BEYOND IN-DISTRIBUTION PRESERVATION: RECOVERING GENERALIZATION IN QUANTIZED VLAS VIA VULNERABILITY-ORIENTED TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization has been shown to preserve VLA performance under standard evaluation conditions, but whether it preserves the full-precision model's robustness and generalization remains underexplored. In this study, we systematically study the robustness and generalization of post-quantized VLA policies under environmental disturbances. Empirical results show that quantized policies can become fragile to subtle environmental variations despite retaining comparable in-distribution performance. We further observe that action discrepancies are concentrated in a small subset of rollout states, while teacher guidance has opposite effects depending on discrepancy: it improves generalization at high-discrepancy states but can degrade it at low-discrepancy states. These findings reveal that effective post-quantization recovery requires selectively intervening on vulnerable states rather than globally distilling the student. We therefore propose Policy-Induced Vulnerability-Oriented Tuning (PIVOT-Q), a vulnerability-aware On-Policy Distillation (OPD) framework that selectively corrects vulnerable states encountered during quantized-student rollouts using the frozen full-precision policy as a teacher. PIVOT-Q identifies vulnerable states using discounted accumulated discrepancies over a short horizon, applies phase-balanced sparse supervision, and uses a Behavioral Anchor to prevent unnecessary changes. Experiments under seven LIBERO-Plus environmental variations demonstrate consistent recovery across multiple VLA backbones and quantization methods. Notably, PIVOT-Q consistently outperforms full-state distillation across all settings while using only 7.4% of its state-level distillation budget.

1 INTRODUCTION

Vision-language-action (VLA) policies have demonstrated strong capabilities in language-guided robotic manipulation, yet their substantial computational and memory footprints remain a major obstacle to efficient deployment (Kim et al., 2025). Post-training quantization (PTQ) provides a practical way to reduce these costs, and recent methods have shown that VLA policies can be substantially compressed while largely preserving their performance under in-distribution evaluation settings (Frantar et al., 2023; Xiao et al., 2023; Zhang et al., 2026; Xu et al., 2026; Wang et al., 2026). However, preserving performance on the original evaluation does not guarantee comparable generalization under environmental variations. To the best of our knowledge, the effect of quantization on VLA generalization under such variations has not been systematically studied.

In this work, we begin from an empirical study of VLA generalization after post-training quantization. Across multiple VLA backbones and quantization methods, we identify a consistent post-quantization generalization gap: quantized policies degrade substantially more than their full-precision counterparts under subtle disturbances in the environment, such as slight change of lighting, initial object poses, etc. For example, on LIBERO-Plus (Fei et al., 2026), QuantVLA reduces the success rate of GR00T-N1.5 from 82.1% to 67.0%, whereas it manages to preserve 97% success rate on its in-distribution LIBERO (Liu et al., 2023) environments. This finding raises a key question: *what behavioral changes induced by quantization underlie this generalization gap?*

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.

To investigate this question, we analyze action discrepancies between quantized and full-precision policies during rollouts under environmental disturbances. Specifically, we roll out the quantized policy and, at each visited state, compare its predicted action with that of its full-precision counterpart. Across the evaluated VLA models, these discrepancies consistently exhibit an approximately log-normal distribution: while most states incur only minor deviations, large discrepancies are concentrated in a sparse tail. This highly non-uniform distribution suggests that a small subset of high-discrepancy states may play a disproportionate role in generalization degradation. More importantly, selectively correcting high-discrepancy states substantially restores generalization, whereas correcting low-discrepancy states surprisingly leads to even worse performance. Effective post-quantization recovery does not require correcting every state, it requires correcting the right states.

Motivated by these observations, we propose Policy-Induced Vulnerability-Oriented Tuning (PIVOT-Q), a vulnerability-aware On-Policy Distillation (OPD) framework that selectively corrects sensitive states encountered by the quantized student under the guidance of the frozen full-precision teacher. Rather than uniformly distilling the student, PIVOT-Q focuses supervision on states where quantization-induced deviations are most consequential, weighting their contribution according to their vulnerability. To prevent targeted recovery from disrupting behaviors already preserved after quantization, we further introduce a Behavioral Anchor that regularizes the student toward the original policy on reference states. Together, these components enable targeted correction where it matters most while preserving behaviors that are already robust to quantization.

We evaluate PIVOT-Q across four model–quantizer settings, including GR00T-N1.5 and $\pi_{0.5}$ with W4A8 QuantVLA, GR00T-N1.5 with W4A4 HoloQ-VLA, and OpenVLA-OFT with QVLA. Entensive experiments are conducted on LIBERO-Plus under seven controlled distribution shifts. PIVOT-Q consistently improves the corresponding quantized policies. It also outperforms full-state distillation in all four settings while using only 7.4% of the state-level distillation budget. These results demonstrate that vulnerability-oriented sparse supervision provides consistent post-quantization recovery across different VLA backbones, quantization schemes, and environmental shifts.

Our contributions include: ♠ We provide a systematic study of environmental generalization after VLA quantization, revealing a consistent generalization gap and a non-uniform distribution of action discrepancies. ♠ We propose PIVOT-Q, a vulnerability-aware post-quantization recovery framework that prioritizes a sparse set of student rollout states, while mitigating unintended policy drift with a Behavioral Anchor. ♠ We demonstrate the effectiveness of PIVOT-Q across multiple VLA backbones and quantization methods. PIVOT-Q consistently outperforms full-state distillation across all our evaluation settings while using only 7.4% of its state-level distillation budget.

2 Related Work

VLA Generalization and Quantization. VLA policies span discrete-action systems such as RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2025), as well as continuous generative policies such as $\pi_{0.5}$ (Black et al., 2025) and GR00T (Bjorck et al., 2025). Their environmental robustness has been evaluated under controlled variations in viewpoint, lighting, language, background, and other factors (Fei et al., 2026). In parallel, post-training quantization reduces inference cost through low-precision weights and activations (Yao et al., 2022; Yuan et al., 2023; Lin et al., 2024). Recent VLA-specific methods, including QuantVLA (Zhang et al., 2026), QVLA (Xu et al., 2026), and HoloQ-VLA (Wang et al., 2026), further improve action preservation after quantization. However, their evaluations primarily focus on the original task distributions, leaving environmental generalization after VLA quantization largely unexplored.

Policy Distillation and On-Policy Adaptation. Knowledge distillation has been widely used for model compression and policy transfer (Hinton et al., 2015; Rusu et al., 2016; Fang, 2025). Recent VLA methods extend this paradigm to action generation: VITA-VLA (Dong et al., 2025) and Act-Distill (Ye et al., 2025) transfer action knowledge to lightweight policies, while VLA-OPD (Zhong et al., 2026) performs distillation on student-visited states to better match the deployment-time state distribution. Quantization-aware approaches such as QAD (Xin et al., 2026) and SQuaT (Lee et al., 2026) further use high-precision teachers to recover low-precision models. However, existing recovery methods do not explicitly account for how the behavioral importance of quantization errors varies across rollout states, particularly under environmental shifts.

3 PROBLEM FORMULATION

Let π^F denote a pretrained full-precision VLA policy and π^Q its post-training quantized counterpart. Given an observation o_t, language instruction x, and stochastic input ξ_t, both policies predict an action chunk, while only the first action is executed in the environment. We denote the corresponding executed actions as $\mathbf{a}_t^F$ and $\mathbf{a}_t^Q$. When comparing the two policies, we use the same stochastic input ξ_t to isolate the behavioral deviation introduced by quantization.

In this work, we focus on post-quantization recovery under environmental variations. Since deviations in π^Q can alter subsequent observations, recovery should account for the state distribution induced by the quantized policy itself rather than relying only on fixed calibration data or teacher-generated trajectories. Accordingly, we consider trajectories collected by executing π^Q and measure the teacher–student action discrepancy at each visited state as:

$$q_t = \frac{1}{D_a} \left\| \mathbf{a}_t^Q - \mathbf{a}_t^F \right\|_2^2, \tag{1}$$

where D_a denotes the action dimensionality. Our goal is to recover the environmental generalization of π^Q using the frozen π^F as a teacher, while limiting supervision to a small subset of student-visited states and avoiding unnecessary changes to behaviors already preserved after quantization.

4 MOTIVATION STUDY

Although quantization improves the deployment efficiency of VLA policies while preserving strong performance under standard evaluation conditions, whether quantized VLA policies remain robust under environmental variations has not been studied. We investigate this question by evaluating post-quantization environmental generalization and analyzing how quantization-induced action discrepancies are distributed across rollout states and whether they are behaviorally consequential.

4.1 POST-QUANTIZATION GENERALIZATION GAP

We first examine the impact of quantization on model generalization by comparing the success rates of the original and quantized policies under environmental shifts. We consider VLA policies trained on the LIBERO suite (Liu et al., 2023) and evaluate their post-quantization generalization on LIBERO-Plus (Fei et al., 2026). By preserving the task objective while introducing controlled environmental shifts, LIBERO-Plus enables a direct comparison of the environmental robustness of full-precision and quantized policies.

As shown in Figure 1(a), quantization consistently reduces success rate across our evaluated models and quantization methods. GR00T-N1.5 with QuantVLA drops from 82.1% to 67.0%, $\pi_{0.5}$ with QuantVLA drops from 78.2% to 72.0%, and GR00T-N1.5 with HoloQ-VLA drops from 82.0% to 69.1%. These results reveal a consistent post-quantization generalization gap, suggesting that strong performance retention under standard evaluation does not imply preservation of model robustness.

4.2 ACTION DISCREPANCY ANALYSIS

The observed generalization gap motivates a detailed examination of how quantization changes policy behavior during execution. We use the quantized policies to sample new rollouts in the generalized environments again. At each visited state, the original and the quantized policies are provided with the same observation and diffusion noise to produce their actions.

Figure 1(b) summarizes the resulting action-discrepancy distributions across the evaluated model–quantizer settings. For each rollout, we group the log-scaled discrepancy into bins and divide the count in each bin by the total number of valid states. These proportions are averaged across rollouts to estimate the distribution of action discrepancy. As shown in Figure 1(b), the discrepancies exhibit approximately log-normal distributions across all evaluated settings, with a pronounced right tail. Detailed analysis is provided in Appendix B. Most visited states lie near the main concentration of the distribution, whereas substantially larger deviations occur at a smaller fraction of rollout states. To evaluate whether large action discrepancies are behaviorally consequential, we conduct an intervention study on GR00T-N1.5 with QuantVLA. During quantized-policy rollouts, we replace

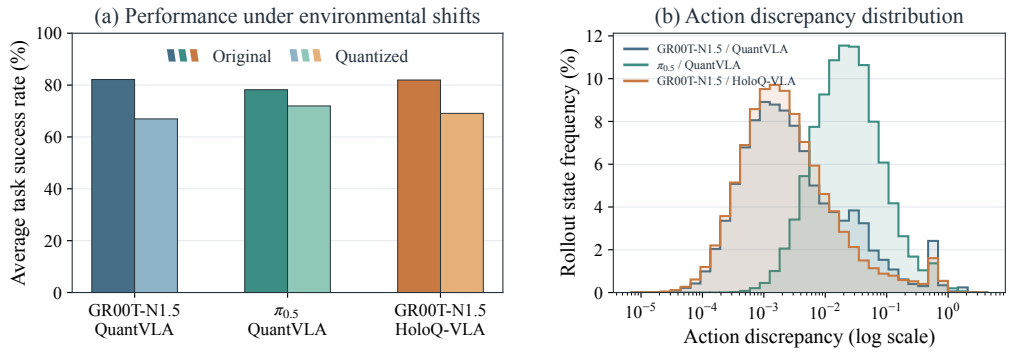

Figure 1: **Generalization degradation after quantization and action discrepancy.** (a) Average success over 560 LIBERO-Plus episodes for each original and quantized policy. (b) Distributions of action discrepancy between each quantized policy and its original counterpart. The horizontal action discrepancy axis is in log-scale.

Table 1: **Effect of intervention at different action-discrepancy levels** Task success rate (%) in the GR00T-N1.5 QuantVLA setting. "Ran. Int.", "HD Int.", and "LD Int." denote interventions at random, high-discrepancy, and low-discrepancy states, respectively. Superscripts indicate absolute changes relative to QuantVLA. Best and second-best results among intervention strategies are shown in **bold** and underline, respectively. The full-precision original policy is shaded as a reference.

Method	Cam.	Init.	Lang.	Light	Back.	Noise	Layout	Mean
Original	85.0	42.5	78.8	95.0	96.2	85.0	88.8	81.6
QuantVLA	61.2	30.0	60.0	92.5	75.0	61.2	85.0	66.4
Ran. Int.	66.3$^{\uparrow 5.1}$	**33.8**$^{\uparrow 3.8}$	61.3$^{\uparrow 1.3}$	93.8$^{\uparrow 1.3}$	75.0$^{0.0}$	60.0$^{\downarrow 1.2}$	**87.5**$^{\uparrow 2.5}$	68.2$^{\uparrow 1.8}$
HD Int.	**67.5**$^{\uparrow 6.3}$	30.0$^{0.0}$	**76.2**$^{\uparrow 16.2}$	**96.2**$^{\uparrow 3.7}$	76.2$^{\uparrow 1.2}$	**87.5**$^{\uparrow 26.3}$	82.5$^{\downarrow 2.5}$	**73.8**$^{\uparrow 7.4}$
LD Int.	62.5$^{\uparrow 1.3}$	32.5$^{\uparrow 2.5}$	58.8$^{\downarrow 1.2}$	86.2$^{\downarrow 6.3}$	**78.8**$^{\uparrow 3.8}$	58.8$^{\downarrow 2.4}$	86.2$^{\uparrow 1.2}$	66.2$^{\downarrow 0.2}$

the quantized action with the corresponding full-precision action at selected states, using high-discrepancy, low-discrepancy, or random states as intervention targets. As shown in Table 1, intervening at high-discrepancy states substantially improves task success. Random intervention yields a mild improvement. In contrast, intervening at low-discrepancy states provides no improvement and even slightly degrades performance. These results indicate that quantization-induced discrepancies are non-uniform in their behavioral impact, motivating recovery to selectively prioritize consequential states rather than uniformly distill all rollout states. *Effective post-quantization recovery does not require correcting every state, it just requires correcting the right ones.*

5 PIVOT-Q: Policy-Induced Vulnerability-Oriented Tuning

To identify the right states for recovery, we propose Policy-Induced Vulnerability-Oriented Tuning (PIVOT-Q), a vulnerability-aware OPD framework that selectively corrects vulnerable states encountered by the quantized student under the guidance of the frozen original policy. PIVOT-Q first performs student-controlled on-policy rollouts while measuring teacher–student discrepancies along the resulting trajectories. It then identifies a sparse set of vulnerable states based on their discounted accumulated discrepancies over a short future horizon, with phase-wise top-k selection to ensure coverage throughout the rollout. The selected states are distilled with vulnerability-aware weighting, while a Behavioral Anchor regularizes the adapted student on reference states from the original distribution to avoid unintended drift. Figure 2 summarizes the overall pipeline.

5.1 Student-Controlled On-Policy Rollouts

Fixed calibration data or teacher-generated trajectories may not represent the states encountered after quantization, as action deviations can alter subsequent observations. We therefore collect recovery trajectories on-policy using the quantized student π^Q. At each visited observation O_t, the student π^Q

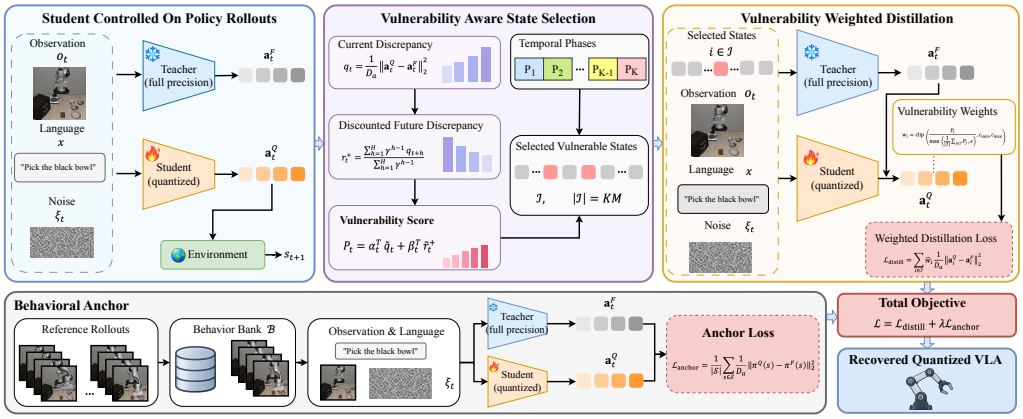

Figure 2: **Overview of PIVOT-Q.** PIVOT-Q identifies vulnerable states from student-controlled rollouts using discounted accumulated discrepancy within a short future segment, performs phase-balanced sparse distillation, and uses a Behavioral Anchor to preserve retained behavior.

and frozen full-precision teacher π^F predict actions $\mathbf{a}_t^Q$ and $\mathbf{a}_t^F$ under the same diffusion noise ξ_t, while only $\mathbf{a}_t^Q$ is executed. This ensures that rollouts follow the student-induced state distribution, while controlling generation stochasticity so that teacher–student discrepancies provide a consistent measure of vulnerability at visited states.

5.2 Vulnerability Aware State Selection

The action discrepancy q_t captures only the immediate teacher–student deviation, while vulnerability may manifest through persistent or emerging disagreement in subsequent steps. Inspired by the discounted return in reinforcement learning, we assess each state by accumulating future discrepancies, emphasizing nearby deviations while reducing the influence of increasingly uncertain distant steps. We define the discounted accumulated discrepancy over future states within a horizon H as:

$$r_t^+ = \frac{\sum_{h=1}^{H} \gamma^{h-1} q_{t+h}}{\sum_{h=1}^{H} \gamma^{h-1}}, \tag{2}$$

where γ is the temporal discount factor. This score assigns greater importance to temporally closer steps while complementing the instantaneous discrepancy q_t.

Greedily selecting the states of the highest discounted future discrepancy globally may concentrate supervision within a narrow temporal region, as neighboring states might exhibit correlated discrepancy patterns. To ensure temporal coverage over the whole trajectory, we partition each trajectory into K contiguous phases and compute within-phase percentile ranks α_t^T, β_t^T for both discrepancy signals. We denote the resulting normalized ranks of the current and discounted future discrepancies as $\tilde{q}_t$ and $\tilde{r}_t^+$. The vulnerability score is finally defined as:

$$P_t = \alpha_t^T \, \tilde{q}_t + \beta_t^T \, \tilde{r}_t^+, \tag{3}$$

where α_t^T and β_t^T control the relative contributions of current discrepancy and discounted future discrepancy within the corresponding phase and superscript T denotes total steps in the phase. Both $\tilde{q}_t$ and $\tilde{r}_t^+$ are computed solely from teacher and student action predictions. No additional annotations such as rewards or task success labels are required.

Since high-vulnerability states might cluster within a short temporal interval inside each phase, directly selecting the top-k states can allocate multiple supervision slots to neighboring states. We therefore greedily select states in descending order of P_t, enforcing a minimum temporal separation between any two selected states within the same phase. We retain M states per phase, resulting in a sparse supervision set $\mathcal{I}$ with $|\mathcal{I}| = KM$. This constraint complements the phase-wise allocation by preventing the supervision budget from concentrating on a narrow temporal region while prioritizing states with high estimated vulnerability.

5.3 Vulnerability Weighted Distillation

Although state selection restricts supervision to vulnerable states, the selected states can still differ substantially in their estimated vulnerability. Uniformly weighting them loses this information, so we assign larger distillation weights to states with higher vulnerability scores P_i. Since the scale of P_i can vary across trajectories and extreme scores may cause a few states to dominate optimization, we normalize each score by the mean vulnerability over the selected set and clip the resulting ratio. This converts the absolute vulnerability score into a bounded relative importance, emphasizing more vulnerable states while maintaining stable contributions from the remaining selected states. The clipped weights are renormalized over the selected set before computing the distillation objective:

$$w_i = \mathrm{clip}\left(\frac{P_i}{\max\left(\frac{1}{|\mathcal{I}|}\sum_{j\in\mathcal{I}} P_j, \epsilon\right)}, c_{\min}, c_{\max}\right), \qquad \bar{w}_i = \frac{w_i}{\sum_{j\in\mathcal{I}} w_j}. \tag{4}$$

Mean normalization makes the weighting comparable across trajectories with different vulnerability scales. Clipping is used to prevent extreme states from dominating optimization while ensuring that every selected state retains a non-negligible contribution.

Using these weights, we distill the quantized student toward the frozen teacher at the selected states:

$$\mathcal{L}_{\mathrm{distill}} = \sum_{i\in\mathcal{I}} \bar{w}_i \frac{1}{D_a} \left\| \mathbf{a}_i^Q - \mathbf{a}_i^F \right\|_2^2. \tag{5}$$

This distillation objective concentrates the recovery update on states with higher estimated vulnerability while retaining sparse supervision over the trajectory.

5.4 Behavioral Anchor

Our selective distillation focuses supervision on vulnerable states visited by the quantized student under environmental shifts. Updating the policy exclusively on these states may alter the policy too much towards the new environment since they provide large supervision signals, leading to instability during learning and forgetting. We therefore introduce a Behavioral Anchor that regularizes the adapted policy toward the behavior of the original full precision policy on reference states. We maintain a small behavior bank $\mathcal{B}$ that includes anchor states from the original dataset used to trained the full-precision policy. The student queries the frozen full-precision teacher on these anchor states to get the resulting teacher targets for optimization, as commonly conducted in continual learning to prevent forgetting (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017). At each optimization step, a subset $\mathcal{S}$ of anchor states is randomly sampled from the memory bank $\mathcal{B}$. To further stabilize the training process and prevent policy drift, we integrate an auxiliary anchor loss:

$$\mathcal{L}_{\mathrm{anchor}} = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{1}{D_a} \left\| \pi^Q(s) - \pi^F(s) \right\|_2^2, \tag{6}$$

The overall recovery objective combines targeted correction with behavioral retention:

$$\mathcal{L} = \mathcal{L}_{\mathrm{distill}} + \lambda \mathcal{L}_{\mathrm{anchor}}, \tag{7}$$

where λ is a hyper-parameter that controls the magnitude of the Behavioral Anchor.

6 Experiments

We evaluate four model–quantizer settings: GR00T-N1.5 with W4A8 QuantVLA, $\pi_{0.5}$ with W4A8 QuantVLA, GR00T-N1.5 with W4A4 HoloQ-VLA, and OpenVLA-OFT with QVLA. Each setting includes an original policy, a quantized policy, a recovered policy with Full State Distillation (FSD), and recovered policy with PIVOT-Q. FSD distills the student on all recorded rollout states. PIVOT-Q partitions each rollout into $K = 4$ phases and selects $M = 4$ vulnerable states per phase, yielding $|\mathcal{I}| = KM = 16$ states per episode based on discounted accumulated discrepancies, with a Behavioral Anchor regularizer. All methods use the same 560-episode LIBERO-Plus protocol, and results are reported as mean $\pm$ standard deviation over three independent seeds.

Table 2: **Performance recovery across models and quantization methods.** We report task success rates (%) for seven LIBERO-Plus distribution shifts. FSD denotes *Full State Distillation*, which distills the student on all recorded rollout states. "#States" denotes the average number of distilled states per episode. Results are reported as mean ± standard deviation across three independent runs with different seeds, with standard deviations shown in gray. Boldface marks the larger value between FSD and PIVOT-Q for each reported shift and the overall average.

Model	Method	Cam.	Init.	Lang.	Light	Back.	Noise	Layout	Avg.↑	#States	Δ ↑
HoloQ-VLA (W4A4)											
GR00T-N1.5	Original	86.3 ± 2.5	35.0 ± 3.3	77.9 ± 1.4	97.5 ± 0.0	97.9 ± 0.7	87.5 ± 0.0	91.7 ± 1.9	82.0 ± 0.2	–	–
	Quantized	67.5 ± 1.2	27.1 ± 3.1	65.0 ± 2.2	95.0 ± 1.2	81.7 ± 2.6	59.2 ± 3.1	82.1 ± 2.6	68.2 ± 0.5	–	−13.8
	FSD	79.2 ± 1.9	32.1 ± 2.6	**75.4** ± 3.8	**96.2** ± 2.5	**98.8** ± 1.2	74.2 ± 1.9	90.0 ± 1.2	78.0 ± 0.4	199.4	−4.0
	PIVOT-Q	**85.0** ± 2.2	**36.2** ± 4.3	71.2 ± 5.7	92.5 ± 3.3	96.7 ± 0.7	**75.4** ± 4.0	**91.2** ± 0.0	**78.3** ± 1.6	16.0	−3.7
QuantVLA (W4A8)											
GR00T-N1.5	Original	84.6 ± 0.7	34.6 ± 2.6	78.3 ± 0.7	96.7 ± 1.9	97.9 ± 0.7	87.9 ± 2.6	90.0 ± 2.2	81.4 ± 0.8	–	–
	Quantized	60.4 ± 2.6	31.7 ± 3.1	61.2 ± 2.2	86.7 ± 3.6	76.7 ± 1.4	66.2 ± 1.2	83.3 ± 0.7	66.6 ± 0.9	–	−14.8
	FSD	**83.3** ± 1.4	33.3 ± 2.9	70.0 ± 2.2	99.2 ± 0.7	99.2 ± 0.7	73.8 ± 1.2	79.6 ± 10.1	76.9 ± 1.6	210.0	−4.5
	PIVOT-Q	77.9 ± 2.6	**37.9** ± 1.9	**79.6** ± 1.9	99.2 ± 0.7	**99.6** ± 0.7	**85.4** ± 3.1	**89.6** ± 0.7	**81.3** ± 0.4	16.0	−0.1
$\pi_{0.5}$	Original	82.1 ± 4.7	72.9 ± 1.4	70.0 ± 2.5	86.7 ± 0.7	89.6 ± 3.1	70.8 ± 3.6	80.4 ± 1.9	78.9 ± 0.6	–	–
	Quantized	57.1 ± 5.1	70.0 ± 2.5	57.9 ± 1.4	86.3 ± 3.3	80.4 ± 0.7	67.5 ± 0.0	84.2 ± 1.4	71.9 ± 0.6	–	−7.0
	FSD	**68.3** ± 5.9	66.3 ± 2.5	**65.8** ± 2.6	83.3 ± 2.6	**82.1** ± 1.9	64.2 ± 0.7	79.2 ± 0.7	72.7 ± 0.5	232.7	−6.2
	PIVOT-Q	67.5 ± 6.3	**71.7** ± 3.8	57.9 ± 4.0	83.3 ± 2.6	80.0 ± 1.3	**66.7** ± 2.6	**82.5** ± 4.3	**72.8** ± 1.3	16.0	−6.1
QVLA (Mixed-bit W-only)											
OpenVLA-OFT	Original	55.0 ± 0.0	21.7 ± 0.7	86.3 ± 0.0	77.1 ± 0.7	68.8 ± 0.0	48.3 ± 0.7	87.5 ± 1.3	63.5 ± 0.3	–	–
	Quantized	50.0 ± 0.0	22.9 ± 0.7	76.3 ± 0.0	40.4 ± 2.9	66.3 ± 2.2	42.5 ± 1.3	70.4 ± 3.1	52.7 ± 0.8	–	−10.8
	FSD	**59.6** ± 0.7	20.0 ± 0.0	**71.2** ± 0.0	**72.9** ± 1.9	**79.6** ± 1.9	**48.8** ± 1.2	70.8 ± 2.6	60.4 ± 0.9	242.0	−3.1
	PIVOT-Q	57.5 ± 0.0	**23.3** ± 0.7	69.6 ± 0.7	70.4 ± 2.6	77.5 ± 2.2	44.2 ± 1.9	**83.3** ± 1.9	**60.8** ± 0.3	16.0	−2.7

6.1 Main Results

Table 2 shows that PIVOT-Q provides consistent post-quantization recovery across model backbones and quantization methods. PIVOT-Q improves over the corresponding policy in all four settings and achieves a higher average success rate than FSD in every case. The gains are obtained with a fixed sparse supervision rule, suggesting that recovery quality depends more on identifying behaviorally consequential states than on increasing the density of state-level supervision.

Recovery under QuantVLA. For GR00T-N1.5 with W4A8 QuantVLA, quantization reduces the average success rate from 81.4% to 66.6%. PIVOT-Q raises the score to 81.3%, recovering nearly the entire quantization gap and approaching the original-policy reference, while FSD reaches only 76.9%. PIVOT-Q is particularly effective under robot initialization, language, background, sensor noise, and object-layout shifts, where it also exceeds FSD; the only category in which FSD is stronger is camera viewpoint. For $\pi_{0.5}$ with W4A8 QuantVLA, the original, quantized, FSD, and PIVOT-Q policies obtain average success rates of 78.9%, 71.9%, 72.7%, and 72.8%, respectively. Although the recovery gap is smaller for such setting, PIVOT-Q still improves over both the quantized policy and FSD, with its largest advantages under sensor noise and layout variations.

Recovery across HoloQ-VLA and QVLA. PIVOT-Q retains the same advantage under alternative quantization pipelines. For GR00T-N1.5 with W4A4 HoloQ-VLA, quantization decreases the average success rate from 82.0% to 68.2%. PIVOT-Q recovers the score to 78.3%, slightly exceeding FSD at 78.0% and recovering most of the lost performance. Its strongest improvements over FSD occur under camera viewpoint, initialization, sensor noise, and layout shifts, whereas FSD remains stronger under language, lighting, and background shifts. For OpenVLA-OFT with QVLA, PIVOT-Q improves the quantized policy from 52.7% to 60.8%, exceeding FSD at 60.4% and approaching the original-policy reference of 63.5%. The improvement is most pronounced for initialization and object layout, showing that the proposed selection mechanism remains effective when the backbone and quantization implementation are changed.

State-selection efficiency results. PIVOT-Q achieves recovery with a substantially smaller state-level supervision budget than FSD. It partitions each rollout into four phases and selects four states per phase, yielding a fixed set of 16 states per episode in every evaluated setting. In contrast, the corresponding FSD runs use, on average, 210.0, 232.7, 199.4, and 242.0 states per episode for GR00T-N1.5 with QuantVLA, $\pi_{0.5}$ with QuantVLA, GR00T-N1.5 with HoloQ-VLA, and OpenVLA-OFT with QVLA, respectively. Comparing the total selected-state count with the total FSD state count

Table 3: **Component ablations and sparse selection baselines.** Task success rate (SR, %) on 560 LIBERO-Plus episodes with GR00T-N1.5 and W4A8 QuantVLA. All recovery methods use the same 16-state supervision budget.

Method	Selection	Anchor	Cam.	Init.	Lang.	Light	Back.	Noise	Layout	Avg.↑
Original	–	–	84.6 ±0.7	34.6 ±2.6	78.3 ±0.7	96.7 ±1.9	97.9 ±0.7	87.9 ±2.6	90.0 ±2.2	81.4 ±0.8
Quantized	–	–	60.4 ±2.6	31.7 ±3.1	61.3 ±2.2	86.7 ±3.6	76.7 ±1.4	66.3 ±1.3	83.3 ±0.7	66.6 ±0.9
Per-Phase Random	Random	✓	70.4 ±3.1	37.1 ±1.4	75.4 ±3.1	99.6 ±0.7	100.0 ±0.0	75.8 ±1.9	90.8 ±1.9	78.5 ±0.8
Per-Phase Uniform	Uniform	✓	82.9 ±4.4	35.0 ±3.3	66.2 ±0.0	100.0 ±0.0	100.0 ±0.0	82.1 ±3.8	89.6 ±3.1	79.4 ±1.0
Globally Greedy	Greedy	✓	74.2 ±3.1	33.3 ±2.1	72.5 ±1.8	98.3 ±1.2	99.6 ±0.6	80.8 ±0.6	92.5 ±0.0	78.8 ±0.7
w/o Future Discrepancy	C only	✓	76.7 ±5.6	31.7 ±4.4	78.8 ±1.2	100.0 ±0.0	100.0 ±0.0	82.9 ±1.4	87.5 ±1.2	79.6 ±1.2
w/o Behavioral Anchor	$C + F$	✗	75.4 ±3.8	31.2 ±2.5	78.8 ±1.2	98.8 ±1.2	92.9 ±3.1	82.5 ±1.2	92.1 ±2.6	78.8 ±1.5
PIVOT-Q (full)	$C + F$	✓	77.9 ±2.6	37.9 ±1.9	79.6 ±1.9	99.2 ±0.7	99.6 ±0.7	85.4 ±3.1	89.6 ±0.7	**81.3** ±0.4

gives budget ratios of 7.6%, 6.9%, 8.0%, and 7.0%, with an average of 7.4% across the four settings. This comparison shows that PIVOT-Q retains or improves recovery performance while using only a small fraction of the supervision required by FSD, demonstrating that vulnerability-oriented state selection is substantially more efficient than dense state-level distillation.

6.2 Ablation Studies

We evaluate PIVOT-Q and its controlled variants with GR00T-N1.5 under W4A8 QuantVLA on LIBERO-Plus. All sparse recovery variants use the same 16-state supervision budget and training protocol. *Per-Phase Random* randomly samples four states from each of the four temporal phases, whereas *Per-Phase Uniform* selects four evenly spaced valid states per phase. *Globally Greedy* ranks all valid rollout states using the vulnerability score and selects the top 16 states without phase-wise allocation, providing a direct test of whether temporal balancing is necessary. *w/o Future Discrepancy* removes the discounted future-discrepancy term from Eq. 3 and estimates vulnerability using only the current discrepancy. *w/o Behavioral Anchor* retains vulnerability-oriented state selection but removes the Behavioral Anchor regularizer. Table 3 summarizes the results.

Phase-balanced sparsity alone is insufficient. Per-Phase Random and Per-Phase Uniform obtain $78.5 \pm 0.8\%$ and $79.4 \pm 1.0\%$, respectively, whereas PIVOT-Q reaches $81.3 \pm 0.4\%$. Thus, under the same phase-wise 16-state budget, PIVOT-Q improves the average success rate by 2.8 and 1.9 percentage points over random and Per-Phase Uniform supervision. These results show that temporal coverage alone does not determine recovery quality: which states receive supervision remains important. Since the random and uniform variants remove vulnerability-based prioritization, this comparison supports vulnerability-oriented sparse supervision as a whole rather than attributing the gain to sparsity alone.

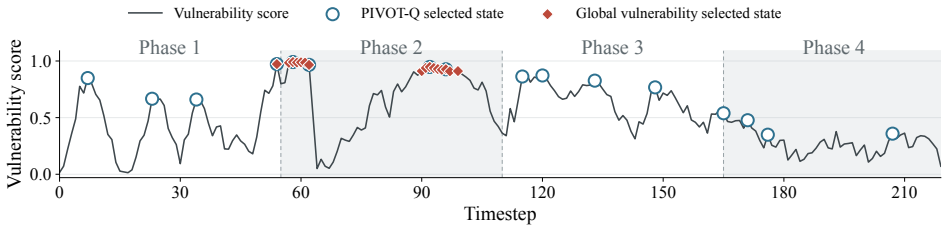

Figure 3: **Global and phase-wise vulnerability selection.** A representative rollout from LIBERO-Spatial under the Camera Viewpoints variation.

Phase-wise balanced allocation improves performance. Figure 3 illustrates a representative rollout in which global vulnerability ranking allocates all 16 selected states to Phase 2. The selected states cluster around neighboring high-vulnerability regions, leaving the other three phases without direct supervision. In Table 3, PIVOT-Q (full) improves average success by 2.5 percentage points over Globally Greedy, which uses the same vulnerability-based ranking but removes phase-wise allocation. This comparison shows that temporal balancing provides additional benefit beyond vulnerability-based prioritization alone. Additional qualitative comparisons of global and phase-wise selections are provided in Appendix D.

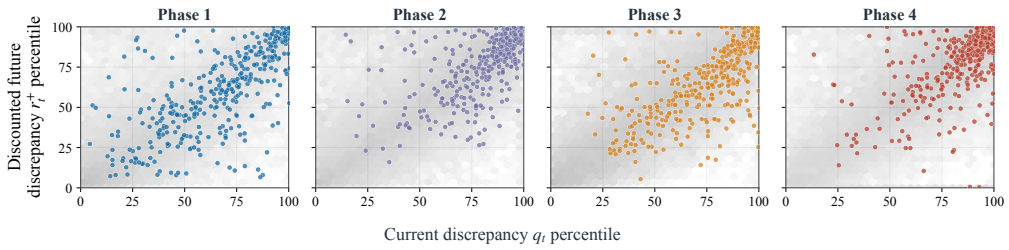

Figure 4: **Phase-wise vulnerability-oriented state selection.** Distribution of candidate and selected states across the four temporal phases for GR00T-N1.5 with W4A8 QuantVLA. Gray hexagonal bins denote candidate-state density, and colored points denote sampled PIVOT-Q selections.

Future discrepancy provides complementary selection cues. Removing the discounted future-discrepancy term decreases the average success rate from $81.3 \pm 0.4\%$ to $79.6 \pm 1.2\%$, a drop of 1.7 percentage points. This controlled variant retains phase-wise sparse selection and the Behavioral Anchor, but ranks vulnerability using only the current teacher–student discrepancy. The degradation indicates that instantaneous disagreement alone does not fully characterize states associated with subsequent behavioral deviation. Short-horizon future discrepancy therefore provides complementary information for identifying recovery states whose effects extend beyond the current timestep.

Behavioral Anchor improves both recovery and stability. Removing the Behavioral Anchor decreases the average success rate from $81.3 \pm 0.4\%$ to $78.8 \pm 1.5\%$. Besides the 2.5% reduction in mean performance, the larger variation across runs suggests less stable recovery when optimization is driven only by vulnerable rollout states. The Behavioral Anchor therefore complements targeted correction by constraining unnecessary policy drift on behaviors that are already preserved after quantization. Together, vulnerability-aware state selection and behavioral anchoring enable PIVOT-Q to recover the quantized policy to 81.3%, closely matching the 81.4% full-precision reference.

6.3 Analysis of Vulnerability-Oriented State Selection

Figure 4 visualizes the states selected by PIVOT-Q across the four temporal phases in the GR00T-N1.5 with W4A8 QuantVLA setting. The selected states are concentrated toward regions with high current discrepancy and high discounted future discrepancy, showing that PIVOT-Q preferentially targets states with pronounced teacher–student disagreement. Meanwhile, the selected states span different combinations of the two signals rather than concentrating along a single discrepancy dimension, suggesting that current and future discrepancies provide complementary cues for identifying vulnerable states. We further analyze their phase-wise relationship in Appendix C. Across all four phases, PIVOT-Q consistently favors high-vulnerability regions while retaining diverse current–future discrepancy profiles. These results show that the proposed vulnerability score effectively focuses sparse supervision on behaviorally informative states throughout the rollout.

7 Conclusion

We systematically study environmental generalization after VLA quantization and reveal a consistent post-quantization generalization gap across different VLA backbones and quantization methods. Our analysis further shows that quantization-induced action discrepancies are highly non-uniform across rollout states, and that correcting high-discrepancy states is substantially more beneficial than indiscriminate intervention. Motivated by these observations, we propose PIVOT-Q, a vulnerability-aware on-policy distillation framework that selectively supervises vulnerable student-visited states through current–future discrepancy prioritization, phase-balanced sparse selection, and behavioral anchoring. Across four model–quantizer settings, PIVOT-Q consistently improves the corresponding quantized policies and outperforms full-state distillation while using only 7.4% of its state-level supervision budget. In the primary GR00T-N1.5 with W4A8 QuantVLA setting, PIVOT-Q recovers the average success rate from 66.6% to 81.3%, closely approaching the 81.4% full-precision reference. These results highlight that effective post-quantization recovery depends not on supervising more rollout states, but on identifying and correcting the states that matter most.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and discovery, including identifying potentially relevant prior work for subsequent manual review. They were also used to support research ideation and aspects of research execution, including discussing and refining the problem formulation, providing feedback on methodological and experimental design, and assisting with the analysis and interpretation of experimental results. In addition, generative AI tools were used to improve the clarity, readability, and presentation of the manuscript.

All suggested references were manually verified against the original sources. All methodological and experimental decisions, implementations, reported results, and scientific interpretations were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including its claims, results, and conclusions.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our method and experiments throughout the paper and appendices. The formulation and complete training objective of PIVOT-Q, including student-controlled on-policy rollouts, vulnerability-aware state selection, vulnerability-weighted distillation, and the Behavioral Anchor, are described in Section 5. Section 6 specifies the evaluated model–quantizer settings, baselines, evaluation protocol, supervision budgets, and multi-seed reporting procedure. Appendix A provides implementation details and hyperparameter settings, including the LoRA configuration, optimization schedule, vulnerability-score parameters, sparse state-selection budget, and Behavioral Anchor configuration. Additional diagnostic and sensitivity analyses are provided in Appendices B–F, covering discrepancy distributions, current–future discrepancy association, temporal coverage of state selection, recovery diagnostics, and vulnerability-score weighting. We will open-source our training and evaluation code, and the scripts used to reproduce our experiments are provided in the supplementary material.

REFERENCES

Johan Bjorck et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, Haoyu Cao, Yang Gao, Xing Sun, Ran He, and Caifeng Shan. VITA-VLA: Efficiently teaching vision-language models to act via action expert distillation. *arXiv preprint arXiv:2510.09607*, 2025.

Shijie et al. Fang. Flex: A framework for learning robot-agnostic force-based skills involving sustained contact object manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4782–4788, 2025. doi: 10.1109/ICRA55743.2025.11127866.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, 2026.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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, pp. 2679–2713, 2025.

HyeonJun Lee, Hyeonsik Jo, Jinwoo Chung, and Jangho Kim. SQuaT: Self-supervised knowledge distillation via student-aware quantized teacher features. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5023–5031. PMLR, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

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 Datasets and Benchmarks Track*, volume 36, 2023.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016.

Xinyu Wang, Mingze Li, Sicheng Lyu, Dongxiu Liu, Kaicheng Yang, Ziyu Zhao, Yufei Cui, Xiao-Wen Chang, and Peng Lu. HoloQ-VLA: Uniform W4A4 quantization of vision-language-action models. *arXiv preprint arXiv:2605.28803*, 2026. URL `https://arxiv.org/abs/2605.28803v3`.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38087–38099, 2023.

Meng Xin, Sweta Priyadarshi, Jingyu Xin, Bilal Kartal, Aditya Vavre, Asma Kuriparambil Thekkumpate, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Ido Shahaf, Akhiad Bercovich, Kinjal Patel, Suguna Varshini Velury, Chenjie Luo, Zhiyu Cheng, Jenny Chen, Chen-Han Yu, Wei Ping, Oleg Rybakov, Nima Tajbakhsh, Oluwatobi Olabiyi, Dusan Stosic, Di Wu, Song Han, Eric Chung, Sharath Turuvekere Sreenivas, Bryan Catanzaro, Yoshi Suhara, Tijmen Blankevoort, and Huizi Mao. Quantization-aware distillation for NVFP4 inference accuracy recovery. *arXiv preprint arXiv:2601.20088*, 2026.

Yuhao Xu, Yantai Yang, Zhenyang Fan, Yufan Liu, Yuming Li, Bing Li, and Zhipeng Zhang. QVLA: Not all channels are equal in vision-language-action model's quantization. In *International Conference on Learning Representations*, 2026.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Wencheng Ye, Tianshi Wang, Lei Zhu, Fengling Li, Guoli Yang, and Hengtao Shen. ActDistill: General action-guided self-derived distillation for efficient vision-language-action models. *arXiv preprint arXiv:2511.18082*, 2025.

Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. RPTQ: Reorder-based post-training quantization for large language models. *arXiv preprint arXiv:2304.01089*, 2023.

Jingxuan Zhang, Yunta Hsieh, Zhongwei Wan, Haokun Lin, Xin Wang, Ziqi Wang, Yingtie Lei, and Mi Zhang. QuantVLA: Scale-calibrated post-training quantization for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39539–39549, 2026.

Zhide Zhong, Haodong Yan, Junfeng Li, Junjie He, Tianran Zhang, and Haoang Li. VLA-OPD: Bridging offline sft and online rl for vision-language-action models via on-policy distillation. *arXiv preprint arXiv:2603.26666*, 2026.

Brianna Zitkovich et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, 2023.

A Implementation Details

We use low rank adaptation (LoRA) adapters with rank 16. For GR00T-N1.5 with QuantVLA, the adapter is applied to the action head; for $\pi_{0.5}$, it is applied to the query, key, and value projections of the action expert; and for HoloQ-VLA, it is applied to the corresponding projections of the quantized action head. Each recovery rollout is followed by five optimizer updates. We use a future horizon of $H = 4$, discount factor $\gamma = 0.9$, and $\alpha = \beta = 0.5$. Each trajectory is divided into $K = 4$ phases, from which $M = 4$ states are selected per phase, yielding 16 supervision states per trajectory. Vulnerability weights are clipped to $[c_{\min}, c_{\max}] = [0.5, 2.0]$. The Behavioral Anchor uses a replay buffer of up to 256 clean states, samples up to four states per optimizer update, and uses $\lambda = 0.1$.

B Additional Analysis of Action Discrepancy

We supplement the main-text analysis of Eq. 1 with distribution fitting and empirical tail-concentration statistics. The diagnostic cohorts contain 560 quantized-policy episodes per setting for GR00T-N1.5 / QuantVLA, $\pi_{0.5}$ / QuantVLA, and GR00T-N1.5 / HoloQ-VLA, with 140 episodes per suite. For OpenVLA-OFT / QVLA, the verified seed-2026 diagnostic cohort covers 140 LIBERO-Spatial episodes. Its distributional results are therefore specific to Spatial.

Distribution fitting. We fit log-normal, Gamma, and Weibull distributions by maximum likelihood to positive, finite discrepancies, fixing the location parameter to zero. For the log-normal fit, $\hat{\mu}$ and $\hat{\sigma}$ are the mean and population standard deviation of $\log q_t$. Pooled Kolmogorov–Smirnov (KS) distance is computed after fitting all states jointly; median episode KS is obtained by fitting each episode separately and taking the median across episodes. The former weights states equally, whereas the latter summarizes episodes equally. We report KS as a descriptive distance, without standard KS-test p-values, because parameters are estimated from the observations and successive states are dependent.

Table 4: **Log-normal discrepancy fits.** N_e/N_s: episodes/positive states. KS: pooled/median episode distance. †Spatial only.

Setting	N_e/N_s	$\hat{\mu}$	$\hat{\sigma}$	KS↓
GR00T-N1.5 / QuantVLA	560/125,569	-5.630	2.079	0.0732/0.1033
$\pi_{0.5}$ / QuantVLA	560/128,739	-3.817	1.366	0.0163/0.0733
GR00T-N1.5 / HoloQ-VLA	560/115,135	-6.126	1.924	0.0720/0.0914
OpenVLA-OFT / QVLA†	140/14,942	-5.675	0.961	0.0243/0.0738

The log-normal family gives the lowest pooled and median episode KS distances among the three candidates in all analyzed cohorts (Table 5). We use it as a descriptive approximation; neither the selection rule nor the recovery objective requires the discrepancies to follow this distribution.

Table 5: **Alternative distribution fits.** Entries report pooled/median episode KS; lower is better. †Spatial only.

Setting	Log-normal	Gamma	Weibull
GR00T-N1.5 / QuantVLA	**0.0732/0.1033**	0.2078/0.2924	0.1176/0.1953
$\pi_{0.5}$ / QuantVLA	**0.0163/0.0733**	0.1285/0.1312	0.0785/0.1104
GR00T-N1.5 / HoloQ-VLA	**0.0720/0.0914**	0.2450/0.2895	0.1330/0.1935
OpenVLA-OFT / QVLA†	**0.0243/0.0738**	0.1146/0.1017	0.0993/0.1006

Empirical tail concentration. We rank all recorded states within a setting by decreasing discrepancy and measure the share of total discrepancy contributed by the largest 1%, 5%, 10%, and 20% of values. These statistics are computed directly from observations, independently of the fitted distribution. The largest 5% of values account for 72.3% and 78.5% of total discrepancy in the two GR00T-N1.5 cohorts, compared with 43.8% for $\pi_{0.5}$ and 34.6% for OpenVLA-OFT on Spatial. This concentration describes the discrepancy signal; the main-text interventions assess its behavioral relevance.

Table 6: **Discrepancy concentration.** Share of total discrepancy (%) contributed by the largest fraction of state discrepancies. †Spatial only.

Setting	Top 1%	Top 5%	Top 10%	Top 20%
GR00T-N1.5 / QuantVLA	24.6	72.3	84.4	93.2
$\pi_{0.5}$ / QuantVLA	16.0	43.8	58.5	73.5
GR00T-N1.5 / HoloQ-VLA	26.3	78.5	88.6	94.5
OpenVLA-OFT / QVLA†	18.3	34.6	45.4	60.1

C Association Between Current and Future Discrepancy

We compute within-trajectory Spearman correlations between the logged current discrepancy q_t and discounted future discrepancy r_t^+ using all valid recovery states. A trajectory of length T is partitioned into four approximately equal contiguous phases according to $\lfloor 4t/T \rfloor$, for zero-based timestep t. These are temporal partitions rather than annotated subtasks. We exclude the terminal state, whose future discrepancy is zero, and omit phase segments with fewer than five valid states or a constant signal.

Each of the first three settings contains 560 recovery trajectories. OpenVLA-OFT provides 558 trajectories: 140 each from Spatial, Object, and Goal, and 138 from Long. All recorded trajectories provide valid correlations in every phase. Table 7 reports the median and a 95% percentile interval from 10,000 trajectory bootstrap resamples stratified by suite. These are descriptive intervals within the recorded runs, not estimates of variability across independent recovery runs.

Table 7: **Current–future discrepancy association.** Median Spearman correlation with a 95% bootstrap interval underneath. N is the trajectory count per phase.

Setting	N	Phase 1	Phase 2	Phase 3	Phase 4
GR00T-N1.5 / QuantVLA	560	0.534 [0.509, 0.551]	0.477 [0.449, 0.498]	0.516 [0.498, 0.546]	0.563 [0.542, 0.589]
$\pi_{0.5}$ / QuantVLA	560	0.486 [0.455, 0.510]	0.468 [0.447, 0.488]	0.534 [0.511, 0.559]	0.467 [0.432, 0.504]
GR00T-N1.5 / HoloQ-VLA	560	0.409 [0.381, 0.426]	0.422 [0.407, 0.440]	0.401 [0.378, 0.424]	0.449 [0.428, 0.468]
OpenVLA-OFT / QVLA	558	0.537 [0.515, 0.553]	0.576 [0.554, 0.593]	0.621 [0.598, 0.644]	0.632 [0.605, 0.662]

Median correlations range from 0.401 to 0.632. The signals are positively associated but induce imperfectly aligned state rankings. Their incremental utility for recovery is assessed by the component ablations in Table 3.

D Temporal Coverage of Vulnerability-Based Selection

Figure 5 compares global and phase-wise vulnerability selection under the same 16-state budget. Global selection ranks candidates across the entire trajectory, while PIVOT-Q allocates four states to each temporal phase. Across the six illustrated rollouts, global selection occupies one phase in panels (a), (c), and (e), and two phases in panels (b), (d), and (f). PIVOT-Q preserves coverage of all four phases while prioritizing vulnerable states within each phase. These examples illustrate temporal concentration rather than estimate its prevalence; the Global Vulnerability ablation evaluates its effect on task success.

E Additional Recovery Analysis

E.1 Discrepancy Distributions Across Recorded Rollouts

We compare episode-balanced discrepancy histograms using 40 common logarithmic bins within each setting. Each episode histogram is normalized before averaging, so long trajectories do not

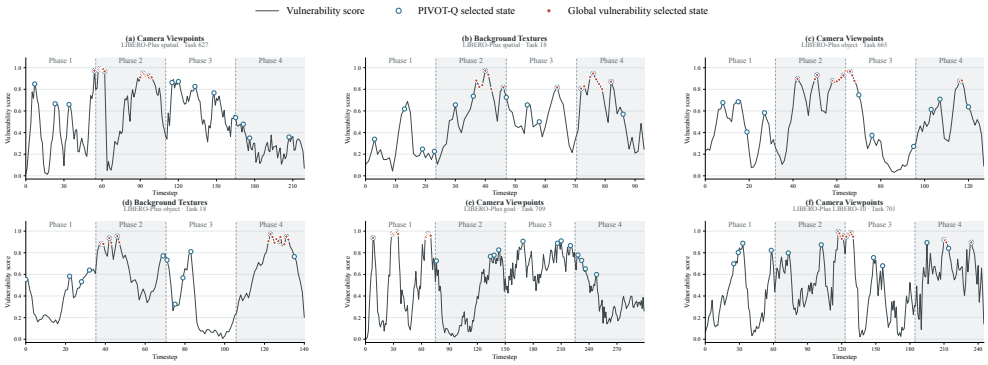

Figure 5: **Temporal coverage of sparse selection.** Six LIBERO-Plus rollouts span camera-viewpoint and background-texture variations across Spatial, Object, Goal, and Long. Red diamonds denote global vulnerability selection; blue circles denote PIVOT-Q. Both select 16 states. Global selection occupies one or two phases in these examples, while PIVOT-Q selects four states per phase.

dominate. The GR00T-N1.5 / QuantVLA comparison uses an archived paired rollout cohort. For $\pi_{0.5}$ / QuantVLA and GR00T-N1.5 / HoloQ-VLA, frozen-quantized diagnostic rollouts are compared with recovery-time rollouts, each covering 560 episodes. OpenVLA-OFT uses 140 paired final-policy diagnostic episodes on Spatial. Recovery-time distributions mix stages of adaptation and do not describe a single final checkpoint; all comparisons also reflect policy-dependent state visitation.

We compute $D_{\mathrm{KL}}(P_Q \| P_R)$ in nats and $D_{\mathrm{JS}}(P_Q, P_R)$ in bits, where P_Q and P_R denote the quantized and recovery discrepancy distributions, respectively. We add 10^{-9} to each bin probability before renormalization. Intervals use 10,000 paired episode bootstrap resamples stratified by suite, with fixed bin edges and pairs aligned by suite and task identity. They describe variation within the recorded cohorts. The smaller divergence for $\pi_{0.5}$ indicates a smaller distributional change in this cohort, but divergence alone specifies neither the direction of change nor its effect on task success.

Table 8: **Recorded discrepancy distribution changes.** KL/JS divergence with 95% bootstrap intervals. N is episodes per method. †Final-policy Spatial evaluations; other rows use archived or recovery-time cohorts.

Setting	N	KL (nats)	JS (bits)
GR00T-N1.5 / QuantVLA	560	0.0997 [0.0818, 0.1333]	0.0287 [0.0238, 0.0357]
$\pi_{0.5}$ / QuantVLA	560	0.0042 [0.0035, 0.0085]	0.0015 [0.0012, 0.0030]
GR00T-N1.5 / HoloQ-VLA	560	0.0567 [0.0468, 0.0712]	0.0187 [0.0155, 0.0229]
OpenVLA-OFT / QVLA†	140	0.0571 [0.0468, 0.0789]	0.0208 [0.0174, 0.0280]

E.2 Selection Diagnostics and Final Outcomes

Figure 6 relates selection diagnostics to final evaluation outcomes. We match the quantized and PIVOT-Q seed-2026 evaluations by suite, task, and initial-state identity, obtaining 560 pairs per setting. Episodes are grouped as unrecovered (failure to failure), recovered (failure to success), regressed (success to failure), or retained (success to success). Each point shows the mean discrepancy over selected states, with color indicating their mean vulnerability score. The selection statistics come from recovery logs for the GR00T-N1.5 and OpenVLA-OFT settings, and from a separate frozen-quantized selection pass for $\pi_{0.5}$. They are associated with final outcomes by task identity, rather than measured on the same final-evaluation trajectory.

For OpenVLA-OFT, we reconstruct the selection score from the logged current and future discrepancies using the implemented within-phase percentile ranks and equal mixture weights. Final outcomes are available for all 560 episodes, while selection records cover 558. Consequently, panel labels use all 560 outcome pairs, and scatter points and medians use only episodes with selection records.

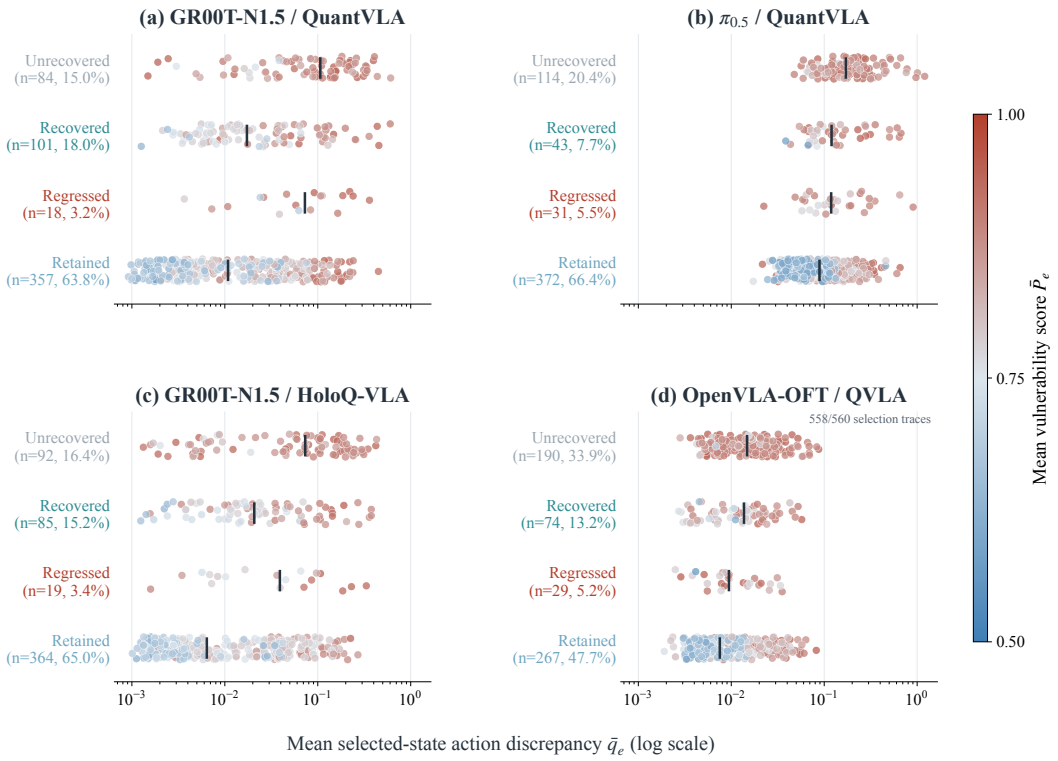

Figure 6: **Selection diagnostics and paired recovery outcomes.** The four panels correspond to GR00T-N1.5 / QuantVLA, $\pi_{0.5}$ / QuantVLA, GR00T-N1.5 / HoloQ-VLA, and OpenVLA-OFT / QVLA. Points show mean selected-state discrepancy on a logarithmic axis; color represents mean vulnerability score using a shared continuous scale from 0.5 to 1.0, and black ticks mark group medians. Counts and percentages use all 560 final-evaluation pairs per setting. OpenVLA-OFT has 558 plotted selection records.

Recovered episodes outnumber regressed episodes in all four settings: 101 versus 18 for GR00T-N1.5 / QuantVLA, 43 versus 31 for $\pi_{0.5}$ / QuantVLA, 85 versus 19 for GR00T-N1.5 / HoloQ-VLA, and 74 versus 29 for OpenVLA-OFT / QVLA. These yield net gains of 14.8, 2.1, 11.8, and 8.0 percentage points, respectively. The selection distributions overlap across outcome groups, so their association with final outcomes should not be interpreted as a causal or deterministic prediction of recovery.

F Sensitivity to Vulnerability Score Weighting

We study the sensitivity of PIVOT-Q to the relative contributions of current and discounted future discrepancy in the vulnerability score. For a state at time t, the score is defined as

$$V_t = \alpha C_t + \beta F_t, \qquad \alpha + \beta = 1, \tag{8}$$

where C_t and F_t denote the normalized current and discounted future discrepancy components used by the selection rule, respectively.

All experiments use GR00T-N1.5 with W4A8 QuantVLA, the same 560-episode LIBERO-Plus evaluation protocol, a fixed 16-state supervision budget, and the same training configuration. The default

$(\alpha, \beta) = (0.5, 0.5)$ configuration corresponds to the canonical PIVOT-Q run. The remaining coefficient settings are single-run sensitivity experiments; they are intended to characterize the effect of score weighting rather than to establish statistically significant differences between coefficients.

Table 9: **Sensitivity to vulnerability-score weighting.** Single-seed task success rate (%) on 560 LIBERO-Plus episodes. All variants use 16 supervision states. Bold and underline denote the highest and second-highest distinct values in each column, including ties.

GR00T-N1.5 / QuantVLA (W4A8)

(α, β)	Cam.	Init.	Lang.	Light	Back.	Noise	Layout	Avg.↑
$(0.3, 0.7)$	73.8	**40.0**	**81.2**	<u>98.8</u>	**100.0**	82.5	<u>88.8</u>	<u>80.7</u>
$(0.4, 0.6)$	<u>78.8</u>	27.5	76.2	**100.0**	**100.0**	<u>86.2</u>	85.0	79.1
$(0.5, 0.5)$	**80.0**	<u>37.5</u>	77.5	<u>98.8</u>	**100.0**	**88.8**	**90.0**	**81.8**
$(0.7, 0.3)$	76.2	30.0	<u>80.0</u>	**100.0**	<u>97.5</u>	82.5	85.0	78.8

The default balanced mixture obtains the highest average success rate among the tested configurations, at 81.8%. The future-weighted $(0.3, 0.7)$ mixture achieves 80.7%, while $(0.4, 0.6)$ and the current-weighted $(0.7, 0.3)$ mixture achieve 79.1% and 78.8%, respectively. Performance therefore does not vary monotonically with either coefficient in this sweep. These results support the balanced mixture as a reasonable default for the evaluated setting. Since the coefficient sweep is based on single-run evaluations, we treat it as a sensitivity analysis rather than evidence for a uniquely optimal weighting.