGSD-attacked SARF. AGSD causes attention to collapse onto the patch region across time, whereas SARF suppresses patch fixation and re-centers attention on task-relevant evidence.

2) *Main robustness gains under strong transferable patches*: Table II reports OpenVLA failure rates (FR; lower is better) under diverse physical patch attacks. Under the strongest stress-test attack, AGSD, the *undefended* policy fails almost always (FR= 100%) on all four suites. In contrast, SARF reduces FR to 17.0% / 14.2% / 26.5% / 56.8% on *Spatial/Object/Goal/Long*, corresponding to absolute drops of 83.0 / 85.8 / 73.5 / 43.2 points, respectively. Averaged across suites, SARF decreases FR from 100.0% to 28.6% under AGSD (a 71.4-point reduction). The higher failure rate on the *Long* suite suggests that minor residual attention jitter can compound over long-horizon trajectories, though the improvement over the undefended baseline remains substantial.

3) *Consistent improvements across attack baselines*: Beyond AGSD, SARF yields consistent robustness gains across prior patch baselines. Averaged over suites, FR decreases from 97.3% (UADA) to 26.8%, from 97.4% (UPA) to 26.0%, and from 100.0% (EDPA) to 27.4%. These results indicate that SARF improves robustness broadly under diverse optimization objectives and transfer settings.

4) *Comparison to adversarial fine-tuning (AF)*: Compared with AF (the EDPA defense), SARF achieves markedly lower FR under strong attacks. For example, under AGSD, AF remains highly vulnerable with an average FR of 96.6%, whereas SARF reduces it to 28.6% (an additional 67.9-point reduction over AF). Similar margins hold for EDPA (65.8% to 27.4%) and UADA (78.3% to 26.8%) on average. This massive performance gap highlights that for end-to-end visuo-motor policies, stabilizing **where** the model looks is far more

effective than broadly aligning **what** the model represents.

5) *Robustness under an adaptive attacker*: Importantly, the AGSD results for SARF are obtained using an *adaptive* setting where the patch is re-optimized against the frozen SARF-tuned model using the same AGSD objective and EOT settings. Even under this stronger threat model, SARF maintains substantially lower FR than both the undefended policy and AF.

6) *Clean performance is preserved*: Robustness gains do not come at the cost of clean-task degradation: on clean inputs, FR changes only marginally from 14.2/11.6/20.8/46.2% (Original) to 14.4/11.8/21.0/46.6% (SARF) across suites (average 23.2% to 23.5%).

7) *Mechanism-level evidence: attention is re-centered under attack*: Fig. 5 provides an attention-level diagnosis consistent with the quantitative improvements. Under AGSD, the original policy exhibits persistent patch fixation where action-to-vision cross-attention collapses onto the patch region across rollout timesteps. After SARF, attention is re-centered on semantically relevant regions under the *same* physical patch, supporting the mechanism-targeted explanation that stabilizing policy-critical attention helps prevent long-horizon failures.

D. Real-Robot Validation on PiPER

1) *Protocol and metric*: We evaluate three real-world manipulation tasks on a PiPER tabletop platform (Table III) and report Success Rate (SR, %). For each task, we run 100 physical trials with randomized initial object poses and randomized trial order. We report clean performance for the undefended policy (Original), and evaluate robustness under a printed AGSD patch by comparing Original, AF [10], and SARF (ours). To reflect realistic sensing variations, we vary the camera viewpoint and distance across trials.

TABLE III

REAL-ROBOT RESULTS ON PiPER UNDER A PRINTED AGSD PATCH (SUCCESS RATE, %). SUCCESS RATES ARE AVERAGED OVER 100 PHYSICAL TRIALS PER CONDITION. WE COMPARE THE UNDEFENDED POLICY (ORIGINAL), THE STRONGEST SIMULATION BASELINE (AF), AND OUR SARF. SARF DECISIVELY RESTORES MANIPULATION CAPABILITIES UNDER THE SAME PHYSICAL PATCH.

Task	Clean	Under AGSD Attack		
	Original	Original	AF [10]	SARF (Ours)
pick & place	79.0	32.0	51.0	74.0
open drawer	71.0	23.0	42.0	63.0
stack	66.0	14.0	29.0	58.0
Average	72.0	23.0	40.7	65.0

2) *SARF recovers real-world performance under a printed AGSD patch*: As shown in Table III, the printed AGSD patch substantially degrades real-world performance, reducing the average success rate of the undefended policy (Original) from 72.0% (clean) to 23.0% under attack. The simulation baseline AF [10] partially improves robustness (40.7% on average), whereas SARF restores manipulation much more effectively under the *same* printed patch, achieving 65.0% average success (a 42.0-point gain over the attacked Original). The gains are

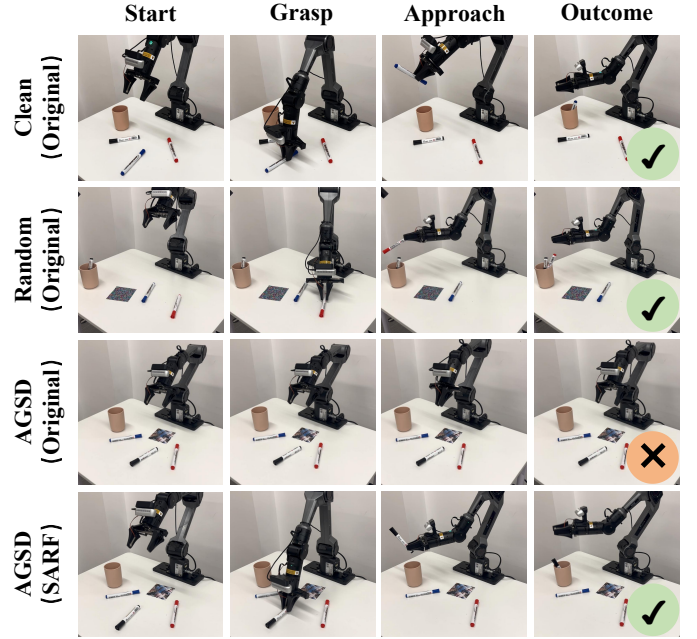


Fig. 6. Real-robot qualitative comparison under physical patch attacks (PiPER tabletop). Four keyframes (Start/Grasp/Approach/Outcome) under Clean and Random baselines (Original), and under an AGSD patch with the Original and SARF policies. The AGSD patch causes failure for the original policy, while SARF restores successful execution under the same printed patch (zero inference overhead).

consistent across tasks: for *pick & place*, SARF reaches 74.0% under attack, close to the clean performance of 79.0%; for the precision-heavy *stack* task, SARF improves success from 14.0% to 58.0% (a 44.0-point gain), although an 8.0-point gap to the clean setting remains. These results suggest that SARF mitigates physical attention hijacking in the real world, although residual errors can still arise from remaining attention leakage to the patch under challenging sensing variations.

E. Ablation Study

To validate the individual contribution of each SARF component towards defending against physical attacks and preserving original capabilities, we conduct an ablation study on LIBERO-Spatial (Table IV). While Full SARF achieves the strongest robustness under Random Patch and AGSD Attack while maintaining near-original clean performance, removing $\mathcal{L}_{\text{pcad}}$ collapses robustness under AGSD (FR spikes to 87.5%).

TABLE IV

SARF ABLATION ON LIBERO-SPATIAL (FAILURE RATE, %). FAILURE RATES OF OPENVLA UNDER CLEAN, RANDOM PATCH, AND AGSD ATTACK. LOWER FR INDICATES BETTER PERFORMANCE.

Method	Failure Rate (FR% ↓)		
	Clean	Random Patch	AGSD Attack
only $\mathcal{L}_{\text{pcad}}$	34.7 ± 2.0	41.1 ± 2.4	54.4 ± 2.5
only $\mathcal{L}_{\text{geo}}$	37.9 ± 1.5	59.8 ± 1.9	81.6 ± 1.2
w/o $\mathcal{L}_{\text{feat}}$	31.6 ± 2.5	34.9 ± 2.8	37.2 ± 3.0
w/o $\mathcal{L}_{\text{pcad}}$	14.2 ± 1.1	44.6 ± 2.2	87.5 ± 1.4
w/o $\mathcal{L}_{\text{geo}}$	14.0 ± 1.3	21.5 ± 1.6	35.8 ± 2.1
Full SARF	14.4 ± 0.6	15.0 ± 0.7	17.0 ± 1.1

Furthermore, dropping $\mathcal{L}_{\text{feat}}$ severely harms clean performance (31.6% FR), and removing $\mathcal{L}_{\text{geo}}$ yields a moderate robustness drop (35.8% FR). These results demonstrate that all three objectives are synergistic and indispensable: policy-critical attention correction drives the core defense, while feature anchoring and geometric consistency ensure the stability of the original visuomotor policy.

V. CONCLUSIONS

We characterized physically realizable patch threats to Vision–Language–Action (VLA) manipulation and pinpointed a mechanism-level failure—policy-critical action-to-vision attention hijacking—that causally links localized visual perturbations to compounding control errors. Building on this diagnosis, we introduced AGSD as a stress-test physical attack that combines attention capture with vision–language semantic disruption, and we proposed SARF, a zero-inference-overhead defense that strengthens robustness by updating only the visual encoder while preserving the deployed policy interface. Across LIBERO suites and on a real PiPER platform, SARF consistently improves robustness under AGSD while maintaining clean-task performance, and attention visualizations corroborate that it re-centers action-conditioned attention onto task-relevant evidence during execution. The ablation results further show that policy-critical attention distillation is the dominant factor for attack robustness, while feature anchoring and geometric consistency help preserve clean-task stability. Overall, our findings suggest that mechanism-targeted robustness—stabilizing the attention pathways that directly drive action generation—offers a practical and deployable route to securing VLA robots against physical attacks without adding test-time complexity. In future work, we will extend SARF to natural clutter and distractors, further reduce residual long-horizon attention jitter, and validate the defense across broader VLA backbones, embodiments, and third-party adaptive attacks.

REFERENCES

- [1] B. Zitkovich, T. Yu, S. Xu, *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.
- [2] M. J. Kim, K. Pertsch, S. Karamcheti, *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [3] O. M. Team, D. Ghosh, H. Walke, *et al.*, “Octo: An open-source generalist robot policy,” *arXiv preprint arXiv:2405.12213*, 2024.
- [4] K. Black, N. Brown, D. Driess, *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [5] T. B. Brown, D. Mané, A. Roy, *et al.*, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017.
- [6] A. Athalye, L. Engstrom, A. Ilyas, *et al.*, “Synthesizing robust adversarial examples,” in *International conference on machine learning*. PMLR, 2018, pp. 284–293.
- [7] H. Cheng, E. Xiao, Y. Wang, *et al.*, “Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models,” *arXiv preprint arXiv:2409.13174*, 2024.
- [8] H. Liu, J. Long, J. Wu, *et al.*, “Eva-vla: Evaluating vision-language-action models’ robustness under real-world physical variations,” *arXiv preprint arXiv:2509.18953*, 2025.
- [9] T. Wang, C. Han, J. Liang, *et al.*, “Exploring the adversarial vulnerabilities of vision-language-action models in robotics,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6948–6958.
- [10] H. Xu, Y. S. Koh, S. Huang, *et al.*, “Model-agnostic adversarial attack and defense for vision-language-action models,” *arXiv preprint arXiv:2510.13237*, 2025.
- [11] B. Liu, Y. Zhu, C. Gao, *et al.*, “Libero: Benchmarking knowledge transfer for lifelong robot learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.
- [12] C. Chi, Z. Xu, S. Feng, *et al.*, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, vol. 44, no. 10–11, pp. 1684–1704, 2025.
- [13] H. Cai, D. Yin, F. Yu, *et al.*, “Mft: Modal fusion transformer for cross-modal fusion in 3d object detection,” in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [14] H. Cai, D. Yin, F. R. Yu, *et al.*, “Dstr: Dual scenes transformer for cross-modal fusion in 3d object detection,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 3064–3073.
- [15] H. Lu, Y. Yu, Y. Yang, *et al.*, “When robots obey the patch: Universal transferable patch attacks on vision-language-action models,” *arXiv preprint arXiv:2511.21192*, 2025.
- [16] N. Zhang, W. Tao, X. Xiao, *et al.*, “Attention-guided patch-wise sparse adversarial attacks on vision-language-action models,” *arXiv preprint arXiv:2511.21663*, 2025.
- [17] J. Li, Y. Zhao, X. Zheng, *et al.*, “Attackvla: Benchmarking adversarial and backdoor attacks on vision-language-action models,” *arXiv preprint arXiv:2511.12149*, 2025.
- [18] A. Madry, A. Makelov, L. Schmidt, *et al.*, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.
- [19] H. Zhang, Y. Yu, J. Jiao, *et al.*, “Theoretically principled trade-off between robustness and accuracy,” in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.
- [20] J. Tobin, R. Fong, A. Ray, *et al.*, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.
- [21] W. Nie, B. Guo, Y. Huang, *et al.*, “Diffusion models for adversarial purification,” *arXiv preprint arXiv:2205.07460*, 2022.
- [22] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-gan: Protecting classifiers against adversarial attacks using generative models,” *arXiv preprint arXiv:1805.06605*, 2018.
- [23] Y. Song, T. Kim, S. Nowozin, *et al.*, “Pixeldefend: Leveraging generative models to understand and defend against adversarial examples,” *arXiv preprint arXiv:1710.10766*, 2017.
- [24] C. Xiang, A. N. Bhagoji, V. Schwag, *et al.*, “{PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2237–2254.
- [25] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *International conference on machine learning*. PMLR, 2018, pp. 274–283.
- [26] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [27] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.
- [28] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” *arXiv preprint arXiv:1612.03928*, 2016.
- [29] Y. Yadav, Z. Zhou, A. Wagenmaker, *et al.*, “Robust finetuning of vision-language-action robot policies via parameter merging,” *arXiv preprint arXiv:2512.08333*, 2025.