Beyond Appearance Shifts: Task-Semantic Action Calibration for VLA Models

Shuaijun Liu[1,2] **Feiyang You**[1] **Chengyu Wu**[1] **Shuyang Hao**[1]
Chenglong Zhang[1] **Jingyao Cai**[3] **Xingwei Chen**[4,5] **Li Sun**[6] **Ningxin Su**[1]*
[1]HKUST (Guangzhou) [2]MUKA Robotics
[3]NCCA, Bournemouth University [4]SJTU [5]EIT, Ningbo [6]BUPT
ningxinsu@hkust-gz.edu.cn Project Page: https://nebulis-lab.com/BAS-VLA

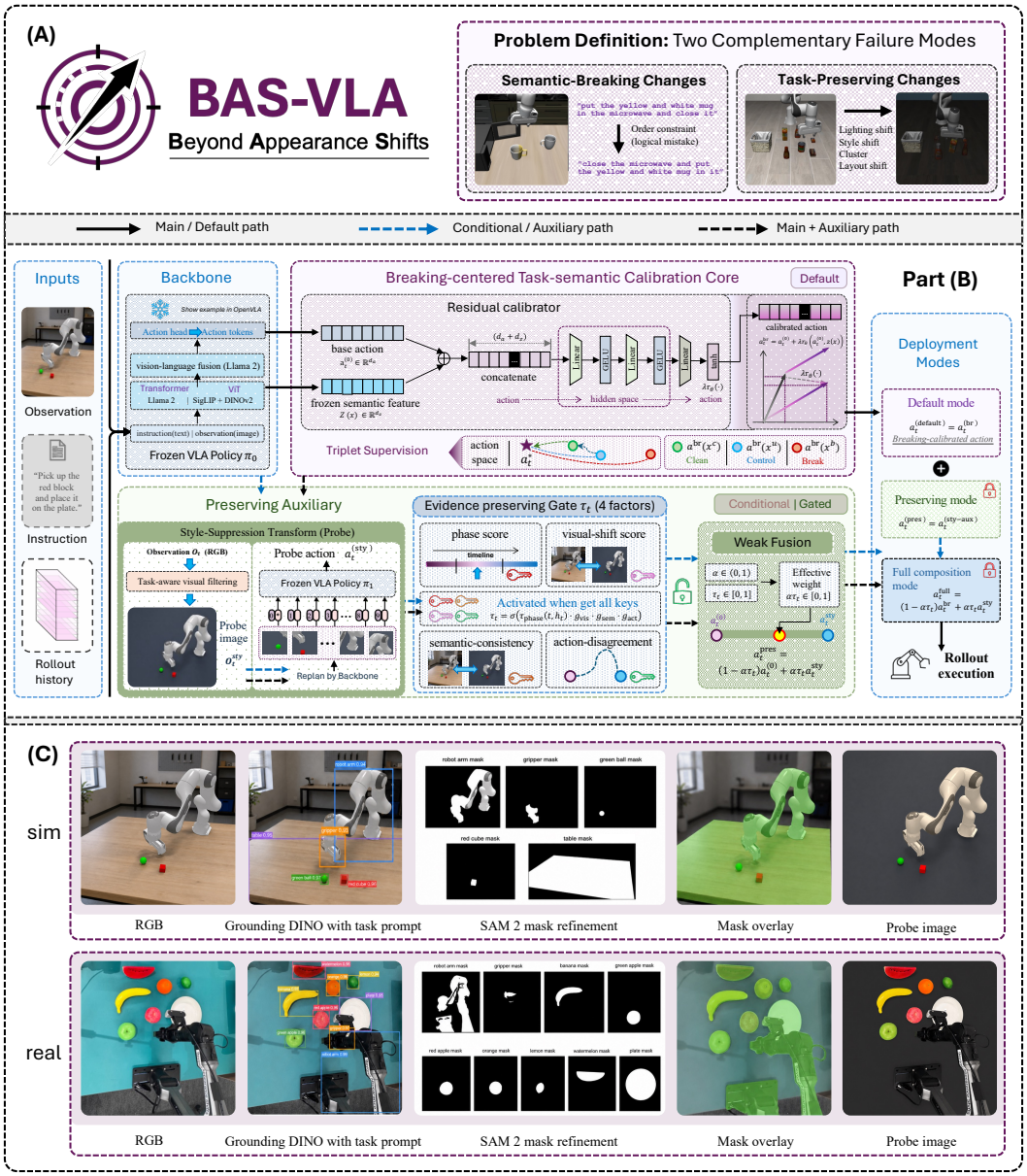


Figure 1: BAS-VLA Overview. Part (A) summarizes the two intervention types studied in this work: task-preserving changes and semantic-breaking changes. Part (B) shows BAS-VLA as an action-calibration layer over a frozen base VLA. Part (C) illustrates the task-aware visual filtering probe, from the input RGB image to text-conditioned grounding, SAM 2 mask refinement, merged task-relevant masks, and the resulting probe image $\tilde{o}_t^{\text{sty}}$.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Abstract

Vision-language-action (VLA) models have achieved strong performance in embodied manipulation, but still struggle to distinguish between changes that should preserve behavior and those that require different actions. We identify two complementary failure modes. Under task-preserving changes, where task semantics remain unchanged but scene appearance varies (e.g., style, illumination, clutter, or paraphrasing), policies often exhibit unnecessary action drift. Conversely, under semantic-breaking changes, where key task semantics such as the target object or constraint are altered, policies frequently continue following the original trajectory instead of adapting to the new task. To address these failures, we propose BAS-VLA, a task-semantic action calibration framework built on top of a frozen base VLA. BAS-VLA adopts a breaking-centered calibration core as its default path, complemented by a selective evidence-gated preserving auxiliary that corrects appearance-induced drift when task-preserving evidence is strong. Across four target-object swap tasks, BAS-VLA improves changed-task success from 2.3% to 68.5% while maintaining strong clean and semantics-preserving performance and suppressing outdated behavior. On validated style-preserving shifts, it further improves success from 42% to 70% without degrading clean performance. These results demonstrate the importance of calibrating VLA policies to remain stable when task semantics are preserved and adapt when task semantics change.

1 Introduction

Vision-language-action (VLA) models have substantially broadened the scope of embodied manipulation by coupling language conditioning with large visuomotor policies [Brohan et al., 2023, Kim et al., 2024]. Yet stronger backbones do not automatically yield *reliable* behavior. A policy may solve a canonical instruction in a canonical scene, but still fail to preserve the right action when semantics are unchanged, or fail to revise the action when semantics genuinely change. This reliability gap is the focus of this work.

We study this gap through two intervention types. *Task-preserving changes* alter style, illumination, clutter, viewpoint, or wording while leaving task semantics unchanged, so the desired behavior should remain in the same action basin. *Semantic-breaking changes* modify a key semantic slot or constraint, so the behavior should separate decisively from the clean trajectory. Current VLAs can fail in both directions: they drift when they should stay stable, and they remain overly similar when they should change.

This problem touches several neighboring lines of work. Embodied grounding and affordance-aware control have improved how instructions connect to objects, skills, and constraints [Ahn et al., 2022, Mees et al., 2023, Huang et al., 2024], while counterfactual reasoning, semantic-intervention benchmarks, and process supervision have emphasized that correct outputs alone are insufficient when meaning changes [Hüyük et al., 2025, Pona et al., 2025, Chen et al., 2025, Lightman et al., 2024, Setlur et al., 2025, Li and Li, 2025]. Our contribution is an explicit, asymmetric action-calibration objective

Table 1: Motivation evidence for reliability gaps before BAS-VLA. *Original* denotes native carrier performance on the unmodified task condition, and *Perturbed* denotes performance under the corresponding preserving or task-changed intervention. Carrier and suite names follow OpenPI-pi0.5 [Black et al., 2025a], OpenVLA-OFT [Kim et al., 2025], and LIBERO [Liu et al., 2023]. Here OpenVLA-OFT denotes an OpenVLA carrier built from the OFT fine-tuning recipe of Kim et al. [2025]; in our study, we adopt that recipe and perform our own fine-tuning for the simulation and real-robot settings used in the paper.

Carrier	**Suite**	**Perturbation**	**Original success** ↑	**Perturbed success** ↑	Δ (pts)
OpenVLA-OFT	LIBERO-Spatial	Perimeter clutter	98/100 (98.0%)	65/100 (65.0%)	−33.0
OpenVLA-OFT	LIBERO-Spatial	Margin noise	97/100 (97.0%)	51/100 (51.0%)	−46.0
OpenPI-pi0.5	LIBERO-Object	Illum. change	98/100 (98.0%)	89/100 (89.0%)	−9.0
OpenPI-pi0.5	LIBERO-Object	Target-object Swap	197/200 (98.5%)	68/200 (34.0%)	−64.5

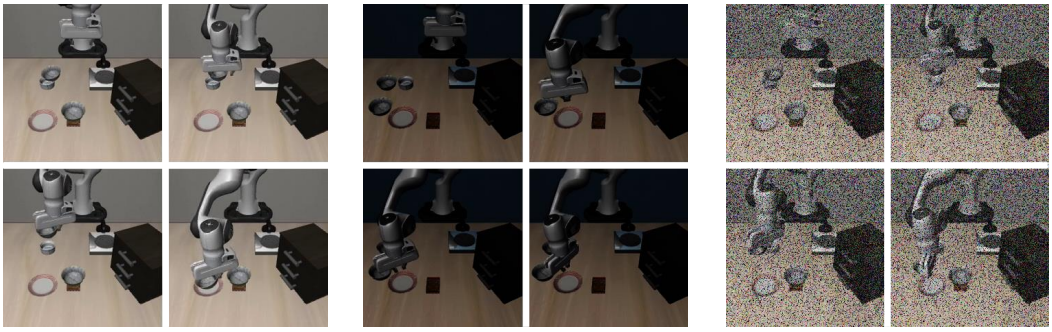


Figure 2: Task-preserving motivation on a single Black-Bowl-to-Plate instruction. The three panels show a clean success, a failure under a darkened illumination shift, and a failure under a semantics-preserving noise perturbation.

for a frozen VLA: which edits should leave behavior invariant, which should force separation, and how should a shared policy be calibrated under both? This complements robustness benchmarks and backbone adaptation by explicitly calibrating action responses to semantics-preserving and semantics-changing interventions.

To address this gap, we propose **BAS-VLA**, a task-semantic action calibration framework built on top of a frozen base VLA. BAS-VLA does not replace the backbone or retrain a new foundation policy; it attaches a lightweight calibration layer to the action output of the shared VLA. Its default deployment path is a *breaking-centered* calibration core that keeps semantically equivalent variants close while pushing genuine semantic breaks away from the clean action basin. In addition, BAS-VLA includes a selective *evidence-gated preserving auxiliary* that activates only when nuisance evidence is strong while task semantics remain consistent, so that a weak stabilizing correction is beneficial.

Figure 2 shows the qualitative preserving-side failure pattern, and Table 1 quantifies the corresponding pre-method gap before BAS-VLA calibration. We evaluate BAS-VLA under a unified intervention protocol with matched backbones, task families, and rollout budgets. On the primary OpenPI-pi0.5 / LIBERO-Object Milk-Swap evaluation, BAS-VLA maintains 98.0% clean success and 97.5% control success while reducing clean-task success under the break condition to 0.0%. On the Bowl-on-Ramekin style-shift setting, it raises changed-condition success from 42.0% to 70.0% without degrading the matched-clean reference row. In summary, we formulate VLA reliability as task-semantic action calibration, introduce BAS-VLA as a breaking-centered core with a selective evidence-gated preserving auxiliary, and provide matched-carrier evidence that reliable VLA behavior requires explicit task-semantic calibration rather than generic appearance robustness alone.

2 Related Work

Large-scale VLA systems provide the substrate on which BAS-VLA operates, but they are not the main intellectual neighbor. Generalist robot policies and multimodal foundation models have shown that broad pretraining improves language-conditioned manipulation across tasks and scenes [Driess et al., 2023, Brohan et al., 2023, Octo Model Team et al., 2024, Kim et al., 2024]. The question here is narrower: once a shared VLA is fixed, how should its action behavior respond to controlled semantic interventions? The closest embodied line concerns language grounding, semantic control, and intervention-aware adaptation. SayCan couples language with affordance-weighted skill selection [Ahn et al., 2022], while later approaches ground language through visual affordances or explicit relational constraints [Mees et al., 2023, Huang et al., 2024]. Counterfactual feedback, semantic-intervention benchmarks, prompt-miscalibration analyses, and inference-time VLA adaptation methods all reinforce the same high-level point: correctness alone is too weak without sensitivity to meaning-preserving versus meaning-changing alternatives [Hüyük et al., 2025, Pona et al., 2025, Chen et al., 2025, Cox et al., 2025, Black et al., 2025b, So et al., 2025, Liang et al., 2026, Ye et al., 2026, Wu et al., 2026, Koo et al., 2026, Jang et al., 2026, Lightman et al., 2024, Setlur et al., 2025, Li and Li, 2025]. BAS-VLA instead targets embodied *action* calibration under matched manipulator rollouts. Appendix A.2.4 gives a compact extended positioning note. LIBERO-Plus and LIBERO-PRO evaluate robustness and generalization under visual, language, object, and task changes [Fei et al., 2025, Zhou et al., 2025]. Evaluation on these external protocols further tests BAS-VLA beyond the intervention settings used in our primary experiments.

Table 2: Intervention taxonomy. Preserving interventions leave task semantics unchanged, whereas semantic-breaking interventions modify action-relevant task semantics. Beige rows provide prompt-edit examples for families that admit an instruction-level realization.

Type	Family	Representative intervention
Preserving	illumination change scene	darkened scene / lighting shift
Preserving	style shift scene	style-transferred scene variant
Preserving	clutter scene	perimeter clutter around the workspace
Preserving	noise scene	margin noise band / hard noise band
Preserving	layout / viewpoint scene	layout shift / view jitter
Preserving	wording control prompt	paraphrase / redundant wording
	prompt example	clean: "put the yellow mug in the microwave" edited: "place the yellow mug into the microwave"
Semantic-breaking	target-object swap mixed	milk ↔ orange juice
	prompt example	clean: "put the milk in the basket" edited: "put the orange juice in the basket"
Semantic-breaking	destination change mixed	back/front compartment swap
	prompt example	clean: "put the book in the back compartment of the caddy" edited: "put the book in the front compartment of the caddy"
Semantic-breaking	spatial-relation change mixed	left-of / behind placement change
	prompt example	clean: "place the black bowl to the left of the plate" edited: "place the black bowl behind the plate"
Semantic-breaking	order change prompt	subgoal order swap
	prompt example	clean: "put the mug in the microwave, then close the door" edited: "close the door, then put the mug in the microwave"
Semantic-breaking	executability change mixed	open/closed state change
	prompt example	clean: "put the bowl in the open drawer" edited: "put the bowl in the closed drawer"

Badge legend. scene indicates an intervention that primarily changes the scene or observation; prompt indicates an instruction-side edit; mixed indicates settings that commonly involve both scene-side and prompt-side changes, while still defining a task-level semantic change rather than a mere rewording.

3 Problem Setup

We study reliability of a frozen vision-language-action policy under *controlled task interventions*. The central question is whether *action behavior* changes in the right way when the task description or scene is edited: preserving edits should leave behavior largely unchanged, whereas semantic-breaking edits should induce a clear behavioral split. Let a rollout state at time t be described by an observation o_t, a natural-language instruction x, and rollout context h_t. A frozen base policy $\pi_0 : (o_t, x, h_t) \mapsto a_t^{(0)}$ maps this tuple to an action or action chunk $a_t^{(0)} \in \mathbb{R}^{d_a}$. We write $\mathcal{T}(o, x)$ for the latent task semantics induced by scene and instruction.

A preserving intervention $g \in \mathcal{G}_{\text{pres}}$ produces $(o', x') = g(o, x)$ with $\mathcal{T}(o', x') = \mathcal{T}(o, x)$; examples include illumination, style, clutter, noise, viewpoint, and semantically equivalent rephrasings. A semantic-breaking intervention $b \in \mathcal{G}_{\text{break}}$ produces $(o', x') = b(o, x)$ with $\mathcal{T}(o', x') \neq \mathcal{T}(o, x)$; examples include target-object swaps, destination changes, order changes, and executability changes.

Table 2 summarizes the intervention families emphasized here. Some define protocol scope, whereas others receive dedicated quantitative evaluation in the main text or appendix. Figure 3 complements this taxonomy with representative simulation and real-deployment views of the same preserving and breaking families.

To distinguish semantic change from harmless variation, we use a three-condition protocol built from a fixed scene and a matched instruction triplet (x^c, x^u, x^b), where x^c is the clean instruction, x^u is a semantics-preserving control variant, and x^b is a semantic-breaking variant. For a fixed scene o, we require $\mathcal{T}(o, x^c) = \mathcal{T}(o, x^u)$ and $\mathcal{T}(o, x^b) \neq \mathcal{T}(o, x^c)$. We refer to the resulting conditions as *clean*, *control*, and *break*. Throughout the paper, the *Break* column reports clean-task success under the break condition, so lower values indicate stronger stale-task suppression and semantic separation rather than direct changed-task success.

The behavioral requirements are intentionally asymmetric: preserving interventions should leave clean behavior and task success largely intact, whereas semantic-breaking interventions should move the policy away from the clean solution. The primary comparison object is therefore a matched task



Figure 3: Representative intervention families in simulation and real deployment. Top: simulation. Bottom: real deployment. Columns show no perturbation, illumination change, style shift, noise, clutter, and spatial-relation change.

triplet under a fixed carrier and rollout interface. Task outcome is the main metric, while action-level diagnostics serve only as supporting evidence for whether break conditions leave the clean basin more decisively than preserving controls do. Appendix A.2.1 summarizes the benchmark-role assignments, rollout budgets, and matched-comparison rules used in the empirical sections.

4 Method

We instantiate BAS-VLA as a lightweight action-calibration layer on top of a frozen vision-language-action policy. It operates as a lightweight action-calibration layer while leaving the underlying VLA unchanged. It acts at the final action level and asks whether a fixed policy substrate can be selectively recalibrated so that it remains stable under task-preserving changes and separates under genuine task-semantic changes.

Our empirical evidence supports a *selective* design rather than a symmetric two-branch architecture that is always active. The deployed form of BAS-VLA therefore contains two components with different roles: a *breaking-centered calibration core*, which serves as the default method, and a *selective evidence-gated preserving auxiliary*, which is activated only when the rollout remains in an early corrective phase, nuisance evidence is strong, and task semantics remain consistent. The full composition is retained for ablation and mechanism analysis, but it is not treated as the default best-performing variant. Figure 1(B) summarizes this action-calibration view, and Figure 1(C) previews the task-aware visual filtering probe used by the preserving auxiliary.

4.1 Overall Formulation

Let the frozen base policy be $\pi_0 : (o_t, x, h_t) \mapsto a_t^{(0)}$, where o_t is the observation at time t, x is the instruction, h_t is rollout context, and $a_t^{(0)} \in \mathbb{R}^{d_a}$ is the predicted action or action chunk. BAS-VLA outputs a calibrated action a_t^{BAS} by composing one or two lightweight modules on top of π_0:

$$a_t^{\mathrm{BAS}} = \mathcal{P}_\phi\Big(\mathcal{B}_\theta(a_t^{(0)}, o_t, x, h_t),\ o_t, x, h_t\Big)\,. \tag{1}$$

Here $\mathcal{B}_\theta$ denotes the breaking-centered core and $\mathcal{P}_\phi$ denotes the preserving auxiliary. In the default deployment, only $\mathcal{B}_\theta$ is enabled. The preserving sidecar is not assigned by perturbation family; it activates only when the evidence gate in Eq. (5) is positive. In preserving-only deployment, the auxiliary fuses a probe action with the frozen base action; in the full composition retained for ablation, the same evidence-gated weak-fusion rule is applied on top of the breaking-centered action. This parameterization reflects the design objective: *selective behavioral calibration*. Preserving interventions should induce small action changes, whereas semantic-breaking interventions should induce large ones.

4.2 Breaking-Centered Task-Semantic Calibration

The breaking-centered core is the default BAS-VLA component. Its purpose is to preserve clean competence while forcing semantically changed instructions to leave the clean/control action basin. We first compute the base action $a_t^{(0)} = \pi_0(o_t, x, h_t)$ and a frozen instruction feature $z(x) \in \mathbb{R}^{d_z}$.

We then define a residual calibrator $r_\theta : \mathbb{R}^{d_a} \times \mathbb{R}^{d_z} \to \mathbb{R}^{d_a}$ and construct

$$a_t^{\mathrm{br}} = a_t^{(0)} + \lambda\, r_\theta\Big(a_t^{(0)}, z(x)\Big), \tag{2}$$

where $\lambda > 0$ controls the residual scale. This residual form preserves the frozen base policy as the clean-task reference and keeps BAS-VLA portable across carriers because the method modifies action output rather than replacing the whole policy stack.

Training is built from matched instruction triplets (x^c, x^u, x^b), where x^c is the clean instruction, x^u is a semantics-preserving control variant, and x^b is a semantic-breaking variant. The objective enforces three properties simultaneously: clean and control should remain close to the clean reference action, clean and control should remain close to one another, and the breaking action should move farther away from the clean basin than the control action does. Let $a_t^\star$ denote the clean action. We write

$$\begin{aligned}
\mathcal{L}_{\mathrm{imit}} &= \|a^{\mathrm{br}}(x^c) - a_t^\star\|_2^2 + \|a^{\mathrm{br}}(x^u) - a_t^\star\|_2^2, \\
\mathcal{L}_{\mathrm{cons}} &= \|a^{\mathrm{br}}(x^c) - a^{\mathrm{br}}(x^u)\|_2^2, \\
\mathcal{L}_{\mathrm{sep}} &= \max\Big(0,\ m - \|a^{\mathrm{br}}(x^b) - a^{\mathrm{br}}(x^c)\|_2 + \|a^{\mathrm{br}}(x^u) - a^{\mathrm{br}}(x^c)\|_2\Big),
\end{aligned} \tag{3}$$

where $m > 0$ is a separation margin. The final breaking-core objective is

$$\mathcal{L}_{\mathrm{br}} = \mathcal{L}_{\mathrm{imit}} + \lambda_c \mathcal{L}_{\mathrm{cons}} + \lambda_s \mathcal{L}_{\mathrm{sep}}. \tag{4}$$

This objective encodes the desired asymmetry directly: preserving rephrasings are contracted toward the clean basin, whereas breaking edits are pushed away from it. During training, the base VLA π_0 remains frozen and only r_θ is optimized. At inference time, the breaking core adds one frozen instruction feature and one small residual forward pass. Appendix C.2 states the direct margin consequence of $\mathcal{L}_{\mathrm{sep}}$, and Appendix C.3 records the corresponding local drift bound induced by the residual form.

4.3 Evidence-Gated Preserving Auxiliary

The preserving side of BAS-VLA is deliberately selective, providing a lightweight correction when nuisance-induced action drift is more plausible than a genuine task change. Given the current observation o_t, we construct a probe view $\tilde{o}_t^{\mathrm{sty}} = \phi_{\mathrm{sty}}(o_t)$, where the subscript "sty" is mnemonic for observation-side visual variation in the input image rather than only the style-shift slice used in one validated benchmark. The transform ϕ_{sty} is a deterministic, non-generative, geometry-preserving appearance-canonicalization map. It suppresses low-level cues such as color cast, illumination bias, and texture-heavy surface variation while preserving object identity, layout, boundaries, and task-relevant geometry. A concrete grounded-mask instantiation of ϕ_{sty}, based on text-conditioned grounding and mask refinement [Liu et al., 2024, Ravi et al., 2024, He et al., 2010], is deferred to Appendix A.2. We then reuse frozen visual features from the carrier's own visual pathway rather than introducing a separate trainable vision module. Let

$$v_t^{\mathrm{mid}} = f_{\mathrm{vis}}^{\mathrm{mid}}(o_t), \qquad \tilde{v}_t^{\mathrm{mid}} = f_{\mathrm{vis}}^{\mathrm{mid}}(\tilde{o}_t^{\mathrm{sty}})$$

denote a frozen mid-level visual representation, and let

$$v_t^{\mathrm{late}} = f_{\mathrm{vis}}^{\mathrm{late}}(o_t), \qquad \tilde{v}_t^{\mathrm{late}} = f_{\mathrm{vis}}^{\mathrm{late}}(\tilde{o}_t^{\mathrm{sty}})$$

denote a frozen later semantic visual representation from the same backbone. The frozen base policy is then queried on the probe:

$$a_t^{\mathrm{sty}} = \pi_0(\tilde{o}_t^{\mathrm{sty}}, x, h_t).$$

Instead of a hard nuisance label, the preserving auxiliary uses an evidence gate

$$\tau_t = \tau_{\mathrm{phase}}(t, h_t)\, g_{\mathrm{vis}}(v_t^{\mathrm{mid}}, \tilde{v}_t^{\mathrm{mid}})\, g_{\mathrm{sem}}(v_t^{\mathrm{late}}, \tilde{v}_t^{\mathrm{late}}, x)\, g_{\mathrm{act}}(a_t^{(0)}, a_t^{\mathrm{sty}}), \tag{5}$$

where $\tau_{\mathrm{phase}} \in [0,1]$ downweights late-stage intervention, $g_{\mathrm{vis}} \in [0,1]$ measures the strength of the visual shift through frozen mid-level feature discrepancy, $g_{\mathrm{sem}} \in [0,1]$ measures semantic consistency through frozen later visual features, and $g_{\mathrm{act}} \in [0,1]$ measures whether the visual change has already perturbed the base action. Concrete score parameterizations are given in Appendix B.3. This makes the auxiliary selective in both time and evidence: it should activate only when there is substantial surface variation at the frozen visual-feature level, preserved higher-level semantics, and action-level

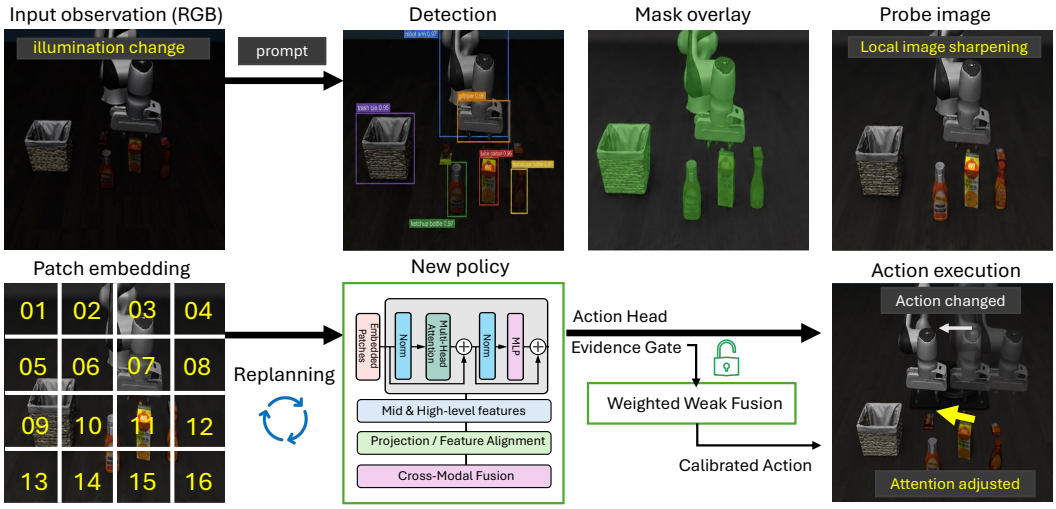


Figure 4: Dark-scene preserving example. The frozen carrier misreads the target under illumination change; the evidence-gated probe path activates and produces a corrected action.

Table 3: Main task-preserving style-shift results on Bowl-on-Ramekin. Each method uses 200 episodes over four seeds; Δ_{style} is measured relative to the baseline style-shift row.

Method	**Clean success** ↑	**Matched-clean subset success** ↑	**Style-shift success** ↑	Δ_{style} ↑
Baseline	194/200 (97.0%)	108/200 (54.0%)	84/200 (42.0%)	—
BAS-VLA	195/200 (97.5%)	108/200 (54.0%)	140/200 (70.0%)	+28.0 pts

disagreement that is still early enough to correct. The auxiliary then applies weak fusion between the base action and the visual-probe action:

$$a_t^{\text{pres}} = (1 - \alpha\tau_t)a_t^{(0)} + \alpha\tau_t a_t^{\text{sty}}, \tag{6}$$

with a small fusion weight $\alpha \in (0, 1)$. When $\tau_t = 0$, this reduces exactly to $a_t^{(0)}$. The choice of weak fusion rather than hard override keeps the frozen base action dominant while allowing a bounded correction when nuisance evidence is present. Figure 4 shows one dark-scene instance in which the probe path is activated after a target-reading error under illumination change. Appendix C.4 makes the resulting gate and drift identities explicit.

4.4 Deployment Modes

This work uses three deployment modes. The default method uses only the breaking-centered core, $a_t^{\text{default}} = a_t^{\text{br}}$, and this is the main variant referred to as BAS-VLA in the semantic-breaking results. For preserving-side stabilization, we deploy the auxiliary through the evidence-gated action

$$a_t^{\text{pres}} = (1 - \alpha\tau_t)a_t^{(0)} + \alpha\tau_t a_t^{\text{sty}}.$$

For ablation and mechanism analysis, we also retain the full composition

$$a_t^{\text{full}} = \mathcal{P}_\phi(a_t^{\text{br}}, o_t, x, h_t) = (1 - \alpha\tau_t)a_t^{\text{br}} + \alpha\tau_t a_t^{\text{sty}}.$$

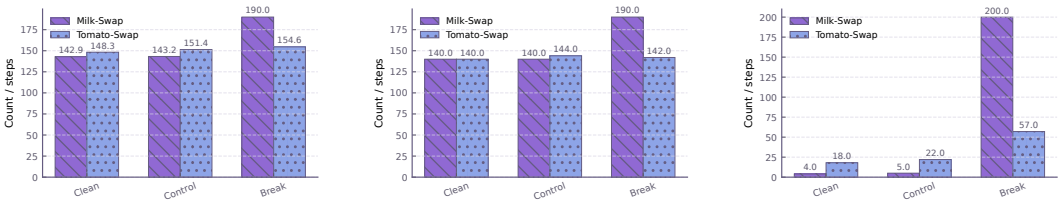
Full receives no perturbation-family label and provides a unified composition for mixed conditions. The selective/default modes provide the strongest setting-specific results in our evaluations. Inference-time cost relations, control-flow details, and scope limits are deferred to Appendix A.2.2, Appendix C.1, and Appendix C.5.

5 Experiments

We evaluate BAS-VLA along three axes: semantic-breaking reliability, task-preserving stability, and matched-carrier inference-time strategy comparisons. We then use harder tasks and supplementary benchmarks to evaluate transfer.

Table 4: Main semantic-breaking results on matched target-object triplets. Each row uses 200 clean, 200 control, and 200 break episodes; lower is better in *Break*.

Task	Setting	Clean success ↑	Control success ↑	Break (clean) ↓	Ctrl.-Break gap ↑
Canonical harder target-object tasks					
BBQ-Swap	BBQ sauce→ketchup	180/200 (90.0%)	185/200 (92.5%)	0/200 (0.0%)	92.5
Ketchup-Swap	ketchup→tomato sauce	176/200 (88.0%)	184/200 (92.0%)	0/200 (0.0%)	92.0
Milk-Swap	milk→orange juice	196/200 (98.0%)	195/200 (97.5%)	0/200 (0.0%)	97.5
Butter-Swap	butter→cream cheese	188/200 (94.0%)	182/200 (91.0%)	113/200 (56.5%)	34.5
OJ-Swap	orange juice→milk	197/200 (98.5%)	193/200 (96.5%)	0/200 (0.0%)	96.5
Stronger interference in the orange-juice→milk setting					
OJ-Swap	Hard noise	184/200 (92.0%)	185/200 (92.5%)	0/200 (0.0%)	92.5
OJ-Swap	Clutter+Noise	192/200 (96.0%)	183/200 (91.5%)	0/200 (0.0%)	91.5



(a) Mean steps. (b) Median steps. (c) 190-step saturation.

Figure 5: Process-side view of the two main semantic-breaking tasks. Milk-Swap keeps clean/control short and unsaturated while break saturates; Tomato-Swap shows a weaker version of same pattern.

5.1 Experimental Setup

Benchmarks, metrics, and scale. The primary evaluation uses LIBERO-Object for semantic breaking, LIBERO-Spatial for task preserving, and LIBERO-10 for corroboration [Liu et al., 2023]; CALVIN, MetaWorld, ManiSkill2, and RoboCasa appear only as supplementary transfer and boundary settings [Mees et al., 2022, Yu et al., 2020, Gu et al., 2023, Nasiriany et al., 2024]. The main evidence comes from two carriers: OpenPI-pi0.5 for semantic-breaking and matched strategy comparisons [Black et al., 2025a], and OpenVLA-OFT for task-preserving main results and OFT corroboration [Kim et al., 2025]. In the latter case, we follow the OFT fine-tuning work of Kim et al. [2025] as the carrier reference and use the same recipe as the basis for our own fine-tuning in the simulation and real-robot settings considered here. Preserving evaluations apply semantics-preserving visual or wording perturbations, whereas the main semantic-breaking evaluations use matched clean/control/break triplets whose *Break* column reports clean-task success under the changed task condition; Table 1 instead reports native pre-method original-versus-perturbed performance.

Task success rate is the primary metric. In preserving settings, the goal is to maintain success under nuisance variation without harming clean-task performance; in semantic-breaking settings, the goal is to keep clean/control success high while adapting to the changed instruction. In the simulation semantic-breaking evaluations, the Break metric reports old-task success under the changed instruction; lower values therefore indicate stronger suppression of stale behavior. Changed-task completion is evaluated separately using mutually exclusive New, Old, and Other/timeout outcomes. Process statistics such as episode length, saturation, and replanning are reported only as supporting evidence. The main preserving campaign evaluates ten LIBERO-Spatial tasks with fifty episodes per task, yielding 500 episodes per variant-condition pair; the main semantic-breaking campaign uses four seeds with fifty episodes per seed, yielding 200 episodes per condition. Appendix A.2.1 summarizes the benchmark-role assignments, triplet interpretation, rollout budgets, and the real-robot deployment setup. Wilson 95% intervals for the principal binary outcomes are reported in Table 16.

5.2 Main Results

We begin with the task-preserving style-shift setting. Table 3 compares BAS-VLA with the uncalibrated baseline on Bowl-on-Ramekin under style shift. Style-shift success increases from 42.0% to 70.0%, while matched-clean performance remains unchanged. Broader preserving-side aggregates and deployment details are deferred to the appendix.

Table 5: Matched-carrier external strategy comparison on Milk-Swap. Each strategy uses 200 clean, 200 control, and 200 break episodes; lower is better in *Break*.

Strategy	Clean success rate ↑	Control success rate ↑	Break (clean) ↓	Ctrl.-Break gap ↑
BAS-VLA	196/200 (98.0%)	195/200 (97.5%)	0/200 (0.0%)	97.5
SGAC-AC (So et al., *NeurIPS'25*)	194/200 (97.0%)	195/200 (97.5%)	2/200 (1.0%)	96.5
AAC (Liang et al., *CVPR'26*)	196/200 (98.0%)	195/200 (97.5%)	2/200 (1.0%)	96.5
RTC (Black et al., *NeurIPS'25*)	197/200 (98.5%)	192/200 (96.0%)	4/200 (2.0%)	94.0
PCD (Wu et al., *ICLR'26*)	176/200 (88.0%)	179/200 (89.5%)	6/200 (3.0%)	86.5
ST4-VLA (Ye et al., *ICLR'26*)	189/200 (94.5%)	190/200 (95.0%)	45/200 (22.5%)	72.5
HAMLET (Koo et al., *ICLR'26*)	88/200 (44.0%)	86/200 (43.0%)	0/200 (0.0%)	43.0

Table 6: Real-robot changed-task completion. Every case uses 80 trials per method (640 changed-condition rollouts in total). T1/T2 include the 17 changed-condition trials per method in Table 11.

Case	Semantic change	n/method	Frozen		BAS-VLA		Δ (pts)
			Successes	SR ↑	Successes	SR ↑	
T1	Block target	80	19	23.8%	47	58.8%	+35.0
T2	Fruit target	80	8	10.0%	31	38.8%	+28.8
T3	Receptacle target	80	13	16.3%	52	65.0%	+48.8
T4	Spatial relation	80	11	13.8%	36	45.0%	+31.3
Aggregate		320	51	15.9%	166	51.9%	+35.9

The main semantic-breaking evidence is consolidated in Table 4. The primary Milk-Swap setting shows the intended high-clean/high-control/low-break pattern. OJ-Swap, BBQ-Swap, and Ketchup-Swap reproduce the same high-clean/high-control/low-break pattern, while Butter-Swap and Tomato-Swap remain harder cases. Butter-Swap in particular retains substantial old-task success, an important negative case consistent with confusion between a visually and semantically similar object pair; we report outcome-level error categories in the appendix. Figure 5 shows the same distinction at the rollout level: Milk-Swap stays short and unsaturated on clean/control but saturates under break, indicating strong suppression of the stale clean-task rollout under semantic change, whereas Tomato-Swap shows a weaker version of the same profile.

We next compare with alternative inference-time strategies under matched rollout conditions. All methods are instantiated on the same frozen OpenPI carrier and evaluated under the same semantic-breaking setting, so the comparison isolates the strategy layer. Table 5 shows that BAS-VLA combines the lowest break-column value with high clean/control performance; the alternative methods either leave more residual clean-task success under semantic breaking or reduce clean/control competence more substantially. The corresponding process-side comparison is deferred to Figure 10 in the appendix. We also evaluate BAS-VLA on four real-robot semantic-change cases under the fixed deployment stack summarized in Table 9. Table 6 reports success on the *new* instruction, using 80 trials per method and case. BAS-VLA improves changed-task completion across all four cases; Table 11 retains the clean/control and behavioral diagnostics for the two initial tasks.

The same semantic-breaking pattern persists across the additional target-object settings and under stronger nuisance interference, including hard noise and combined clutter-plus-noise conditions. These results indicate that the observed separation is not tied to a single favorable task instance; Appendix B and Figure 8 provide the corresponding scope and visual context. Table 7 evaluates changed-task completion directly. Across the four easy swaps, BAS-VLA reaches 68.5% new-task success, while both the frozen carrier and the norm-matched random residual remain below 3%. This comparison distinguishes instruction-aligned redirection from residual magnitude alone. Butter-Swap remains the harder case, and the mixed visual-interference setting retains changed-task completion alongside stale-task suppression.

External zero-shot evaluation on LIBERO-Plus and LIBERO-PRO tests transfer beyond our constructed interventions, using matched carriers without tuning on either protocol [Fei et al., 2025, Zhou et al., 2025]. Table 8 shows consistent success-rate gains on both protocols. Non-object semantic changes additionally cover destination, spatial relation, and order (Table 13).

Table 7: Changed-task outcome decomposition. New, Old, and Other/timeout are mutually exclusive. Easy swaps aggregate Milk/OJ/BBQ/Ketchup (four tasks × four seeds × 50 episodes); each harder setting uses four seeds × 50 episodes.

Setting	Method	n	New ↑	Old ↓	Other/timeout ↓
Easy target-object swaps: Milk / OJ / BBQ / Ketchup					
Four-task aggregate	Frozen	800	18 (2.3%)	712 (89.0%)	70 (8.8%)
Four-task aggregate	BAS-VLA	800	548 (68.5%)	0 (0.0%)	252 (31.5%)
Four-task aggregate	Random residual	800	12 (1.5%)	708 (88.5%)	80 (10.0%)
Harder target-object and visual-interference settings					
Butter→cream cheese	BAS-VLA	200	53 (26.5%)	113 (56.5%)	34 (17.0%)
OJ→milk + vis. interf.	BAS-VLA	200	127 (63.5%)	0 (0.0%)	73 (36.5%)

Table 8: External-protocol success without benchmark-specific tuning. LIBERO-Plus uses four categories (light/background/noise/language) × ten cases × 120 rollouts per method; LIBERO-PRO uses ten task-perturbation cases × 120 rollouts per method.

Protocol	Cases	n/method	Baseline		BAS-VLA		Δ (pts)
			Successes	SR ↑	Successes	SR ↑	
LIBERO-Plus	4 × 10	4,800	2,632	54.8%	3,741	77.9%	+23.1
LIBERO-PRO	10	1,200	443	36.9%	857	71.4%	+34.5

5.3 Ablations

The main ablation question in this work is whether the individual calibration mechanisms and deployment modes remain distinguishable under protocol-identical comparison. Table 19 therefore focuses on the semantic-breaking setting and compares the frozen carrier baseline, the offline instruction-margin variant, the online semantic-margin variant, the default BAS-VLA core, and the monolithic full composition under the same OpenPI-pi0.5/LIBERO-Object protocol on two target-object settings. The frozen carrier baseline leaves substantial residual clean-task success under semantic-breaking interventions; the instruction-margin-only and semantic-margin-only variants improve separation; and the default BAS-VLA core is the strongest protocol-identical deployment. The full composition does not dominate the default core, which is consistent with the intended selective role of the preserving auxiliary. The preserving-side deployment comparison and the focused action-space support view are reported in Tables 18 and 22 in the appendix.

Full receives no perturbation-family label. The fixed-policy gate diagnostics, controlled mask-corruption study, and end-to-end latency measurements are reported in Tables 14, 15, and 12. These measurements distinguish the default breaking core from the optional preserving sidecar. Clean and semantic-break false activations are 7.5% and 8.0%, respectively, while the preserving-shift false-negative rate is 23.5%. Oracle-routed Full retains 91.5% task success and is reported separately as an upper-bound diagnostic. Mask-corruption degradation increases from 2.0 points at IoU ≥ 0.7 to 18.5 points with missed-target or random masks. The breaking core adds 47–53 ms at p50; preserving and Full modes have end-to-end p50/p95 latencies of 924/1,048 and 972/1,112 ms, respectively, including the additional frozen-policy query.

Additional evaluations and boundary cases are deferred to Appendix A.2.3.

6 Conclusion

This work identifies a reliability gap in vision-language-action policies that generic robustness does not capture. BAS-VLA addresses it with a breaking-centered frozen-policy calibration core and a selective evidence-gated preserving auxiliary. BAS-VLA maintains high clean/control success under semantic-breaking evaluation and improves task-preserving style-shift performance. Stale-task suppression and changed-task completion are distinct outcomes, measured separately in the dual-goal evaluations. Overall, the results support task-semantic action calibration as a complementary mechanism for controlling when VLA behavior should remain stable and when it should change under task-semantic interventions.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding or support for this work and declare no competing interests.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, 2022.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 2025a.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. In *Advances in Neural Information Processing Systems*, 2025b.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 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*, pages 2165–2183. PMLR, 2023.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. Counterbench: Evaluating and improving counterfactual reasoning in large language models. *arXiv preprint arXiv:2502.11008*, 2025.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

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: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao

Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *European Conference on Computer Vision*, 2010.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Fei-Fei Li. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

Alihan Hüyük, Xinnuo Xu, Jacqueline Maasch, Aditya V. Nori, and Javier Hernandez. Reasoning elicitation in language models via counterfactual feedback. In *International Conference on Learning Representations*, 2025.

Suhyeok Jang, Dongyoung Kim, Changyeon Kim, Youngsuk Kim, and Jinwoo Shin. Verifier-free test-time sampling for vision language action models. In *International Conference on Learning Representations*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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 *8th Annual Conference on Robot Learning*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems*, 2025.

Myungkyu Koo, Daewon Choi, Taeyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. Hamlet: Switch your vision-language-action model into a history-aware policy. In *International Conference on Learning Representations*, 2026.

Wendi Li and Yixuan Li. Process reward model with q-value rankings. In *International Conference on Learning Representations*, 2025.

Yuanchang Liang, Xiaobo Wang, Kai Wang, Shuo Wang, Xiaojiang Peng, Haoyu Chen, David Kim Huat Chua, and Prahlad Vadakkepat. Adaptive action chunking at inference-time for vision-language-action models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 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*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Yuankai Luo, Woping Chen, Tong Liang, Baiqiao Wang, and Zhenguo Li. Simvla: A simple vla baseline for robotic manipulation. *arXiv preprint arXiv:2602.18224*, 2026.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022.

Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *IEEE International Conference on Robotics and Automation*, 2023.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

Edoardo Pona, Milad Kazemi, Yali Du, David Watson, and Nicola Paoletti. Abstract counterfactuals for language model agents. In *Advances in Neural Information Processing Systems*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025.

Junhyuk So, Chiwoong Lee, Shinyoung Lee, Jungseul Ok, and Eunhyeok Park. Improving generative behavior cloning via self-guidance and adaptive chunking. In *Advances in Neural Information Processing Systems*, 2025.

Shihan Wu, Xu Luo, Ji Zhang, Junlin Xie, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Policy contrastive decoding for robotic foundation models. In *International Conference on Learning Representations*, 2026.

Jinhui Ye, Fangjing Wang, Ning Gao, Junqiu Yu, Yangkun Zhu, Bin Wang, Jinyu Zhang, Weiyang Jin, Yanwei Fu, Feng Zheng, Yilun Chen, and Jiangmiao Pang. Spatially guided training for vision-language-action model. In *International Conference on Learning Representations*, 2026.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020.

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.

A Supplementary Results and Visual Evidence

This appendix provides supplementary quantitative and qualitative results, including perturbation inventories, additional benchmark results, and additional visual evidence from completed experiments.

Licensing. BAS-VLA code is released under the MIT license. OpenPI and SAM 2 use Apache 2.0; LIBERO and OpenVLA-OFT use MIT. Other third-party assets retain the licenses and terms specified in their upstream repositories.

A.1 Compute and Real-Robot Setup

Table 9 summarizes the fixed compute and deployment hardware used for the simulation and real-robot experiments. The compute row records the server configuration used for the main experiments, and the remaining rows summarize the shared robot and camera stack used by the real-robot deployment setup.

Table 9: Compute and real-robot setup. Depth streams are retained only for recording and diagnosis; model inputs use RGB only.

Component	Configuration
Compute server	8× NVIDIA RTX 5880 Ada GPUs (48GB VRAM each); single-GPU execution unless noted
Robot platform	Songling PiPER, single-arm
Robot control stack	`piper_sdk` over CAN
Cameras	2× Orbbec RGB-D input cameras + 1 record-only overview camera
Model input modality	RGB only
Camera SDK	OrbbecSDK v1

Table 10: Real-robot deployment latency diagnostics. Chunk p50 and p95 report per-chunk latency, while Wall-clock p50 reports full-episode duration.

Task	Method	Chunk p50	Chunk p95	Underrun rate	Wall-clock p50
Color-block grouping	baseline	345 ms	620 ms	0.18	34.5 s
Color-block grouping	**BAS-VLA default**	392 ms	710 ms	0.27	41.8 s
Fruit-to-plate	baseline	368 ms	655 ms	0.22	43.2 s
Fruit-to-plate	**BAS-VLA default**	421 ms	768 ms	0.31	52.6 s

Extended real-robot evaluation and mode costs. The four changed-task cases in Table 6 comprise 640 rollouts, with 80 trials per method and case. T1 and T2 include the 17 changed-condition trials per method reported in Table 11; the latter also gives the clean/control and behavioral diagnostics. T3 and T4 add receptacle-target and spatial-relation changes.

Table 11: Real-robot semantic-adaptation results on two deployment tasks. Each method-task pair uses 16 clean, 17 control, and 17 break episodes. Here *Break* reports success on the changed instruction, and the separation score averages old-target suppression and new-target first-commit.

Task	Method	n (c/ctrl/br)	Clean	Control	Break	Old-target suppr.	New-target first-commit	Sep. score
T1: group the specified colored block with its target set								
baseline	Frozen	16/17/17	9/16 (56.3%)	9/17 (52.9%)	4/17 (23.5%)	7/17 (41.2%)	6/17 (35.3%)	38.2
BAS-VLA default	BAS-VLA	16/17/17	10/16 (62.5%)	10/17 (58.8%)	7/17 (41.2%)	12/17 (70.6%)	11/17 (64.7%)	67.6
T2: place the specified fruit onto the plate								
baseline	Frozen	16/17/17	8/16 (50.0%)	8/17 (47.1%)	3/17 (17.6%)	6/17 (35.3%)	5/17 (29.4%)	32.4
BAS-VLA default	BAS-VLA	16/17/17	9/16 (56.3%)	9/17 (52.9%)	6/17 (35.3%)	11/17 (64.7%)	10/17 (58.8%)	61.8

The breaking core adds 47–53 ms at p50. Preserving and Full modes compute the probe action before the action-disagreement factor g_{act}, so the gate controls the probe's influence on execution while the additional policy query remains included in the measured latency.

Table 12: End-to-end chunk latency by deployment mode. Sidecar and Full totals include grounding, mask refinement, the probe image, and the additional frozen-policy query, measured over at least 200 chunks per task and three timing passes.

Mode	T1 p50/p95 (ms)	T2 p50/p95 (ms)
Frozen carrier	345/620	368/655
Default breaking core	392/710	421/768
Preserving sidecar	924/1,048	924/1,048
Full composition	972/1,112	972/1,112

A.2 Protocol and Extended Results

A.2.1 Protocol Details

The main semantic-breaking evaluations use matched clean/control/break triplets. Clean denotes the original task condition, control denotes a semantics-preserving variant of that task, and break denotes a task-semantic change. Unless explicitly stated otherwise, the *Break* column in the simulation semantic-breaking tables reports clean-task success under the changed task condition, so lower values indicate stronger semantic separation. Task outcome is the primary metric, and action-level diagnostics are reported only as supporting evidence for whether break conditions move farther from the clean action basin than preserving controls do.

The benchmark and carrier roles are likewise fixed. OpenPI-pi0.5 with LIBERO-Object is the main semantic-breaking and matched external-strategy setting; OpenVLA-OFT with LIBERO-Spatial is the main task-preserving setting; and LIBERO-10 provides supplementary corroboration. CALVIN, MetaWorld, ManiSkill2, and RoboCasa appear only as boundary or transfer settings. The primary preserving campaign evaluates ten LIBERO-Spatial tasks with fifty episodes per task, yielding 500 episodes per variant-condition pair. The primary semantic-breaking campaign evaluates four seeds with fifty episodes per seed, yielding 200 episodes per condition. External strategies are compared only when carrier, benchmark slice, rollout interface, and evaluation budget are matched.

These comparisons are intentionally anchored to strong in-domain carriers rather than weak generic zero-shot policies. OpenPI-pi0.5 is already adapted to LIBERO object manipulation, and OpenVLA-OFT is already adapted to LIBERO-Spatial. The matched-carrier protocol therefore isolates action calibration on competent frozen substrates rather than conflating it with gross backbone mismatch.

The same appendix also records details that are compressed in the main paper. The primary metric is task success rate. In preserving settings, the goal is to maintain success under nuisance variation without harming clean-task performance; in semantic-breaking settings, the goal is to keep clean/control success high while reducing clean-task success under the changed task condition. Process statistics such as episode length, saturation, replanning, and action-gap diagnostics are reported only as supporting evidence. The real-robot deployment study uses the same BAS-VLA interface and is summarized by Tables 9 and 10.

Changed-task and non-object evaluation. The dual-goal evaluator assigns each changed-condition episode to mutually exclusive New, Old, and Other/timeout outcomes. Table 7 reports these outcomes together with the norm-matched random-residual control. Table 13 extends evaluation beyond object substitutions; each family uses two tasks, four seeds, and fifty episodes per seed.

Unified gate and probe robustness. A fixed Full policy is evaluated in a 2×2 design crossing preserved/changed semantics with absent/present visual shift. The gate target is whether the probe action is closer than the base action to the expert/reference action. Standard Full receives no oracle family label. Table 14 reports condition-level false activation and false negatives, with oracle-routed task retention reported separately as an upper-bound diagnostic.

Table 13: Changed-task completion on non-object semantic interventions.

Semantic change	Evaluation volume	Metric	BAS-VLA result
Destination	$2 \times 4 \times 50 = 400$	New-task SR	224/400 (56.0%)
Spatial relation	$2 \times 4 \times 50 = 400$	New-task SR	168/400 (42.0%)
Subgoal order	$2 \times 4 \times 50 = 400$	New-task SR	128/400 (32.0%)

Table 14: Full-mode gate diagnostics. Every row uses four seeds and 50 episodes per seed.

Gate diagnostic	Evaluation volume	Result
Clean false activation	200	15/200 (7.5%)
Preserving-shift false negative	200	47/200 (23.5%)
Semantic-break false activation	200	16/200 (8.0%)
Oracle-routed Full task retention (upper bound)	200	183/200 (91.5%)

Table 15: Preserving-SR change relative to the normal-mask reference. Each corruption level uses four seeds and 50 episodes per seed.

Mask condition	Evaluation volume	Preserving-SR change
IoU ≥ 0.7	200	$-4/200$ (-2.0 pts)
IoU $= 0.5$	200	$-19/200$ (-9.5 pts)
Missed target or random mask	200	$-37/200$ (-18.5 pts)

Controlled mask corruption measures how localization quality propagates through evidence-gated fusion. Table 15 reports preserving-SR changes relative to normal masks. Moderate localization errors cause smaller degradation than missed-target or random masks. The identity $\tau = 0 \Rightarrow a^{\text{pres}} = a^{(0)}$ describes fallback at a low gate, while these measurements quantify error propagation when the probe influences the action.

Statistical uncertainty. Table 16 reports Wilson 95% confidence intervals computed from aggregate binary episode counts. The intervals quantify uncertainty in both the preserving improvement and the high-clean/high-control/low-old-task pattern without requiring episode pairing.

A.2.2 Complexity, Scope, and Boundaries

The BAS-VLA breaking core remains lightweight relative to the underlying policy, adding one instruction feature and one residual module. Preserving and Full modes require an additional base-policy query to compute the probe action before evaluating the action-disagreement gate; the gate controls the influence of that action. Appendix C.1 summarizes the inference-time control flow, and Appendix C.5 records the corresponding per-mode cost relations and exact identity relations.

The scope claim is deliberately narrow. BAS-VLA is not presented as a universal preserving branch or a general-purpose robustness layer. The strongest evidence comes from breaking-centered calibration, with selective preserving gains on validated nuisance families. Additional evaluations on other carriers, suites, and local protocols are summarized in Table 20. Boundary and negative cases, including executability changes and supplementary benchmarks such as MetaWorld, ManiSkill2, RoboCasa, and CALVIN, are summarized in Table 21 and the extended appendix inventories.

A.2.3 Analysis and Boundary Summaries

The main text focuses on the preserving style-shift setting, the primary semantic-breaking rows, and the matched strategy comparison. The appendix collects the broader scope checks: additional evaluations on other carriers and suites, supplementary harder-task or interference results, boundary cases, and negative cases. These rows are included to delimit where the current deployment is supported, where it remains only partially supported, and which benchmark transfers should still be interpreted cautiously.

Table 16: Wilson 95% confidence intervals for the principal binary outcomes.

Result	Success rate	Wilson 95% CI
Baseline style	84/200 (42.0%)	[35.4, 48.9]
BAS-VLA style	140/200 (70.0%)	[63.3, 75.9]
Milk clean	196/200 (98.0%)	[95.0, 99.2]
Milk control	195/200 (97.5%)	[94.3, 98.9]
Milk old-task under break	0/200 (0.0%)	[0.0, 1.9]
Butter old-task under break	113/200 (56.5%)	[49.6, 63.2]

Table 17: Task-preserving perturbation inventory. Baseline-only rows report changed-condition success without BAS-VLA calibration; calibrated rows report the corresponding method-level changed-condition success.

Carrier	Suite	Task slice	Perturbation	Best method	Baseline ↑	Calibrated ↑	Δ ↑ (pts)
Limited-recovery appearance family							
Pi0.5	LIBERO-Object	OJ-Swap×200	illumination change	baseline-only	192/200 (96.0%)	—	—
Moderate support under stronger nuisance							
OFT	LIBERO-Spatial	all10×20	clutter+noise	PARR	165/200 (82.5%)	174/200 (87.0%)	+4.5
OFT	LIBERO-Spatial	all10×20	hard noise band	PARR	171/200 (85.5%)	180/200 (90.0%)	+4.5
Style-sensitive families with strongest BAS-VLA gains							
OFT	LIBERO-Spatial	all10×50	style shift	BAS-VLA (style-aux)	470/500 (94.0%)	482/500 (96.4%)	+2.4
OFT	LIBERO-Spatial	Bowl-on-Ramekin ×200	style shift	BAS-VLA (style-aux)	84/200 (42.0%)	140/200 (70.0%)	+28.0
OFT	LIBERO-Spatial	Bowl-next-to-Plate×200	style shift	BAS-VLA (style-aux)	96/200 (48.0%)	138/200 (69.0%)	+21.0
OFT	LIBERO-Spatial	Bowl-on-Cabinet×200	style shift	BAS-VLA (style-aux)	101/200 (50.5%)	147/200 (73.5%)	+23.0
OFT	LIBERO-Spatial	aggregation	style shift	PSSC	190/200 (95.0%)	200/200 (100.0%)	+5.0

The headline semantic-breaking rows should not be read as a generic robustness objective. A uniformly robust policy would preserve clean-task success across clean, control, and break alike. Our protocol instead treats low clean-task success under a genuine break as evidence of the intended behavioral split. The process-side panels and matched strategy comparisons are included for the same reason: they show that semantic separation is visible not only in final task outcomes, but also in rollout duration and strategy-level budget use.

A.2.4 Extended Related Positioning

The related-work positioning in the main text emphasizes only the closest conceptual neighborhoods. The additional policy-side context is that recent inference-time VLA methods differ primarily in what they optimize at test time: RTC and adaptive chunking methods target execution freshness and latency, ST4VLA strengthens spatial grounding, PCD contrasts original and masked observations, and HAMLET or MG-Select emphasize history-aware prediction or selection. These methods are close to BAS-VLA in deployment regime, but not in comparison object: BAS-VLA is organized around controlled preserving and semantic-breaking interventions rather than around generic test-time improvement.

The additional reasoning-side context is that counterfactual feedback, semantic-intervention benchmarks, and process supervision collectively motivate intervention-sensitive evaluation rather than fixed-answer evaluation alone. BAS-VLA imports that logic into embodied control by asking whether action behavior remains invariant when meaning is unchanged and separates when meaning truly changes.

Table 17 reports task-preserving results across perturbation families and task settings. The inventory is organized to distinguish limited-recovery appearance families, moderate support under stronger nuisance settings, and the strongest style-sensitive BAS-VLA gains. The OpenPI-pi0.5 illumination row serves as an appearance-family reference, while the style-shift rows emphasize the task-specific slices on which the preserving auxiliary is most effective.

A.3 Supplementary Visual Evidence

This subsection groups the screenshot-based supplementary evidence into three roles. The first role is preserving-side motivation and nuisance robustness, where clean rollouts are juxtaposed with semantics-preserving perturbations to show when the visual change should leave the task unchanged but still induces behavioral drift or failure. The second role is semantic-breaking and harder-setting

Table 18: Protocol-identical preserving-side and deployment ablations on the Bowl-on-Ramekin style-shift setting, with the corresponding OFT-LIBERO replay row. Each Bowl-on-Ramekin row uses 200 episodes.

Variant	Role	Bowl-on-Ramekin clean ↑	Bowl-on-Ramekin style-shift ↑	Δ_{style} vs. baseline ↑	OFT-LIBERO replay
Carrier baseline	frozen-policy ref.	108/200 (54.0%)	84/200 (42.0%)	—	native 20/20
Semantic-margin only	break-only cal.	108/200 (54.0%)	83/200 (41.5%)	−0.5 pts	20/20
Style-aux only	preserving-only aux.	107/200 (53.5%)	132/200 (66.0%)	+24.0 pts	19/20
BAS-VLA (style-aux)	selective preserve.	108/200 (54.0%)	140/200 (70.0%)	+28.0 pts	20/20
BAS-VLA (full)	full composition	106/200 (53.0%)	134/200 (67.0%)	+25.0 pts	20/20

Table 19: Protocol-identical breaking-side ablations on two target-object settings. Each clean/control/break column uses 200 episodes; lower is better in *Break*.

Variant	Role	Milk-Swap setting				OJ-Swap setting			
		Clean ↑	Control ↑	Break ↓	Gap ↑	Clean ↑	Control ↑	Break ↓	Gap ↑
Carrier baseline	●	196/200 (98.0%)	194/200 (97.0%)	180/200 (90.0%)	7.0	197/200 (98.5%)	193/200 (96.5%)	176/200 (88.0%)	8.5
Instruction-margin only	●	196/200 (98.0%)	195/200 (97.5%)	72/200 (36.0%)	61.5	197/200 (98.5%)	192/200 (96.0%)	80/200 (40.0%)	56.0
Semantic-margin only	●	196/200 (98.0%)	195/200 (97.5%)	16/200 (8.0%)	89.5	196/200 (98.0%)	192/200 (96.0%)	22/200 (11.0%)	85.0
BAS-VLA (core)	●	196/200 (98.0%)	195/200 (97.5%)	0/200 (0.0%)	97.5	197/200 (98.5%)	193/200 (96.5%)	0/200 (0.0%)	96.5
BAS-VLA (full)	●	195/200 (97.5%)	193/200 (96.5%)	8/200 (4.0%)	92.5	196/200 (98.0%)	191/200 (95.5%)	10/200 (5.0%)	90.5

● frozen-policy reference ● mechanism-only ablation ● BAS-VLA deployment variant

evidence, where matched clean and break rollouts are extended to stronger interference conditions so that the clean/control/break distinction remains visible at the rollout level. The third role is external-strategy qualitative evidence, whose purpose is not to replace the matched-carrier quantitative tables but to make the behavioral differences between comparison strategies easier to inspect.

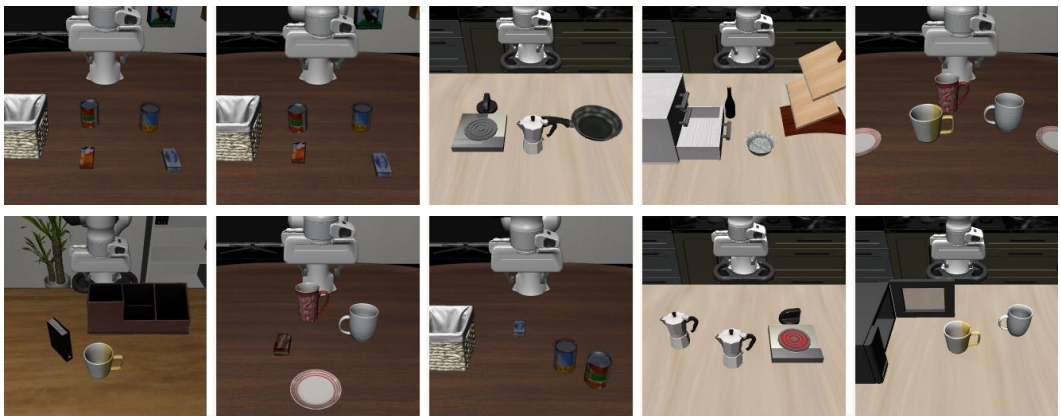


Figure 6: Main simulation task scenes. The grid shows the ten task layouts used in the primary simulation campaigns.

Butter→cream cheese yields 53/200 New, 113/200 Old, and 34/200 Other/timeout outcomes. Stale-task completion is the dominant failure outcome. The rollout annotations distinguish stale-target grasp, correct-target reach followed by grasp failure, third-object commitment, oscillation, and timeout, separating target confusion from downstream control failure.

A.4 Concrete Probe Instantiation

The main text defines the preserving-side probe transform ϕ_{sty} functionally, as a deterministic, non-generative, geometry-preserving appearance-canonicalization map. Here the subscript "sty" denotes generic observation-side visual variation in the input image rather than only a narrow style-transfer setting. A concrete instantiation consistent with that definition is a grounded-foreground probe. Given the current observation o_t and instruction x, we first use text-conditioned grounding to localize the robot arm, gripper, target object, and task-relevant receptacle cues implied by x. We then refine these grounded regions into a merged foreground mask and attenuate only the complement region. The resulting probe leaves task-relevant foreground geometry intact while weakening nuisance-heavy background evidence.

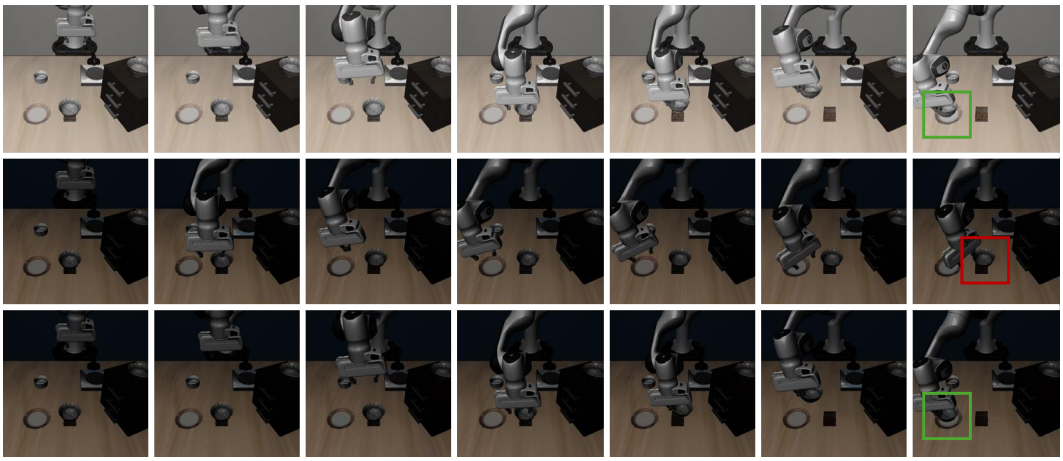


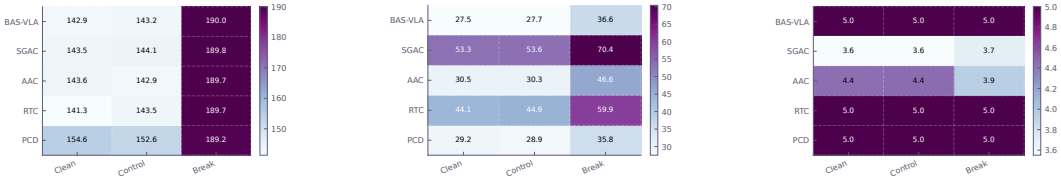
Figure 7: Dark-scene preserving comparison. Top: clean success. Middle: illumination-change failure under the frozen baseline. Bottom: the same darkened condition succeeds with BAS-VLA.



Figure 8: Representative OJ-Swap under clutter+noise; BAS-VLA completes the intended task.



Figure 9: Butter-Swap failure sequence. The rollout continues toward the stale clean-task object rather than committing to the swapped target.



(a) Mean steps ↓. (b) Mean replans ↓. (c) Mean selected replan steps ↑.

Figure 10: Process-side comparison among external strategies with a BAS-VLA reference row. Rows are strategies and columns are clean, control, and break.

Table 20: Support and corroboration results across additional suites and carriers. Triplet rows report clean, control, and break directly; matched-support and native-corroboration rows state their local comparison explicitly.

Slice	Carrier	Suite	Protocol	Result
Direct triplet corroboration				
SimVLA	SimVLA [Luo et al., 2026]	LIBERO	triplet	20/20, 20/20, 0/20
Executability	OpenPI-pi0.5	LIBERO-10	triplet	19/20, 19/20, 1/20
Destination (hard)	OpenPI-pi0.5	LIBERO-10	triplet	18/20, 19/20, 4/20
Matched support comparisons				
SSAC Order	OpenVLA-OFT	OFT-LIBERO	matched support	baseline 0.80, 1.00, 1.00; support 1.00, 1.00, 1.00
Bowl-on-Ramekin	OpenVLA-OFT	LIBERO-Spatial	matched support	baseline 108/200, 84/200; support 108/200, 140/200
Noise-band support	OpenVLA-OFT	LIBERO-Spatial	matched support	baseline 187/200; support 191/200
Native corroboration under local metrics				
LIBERO-10 TargetSwap	OpenPI-pi0.5	LIBERO-10	native corroboration	0/20, 0/20, 0/20; gap +0.113
Second-carrier corr.	OpenVLA-OFT	LIBERO-10	native corroboration	0/20, 0/20, 0/20; gap +0.005

One explicit form is

$$B_t = \mathrm{GroundDINO}(o_t, x), \qquad M_t = \mathrm{SAM2}(o_t, B_t), \qquad \tilde{o}_t^{\mathrm{sty}} = M_t \odot o_t + (1 - M_t) \odot \psi_{\mathrm{bg}}(o_t),$$

where B_t denotes grounded boxes, M_t denotes the merged foreground mask, and ψ_{bg} denotes deterministic background attenuation on the mask complement. In this work, ψ_{bg} is intended to reduce color cast, illumination bias, and texture-heavy surface variation without altering object layout or semantic structure.

A.5 Support and Boundary Analyses

This subsection complements the main-text results with additional evaluations that clarify scope. We separate them into two groups. The first group consists of *support and corroboration* evaluations from additional suites and carriers. These evaluations report whether the same calibration logic remains visible once the benchmark family, carrier, or local success criterion changes. The second group consists of *boundary and negative* evaluations covering benchmark-specific, protocol-specific, or integration-heavy cases.

Table 20 reports additional results beyond the main target-object protocol. The table is organized into direct triplet corroboration on additional semantic-breaking families, matched support comparisons on preserving-side settings, and native corroboration rows reported under local benchmark metrics. These rows are supportive rather than protocol-identical to the main comparison setting, and are included to show whether the qualitative direction of the BAS-VLA effect remains visible once the carrier, suite, or local metric changes.

The matched-support block uses local comparator names from the supplementary evaluations. In particular, *SSAC Order* denotes the order-support row from the SSAC comparator family on the OFT-LIBERO carrier: the evaluation reports a baseline-to-support clean/control/break improvement under the local OFT order protocol rather than a new protocol-identical BAS-VLA deployment row.

Table 21 collects selected boundary cases in three groups: harder semantic-breaking families, stronger preserving-side stress settings, and cross-benchmark mismatch cases. Some rows retain the triplet structure used in the main semantic-breaking evaluations, while others are reported in suite-native formats such as changed-only stress testing or clean-only evaluation. For the non-triplet rows, the Result column states the comparison explicitly rather than reusing triplet shorthand.

Table 22 reports a focused action-space view of the BAS-VLA Full mechanism variant on the preserving anchor and the evaluated support families under a matched OpenVLA-OFT protocol.

Table 21: Selected boundary and negative cases across additional benchmarks and protocols. Triplet rows report clean, control, and break directly; non-triplet rows use their suite-native comparison format.

Slice	Carrier	Suite	Protocol	Result
Semantic-breaking boundary cases				
Executability change	OpenPI-pi0.5	LIBERO-10	triplet	$0/20, 0/20, 1/20$
Spatial relation (hard)	OpenPI-pi0.5	LIBERO-10	triplet	$17/20, 18/20, 9/20$
Destination change (hard)	SimVLA [Luo et al., 2026]	LIBERO	triplet	$18/20, 17/20, 6/20$
Preserving-side stress settings				
Clutter+noise stress	OpenVLA-OFT	LIBERO-Spatial	changed only	baseline 82.5%; stress 70.0%; BAS-VLA 74.5%
Viewpoint/layout stress	OpenVLA-OFT	LIBERO-Spatial	changed only	baseline 78.0%; BAS-VLA 79.5%
Cross-benchmark mismatch				
MetaWorld Door-Lock	OpenPI-pi0.5	MetaWorld	triplet	$17/30, 16/30, 17/30$

Table 22: Focused action-space diagnostics on the preserving anchor and evaluated support families under BAS-VLA Full. All rows use the matched OpenVLA-OFT ext6 protocol with five trials per family and a 24-step comparison prefix. Larger positive values indicate clearer semantic separation.

Block	Family	Break-sep. rate ↑	Clean-ref. gap ↑	Initial-chunk sep. ↑	Executed-prefix sep. ↑
Preserving anchor	Semantic-preserving control	20.0	-0.019	-0.019	-0.017
Retained support families					
Retained support	Spatial relation	100.0	$+0.055$	$+0.055$	$+0.079$
Retained support	Target object	60.0	$+0.001$	$+0.001$	$+0.002$
Retained support	Destination	60.0	$+0.011$	$+0.011$	$+0.008$

Unlike the main semantic-breaking tables, which are organized around task-level success, this table isolates whether the action-space diagnostics follow the expected sign structure for semantic separation. Break-sep. rate denotes the fraction of episodes in which the break rollout is farther from the clean reference than the control rollout, so higher values indicate clearer separation. Clean-reference gap denotes the average difference $\mathrm{dist}(\mathrm{break}, \mathrm{clean}) - \mathrm{dist}(\mathrm{control}, \mathrm{clean})$. Initial-chunk separation and executed-prefix separation denote the analogous break-minus-control differences computed on the first action chunk and on the executed comparison prefix, respectively. Under this interpretation, the preserving anchor should remain near zero or negative in the three distance-difference columns, whereas evaluated breaking families should yield positive values.

The reported signs are consistent with this interpretation: the preserving anchor remains negative, spatial relation shows the clearest positive separation, and the target-object and destination rows remain positive but smaller in magnitude.

A.6 Uncertainty and Trial-Level Stability Notes

Table 16 complements exact rollout counts with Wilson 95% intervals. The style-shift intervals support the preserving improvement, while the Milk-Swap intervals show high clean/control retention and low old-task success under semantic change. These intervals describe uncertainty in aggregate binary outcomes; repeated trials and the matched evaluation protocol provide the accompanying experimental context.

A.7 Comparison to Alternative Calibration Strategies

Table 23 compares BAS-VLA to several related calibration strategies and deployment variants under the local settings used to probe each mechanism. The left block summarizes preserving-side strategies, ranging from generic appearance consensus to more targeted auxiliaries such as Peripheral Anomaly-Triggered Robust Action Rerouting (PARR) and Phase-Scoped Scene-Style Consistency (PSSC), while the right block summarizes breaking-side or deployment-level strategies such as the semantic-margin-only and instruction-margin calibrators. Arrow notation denotes before/after percentage changes for the corresponding strategy; it is not a clean/control/break tuple.

These strategies are reported as mechanism probes rather than as direct substitutes for the final method. Generic appearance consensus serves as a baseline appearance-averaging strategy for illumination-change settings. Peripheral Anomaly-Triggered Robust Action Rerouting (PARR) is an inference-time dual-branch rerouting strategy that compares actions from the original observation and

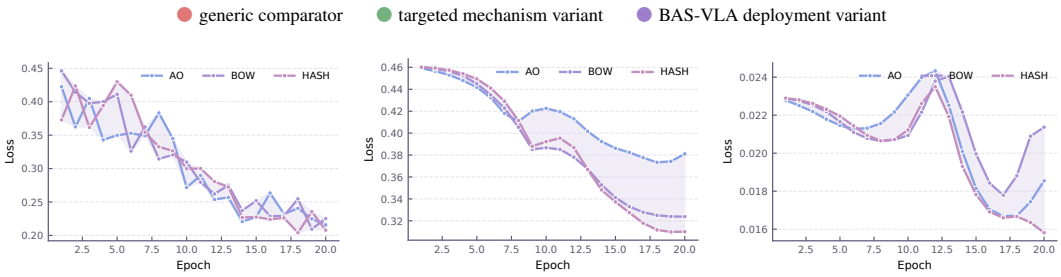
Table 23: Comparison to alternative calibration strategies for BAS-VLA. The left table reports preserving-side strategies and preserving-oriented deployments, and the right table reports breaking-side or deployment-level strategies. Arrow notation denotes before/after percentage changes rather than clean/control/break tuples.

Preserving-side strategies

Variant	Type	Slice	Result
Appearance consensus	●	illumi-nation	$0.70 \rightarrow 0.82$; $0.50 \rightarrow 0.74$
PARR	●	clutter/ noise	$46/50 \rightarrow 47/50$; $46/50 \rightarrow 48/50$; $165/200 \rightarrow 174/200$
PSSC	●	style shift	$95.0 \rightarrow 100.0$; $84/200 \rightarrow 126/200$
Style-aux only	●	Bowl-on-Ramekin style	$84/200 \rightarrow 132/200$; clean held at $107/200$
BAS-VLA (style-aux)	●	Bowl-on-Ramekin style	$84/200 \rightarrow 140/200$; clean held at $108/200$

Breaking-side strategies

Variant	Type	Slice	Result
Instruction-margin	●	positive rate	$0.50 \rightarrow 0.85$
Semantic-margin only	●	Milk-Swap	$196/200$, $195/200$, $16/200$; gap 89.5
BAS-VLA (core)	●	Milk-Swap	$196/200$, $195/200$, $0/200$; gap 97.5
BAS-VLA (core)	●	OJ-Swap	$197/200$, $193/200$, $0/200$; gap 96.5
BAS-VLA (full)	●	full deploy.	native $20/20$; full $20/20$; matched break row weaker than core



(a) Train total. (b) Val total. (c) Val consistency.

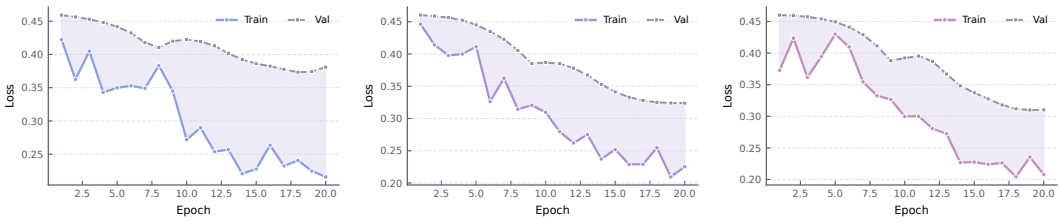
Figure 11: Metric-wise offline instruction comparator losses. Shaded bands show the per-epoch range across trainable variants.

a periphery-suppressed observation, then switches only when both the periphery anomaly score and the branch disagreement are high. Phase-Scoped Scene-Style Consistency (PSSC) is an early-phase consistency gate that activates only under a style-specific cue and only during the first few replanning steps. On the breaking side, the semantic-margin-only row isolates the online semantic-calibration mechanism without the preserving auxiliary, while the instruction-margin row moves the intervention to an offline instruction-level calibration stage rather than the online action-calibration stage used by BAS-VLA. The BAS-VLA core rows provide matched Milk-Swap and OJ-Swap references under the main breaking protocol, and the BAS-VLA (full) row reports the effect of applying the full composition monolithically instead of the selective deployment adopted in the main text.

Figures 11–15 extend this subsection with process and diagnostic views for the offline instruction comparator family. The loss figures should be read as training-side mechanism evidence rather than as headline BAS-VLA training curves: they summarize how several comparator variants optimize their offline objectives. The action-gap figures then show what these variants do in action space, including how strongly break conditions separate from preserving references and how family-specific distances differ across target-object, ordering/executability, and destination/relation groups.

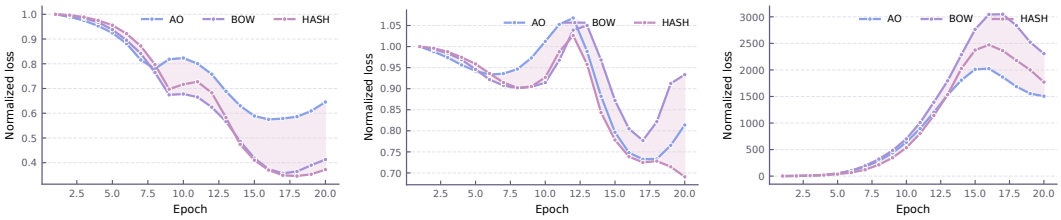
Figures 11–13 focus on optimization dynamics. Across the trainable variants, the total loss decreases over training, but the validation-side trajectories remain separated and the auxiliary objectives retain visible variant-dependent differences. The normalized views in Figure 13 isolate trajectory shape from absolute scale and show that the trainable variants share a common early descent before diverging more clearly later in training.

Figures 14 and 15 shift from training dynamics to action-space behavior. In the aggregate summaries, preserve-to-clean and positive-to-expert distances remain larger than the corresponding break-to-reference distances across the plotted variants, so the offline comparator family organizes these action relations differently from the main online calibration results. The family-wise breakdown further shows that break-to-clean distances are largest on target-object interventions, smaller on



(a) AO. (b) BOW. (c) HASH.

Figure 12: Per-variant train/validation trajectories for the offline instruction comparator family. Shaded regions mark train–validation gaps.



(a) Margin. (b) Consistency. (c) Anchor.

Figure 13: Normalized loss profiles for the offline instruction comparator family. Shaded bands show per-epoch ranges across trainable variants.

destination/relation slices, and smallest on ordering/executability slices, whereas the control-side distances are dominated by format and redundant-wording controls rather than by paraphrase.

B Limitations and Discussion

The evidence supports a narrower claim than "universal VLA robustness." BAS-VLA is best interpreted as a *task-semantic calibration* method for cases in which task-relevant meaning truly changes. The strongest support comes from the semantic-breaking setting, where the method preserves clean/control competence while suppressing stale-task success under target change. Matched external comparisons support this interpretation, but they do not justify a claim that BAS-VLA dominates every inference-time strategy on every operational metric.

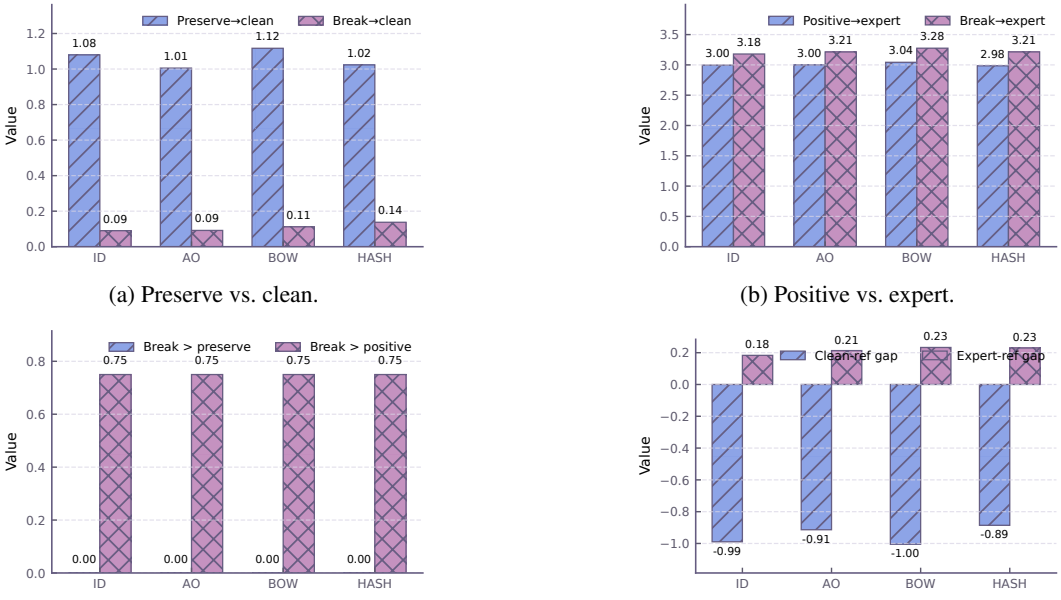
The preserving side is more limited. The auxiliary itself is defined by evidence-gated activation rather than by a family-specific switch, but the strongest preserving-side evidence currently comes from the validated style-shift slice rather than from a universal nuisance-stabilization regime. The full composition is not the preferred deployment under the present evidence. The appendix-level action-space diagnostics are likewise mechanism evidence rather than a replacement for task-level evaluation.

Supplementary benchmarks such as MetaWorld, CALVIN, ManiSkill2, and RoboCasa are best read as boundary or transfer evidence rather than as primary support for the central claim. Their role is to show where the same qualitative calibration logic remains visible and where it weakens under benchmark-specific interfaces or success criteria.

More broadly, the study remains concentrated on controlled intervention families and frozen-policy adaptation. Extending the trusted preserving families, handling more ambiguous semantic-breaking settings, and strengthening efficiency, uncertainty, and verification analyses remain open next steps.

C Appendix. Pseudocode and Formal Properties of BAS-VLA

This appendix records an inference-time pseudocode summary together with several formal properties of the current BAS-VLA parameterization. The pseudocode fixes the deployed control flow and notation used throughout the paper, while the propositions below make explicit the local action-space implications of the breaking objective, the residual calibration form, the preserving-side gate and weak-fusion rule, and the resulting deployment-mode relations.



(a) Preserve vs. clean.

(b) Positive vs. expert.

(c) Break separation.

(d) Clean/expert gaps.

Figure 14: Aggregate action-gap diagnostics for the offline instruction comparator family.

C.1 Inference-Time Deployment Pseudocode

Table 24 summarizes inference-time deployment under the three modes introduced in Section 4: the breaking-centered default mode, the preserving mode, and the full composition retained for ablation. The pseudocode uses exactly the same symbols as the main text and makes the evidence-gated control flow explicit.

Concrete probe instantiation note. The grounded-mask construction above is included as a concrete probe instantiation rather than as a new learning component. It enforces the same modeling intent used in the main text: preserve foreground entities that are semantically tied to the task, and attenuate only the nuisance-heavy complement region. The grounding and segmentation operators therefore define *where* the probe may modify the observation, while the deterministic background attenuator ψ_{bg} defines *how* the complement is weakened.

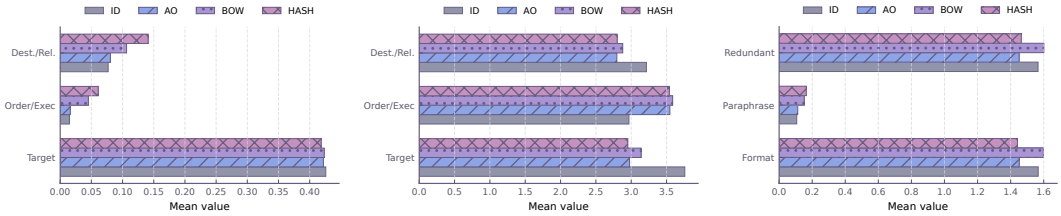
C.2 Formal Action-Space Properties of the Breaking Objective

The breaking-centered objective is built around three terms: imitation on the clean and semantics-preserving control conditions, contraction of clean and control actions toward one another, and a margin-style separation term for semantic-breaking variants. The separation term is

$$\mathcal{L}_{\mathrm{sep}} = \max\Big(0,\ m - \|a^{\mathrm{br}}(x^b) - a^{\mathrm{br}}(x^c)\|_2 + \|a^{\mathrm{br}}(x^u) - a^{\mathrm{br}}(x^c)\|_2\Big).$$

Proposition 1 (Margin implication). Assume $\mathcal{L}_{\mathrm{sep}} = 0$ for a triplet (x^c, x^u, x^b). Then

$$\|a^{\mathrm{br}}(x^b) - a^{\mathrm{br}}(x^c)\|_2 \geq \|a^{\mathrm{br}}(x^u) - a^{\mathrm{br}}(x^c)\|_2 + m.$$



(a) Break-clean by family.

(b) Break-expert by family.

(c) Control invariance by family.

Figure 15: Family-wise action-distance diagnostics for the offline instruction comparator family.

Table 24: Inference-time deployment of BAS-VLA. The preserving-side gate uses the four factors from Eq. (5); the concrete probe instantiation shown here uses text-conditioned grounding, mask refinement, and deterministic background attenuation.

Algorithm 1: Inference-time deployment of BAS-VLA

Input: observation o_t, instruction x, rollout context h_t, step index t, deployment mode $\mathrm{mode} \in \{\textsc{default}, \textsc{pres}, \textsc{full}\}$, frozen base policy π_0, frozen instruction feature extractor $z(\cdot)$, frozen visual feature maps $f_{\mathrm{vis}}^{\mathrm{mid}}$, $f_{\mathrm{vis}}^{\mathrm{late}}$, residual calibrator r_θ, style-suppression map ϕ_{sty}, gate factors $\tau_{\mathrm{phase}}, g_{\mathrm{vis}}, g_{\mathrm{sem}}, g_{\mathrm{act}}$, scales $\kappa_{\mathrm{vis}}, \kappa_{\mathrm{sem}}$, residual scale λ, and fusion weight α.
Output: deployed BAS-VLA action a_t^{deploy}.

Compute the frozen base action $a_t^{(0)} \leftarrow \pi_0(o_t, x, h_t)$.
if $\mathrm{mode} = \textsc{pres}$ **then**
  Ground task-relevant entities: $B_t \leftarrow \mathrm{GroundDINO}(o_t, x)$ for the robot arm, gripper, target object, and receptacle cues implied by x.
  Refine a merged foreground mask $M_t \leftarrow \mathrm{SAM2}(o_t, B_t)$ and construct $\tilde{o}_t^{\mathrm{sty}} \leftarrow M_t \odot o_t + (1 - M_t) \odot \psi_{\mathrm{bg}}(o_t)$ with deterministic background attenuation ψ_{bg}.
  Extract frozen mid/late visual features $v_t^{\mathrm{mid}}, \tilde{v}_t^{\mathrm{mid}}, v_t^{\mathrm{late}}, \tilde{v}_t^{\mathrm{late}}$ from the original and probe views.
  Query the frozen policy on the probe view: $a_t^{\mathrm{sty}} \leftarrow \pi_0(\tilde{o}_t^{\mathrm{sty}}, x, h_t)$.
  Compute $g_{\mathrm{vis}} \leftarrow \mathrm{clip}(\|v_t^{\mathrm{mid}} - \tilde{v}_t^{\mathrm{mid}}\|_2 / \kappa_{\mathrm{vis}}, 0, 1)$ and $g_{\mathrm{sem}} \leftarrow \exp(-\|v_t^{\mathrm{late}} - \tilde{v}_t^{\mathrm{late}}\|_2^2 / \kappa_{\mathrm{sem}}^2)$.
  Compute $\tau_t \leftarrow \tau_{\mathrm{phase}}(t, h_t)\, g_{\mathrm{vis}}\, g_{\mathrm{sem}}\, g_{\mathrm{act}}(a_t^{(0)}, a_t^{\mathrm{sty}})$.
  return $a_t^{\mathrm{deploy}} \leftarrow (1 - \alpha\tau_t) a_t^{(0)} + \alpha\tau_t a_t^{\mathrm{sty}}$
end if
Compute the frozen instruction feature $z_t \leftarrow z(x)$.
Compute the breaking-centered calibrated action $a_t^{\mathrm{br}} \leftarrow a_t^{(0)} + \lambda\, r_\theta(a_t^{(0)}, z_t)$.
if $\mathrm{mode} = \textsc{default}$ **then return** $a_t^{\mathrm{deploy}} \leftarrow a_t^{\mathrm{br}}$
Ground task-relevant entities, refine the foreground mask, and construct $\tilde{o}_t^{\mathrm{sty}}$ through the same ϕ_{sty} pipeline; then query $a_t^{\mathrm{sty}} \leftarrow \pi_0(\tilde{o}_t^{\mathrm{sty}}, x, h_t)$.
Extract frozen mid/late visual features from o_t and $\tilde{o}_t^{\mathrm{sty}}$, compute g_{vis} and g_{sem}, and then compute $\tau_t \leftarrow \tau_{\mathrm{phase}}(t, h_t)\, g_{\mathrm{vis}}\, g_{\mathrm{sem}}\, g_{\mathrm{act}}(a_t^{(0)}, a_t^{\mathrm{sty}})$.
return $a_t^{\mathrm{deploy}} \leftarrow (1 - \alpha\tau_t) a_t^{\mathrm{br}} + \alpha\tau_t a_t^{\mathrm{sty}}$

Proof. By definition of the hinge loss, $\mathcal{L}_{\mathrm{sep}} = 0$ implies that the hinge argument is non-positive:

$$m - \|a^{\mathrm{br}}(x^b) - a^{\mathrm{br}}(x^c)\|_2 + \|a^{\mathrm{br}}(x^u) - a^{\mathrm{br}}(x^c)\|_2 \le 0.$$

Rearranging terms yields

$$\|a^{\mathrm{br}}(x^b) - a^{\mathrm{br}}(x^c)\|_2 \ge \|a^{\mathrm{br}}(x^u) - a^{\mathrm{br}}(x^c)\|_2 + m,$$

which is the desired inequality. □

This proposition states the basic local implication of the triplet loss: the semantic-breaking action must be separated from the clean action by at least margin m beyond the clean-control discrepancy. The statement is local to the optimized action representation and does not by itself imply rollout-level correctness.

C.3 Residual Drift Bounds Around the Frozen Base Policy

The breaking-centered core uses residual calibration,

$$a_t^{\mathrm{br}} = a_t^{(0)} + \lambda\, r_\theta\Big(a_t^{(0)}, z(x)\Big),$$

where $a_t^{(0)}$ is the frozen base-policy action and r_θ is the learned residual.

Proposition 2 (Residual drift bound). Suppose $\|r_\theta(a, z)\|_2 \le \varepsilon$ on a region of interest. Then the calibrated action satisfies

$$\|a_t^{\mathrm{br}} - a_t^{(0)}\|_2 \le \lambda\varepsilon.$$

Proof. Starting from the residual parameterization,

$$\|a_t^{\mathrm{br}} - a_t^{(0)}\|_2 = \|\lambda r_\theta(a_t^{(0)}, z(x))\|_2 = \lambda \|r_\theta(a_t^{(0)}, z(x))\|_2 \le \lambda\varepsilon.$$

This proves the claim. □

The bound shows that residual calibration stays within a radius controlled jointly by the residual magnitude and the scale parameter λ. An immediate consequence is that whenever $r_\theta(a_t^{(0)}, z(x)) = 0$ on a subset of states, BAS-VLA reduces exactly to the frozen base policy on that subset.

C.4 Gate and Stability Properties of the Preserving Auxiliary

The preserving auxiliary uses the evidence gate

$$\tau_t = \tau_{\text{phase}}(t, h_t)\, g_{\text{vis}}(v_t^{\text{mid}}, \tilde{v}_t^{\text{mid}})\, g_{\text{sem}}(v_t^{\text{late}}, \tilde{v}_t^{\text{late}}, x)\, g_{\text{act}}(a_t^{(0)}, a_t^{\text{sty}}),$$

with each factor in $[0, 1]$. In the current parameterization,

$$g_{\text{vis}}(v, \tilde{v}) = \text{clip}\left(\frac{\|v - \tilde{v}\|_2}{\kappa_{\text{vis}}},\, 0,\, 1\right), \qquad g_{\text{sem}}(v, \tilde{v}, x) = \exp\left(-\frac{\|v - \tilde{v}\|_2^2}{\kappa_{\text{sem}}^2}\right),$$

where the first term is evaluated on a frozen mid-level visual representation and the second on a frozen later semantic visual representation from the same backbone. Under the concrete probe instantiation above, these features are extracted from the original image and from the grounded-mask probe image produced by deterministic background attenuation outside the retained foreground mask. As in the main text, the subscript "sty" denotes generic observation-side visual variation rather than only the validated style-shift slice. The preserving-only weak-fusion rule is

$$a_t^{\text{pres}} = (1 - \alpha\tau_t)a_t^{(0)} + \alpha\tau_t a_t^{\text{sty}}, \qquad \alpha \in (0, 1).$$

The full composition retained for ablation applies the same gate and fusion weight on top of the breaking-centered action,

$$a_t^{\text{full}} = (1 - \alpha\tau_t)a_t^{\text{br}} + \alpha\tau_t a_t^{\text{sty}}.$$

Proposition 3 (Gate range and identity cases). If $\tau_{\text{phase}}, g_{\text{vis}}, g_{\text{sem}}, g_{\text{act}} \in [0, 1]$, then $\tau_t \in [0, 1]$. Moreover,

$$\tau_t = 0 \Longrightarrow a_t^{\text{pres}} = a_t^{(0)} \quad \text{and} \quad a_t^{\text{full}} = a_t^{\text{br}}.$$

Proof. The product of numbers in $[0, 1]$ remains in $[0, 1]$, hence $\tau_t \in [0, 1]$. If $\tau_t = 0$, then

$$a_t^{\text{pres}} = (1 - \alpha \cdot 0)a_t^{(0)} + \alpha \cdot 0\, a_t^{\text{sty}} = a_t^{(0)}.$$

Applying the same substitution to the full-composition rule yields

$$a_t^{\text{full}} = (1 - \alpha \cdot 0)a_t^{\text{br}} + \alpha \cdot 0\, a_t^{\text{sty}} = a_t^{\text{br}}.$$

□

Proposition 4 (Gate-modulated drift identities). For every step,

$$a_t^{\text{pres}} - a_t^{(0)} = \alpha\tau_t\big(a_t^{\text{sty}} - a_t^{(0)}\big),$$

and therefore

$$\|a_t^{\text{pres}} - a_t^{(0)}\|_2 = \alpha\tau_t\|a_t^{\text{sty}} - a_t^{(0)}\|_2.$$

In the full-composition mode,

$$a_t^{\text{full}} - a_t^{\text{br}} = \alpha\tau_t\big(a_t^{\text{sty}} - a_t^{\text{br}}\big),$$

and therefore

$$\|a_t^{\text{full}} - a_t^{\text{br}}\|_2 = \alpha\tau_t\|a_t^{\text{sty}} - a_t^{\text{br}}\|_2.$$

Proof. Subtracting $a_t^{(0)}$ from the preserving-only weak-fusion equation gives

$$a_t^{\text{pres}} - a_t^{(0)} == \big((1-\alpha\tau_t) - 1\big)a_t^{(0)} + \alpha\tau_t a_t^{\text{sty}} = \alpha\tau_t\big(a_t^{\text{sty}} - a_t^{(0)}\big),$$

and taking the ℓ_2 norm yields

$$\|a_t^{\text{pres}} - a_t^{(0)}\|_2 = \|\alpha\tau_t(a_t^{\text{sty}} - a_t^{(0)})\|_2 = \alpha\tau_t\|a_t^{\text{sty}} - a_t^{(0)}\|_2,$$

because $\alpha\tau_t \geq 0$ is scalar. The full-composition identities follow by the same algebra after replacing the anchor action $a_t^{(0)}$ with a_t^{br}. □

Two immediate consequences are useful. First, on every active step,

$$a_t^{\text{pres}} - a_t^{\text{sty}} = (1-\alpha\tau_t)\big(a_t^{(0)} - a_t^{\text{sty}}\big),$$

and therefore

$$\|a_t^{\text{pres}} - a_t^{\text{sty}}\|_2 = (1-\alpha\tau_t)\,\|a_t^{(0)} - a_t^{\text{sty}}\|_2.$$

Together, these identities show that the preserving auxiliary remains on the line segment between its anchor action and the probe action, with displacement scaled continuously by τ_t. The anchor is $a_t^{(0)}$ in preserving-only deployment and a_t^{br} in the full composition.

C.5 Deployment-Mode Complexity and Identity Relations

This work uses three deployment modes:

$$a_t^{\text{default}} = a_t^{\text{br}}, \qquad a_t^{\text{pres}} = (1-\alpha\tau_t)a_t^{(0)} + \alpha\tau_t a_t^{\text{sty}}, \qquad a_t^{\text{full}} = (1-\alpha\tau_t)a_t^{\text{br}} + \alpha\tau_t a_t^{\text{sty}}.$$

Let C_{π_0} denote the cost of one frozen base-policy query, let C_r denote the cost of one residual forward pass through r_θ, and let C_g collect the trigger/fusion overhead that does not invoke an additional base-policy query.

Proposition 5 (Coarse-grained per-mode overhead). Measured relative to a single frozen base-policy query, and under the convention that the mode-selection logic is absorbed into C_g, the additional inference cost of the three deployment modes satisfies

$$\Delta C_{\text{default}} = C_r + C_g, \qquad \Delta C_{\text{pres}} = C_{\pi_0} + C_g, \qquad \Delta C_{\text{full}} \geq C_r + C_g,$$

where the preserving-side gate changes the fused action continuously but does not remove the need for the auxiliary probe query.

Proof. The default mode adds one residual pass and the mode-selection overhead on top of the frozen-policy output, hence $\Delta C_{\text{default}} = C_r + C_g$. The preserving sidecar computes the probe action together with the gate and fusion overhead, hence $\Delta C_{\text{pres}} = C_{\pi_0} + C_g$. The full mode contains the residual pass and the full-composition overhead and is therefore bounded below by $C_r + C_g$. □

The current parameterization also contains several exact identity relations:

$$r_\theta(\cdot,\cdot) \equiv 0 \Longrightarrow a_t^{\text{br}} = a_t^{(0)},$$

$$\tau_t = 0 \Longrightarrow a_t^{\text{pres}} = a_t^{(0)} \quad \text{and} \quad a_t^{\text{full}} = a_t^{\text{br}},$$

$$\alpha = 0 \Longrightarrow a_t^{\text{pres}} = a_t^{(0)} \quad \text{and} \quad a_t^{\text{full}} = a_t^{\text{br}} \quad \text{even on triggered steps.}$$

These identities clarify that BAS-VLA remains a calibration layer around the frozen policy and that exact recovery of the base action is contained in the parameterization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and introduction state a breaking-centered task-semantic calibration claim together with a selective preserving-side claim, and this scope matches the experimental and formal evidence reported in the main text and appendices. See the abstract, Sections 1, 4, and 5, and the appendix summaries.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper includes an explicit limitations discussion covering the narrower supported regime, the selective preserving auxiliary, the non-default status of the full composition, and the boundary role of supplementary benchmarks. See Appendix B.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The formal properties include numbered propositions, explicit assumptions, and complete proofs. See Appendices C.2–C.5.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Sections 3–5 and Appendix A.2.1 describe the calibration method, controlled intervention protocols, carriers, metrics, and rollout budgets; the public repository provides training and evaluation code for selected settings. Complete carrier-matched training inputs and evaluation manifests for all main results are not included in the release.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The project website (https://nebulis-lab.com/BAS-VLA) links to code for selected paper settings (https://github.com/NEBULIS-Lab/BAS-VLA) and the OpenVLA/LIBERO-10 offline adapter and matched records (https://huggingface.co/Shuaijun/BAS-VLA). The main online evaluations additionally require carrier-specific weights and task manifests, and real-robot evaluation requires a configured hardware backend and calibration.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [No]

Justification: Sections 3–5 and Appendix A.2.1 specify deployment modes, evaluation metrics, seed counts, rollout budgets, and task settings, while the Hugging Face package documents the training configuration and data splits for the released offline adapter. Complete training configurations and hyperparameter-selection procedures for all main-experiment carriers are not included.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports Wilson 95% confidence intervals for principal binary outcomes, computed from aggregate episode counts. See Table 16 and the accompanying uncertainty discussion.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Table 9 reports GPU type and memory and the robot hardware, while Tables 10 and 12 report deployment timing. Total project compute and training/evaluation time for every run are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work studies task-semantic action calibration through controlled simulation and real-robot manipulation experiments without human subjects. We have reviewed the NeurIPS Code of Ethics and are not aware of a deviation from it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on task-semantic action calibration and its empirical evaluation; it does not include an explicit discussion of positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released assets comprise calibration code, a lightweight residual adapter, and simulation-derived records; they do not introduce high-risk generative models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the benchmark suites and policy carriers, and the supplementary appendix identifies the licenses of BAS-VLA, OpenPI, SAM 2, LIBERO, and OpenVLA-OFT. The released packages document third-party sources and retain the applicable license notices.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public code and Hugging Face release linked above document the release scope, dependencies, training configuration, action conventions, data alignment, and usage alongside the code, residual adapter, and matched LIBERO records. The packages include the BAS-VLA MIT license and the applicable upstream notices.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human subjects or crowdsourcing experiments, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the core BAS-VLA method or of the experimental methodology reported in this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.