DOES ADVERSARIAL TRAINING IMPROVE GENERALIZATION IN MULTI-VIEW VLAS? REVEALING AND MITIGATING VIEW COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models adapt pretrained vision-language models (VLMs) for closed-loop robot control, transferring their perceptual and semantic capabilities to action prediction. Despite strong in-distribution performance, however, VLAs often degrade under deployment shifts. Adversarial training (AT) offers a model-adaptive approach to robustness without explicitly anticipating individual shifts, but its effect on natural distribution-shift generalization in multi-view VLAs remains unclear. We study this question using a multi-view VLA directly adapted from a pretrained VLM and evaluate generalization across seven LIBERO-Plus shift axes. Direct AT substantially improves Camera Viewpoint and Sensor Noise, the two shifts affecting only the third-person view, yet produces mixed or negative effects on other shifts. Controlled view interventions reveal a surprising failure mode that we term *view collapse*: Direct AT can shift cross-view reliance so strongly that the policy becomes dominated by the wrist view. This exposes a *robustness shortcut*: apparent robustness to a shifted view can arise from reduced use of that view rather than more robust perception of it. This motivates a distinction between *robust perception*, extracting reliable information under within-view shifts, and *robust fusion*, adapting reliance across views according to their reliability. To reduce fixed view reliance, we use a simple View Swap intervention and then re-evaluate AT. With View Swap, AT further improves Camera Viewpoint, Sensor Noise, and Robot Initial State, while its effects remain mixed on other shifts. Our results show that multi-view robustness requires separating improved perception from changes in cross-view reliance, and that AT provides selective rather than generic distribution-shift benefits.

1 INTRODUCTION

Vision-language-action (VLA) models transfer the perceptual and semantic capabilities of pretrained vision-language models (VLMs) to closed-loop robot control and achieve high success on standardized manipulation benchmarks (Kim et al., 2025b;a; Physical Intelligence, 2025; StarVLA Community, 2026). High in-distribution performance, however, does not ensure reliable deployment. Recent evaluations show substantial degradation under deployment shifts, such as camera viewpoints, sensor noise, lighting, and object layouts (Fei et al., 2026b; Zhou et al., 2025). A key challenge is therefore to develop VLA policies that remain reliable beyond the conditions represented in their limited robot demonstrations.

One effective approach to improving VLA robustness is to incorporate anticipated environmental variations into training through expanded datasets or predefined augmentations (Fei et al., 2026b; Guo et al., 2026). However, this approach requires specifying which variations to model and how to generate them. As environments and sensors vary, anticipating and designing training examples for each potential deployment shift becomes increasingly difficult and costly. *Can we improve distribution-shift generalization while reducing the need to explicitly model individual environmental changes?* Adversarial training (AT) offers one possible approach. Rather than defining a fixed set of augmentations, AT generates model-adaptive perturbations that expose weaknesses of the current policy during optimization. In computer vision, adversarially generated examples have sometimes improved robustness beyond the perturbation used for training, including generalization to unseen

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

domains and common corruptions (Volpi et al., 2018; Xie et al., 2020; Rusak et al., 2020; Kireev et al., 2022). These benefits are not universal, however, as AT can alter feature reliance and amplify alternative shortcuts (Moayeri et al., 2022). This motivates our central question:

Does adversarial training improve distribution-shift generalization in VLA models?

We investigate this question in the multi-view direct VLM-to-VLA setting, where pretrained VLMs are adapted directly to robot policies without intermediate large-scale robot-action pretraining (StarVLA Community, 2026; Ye et al., 2026; Zhang et al., 2026; Wang et al., 2026). This *direct VLM-to-VLA* setting is attractive both practically and experimentally: it enables flexible adoption of new VLM backbones without repeating large-scale robot-action pretraining, while removing such pretraining as a confounding factor when studying downstream generalization. Following common multi-view VLA designs (Kim et al., 2025a; Physical Intelligence, 2025; StarVLA Community, 2026), our policy receives a global third-person view and a close-range wrist view, and we evaluate generalization across seven LIBERO-Plus shift axes without generating the corresponding evaluation conditions during training.

Surprisingly, applying AT directly from the pretrained VLM substantially improves Camera Viewpoint and Sensor Noise, both of which affect only the third-person stream, while producing mixed effects across the remaining shifts. Controlled view interventions reveal a *view-collapse failure mode* under Direct AT: cross-view reliance can shift so strongly that the policy becomes dominated by the wrist view, showing little sensitivity to degradation of the third-person view. This exposes a *robustness shortcut*: gains under shifts affecting one camera can arise from reduced use of that camera rather than more robust use of it. This observation motivates a distinction between two complementary capabilities: (1) **robust perception**, extracting task-relevant information despite within-view distribution shifts, and (2) **robust fusion**, adjusting reliance across views according to their reliability. AT may support robust perception by perturbing images, but it does not require the policy to retain useful information from both views. It therefore permits solutions that rely predominantly on one visual source when that is sufficient for minimizing the action loss.

To mitigate this failure mode, we first initialize from an SFT policy and use View Swap, a simple intervention that randomly replaces one camera view with the corresponding view from another training sample, making that view unreliable while preserving the original action target. View Swap substantially improves Camera Viewpoint and Sensor Noise, the two shifts affecting only the third-person view, while preserving original LIBERO performance. On top of View Swap, which reduces fixed view reliance, we re-evaluate the incremental effect of AT. Notably, the benefits of AT extend beyond image corruptions: performance under Robot Initial State improves consistently across settings, while Sensor Noise and Camera Viewpoint further improve in most cases. However, gains are smaller for Object Layout and can be negative for lighting, backgrounds, and language, showing that AT provides complementary, yet strongly shift-dependent, generalization benefits.

Our contributions are threefold:

- We identify *view collapse* as a failure mode of AT in VLM-to-VLA adaptation: gains under some shifts can arise from persistent single-view reliance rather than more robust perception.
- We distinguish *robust perception* from *robust fusion* as complementary aspects of multi-view robustness, and use View Swap as a simple intervention to reduce fixed view reliance.
- We characterize the selective transfer of AT after mitigating fixed view reliance, finding consistent gains on Camera Viewpoint, Sensor Noise, and Robot Initial State across direct VLM-to-VLA configurations, but mixed effects on other shifts.

2 Related Work

VLA generalization under distribution shifts. Recent benchmarks expose substantial VLA performance degradation beyond standard evaluation conditions (Pumacay et al., 2024; Zhou et al., 2025), with LIBERO-Plus providing a fine-grained evaluation across seven visual, linguistic, and robot-state shift axes (Fei et al., 2026b). Existing approaches improve such generalization through several routes. LIBERO-Plus uses shift-specific training data (Fei et al., 2026b), while RobustVLA (Guo et al., 2026) combines predefined augmentations with adversarial training. Model-centric methods improve visual

processing or preserve pretrained representations through recurrent attention, backbone restoration, or architectural design (Xiao et al., 2026; Dey et al., 2025; Fu et al., 2026), while other work expands policy experience through online reinforcement learning or heterogeneous co-training (Fei et al., 2026a; Ye et al., 2026). In contrast, we investigate AT as a model-adaptive training strategy that requires neither shift-specific training data nor modifications to the underlying VLA architecture.

Adversarial training for generalization. Adversarial training (AT) (Goodfellow et al., 2015; Madry et al., 2018) optimizes models against worst-case, norm-bounded input perturbations. Beyond adversarial defense, model-adaptive perturbations have been used to expand source domains, improve image recognition, and generalize to unseen corruptions (Volpi et al., 2018; Xie et al., 2020; Rusak et al., 2020). Appropriately tuned ℓ_p-bounded AT can also improve accuracy and calibration under common image corruptions (Kireev et al., 2022). However, adversarial robustness can substantially alter the features learned and used by a model (Tsipras et al., 2019; Ilyas et al., 2019), and such changes do not necessarily improve all forms of natural robustness (Moayeri et al., 2022). Recent VLA work has incorporated AT into robustness-oriented training (Guo et al., 2026), while other studies focus on adversarial attacks and defenses (Wang et al., 2025; Xu et al., 2025). In contrast, our focus is diagnostic: we study how AT affects natural distribution-shift generalization and reshapes multi-view reliance in VLAs.

Multi-view policy learning. Multiple camera views provide complementary information that reduces occlusion and perceptual ambiguity in manipulation (Akinola et al., 2020). Wrist cameras offer stable hand-centric observations but have limited field of view, motivating their combination with global views (Hsu et al., 2022). Prior work addresses sensor failure by randomly masking views or sensor modalities during training (Neverova et al., 2015; Akinola et al., 2020; Skand et al., 2025). Related work in multimodal learning has also shown that models can favor one modality over others (Wu et al., 2022; Du et al., 2023). However, how robustness-oriented training reshapes view reliance in closed-loop multi-view policies remains less understood.

3 Study Setup

3.1 Multi-View VLA Training

Let f_θ denote a multi-view VLA policy. At each timestep, it receives a third-person observation x^g, a wrist-camera observation x^w, and a language instruction ℓ, and predicts an action or action chunk

$$\hat{a} = f_\theta(x^g, x^w, \ell). \tag{1}$$

The two visual streams provide complementary observations of the same manipulation episode: the third-person view captures global scene state, whereas the wrist view provides a close-range observation of the end effector and nearby objects.

Let $\mathcal{D}$ denote a demonstration dataset and a the target action or action chunk. Supervised fine-tuning (SFT) minimizes

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}_{(x^g, x^w, \ell, a)\sim\mathcal{D}} \left[\ell_{\mathrm{act}}\big(f_\theta(x^g, x^w, \ell), a\big) \right], \tag{2}$$

where ℓ_{act} denotes the training objective associated with the policy's action decoder. This notation covers both direct action-regression objectives (Kim et al., 2025a) and generative objectives such as flow matching (Black et al., 2025; Physical Intelligence, 2025); our analysis does not depend on a particular action-decoding architecture.

3.2 Adversarial Training

Adversarial training (AT) replaces the visual inputs with model-adaptive, norm-bounded perturbations. For both views, the inner maximization is

$$(\delta^{g\star}, \delta^{w\star}) \in \arg\max_{\|\delta^g\|_\infty \le \epsilon, \|\delta^w\|_\infty \le \epsilon} \ell_{\mathrm{act}}\big(f_\theta(x^g + \delta^g, x^w + \delta^w, \ell), a\big), \tag{3}$$

yielding the training objective

$$\mathcal{L}_{\mathrm{AT}}(\theta) = \mathbb{E}_{\mathcal{D}} \left[\ell_{\mathrm{act}}\big(f_\theta(x^g + \delta^{g\star}, x^w + \delta^{w\star}, \ell), a\big) \right]. \tag{4}$$

Our central question is whether optimizing Eq. 4 on the training distribution improves policy success under natural distribution shifts that are neither observed during training nor used to construct the adversarial perturbations.

Table 1: **Adversarial Training (AT) preserves original-task performance and improves LIBERO-Plus performance, but its effects are highly shift-dependent.** It strongly improves Camera (camera viewpoint) and Noise (sensor noise), the two shifts affecting only the third-person view, while degrading several others. Values are four-suite success rates (%).

	Original	LIBERO-Plus							
	Total	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
SFT	92.7	39.8	55.7	64.4	84.8	82.2	53.3	71.7	62.7
Direct AT	92.9	88.4	36.9	69.2	68.2	66.7	86.8	68.6	69.6
	↑ 0.2	↑ 48.6	↓ 18.8	↑ 4.8	↓ 16.6	↓ 15.5	↑ 33.5	↓ 3.1	↑ 6.9

3.3 Experimental Instantiation

Model and adaptation. We instantiate this formulation through *direct VLM-to-VLA adaptation* using the public StarVLA implementation (StarVLA Community, 2026), following the simple architecture studied in StarVLA-α (Ye et al., 2026). Specifically, we directly adapt Qwen3.5-0.8B (Qwen Team, 2026) to robot control without intermediate action-specific robot pretraining. We adapt the VLM with LoRA (rank 32) and attach an OFT-style lightweight MLP head that regresses eight-step continuous action chunks (Kim et al., 2025a; Ye et al., 2026). A single policy is trained jointly across all four LIBERO suites: Spatial, Object, Goal, and Long. Further training details are provided in Appendix A.

Adversarial optimization. Multi-step PGD-AT (Madry et al., 2018) substantially increases the cost of fine-tuning large VLA policies. We therefore use Fast-AT (Wong et al., 2020), which approximates the inner maximization with one gradient step from a random initialization within the perturbation set. Unless otherwise specified, we use an ℓ_∞ radius of $\epsilon = 1/255$, step size ϵ, and a uniform random start within the ϵ-ball. Perturbations are optimized jointly for both visual streams.

Evaluation. We evaluate rollout success on original LIBERO (Liu et al., 2023) and on the seven unseen shift axes of LIBERO-Plus (Fei et al., 2026b): Background, Robot Initial State, Camera Viewpoint, Language, Sensor Noise, Object Layout, and Lighting. We report each shift axis and the overall LIBERO-Plus success rate (Total). Further details are provided in Appendix A.

4 Adversarial Training Can Lead to View Collapse

This section examines whether pixel-space AT improves generalization to distribution shifts that are not explicitly modeled during training, and analyzes how AT affects multi-view information use. For this analysis, both SFT and Direct AT are trained from the pretrained VLM initialization.

4.1 Selected Shift Gains Do Not Imply Uniform Generalization

Table 1 summarizes the resulting performance profile. Direct AT substantially improves Camera Viewpoint (+48.6 pp) and Sensor Noise (+33.5 pp), while degrading Robot Initial State (-18.8 pp), Lighting (-16.6 pp), and Background (-15.5 pp). Despite these trade-offs, the overall LIBERO-Plus success improves from 62.7% to 69.6%, showing that aggregate gains can coexist with highly uneven shift-specific behavior. Additional threat-model and budget ablations are provided in Table 12 in the Appendix, showing that similar shift-dependent behavior appears across multiple perturbation norms and budgets, although the magnitude varies with the AT configuration.

The two largest gains share an important property: Camera Viewpoint and Sensor Noise perturb only the third-person stream, while leaving the wrist stream unchanged. This creates a key ambiguity: *did AT make perception from the third-person view more robust, or did the policy learn to solve these shifts by largely ignoring that view?*

4.2 The Selected Gains Can Arise From View Collapse

To distinguish (i) more robust third-person perception from (ii) reduced use of the third-person view, we measure the trained policy's dependence on each visual stream. We first black out one view

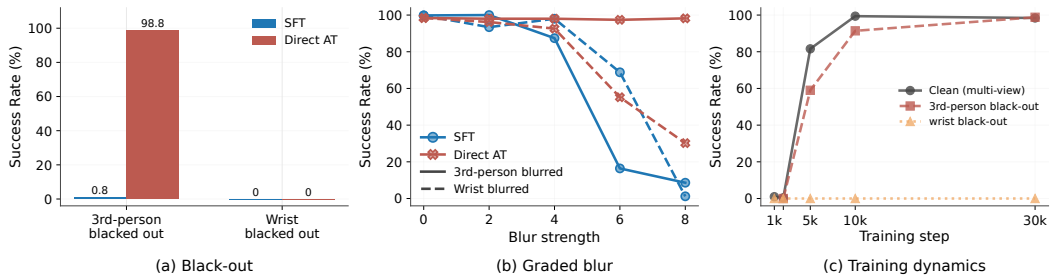


Figure 1: **Controlled interventions reveal wrist-dominant view collapse.** (a) Direct AT remains functional with only the wrist view but fails with only the third-person view. (b) Graded blur confirms the same asymmetric reliance under progressive degradation. (c) Intermediate checkpoints show that wrist-dominant reliance can emerge together with task competence. All panels use LIBERO-Object.

while leaving the other unchanged, and additionally apply graded blur to avoid relying on an extreme out-of-distribution intervention. We evaluate interventions in Fig. 1 on 500 LIBERO-Object episodes.

Direct AT can exhibit wrist-dominant view collapse. The interventions reveal a striking pattern: the Direct AT policy largely ignores the third-person view and relies on the wrist stream. As shown in Fig. 1(a), it retains 98.8% success when only the wrist view remains, but achieves 0.0% when only the third-person view remains. In contrast, SFT fails when either view is removed, indicating reliance on both streams. Graded blur confirms the same asymmetry (Fig. 1(b)): corrupting the third-person stream has almost no effect, whereas corrupting the wrist stream sharply degrades performance. We term this behavior *view collapse*: extreme single-view reliance in which one view dominates the policy. This resembles previously observed modality-imbalance phenomena in multimodal learning (Wu et al., 2022; Du et al., 2023). The gains under Camera Viewpoint and Sensor Noise, which perturb only the third-person stream, can therefore arise from reduced reliance on the affected view rather than more robust use of it. View collapse arises in three of the four AT configurations we test (ℓ_∞, $\epsilon \in \{1, 2\}/255$; ℓ_2, $\epsilon = 2$), but not for ℓ_2, $\epsilon = 1$. Additional 3k-step Direct AT experiments show wrist-biased reliance across model scales, action decoders, and a robot-pretrained policy, although the severity varies (Appendix D).

View collapse can emerge with task competence. Intermediate Direct AT checkpoints show that, in this run, the policy is already wrist-dominant at the earliest checkpoint with non-trivial task success (Figure 1(c)): performance with the third-person view removed closely tracks clean performance, while success remains near zero when the wrist view is removed. This pattern persists as task performance improves, showing that view collapse can emerge together with task competence.

Matched single-view SFT confirms that both views provide useful task information: neither single-view policy matches multi-view performance (Table 2). These results expose a *robustness shortcut*: by relying largely on the wrist stream, Direct AT can remain robust to third-person-only shifts without learning to robustly use that stream.

Table 2: **Matched single-view SFT.**

Views	Success (%)
Multi-view	**92.7**
wrist only	80.6
third-person only	72.2

What drives view collapse under AT? To first order, the ℓ_p-robust objective with independent per-view budget ϵ adds a sensitivity penalty $\epsilon \sum_v \|\nabla_{x^v} \ell_{\text{act}}\|_q$ ($q = 1$ for $p = \infty$ and $q = 2$ for $p = 2$). However, reducing this penalty does not necessarily require robustifying every view. If one view contributes disproportionately to the adversarial objective, AT may instead suppress reliance on that view. We test this hypothesis with *3rd-person-only AT* and *wrist-only AT*, perturbing only the named view while retaining both inputs. Interestingly, 3rd-person-only AT effectively discards the third-person view, while wrist-only AT effectively discards the wrist view. Under the black-out evaluation, 3rd-person-only AT relies on the wrist view (88.0% vs. 0.0%), whereas wrist-only AT relies on the third-person view (50.0% vs. 0.0%), already clear at the 3k-step checkpoint. Thus, view collapse is not specific to the third-person camera; AT may suppress reliance on a view when doing so is easier than learning to use that view robustly. We also find asymmetric sensitivity at initialization: attacking the third-person view increases the loss more than attacking the wrist view, which may bias multi-view AT toward suppressing the third-person view (Appendix C).

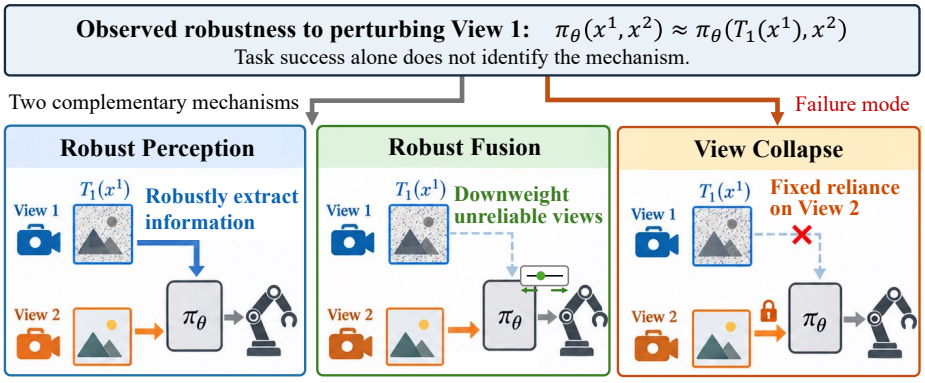


Figure 2: **Observed robustness to perturbing one view is mechanistically ambiguous.** Such robustness may arise from *robust perception*, *robust fusion*, or the failure mode of *view collapse*. Here, fixed reliance means reliance that does not adapt to relative view reliability.

4.3 Robust Multi-View Generalization Involves Both Perception and Fusion

The preceding experiments expose two conceptually distinct but behaviorally intertwined aspects of multi-view robustness. A policy must continue extracting useful information when observations within a view change, while also adapting how it combines multiple views as their relative reliability changes. We refer to these aspects as *robust perception* and *robust fusion*, respectively.

(1) Robust perception. The ability to extract reliable, task-relevant information from a view under within-view distribution shifts, while that view remains task-aligned and potentially informative.

(2) Robust fusion. The ability to adaptively integrate information across views as their reliability changes. When one view becomes unreliable, the policy should rely more on the remaining view without developing permanent dependence on either source.

View collapse illustrates that robustness to a perturbed view does not necessarily imply robust perception of that view. When such collapse occurs, apparent robustness can result from fixed view reliance rather than improved use of the affected stream.

Pixel-space AT may contribute to robust perception by introducing perturbations within each view, but its action-level objective does not explicitly constrain cross-view utilization. It therefore permits solutions that concentrate reliance on a single stream. This motivates considering robust perception and robust fusion as two complementary aspects of multi-view robustness.

5 Mitigating View Collapse in Adversarial Training

Sec. 4 shows that Direct AT can improve robustness to selected shifts through strong single-view reliance rather than robust multi-view utilization. How does AT behave after reducing the fixed single-view reliance? We investigate two simple strategies: SFT initialization and View Swap.

5.1 Mitigating View Collapse: SFT Initialization and View Swap

First, we consider *SFT initialization*. Figure 1(c) suggests that wrist-dominant reliance emerges together with task competence under Direct AT. Motivated by this observation, we first establish multi-view task behavior with standard SFT and then continue training with AT (SFT→AT), aiming to reduce extreme view collapse while retaining the potential robustness benefits of AT.

Second, we consider *View Swap* as a direct intervention against fixed view reliance. For a training sample i, we randomly select one view $s \in \{g, w\}$ and replace it with the corresponding view from another sample in the minibatch:

$$\tilde{x}_i^s := x_{\sigma(i)}^s, \qquad \tilde{x}_i^{\bar{s}} := x_i^{\bar{s}}, \tag{5}$$

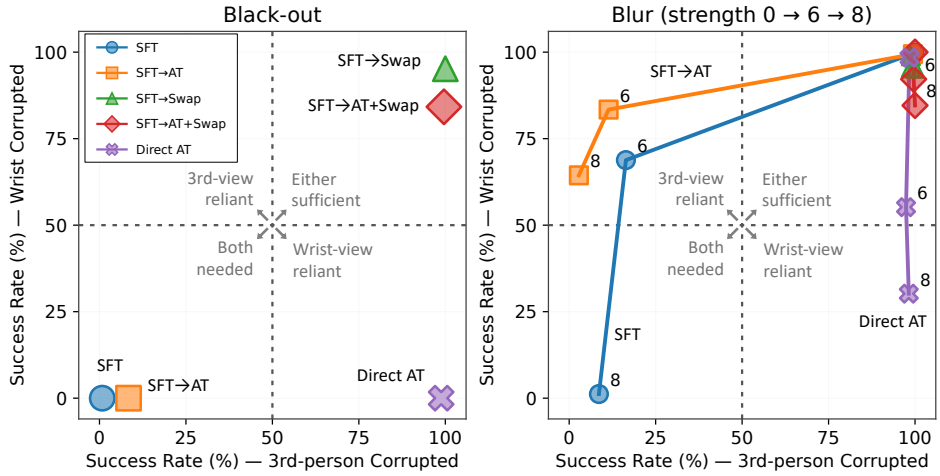


Figure 3: **Controlled view interventions reveal patterns of view reliance.** Left: black-out removes either the third-person or wrist view. Right: graded blur progressively degrades either view (strengths 0, 6, and 8). The horizontal and vertical axes report success under third-person- and wrist-view corruption, respectively, with the upper-right indicating flexible fallback on either view.

where σ denotes a random minibatch permutation and $\bar{s}$ denotes the unchanged view. The language instruction and action target are kept unchanged. The swapped view therefore remains visually plausible but becomes unreliable for the current task. Because either view can be swapped, the policy cannot consistently treat a particular visual source as reliable.

From the shared SFT initialization, we define four matched training variants: continued SFT, adversarial training (**SFT→AT**), View Swap (**SFT→Swap**), and their combination (**SFT→AT+Swap**). All variants start after 3,750 SFT steps and continue for 10k steps. View Swap is applied on 25% of steps; SFT→AT+Swap uses AT on 75% and View Swap on 25%. Qualitative examples of the adversarial perturbations and View Swap intervention are provided in Appendix G.

5.2 View Interventions Reveal How Training Reshapes View Reliance

We first examine how the training variants reshape reliance on each visual source using the same controlled view interventions as in Sec. 4. Figure 3 summarizes the results as a two-dimensional view-reliance map, with Direct AT included for comparison.

SFT-initialized AT does not exhibit the extreme collapse of Direct AT. Direct AT retains 98.8% success when the third-person view is blacked out but fails completely when the wrist view is removed (0.0%). In contrast, SFT→AT achieves 8.4% and 0.0%, respectively. Thus, the SFT-initialized policy does not exhibit the extreme wrist-dominant collapse observed under Direct AT, although its view reliance remains asymmetric.

When introduced after SFT, AT can amplify existing view-reliance asymmetry. Under graded blur, SFT is more sensitive to third-person-view degradation than to wrist-view degradation, and SFT→AT further accentuates this asymmetry. This suggests that AT does not induce a universal preference for a particular camera, but can reinforce view reliance already present in the policy.

View Swap enables flexible fallback across views. The behavior changes substantially when View Swap is introduced. Under black-out, SFT→Swap and SFT→AT+Swap remain successful when either view is removed, moving from the "both needed" or single-view-reliant regimes toward the "either sufficient" regime. The same pattern persists under graded blur: both variants remain robust as either visual source is progressively degraded. These results indicate that View Swap mitigates fixed view reliance and enables flexible fallback across views.

Table 3: **Comparison of robust adaptation strategies for Qwen3.5-0.8B-OFT under matched training steps.** Four-suite success rates (%); green/red mark gains/losses over the SFT baseline.

	Original	LIBERO-Plus							
Training	**Total**	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
SFT	95.6	38.8	55.7	71.0	**88.9**	88.5	59.1	69.8	65.3
SFT→AT	**97.8**	35.8	55.9	**71.8**	86.2	86.8	56.8	75.7	65.0
	↑ 2.2	↓ 3.0	↑ 0.2	↑ 0.8	↓ 2.7	↓ 1.7	↓ 2.3	↑ 5.8	↓ 0.3
SFT→Swap	96.3	77.9	57.9	70.5	87.5	**88.8**	85.7	77.0	77.0
	↑ 0.7	↑ 39.0	↑ 2.3	↓ 0.5	↓ 1.4	↑ 0.3	↑ 26.6	↑ 7.2	↑ 11.7
SFT→AT+Swap	96.5	**84.1**	**65.2**	68.4	88.8	87.7	**88.3**	**79.5**	**79.7**
	↑ 0.9	↑ 45.2	↑ 9.5	↓ 2.5	↓ 0.1	↓ 0.7	↑ 29.2	↑ 9.6	↑ 14.3

5.3 Distribution-Shift Effects of View Swap and Adversarial Training

We next evaluate how the observed changes in view reliance affect natural distribution-shift generalization on LIBERO-Plus. Table 3 compares the four variants across seven shift categories and the original LIBERO tasks.

SFT initialization alone is insufficient for AT to improve overall generalization. The SFT→AT policy does not exhibit the extreme view collapse observed under Direct AT, but does not improve aggregate LIBERO-Plus performance. Its effects remain shift-dependent, with a gain on Layout (+5.8 pp) but degradations on Camera (-3.0 pp), Light (-2.7 pp), and Noise (-2.3 pp). Thus, the absence of extreme collapse alone is insufficient for AT to yield broader distribution-shift gains.

View Swap improves robustness to third-person-view perturbations. SFT→Swap substantially improves Camera (+39.0 pp) and Noise (+26.6 pp), the two shifts affecting only the third-person view, while preserving original LIBERO performance. Together with the view-intervention results in Sec. 5.2, these gains are consistent with more flexible reliance across views. With AT, a matched View Dropout baseline is less effective (Appendix E); one possible factor is that View Swap preserves realistic per-view image statistics, whereas View Dropout introduces an extreme missing-view pattern.

AT provides complementary but shift-dependent gains with View Swap. Relative to SFT→Swap, SFT→AT+Swap improves Camera (+6.2 pp), Robot Initial State (+7.3), Noise (+2.6), and Layout (+2.5), preserving original LIBERO performance but yielding small or negative effects elsewhere. Camera, Robot Initial State, and Noise improve across all four direct VLM-to-VLA configurations (Tables 3 and 4). Noise gains and Lighting degradations may partly reflect AT's frequency-dependent trade-offs: improved robustness to several blur corruptions but reduced robustness to fog and contrast (Yin et al., 2019). Camera and Robot Initial State gains are less directly explained by pixel-level robustness, which need not transfer to spatial transformations (Engstrom et al., 2019). One possibility is that AT learns more robust task-relevant features, such as hand–object spatial cues, that support control under these shifts.

A lightweight direct VLM-to-VLA model achieves competitive overall robustness. Despite these shift-dependent effects, SFT→AT+Swap reaches 79.7% overall success on LIBERO-Plus, competitive with existing LIBERO-finetuned VLAs. Notably, this performance comes from directly adapting a 0.8B VLM without large-scale robot pretraining; full comparisons appear in Appendix F.

5.4 Generality of View Swap and Selective AT Transfer Across Models

We next test whether View Swap and the selective transfer of AT generalize across action decoders, model scales, VLM backbones, and pretraining regimes. Beyond the primary Qwen3.5-0.8B-OFT setting, we evaluate Qwen3.5-0.8B with a PI-style decoder, Qwen3.5-2B and PaliGemma (Beyer et al., 2024) with the same OFT-style decoder, and the robot-pretrained π_0 policy (Table 4).

View Swap consistently improves Camera Viewpoint and Sensor Noise across the direct VLM-to-VLA configurations. These improvements hold across Qwen model scales and action decoders as well as the PaliGemma backbone, consistent with the link observed in the primary setting between reducing fixed view reliance and robustness to view-specific shifts.

Table 4: **Comparison of adaptation strategies across model configurations.** Values are four-suite success rates (%); green/red mark gains/losses over matched SFT baselines.

	Original	LIBERO-Plus							
Training	**Total**	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
Direct VLM-to-VLA adaptation									
Qwen3.5-0.8B-PI									
SFT	95.5	27.7	51.5	**70.5**	80.5	61.2	54.3	71.3	58.4
SFT→Swap	97.2	43.2	56.9	55.4	**85.6**	47.7	77.5	73.4	62.5
	↑ 1.7	↑ 15.4	↑ 5.4	↓ 15.1	↑ 5.1	↓ 13.6	↑ 23.1	↑ 2.2	↑ 4.1
SFT→AT+Swap	**98.1**	**68.8**	**64.8**	66.7	71.8	**65.9**	**82.7**	**78.0**	**71.5**
	↑ 2.6	↑ 41.1	↑ 13.3	↓ 3.8	↓ 8.7	↑ 4.6	↑ 28.4	↑ 6.7	↑ 13.1
Qwen3.5-2B-OFT									
SFT	**96.6**	42.0	54.3	**84.0**	**90.9**	**91.0**	70.3	81.0	71.6
SFT→Swap	95.8	81.2	55.1	73.9	86.2	84.5	85.3	**81.6**	77.7
	↓ 0.8	↑ 39.2	↑ 0.8	↓ 10.1	↓ 4.7	↓ 6.5	↑ 15.0	↑ 0.6	↑ 6.1
SFT→AT+Swap	95.8	**85.0**	**57.6**	68.8	83.9	85.2	**90.1**	79.2	**78.1**
	↓ 0.8	↑ 43.0	↑ 3.3	↓ 15.2	↓ 7.0	↓ 5.8	↑ 19.8	↓ 1.8	↑ 6.5
PaliGemma-OFT									
SFT	80.9	12.3	28.7	**36.1**	**61.6**	30.8	17.4	48.4	32.4
SFT→Swap	84.2	31.3	33.4	34.3	57.7	43.3	30.7	55.2	39.9
	↑ 3.3	↑ 19.0	↑ 4.7	↓ 1.8	↓ 3.9	↑ 12.5	↑ 13.3	↑ 6.8	↑ 7.5
SFT→AT+Swap	**93.6**	**74.7**	**39.1**	36.0	27.8	**56.8**	**68.6**	**65.2**	**53.6**
	↑ 12.7	↑ 62.4	↑ 10.4	↓ 0.1	↓ 33.8	↑ 26.0	↑ 51.2	↑ 16.8	↑ 21.2
Robot-pretrained VLA									
π_0									
SFT	**93.9**	9.1	5.0	**62.3**	**88.4**	**85.2**	12.9	**73.8**	**44.3**
SFT→Swap	92.3	**16.3**	5.1	51.8	83.5	77.0	**14.4**	66.4	41.5
	↓ 1.6	↑ 7.2	↑ 0.1	↓ 10.5	↓ 4.9	↓ 8.2	↑ 1.5	↓ 7.4	↓ 2.8
SFT→AT+Swap	93.2	13.9	**7.6**	50.1	85.3	79.7	14.0	68.3	41.9
	↓ 0.7	↑ 4.8	↑ 2.6	↓ 12.2	↓ 3.1	↓ 5.5	↑ 1.1	↓ 5.5	↓ 2.4

AT shows consistent but selective transfer across direct VLM-to-VLA configurations. On top of View Swap, AT further improves Camera Viewpoint, Sensor Noise, and Robot Initial State in all four configurations, while effects on the remaining shifts are mixed. PaliGemma shows particularly large Camera and Noise gains but a substantial Lighting degradation.

Pretrained π_0 exhibits a different pattern. At baseline, π_0 is already substantially weaker on Camera Viewpoint and Sensor Noise. View Swap partially improves these shifts, but the gains are modest, and adding AT provides no further improvement. This suggests that substantially lower starting robustness may limit recovery from View Swap and the additional benefit of AT.

6 Discussion and Limitations

Training trajectory and view reliance. Our results suggest that *when* AT is introduced can affect view reliance, with extreme collapse arising under some configurations. Fully explaining the optimization dynamics underlying these view-reliance patterns remains future work.

Robust fusion and source utilization. View Swap mitigates persistent single-view reliance. However, robust fusion more broadly requires adapting reliance to the relative reliability of both views, which remains to be characterized more directly.

Scope of evaluation. Our controlled view-reliance analysis focuses on direct VLM-to-VLA adaptation in LIBERO/LIBERO-Plus. The different behavior of robot-pretrained π_0 suggests that pretraining may affect downstream AT benefits; broader VLAs and real-world settings remain future work.

7 Conclusion

We studied AT for natural distribution-shift generalization in multi-view VLAs. Direct AT can reshape cross-view reliance and, in extreme cases, produce view collapse, so gains on view-specific shifts need not reflect robust use of the perturbed view. This motivates distinguishing robust perception from robust fusion. With View Swap reducing fixed view reliance, AT provides additional but strongly shift-dependent gains. Multi-view robustness should therefore consider not only task success, but also how policies use and combine views.

Reproducibility Statement

We provide the experimental setup and major training configurations in the main paper and Appendix, including model adaptation details, adversarial-training configurations, training schedules, evaluation protocols, and hardware/GPU settings. Additional implementation details and ablations are also reported in the Appendix. We will publicly release the training and evaluation code, together with the corresponding configurations, to facilitate reproduction of our results.

Ethics Statement

This work studies robustness and failure modes of Vision-Language-Action policies in simulated robotic environments and does not involve human participants, personal data, or sensitive information. Although we use adversarial perturbations, our goal is to diagnose robustness failures and improve the reliability of embodied AI systems rather than to enable harmful attacks. Our results are obtained in simulation and should not be interpreted as guaranteeing safety in real-world robotic deployment.

AI Use Statement

Generative AI tools were used to assist with language editing, code development and debugging, and figure-layout ideation. All generated content was reviewed and revised by the authors. AI tools were not used to generate experimental results or determine the scientific conclusions of the paper, and all reported results, claims, and citations were verified by the authors.

References

Iretiayo Akinola, Jacob Varley, and Dmitry Kalashnikov. Learning precise 3D manipulation from multiple uncalibrated cameras. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020. doi: 10.1109/ICRA40945.2020.9197181. URL https://ieeexplore.ieee.org/document/9197181.

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, June 2025. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

Sombit Dey, Jan-Nico Zaech, Nikolay Nikolov, Luc Van Gool, and Danda Pani Paudel. ReVLA: Reverting visual domain limitation of robotic foundation models. In *IEEE International Conference on Robotics and Automation*, 2025. URL https://arxiv.org/abs/2409.15250.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pp. 1802–1811. PMLR, 2019.

Senyu Fei, Siyin Wang, Li Ji, Ao Li, Shiduo Zhang, Liming Liu, Jinlong Hou, Jingjing Gong, Xianzhong Zhao, and Xipeng Qiu. SRPO: Self-referential policy optimization for vision-language-action models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6718–6728, June 2026a.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fei_LIBERO-Plus_A_Progressive_Robustness_Benchmark_for_Visual-Language-Action_Models_CVPR_2026_paper.html.

Yuxia Fu, Zhizhen Zhang, Yuqi Zhang, Zijian Wang, Zi Huang, and Yadan Luo. MergeVLA: Cross-skill model merging toward a generalist vision-language-action agent. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, et al. On robustness of vision-language-action model against multi-modal perturbations. In *International Conference on Learning Representations*, volume 2026, pp. 70248–70272, 2026.

Kyle Hsu, Moo Jin Kim, Rafael Rafailov, Jiajun Wu, and Chelsea Finn. Vision-based manipulators need to also see from their hands. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RJkAHKp7kNZ.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1012–1021. PMLR, 2022. URL https://proceedings.mlr.press/v180/kireev22a.html.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1939. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit tradeoffs between adversarial and natural distributional robustness. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2809. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fd62b65606f0f0d2af2c01623a224258-Abstract.html.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2015.

Physical Intelligence. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *Proceedings of The 9th Conference on Robot Learning*, 2025.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation. *Proceedings of Robotics: Science and Systems*, July 2024. doi: 10.15607/RSS.2024.XX.133.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pp. 53–69. Springer, 2020. doi: 10.1007/978-3-030-58580-8_4. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123480052.pdf.

Skand Skand, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. In *Proceedings of the Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4463–4482. PMLR, 2025. URL https://proceedings.mlr.press/v270/skand25a.html.

Skywalker-yqz. Reproducibility: Official openpi π_0 gets 16–18% on sensor noise, versus 79.0% on the leaderboard. GitHub Issue #61, LIBERO-plus, 2026. URL https://github.com/sylvestf/LIBERO-plus/issues/61.

StarVLA Community. StarVLA: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026. doi: 10.48550/arXiv.2604.05014. URL https://arxiv.org/abs/2604.05014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5334–5344. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1d94108e907bb8311d8802b48fd54b4a-Abstract.html.

Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6948–6958, 2025. doi: 10.1109/ICCV51701.2025.00653. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Exploring_the_Adversarial_Vulnerabilities_of_Vision-Language-Action_Models_in_Robotics_ICCV_2025_paper.html.

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 18638–18646, 2026.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJx040EFvH.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Lei Xiao, Jifeng Li, Juntao Gao, Feiyang Ye, Yan Jin, Jingjing Qian, Jing Zhang, Yong Wu, and Xiaoyuan Yu. AVA-VLA: Improving vision-language-action models with active visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2511.18960.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00090. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Xie_Adversarial_Examples_Improve_Image_Recognition_CVPR_2020_paper.html.

Haochuan Xu, Yun Sing Koh, Shuhuai Huang, Zirun Zhou, Di Wang, Jun Sakuma, and Jingfeng Zhang. Model-agnostic adversarial attack and defense for vision-language-action models. *arXiv preprint arXiv:2510.13237*, 2025. doi: 10.48550/arXiv.2510.13237. URL https://arxiv.org/abs/2510.13237.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. StarVLA-α: Reducing complexity in vision-language-action systems. *European Conference on Computer Vision (ECCV)*, 2026.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026.

Jianke Zhang, Xiaoyu Chen, Yanjiang Guo, Yucheng Hu, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tc2UsBeODW.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025. URL https://arxiv.org/abs/2510.03827.

A Implementation Details

A.1 Training Hyperparameters

Tables 5 and 6 summarize the implementation details for our primary Qwen3.5-0.8B-OFT experiments.

Table 5: **Model and optimization configuration for the primary setting.**

Component	Configuration
VLM backbone	Qwen3.5-0.8B
Action head	OFT-style lightweight MLP head
Action chunk	8 steps
Input resolution	224×224 raw frames; internally resized to 256×256
VLM adaptation	LoRA
LoRA rank / α / dropout	32 / 64 / 0.05
Optimizer	AdamW
VLM LoRA learning rate	1×10^{-4}
Action-head learning rate	2×10^{-5}
Effective batch size	128
LR schedule	Cosine decay
Warmup	300 steps
Precision	bfloat16 mixed precision

LoRA is applied within `qwen_vl_interface` to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`, `qkv`, `proj`, `linear_fc1`, and `linear_fc2`. The action head is optimized separately.

Table 6: **Training configurations for Qwen3.5-0.8B-OFT.** Sec. 4 compares SFT and Direct AT from the pretrained VLM initialization, whereas Sec. 5 compares matched continuations from a shared SFT checkpoint.

Experiment	Initialization	Objective	Steps	AT norm	ϵ	AT / Swap
Sec. 4: SFT	Pretrained VLM	SFT	30k	–	–	–
Sec. 4: Direct AT	Pretrained VLM	Fast-AT	30k	ℓ_∞	$1/255$	100% / 0%
Sec. 5: SFT	SFT (3,750 steps)	SFT	+10k	–	–	–
Sec. 5: SFT→AT	SFT (3,750 steps)	Fast-AT	+10k	ℓ_∞	$1/255$	100% / 0%
Sec. 5: SFT→Swap	SFT (3,750 steps)	View Swap	+10k	–	–	0% / 25%
Sec. 5: SFT→AT+Swap	SFT (3,750 steps)	Fast-AT + View Swap	+10k	ℓ_∞	$1/255$	75% / 25%

For all Fast-AT experiments, we use one gradient step from a uniform random initialization within the perturbation set, with step size equal to ϵ. Unless otherwise specified, perturbations are optimized jointly over both visual views, and the perturbation budget is linearly ramped from zero to its target value over the first 500 training steps.

The effective batch size of 128 is kept fixed across the two-GPU configuration (32 samples per GPU with two-step gradient accumulation) and the four-GPU configuration (32 samples per GPU without gradient accumulation).

Model-specific training schedules. For the matched comparisons in Sec. 5, the number of initial SFT optimizer updates varies across model configurations (Table 7). Within each configuration, all compared variants share the same initial checkpoint and undergo 10k further optimizer updates with the same learning-rate schedule. Thus, training budgets are matched within each configuration, while initial SFT budgets differ across configurations.

Table 7: **Training schedules for the matched adaptation comparisons.** All counts denote optimizer updates. Initial SFT refers to LIBERO training before the compared variants begin.

Model	Initial SFT updates	Further updates
Qwen3.5-0.8B-OFT	3,750	10,000
Qwen3.5-0.8B-PI	7,500	10,000
Qwen3.5-2B-OFT	15,000	10,000
PaliGemma-OFT	20,000	10,000
π_0	30,000	10,000

For π_0, we use the StarVLA-converted OpenPI LIBERO checkpoint; its official fine-tuning configuration specifies 30k updates. Differences in improvement magnitudes across configurations should therefore not be attributed solely to architecture or model scale, since their training histories also differ.

A.2 Evaluation Protocol

We evaluate a single policy jointly trained across the four LIBERO suites: Spatial, Object, Goal, and Long. Original LIBERO performance and LIBERO-Plus performance are reported separately. For LIBERO-Plus, we evaluate the seven shift categories used throughout the main paper: Camera Viewpoint, Robot Initial State, Language, Lighting, Background, Sensor Noise, and Object Layout. Each shift-specific score pools rollouts across the four suites, and Total is the success rate pooled over all evaluated rollouts across suites and shifts, rather than an unweighted average of the seven shift-specific scores.

For the controlled view-intervention experiments, we independently perturb the third-person and wrist views while keeping the other visual source unchanged. Black-out replaces the selected view entirely, whereas graded blur progressively degrades its visual information. All evaluation settings, random seeds, and rollout counts are fixed across the compared methods.

B Computational Cost

Experiments were conducted on NVIDIA A100, RTX 6000 Ada, and H100 GPUs. We report approximate compute in A100-hours (A100-h), using throughput-based conversion factors of $1.0\times$, $0.9\times$, and $2.0\times$, respectively. Costs are estimated from scheduler logs and exclude failed runs and preliminary tests.

Table 8: **Approximate computational cost for the primary Qwen3.5-0.8B-OFT experiments.** Training costs are reported per run and evaluation costs per checkpoint, in A100-hours.

Run / Evaluation	A100-h
Training	
SFT (30k steps)	88
Direct AT (30k steps)	181
SFT initialization (3,750 steps)	11
SFT→AT (+10k)	49
SFT→Swap (+10k)	28
SFT→AT+Swap (+10k)	47
Evaluation	
Original LIBERO (4 suites)	∼3.6
LIBERO-Plus (4 suites, 7 shifts)	∼45

*Estimated from the median of evaluations with complete timing records.

The SFT-initialized variants share the same 3,750-step SFT checkpoint; its cost is counted only once in the aggregate compute below.

Table 9: **Approximate aggregate computational cost of the experiments.**

Category	A100-h
Training	2,167
Original LIBERO evaluation	115
LIBERO-Plus evaluation	1,536
Additional evaluations	1,669
Total	**5,487**

In total, the experiments summarized above required approximately 5,487 A100-h (229 A100-days), with training accounting for approximately 39% of the total compute.

C Probing the Mechanism of View Collapse

Joint Direct AT eventually produces a policy that relies almost exclusively on the wrist view. We investigate how this direction is selected and how the imbalance develops during training. Specifically, we test whether the collapse reflects an intrinsic preference for the wrist view or depends on which view is attacked, and examine whether joint AT begins from an asymmetric state.

Attack-target intervention. We first intervene on the attacked view during from-scratch AT. We train two variants that differ only in attack placement: `3rd-only-AT` perturbs only the third-person image, whereas `wrist-only-AT` perturbs only the wrist image. Both use single-step ℓ_∞ PGD with $\epsilon = 2/255$ and the 10k-step schedule. At steps 1000, 2000, and 3000, we evaluate each checkpoint on LIBERO-Object under clean input, wrist black-out, and third-person black-out, using 50 trials per task across 10 tasks.

Table 10: **Attack placement reverses view reliance.** Success rates (%) on LIBERO-Object. Each policy remains effective when the attacked view is removed, but fails when the unattacked view is removed.

Training	Step	Clean	Wrist black-out	3rd black-out
`3rd-only-AT`	1000	83.0	0.0	78.4
	2000	87.4	0.0	87.6
	3000	93.4	0.0	88.0
`wrist-only-AT`	1000	21.4	23.2	0.0
	2000	52.2	51.6	0.0
	3000	49.2	50.0	0.0

Table 10 shows a mirrored reliance pattern. When only the third-person view is attacked, removing that view has little effect, whereas removing the unattacked wrist view reduces success to zero. Attacking only the wrist view produces the opposite pattern. The direction remains stable across all three checkpoints. Thus, attack placement causally affects which view is suppressed during from-scratch AT, ruling out an unconditional tendency to select the wrist view. The wrist-reliant variant nevertheless learns faster and reaches higher clean success, suggesting that wrist observations may still offer an optimization or task-information advantage; this advantage alone, however, does not determine the direction of collapse.

Per-view adversarial sensitivity. The intervention above explains how attack placement can control the selected view, but joint Direct AT attacks both views. To examine why it initially moves toward wrist reliance, we measure the loss increase induced by attacking each view separately:

$$\Delta\ell_v = \ell_\theta(x + \hat{\delta}_v) - \ell_\theta(x), \qquad v \in \{g, w\}, \tag{6}$$

where g and w denote the third-person and wrist views. The perturbation $\hat{\delta}_v$ is generated using the same implementation as in training, with one step, $\epsilon = 2/255$, step size ϵ, and a random start, but is restricted to view v. Because this is an approximate attack rather than exact inner maximization, individual loss changes can be slightly negative.

We sample mid-episode frames from the LIBERO training data, stratified across the four suites. Each checkpoint is evaluated on three fixed probe-set draws of 128 examples, giving 384 pooled evaluations. We report the mean difference $d = \Delta\ell_g - \Delta\ell_w$, its paired t-statistic, and the fraction of examples satisfying $\Delta\ell_g > \Delta\ell_w$. The initialization checkpoint consists of the pretrained VLM with newly initialized LoRA parameters and action head, before any LIBERO-specific updates.

Table 11: **Per-view adversarial sensitivity before and during joint Direct AT.** Positive $\bar{d}$ indicates a larger loss increase from perturbing the third-person view. The initial ordering disappears by step 1000 and clearly reverses by steps 2000–3000.

Checkpoint	$\overline{\Delta\ell_g}$	$\overline{\Delta\ell_w}$	$\bar{d}$	t	$\Delta\ell_g > \Delta\ell_w$
Initialization	0.01610	0.00750	$\mathbf{+0.00860}$	$\mathbf{+21.15}$	86%
Direct AT, step 1000	−0.00002	0.00003	−0.00005	−2.11	45%
Direct AT, step 2000	0.00022	0.00267	$\mathbf{-0.00245}$	$\mathbf{-9.87}$	15%
Direct AT, step 3000	0.00037	0.00334	$\mathbf{-0.00298}$	$\mathbf{-10.76}$	11%

An initial asymmetry aligns with the collapse direction. Before LIBERO-specific training, perturbing the third-person view induces a substantially larger loss increase than perturbing the wrist view: $\bar{d} = 0.0086$, with the same ordering on 86% of probe examples. This asymmetry is already present when joint Direct AT begins and aligns with the subsequent suppression of the third-person stream. Because the measurement includes the newly initialized action head and LoRA parameters, we do not attribute it solely to the pretrained VLM backbone.

The ordering changes rapidly during training. By step 1000, the mean difference is close to zero. By steps 2000–3000, perturbing the wrist view induces a clearly larger loss increase, and the magnitude of this reversed imbalance grows. A companion gradient analysis shows the same transition: the third-person share of the total visual input-gradient magnitude decreases from 0.435 at step 1000 to 0.146 at step 2000. Task success remains approximately zero at both checkpoints and begins to recover only around step 5000. The shift in view influence therefore emerges before successful task behavior and becomes more pronounced as training proceeds.

Selection followed by amplification. The combined results are consistent with a two-stage account of view collapse. First, the attack-target intervention shows that the attacked view can determine the initial direction of view selection. Under joint attacks, the initialization-level sensitivity imbalance provides a plausible bias toward suppressing the third-person view. Second, once a wrist-side bias emerges, subsequent training increasingly concentrates gradients and adversarial loss sensitivity on the wrist stream, consistent with amplification of the selected direction.

This interpretation does not treat a larger $\Delta\ell_v$ as evidence that view v is intrinsically harder to robustify. The quantity reflects the current policy's sensitivity and therefore depends on both the robustness of the extracted features and how strongly the policy uses that view. As wrist reliance develops, attacking the wrist stream has a larger effect, whereas attacking the increasingly suppressed third-person stream has little effect. Accordingly, the intervention establishes attack placement as a causal factor in view selection, while the initial sensitivity asymmetry and its evolution provide evidence consistent with, but do not fully identify, the subsequent amplification mechanism.

D Additional Analyses of Direct Adversarial Training

Table 12: **The shift-specific robustness pattern persists across adversarial threat models and perturbation budgets.** Across multiple configurations, Direct AT produces pronounced gains on Camera Viewpoint and Sensor Noise, together with mixed gains and degradations across the remaining shifts. Values are four-suite success rates (%).

	Original	LIBERO-Plus							
Training	**Total**	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
SFT	92.7	39.8	55.7	64.4	84.8	82.2	53.3	71.7	62.7
Direct AT (ℓ_∞, ϵ=2/255)	92.1	75.7	35.5	59.3	13.0	29.8	83.0	60.7	53.8
	↓ 0.6	↑ 35.9	↓ 20.2	↓ 5.1	↓ 71.8	↓ 52.4	↑ 29.7	↓ 11.0	↓ 8.9
Direct AT (ℓ_∞, ϵ=1/255)	92.9	88.4	36.9	69.2	68.2	66.7	86.8	68.6	69.6
	↑ 0.2	↑ 48.6	↓ 18.8	↑ 4.8	↓ 16.6	↓ 15.5	↑ 33.5	↓ 3.1	↑ 6.9
Direct AT (ℓ_2, ϵ=2)	89.9	90.0	33.9	61.6	57.5	58.2	91.3	60.3	65.6
	↓ 2.8	↑ 50.2	↓ 21.8	↓ 2.8	↓ 27.3	↓ 24.0	↑ 38.0	↓ 11.4	↑ 2.9
Direct AT (ℓ_2, ϵ=1)	96.9	31.0	39.8	64.7	74.3	77.1	67.0	75.9	60.0
	↑ 4.2	↓ 8.8	↓ 15.9	↑ 0.3	↓ 10.5	↓ 5.1	↑ 13.7	↑ 4.2	↓ 2.7

Threat-model and budget ablations. To examine whether the highly shift-dependent behavior of Direct AT is specific to the main ℓ_∞, ϵ=1/255 configuration, we additionally evaluate multiple perturbation norms and budgets. Table 12 reports the complete seven-axis LIBERO-Plus results.

The results show that this shift-specific behavior persists across multiple adversarial settings, but is not universal. Increasing the ℓ_∞ budget to ϵ=2/255 produces a more severe trade-off: Camera Viewpoint and Sensor Noise remain substantially improved, while several other shifts degrade much more strongly. A similar pattern also appears under ℓ_2 AT with ϵ=2, which again yields large gains on Camera Viewpoint and Sensor Noise at the cost of degradations on other shifts. In contrast, the weaker ℓ_2, ϵ=1 configuration does not exhibit the same Camera Viewpoint gain and shows a qualitatively different generalization profile. Thus, the shift-specific effects of Direct AT are reproducible across different threat models and perturbation budgets, while their magnitude and failure profile depend strongly on the AT configuration.

View reliance across threat models. We next test whether the view collapse identified for the main ℓ_∞, ϵ=1/255 configuration also appears under alternative AT settings. Figure 4 shows the same near-complete wrist-dominant collapse for ℓ_∞, ϵ=2/255 and ℓ_2, ϵ=2, both of which substantially improve Camera Viewpoint. In contrast, ℓ_2, ϵ=1, which does not improve Camera Viewpoint, shows a qualitatively different reliance pattern and is more sensitive to third-person degradation under graded blur. Together, these results show that the severity of single-view reliance depends on the AT configuration, with several settings reaching near-complete collapse.

View reliance across model configurations. We extend the view-intervention analysis to additional model configurations after a short 3k-step Direct AT. Under graded blur, all tested configurations are more sensitive to wrist-view degradation than to third-person-view degradation, indicating wrist-biased reliance (Figure 5). Black-out responses vary across models, and not all configurations exhibit the near-complete view collapse observed in our primary setting. These results suggest that wrist-biased reliance under Direct AT extends beyond the primary model, although its severity varies across configurations.

E View Swap versus Matched View Dropout

We further test whether simply removing one visual source is sufficient to obtain the benefit of View Swap. As a matched baseline, we use View Dropout, which blacks out one randomly selected view at the same 25% training frequency. Both variants use the same SFT initialization and adversarial-training configuration; only the view intervention is changed.

Our hypothesis is that View Swap provides a more informative reliability intervention than View Dropout. A blacked-out view is an obvious out-of-distribution signal that the source is unavailable, allowing the policy to detect and ignore it directly. In contrast, a swapped view remains visually

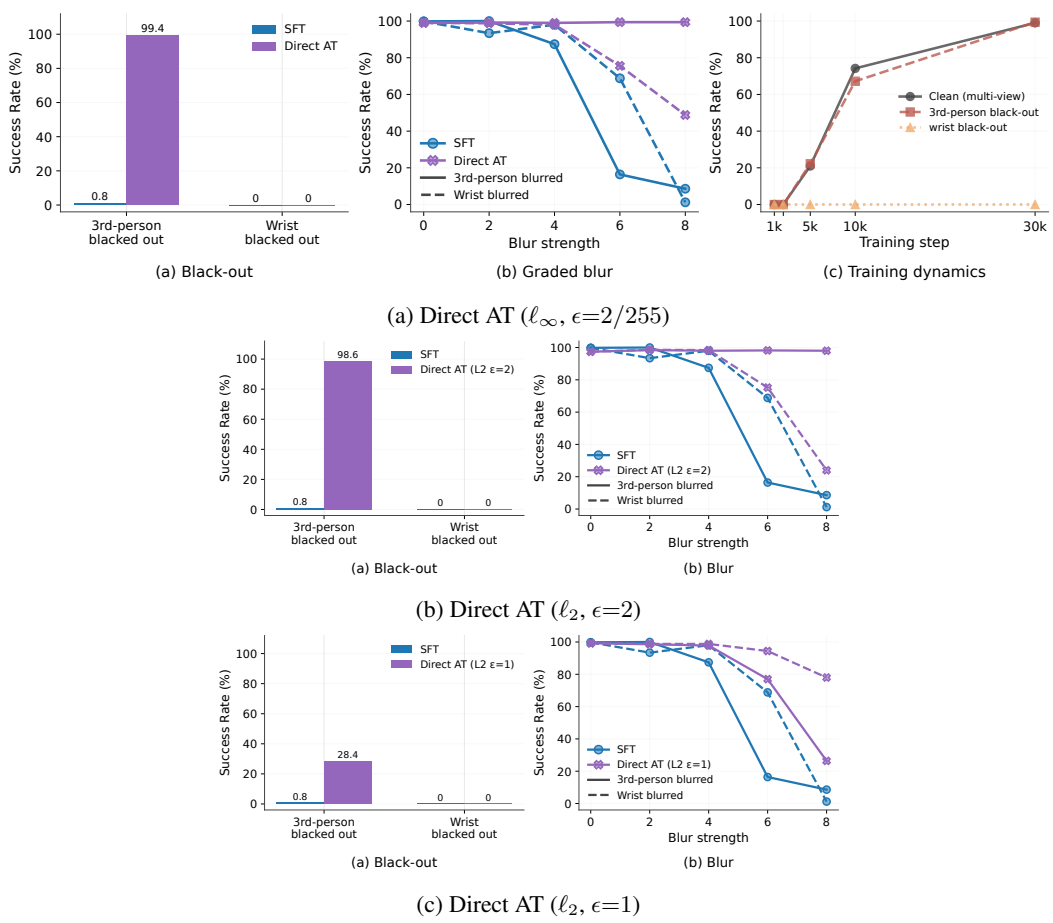


Figure 4: **View-intervention results for additional Direct AT configurations.** The two settings that substantially improve Camera Viewpoint exhibit near-complete wrist-dominant collapse, whereas ℓ_2, ϵ=1 shows a qualitatively different pattern with greater sensitivity to third-person degradation.

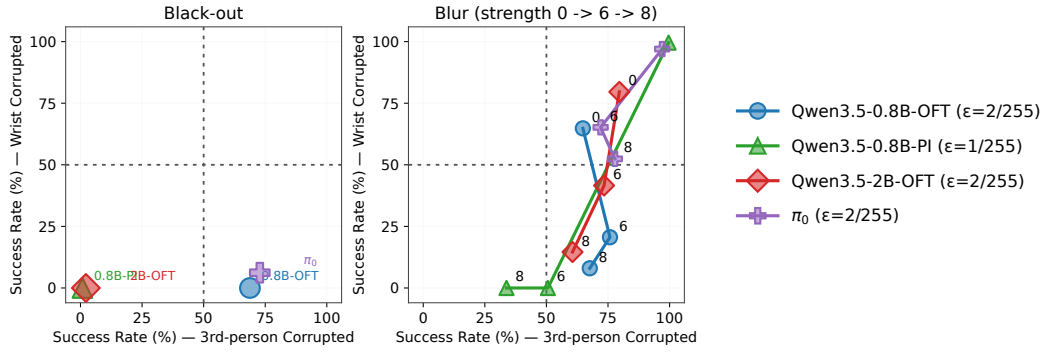


Figure 5: **View reliance after 3k steps of Direct AT across model configurations.** Success rates under third-person-view corruption (horizontal axis) and wrist-view corruption (vertical axis). Under graded blur, all configurations show greater sensitivity to wrist-view degradation, while black-out responses vary. Perturbation budgets are shown in the legend.

plausible and preserves the appearance statistics of the corresponding camera, but its content is inconsistent with the current task context. The policy must therefore learn that an apparently valid view can still be unreliable.

Table 13: **View Swap versus matched View Dropout on LIBERO-Plus.** Both variants use the same SFT initialization, adversarial-training configuration, and view-intervention frequency. Values are four-suite success rates (%).

	Original	LIBERO-Plus							
Method	**Total**	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
SFT→AT+Dropout	**97.3**	63.6	60.4	**72.3**	86.1	**91.0**	85.1	76.9	75.4
SFT→AT+Swap	96.5	**84.1**	**65.2**	68.4	**88.8**	87.7	**88.3**	**79.5**	**79.7**
Δ (Swap − Dropout)	-0.8	+20.5	+4.8	-3.8	+2.7	-3.3	+3.1	+2.6	+4.3

As shown in Table 13, View Swap provides stronger overall robustness than matched View Dropout while maintaining comparable original-task performance. This result supports our hypothesis that training on plausible-but-unreliable views provides a more useful signal than exposing the policy only to completely missing visual inputs.

F Comparison with Prior Methods on LIBERO-Plus

To contextualize our results, we compare our primary Qwen3.5-0.8B-OFT model with representative methods evaluated on LIBERO-Plus. We distinguish direct VLM-to-VLA adaptation from models with prior robot-action pretraining; all models except OpenVLA-OFT + PT are evaluated zero-shot on LIBERO-Plus without training on its distribution shifts.

Table 14: **Comparison with prior methods on LIBERO-Plus.** Values are four-suite pooled success rates (%). Bold indicates the best result among models directly adapted from a VLM using only standard LIBERO training data.

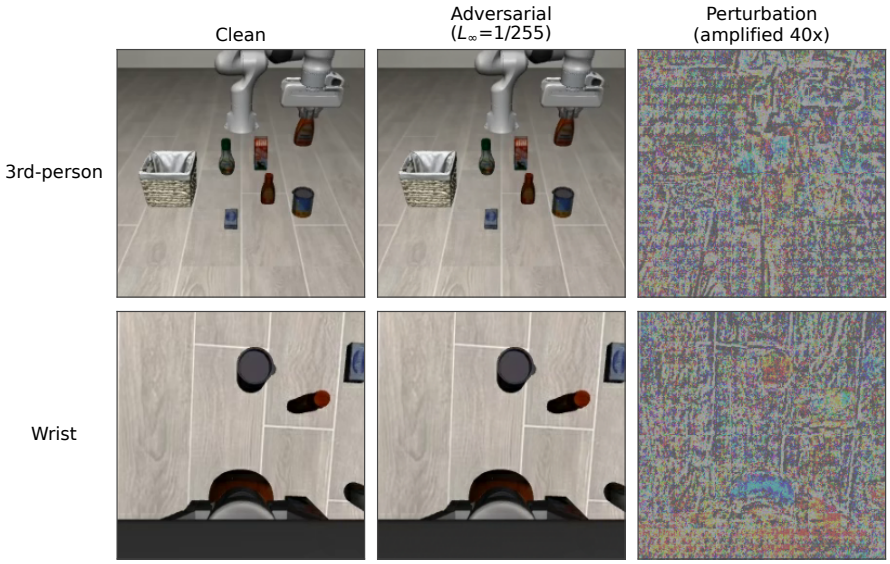
Method	**Size**	**Camera**	**Robot**	**Language**	**Light**	**Background**	**Noise**	**Layout**	**Total**
Direct VLM-to-VLA adaptation on standard LIBERO									
Qwen3-VL-OFT	4B	47.0	60.1	**87.0**	**96.3**	**95.3**	73.1	79.2	75.0
Qwen3-VL-PI	4B	64.3	57.2	82.8	94.2	94.0	79.6	78.2	77.0
Qwen3.5-0.8B-OFT (SFT)	0.8B	38.8	55.7	71.0	88.9	88.5	59.1	69.8	65.3
Qwen3.5-0.8B-OFT (SFT→AT+Swap)	0.8B	**84.1**	**65.2**	68.4	88.8	87.7	**88.3**	**79.5**	**79.7**
Robot-pretrained VLAs finetuned on standard LIBERO									
OpenVLA-OFT (Fei et al., 2026b)	7B	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0 (public repro.)‡	3.3B	17.3	6.7	63.6	86.5	82.1	17.6	74.2	46.3
π_0-Fast (Fei et al., 2026b)	3.3B	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
ABot-M0 (Yang et al., 2026)	4B	60.4	67.9	86.4	96.2	91.6	86.4	82.6	80.5
$\pi_{0.5}$ (Physical Intelligence, 2025)	3.3B	78.4	73.6	80.8	96.2	94.1	89.0	84.5	84.4
Qwen-RobotManip (Yuan et al., 2026)	4B	87.2	75.5	85.6	96.6	97.7	97.7	87.3	89.0
Training with generalized LIBERO data									
OpenVLA-OFT + PT† (Fei et al., 2026b)	7B	92.8	30.3	85.8	94.9	93.9	89.3	77.6	79.5

Model size denotes the nominal model/backbone scale; auxiliary action-head parameters are not consistently reported. †PT denotes post-training with more than 20K additional generalized trajectories generated through the LIBERO-Plus pipeline. LIBERO-Plus results for $\pi_{0.5}$ are taken from Yuan et al. (2026). ‡For π_0, we use the public full-benchmark reproduction with the official OpenPI checkpoint (Skywalker-yqz, 2026), which reports 17.6% Sensor Noise versus 79.0% in the original leaderboard.

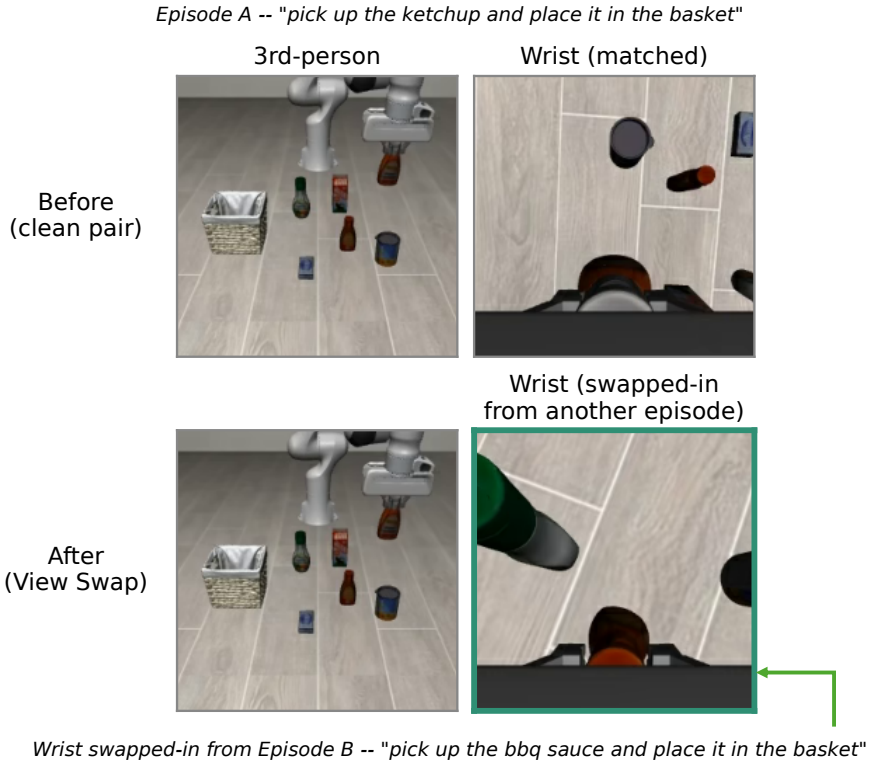
Qwen3-VL-OFT and Qwen3-VL-PI results are taken from the StarVLA LIBERO-Plus evaluation. Our 0.8B direct VLM-to-VLA baseline improves from 65.3% to 79.7% overall success with SFT→AT+Swap. The resulting model achieves the highest average among the direct VLM-to-VLA models in this comparison, with particularly strong Camera Viewpoint (84.1%) and Sensor Noise (88.3%) robustness, while remaining competitive with several substantially larger robot-pretrained VLAs.

G Visualization of Training Perturbations

Figure 6 provides qualitative examples of the two input perturbations used in our training. Pixel-space AT applies small adversarial perturbations to both camera views, whereas View Swap explicitly breaks the correspondence between views by replacing one view with a frame from another sample.



(a) **Adversarial perturbation.** Clean and adversarial observations under 1-step PGD with ℓ_∞, ϵ=$1/255$. The right column visualizes the perturbation amplified by $40\times$.



(b) **View Swap.** The third-person observation is kept unchanged, while the wrist observation is replaced by a frame from another sample, producing a mismatched multi-view pair.

Figure 6: **Visualization of the training perturbations.** Adversarial training perturbs pixels within each view, while View Swap perturbs the correspondence across views.