DIAGNOSING THE STRUCTURE AND LOCAL BEHAVIORAL EFFECTS OF VIEWPOINT-INDUCED ACTION DISTORTION IN ROBOT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Changing the external camera position can alter a visual robot policy's control commands even when the task, robot configuration, and action coordinate frame remain fixed. However, whole-episode success alone cannot separate this immediate viewpoint response from state drift accumulated after execution. To isolate the response, we query frozen policies at identical physical and policy states before execution, changing only the external-camera extrinsics. We then compare geometric, affine, information-matched linear, and conditional nonlinear mappings to characterize the resulting action differences. Bidirectional one-step action substitutions test the local behavioral effects of these differences. Our evaluation covers multiple policies, with detailed diagnosis of OpenVLA and Fast-WAM, spanning autoregressive vision-language-action and video-based world-action architectures, and complementary viewpoint evaluations on ACT and π_0. On an independent OpenVLA primary test set, a task-specific affine map raises shifted-view success from 14.0% to 49.2%, and a conditional predictor further reaches 62.4%, still below the familiar-view reference of 82.0%. At pre-contact states, injecting one shifted-view action reduces success under familiar-view continuation by 15.2 percentage points. Conversely, reverse-inserting one familiar-view action improves shifted-view continuation by 10.6 percentage points. Independent evaluation on Fast-WAM also shows one-step intervention effects and partial recovery, supporting recurrence across these two architectures. These results reveal that viewpoint changes can induce predictable action distortion before trajectories diverge, and that such changes at pre-contact interactions alter subsequent success. Our work jointly diagnoses this output structure and its behavioral effects, and provides direct evidence for locating the source of viewpoint failure and identifying the effective range of calibration.

1 INTRODUCTION

Why can moving the camera cause a robot policy to fail when the robot already knows what to manipulate and the objects have not moved? This problem is common in visual robot control. Vision-language-action models map images, language, and robot states to control commands. A camera change alters image locations, occlusion, and visual geometry. Existing benchmarks show that visual perturbations, particularly viewpoint changes, can substantially affect manipulation success (Liu et al., 2023; Pumacay et al., 2024; Fei et al., 2026). View synthesis, observation adaptation, and consistency learning continue to improve robustness (Tian et al., 2025; Heo et al., 2026; Huang et al., 2026; Gu et al., 2026). These advances motivate a more specific question. At identical physical and policy states, how does a camera change alter the selected action? Is that change sufficient to affect subsequent behavior?

Whole-episode outcomes cannot answer this question alone. Once the first action is executed, the robot configuration, contact state, and next observation can all change. Subsequent errors combine the response to the current view with state-distribution changes and control feedback caused by earlier actions. Analyses of distribution shift in imitation learning describe this accumulation (Ross et al., 2011). The same failure rate can therefore arise through different mechanisms. Actions may be systematically displaced from the outset, or execution may become sensitive near contact. Feedback

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

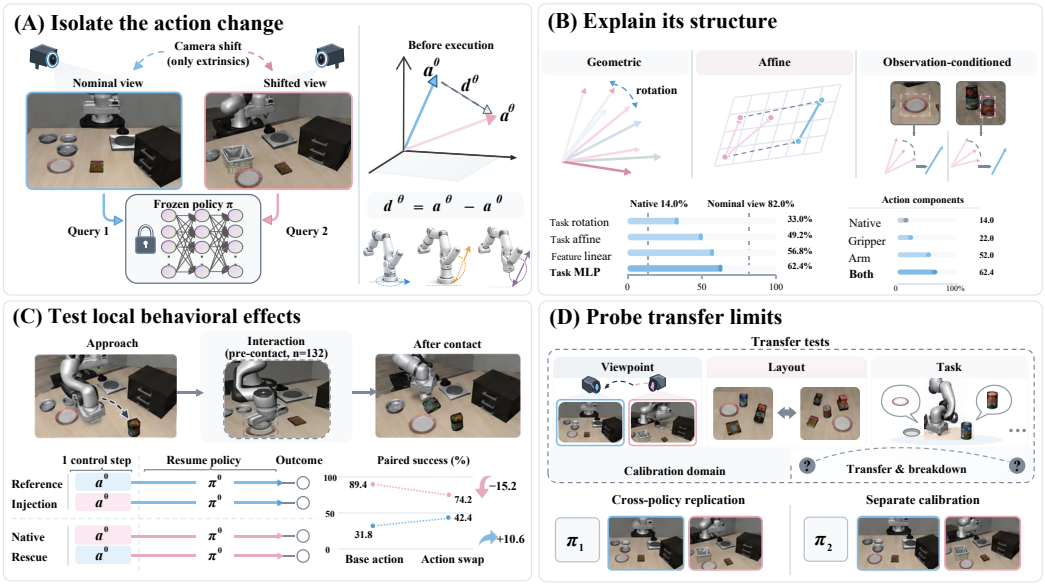


Figure 1: Overview of the diagnostic. (A) Independent frozen-policy queries at identical restored states change only external-camera extrinsics. (B) Geometric and observation-conditioned mappings characterize action displacement. (C) Finite injection and rescue test local effects and phase dependence. (D) Independent views, layouts, and tasks assess transfer; separate policy calibration assesses repeatability. Scene and action graphics illustrate the protocol.

may absorb a large free-space error, while a smaller pre-contact error may determine whether the robot grasps its target. Action differences, their predictability, and their relationship to failure must be examined separately.

We construct matched-state view pairs by restoring simulator and policy snapshots. Before either action is executed, output differences induced solely by external-camera extrinsics are measured. We then fit action mappings of increasing complexity and apply their predicted corrections in closed-loop execution, thereby testing whether predictable output structure translates into behavioral recovery.

Behavioral validation is performed through bidirectional, finite action interventions. At a restored state, a single control action is replaced, and execution then resumes under the prescribed continuation policy. Phase comparisons and magnitude-matched direction controls identify the conditions under which the effect emerges. Independent-distribution evaluations then probe how far calibration transfers. Figure 1 summarizes the overall diagnostic pipeline.

Our contributions are threefold:

- We establish a pre-execution action diagnostic that restores both physical and policy states, separating camera interventions from subsequent trajectory feedback.
- We connect predictable action differences to local behavioral effects through progressively stronger structural controls and bidirectional finite substitutions, distinguishing the roles of action direction, component, and interaction phase.
- We evaluate the repeatability of these findings and the scope of calibration transfer on independent test sets, different task suites, and two action interfaces.

2 Related Work

Viewpoint robustness and evaluation. LIBERO studies knowledge transfer across manipulation suites with varying spatial relations, objects, and goals (Liu et al., 2023). THE COLOSSEUM systematically evaluates generalization under changes to cameras, backgrounds, and object attributes (Pumacay et al., 2024). LIBERO-Plus adds controlled perturbations to vision-language-action

evaluation (Fei et al., 2026). These benchmarks primarily characterize robustness through whole-episode outcomes. We retain task success as the behavioral endpoint and add paired pre-execution measurements and local interventions under fixed continuation policies.

Cross-view learning and adaptation. VISTA introduces visual diversity through novel-view synthesis (Tian et al., 2025). AnyCamVLA adapts observations to improve frozen-policy execution under camera changes (Heo et al., 2026). VistaBot combines geometric estimation, video generation, and action representations for novel-view control (Gu et al., 2026). Closest to our measurement setting, Cross-View Action Consistency uses same-state view pairs and an action-consistency objective for policy learning (Huang et al., 2026). We use these pairs to analyze frozen-policy outputs, compare geometric, linear, and conditional nonlinear explanations, and test behavioral relevance through finite action substitutions. Equivariant Diffusion Policy and EquiBot build geometric symmetries into policy learning (Wang et al., 2025; Yang et al., 2025), whereas our camera-only interventions diagnose frozen policies in a fixed action frame.

Interventional policy diagnosis. Observational correlations alone may not identify which task-relevant factors an imitation policy uses; controlled interventions help distinguish explanations (de Haan et al., 2019). Embodied Interpretability studies causal understanding and generalization through visual interventions and policy responses (Zhang et al., 2026). Internal activation interventions also analyze and improve world action model robustness (Hong et al., 2026). We intervene on executed actions at the controller interface and measure their consequences in paired experiments with fixed continuation policies. OpenVLA and Fast-WAM provide stepwise and chunked action interfaces, respectively, for testing this diagnostic under different action-generation mechanisms (Kim et al., 2025b; Yuan et al., 2026).

3 Matched-State Action Diagnosis Before Execution

Controlled view pairs and restoration. Let s denote physical state excluding camera extrinsics, q proprioception, c the instruction or known task condition, and η policy history, random-number state, and queued actions. External-camera extrinsics κ_θ give

$$o^\theta = \mathcal{R}(s; \kappa_\theta), \qquad a^\theta = \pi_\psi(o^\theta, q, c; \eta), \qquad \psi \text{ frozen.} \tag{1}$$

Queries independently restore (s, q, c, η) from one snapshot and refresh their observations before inference. Intrinsics, lighting, wrist observations, action decoding, and controller settings remain fixed; stochastic policies use coupled random draws.

Audits cover generalized positions/velocities, actuators, object/contact states, simulation time, task/controller states, random generators, history, and queues. Same-view repetitions check restoration and inference consistency (Section 7.1).

Displacement and physical components. In decoded action coordinates,

$$d_\theta(s, c) = a^\theta - a^0, \qquad r_\theta(s, c) = a^0 - a^\theta. \tag{2}$$

Here a^0 is the familiar-view reference, not assumed correct or optimal, and r_θ a cancellation direction for interventions.

The control action $a = (v, Q, b)$ comprises translation increment $v \in \mathbb{R}^3$, rotation increment $Q \in \mathrm{SO}(3)$, and gripper command b, with

$$d_v = v^\theta - v^0, \qquad d_Q = \left[\log((Q^0)^\top Q^\theta)\right]^\vee, \qquad d_b = \mathbf{1}[b^\theta \neq b^0]. \tag{3}$$

We report translation magnitudes in millimeters, rotations in degrees, and gripper disagreement rates separately, without mixing physical units.

State dependence and geometry. A fixed rotation $v^\theta = Rv^0$ gives $d_v = (R - I)v^0$: state-independent R produces action-dependent displacement. Thus, state dependence alone does not establish a need for complex conditioning. We test geometric and affine explanations before observation conditioning, then assess behavioral utility through interventions.

4 Explaining the Structure of Action Displacement

Geometric mappings and information-matched controls. Mappings share paired calibration data and familiar-view offline targets. Translation models comprise task-specific constant offsets, rotations, scaled rotations, and affine mappings:

$$f_{\text{const}}(v^\theta, c) = v^\theta + b_c, \qquad f_{\text{rot}}(v^\theta, c) = R_c v^\theta,$$
$$f_{\text{scale}}(v^\theta, c) = \alpha_c R_c v^\theta, \quad f_{\text{affine}}(v^\theta, c) = A_c v^\theta + b_c. \tag{4}$$

Orthogonal Procrustes (Schönemann, 1966) fits $R_c \in \mathrm{SO}(3)$; scales are nonnegative. Affine parameters are shared or task-specific. The robot control frame stays fixed: these are empirical output mappings, not required coordinate transformations.

Rotation residuals use a common local parameterization and rotation composition. Grippers are separate discrete decisions. Rotation and scaled-rotation models retain original gripper commands; gripper-predicting models report agreement.

To separate input-information gains from nonlinearity, an information-matched linear model receives the conditional predictor's visual features, proprioception, original action, and task identity, with task-specific parameters allowed.

Conditional and direct predictors. The conditional predictor uses shifted-view inputs:

$$u^\theta = [e^\theta, q, a^\theta], \qquad \hat{r} = g_\phi(u^\theta, e_c), \qquad \tilde{a} = \mathcal{C}(a^\theta, \hat{r}). \tag{5}$$

Here e^θ is a frozen visual feature, e_c known task identity, and $\mathcal{C}$ combines predictions using controller conventions. Visual PCA (64 dimensions) and input standardization are fitted only on training data. The predictor is a two-layer, width-128 GELU MLP.

A direct familiar-view action predictor retains identical inputs, hidden layers, optimizer, and data budget, removing only the residual skip connection. Learned calibrators execute using shifted-view inputs only; familiar-view outputs serve as offline targets.

5 Finite Action Interventions and Local Behavioral Effects

Bidirectional injection and rescue. Full-episode replacement would follow the reference policy, obscuring individual effects. We replace finite control steps, then resume the designated policy. Let $Y_s(a; \pi^v)$ be binary success after action a from restored state s, followed by frozen-policy continuation under view v:

$$\tau_{\text{inject}} = \mathbb{E}_s\left[Y_s(a^\theta; \pi^0) - Y_s(a^0; \pi^0)\right], \tag{6}$$
$$\tau_{\text{rescue}} = \mathbb{E}_s\left[Y_s(a^0; \pi^\theta) - Y_s(a^\theta; \pi^\theta)\right]. \tag{7}$$

Injection substitutes a shifted-view first action into familiar-view continuation (Reference); Rescue substitutes a familiar-view first action into shifted-view continuation (Native). Negative injection indicates harm; positive rescue indicates improvement. Expectations use each experiment's sampled states. Pairs fix restored state, task condition, policy state, and continuation policy. Branches then follow their realized states, including downstream feedback in the effect.

Interventions last one or four actual control steps (0.05/0.20 s at 20 Hz). For four steps, OpenVLA re-queries at each branch state; Fast-WAM uses an independently queried other-view chunk's first four slots, then resumes the designated queue.

Policy history and action queues. At decision boundaries, both views are queried from cloned policy states; output sequences and post-query states are saved. Reference and Injection share the familiar-view continuation state and queue index, while Native and Rescue share the corresponding shifted-view state and queue index. Replacements consume normal slots; executed-action histories record actual commands. Other-view queries neither overwrite continuation history nor cause condition-specific replanning.

We log both the actions proposed by the policy and the commands actually executed, together with any clipping, execution duration, and the source episode. We also verify that matched conditions follow the same rules for queue advancement and replanning, ensuring that action replacements consume the intended queue slots without triggering condition-specific replanning.

Phase and direction controls. We sample approach, pre-contact, and post-contact states along fixed familiar-view roll-ins. Each task uses a designated grasp or placement event: approach is a decision boundary after entering the interaction-distance band; pre-contact is the last available boundary in a narrow proximity band before the event; post-contact is the first eligible boundary after the event and before termination. Geometry and events define phases, including both ultimately successful and failed source episodes.

Direction controls affect translation only. For nonzero predicted correction $\hat{r}_v$, let $u = \hat{r}_v/\|\hat{r}_v\|$. We compare u, $-u$, and three unit directions w_j in its orthogonal plane, with $u^\top w_j = 0$ and $\|w_j\| = 1$; the w_j need not be mutually orthogonal. Zero residuals remain in the evaluation as zero-magnitude interventions. All directions at a state share the feasible magnitude

$$\alpha(s) = \min\left\{\alpha_{\text{val}},\ \min_{z\in\mathcal{Z}} \sup\{\alpha \geq 0 : v^\theta + \alpha z \in \mathcal{A}_v\}\right\}, \qquad \mathcal{Z} = \{u, -u, w_1, w_2, w_3\}. \tag{8}$$

The controller executes $v^\theta + \alpha(s)z$, retaining Native rotation and gripper commands. Matching feasible magnitudes at the control interface avoids direction-dependent magnitudes after clipping.

6 Experimental Setup

Policies and tasks. Our primary experiments evaluate frozen OpenVLA and Fast-WAM on LIBERO. OpenVLA's primary test uses Spatial tasks 0, 2, 3, 6, and 7, while Fast-WAM's main test covers ten tasks. The primary comparison is familiar view 0° versus shifted view $+30^\circ$, with separate paired calibration data for each policy.

To broaden our empirical validation, we independently evaluate ACT (Zhao et al., 2023) and π_0 (Black et al., 2025) on eight RLBench tasks (James et al., 2020), comparing their baseline configurations with their VistaBot-enhanced variants (Gu et al., 2026) (Figure 4(g,h)). These experiments further assess viewpoint sensitivity and performance recovery across policy architectures, complementing the matched-state diagnosis and finite action interventions on OpenVLA and Fast-WAM.

We partition by source episode to separate fitting, output measurement, and behavioral testing. Structural-test snapshots, main-test initial states, and finite-intervention source episodes are distinct (Table 1). Calibration mixes familiar- and shifted-view roll-ins per task, including failed trajectories, reducing bias from familiar-view-only coverage. These calibration splits are separate from potential exposure in base-model pretraining. Recent benchmark analysis likewise emphasizes initialization, data exposure, and evaluation protocols (Jiang et al., 2026).

Table 1: Main data splits. Episodes or initial states are the sampling units for each role. OpenVLA's structural test contains 22 snapshots from each of 100 source episodes (2,200 paired states). Fast-WAM's local interventions use 200 pre-contact decision states.

Data role	OpenVLA	Fast-WAM
Calibration training episodes	100	200
Calibration validation episodes	50	100
Independent structural-test episodes	100	—
Main-test initial states	500	500
Local-intervention source episodes	150	200
Main-test tasks	5	10

Metrics and model selection. Offline translation and rotation errors are reported as NMSE, each normalized by the corresponding uncalibrated error, with gripper agreement reported separately to avoid conflating continuous and discrete components. Closed-loop evaluation compares full-episode success from independent initial states, separating offline fit from behavioral utility. For fair comparison, all mappings share source-episode splits and validation budgets, task-conditioned models receive matched task information, and the information-matched linear model and conditional predictor receive identical observation inputs. Checkpoints and hyperparameters are selected on validation data only. The main conditional predictor is fitted with three optimization seeds on the same environment starts to characterize optimization sensitivity.

Table 2: OpenVLA offline fit on independent structural snapshots and closed-loop success on 500 starts per policy. NMSE is normalized by Native component error. Success cells give count (percent). Policies are calibrated independently and cover different tasks; methods within each policy share starts. Fast-WAM dashes denote unevaluated mappings, not failures.

	OpenVLA NMSE ↓		OpenVLA	Success /500 (%) ↑	
Mapping or condition	Trans.	Rot.	Grip agree. (%) ↑	OpenVLA	Fast-WAM
Native	1.00	1.00	85.1	70 (14.0)	300 (60.0)
Task constant	0.92	0.95	86.2	90 (18.0)	—
Task rotation	0.66	0.79	85.1	165 (33.0)	—
Task scaled rotation	0.59	0.72	85.1	181 (36.2)	—
Shared affine	0.49	0.60	89.4	215 (43.0)	—
Task affine	0.42	0.52	90.2	246 (49.2)	350 (70.0)
Task feature-linear	0.35	0.44	92.1	284 (56.8)	385 (77.0)
Shared MLP	0.36	0.43	91.7	236 (47.2)	—
Conditional predictor	0.29	0.36	94.0	312 (62.4)	395 (79.0)
Direct action predictor	0.30	0.37	93.8	304 (60.8)	—
Familiar-view reference	—	—	—	410 (82.0)	455 (91.0)

Statistical units and independent evaluation. For method m and Native at the same initial state, let $D_i = Y_{i,m} - Y_{i,0}$. Mean gain across tasks is

$$\hat{\Delta} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} D_i, \tag{9}$$

where K is the number of tasks, I_k indexes task k's starts, and $n_k = |I_k|$. Equal per-task sample sizes make main-test macro-averages equal pooled success rates. Main-test gains use 10,000 paired episode bootstrap replicates preserving task strata, so that uncertainty is estimated within tasks rather than dominated by task-level variation. Local experiments cluster by source episode, jointly resampling its phases and repeated directions to respect their dependence. The three orthogonal directions are aggregated within each state before state-level comparisons, avoiding treating repeated directions as independent. Main tables report the validation-selected primary checkpoint, and Appendix A.5 reports optimization-seed sensitivity. Percentage points denote success-rate differences, and bracketed intervals are 95% confidence intervals. Primary comparisons are conditional versus affine and information-matched linear predictors, pre-contact single-step injection and rescue, and single-step replication in the second policy. Supplementary action measurements use 20,000 source-episode bootstrap replicates under their own estimation settings, as they involve different sampling units.

7 Results

7.1 Viewpoint sensitivity depends on direction and task

Viewpoint sensitivity depends on task and direction: OpenVLA exhibits abrupt failures and asymmetric responses to signed camera shifts, while Fast-WAM varies non-monotonically across azimuths (Figure 4(a–f)). ACT and π_0 also benefit from observation-side view restoration on RLBench (Figure 4(g,h); Appendix C), providing complementary outcome-level evidence of recoverability.

Matched-state measurements locate sensitivity before execution. Across 2,200 structural-test snapshots, mean translation and rotation discrepancies are 11.68 mm and 3.03°, with 14.9% gripper disagreement. A separate restoration audit finds zero non-camera state difference alongside task-dependent signed action shifts. In 75 rollout pairs selected for familiar-view success and shifted-view failure, the recorded divergence threshold is crossed during approach, often before object motion (Appendix D.2). This localizes divergence within the selected pairs; Section 7.3 tests behavioral sensitivity by phase.

7.2 Low-complexity maps recover substantial performance

Geometric and affine maps recover substantial performance beyond task-constant correction (Table 2; Figure 2). The conditional predictor adds 13.2 percentage points (pp) $[8.8, 17.6]$ over task affine, but 5.6 pp $[1.2, 9.8]$ over the information-matched feature-linear model: additional inputs explain part of the gain, with a remaining nonlinear advantage. Recovery is incomplete, and residual prediction has no clear advantage over direct action prediction. Supplementary CNN-based controls show the value of known task identity and a mismatch between offline error and behavioral utility (Appendices A.3 and D.3).

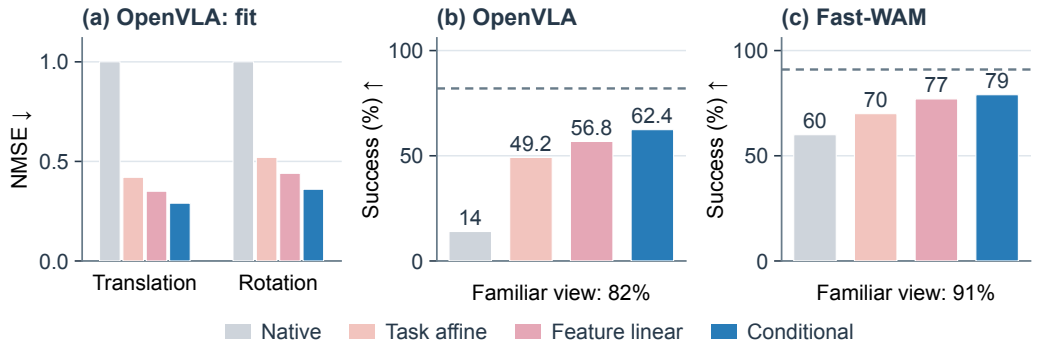


Figure 2: Structure and closed-loop recovery. (a) OpenVLA translation and rotation NMSE on the independent structural test. (b,c) Success on the independent 500-start tests, with separate calibration per policy. Dashed lines mark familiar-view references. Bars show point estimates; Table 2 reports all mappings.

7.3 One action substitution changes subsequent behavior

A single pre-contact action substitution harms familiar-view continuation or improves shifted-view continuation, depending on its direction (Figure 3(a,b)). Effects are smaller during approach and after contact, and four-step substitutions strengthen the pre-contact effects. Thus, action differences can alter subsequent success, with sensitivity concentrated near interaction. Phase comparisons remain descriptive because reachable-state coverage differs (Appendix B.3).

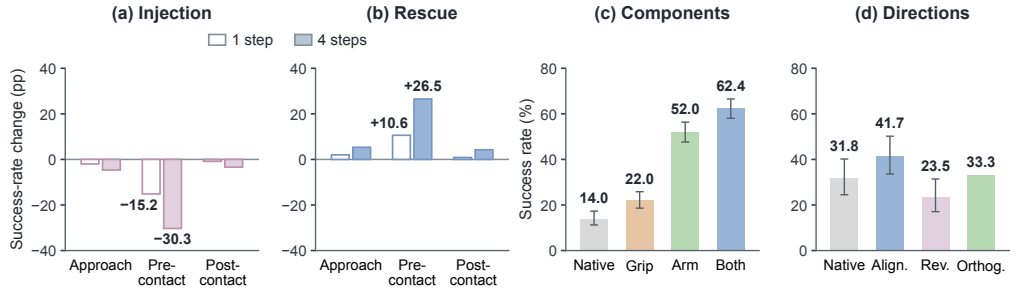


Figure 3: Phase, component, and direction effects. (a,b) One-step (open) and four-step (filled) injection and rescue gains against matched continuation controls (Table A7); phase denominators differ. (c) Gripper, arm, or joint full-rollout correction on 500 shared starts. (d) One-step translation on 132 pre-contact states with equal feasible executed magnitudes; rotation and gripper remain Native. Orthog. averages three directions. Panels (c,d) show absolute success. Whiskers are marginal 95% Wilson intervals, not paired-gain intervals; none is shown for the orthogonal mean.

7.4 Translation direction and action components affect recovery

Arm and gripper correction both contribute, with joint correction yielding the highest recovery (Figure 3(c)). In the separate 132-state single-step test, aligned translation outperforms reversed and orthogonal controls at matched final executed magnitudes, while rotation and gripper remain

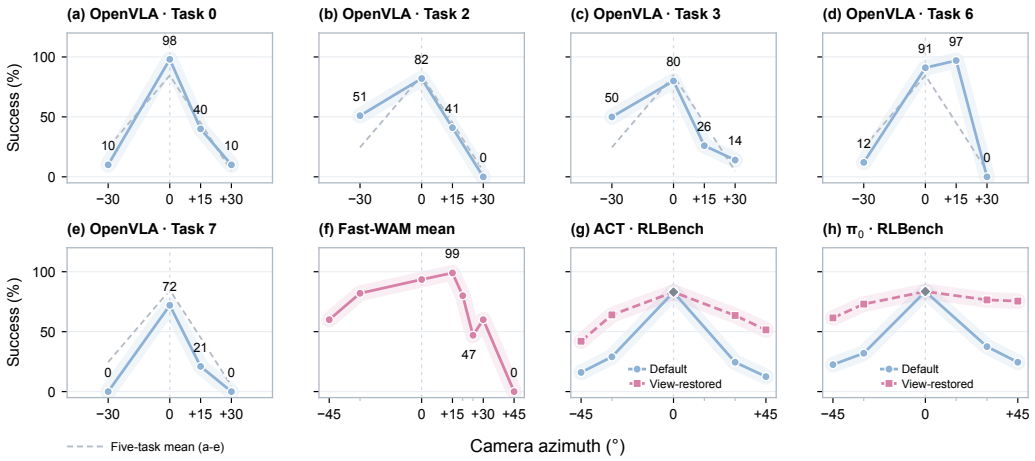


Figure 4: Task- and direction-dependent viewpoint sensitivity. (a–e) OpenVLA success on Spatial tasks 0, 2, 3, 6, and 7, respectively, with 100 rollouts per task–azimuth pair. (f) Fast-WAM success macro-averaged over ten tasks at eight camera azimuths. These LIBERO scans are separate from the primary calibration tests (Appendix A.1). (g,h) Our independent evaluations of ACT and π_0 on eight RLBench tasks, following the experimental protocol of Gu et al. (2026), with 25 evaluations per task at each novel azimuth. Default and VistaBot denote the baseline policies and their VistaBot-enhanced variants, respectively. Both curves share the clean 0° reference within each policy (Appendix C).

Native (Figure 3(d)). Its advantage over the mean orthogonal control is 8.3 pp $[3.3, 13.6]$, supporting behavioral value of the predicted direction beyond intervention magnitude (Appendix B.4).

7.5 Partial recovery and single-step effects recur across policies

With separate calibration, Fast-WAM recovers 19.0 pp $[15.0, 23.0]$ over Native. In this interface, feature-linear and nonlinear predictors perform comparably (2.0 pp $[-1.4, 5.4]$), indicating that nonlinear structure is not universally required for recovery. At 200 additional pre-contact states, one-step injection changes familiar-view success by -12.0 pp $[-17.0, -7.0]$, while rescue improves shifted-view success by 12.0 pp $[5.5, 18.5]$. Partial recovery and local action effects therefore recur across both policy interfaces, while the benefit of nonlinear mapping is itself interface-dependent.

7.6 Which distributions support calibration transfer?

Calibration at $+30^\circ$ transfers to the unseen $+15^\circ$ view and new object layouts, while corrected success remains low at $+45^\circ$ (Table 3). Opposite-direction transfer and leave-one-task-out gains remain limited. The latter uses a shared predictor trained on the remaining tasks, since an unseen task has no trained task embedding.

Transfer to new layouts shows that recovery extends beyond the exact object configurations used for calibration within the evaluated tasks. The contrast between the two unseen camera angles reveals a limited range of effective correction, with larger viewpoint changes remaining difficult under the same calibration. Cross-task transfer places an additional demand on the predictor, which must capture action distortions associated with behaviors absent from its training tasks. These distinctions matter because generalization across scene configurations, camera angles, and tasks probes different aspects of the learned mapping. Recovery is therefore governed by the calibration distribution.

7.7 Sustained directional interventions and suite replication

Recovery recurs across independent Spatial, Object, and Goal banks, with the Object suite showing substantial sensitivity to optimization seed (Table 4). We next isolate the contribution of direct visual input by removing the auxiliary predictor's RGB branch, and Spatial performance shows no clear degradation. This result reflects the predictor's role as a correction module on top of a visual base

Table 3: Calibration transfer to independent distributions. View and layout tests each use 200 new starts and the task-conditional predictor calibrated at $+30^\circ$. Leave-one-task-out evaluation uses 500 starts and shared predictors trained on the remaining tasks. Gains and intervals are in percentage points.

Test condition	Starts	Native (%)	Calibrated (%)	Gain [95% CI]
Unseen $+15^\circ$	200	52.0	72.0	$+20.0\ [13.5, 26.5]$
Unseen $+45^\circ$	200	6.0	25.0	$+19.0\ [13.0, 25.0]$
Unseen -30°	200	24.0	19.0	$-5.0\ [-10.5, 0.5]$
New object layouts	200	16.0	54.0	$+38.0\ [31.0, 45.0]$
Leave-one-task-out	500	14.0	16.2	$+2.2\ [-1.0, 5.4]$

Table 4: Independent-state and suite replications. Each row reuses 100 starts across three optimization seeds; the final column gives mean success and sample standard deviation.

Suite / condition	Native /100	Calibrated /100	Mean $\pm$ SD (%)
Spatial, task-conditional	2	60, 57, 54	57.0 ± 3.0
Spatial, no predictor RGB	2	66, 58, 63	62.3 ± 4.0
Object, task-conditional	0	72, 40, 47	53.0 ± 16.8
Goal, task-conditional	7	59, 59, 48	55.3 ± 6.4

policy, in that the base policy already consumes images and its Native action can therefore carry visual information.

8 Discussion

From action consistency to behavioral effectiveness. Closer agreement with familiar-view outputs need not yield proportional behavioral gains, because task success also depends on when action errors occur, their direction, and the feedback that follows. Offline measurements and behavioral interventions therefore provide complementary evidence: the former characterize predictable output differences, while finite substitutions test how those differences affect subsequent execution. Continuous calibration then assesses whether correcting this structure throughout a rollout produces sustained recovery. Since the predictor is trained on familiar-view targets, part of this recovery reflects approximation of the reference policy. Its value lies in revealing correctable action structure and establishing its behavioral relevance.

Implications for method design. These findings suggest starting with simple linear mappings and then adding observation conditioning to address remaining deviations. The stronger effects near pre-contact interactions further motivate concentrating calibration or re-observation resources around these stages. This timing implication follows from the sensitivity observed across sampled phase populations, with local effects measured under specified continuation policies. Separate calibration on OpenVLA and Fast-WAM supports the repeatability of the diagnostic across the two interfaces. The current local experiments primarily characterize states reached by familiar-view roll-ins; extending the same analysis to states reached after substantial drift would clarify how correction requirements evolve as execution departs from this operating regime.

9 Conclusion

Matched-state measurements, structural controls, and bidirectional finite interventions examine immediate action responses to viewpoint changes separately from subsequent control feedback. Viewpoint-induced action distortion has predictable structure and can change subsequent success at pre-contact states. This behavioral phenomenon recurs in OpenVLA and Fast-WAM. Independent-distribution evaluations further delimit calibration transfer. The diagnostic provides action-level evidence for understanding viewpoint failure and informs the choice of calibration models and intervention timing.

AI USE STATEMENT

Generative AI tools (ChatGPT and Codex) were used solely as assistive aids for language polishing, reference formatting and typesetting, LaTeX typesetting, and debugging code errors. They did not contribute to the experimental design, data analysis, results, or scientific conclusions. We reviewed and verified all AI-assisted outputs and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

Sections 3–6 describe state restoration, structural mappings, finite interventions, data splits, and statistical units. Appendix B provides calibration, execution, and resampling details. Appendix A reports supplemental counts and optimization-seed results. Appendix C gives ACT/π_0 view-restoration protocols and task-level rates; Appendix D provides the supplementary audit, predictor, and controller details. Appendix E records additional benchmark aggregates.

REFERENCES

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

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025. URL https://arxiv.org/abs/2506.21539.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/947018640bf36a2bb609d3557a285329-Abstract.html.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fei_LIBERO-Plus_A_Progressive_Robustness_Benchmark_for_Visual-Language-Action_Models_CVPR_2026_paper.html.

Songen Gu, Yuhang Zheng, Weize Li, Yupeng Zheng, Yating Feng, Xiang Li, Yilun Chen, Pengfei Li, and Wenchao Ding. VistaBot: View-robust robot manipulation via spatiotemporal-aware view synthesis. *arXiv preprint arXiv:2604.21914*, 2026. URL https://arxiv.org/abs/2604.21914.

Hyeongjun Heo, Seungyeon Woo, Sang Min Kim, Junho Kim, Junho Lee, Yonghyeon Lee, and Young Min Kim. AnyCamVLA: Zero-shot camera adaptation for viewpoint robust vision-language-action models. *arXiv preprint arXiv:2603.05868*, 2026. URL https://arxiv.org/abs/2603.05868.

Jihoon Hong, Julian Skifstad, Qiyue Dai, Alice Chan, and Glen Chou. Steering robustness into world action models via mechanistic interpretability and optimal control. *arXiv preprint arXiv:2607.14943*, 2026. URL https://arxiv.org/abs/2607.14943.

Bingqi Huang, Bingchuan Wei, Xuan Wang, Yingkai Cai, and Zhaokui Wang. Cross-view action consistency for camera-robust vision-language-action policies. *arXiv preprint arXiv:2608.06965*, 2026. URL https://arxiv.org/abs/2608.06965.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U-Xuan Tan, Navonil Majumder, and Soujanya Poria. NORA: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025. URL https://arxiv.org/abs/2504.19854.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. URL https://arxiv.org/abs/1909.12271.

Tianchong Jiang, Xiangshan Tan, Samuel Wheeler, Luzhe Sun, Tewodros W. Ayalew, and Matthew Walter. What are we actually benchmarking in robot manipulation? *arXiv preprint arXiv:2606.04233*, 2026. URL https://arxiv.org/abs/2606.04233.

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

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

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

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.012. URL https://www.roboticsproceedings.org/rss21/p012.html.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. THE COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.133. URL https://www.roboticsproceedings.org/rss20/p133.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025. URL https://arxiv.org/abs/2505.17016.

Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Campagnolo Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis.

In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1173–1193. PMLR, 2025. URL https://proceedings.mlr.press/v270/tian25a.html.

Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 48–69. PMLR, 2025. URL https://proceedings.mlr.press/v270/wang25a.html.

Jingyun Yang, Ziang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. EquiBot: SIM(3)-equivariant diffusion policy for generalizable and data efficient learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1048–1068. PMLR, 2025. URL https://proceedings.mlr.press/v270/yang25a.html.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Hanxin Zhang, Mingshuo Xu, Abdulqader Dhafer, Shigang Yue, Hongbiao Dong, and Zhou Daniel Hao. Embodied interpretability: Linking causal understanding to generalization in vision-language-action models. *arXiv preprint arXiv:2605.00321*, 2026. URL https://arxiv.org/abs/2605.00321.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016. URL https://roboticsproceedings.org/rss19/p016.html.

SUPPLEMENTARY MATERIAL

This supplement separates the primary calibration and finite-intervention tests from the additional matched-state, sustained-control, and benchmark protocols. Appendix A gives complete counts and supplementary response profiles; Appendix B records the primary implementation and statistical conventions. Appendix C contains the complete ACT/π_0 view-restoration evaluation. Appendix D provides camera and restoration audits, the supplementary predictor, independent-state and cross-suite details, and sustained-control mechanics. Appendix E records the separately sourced robustness benchmarks. Sample sizes, action units, and intervention durations remain specific to each experiment.

A COHORTS AND COMPLETE RESULTS FOR SUPPLEMENTARY EXPERIMENTS

A.1 SIGNED VIEWPOINT SWEEPS FOR OPENVLA AND FAST-WAM

Table A1: OpenVLA viewpoint sweep over five tasks. Each task entry gives the number of successes out of 100 trials, with 500 episodes per angle. This sweep is independent of both the main test set and the directional intervention test set.

Task	0°	$+15^\circ$	$+30^\circ$	-30°
0	98	40	10	10
2	82	41	0	51
3	80	26	14	50
6	91	97	0	12
7	72	21	0	0
Total /500	423	225	24	123
Success rate (%)	84.6	45.0	4.8	24.6

For Fast-WAM, the macro-averaged success rates across ten tasks at -45°, -30°, 0°, $+15^\circ$, $+20^\circ$, $+25^\circ$, $+30^\circ$, and $+45^\circ$ are 60.0%, 82.0%, 93.5%, 99.0%, 80.0%, 47.0%, 60.0%, and 0%, respectively. Success neither decreases monotonically with the absolute angle nor is symmetric between positive and negative angles. These macro-averaged rates describe the corresponding sweep protocol and are not pooled with the failure-enriched intervention cohort of 300 states.

In the Fast-WAM sweep, Task 7 differs by 82.5 percentage points between -30° and $+30^\circ$. The aggregate and task-specific differences illustrate sign sensitivity under this scan protocol.

A.2 SIGNED DISPLACEMENTS IN NORMALIZED ACTION COORDINATES

Table A2: Signed displacements in a supplementary set of 1,100 matched-state snapshots. Each of the five tasks contributes ten source episodes, with 22 snapshots per episode. Values are normalized action displacements, first averaged within each episode and then averaged with equal episode weights; they are not measured in millimeters or degrees.

Task	Δx	Δy	Δz	$\Delta\omega_x$	$\Delta\omega_y$	$\Delta\omega_z$	Grip
0	-0.139	-0.005	-0.169	0.001	-0.012	0.033	0.068
2	-0.041	-0.010	0.128	-0.003	0.009	0.018	-0.149
3	-0.052	-0.012	0.323	-0.011	0.017	0.011	-0.072
6	-0.036	0.017	0.002	-0.008	-0.009	0.016	-0.018
7	-0.158	-0.131	-0.133	-0.004	0.000	0.013	0.009

The mean paired seven-dimensional distances for tasks 0, 2, 3, 6, and 7 are 0.812, 0.676, 0.703, 0.687, and 0.680, respectively. These measurements retain the same policy action decoding and normalization conventions and serve only to illustrate the task-dependent structure of output changes. The physical-unit evaluation in the main text reports translation, rotation, and gripper outcomes separately.

Figures A1 and A2 show information that is hidden by task averages: the dispersion of state-level distances, variation between episodes, and the signs and uncertainty of individual coordinates. The

distance panels use the same 1,100 supplementary snapshots as Table A2; they are separate from the primary structural test. Magnitude variation alone does not determine which mapping family is required, so the structural comparisons retain their own held-out evaluation.

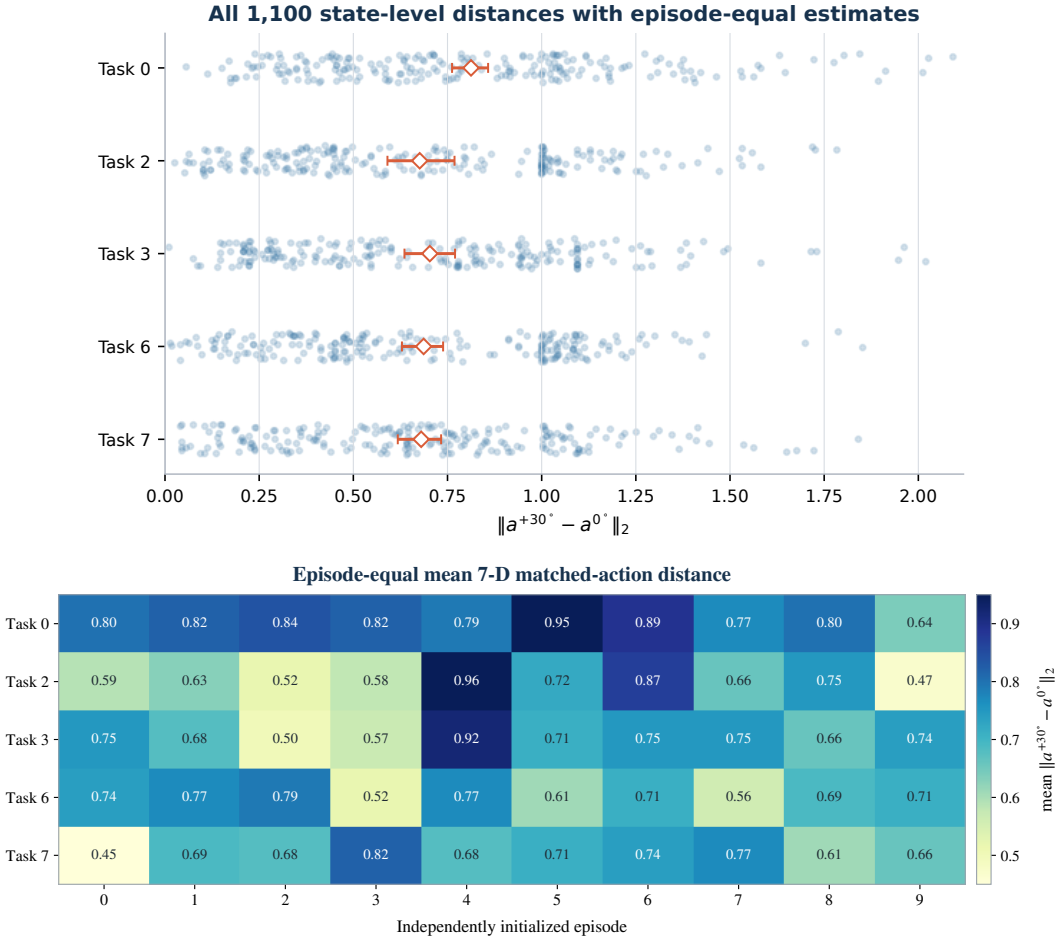


Figure A1: Supplementary matched-state action distances in normalized output coordinates. Top: all 1,100 snapshots; diamonds and whiskers show episode-equal task means and 95% episode-bootstrap intervals. Bottom: the mean of 22 snapshots within each of 50 source episodes. Points visualize state-level dispersion, while uncertainty keeps snapshots from the same episode together. These normalized seven-dimensional distances are not physical-unit errors.

The coordinatewise display contains 35 unadjusted 95% episode-bootstrap intervals, of which 19 exclude zero. The intervals describe the signed response profile; they are exploratory and do not constitute 35 multiplicity-adjusted tests. For example, the mean z displacement changes sign between Tasks 0 and 3. The familiar-view action remains a policy reference rather than an expert target.

A.3 Roles of Known Task Identity and Auxiliary Inputs

These supplementary controls use the CNN-based residual predictor described in Appendix D.3. Its architecture, normalized action targets, and validation-MAE checkpoint selection are specific to this cohort and differ from the primary calibration models in Section 4 and Appendix B.1.

The difference between the per-task constant and Native yields $p = 0.453$ in an exact two-sided McNemar test. After known task identity is added, the improvement for each matched seed satisfies $p \leq 8.21 \times 10^{-4}$. Further removing the original action from the no-RGB condition increases MAE from 0.0958 to 0.1359, a 41.9% increase, without a corresponding decrease in closed-loop success. This again indicates that offline error and behavioral utility measure different aspects.

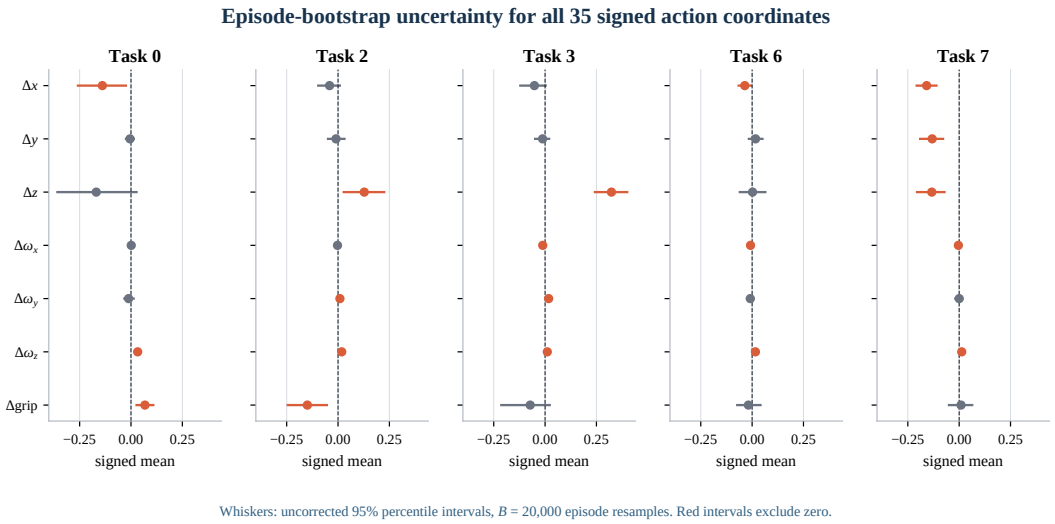


Figure A2: Signed-coordinate uncertainty for the supplementary 1,100-state cohort. Each panel corresponds to one task and shows the seven coordinate means with 95% episode-bootstrap intervals (20,000 resamples). Red intervals exclude zero; gray intervals include zero. All coordinates use the archived normalized action convention, and the intervals are unadjusted.

Table A3: Auxiliary predictor controls on a fixed structural cohort of 100 states. Source episodes 0–34 are used for training, 35–39 for model selection, and 40–49 for evaluation. All conditions and seeds reuse the same evaluation states. This cohort differs from the independent Spatial test set in Table 4.

Condition	Successes /100 (three seeds)	Mean success rate (%)	Residual MAE
Native $+30^\circ$	2	2.0	—
Per-task constant	5	5.0	—
No task conditioning, with RGB	20, 15, 28	21.0 ± 6.6	0.0980
No task conditioning, without RGB	24, 26, 27	25.7 ± 1.5	0.0958
Without RGB or original action	31, 26, 25	27.3 ± 3.2	0.1359
Shared task-conditioned predictor	58, 46, 48	50.7 ± 6.4	0.0950
Separate predictor for each task	64, 52, 62	59.3 ± 6.4	0.0985
Familiar-view reference	75	75.0	—

Figure A3 compares these two measurements directly. Predictors with similar residual MAE can have different closed-loop outcomes, and removing the Native action can increase MAE without lowering success on this cohort. This comparison concerns the supplementary CNN-based predictors; their checkpoints are still selected using the validation-MAE rule documented in Appendix D.3.

For the independent Spatial test in Table 4, paired comparisons between the full predictor and the predictor without its direct RGB branch yield $p \geq 0.136$ for all three seeds. These ablations modify the auxiliary predictor's input pathways; the base policy still receives images. The full predictor obtains 0/10 successes on task 1 for every seed. A predictor trained only on task 6 changes Native's 3/50 successes to 0/50 when applied to 50 states from five other tasks. In a text-conditioned experiment trained on four tasks with task 7 held out, none of the five seeds succeeds. Together with the leave-one-task-out evaluation in the main text, the latter two results characterize the transfer range of conditioning on known tasks.

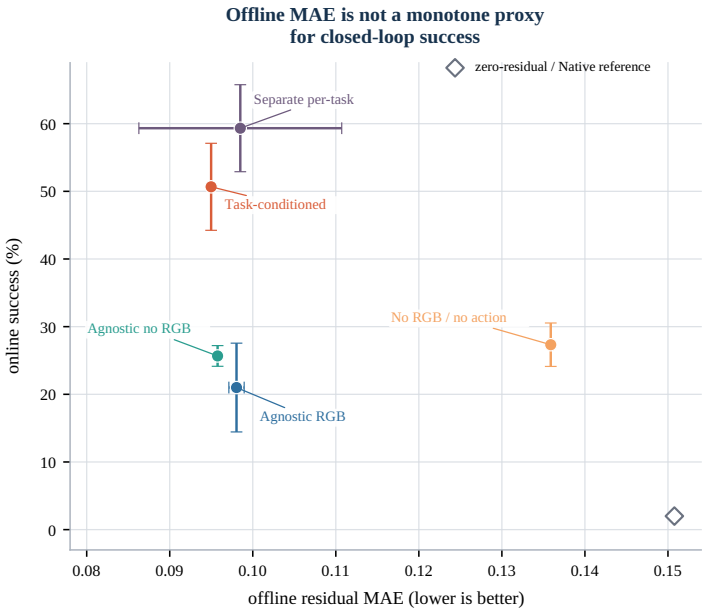


Figure A3: Offline residual MAE versus closed-loop success for the supplementary strict 100-state cohort. Each learned condition summarizes seeds 100/101/102; whiskers show sample standard deviations across these fits. The Native diamond is the zero-residual reference. MAE measures normalized output agreement, whereas success measures completed rollouts; their rankings need not coincide.

A.4 Complete Counts for Sustained Directional Interventions

Appendix D.5 specifies the supplementary OpenVLA transformation and clipping audit; Appendix D.6 specifies the sustained Fast-WAM controller. These implementations are separate from the primary finite-intervention procedure in Appendix B.2.

Table A4: OpenVLA directional intervention cohort of 50 fixed states. Aligned, Reverse, Orthogonal, and Delay modify the direction or timing of the predicted residual throughout the rollout. The three seeds reuse the same 50 environment states.

Condition	Successes /50 (three seeds)	Mean success rate (%)	Action-boundary incidence (%)
Native	2	4.0	—
Aligned	41, 41, 41	82.0	3.3
Reverse	0, 0, 0	0.0	30.9
Orthogonal	0, 1, 0	0.7	4.4
One-step delay	39, 39, 37	76.7 ± 2.3	10.1

Each paired comparison of Aligned with Native yields an exact two-sided McNemar $p = 3.64 \times 10^{-12}$. Orthogonality and norm constraints apply to the six-dimensional arm subspace before clipping, with the gripper residual set to zero. This protocol differs from the feasible, single-step, translation-only control in Figure 3(d) of the main text.

The aggregate OpenVLA direction comparison also varies by task (Figure A4). Each cell retains the three optimization seeds on the same ten starts, making both repeatability and unresolved tasks visible. The delay condition shifts the requested residual by one control slot throughout execution; it is not the primary one-step replacement experiment.

Table A5: Fast-WAM failure-enriched cohort of 300 states. Every condition uses the same 300 fixed starting states, with 30 states for each of ten tasks. This table tests direction and phase; it does not represent the overall success rate of unfiltered episodes.

Condition	Successes /300	Success rate (%)
Native	2	0.7
Phase-matched Aligned	297	99.0
Reverse	0	0.0
Orthogonal	3	1.0
Correction during lifting	130	43.3

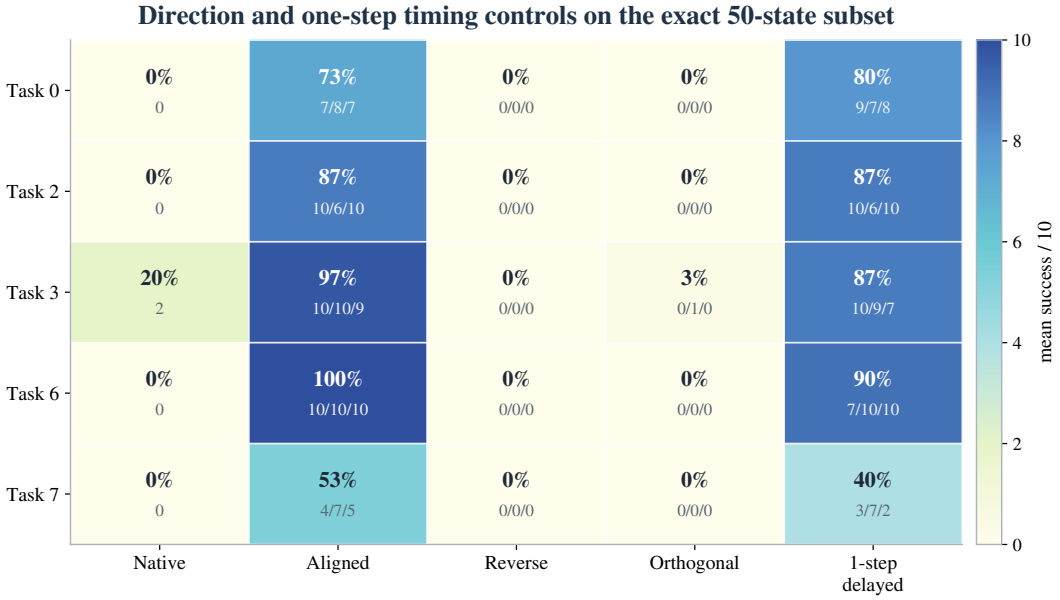


Figure A4: Task-level sustained direction and delay controls on the supplementary 50-state OpenVLA subset. Percentages average the three optimization seeds; slash-separated entries are counts out of ten starts for seeds 100/101/102. Native is evaluated once on the shared starts. Orthogonal changes the six-dimensional arm residual before clipping and sets its gripper residual to zero, as specified in Appendix D.5.

A.5 Per-Task Main-Test Results and Optimization Seeds

Table A6: Per-task success counts on the independent OpenVLA main test. Each task has 100 starting states, which are reused across methods within that task.

Condition	Task 0	Task 2	Task 3	Task 6	Task 7	Total
Native	20	15	18	5	12	70
Task-specific affine	56	49	56	36	49	246
Feature-linear	63	56	64	41	60	284
Shared MLP	53	46	58	32	47	236
Task-conditioned MLP	70	62	68	46	66	312
Direct prediction	67	60	67	45	65	304
Arm only	59	53	57	36	55	260
Gripper only	28	23	25	12	22	110
Familiar-view reference	88	82	84	76	80	410

The main checkpoint achieves 312/500 successes, while the other two optimization seeds achieve 306/500 and 318/500. The mean and sample standard deviation are 62.4% $\pm$ 1.2 percentage points. The environment denominator remains the same set of 500 starting states for all three seeds. Relative

to the task-specific affine model, the main conditional predictor recovers 97 failures and loses 31 successes, yielding a net gain of 66/500, or 13.2 percentage points.

A.6 Complete Counts for Finite Action Interventions

Table A7 supplies the raw counts underlying Figure 3(a,b). Conditions within each phase share restored snapshots. The three phase populations differ in coverage, as detailed in Appendix B.3.

Table A7: Success counts under finite interventions. Injection is paired with familiar-view reference continuation; rescue is paired with Native continuation. Conditions within a row share snapshots; phase coverage differs. Subscripts denote executed control steps.

Phase (n)	Reference	Inject$_1$	Inject$_4$	Native	Rescue$_1$	Rescue$_4$
Approach (150)	126	123	119	45	48	53
Pre-contact (132)	118	98	78	42	56	77
Post-contact (118)	108	107	104	52	53	57

B Implementation and Statistical Details

B.1 Calibration Data and Training

This subsection describes the primary calibration models and tests. The supplementary CNN-based predictor and its separate cohorts are documented in Appendix D.3.

The visual PCA projection and input normalization are fitted only on calibration-training source episodes and then applied unchanged to validation and testing. Controls that require visual inputs share the same projection. Training covers both familiar-view and shifted-view roll-ins, with snapshot positions fixed in advance for each source episode. Missing positions due to termination or unavailable states are not determined by intervention outcomes. Different mappings receive the same target pairs, and the main test uses independent starting states.

Both the shared and task-conditioned MLPs use a two-layer GELU backbone with width 128. The direct predictor uses the same backbone and training budget. Errors for continuous components are scaled using training-set component statistics; the gripper uses a separate classification objective and a validation-selected threshold. The learning-rate search is $\{10^{-4}, 3 \times 10^{-4}\}$, with at most 100 training epochs. Ridge coefficients for the feature-linear model are selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Validation success rate is used for model selection, with residual error breaking ties. Test results are not used to select residual strength or the best seed.

Rotation targets are defined in a consistent tangent space and executed through rotation composition. A fixed rotation mapping can be written as $Q \mapsto RQR^\top$; scaled and affine rotation models are constructed under the same local convention. Arm and gripper outputs are handled separately, so an unconstrained seven-dimensional affine matrix cannot convert between components with different units.

B.2 Four-Branch Execution Procedure

For each finite-intervention state, execution proceeds as follows. First, we restore the source snapshot and query the familiar and shifted views using two independently cloned policy states. We save the action sequences and post-query policy states. We then create four independent environment branches: Reference, Injection, Native, and Rescue. Reference and Injection share the familiar-view continuation configuration, while Native and Rescue share the shifted-view continuation configuration. Each branch consumes one control slot, executes its corresponding original or replacement action, and then continues using its own new observations from the specified view.

Action replacement does not change how the action queue advances. Queue retention, replanning times, and executed-action history follow the same convention within each paired comparison. The other view is used only to obtain candidate replacement actions. The four-step version preserves

each policy's interface definition, so OpenVLA's stepwise re-querying and Fast-WAM's action-chunk prefix replacement are interpreted separately.

B.3 Phase Coverage and Repeated Observations

All 150 source episodes for OpenVLA's local interventions have an approach state; 132 have an eligible pre-contact state, and 118 have an eligible post-contact state. The 18 episodes without a pre-contact state comprise eight reference successes and ten reference failures. The 32 episodes without a post-contact state comprise 18 reference successes and 14 reference failures. The reference-success label is consistent across phase snapshots from the same source episode, and phase coverage determines the denominator in each row.

The primary pre-contact minus approach contrasts are $+8.6$ pp $[0.8, 16.8]$ for rescue and -13.2 pp $[-20.5, -6.0]$ for injection. These contrasts describe the sampled phase populations rather than the causal effect of changing phase itself.

Multiple phases, snapshots, and repeated orthogonal directions from the same source episode are statistically dependent. The main analysis clusters by source episode, with orthogonal outcomes first averaged within each source state. Different optimization seeds describe repeated fits and are not treated as additional environment samples. Cross-phase differences are estimated by resampling complete source episodes together, preserving dependence between phases.

B.4 Executed Magnitudes and Control Conventions

The main finite-intervention configuration uses a control frequency of 20 Hz: one control step corresponds to 0.05 s, and four steps to 0.20 s. The translation-direction control defines the feasible set using per-axis action limits of ± 25 mm. These settings apply to the main single-step and short-duration intervention experiments; the sustained directional intervention cohorts are reported under their own control interfaces. The translation-direction experiment computes the common magnitude from the controller's final feasible set and records both proposed and executed vectors. The nominal magnitude is selected on the validation set, and the actual magnitude satisfies the joint feasibility constraints across all compared directions. Norm and direction are checked at the final controller input, and zero-magnitude states remain in the support.

Full-episode recovery, single-step rescue, and sustained directional interventions use their stated treatment durations and action subspaces. In particular, orthogonality in the six-dimensional arm subspace, orthogonality in three-dimensional translation, and complete action replacement have different experimental meanings. All counts, denominators, intervals, and comparisons are reported under their corresponding protocols.

For the primary component ablation in Figure 3(c), arm-only correction gains 38.0 pp $[33.8, 42.4]$ over Native and gripper-only correction gains 8.0 pp $[4.6, 11.4]$. In the separate 132-state translation-direction test, the validation-selected nominal limit is 3 mm and every direction has mean executed magnitude 2.86 mm, with no further clipping. The three orthogonal directions succeed on 44, 43, and 45 of the same 132 states.

B.5 Interpreting Paired Outcomes

At the same starting state, calibration can turn failure into success, but it can also turn an original success into failure. For OpenVLA, the main predictor recovers 244 failures and loses two successes relative to Native, together yielding a net gain of 48.4 percentage points. Relative to the task-specific affine model, it recovers 97 failures and loses 31 successes, yielding 13.2 percentage points. For Fast-WAM, 109 recoveries and 14 losses yield 19.0 percentage points. Reporting these transitions distinguishes an aggregate improvement from an improvement at every starting state.

The primary OpenVLA residual predictor exceeds direct action prediction by 1.6 pp $[-2.6, 5.8]$, so the data do not establish a residual-parameterization advantage. On the separate 200-state Fast-WAM pre-contact test, familiar-view success changes from 88.0% to 76.0% under injection, and shifted-view success changes from 60.0% to 72.0% under rescue. These finite interventions use their specified continuation policies and are distinct from full-episode calibration.

Paired intervals respect the task and source-episode structure. Exact McNemar tests for independent paired binary outcomes are not applied to a flattened list of correlated snapshots. Intervals for the reverse viewpoint, leave-one-task-out transfer, direct versus residual prediction, and nonlinear versus feature-linear prediction in Fast-WAM all include zero. Together, these results delimit the gains supported by the current evidence.

C ACT AND π_0 UNDER THE VISTABOT VIEW-RESTORATION PROTOCOL

C.1 TASKS, CONDITIONS, AND EVALUATION UNITS

The RLBench evaluation (James et al., 2020) contains eight tasks: close box, close laptop, meat on grill, open box, push buttons, stack wine, take lid off, and toilet seat down. ACT (Zhao et al., 2023) and π_0 (Black et al., 2025) are trained from 50 familiar-view trajectories per task under the VistaBot protocol (Gu et al., 2026). Each novel azimuth in $\{-45°, -30°, +30°, +45°\}$ is evaluated 25 times per task, giving 200 evaluations per policy–condition–novel-angle combination.

Default uses the novel-view observation; View-restored applies the VistaBot observation-side pipeline (Gu et al., 2026). This restoration acts on observations supplied to the policy, whereas the primary calibration in this paper acts on policy outputs. Figure 4(g,h) shows equal-task macro-averages. The two curves share one clean $0°$ reference within each policy; the stated 25-trial denominator applies to the novel-view entries.

These results are reported at task–angle aggregate level. The policies, tasks, demonstrations, action interfaces, and evaluation starts differ from the OpenVLA matched-state cohorts. They supply complementary evidence of outcome-level viewpoint sensitivity and robustness improvements ability and are not pooled with the matched-state or action-intervention experiments. Task-level rates do not establish paired outcome transitions or identify an internal policy mechanism.

C.2 COMPLETE RATES AND VIEWPOINT-GENERALIZATION SCORE

Table A8 gives the aggregates plotted in Figure 4(g,h), and Tables A9 and A10 give all task-level rates. For task τ and novel-view set Θ, the viewpoint-generalization score normalizes by the task's own clean-view rate before averaging:

$$\mathrm{VGS} = \frac{1}{|\mathcal{T}||\Theta|} \sum_{\tau \in \mathcal{T}} \sum_{\theta \in \Theta} \frac{S_\tau(\theta)}{S_\tau(0)}. \tag{10}$$

Here $S_\tau(\theta)$ is the task success rate under the stated condition. Clean-view rates span 0.60–1.00 for ACT and 0.56–0.96 for π_0. VGS therefore generally differs from the ratio of two rounded macro-averages.

Table A8: VistaBot-protocol RLBench summary. Success rates are percentages macro-averaged over eight tasks; VGS is dimensionless and follows Equation 10. The clean reference is shared between conditions.

Policy	Condition	Clean $0°$	$-45°$	$-30°$	$+30°$	$+45°$	VGS
ACT	Default	83.0	16.0	29.0	24.5	12.5	.242
ACT	View-restored	83.0	42.0	64.0	63.5	51.5	.671
π_0	Default	83.5	22.5	32.0	37.5	24.5	.333
π_0	View-restored	83.5	61.5	73.0	76.5	75.5	.873

At $-45°, -30°, +30°, +45°$, the View-restored minus Default differences are 26.0, 35.0, 39.0, and 39.0 percentage points for ACT and 39.0, 41.0, 39.0, and 51.0 percentage points for π_0. VGS increases by 0.429 and 0.540, respectively. These are descriptive differences between the reported aggregate rates; no paired significance test is inferred from the marginals.

Table A9: Complete ACT success rates under the VistaBot-protocol RLBench evaluation. Values are proportions. Every novel-view task–condition cell uses 25 evaluations; Clean is the shared familiar-view reference.

	Clean	Default				View-restored			
Task	0°	-45°	-30°	$+30^\circ$	$+45^\circ$	-45°	-30°	$+30^\circ$	$+45^\circ$
Close box	0.84	0.20	0.32	0.56	0.28	0.40	0.76	0.60	0.56
Close laptop	0.60	0.20	0.20	0.20	0.16	0.36	0.44	0.44	0.40
Meat on grill	0.84	0.00	0.00	0.04	0.00	0.40	0.40	0.52	0.32
Open box	0.96	0.08	0.32	0.08	0.08	0.32	0.72	0.76	0.28
Push buttons	0.64	0.08	0.32	0.08	0.08	0.52	0.64	0.68	0.64
Stack wine	0.80	0.00	0.08	0.04	0.00	0.12	0.52	0.48	0.40
Take lid off	0.96	0.32	0.40	0.32	0.08	0.32	0.64	0.68	0.60
Toilet seat down	1.00	0.40	0.68	0.64	0.32	0.92	1.00	0.92	0.92
Macro-average	0.830	0.160	0.290	0.245	0.125	0.420	0.640	0.635	0.515

Table A10: Complete π_0 success rates under the VistaBot-protocol RLBench evaluation. Values are proportions. Every novel-view task–condition cell uses 25 evaluations; Clean is the shared familiar-view reference.

	Clean	Default				View-restored			
Task	0°	-45°	-30°	$+30^\circ$	$+45^\circ$	-45°	-30°	$+30^\circ$	$+45^\circ$
Close box	0.96	0.24	0.20	0.72	0.60	0.44	0.76	0.92	0.76
Close laptop	0.84	0.24	0.32	0.60	0.40	0.60	0.60	0.56	0.72
Meat on grill	0.84	0.04	0.24	0.20	0.08	0.56	0.68	0.76	0.72
Open box	0.96	0.28	0.24	0.12	0.16	0.36	0.52	0.92	0.88
Push buttons	0.68	0.16	0.20	0.32	0.04	0.76	0.72	0.80	0.76
Stack wine	0.56	0.00	0.12	0.00	0.08	0.48	0.76	0.48	0.52
Take lid off	0.92	0.32	0.56	0.64	0.24	0.80	0.80	0.72	0.84
Toilet seat down	0.92	0.52	0.68	0.40	0.36	0.92	1.00	0.96	0.84
Macro-average	0.835	0.225	0.320	0.375	0.245	0.615	0.730	0.765	0.755

C.3 Task-Level Variation in Observation-Side Recovery

The task-level rates show why aggregate recovery should be interpreted together with the complete profiles. For ACT at $+45^\circ$, restoration improves meat on grill and stack wine, but their success rates remain 0.32 and 0.40, below the respective clean references of 0.84 and 0.80. For π_0 on close laptop at $+30^\circ$, the restored rate is 0.56 versus 0.60 under Default. Thus, improved policy-level averages do not imply improvement at every task–angle combination. These are descriptive differences between aggregate cells, without an inferred paired significance test.

VGS accounts for different clean-view task rates by normalizing within each task before averaging. It complements absolute success: a task with low clean-view success can have a relatively high normalized score while remaining difficult to complete. The Default and View-restored columns are therefore retained alongside VGS, so the magnitude and consistency of observation-side recovery can be assessed directly.

D SUPPLEMENTARY AUDIT AND SUSTAINED-INTERVENTION PROTOCOLS

D.1 COHORT RELATIONS AND EVIDENCE SCOPE

The supplementary experiments below retain their own datasets, predictors, action coordinates, and execution interfaces. Their training details do not replace the primary-test calibration protocol in Appendix B.1. The strict structural cohort and independent Spatial bank each have Native success $2/100$, but comprise different environmental states. The 50-state directional bank is an exact subset of the independent Spatial bank; both Native successes fall in that subset, giving $2/50$. Optimization seeds reuse environmental starts and are not additional independent state samples. Table A11 records the relationships needed to interpret these results.

Table A11: Sampling units and relationships of supplementary cohorts. These cohorts remain separate from the primary structural and finite-intervention tests.

Evidence	Unit and coverage	Relationship and role
Fresh/fresh audit	50 traces; 1,100 selected snapshots	Five Spatial tasks; identifies pre-execution action differences.
Strict structure	100 states; ten per Spatial task	Common held-out states for all structural controls in Table A3.
Independent Spatial	100 newly instantiated states	Separate from strict structure; parent of the directional subset.
Sustained direction	50 starts; 650 traces	Exact five-task subset of independent Spatial; Native shared across seeds.
Object and Goal	100 online states per suite	Separate policies, paired datasets, probes, and state banks.
Trajectory timing	75 paired rollouts	Selected familiar-success/shifted-failure pairs; descriptive downstream timing.
Fast-WAM sustained	300 starts per condition	Separate from the signed sweep and primary finite interventions.

Task identities. The ten-task Spatial experiments use the task IDs in Table A12. The matched-state and directional subsets use Tasks 0, 2, 3, 6, and 7. These IDs are suite-local; Object and Goal use their own policies and task definitions.

Table A12: LIBERO-Spatial task mapping. Each instruction is "Pick up the black bowl [location] and place it on the plate."

Task ID	Location phrase
0	between the plate and the ramekin
1	next to the ramekin
2	from table center
3	on the cookie box
4	in the top drawer of the wooden cabinet
5	on the ramekin
6	next to the cookie box
7	on the stove
8	next to the plate
9	on the wooden cabinet

D.2 FRESH/FRESH RESTORATION AUDIT AND TRAJECTORY TIMING

Frozen run and camera treatment. The supplementary OpenVLA audit uses `openvla/openvla-7b-finetuned-libero-spatial`, repository commit `c8f03f48af69`, inference seed 7, and center cropping. Rollouts have at most 220 action steps after ten settling steps. This audit is indexed by control step rather than a fixed wall-clock frequency. The camera is rotated from its base pose about the world z axis and a derived look-at center; its look-at quaternion is recomputed and observations are refreshed. Robot and object states, action coordinates, controller frame, and task

remain unchanged. The exact camera poses, with quaternion order (w, x, y, z), are

$$
\begin{aligned}
p_0 &= (0.65861317, 0, 1.61035002),\\
q_0 &= (0.63801772, 0.30484972, 0.30484984, 0.63801772),\\
p_{30} &= (0.57037574, 0.32930659, 1.61035002),\\
q_{30} &= (0.45114666, 0.21556135, 0.37336320, 0.78140893).
\end{aligned}
$$

Figures A5 and A6 make the camera treatment and action contrast visible. Their examples come from a separate initial-state figure-asset bank of 50 records, ten for each of Tasks 0, 2, 3, 6, and 7. Within each task, the displayed record is the one nearest the task median action distance in that bank, selected independently of rollout outcome. This bank is separate from the 1,100 along-trajectory snapshots. Same-view repeated actions agree exactly in all 50 asset records.

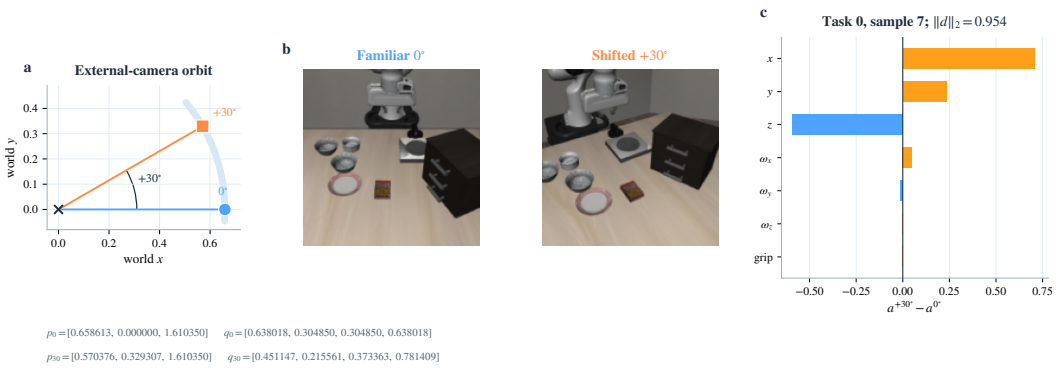


Figure A5: Camera geometry and one supplementary paired query. (a) World-frame camera positions at $0°$ and $+30°$. (b) The familiar and shifted observations at one restored Task-0 state. (c) The signed decoded-action difference before execution. Camera poses use the MuJoCo (w, x, y, z) quaternion convention; action bars use normalized policy-output coordinates. The example is selected from the separate 50-record figure-asset bank.

Independent queries and audit invariants. For each branch, the evaluator restores the manipulation state, controller state, task input, policy inference initialization, and relevant random state; sets exactly one camera pose; refreshes the renderer and observation; and issues a fresh frozen-policy query. It then discards transient objects and restores the original snapshot before querying the other view. Neither branch reuses the other branch's decoded action. All 50 traces pass the audit, covering 11,000 control steps. Each contributes the 22 fixed indices

$$
\begin{aligned}
&0, 10, 21, 31, 42, 52, 63, 73, 83, 94, 104,\\
&115, 125, 136, 146, 156, 167, 177, 188, 198, 209, 219,
\end{aligned}
$$

for 1,100 selected matched snapshots. Table A13 reports the restoration and execution checks.

Table A13: Fresh/fresh audit checks. Image mean absolute differences (MAD) use the saved image scale; other numeric checks use stored floating-point coordinates.

Check	Maximum difference or count
Expected / discovered / passed / failed traces	50 / 50 / 50 / 0
Paired non-camera simulator state	0
Executed raw versus fresh shifted-view action	0
Stored versus recomputed paired residual	0
Camera restoration position / quaternion	0 / 0
Restored Native image MAD	1.5258789×10^{-5}
Native versus canonical image MAD (minimum)	29.0081
Exception-log hits / discovery errors	0 / 0

Recomputed signed coordinates agree with $a^{+30} - a^0$ to 2.22×10^{-16}, and norms agree to 4.44×10^{-16}. The restored-image MAD is below the specified 2×10^{-5} tolerance. Action coordinates are

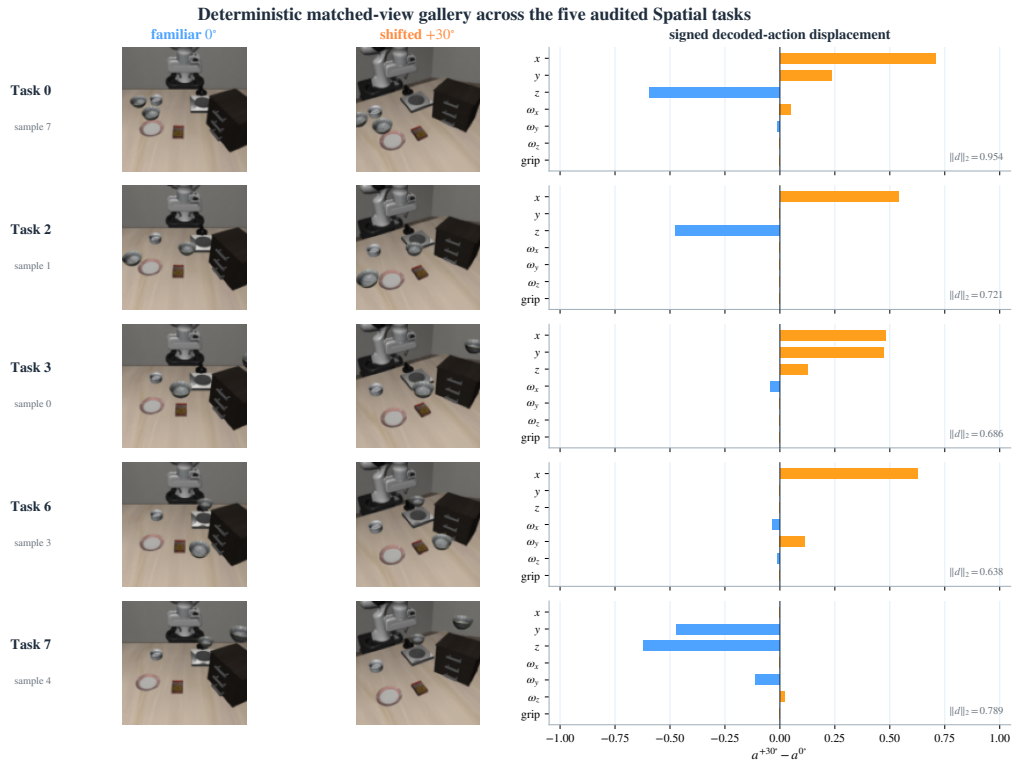


Figure A6: Matched-view examples from the five audited Spatial tasks. Each row contains familiar and shifted images from the same restored state and the resulting signed seven-dimensional output difference. One record per task is selected by the median-distance criterion in the separate 50-record asset bank. The images illustrate camera-only changes; no rollout success or physical-unit action error is inferred from these examples.

ordered $(x, y, z, \omega_x, \omega_y, \omega_z, \text{grip})$ in normalized policy-output space. The mean seven-dimensional distance is 0.71168 (95% interval $[0.68247, 0.74091]$), the snapshot median is 0.67816, and the range is $[0.01035, 2.08669]$. These values do not combine physical millimeters and degrees. Table A14 supplements the signed means in Table A2 with episode-bootstrap uncertainty.

Table A14: Supplementary matched-action distances in normalized coordinates. Intervals use 20,000 episode-block percentile resamples; all 22 snapshots remain grouped. The overall interval is task-stratified with equal task weights. RNG seed: 20260727.

Task	Mean $\|a^{+30} - a^{0}\|_2$	95% interval
0	0.81244	[0.76210, 0.85781]
2	0.67618	[0.59056, 0.76862]
3	0.70303	[0.63577, 0.77006]
6	0.68660	[0.62907, 0.73833]
7	0.68012	[0.61799, 0.73305]

Selected trajectory pairs. The separate timing audit contains 75 pairs selected for familiar-view success and shifted-view failure. All first sustained end-effector deviations exceeding 10 mm occur during early approach; 58/75 precede detected target-object motion. After the ten-step settling prefix, the recorded crossing indices 2, 3, 4, 6, and 7 occur in 45, 23, 4, 2, and 1 pairs, respectively. Shifted-view failure classes are before target motion (58), carried then failed (7), disturbed without lifting (6), and lifted without carrying (4). The archived derived summary retains the prespecified crossing index and failure class, but does not support re-estimation under alternative thresholds. This outcome-selected audit localizes divergence within these pairs; it is not an estimate of failure prevalence and is not pooled with the fresh/fresh identification cohort.

Figure A7 visualizes the recorded temporal indices and failure categories. These measurements describe divergence after actions have been executed, complementing the pre-execution paired queries. Because pairs were selected by their different outcomes, this analysis cannot estimate the frequency of such failures among unfiltered starts.

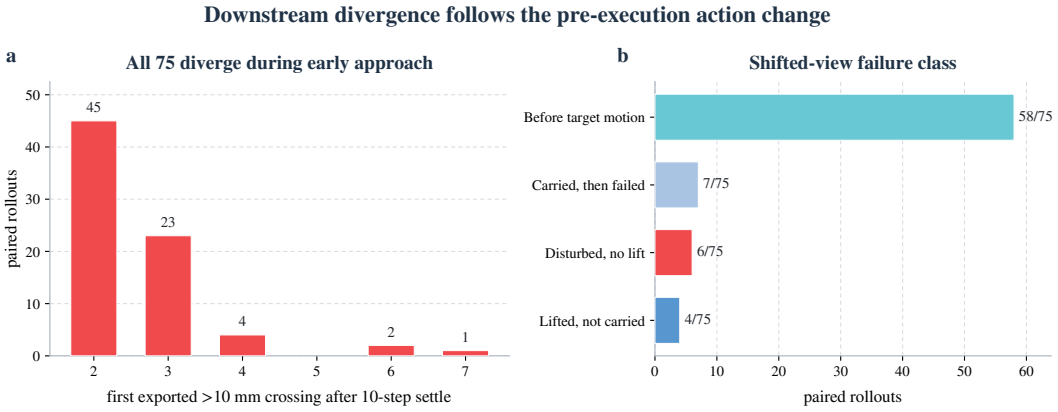


Figure A7: Temporal localization within the separate 75-pair familiar-success/shifted-failure audit. Left: the first recorded sustained end-effector separation above 10 mm, indexed after the ten-step settling prefix. Right: the recorded shifted-view failure category. Counts apply to these outcome-selected pairs only; they do not estimate population failure prevalence or compare alternative divergence thresholds.

D.3 Predictor Protocol for Supplementary Structural Controls

This subsection documents the supplementary strict-cohort predictor used in Table A3, together with the associated independent-bank and sustained-intervention analyses. Its architecture and selection rule are distinct from the primary mappings in Appendix B.1.

Targets and inputs. Targets are offline fresh-query residuals $y_t = a_t^0 - a_t^{+30}$. Deployment inputs exclude the canonical image, a_t^0, object ground truth, future states, expert actions, and online counterfactual queries. The 18 base numeric inputs are end-effector position (3), end-effector quaternion (4), gripper joint positions (2), shifted-view raw Native action (7), control-step index divided by the maximum steps (1), and camera angle divided by $45°$ (1). The task-conditioned predictor appends a ten-dimensional task one-hot vector for 28 inputs. Standardization uses training-set statistics only. The setup assumes known suite-local task identity.

Architecture and fitting. The RGB encoder has four convolution–GroupNorm–SiLU blocks with channels $3 \to 32 \to 64 \to 96 \to 128$ and global average pooling. The numeric branch is $28 \to 128 \to 128$. Their 256-dimensional concatenation feeds a $256 \to 256 \to 128 \to 7$ head with LayerNorm, SiLU, dropout 0.10, and final tanh. The task-conditioned predictor has 308,263 trainable parameters. Images are 128×128 pixels.

The balanced supplementary dataset contains 124,840 samples: 109,170 training and 15,670 validation samples (10,917 and 1,567 per task). Episodes 0–34 supply training, 35–39 supply validation, and 40–49 are reserved for the strict online cohort. Seeds 100/101/102 train for 35 epochs using AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256, cosine decay, and gradient-norm clipping at 5. Smooth-L1 coordinate weights are $(1, 1, 1, 0.6, 0.6, 0.6, 2)$. Targets are scaled by training-set 95th absolute percentiles, lower bounded by $(0.05, \ldots, 0.05, 0.25)$. Selection minimizes mean validation MAE without online success feedback. Constants use per-task training residual means; task-agnostic, shared task-conditioned, and separate task-specific predictors use the stated input restrictions. None is fitted to the strict online test residuals.

Figure A8 summarizes the supplementary predictor's training and deployment inputs. The familiar-view action enters only as an offline target. At deployment, the correction uses the shifted image, Native action, available robot state, and known task identity. The no-RGB ablation removes the auxiliary visual branch while leaving the frozen base policy's image input intact.

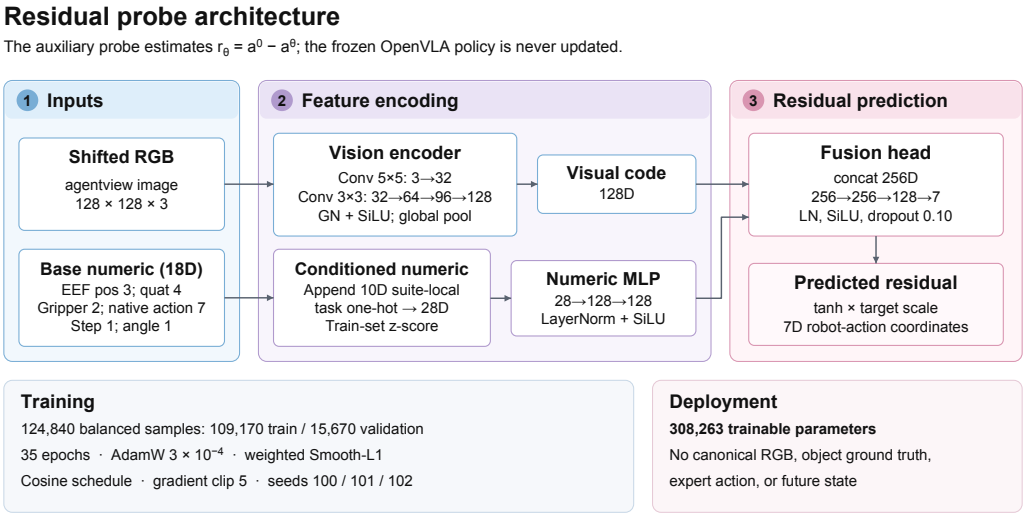


Figure A8: Information flow of the supplementary CNN-based residual predictor. Shifted-view RGB and 28 numeric inputs (including a ten-dimensional task indicator) feed the residual head. Familiar-view observations and actions, object ground truth, expert actions, and future state are excluded from deployment inputs. This architecture is used for the supplementary strict and independent banks; it is distinct from the primary PCA-feature MLP.

Task heterogeneity and auxiliary RGB. Table A15 expands the aggregate structural comparison. The no-RGB independent-bank ablation zeroes and freezes the auxiliary visual branch in training and evaluation; the frozen base policy continues to receive images. Table A16 reports its paired comparison to the full predictor. These results show no detectable paired advantage for the full predictor on this bank; they do not establish equivalence or RGB independence of the base policy.

Table A15: Strict-cohort successes out of ten starts per task. Triples are seeds 100/101/102. A-RGB and A-noRGB are task-agnostic predictors with and without direct RGB; No action additionally removes the Native action. TC is shared task-conditioned, and PT fits a separate predictor per task.

Task	Native	Constant	A-RGB	A-noRGB	No action	TC	PT	Familiar
0	0	0	2/4/2	0/1/1	3/5/1	9/6/7	9/5/9	9
1	0	0	0/0/0	0/0/0	0/0/0	1/0/0	8/2/1	10
2	0	0	1/2/3	4/2/2	4/2/1	7/3/7	7/6/9	8
3	1	5	5/3/5	5/6/5	6/6/5	8/9/10	10/10/9	7
4	0	0	0/0/3	1/2/1	6/7/5	3/3/1	3/0/6	6
5	0	0	1/0/2	0/0/1	4/1/3	4/3/1	4/5/2	4
6	1	0	4/1/4	6/5/6	2/1/4	9/8/8	9/9/8	10
7	0	0	1/1/0	3/3/4	0/0/0	4/3/4	7/5/5	8
8	0	0	5/4/6	4/4/3	2/1/2	7/6/6	5/7/9	6
9	0	0	1/0/3	1/3/4	4/3/4	6/5/4	2/3/4	7

Table A16: Paired full-versus-no-RGB outcomes on the independent 100-state Spatial bank. Only columns give discordant outcomes; p is the exact two-sided McNemar value.

Seed	Full	No RGB	Full only	No RGB only	p
100	60	66	10	16	0.32694
101	57	58	15	16	1.00000
102	54	63	10	19	0.13605

For the separate Object suite, gains/losses relative to Native are 72/0, 40/0, and 47/0. Goal gains/losses are 52/0, 55/3, and 43/2, including losses of some Native successes despite positive net effects. These

suite-specific counts explain why recovery is reported separately in Table 4 rather than pooled across policies.

D.4 Independent-State and Cross-Suite Details

This section expands the independent-bank results in Table 4. The Spatial, Object, and Goal online evaluations each use 100 fixed environmental starts, with ten starts per task. All three optimization seeds reuse the same starts within a suite. These are supplementary cohorts, separate from the primary 500-start tests; the independent Spatial bank also differs from the strict structural cohort in Table A3. The suite-specific predictors follow the supplementary protocol in Appendix D.3. Separate policies, checkpoints, calibration data, and task definitions are used for Object and Goal, so cross-suite recurrence does not constitute transfer of one fitted predictor between suites.

D.4.1 Initial-State Paired Queries Across Suites

Before evaluating closed-loop correction, each suite provides a separately instantiated 50-state paired-query bank: ten states for each of suite-local Tasks 0, 2, 3, 6, and 7. Restoring the same state and repeating the canonical-view query produces zero action difference in all three banks, and all state-restoration checks pass. Changing only the camera produces the nonzero distances in Table A17. These initial-state measurements support recurrence of the immediate viewpoint response across the three evaluated policies. They do not measure downstream success or determine which action components drive failure. Distances are in normalized seven-dimensional policy-output coordinates and should not be interpreted as physical translation or rotation errors.

Table A17: Mean initial-state paired action distance in each supplementary 50-state query bank. Each task contributes ten matched states. Task IDs are suite-local; the overall mean gives equal weight to the five tasks.

Suite	Overall	Task 0	Task 2	Task 3	Task 6	Task 7
Spatial	0.754	0.837	0.725	0.665	0.741	0.800
Object	0.732	0.669	0.824	0.639	1.042	0.485
Goal	0.508	0.278	0.383	0.444	0.655	0.779

D.4.2 Task-Level Recovery and Optimization Variability

Figure A9 retains the task-level variation hidden by suite averages. Spatial recovery is consistent on some tasks, particularly Task 6, while Task 1 remains unsuccessful under the full predictor for all seeds. Object has larger optimization sensitivity: for example, Task 9 ranges from zero to ten successes on the same ten starts. Goal also contains persistent difficult cases despite positive aggregate recovery. Because task IDs have different meanings across suites, horizontal positions can be compared as success rates, but matching row numbers do not identify matching tasks.

Independent Spatial bank. Table A18 gives the complete task-level counts. For the full predictor, paired gains/losses relative to Native are 58/0, 55/0, and 52/0 for seeds 100/101/102. The corresponding exact two-sided McNemar values are 6.939×10^{-18}, 5.551×10^{-17}, and 4.441×10^{-16}. The archived hierarchical task/state/seed bootstrap reports a mean paired gain of 55 pp with a 95% interval of $[35.33, 73.67]$ pp. This interval pertains to the supplementary hierarchy and is distinct from the main-test bootstrap protocol. The ten fixed states for each of Tasks 0, 2, 3, 6, and 7 also form the 50-state sustained-direction subset in Appendix A.4; those intervention runs are not an additional independent replication of this bank.

Direct RGB input to the auxiliary predictor. The no-probe-RGB variant keeps the full model's task-conditioned numeric inputs and zeroes and freezes its visual branch during both training and evaluation. The base OpenVLA policy continues to use images, and its Native action remains an input to the predictor. Consequently, this ablation tests the utility of an additional direct RGB pathway, rather than whether the complete policy is independent of vision. Table A16 reports the exact paired comparison with the full predictor; none of the three seeds yields a detectable difference, which does not establish equivalence or superiority. Table A19 records the corresponding offline validation



Figure A9: Task-level outcomes on the supplementary independent 100-state banks. Each panel uses ten fixed starts per task and a separately trained suite-specific predictor. Diamonds indicate shifted-view Native success, colored dots indicate seeds 100/101/102, and red ticks indicate their mean. Seed variation reflects optimization on shared environmental starts, not additional independent test states. These cohorts are separate from the primary 500-start tests.

Table A18: Success counts on the supplementary independent Spatial bank. Each task has ten fixed starts shared by Native, the full task-conditioned predictor, and its no-probe-RGB variant. Columns 100/101/102 identify optimization seeds. The no-probe-RGB condition removes only the auxiliary predictor's direct image input.

		Full predictor			No probe RGB		
Task	Native	100	101	102	100	101	102
0	0	7	8	7	3	4	1
1	0	0	0	0	1	0	0
2	0	10	6	10	10	10	10
3	2	10	10	9	10	10	10
4	0	4	3	4	9	9	8
5	0	1	2	2	2	1	3
6	0	10	10	10	10	9	10
7	0	4	7	5	6	5	8
8	0	8	6	4	7	4	8
9	0	6	5	3	8	6	5
Total /100	2	60	57	54	66	58	63

values. Relative to Native, the no-probe-RGB gains/losses are 64/0, 56/0, and 61/0, with exact paired $p = 1.0842 \times 10^{-19}$, 2.7756×10^{-17}, and 8.6736×10^{-19}, respectively.

Table A19: Validation-selected no-probe-RGB checkpoints for the independent Spatial evaluation. MAE is measured on the supplementary normalized residual target; the baseline predicts zero residual. Reductions are relative to that common validation baseline.

Seed	Selected epoch	MAE	Baseline MAE	Reduction
100	32	0.09078995	0.15077410	39.7841%
101	34	0.09027495	0.15077410	40.1257%
102	35	0.09012678	0.15077410	40.2240%

Object and Goal. Table A20 expands the suite aggregates into task-level counts. Object shows recovery from an all-failure Native baseline, but its wide seed spread limits claims of stable performance across training runs. Goal's positive net recovery includes losses of some Native successes, as shown by the paired counts in Appendix D.3. These observations support reporting the suites separately and examining task and seed heterogeneity alongside their means. We use absolute shifted-view success for this comparison, without normalizing by suite-specific familiar-view performance.

Table A20: Task-level success counts on the separate supplementary Object and Goal banks, with ten fixed starts per task. Object Native succeeds on zero starts for every task. All corrected conditions use suite-specific task-conditioned predictors. Seed columns reuse the same environmental starts within each suite.

	Object aligned			Goal			
Task	100	101	102	Native	100	101	102
0	5	5	8	0	3	2	1
1	0	1	2	0	10	10	10
2	8	1	4	2	7	8	7
3	4	2	1	0	0	1	0
4	7	10	4	1	9	10	6
5	10	4	4	0	7	8	6
6	10	2	7	0	1	3	0
7	9	8	7	4	9	7	10
8	9	7	8	0	8	9	8
9	10	0	2	0	5	1	0
Total /100	72	40	47	7	59	59	48

D.4.3 Held-Out Tasks and Interpretation Limits

The supplementary transfer settings in Appendix A.3 further delimit these gains. A predictor trained only on Spatial Task 6 changes task-level Native counts $1, 0, 1, 1, 0$ to $0, 0, 0, 0, 0$ when deployed without retraining on Tasks 0, 2, 3, 4, and 7, respectively (ten starts per task). A separate text-conditioned predictor trained on four tasks obtains $0/10$ on held-out Task 7 for each of five optimization seeds, reusing the same ten states. The available records support these aggregate counts, but not additional row-level paired inference. Together with the persistent Spatial Task 1 failures, they show that recurrence of recoverable distortion on calibrated tasks does not imply successful correction on unseen tasks.

The unit of environmental replication throughout this section is the fixed start. Optimization seeds characterize training sensitivity and do not triple the environmental sample size. Exact paired tests condition on discordant outcomes for the same starts; the reported comparisons are exploratory and are not adjusted for multiplicity. Seed standard deviations and the task/state/seed bootstrap answer different questions from the primary test's task-stratified episode bootstrap, so their uncertainty summaries are retained within their own cohorts.

D.5 SUSTAINED OPENVLA DIRECTION, TIMING, AND CLIPPING AUDIT

The 50-state directional subset contains ten starts from each of Tasks 0, 2, 3, 6, and 7. Native is executed once and each of four interventions for three optimization seeds, giving $50+4\times3\times50 = 650$ traces. Every trace passes state-identity, schema, and completeness checks. Let $\widehat{r}_t$ be the predicted seven-dimensional residual. Native, Aligned, Reverse, and Delay request 0, $\widehat{r}_t$, $-\widehat{r}_t$, and $\widehat{r}_{t-1}$, respectively, with $\widehat{r}_{-1} = 0$.

For Orthogonal, let $b = \widehat{r}_{t,1:6}$. When $\|b\|_2 > 10^{-12}$, basis indices are visited in ascending order of $|b_j|$. The first basis vector with

$$v = e_j - \frac{e_j^\top b}{b^\top b} b, \qquad \|v\|_2 > 10^{-10}$$

sets the arm request to $\|b\|_2 v/\|v\|_2$ and the gripper request to zero. This tests one deterministic perpendicular direction with the aligned pre-clipping arm norm. All modes execute

$$a^{\text{exec}}_{t,1:6} = \text{clip}(a^{+30}_{t,1:6} + u_{t,1:6}, -1, 1),$$

with raw gripper values clipped to $[0, 1]$ before environment-specific normalization and sign inversion. This sustained six-dimensional control differs from the primary one-step translation experiment with equal final executed magnitudes.

The fidelity audit covers the first 50 control steps per intervention trace. Reconstructing executed arm actions from the stored Native action, applied residual, and clipping yields zero error. An induced-boundary event occurs when an executed arm coordinate is within 10^{-8} of ±1 and its Native counterpart was not already on the boundary. Table A21 distinguishes pre-clipping geometry from delivered perturbations.

Table A21: Sustained-control execution audit. Delivered arm, translation, and rotation entries are mean norms in normalized output coordinates after clipping; they are not physical-unit errors.

Mode	Pre-clip cosine	Norm ratio	Delivered arm	Translation	Rotation	Boundary (%)
Aligned	1.0000	1.0000	.369741	.364412	.046922	3.267
Reverse	-1.0000	1.0000	.372108	.367508	.046092	30.893
Orthogonal	≈ 0	1.0000	.441900	.049567	.401980	4.387
Delay	.685308	1.196562	.344149	.338695	.046705	10.120

Reverse changes saturation exposure as well as direction. Orthogonal has a boundary fraction closer to Aligned, but redistributes the perturbation across action components and does not preserve final executed magnitudes. The task/state hierarchical-bootstrap 95% intervals for Aligned minus Native, Reverse, Orthogonal, and Delay are $[61.33, 93.33]$, $[64.00, 98.00]$, $[64.00, 96.67]$, and $[-5.33, 16.00]$ percentage points, respectively (20,000 resamples). Thus the delay comparison remains compatible with one-step tolerance and does not establish exact timing necessity. Repeated seeds reuse the same starts.

Figure A10 shows how the nominal control construction translates into executed actions. In particular, exact pre-clipping orthogonality does not imply equality of the post-clipping arm norm or its translation/rotation allocation. These checks limit the interpretation of the sustained controls and complement the primary translation-only experiment.

Measured common-start trajectory. Figure A11 shows the archived Task-0 state 000 with predictor seed 100. Native fails at the 220-step horizon, whereas Aligned succeeds at step 75. All modes start from the same recorded end-effector position. The world-xy and world-xz paths use measured simulator coordinates, and separation from Native is computed only over the shared recorded prefix. Successful traces stop at their recorded terminal state; they are neither extrapolated nor padded to the longer horizon.

The example is fixed by the archived task/state key and seed, rather than chosen from the rendered paths. It illustrates how the audited output intervention changes one common-start trajectory. It contributes no additional sample to the success-rate estimates and is not assumed to represent the other 49 starts. Cross-suite trajectory candidates with failed same-start checks are excluded from this illustration; their mismatch is not treated as an intervention effect.

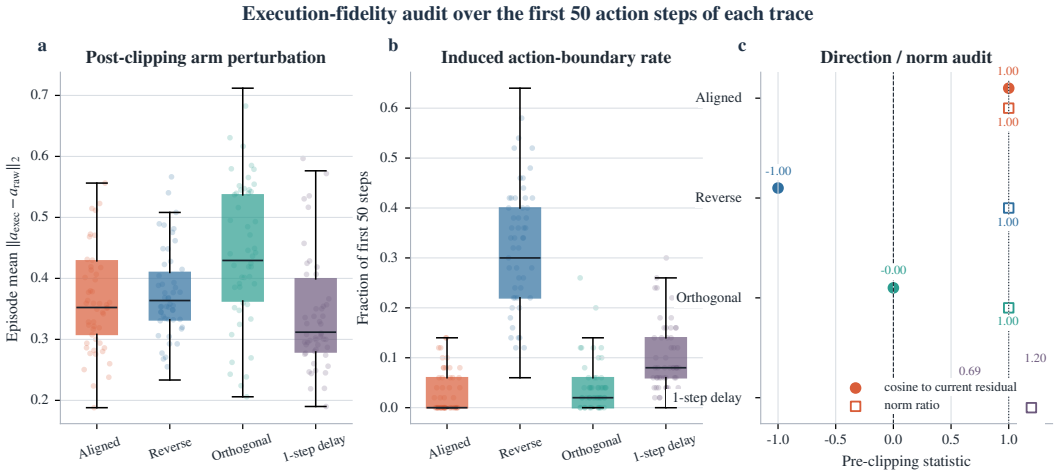


Figure A10: Execution audit over the first 50 control steps of the supplementary sustained-intervention traces. Panels show post-clipping arm perturbations, induced action-boundary incidence, and the requested direction/norm statistics. Aligned, Reverse, and Orthogonal match their pre-clipping constructions. Reverse has greater boundary exposure; Orthogonal changes component allocation and does not preserve final executed norms. These are supplementary six-dimensional arm controls.

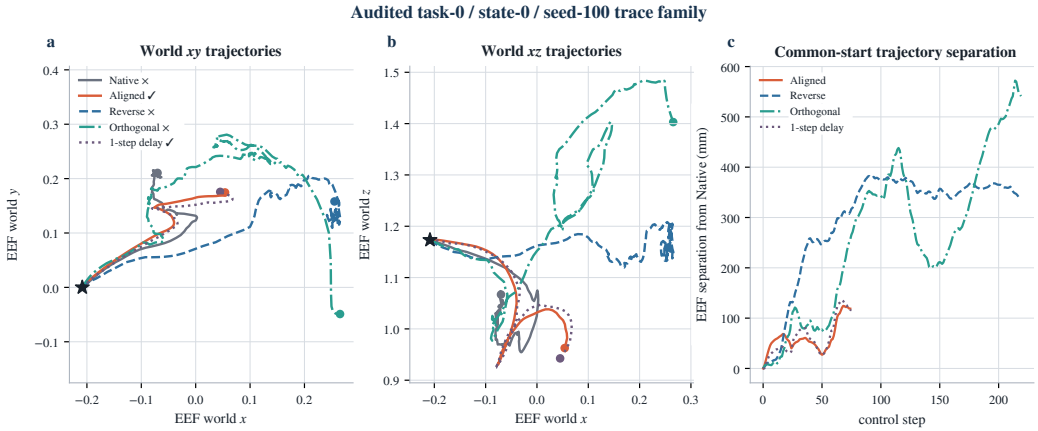


Figure A11: Measured end-effector trajectories from the fixed supplementary Task-0 state 000, seed 100. Left and middle: world-xy and world-xz paths for Native and sustained interventions. Right: separation from Native over each shared recorded prefix. Endpoints mark the last available sample; successful paths terminate at the environment success signal. All conditions share the same audited start, and this single example is separate from aggregate success estimation.

D.6 Fast-WAM Sustained Correction Mechanics

This controller audit applies to the supplementary 300-start direction/phase aggregate in Table A5; it is separate from the primary 500-start test and 200-state local interventions. Implementation-level mechanics and condition-level outcomes are distinct evidence: aggregate success totals do not supply per-state pairings, per-task outcomes, clipping frequencies, or uncertainty intervals.

Correction and conditions. At verified lift, the evaluated v3 implementation freezes a predicted world-xy bias and requests $-\alpha$ times that bias, rotated by the condition's world-xy angle. Table A22 defines the conditions. The orthogonal scale and effective 4.0231-mm step limit match the action-layer norm after world/action calibration. A separate Task-7 resolver configuration is not pooled with this v3 matrix.

Table A22: Fast-WAM sustained-control definitions in the evaluated v3 implementation. Phase-matched Aligned uses descent; the lift condition delivers the same aligned request earlier.

Condition	α	Rotation	Trigger	Interpretation
Native	0	0°	descent	Zero-dose hook control
Aligned	$+1$	0°	descent	Predicted cancellation
Reverse	-1	0°	descent	Opposite nominal direction
Orthogonal	$+1.005779$	90°	descent	Action-norm-matched request
Lift phase	$+1$	0°	lift	Immediate delivery after lift

Phase detector and delivered motion. The moved object is chosen by maximum observed three-dimensional displacement, and its identity is frozen once motion exceeds 50 mm. Correction is frozen once at verified lift. The descent trigger waits for the end effector to fall at least 20 mm below its post-lift peak and to decrease in height by more than 0.1 mm on each of three consecutive control steps. This lift-versus-descent comparison differs from OpenVLA's one-step delay.

Each intervention step has a nominal 4-mm world-xy limit, a 50-mm total requested-correction cap, and at most 20 intervention steps. The controller maps requests to action coordinates, clips action xy to $[-1, 1]$, maps the delivered change back to world coordinates, and subtracts delivered rather than requested motion from the remainder. The calibration is

$$\begin{bmatrix}\Delta x_{\text{world}}\\ \Delta y_{\text{world}}\end{bmatrix} = \begin{bmatrix}.0120438824 & .0001486760\\ .0008958524 & .0119420386\end{bmatrix}\begin{bmatrix}\Delta x_{\text{action}}\\ \Delta y_{\text{action}}\end{bmatrix}.$$

It has condition number 1.09147 and determinant 1.43695×10^{-4}. A clipping flag marks a requested-versus-delivered world difference above 10^{-5} m; completion tolerance is 10^{-6} m. Execution stops at the first environment termination. A successful terminal-partial delivery remains a success, with its requested/delivered/remaining snapshot retained. The compatibility fields $k = 10$ and gate 0.12 m are inactive in the evaluated `pred_cancel` configuration; the replanning field is separate, and these compatibility values do not specify the active chunk horizon or activation threshold.

D.7 Supplementary Inference and Evidence Scope

The supplementary action summaries are nested by task, episode, and snapshot. Continuous intervals resample complete source episodes, preserving all 22 selected snapshots; task-stratified summaries weight tasks equally. The 35 signed-coordinate intervals are descriptive and unadjusted. For paired binary outcomes, the same environmental state identifies each pair, and discordant success counts determine the exact two-sided McNemar test when the required state-level pairs are available. Optimization-seed standard deviations quantify variation between fitted predictors, rather than uncertainty from additional environmental starts.

The strict and independent Spatial banks, the initial-state figure assets, the selected 75-pair timing audit, and the signed-angle sweeps remain separate. Object and Goal instantiate the procedure with their own policies and fitted predictors, so recurrence across these suites does not demonstrate transfer of one frozen calibration map. Benchmark and Fast-WAM aggregate records are reported at their available granularity; marginal success totals alone do not reconstruct paired transitions. The supplementary direction controls act throughout execution and use their documented clipping and phase conventions, whereas the primary finite interventions have explicit continuation policies and short replacement durations.

These distinctions also guide interpretation of negative results: an unresolved task is retained, absence of a detectable full-versus-no-RGB difference is not equivalence, and a delay interval containing zero does not establish precise timing necessity. Familiar-view behavior is a policy-relative reference throughout, not an oracle or a claim of global optimality.

E Additional Outcome-Level Benchmark Context

Table A23 records the reported LIBERO-Plus-protocol aggregates (Fei et al., 2026). They provide broader context for viewpoint sensitivity across policy variants. These benchmark rates use their own original condition and perturbation protocol; they are separate from the LIBERO signed-azimuth scans, the primary 500-start calibration tests, and the supplementary matched-state banks. They do not establish paired action differences or the effects of our action interventions.

Model definitions follow the original publications for OpenVLA-OFT (Kim et al., 2025a), π_0-FAST (Pertsch et al., 2025), NORA (Hung et al., 2025), WorldVLA (Cen et al., 2025), UniVLA (Bu et al., 2025), and RIPT-VLA (Tan et al., 2025); the OFTw/OFTm rows retain the variant labels used by LIBERO-Plus. These citations identify model origins. All numerical values in Table A23 remain attributed to the LIBERO-Plus evaluation (Fei et al., 2026), rather than to a newly conducted experiment in this paper.

Table A23: Reported LIBERO-Plus-protocol success rates (%). Each perturbation column denotes one evaluation family. These outcome-level aggregates are retained as benchmark context.

Model	Original	Camera	Robot	Language	Light	Background	Noise	Layout
OpenVLA	76.5	1.1	4.1	26.8	4.4	25.3	19.3	31.6
OpenVLA-OFT	97.1	59.7	37.2	81.5	85.8	92.4	76.7	77.1
OpenVLA-OFTw	95.3	16.8	43.7	73.2	68.2	92.5	51.4	72.3
OpenVLA-OFTm	97.6	57.9	30.6	83.6	91.6	83.6	76.3	73.2
π_0	94.2	15.8	6.6	61.0	79.6	78.5	79.4	70.4
π_0-fast	85.5	66.4	24.8	63.3	73.0	67.7	75.8	70.3
Nora	87.9	4.0	41.1	67.0	31.0	50.5	17.6	63.9
WorldVLA	79.1	0.3	30.2	44.2	29.4	14.5	12.2	39.4
UniVLA	95.2	4.3	50.3	71.8	59.1	80.0	25.3	34.3
RIPT-VLA	97.5	58.3	36.7	80.1	87.9	90.4	73.8	76.5

Across the ten rows, the Original minus Camera drops are 75.4, 37.4, 78.5, 39.7, 78.4, 19.1, 83.9, 78.8, 90.9, and 39.2 percentage points. The variation motivates policy-specific diagnosis rather than treating viewpoint robustness as a uniform model property.