LIBERO-MAX: DO ROBOT POLICIES ADAPT WHEN THE WORLD CHANGES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Robots must often continue a task after a target moves, the viewpoint shifts, or an obstacle appears, even though their earlier observations and committed actions reflect the previous scene. Many simulation robustness benchmarks fix external conditions at reset, leaving this temporal challenge underexamined. We introduce LIBERO-MAX, a benchmark of 8,000 paired cases spanning eight types of changes to geometry, observations, appearance, clutter, and paths. Each pair compares task execution with and without a mid-task event, holding the task, initial state, policy seed, and pre-event action sequence fixed. This controlled comparison distinguishes event-associated regressions from failures already present without the change. Across fourteen current VLA, hybrid, and world-action policies, events reduce success by 11.0–25.7 percentage points. Event profiles reveal shared vulnerabilities to geometry and observation changes, while policy-family rankings interleave. Camera controls show that robustness reflects both competence under the changed conditions and the trajectory from which they are encountered; varying query cadence does not eliminate the gap. Together, the paired protocol and temporal diagnostics establish LIBERO-MAX as a reproducible testbed for diagnosing failures under mid-execution changes and measuring progress toward robot policies that remain effective as the world changes.

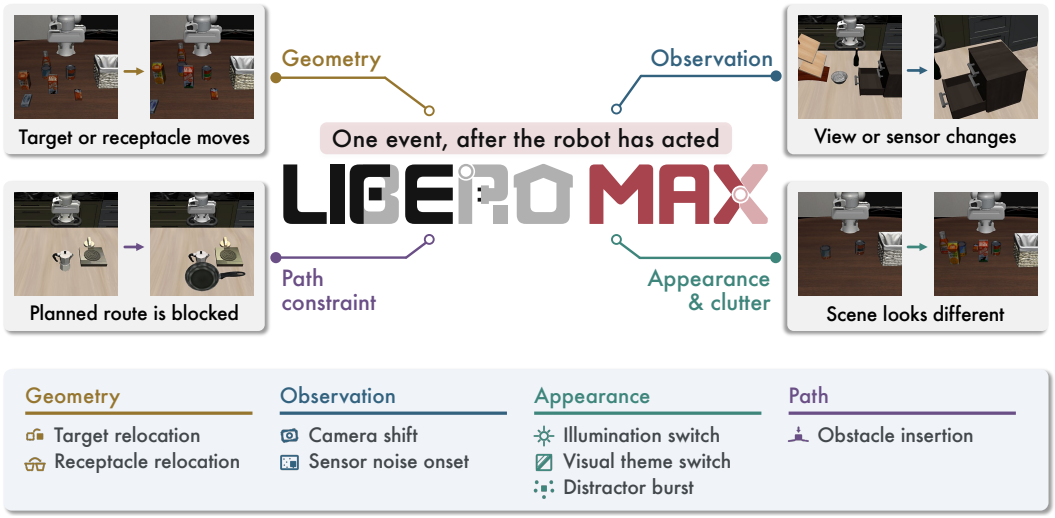


Figure 1: **LIBERO-MAX evaluates whether robot policies can complete tasks when the world changes during execution.** Each case pairs two rollouts with identical pre-event actions: Base continues without an added event, while Dynamic receives one mid-task change. Before-and-after images illustrate four event families: geometry, observation, appearance and clutter, and path constraints. The lower panel lists all eight event types.

1 INTRODUCTION

Reliable robot control requires completing tasks even when the conditions that guided earlier actions no longer hold. A target may move during a reach, a camera may shift, or an obstacle may

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.

appear along the intended path. Simulation benchmarks provide controlled, repeatable settings for evaluating robot behavior (Yu et al., 2020; James et al., 2020; Mees et al., 2022; Mu et al., 2021; Gu et al., 2023). Yet, robustness evaluations often configure external conditions before the first action and retain them throughout the rollout. This leaves a distinct question unresolved: *How reliably can current robot policies complete a task when the world changes after they have begun to act?*

The timing of a change matters because it determines the history from which the policy must respond. A change introduced at reset tests whether a policy can operate under altered conditions from its first observation. A mid-task event requires the policy to continue from a state reached using earlier observations and actions. A camera shift changes the visual evidence available for control, while target relocation can invalidate a reach already in progress. Progress made before the change can also leave the robot closer to completing the task. For chunked policies, previously committed actions may continue to execute before the policy receives a post-event observation (Black et al., 2025b; Kim et al., 2025). Moreover, no fixed action-chunk length is uniformly optimal across tasks (Liang et al., 2026). Task success therefore reflects both competence under the changed conditions and the state and action commitments inherited from earlier execution.

Existing benchmarks reveal failures under visual, geometric, linguistic, and semantic variations (Fei et al., 2025; Zhou et al., 2025b; Wang et al., 2026), while DynamicVLA and DOMINO directly study moving-object manipulation and reactive control (Xie et al., 2026; Fang et al., 2026b). These settings motivate a complementary evaluation: measuring the effect of a specific mid-execution event by comparing task completion with and without that event after the same observed and executed history. Independently sampled rollouts can diverge before the change, confounding its contribution with reset difficulty, policy randomness, and earlier actions. A controlled comparison must therefore hold the pre-event history fixed and measure whether task completion is preserved.

In this work, we introduce **LIBERO-MAX**, a benchmark for controlled evaluation of robot policies under mid-execution environmental changes. LIBERO-MAX comprises 8,000 rollout pairs, derived from LIBERO-Plus and LIBERO-PRO and balanced across eight event types. Each pair shares the task, initial simulator state, instruction, policy seed, and all actions executed before the event. This shared action sequence, the *executed prefix*, establishes a common history for two continuations: **Base** proceeds without an added event, while **Dynamic** receives one event change and continues in the altered scene. For example, both rollouts approach the same target, which is then relocated only in Dynamic. Comparing their outcomes isolates the effect of introducing the event on task completion, holding the preceding execution fixed. Fig. 1 illustrates the changes, and Fig. 2 previews their aggregate effect.

We evaluate fourteen pretrained policies spanning vision-language-action (VLA) models, hybrid predictive, and world-action model (WAM) designs. Every policy loses 11.0–25.7 percentage points in success rate, and 20.8%–56.1% of episodes solved in Base fail after the event. MolmoAct2 and $\pi_{0.5}$ achieve the highest Dynamic success rates, but the remaining family rankings interleave: current WAMs are neither uniformly stronger nor uniformly weaker than VLAs. Relocation and camera/sensor changes produce the largest shared losses. On the same 800 paired cases, we vary how many actions policies execute before receiving a new observation and predicting their next actions. Changing this interval affects performance, but at every tested setting, success remains lower when an event occurs than when it does not. More frequent feedback therefore does not automatically eliminate the failures. Our contributions can be summarized as follows:

- **(1) A paired benchmark for robustness during execution.** We introduce LIBERO-MAX, comprising 8,000 rollout pairs across diverse event types. Its Base/Dynamic protocol holds the executed action prefix fixed, isolating the effect of environmental changes introduced after action commitment.
- **(2) A systematic study of how online changes disrupt task success.** Across pretrained VLA, hybrid, and WAM policies, every paired success-rate gap is negative and statistically resolved. Paired outcome transitions quantify how often episodes that succeed in Base fail after the event.
- **(3) Diagnostics that connect event sensitivity with response timing.** We analyze event profiles and paired outcome transitions with trigger/response coverage, stale-action exposure, and query-cadence sweeps. These analyses reveal shared vulnerabilities to geometry and observation changes, interleaved VLA and WAM rankings, and persistent performance gaps across all tested feedback intervals.

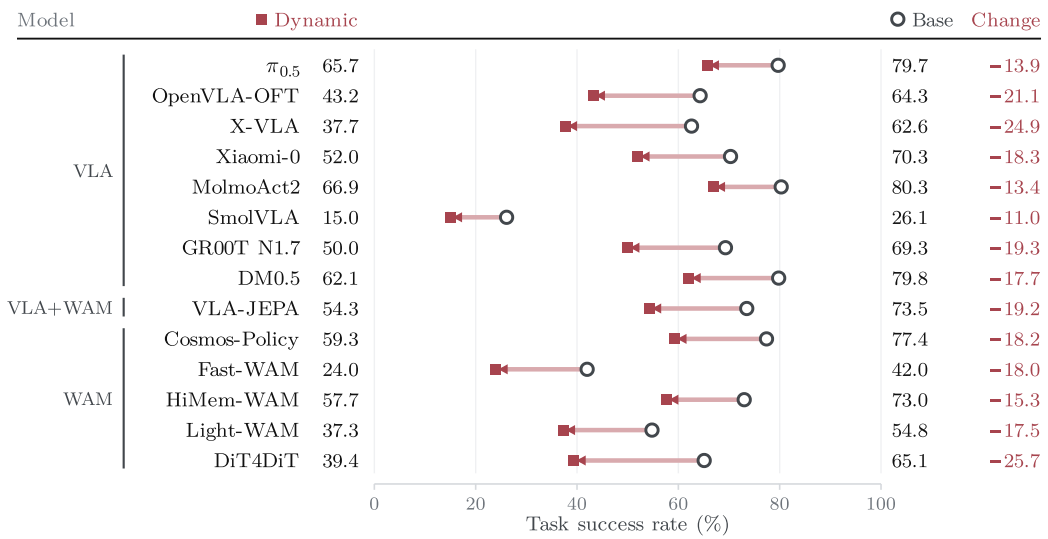


Figure 2: **One controlled mid-task change reduces success for every policy.** We evaluate each policy on the same 8,000 pairs using its released inference settings. Arrows connect Base (open circle) to Dynamic (filled square); columns report success rates (%) and paired Dynamic minus Base changes in percentage points. Each Base/Dynamic pair shares the task, initial state, policy seed, and executed prefix, so the difference measures the effect of adding the event.

2 Related Work

Robot benchmarks and online changes. Simulation benchmarks support robot learning through controlled, repeatable evaluations of transfer and language-conditioned manipulation across diverse tasks (Yu et al., 2020; James et al., 2020; Mees et al., 2022; Gu et al., 2023; Nasiriany et al., 2024). LIBERO and its Plus and PRO extensions evaluate transfer, OOD robustness, and memorization (Liu et al., 2023; Fei et al., 2025; Zhou et al., 2025b); related stress tests examine visual and embodiment variation (Pumacay et al., 2024; Wang et al., 2025b; Zhang et al., 2024; Li et al., 2024c). ORION studies generalization from one human video to novel objects and varied scenes (Zhu et al., 2026). These evaluations usually set external conditions at reset. DynamicVLA and DOMINO study moving-object manipulation and reactive control (Xie et al., 2026; Fang et al., 2026b). A matched comparison can distinguish the effect of a specific mid-task event from differences in the preceding execution. LIBERO-MAX provides this comparison by pairing no-event and event rollouts after an identical executed prefix across abrupt nuisance, observation, geometry, and path changes.

Policies, prediction, and action commitment. VLA policies combine language and visual representations with action prediction (Kim et al., 2024; 2025; Black et al., 2025b;a; Zheng et al., 2026). World-action and hybrid policies further connect prediction to control through video generation, latent actions, joint training, or memory (Hu et al., 2025; Cen et al., 2025; Bi et al., 2025; Sun et al., 2026b; Li et al., 2026b). Chunked policies execute several predicted actions before observing and querying again, trading fewer policy calls for delayed feedback. Adaptive chunking adjusts this commitment across tasks (Liang et al., 2026). Success under conditions fixed at reset leaves open whether policies remain effective after changes during execution. LIBERO-MAX compares task completion after an external event across these policy designs and measures how query cadence changes outcomes. Its fixed event boundary supports controlled evaluation of future methods that cancel pending actions or query again after a change. Appendix A covers additional related work.

3 The LIBERO-MAX Benchmark

3.1 Paired Evaluation

LIBERO-MAX measures the effect of a mid-task event by comparing two rollouts with the same *executed prefix*, the sequence of actions performed before the event. For a fixed policy, each case specifies a task, instruction, initial state, policy seed, and event. We run each case in two conditions.

Table 1: **Benchmark scope and evaluation unit.** ✓ marks a directly evaluated property. LIBERO scores 2,000 task episodes, LIBERO-Plus 10,030 generated variants, LIBERO-PRO 10,000 task episodes, and LIBERO-MAX 8,000 Dynamic episodes with one no-event Base control per case. Only LIBERO-MAX applies an online change after an identical executed prefix.

Benchmark	Static OOD	Online change	Paired control	Exact prefix	Dynamic metrics	Scored episodes
LIBERO	×	×	×	×	×	**2,000**
LIBERO-Plus	✓	×	×	×	×	**10,030**
LIBERO-PRO	✓	×	×	×	×	**10,000**
LIBERO-MAX	✓	✓	✓	✓	✓	**8,000**

Base retains the original source configuration without an additional mid-task event, including any variations inherited from LIBERO-Plus or LIBERO-PRO. Dynamic replays the Base actions up to trigger time t_i, then receives the event and continues in the changed environment.

Let $Y_i^0, Y_i^1 \in \{0, 1\}$ denote task completion in Base and Dynamic. Over N pairs, the average success-rate change is $\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^1 - Y_i^0)$. We report $100\widehat{\Delta}$ in percentage points. We retain every assigned case and its observed task outcome, including cases that do not reach the event trigger or obtain a post-event query. The four outcomes are preserved success $(1, 1)$, change-associated success $(0, 1)$, event-associated regression $(1, 0)$, and persistent failure $(0, 0)$.

The shared executed prefix is essential: independently sampled rollouts can approach the same task differently, whereas replay fixes the actions preceding the event. Beyond success rates, we report the fraction of all assigned cases that trigger the event and the fraction with a subsequent policy query. These measures distinguish reaching the change from obtaining post-event feedback. If a policy never reaches the trigger region, its task outcome still contributes to the success rate, while trigger coverage records that no event occurred. Figs. 1 and 3 show pairing and case construction.

3.2 Construction and evaluation tracks

Max contains 5,600 cases from seven LIBERO-Plus categories and 2,400 from ten LIBERO-PRO categories, giving 16,000 Base/Dynamic rollouts per policy. Each of the eight event types has 1,000 pairs, split 700/300 between sources, with two fixed parameter variants. We select cases without using policy outcomes and replace invalid configurations within the same source category, event type, and parameter variant. Before evaluation, we check simulator validity, including support, workspace, visibility, and collision constraints for geometry changes, and retain only configurations that pass. Table 1 compares evaluation scope, and Fig. 3 shows the construction.

For rapid iteration, **LIBERO-MAX Lite** fixes 800 pairs selected without policy outcomes, preserving Max's event and source proportions. It requires 1,600 rollouts per policy, one tenth of Max. Across fourteen policies, Base and Dynamic success rates and paired gaps stay within 2.4 percentage points of Max. We retain Max for detailed event and policy comparisons (Appendix C).

3.3 Events and temporal protocol

The eight events are target and receptacle relocation, camera shift, illumination switch, sensor-noise onset, visual-theme switch, distractor burst, and obstacle insertion. Each is introduced once and is not reversed by the evaluator before the rollout ends; this is a *persistent event*. For example, a camera shift retains the new viewpoint, while objects continue to obey the simulator's dynamics. An illumination switch leaves task geometry fixed; target relocation changes where the robot must reach, and obstacle insertion changes the approach region. These events require different responses, from continuing under altered observations to revising the target or path. Every event uses the deterministic `pre_grasp_proximity` trigger, which fires when the end effector first enters the designated entity's neighborhood during approach. Examples are shown in Appendix B.5.

Base records policy queries and actions; Dynamic restores the same initial state and replays the exact prefix before applying the event. The evaluator checks prefix identity and records the trigger and first post-event query. A chunked policy predicts several actions at once, so it may execute actions queued before the event before seeing a new observation. We call their number *stale-action exposure*. All cases count toward success rates, including those without an event trigger or a post-event query. Protocol and validity details are in Appendix B.

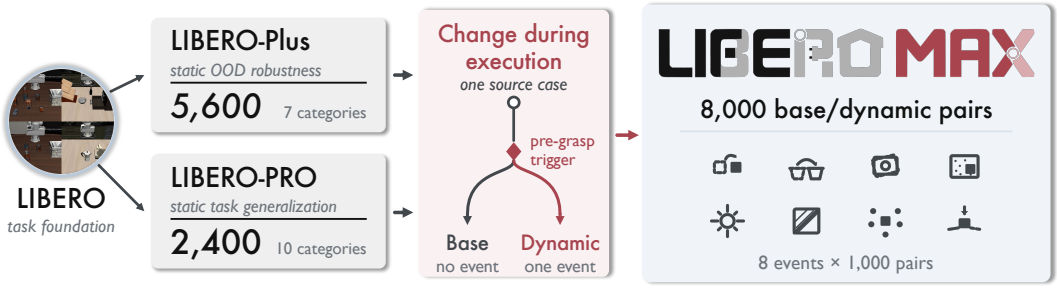


Figure 3: **From static source tasks to changes during execution.** LIBERO-Plus and LIBERO-PRO contribute 5,600 and 2,400 source cases. Each becomes a Base/Dynamic pair with the same actions before one mid-task event. Max contains 8,000 pairs, balanced at 1,000 per event.

4 Results

4.1 Evaluation protocol

We evaluate fourteen pretrained policies on the same 8,000 pairs: eight VLAs, VLA-JEPA with a predictive objective, and five WAMs. Each uses its released serving protocol, so scale, training, and action interfaces remain part of the policy comparison. Mimic-Video is evaluated separately on its supported partition. We report Base and Dynamic success rates, paired changes, conditional regression among Base successes, and event-family breakdowns. Dynamic success measures completion over all assigned cases; conditional regression measures the share of each policy's Base successes lost after the event. Paired changes use 95% bootstrap intervals, and exact McNemar tests compare discordant outcomes. Cross-policy Dynamic contrasts use common case IDs with Holm correction. Model and evaluation details are in Appendix D.

4.2 Online events break previously successful episodes

Every policy is less successful after the event (Table 2, Fig. 2). Even policies with strong Base performance lose success: MolmoAct2 falls from 80.3% to 66.9%, and $\pi_{0.5}$ from 79.7% to 65.7%. Fig. 4 separates cases that fail only after the event from those that already fail in Base. SmolVLA has the smallest aggregate gap, but only 26.1% Base success; the event breaks 56.1% of those successes. A few cases succeed only in Dynamic $(0, 1)$, partly offsetting these losses in the average. Conditional regression counts the share of Base successes lost after the event, making these failures visible even when the average gap is small. Full paired diagnostics are in Appendix E.1.

MolmoAct2 has the highest Dynamic success rate, yet still loses many cases it solves in Base. It preserves 5,091 Base successes but loses another 1,336 after the event; its 264 change-associated successes partially offset those losses in the average. Thus its 13.4-point net decline corresponds to the loss of 20.8% of previously solved episodes. A higher Dynamic rate can come from preserving more Base successes, solving different cases, or both; paired outcomes distinguish these possibilities. Trigger and response coverage range from 87.7% to 96.6% across policies. Most cases therefore reach the event and receive a new observation, yet many still fail. Cases that never reach the trigger still count toward the success rate.

Takeaway 1: Online events break policies that otherwise succeed.

Across policies, events reduce success by 11.0 to 25.7 percentage points and turn 20.8% to 56.1% of Base successes into failures. SmolVLA loses the fewest percentage points overall but the largest share of its Base successes because its Base success rate is low.

4.3 Policy families have no consistent ordering

Dynamic rankings and paired losses interleave across policy families. MolmoAct2 and $\pi_{0.5}$ lead Dynamic success at 66.9% and 65.7%, followed by DM0.5 (62.1%), Cosmos-Policy (59.3%), and HiMem-WAM (57.7%).

Table 2: **Every policy loses success after a controlled online change.** All fourteen rows contain the same 8,000 Base/Dynamic pairs. Overall Δ and the four event-family columns report Dynamic minus Base success in percentage points. Released serving protocols differ across rows, so the table compares policies rather than isolating an architecture-family effect.

Family	Model	Overall success rate (%)			Change family (paired Δ, pp)			
		Base SR ↑	Dynamic SR ↑	Δ	Obs.	Geom.	App. + clutter	Path
VLA	$\pi_{0.5}$	79.7	65.7	−13.9	−23.1	−24.2	−4.2	−4.2
	OpenVLA-OFT	64.3	43.2	−21.1	−32.0	−33.6	−10.3	−6.8
	X-VLA	62.6	37.7	−24.9	−43.0	−52.4	−1.2	−5.1
	Xiaomi-Robotics-0	70.3	52.0	−18.3	−25.2	−37.7	−5.7	−3.9
	MolmoAct2	80.3	66.9	−13.4	−19.3	−26.6	−4.6	−1.8
	SmolVLA	26.1	15.0	−11.0	−16.6	−21.2	−3.3	−2.7
	GR00T N1.7	69.3	50.0	−19.3	−24.6	−45.0	−2.2	−8.3
	DM0.5	79.8	62.1	−17.7	−29.0	−31.3	−5.2	−5.7
VLA+WAM	VLA-JEPA	73.5	54.3	−19.2	−30.3	−33.7	−7.3	−4.0
WAM	Cosmos-Policy	77.4	59.3	−18.2	−25.2	−35.8	−6.3	−4.6
	Fast-WAM	42.0	24.0	−18.0	−35.6	−27.5	−3.9	−6.1
	HiMem-WAM	73.0	57.7	−15.3	−13.3	−39.8	−4.3	−3.3
	Light-WAM	54.8	37.3	−17.5	−21.7	−39.9	−5.0	−2.0
	DiT4DiT	65.1	39.4	−25.7	−49.7	−26.6	−16.2	−4.5

Takeaway 2: Current WAMs do not consistently outperform or underperform VLAs.

The three highest Dynamic success rates belong to VLAs, followed by two WAMs. The remaining rankings and the size of the losses show no consistent VLA–WAM ordering. These policies also differ in training, scale, and inference settings, so the comparison cannot isolate the effect of architecture.

4.4 Geometry and observation changes cause the largest losses

Event profiles reveal where these robustness losses are concentrated. Fig. 5 shows a shared event ordering, with mean pairwise Spearman correlation 0.80 across policies. Geometry losses span 21.2–52.4 points and observation losses 13.3–49.7, exceeding most appearance-and-clutter and path losses. Individual profiles differ: Fast-WAM falls from 43.4% Base success to 0.3% under sensor noise, whereas X-VLA is most sensitive to camera shift. Illumination and theme changes generally preserve more success than relocating the target or destination, but the exceptions matter: sensor noise dominates Fast-WAM and DiT4DiT, while HiMem-WAM is comparatively camera robust. Thus event profiles help identify which changed conditions a policy handles poorly, beyond a single family ranking. Full event-level results are in Appendix E.2.

Source and task variation. The gap persists when source categories receive equal weight: category-macro losses range from 10.7 to 23.4 points across all fourteen policies, and every policy loses success in each of the four LIBERO task suites. The hardest suite nevertheless depends on the policy. MolmoAct2 loses 16.0 points on LIBERO-10 versus 9.6 on LIBERO-Object, whereas Light-WAM loses 25.3 points on LIBERO-Spatial versus 12.1 on LIBERO-10. Breaking down results by event and task shows where each policy struggles, even when its overall success rate is high. Full breakdowns are in Appendix E.4.

Base success also matters when comparing source categories. PRO-derived cases have lower Dynamic success but also lower Base success and smaller paired losses than Plus-derived cases. SmolVLA succeeds on only 0.4% of PRO position cases in Base, so its success rate can barely fall further. Base success and conditional regression help distinguish cases a policy already struggles with from cases it fails only after the event (Appendix E.3).

Event parameters expose further sensitivity. Across policies, the largest within-event difference in Dynamic success between the two fixed parameter variants ranges from 12.2 to 36.4 points; the difference reaches 35.6 points for $\pi_{0.5}$ under camera shift and 36.4 points for HiMem-WAM under receptacle relocation. Success can therefore vary substantially within the same event type. Because

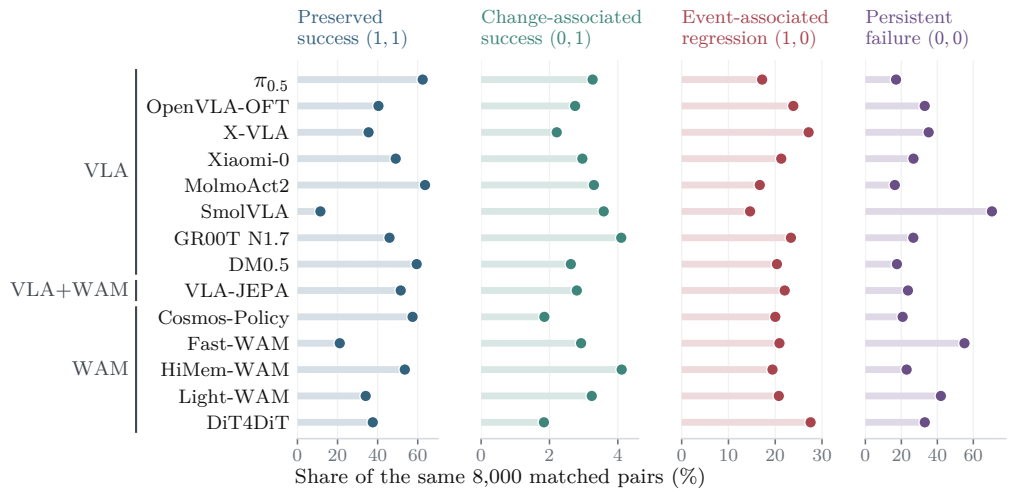


Figure 4: **Paired outcomes separate post-change regression from low Base competence.** Panels partition all 8,000 pairs into the four binary outcomes. Points show episode shares; independently scaled axes retain visibility of rare change-associated successes $(0, 1)$. The $(1, 0)$ group counts cases solved in Base but failed in Dynamic; $(0, 1)$ counts the reverse outcome.

the variants contain different source instances, the comparison does not isolate the effect of changing the parameter. Each policy is evaluated on the same cases from both parameter variants. Reporting the groups separately shows whether an improvement extends across the predefined configurations or is concentrated in one of them. Medium/high severity comparisons also combine different event types and should be read as descriptions of those groups (Appendix E.4).

Takeaway 3: Target relocation causes the largest loss for 11 of 14 policies.

On LIBERO-MAX, target relocation is the most damaging tested event for eleven policies. X-VLA instead loses the most under camera shift, while Fast-WAM and DiT4DiT lose the most under sensor noise. Illumination and visual-theme changes are generally milder.

4.5 Camera controls reveal the roles of viewpoint and execution history

The large losses under observation changes raise a question: how much of the camera-shift difficulty is already present when a policy starts in the changed view? We run an independent control experiment on all 1,000 camera cases for X-VLA, $\pi_{0.5}$, and HiMem-WAM (700 Plus and 300 PRO cases per policy; 9,000 rollouts). Base retains the original camera. Reset applies the camera change before the first policy input. Mid-task follows the Dynamic protocol, applying the same change after replaying the Base actions up to the trigger. Across these three conditions, the cases, initial states, checkpoints, seeds, and success criteria remain fixed.

Table 3: **Changed views can impair success from reset.** SR (%) on all 1,000 cases per policy. Bold: estimates; italic below: 95% source-stratified bootstrap CIs.

Policy	Base	Reset	Mid-task
X-VLA	**68.2** *[65.4, 71.0]*	**1.3** *[0.6, 2.0]*	**12.4** *[10.4, 14.5]*
$\pi_{0.5}$	**79.5** *[77.1, 81.9]*	**51.3** *[48.2, 54.4]*	**55.3** *[52.2, 58.4]*
HiMem-WAM	**72.5** *[69.7, 75.2]*	**68.8** *[65.9, 71.6]*	**72.6** *[69.8, 75.3]*

The changed view can be difficult even from the start (Table 3). X-VLA falls from 68.2% Base success to 1.3% at Reset and reaches 12.4% at Mid-task. Mid-task exceeds Reset by 11.1, 4.0, and 3.8 points for X-VLA, $\pi_{0.5}$, and HiMem-WAM, respectively. Starting in the original view can therefore help a policy make progress before the change. Reset and Mid-task differ in visited states, time spent in the changed view, and remaining task budget, whereas Base and Mid-task share the same executed prefix.

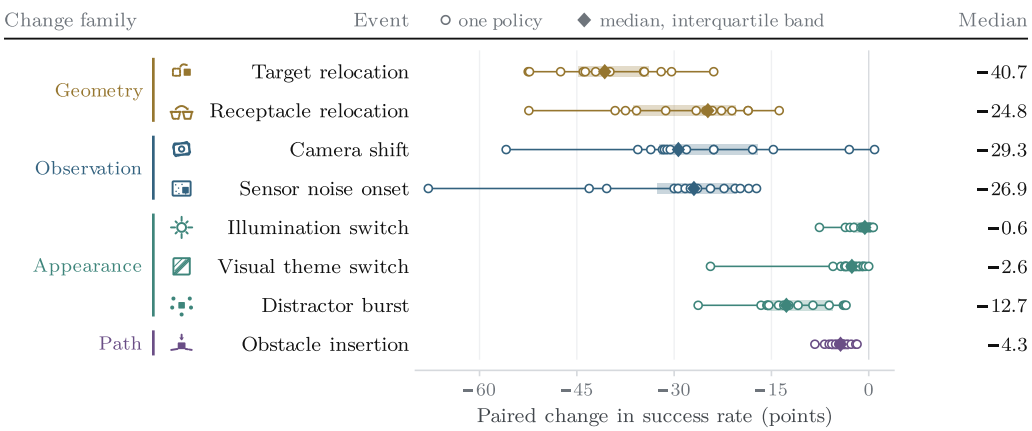


Figure 5: **Geometry and observation changes produce the largest shared losses.** Rows group eight events into four families. Open circles show fourteen policies per 1,000-pair event; diamonds, shading, and thin segments mark the median, interquartile range, and full range.

Consistency across sources and task groups. All three policies have higher Mid-task than Reset success in both the Plus and PRO groups. When we combine the sources, the advantage remains statistically supported after resampling the 40 shared reference tasks, source-specific task groups, or initial-configuration groups. The smaller PRO group gives less precise estimates: its results alone do not establish an advantage for $\pi_{0.5}$ or HiMem-WAM (Appendix F).

Which trajectories lose success? Similar success rates can hide different outcomes on individual cases. HiMem-WAM's nearly unchanged Base/Mid success (72.5%/72.6%) combines 66 cases that lose success with 67 that gain it. Across the three policies, 84 model–case instances succeed in both Base and Reset but fail in Mid-task. All receive a new policy query, and 13 execute no old queued actions after the event. Thus some mid-task failures occur even when the policy can solve the case from the changed view at reset, and stale actions alone cannot explain all of them.

We also check whether the robot reaches the trigger's 18 cm neighborhood. Among the 106 X-VLA cases solved in Base and Mid-task but failed in Reset, 74 Reset trajectories never enter this region; the corresponding count is only 2 of 59 for HiMem-WAM. In these cases, X-VLA often struggles to approach the target from the changed view, whereas HiMem-WAM usually fails after reaching the neighborhood. Entering this region marks approach, not a completed grasp.

The timing advantage persists at milder camera shifts. On the same 1,000 cases, reducing camera magnitude to 25% raises X-VLA Reset success to 36.0% and Mid-task success to 51.7%, compared with 68.2% in Base. Both improve but remain below Base. Mid-task still outperforms Reset when Reset success is well above its full-strength near-zero level. The timing gap peaks at 50% strength rather than increasing monotonically with magnitude. At the milder settings, most Reset failures among cases solved in Base and Mid-task occur after reaching the approach threshold. These controls show that both the viewpoint and the preceding trajectory matter for completing the task (Appendix F).

4.6 Query cadence changes performance without closing the gap

We next examine whether changing the frequency of policy feedback can recover Dynamic success. The action horizon H is the number of actions returned by a policy call; the query interval Q is the number executed before a new observation and call. We vary Q on the fixed 800-pair Lite pool (Fig. 6). X-VLA, GR00T N1.7, and Fast-WAM use $Q \in \{2, 4, 8, 12, 16\}$; $\pi_{0.5}$ uses $\{2, 4, 8\}$ because its $H = 10$ horizon excludes longer prefixes. Fast-WAM fixes $H = 32$, while X-VLA and GR00T N1.7 use $H = Q$. Thus the sweep measures each checkpoint's practical serving protocol.

X-VLA and $\pi_{0.5}$ improve as Q increases over their valid ranges; GR00T N1.7 and Fast-WAM peak in Dynamic success at $Q = 12$ and $Q = 8$. Across all 18 settings, Dynamic remains 11.1–23.0 points below Base. This policy-specific pattern agrees with evidence that the best fixed chunk length varies across tasks (Liang et al., 2026). In the primary evaluation, policies often continue executing queued actions after the event. The median is two actions for $\pi_{0.5}$ and seven for Cosmos-Policy and



Figure 6: **Dynamic remains below Base at every valid query cadence.** Panels use the same 800 matched pairs per policy. Open circles are Base and filled squares Dynamic; Q is the executed actions per query. Only $Q \leq H$ is valid, so $\pi_{0.5}$ stops at $Q = 8$.

Fast-WAM. These actions delay the next observation, but no tested query cadence closes the gap (Appendix G).

For X-VLA, Dynamic success rises from 10.8% at $Q = 2$ to 39.2% at $Q = 16$, although the Base–Dynamic gap grows. A smaller gap alone therefore does not imply better task completion.

Takeaway 4: More frequent queries alone do not close the response gap.

For the six policies with execution traces, the average number of queued actions executed after an event before the next observation ranges from 2.02 to 7.47. On the fixed 800-pair sweep, Dynamic success stays below Base at every tested query interval. The best Q depends on the policy, and changing it improves some results without eliminating the loss.

A candidate response to sensor corruption. We test whether restoring corrupted images can reduce the sensor-noise loss on 300 fixed cases. Before querying X-VLA, a quality gate checks the images and applies inpainting and denoising when needed; the policy itself is not retrained. The gate uses only the current images and receives neither an event flag nor a clean reference image. Table 4 compares native observations, always-on restoration, and quality-gated restoration. The gated response raises Dynamic success from 40.7% to 47.7%, while Base changes from 66.7% to 67.0%. The Base/Dynamic gap therefore narrows from 26.0 to 19.3 points, a paired reduction of 6.7 points (95% CI: [2.3, 11.3]). At the case level, gated restoration converts 33 native Dynamic failures to successes while losing 12 native successes. Always-on restoration also improves Dynamic success, and the data do not establish an advantage for gating over always-on processing. The benchmark can thus measure whether a targeted intervention reduces the loss after a change. Experimental details are in Appendix H.

Table 4: **Restoration recovers part of the sensor-noise loss.** X-VLA on 300 cases. Bold: SR (%) and Dynamic gain (pp) over Native; italic below: paired 95% bootstrap CIs for gains.

Processing	Base	Dynamic	Gain
Native	**66.7**	**40.7**	—
Always-on	**68.3**	**47.3**	**+6.7** *[2.3, 11.0]*
Quality-gated	**67.0**	**47.7**	**+7.0** *[3.0, 11.3]*

5 Conclusion

LIBERO-MAX measures whether robot policies can complete tasks after their environment changes during execution. Its paired Base/Dynamic protocol holds the task, initial state, policy seed, and executed action prefix fixed, allowing direct comparison with a no-event rollout. Across eight event types and fourteen current policies, online changes cause substantial failures, with the largest shared losses under geometry and observation changes. Camera controls show that the changed viewpoint and preceding trajectory both affect success. Paired outcomes reveal which previously successful cases fail, while execution traces show when the policy receives new observations. No policy family consistently outperforms the others, and changing query cadence leaves substantial losses at every tested setting. A simple image-restoration module recovers part of the sensor-noise loss. LIBERO-MAX can therefore support both identifying the changes that disrupt a policy and testing methods that help it continue the task.

AI USE STATEMENT

We used a Large Language Model to assist with language polishing and improving the readability of this manuscript. The authors are fully responsible for all research ideas, experimental results, and claims presented in this paper.

ETHICS STATEMENT

This work evaluates robot policies under controlled changes in simulated manipulation environments. Our aim is to expose robustness failures before deployment on physical robots. Benchmark success does not establish real-world safety: the evaluation does not measure harmful contacts or other safety violations. We use existing simulation assets and pretrained policies with attribution to their original sources.

REPRODUCIBILITY STATEMENT

Appendices B and D document benchmark construction and evaluation protocols. The supplementary results include complete paired outcomes, temporal diagnostics, camera controls, and the sensor-restoration study.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025a. URL https://arxiv.org/abs/2504.16054.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael 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, 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 (RSS)*, 2025b.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022. URL https://api.semanticscholar.org/CorpusID:254591260.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Krzysztof Choromanski, Tianli Ding, Danny Driess, Kumar Avinava Dubey, Chelsea Finn, Peter R. Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Sergey Levine, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Ho Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Ted Xiao, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818, 2023. URL https://api.semanticscholar.org/CorpusID:260293142.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. URL https://arxiv.org/abs/2402.15391.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-Robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026. URL https://arxiv.org/abs/2602.12684.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025a.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19752–19763, October 2025b.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv preprint arXiv:1805.12114*, 2018. URL https://arxiv.org/abs/1805.12114. NeurIPS 2018 (PETS).

Rongxu Cui, Zongzheng Zhang, Jingrui Pang, Haohan Chi, Jinbang Guo, Saining Zhang, Shaoxuan Xie, Xin Jin, Yao Mu, Jiaolong Yang, Guocai Yao, Xianyuan Zhan, Ya-Qin Zhang, and Hao Zhao.

Libero-safety: A comprehensive benchmark for physical and semantic safety in vision-language-action models. *arXiv preprint arXiv:2606.23686*, 2026.

Guangyuan Dong, Chuang Liu, Yangchen Zeng, Haoyu Wang, Xiaoyang Yu, Pinlong Zhao, Yuchao Hou, Ziwei Li, and Zheng Lin. When generated images look right and retrieve wrong: Coverage-guided cross-scale re-indexing for knowledge-faithful generative perception. *arXiv preprint arXiv:2608.20810*, 2026. URL https://arxiv.org/abs/2608.20810.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8469–8488, 2023.

Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, Shanli Xing, Jaemin Cho, Jae Sung Park, Ainaz Eftekhar, Peter Sushko, Karen Farley, Angad Wadhwa, Cole Harrison, Winson Han, Ying-Chun Lee, Eli VanderBilt, Rose Hendrix, Suveen Ellawela, Lucas Ngoo, Joyce Chai, Zhongzheng Ren, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoAct2: Action reasoning models for real-world deployment. *arXiv preprint arXiv:2605.02881*, 2026a. URL https://arxiv.org/abs/2605.02881.

Heng Fang, Shangru Li, Shuhan Wang, Xuanyang Xi, Dingkang Liang, and Xiang Bai. Towards generalizable robotic manipulation in dynamic environments. *arXiv preprint arXiv:2603.15620*, 2026b. URL https://arxiv.org/abs/2603.15620.

Irving Fang, Juexiao Zhang, Shengbang Tong, and Chen Feng. From intention to execution: Probing the generalization boundaries of vision-language-action models. *arXiv preprint arXiv:2506.09930*, 2025.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. URL https://arxiv.org/abs/2510.13626.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. *arXiv preprint arXiv:1610.00696*, 2017. URL https://arxiv.org/abs/1610.00696. ICRA 2017.

Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8996–9002. IEEE, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104. DreamerV3 (arXiv).

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2024. URL https://arxiv.org/abs/2310.16828. ICLR 2024.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=c0dhw1du33.

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.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *arXiv preprint arXiv:1906.08253*, 2019. URL https://arxiv.org/abs/1906.08253. NeurIPS 2019.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, 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 *Conference on Robot Learning*, 2024. URL https://arxiv.org/abs/2406.09246.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. URL https://arxiv.org/abs/2502.19645.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wPEIStHxYH.

Lingdong Kong, Ao Liang, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Xian Sun, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, Dongyue Lu, Youquan Liu, Guangfeng Jiang, Linfeng Li, Xiangtai Li, Long Zhuo, Lai Xing Ng, Benoit R. Cottereau, Changxin Gao, Liang Pan, Wei Tsang Ooi, and Ziwei Liu. Is your driving world model an all-around player? *arXiv preprint arXiv:2605.10858*, 2026. URL https://arxiv.org/abs/2605.10858.

David Kube, Simon Hadwiger, and Tobias Meisen. Beyond performance: Explaining generalisation failures of robotic foundation models in industrial simulation. *Biomimetic Intelligence and Robotics*, pp. 100249, 2025.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024b.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024c.

Ziang Li, Dongzhou Cheng, Yibin Wang, Shiyue Wang, Xiaoyang Xu, Lingxuan Weng, Juan Wang, and Jiaqi Wang. Light-WAM: Efficient world action models with state-fusion action decoding. *arXiv preprint arXiv:2606.08242*, 2026b. URL https://arxiv.org/abs/2606.08242.

Yuanchang Liang, Xiaobo Wang, Kai Wang, Shuo Wang, Xiaojiang Peng, Haoyu Chen, David Kim Huat Chua, and Prahlad Vadakkepat. Adaptive action chunking at inference-time for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20802–20811, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liang_Adaptive_Action_Chunking_at_Inference-time_for_Vision-Language-Action_Models_CVPR_2026_paper.html.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. URL https://arxiv.org/abs/2306.03310.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. Dual-pathway circuits of object hallucination in vision-language models. *arXiv preprint arXiv:2605.13156*, 2026a. URL https://arxiv.org/abs/2605.13156.

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. *arXiv preprint arXiv:2505.19789*, 2025.

Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, Fangqiang Ding, and Ismini Lourentzou. PALM: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28096–28110, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_PALM_Progress-Aware_Policy_Learning_via_Affordance_Reasoning_for_Long-Horizon_Robotic_CVPR_2026_paper.html.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. DiT4DiT: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026. URL https://arxiv.org/abs/2603.10448.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Jorge A Mendez, Marcel Hussing, Meghna Gummadi, and Eric Eaton. Composuite: A compositional reinforcement learning benchmark. *arXiv preprint arXiv:2207.04136*, 2022.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

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

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint 2512.15692*, 2025. URL https://arxiv.org/abs/2512.15692.

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. *ArXiv*, abs/2501.09747, 2025. URL https://api.semanticscholar.org/CorpusID:275570494.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019. URL https://arxiv.org/abs/1911.08265. MuZero.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James M. Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17929–17960, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1a17a06de88cf77f25cda0da91615a54-Abstract-Conference.html.

M Shukor, D Aubakirova, F Capuano, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026a. URL https://arxiv.org/abs/2602.10098.

L Sun, B Xie, Y Liu, et al. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

Xiaoquan Sun, Ruijian Zhang, Chen Cao, Yihan Sun, Jiahui Chen, Zetian Xu, Bo Chen, Haijier Chen, Zhen Yang, Jiarun Zhu, Yijun Hong, JingZhe Xu, Jingrui Pang, Mingqi Yuan, and Jiayu Chen. HiMem-WAM: Hierarchical memory-gated world action models for robotic manipulation. *arXiv preprint arXiv:2606.10363*, 2026b. URL https://arxiv.org/abs/2606.10363.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Guodong Wang, Chenkai Zhang, Qingjie Liu, Jinjin Zhang, Jiancheng Cai, Junjie Liu, and Xinmin Liu. LIBERO-X: Robustness litmus for vision-language-action models. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.097.

Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics, 2025a. URL https://arxiv.org/abs/2411.13587.

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Vlatest: Testing and evaluating vision-language-action models for robotic manipulation. *Proceedings of the ACM on Software Engineering*, 2(FSE):1615–1638, 2025b.

Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*, 2025.

Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, and Ziwei Liu. Dynamicvla: A vision-language-action model for dynamic object manipulation. *arXiv preprint arXiv:2601.22153*, 2026. URL https://arxiv.org/abs/2601.22153.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Yifan Ye, Yankai Fu, Yaoxu Lv, Bohan Hou, Jun Cen, Lingdong Kong, Duo Zheng, Tianxing Chen, Jiaming Liu, Ziang Cao, Yunfan Lou, Wei Chow, Xian Sun, Yingshuo Wang, Kuangzhi Ge, Xiaowei Chi, Xidong Zhang, Zhibo Pang, Yiwu Zhong, Sirui Han, Zhihe Lu, Weihao Yuan, Qifeng Chen, Michael Yu Wang, Yao Mu, Ziwei Liu, Jianfei Yang, Ping Luo, and Shanghang Zhang. Data pyramid for embodied manipulation: A survey. *arXiv preprint arXiv:2607.24744*, 2026b. URL https://arxiv.org/abs/2607.24744.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

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.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024.

Q Zhao, Y Lu, M J Kim, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Tai Wang, Ya-Qin Zhang, Jingjing Liu, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kt51kZH4aG.

Jiaming Zhou, Ke Ye, Jiayi Liu, Teli Ma, Zifan Wang, Ronghe Qiu, Kun-Yu Lin, Zhilin Zhao, and Junwei Liang. Exploring the limits of vision-language-action manipulations in cross-task generalization. *arXiv preprint arXiv:2505.15660*, 2025a.

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

Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *Autonomous Robots*, 50(2):27, 2026. doi: 10.1007/s10514-026-10253-8. URL https://doi.org/10.1007/s10514-026-10253-8.

APPENDIX

A Additional Related Work 18
- A.1 Robot benchmarks and dynamic manipulation 18
- A.2 Object-centric generalization and changes during execution 18
- A.3 Vision-language-action policies and action commitment 18
- A.4 World models and world-action policies 18

B Benchmark Construction and Intervention Examples 19
- B.1 Design requirements 19
- B.2 Source selection and balancing 19
- B.3 Events, triggers, and prefix replay 20
- B.4 Validity and treatment of failures 20
- B.5 Qualitative intervention examples 21

C LIBERO-MAX Lite 22
- C.1 Outcome-independent subset 22
- C.2 Agreement with the full benchmark 22

D Evaluation Protocol and Policy Details 22
- D.1 Metrics and statistical comparisons 22
- D.2 Released checkpoints and training 23
- D.3 Action prediction and observation cadence 23
- D.4 Partial-coverage Mimic-Video diagnostic 25

E Paired Outcomes and Robustness Breakdowns 25
- E.1 Paired outcomes and mechanism diagnostics 25
- E.2 Event-level robustness 26
- E.3 Robustness across source categories 27
- E.4 Task suites, severity metadata, and parameter sensitivity 29

F Camera Controls: Viewpoint and Execution History 30
- F.1 Design and paired comparisons 30
- F.2 Timing effects and source consistency 31
- F.3 Paired outcomes and trajectory diagnostics 32
- F.4 Milder camera shifts 33

G Query Cadence and Action Commitment 34
- G.1 Sweep design and complete results 34
- G.2 Native stale-action exposure 35

H Targeted Recovery from Sensor Corruption 35
- H.1 Study design and restoration module 35
- H.2 Paired effects 36
- H.3 Coverage and processing cost 36

I Discussion and Future Directions 36
- I.1 Interpreting progress on the benchmark 36
- I.2 Scope and limitations 37
- I.3 Learning and evaluating responses to change 37

A Additional Related Work

A.1 Robot benchmarks and dynamic manipulation

Existing environments measure transfer, language-conditioned manipulation, and compositional skills (Yu et al., 2020; James et al., 2020; Mees et al., 2022; Mu et al., 2021). Other platforms emphasize simulation scale and controlled intervention (Gu et al., 2023; Mendez et al., 2022; Ahmed et al., 2020), while household, bimanual, and demonstration resources broaden task and scene diversity (Nasiriany et al., 2024; Li et al., 2023; Chen et al., 2025a; Walke et al., 2023). LIBERO, LIBERO-Plus, and LIBERO-PRO provide controlled studies of transfer, OOD robustness, and memorization (Liu et al., 2023; Fei et al., 2025; Zhou et al., 2025b); LIBERO-Safety evaluates physical and semantic safety (Cui et al., 2026). Related stress tests expose visual, embodiment, distractor, and adversarial weaknesses (Pumacay et al., 2024; Wang et al., 2025b; Zhang et al., 2024; Li et al., 2024c; Zhou et al., 2025a; Liu et al., 2025; Fang et al., 2025; Garcia et al., 2025; Kube et al., 2025; Wang et al., 2025a). These robustness benchmarks normally set external conditions before the first policy action. DynamicVLA and DOMINO directly study moving-object manipulation and low-latency reactive control (Xie et al., 2026; Fang et al., 2026b). Their focus on continuous target motion complements LIBERO-MAX, which pairs no-event and event rollouts with identical executed prefixes to test abrupt nuisance, geometry, and path changes.

LIBERO-X (Wang et al., 2026) evaluates progressively combined spatial, object, and instruction perturbations through a hierarchy of difficulty levels, and provides diverse human demonstrations collected through teleoperation. This complements LIBERO-MAX, which introduces an event after an identical executed action prefix to compare task outcomes with and without a change during execution.

A.2 Object-centric generalization and changes during execution

ORION (Zhu et al., 2026) constructs a manipulation policy from a single human video using object graphs that represent task-relevant states and interaction relations. Its evaluation includes generalization to novel object instances and varied scenes. A complementary extension of LIBERO-MAX would test changes to a particular object during an ongoing task while holding its pre-event history fixed.

A.3 Vision-language-action policies and action commitment

The VLA lineage ranges from RT-1, RT-2, and PaLM-E (Brohan et al., 2022; 2023; Driess et al., 2023) to open and generalist policies such as OpenVLA, OpenVLA-OFT, π_0, $\pi_{0.5}$, FAST, and Octo (Kim et al., 2024; 2025; Black et al., 2025b;a; Pertsch et al., 2025; Octo Model Team et al., 2024). Recent variants study cognition, diffusion decoding, unified tokenization, geometry, compact policies, and language grounding (Li et al., 2024a; Wen et al., 2025; Bu et al., 2025; Zheng et al., 2026; Sun et al., 2025; Zhao et al., 2025; Shukor et al., 2025; Li et al., 2024b). Complementary work studies whether visual content remains semantically faithful through cross scale reindexing and mechanistic analysis of object hallucination pathways in VLMs (Dong et al., 2026; Liu et al., 2026a). Methods for progress-aware manipulation and fine-grained spatial VLM reasoning address complementary failure sources by tracking subtask state and improving visual grounding (Liu et al., 2026b; Shen et al., 2025; Ye et al., 2026b). These methods strengthen internal progress or perception, whereas LIBERO-MAX tests response to an external change after execution has begun.

A.4 World models and world-action policies

World models support visual foresight, model-predictive control, and latent planning (Ha & Schmidhuber, 2018; Finn & Levine, 2017; Chua et al., 2018; Hafner et al., 2020; Janner et al., 2019; Schrittwieser et al., 2019; Hansen et al., 2022; Hafner et al., 2023; Hansen et al., 2024). Video foundation models and predictive representations extend these ideas to large-scale spatiotemporal priors (Bruce et al., 2024; Assran et al., 2025; Agarwal et al., 2025; Kong et al., 2026). World-action and hybrid policies connect prediction to control through video generation, latent actions, joint training, memory, or causal interleaving (Hu et al., 2025; Ma et al., 2026; Chen et al., 2025b; Cen et al., 2025; Bi et al., 2025; Liao et al., 2025; Li et al., 2026a; Ye et al., 2026a; Sun et al., 2026b; Li et al., 2026b).

Table 5: **Outcome-independent selection balances all eight events.** Cells combine source category, event, and parameter variant. Every selected case receives one Base and one Dynamic rollout, yielding 16,000 rollouts per policy.

Source	Categories	Events	Variants	Cases/cell	Total cases
LIBERO-Plus	7	8	2	50	5,600
LIBERO-PRO	10	8	2	15	2,400
Combined	17	8	2	—	8,000

Our evaluation measures whether these policies still complete the task after an external event, rather than the accuracy of their predicted video or latent state. The fixed event boundary can also support future adaptive methods that use prediction error to decide when to cancel pending actions and query the policy again.

B Benchmark Construction and Intervention Examples

This section describes how source tasks become matched Base/Dynamic cases, how events are applied, and which checks determine eligibility. All construction decisions precede policy evaluation.

B.1 Design requirements

Four requirements define the comparison. **Temporal placement:** an event occurs after execution begins. **Pre-event identity:** Base and Dynamic share the task, initial state, instruction, policy seed, and executed action prefix. **Continued feasibility:** geometry screening checks whether interventions satisfy support, workspace, visibility, and collision constraints. **Denominator integrity:** cases remain in the score even if they never reach the trigger or receive a later query.

These requirements let us distinguish whether a policy can complete the Base task, reaches the event, and completes the task after the change. The primary score includes every assigned case; paired transitions and coverage diagnostics report what happens at each stage. The camera controls also compare encountering the same changed view at reset and during execution.

B.2 Source selection and balancing

A selection cell combines a source category, event type, and parameter variant. Each Plus cell contains 50 unique source tasks, and each PRO cell contains 15 task/initial-state configurations. The resulting 5,600/2,400 split gives every event 700 Plus-derived and 300 PRO-derived cases. Two fixed variants per event probe direction, magnitude, transform, or placement. Table 5 summarizes the allocation.

Plus-derived cases. The source catalog contains 10,030 variants across seven categories and five difficulty levels (Fei et al., 2025). Image-space events require a valid trigger entity. Relocation requires a movable target or receptacle on planar support. Distractor insertion requires five eligible objects with declared supports, while obstacle insertion requires an unused object sharing the target's support. Capacity-constrained matching fills 112 cells with 50 unique tasks each. The five difficulty levels contribute 1,075, 1,137, 1,135, 1,135, and 1,118 cases, respectively. Event direction, magnitude, object identity, and placement are fixed before evaluation.

PRO-derived cases. We retain semantic, object, position, task, visual noise and glare, camera viewpoint, object texture, view occlusion, object shape, and initial pose categories (Zhou et al., 2025b). Each contains 40 source variants, one for each task in the four LIBERO suites, with ten candidate initial states. Runtime object move is excluded because it already changes the world during execution. The environment category lacks complete scene and initialization assets. Each of the 160 retained cells receives 15 unique task/initial-state configurations, with selection favoring balanced task use. A source task can occur under several events, so configurations are the sampling unit rather than distinct task identities.

Table 6: **Eight events test distinct responses to a post-commitment change.** Parameters take one of two fixed variants. *Continue* means that task semantics remain unchanged, although the policy may still need perceptual correction; *replan* means that the target, destination, or path has changed.

Event	Family	Expected response	Frozen intervention
Target relocation	Geometry	Replan	Move the task target on its valid support after approach
Receptacle relocation	Geometry	Replan	Move the destination while keeping the task feasible
Camera shift	Observation	Continue or correct	Shift camera position, yaw, and field of view
Illumination switch	Appearance and clutter	Continue	Change scene lighting by a frozen scale
Sensor noise onset	Observation	Continue or correct	Add image corruption and a fixed occluded fraction
Visual theme switch	Appearance and clutter	Continue or correct	Apply one fixed color and channel transform
Distractor burst	Appearance and clutter	Continue	Insert five irrelevant objects on the active support
Obstacle insertion	Path constraint	Replan	Insert an object in the target-support approach region

We select cases and replace invalid configurations within each cell using only source metadata and simulator checks. Replacements occur before any policy outcomes are observed and preserve all category, event, and parameter quotas. Every policy uses the same final 8,000 cases. The primary micro average weights all cases equally; a secondary 17-category macro average gives equal weight to source categories of different sizes.

B.3 Events, triggers, and prefix replay

The eight events span changes to the target, destination, visual input, clutter, and approach path. Each is applied once and is not reversed by the evaluator before termination. Table 6 defines the intervention and intended response. Continuing can require perceptual correction even when the task goal remains unchanged.

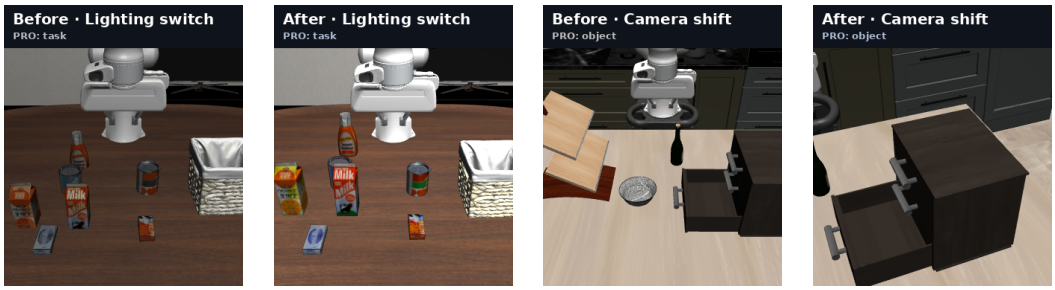
All events use the deterministic `pre_grasp_proximity` trigger, which fires when the end effector first enters the designated entity's neighborhood during approach. Base records policy queries and actions. Dynamic restores the same initial state, replays the executed prefix, and applies the event at the trigger. The evaluator verifies identical pre-event actions, one event per triggered case, and the intended change after onset.

For chunked policies, the event can arrive while actions predicted before onset remain queued. The number of these queued actions executed after the event is the *stale-action exposure*. Trigger coverage reports the fraction of assigned cases that reach the event; response coverage reports the fraction that receive a subsequent policy query. Both use all assigned cases, as does the primary success rate.

B.4 Validity and treatment of failures

All 5,600 Plus-derived and 2,400 PRO-derived cases pass automated simulator preflight. Checks require a finite, stable state, valid support and workspace placement, visibility, a nonzero observation change, and no new unrelated contact. Scene-construction or rendering failures are repaired before final scoring. All fourteen primary policy evaluations contain outcomes for the same 8,000 cases.

We score every case by its observed task outcome, including cases that never reach the trigger or receive a later query. Coverage reports how often the event occurs and the policy receives a new observation. Automated validity checks do not establish human-perceived naturalness or demonstrate successful completion after every intervention. The human feasibility audit remains incomplete.



Illumination switch: Base and Dynamic **Camera shift**: Base and Dynamic



Visual theme switch: Base and Dynamic **Sensor corruption**: Base and Dynamic

Figure 7: **Paired observation changes.** Each Base image is the exact prechange reference for the adjacent Dynamic image. The intervention changes only the named factor at the frozen trigger.



Receptacle relocation: Base and Dynamic **Distractor burst**: Base and Dynamic



Target relocation: Base and Dynamic **Obstacle insertion**: Base and Dynamic

Figure 8: **Paired geometry, clutter, and path changes.** Each Base image is the exact prechange reference for the adjacent Dynamic image. The task instruction remains fixed while one declared physical factor changes.

B.5 Qualitative intervention examples

Figs. 7 and 8 show matched Base and Dynamic images for all eight events. Observation events alter visual evidence without moving task objects. Geometry, clutter, and path events modify the physical state and are designed to preserve a feasible completion route.

Table 7: **LIBERO-MAX Lite provides a rapid 800-pair estimate of Max on fourteen policies.** Every value is success rate or Dynamic minus Base success in percentage points.

Model	Max Base	Lite Base	Max Dynamic	Lite Dynamic	Max Gap	Lite Gap
$\pi_{0.5}$	79.7	79.3	65.7	65.1	−13.9	−14.1
OpenVLA-OFT	64.3	64.3	43.2	45.0	−21.1	−19.3
X-VLA	62.6	62.0	37.7	39.4	−24.9	−22.6
Xiaomi-Robotics-0	70.3	70.6	52.0	53.5	−18.3	−17.1
MolmoAct2	80.3	79.5	66.9	68.5	−13.4	−11.0
SmolVLA	26.1	25.6	15.0	15.3	−11.0	−10.4
GR00T N1.7	69.3	69.3	50.0	51.4	−19.3	−17.9
DM0.5	79.8	79.4	62.1	63.6	−17.7	−15.8
VLA-JEPA	73.5	72.5	54.3	53.4	−19.2	−19.1
Cosmos-Policy	77.4	76.1	59.3	58.6	−18.2	−17.5
Fast-WAM	42.0	41.4	24.0	24.0	−18.0	−17.4
HiMem-WAM	73.0	72.5	57.7	58.0	−15.3	−14.5
Light-WAM	54.8	55.3	37.3	38.5	−17.5	−16.8
DiT4DiT	65.1	65.5	39.4	38.5	−25.7	−27.0

C LIBERO-MAX LITE

C.1 OUTCOME-INDEPENDENT SUBSET

LIBERO-MAX Lite uses a fixed subset of 800 released Max case IDs. A persistent pseudorandom generator with seed `20260830` selects 70 Plus derived and 30 PRO derived cases for each event using only the frozen manifest, event identity, and source label. Selection is independent of checkpoint identity, Base success, Dynamic success, trigger coverage, and query cadence. The resulting Lite track contains 800 matched pairs and 1,600 scored rollouts per checkpoint. This strict subset preserves Max's equal allocation across all eight events and 70:30 source composition. The action cadence analysis uses the same fixed pool.

C.2 AGREEMENT WITH THE FULL BENCHMARK

Table 7 compares Lite with Max for fourteen checkpoints whose per-case outcomes can be scored on the fixed subset. All Lite rates use the same 800-pair denominator, so a missing terminal result remains a failure under the benchmark scoring rule. The largest absolute deviation is 2.4 points across all reported Base rates, Dynamic rates, and paired gaps. Lite preserves 88 of 91 pairwise Dynamic orderings. The three differences involve VLA-JEPA versus Xiaomi-Robotics-0 and DiT4DiT versus X-VLA and Light-WAM; these pairs are separated by at most 2.3 points on Max, and DiT4DiT and Light-WAM tie on Lite.

Lite supports integration, debugging, ablation, and early comparison. Max remains the primary reporting track because its larger event cells support more detailed analysis with lower variance. DiT4DiT and DM0.5 each have recorded outcomes for all 800 fixed Lite case IDs; none are reconstructed from aggregate results.

D EVALUATION PROTOCOL AND POLICY DETAILS

D.1 METRICS AND STATISTICAL COMPARISONS

We evaluate all fourteen primary policies on the same 8,000 Base/Dynamic pairs. We report success rates over all assigned cases, paired change $\Delta = \mathrm{SR}_{\mathrm{Dynamic}} - \mathrm{SR}_{\mathrm{Base}}$ in percentage points, and event-family breakdowns. Conditional regression is the fraction of Base-successful cases that fail in Dynamic. We also report retained performance, $\mathrm{SR}_{\mathrm{Dynamic}}/\mathrm{SR}_{\mathrm{Base}}$, and trigger and response coverage. Cases that do not reach the trigger or a post-event query remain in the primary denominator.

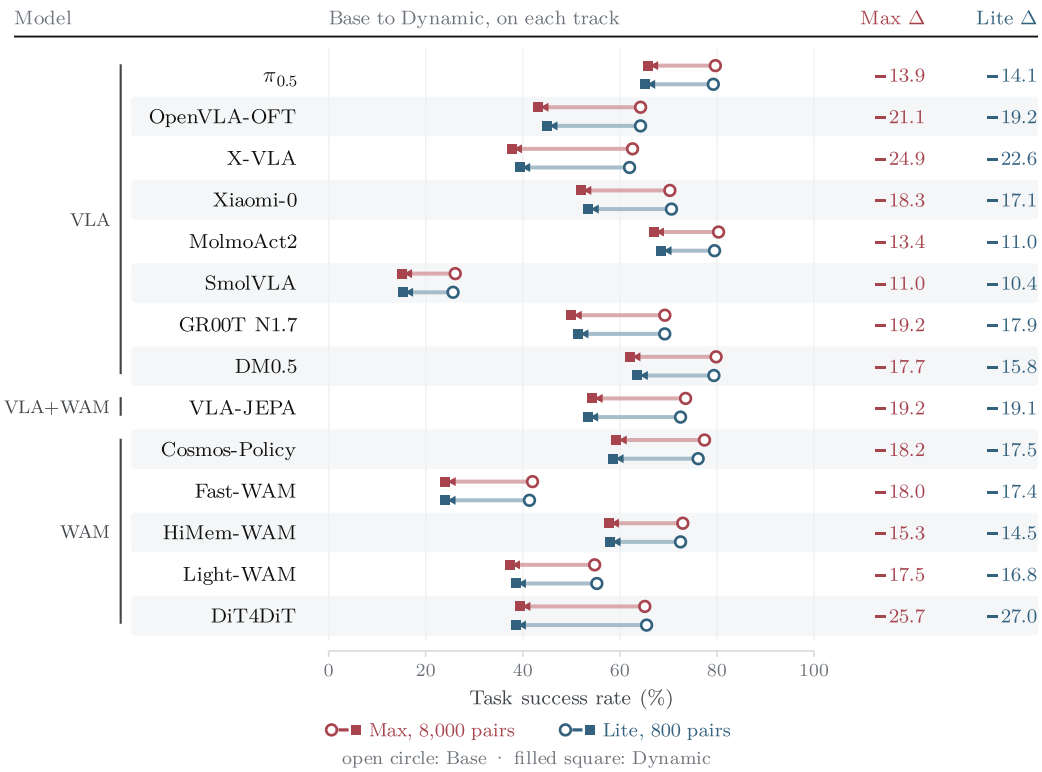


Figure 9: **The 800-pair Lite track closely reproduces Max across fourteen policies.** Each policy has an upper Max track over 8,000 pairs and a lower Lite track over 800 pairs. Open circles are Base, filled squares Dynamic, and arrows connect them. The two paired changes are printed at right.

Binomial proportions use 95% intervals, and paired changes use paired bootstrap intervals. Within each policy, the exact McNemar test compares event-associated regression $(1, 0)$ with change-associated success $(0, 1)$ under the null that these discordant outcomes are equally likely. Cross-policy Dynamic contrasts use common case IDs, with Holm correction across the resulting multiple comparisons. These comparisons describe released policies, whose scale, training data, and serving protocols differ.

We also break down results by event type, task suite, source benchmark, source category, severity, and parameter variant. These comparisons are descriptive unless matched on case ID. Plus-derived and PRO-derived cases are disjoint task instances, so their aggregate difference does not estimate the causal effect of adding PRO. The 17-category macro average gives categories equal weight and remains secondary to the result over all assigned cases. Uncertainty intervals quantify evaluation precision; they do not account for differences in checkpoint training.

D.2 Released checkpoints and training

The primary comparison includes eight VLAs, the predictive VLA-JEPA policy, and five WAMs. Table 8 lists each evaluated LIBERO checkpoint, its training signal, and its official source through the linked model name. Reported scales refer to the principal trainable model or the components named by the model authors; "not reported" means that a single comparable total is unavailable.

D.3 Action prediction and observation cadence

Throughout each primary evaluation, the policy receives new observations and predicts action chunks until the episode ends. The evaluator executes Q actions from each chunk before the next observation and query. Each query returns at most H actions, with $Q \leq H$; when $Q < H$, the

Table 8: **Released checkpoints cover three policy families.** Scale distinguishes reported totals from separately reported components. The final column summarizes architecture and training of the evaluated LIBERO checkpoint.

Model	Reported scale	Backbone and checkpoint training
Vision-language-action policies (VLA)		
$\pi_{0.5}$	2B VLM + 300M action expert	PaliGemma with a flow-matching action expert; heterogeneous robot, semantic, and web pretraining followed by LIBERO adaptation (Black et al., 2025a).
OpenVLA-OFT	7B VLA + lightweight heads	Prismatic OpenVLA with continuous regression, parallel decoding, proprioception, and wrist images; combined checkpoint LoRA-optimized on the four standard LIBERO suites (Kim et al., 2024; 2025).
X-VLA	0.9B	Cross-embodiment VLA with embodiment-specific soft prompts and a flow-matching decoder; released LIBERO checkpoint without MAX-specific adaptation (Zheng et al., 2026).
Xiaomi-Robotics-0	5B	Cross-embodiment and vision-language pretraining with asynchronous action-chunk deployment; released LIBERO checkpoint with its native query cadence (Cai et al., 2026).
MolmoAct2	Not reported	Embodied-reasoning VLM with a flow-matching action expert conditioned through per-layer key-value caches; released LIBERO policy with continuous action decoding (Fang et al., 2026a).
SmolVLA	0.45B variant	SmolVLM2 with flow-matching action generation and asynchronous robot-data training; released `HuggingFaceVLA/smolvla_libero` without MAX-specific adaptation (Shukor et al., 2025).
GR00T N1.7	3B	Cosmos-Reason2-2B with proprioception and a flow-matching action transformer; NVIDIA's official LIBERO post-trained checkpoint without MAX-specific adaptation (Bjorck et al., 2025).
DM0.5	Not reported	Dexmal's multimodal flow-matching VLA for open-world control; official `DM05-libero` checkpoint without MAX-specific adaptation.
Predictive VLA (VLA+WAM)		
VLA-JEPA	2B VLM + predictive encoders	Qwen3-VL-2B with a V-JEPA2 teacher and latent future-state alignment; human video and single-embodiment robot pretraining followed by end-to-end LIBERO adaptation (Sun et al., 2026a; Assran et al., 2025).
World-action policies (WAM)		
Cosmos-Policy	2B	Cosmos-Predict2-2B adapted without a separate policy backbone; single-stage post-training on target LIBERO demonstrations (Agarwal et al., 2025; Kim et al., 2026).
Fast-WAM	Approximately 6B	Wan2.2-TI2V-5B plus ActionDiT, jointly trained on video and actions from clean LIBERO demonstrations; inference omits future video generation (Yuan et al., 2026).
HiMem-WAM	Not reported	Low-level motion and high-level skill latents with boundary-triggered task memory; released LIBERO checkpoint with native memory-gated inference (Sun et al., 2026b).
Light-WAM	0.44B trainable	Compact video backbone with latent future-video supervision and state-fusion action decoding; released LIBERO checkpoint with its native query cadence (Li et al., 2026b).
DiT4DiT	2B video backbone + action DiT	Coupled video-dynamics and action diffusion transformers with dual flow-matching objectives; released LIBERO checkpoint without MAX-specific adaptation (Ma et al., 2026).

Table 9: **Primary evaluations retain each policy's serving protocol.** H is the maximum number of actions returned by one query; Q is the number executed before the next observation and query.

Family	Model	H	Q
VLA	$\pi_{0.5}$	50	5
VLA	OpenVLA-OFT	8	8
VLA	X-VLA	30	30
VLA	Xiaomi-Robotics-0	30	10
VLA	MolmoAct2	10	10
VLA	SmolVLA	50	10
VLA	GR00T N1.7	40	8
VLA	DM0.5	10	10
VLA+WAM	VLA-JEPA	7	7
WAM	Cosmos-Policy	16	16
WAM	Fast-WAM	32	16
WAM	HiMem-WAM	32	10
WAM	Light-WAM	32	10
WAM	DiT4DiT	8	8

policy receives a new observation before all predicted actions have been executed. Table 9 gives the settings used for the primary 8,000-pair comparison.

SmolVLA uses ten flow-matching integration steps, and DM0.5 uses ten diffusion integration steps. A separate 200-episode standard-LIBERO integration check for SmolVLA yields 85.0% success,

Table 10: **Mimic-Video loses success across all three supported suites.** Success rates are percentages, and the gap is Dynamic minus Base in percentage points. The supported total contains 5,972 pairs; the 2,028 LIBERO-10 pairs are unavailable. This diagnostic is separate from the primary 8,000-pair comparison.

Suite	Pairs	Base SR	Dynamic SR	Gap
LIBERO-Goal	1,656	54.2	31.4	−22.8
LIBERO-Object	2,448	54.5	34.2	−20.3
LIBERO-Spatial	1,868	61.8	29.2	−32.5
Supported total	**5,972**	**56.7**	**31.9**	−24.8
LIBERO-10	2,028	N/A	N/A	N/A

Table 11: **Paired outcomes separate post-change regression from Base competence.** Counts partition the same 8,000 pairs for each of the fourteen primary policies into preserved success, change-associated success, event-associated regression, and persistent failure.

Model	Preserved $(1,1)$	Gained $(0,1)$	Regressed $(1,0)$	Failed $(0,0)$	Total
Cosmos-Policy	4,594	148	1,601	1,657	8,000
$\pi_{0.5}$	4,998	261	1,376	1,365	8,000
OpenVLA-OFT	3,232	220	1,909	2,639	8,000
X-VLA	2,838	177	2,171	2,814	8,000
Xiaomi-Robotics-0	3,921	237	1,703	2,139	8,000
MolmoAct2	5,091	264	1,336	1,309	8,000
SmolVLA	916	287	1,170	5,627	8,000
GR00T N1.7	3,673	328	1,868	2,131	8,000
DM0.5	4,757	210	1,629	1,404	8,000
VLA-JEPA	4,121	224	1,762	1,893	8,000
Fast-WAM	1,688	234	1,672	4,406	8,000
HiMem-WAM	4,284	329	1,553	1,834	8,000
Light-WAM	2,723	259	1,661	3,357	8,000
DiT4DiT	3,006	147	2,205	2,642	8,000

compared with the 87.3% reported for its 0.45B variant. The primary $\pi_{0.5}$ setting is $H = 50$, $Q = 5$; the action-cadence sweep uses a different action-horizon setting, documented with that experiment.

D.4 Partial-coverage Mimic-Video diagnostic

Mimic-Video uses frozen Cosmos-Predict2 video features and a flow-matching inverse-dynamics decoder (Pai et al., 2025). Its public release provides Spatial, Object, and Goal checkpoints but no LIBERO-10 checkpoint. We evaluate all 5,972 supported pairs with the released $H = 16$, $Q = 5$ protocol. The unsupported 2,028 LIBERO-10 pairs are unavailable, so Mimic-Video is excluded from primary tables, figures, rankings, and cross-model statistics.

E Paired Outcomes and Robustness Breakdowns

We first distinguish cases that fail after the change from those that already fail in Base, then compare events, sources, task suites, and parameter variants. Unless stated otherwise, these results use all fourteen primary policies and the same 8,000 pairs.

E.1 Paired outcomes and mechanism diagnostics

Table 11 separates cases that fail after the event from those that already fail without it. It assigns each pair to one of four outcomes: preserved success, change-associated success, event-associated regression, or persistent failure.

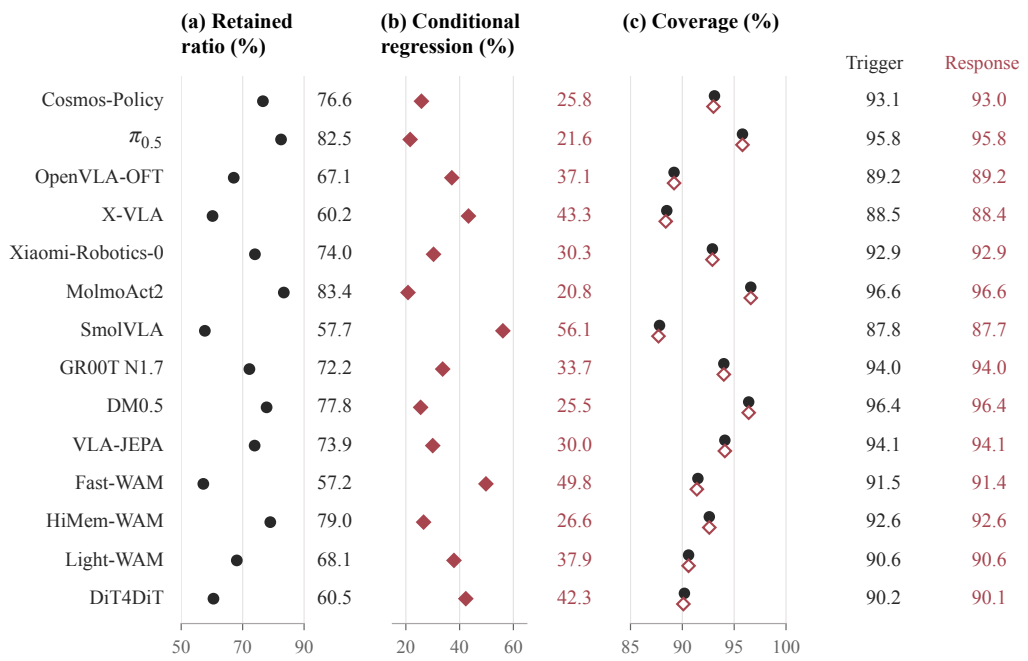


Figure 10: **Most cases reach the event and a later policy query, but many Base successes are lost.** For each of fourteen policies, retained ratio is Dynamic success rate divided by Base success rate, including change-associated successes; conditional regression is the fraction of Base successes that become Dynamic failures. Trigger and response coverage each use all 8,000 evaluated cases per policy; response requires a policy query after the event. Upper filled circles show trigger coverage and lower hollow diamonds show response coverage. All values are percentages, and the three panels use separate horizontal scales.

MolmoAct2 retains the most successes, DiT4DiT has the most event-associated regressions, and SmolVLA has the most persistent failures. Fig. 10 reports these results relative to Base performance, alongside trigger and response coverage. Coverage ranges from 87.7% for SmolVLA to 96.6% for MolmoAct2. Cases that fail before the event or before a later query still contribute to the primary success rate.

The online event breaks 20.8% to 56.1% of Base successes across the fourteen checkpoints. SmolVLA has the smallest overall gap, at 11.0 points, but loses the largest fraction of Base successes, at 56.1%; only 26.1% of its Base episodes succeed. MolmoAct2 at 20.8%, $\pi_{0.5}$ at 21.6%, DM0.5 at 25.5%, Cosmos-Policy at 25.8%, and HiMem-WAM at 26.6% have the lowest conditional regression rates.

E.2 Event-level robustness

Policies lose less success under some nuisance changes than under changes to perception, the task target, or the path (Fig. 5; full values in Table 12). Illumination and visual-theme changes cause comparatively small losses, consistent with stability under nuisance appearance changes. Target or receptacle relocation causes the largest loss for eleven policies. Fast-WAM instead falls from 43.4% Base success to 0.3% under sensor-noise onset, a 43.1-point drop, while X-VLA is most affected by camera shift. Camera and geometry changes remain consistently harmful, even though they call for different responses: perceptual correction and physical replanning. Final outcomes show how sensitivity varies by event; they do not reveal whether a policy detected the event internally or attempted the intended response.

Policies tend to struggle with the same events. Across the eight event losses, the mean pairwise Spearman correlation is 0.80 over the fourteen policies, although individual pairs range from 0.29 to 1.00. Geometry losses range from 21.2 to 52.4 points and observation losses range from 13.3

Table 12: **Event sensitivity varies across policies.** Entries show Dynamic minus Base success in percentage points for every policy and event. Every row has 1,000 pairs per event.

Model	Camera	Distractor	Light	Obstacle	Receptacle	Sensor	Target	Theme
Cosmos-Policy	−23.9	−15.7	−0.3	−4.6	−24.1	−26.4	−47.5	−2.8
$\pi_{0.5}$	−23.9	−10.9	+0.2	−4.2	−13.8	−22.3	−34.6	−2.0
OpenVLA-OFT	−35.6	−26.3	−1.2	−6.8	−25.1	−28.3	−42.1	−3.5
X-VLA	−55.9	−3.5	+0.7	−5.1	−52.4	−30.0	−52.3	−0.9
Xiaomi-Robotics-0	−30.6	−15.4	−0.1	−3.9	−31.3	−19.8	−44.0	−1.5
MolmoAct2	−17.9	−13.9	+0.6	−1.8	−21.1	−20.6	−32.0	−0.5
SmolVLA	−14.7	−3.9	−2.3	−2.7	−18.6	−18.5	−23.9	−3.7
GR00T N1.7	−31.8	−6.1	+0.3	−8.3	−37.5	−17.3	−52.5	−0.8
DM0.5	−33.6	−12.3	−0.9	−5.7	−22.7	−24.4	−39.9	−2.4
VLA-JEPA	−31.1	−15.6	−2.9	−4.0	−26.6	−29.4	−40.7	−3.5
Fast-WAM	−28.1	−3.7	−3.6	−6.1	−24.5	−43.1	−30.4	−4.3
HiMem-WAM	+0.9	−13.1	+0.1	−3.3	−35.8	−27.5	−43.7	0.0
Light-WAM	−3.0	−8.6	−0.9	−2.0	−39.1	−40.4	−40.7	−5.5
DiT4DiT	−31.5	−16.6	−7.6	−4.5	−18.6	−67.9	−34.7	−24.4

to 49.7 points. Both are generally larger than appearance-and-clutter losses of 1.2 to 16.2 points and path-constraint losses of 1.8 to 8.3 points. Sensor noise causes the largest losses for Fast-WAM and DiT4DiT, camera shift for X-VLA; HiMem-WAM remains unusually robust to camera changes. GR00T N1.7 and DM0.5 are particularly sensitive to geometry changes: target relocation reduces their success by 52.5 and 39.9 points, respectively, while appearance-and-clutter losses remain 2.2 and 5.2 points. These differences show why event-specific results are more informative than a family-wide ranking.

Table 12 reports paired losses for all eight events. Each event receives equal weight, so comparing rows requires no correction for event frequency.

E.3 Robustness across source categories

Source subsets and low Base success. All fourteen evaluated checkpoints have lower Dynamic success and a smaller paired drop on PRO derived cases than on Plus derived cases (Table 13). The two subsets contain different task instances, and lower PRO derived Base success leaves less room to fall. Their aggregate difference therefore does not establish that PRO increases sensitivity to online events.

Source-category patterns. Fig. 11 shows paired changes across all 17 source categories. Every policy loses success across many Plus derived categories; several PRO derived categories already have low Base success, which limits how far success can fall. Policies also differ in which categories cause the largest losses: X-VLA loses 37.2 points on Plus background texture and 47.9 on PRO noise and glare, whereas MolmoAct2 has smaller aggregate losses but remains sensitive to sensor noise and object shape.

A small or positive gap can coexist with frequent failures when Base success is already low. For example, SmolVLA scores only 0.4% on the PRO position Base cases, and Fast-WAM scores 0% on visual noise and glare in both trajectories. Figs. 12 and 13 therefore show both Base and Dynamic success alongside the heatmap of paired changes. The average loss remains negative when all 17 source categories receive equal weight: macro deltas range from −10.7 to −23.4 across all fourteen checkpoints.

Figs. 12 and 13 compare Base and Dynamic success directly for every model and source category. The paired bars show how much success falls and how often the policy succeeds in Base. A small gap supports a claim of robustness only when the policy can already solve the Base tasks.

Across 98 Plus source comparisons, most gaps are negative, with a small number of checkpoint-specific exceptions. Depending on the checkpoint, the largest Plus loss occurs in background texture, light condition, sensor noise, or language. PRO results vary more across categories. Several have Base success below 20%, so small positive gaps do not establish a successful post-change response. Categories with high Base success show large losses on object shape, object texture, semantics, and visual noise and glare.

Table 13: **Success falls after the change on both Plus and PRO cases.** Plus contains 7 categories and 5,600 cases; PRO contains 10 categories and 2,400 cases. Base and Dynamic success rates use all assigned cases, showing where low Base success limits the size of the loss. The final columns identify the source categories with the lowest Base success and the largest loss.

Model	Source	Base	Dynamic	Gap	Lowest Base	Largest loss
Cosmos-Policy	Plus	83.0	63.2	−19.8	Robot initial state	Background texture
	PRO	64.5	50.2	−14.4	Position	Noise and glare
$\pi_{0.5}$	Plus	86.1	70.3	−15.8	Camera viewpoint	Light condition
	PRO	64.8	55.1	−9.7	Initial pose	Noise and glare
OpenVLA-OFT	Plus	68.3	45.2	−23.0	Robot initial state	Light condition
	PRO	55.0	38.3	−16.7	Initial pose	Noise and glare
X-VLA	Plus	67.9	41.2	−26.6	Camera viewpoint	Background texture
	PRO	50.4	29.5	−20.9	Camera angle	Noise and glare
Xiaomi-Robotics-0	Plus	74.5	54.4	−20.1	Camera viewpoint	Sensor noise
	PRO	60.5	46.4	−14.1	Position	Noise and glare
MolmoAct2	Plus	86.6	71.8	−14.9	Robot initial state	Sensor noise
	PRO	65.7	55.7	−10.0	Camera angle	Object shape
SmolVLA	Plus	26.8	15.4	−11.4	Sensor noise	Light condition
	PRO	24.4	14.3	−10.1	Position	Object shape
GR00T N1.7	Plus	75.3	54.1	−21.2	Robot initial state	Light condition
	PRO	55.2	40.5	−14.7	Initial pose	Semantic
DM0.5	Plus	86.4	66.9	−19.5	Camera viewpoint	Light condition
	PRO	64.6	50.9	−13.7	Initial pose	Noise and glare
VLA-JEPA	Plus	79.0	58.3	−20.8	Sensor noise	Background texture
	PRO	60.8	45.1	−15.6	Initial pose	Noise and glare
Fast-WAM	Plus	42.8	24.4	−18.4	Camera viewpoint	Light condition
	PRO	40.2	23.1	−17.0	Noise and glare	Object shape
HiMem-WAM	Plus	76.8	60.5	−16.3	Robot initial state	Background texture
	PRO	64.0	51.1	−12.9	Initial pose	Noise and glare
Light-WAM	Plus	56.2	38.3	−17.9	Robot initial state	Light condition
	PRO	51.5	34.8	−16.6	Position	Object shape
DiT4DiT	Plus	70.7	42.5	−28.2	Background texture	Light condition
	PRO	52.2	32.2	−20.0	Initial pose	Object texture

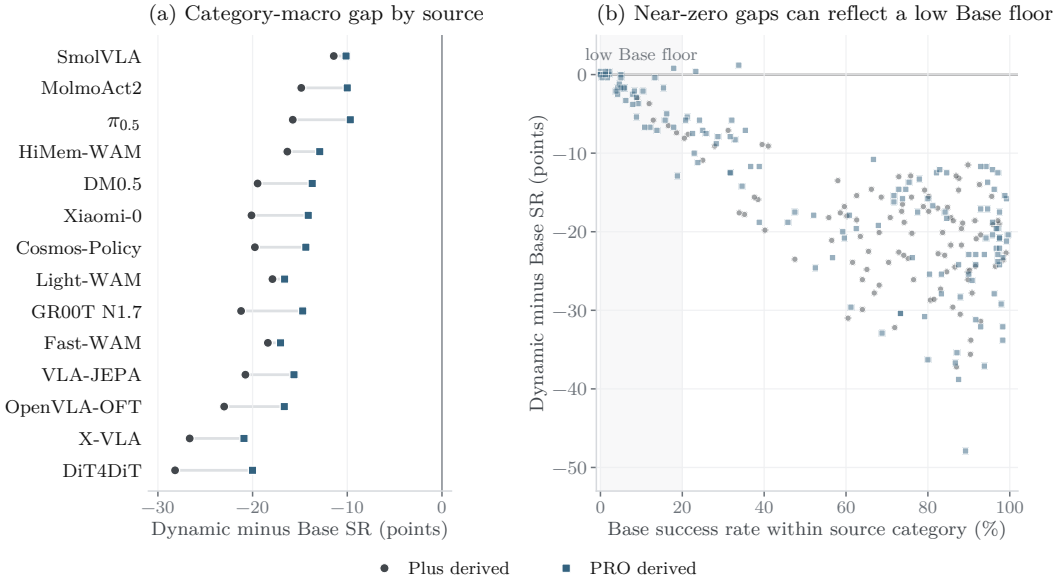


Figure 11: **Paired change across all 17 source categories.** (a) Category-macro Dynamic minus Base success for the seven Plus categories (circles) and ten PRO categories (squares). (b) Each model-category gap plotted against its Base success: near-zero gaps cluster where Base success is already below 20%, leaving little room for success to fall.



Figure 12: **Base and Dynamic success across the seven Plus derived source categories.** Most of the 98 model-category comparisons show losses; some categories already have low Base success. Every model-category cell contains 800 pairs.

E.4 Task suites, severity metadata, and parameter sensitivity

Table 14 tests whether the main gap is concentrated in one task suite or one severity label. It also records the largest difference between the two frozen parameter variants for each checkpoint.

Event parameters and task suites. Success can differ substantially between the two fixed parameter variants of the same event. Across fourteen checkpoints, the largest within-event Dynamic SR split ranges from 12.2 to 36.4 points. A benchmark's choice of event parameters can therefore materially affect measured robustness. LIBERO-MAX balances both variants and identifies them in the results as parameter variants, not independent model seeds.

The largest losses also occur in different task suites for different policies. LIBERO-10 produces the largest drop for several checkpoints, including GR00T N1.7, while LIBERO-Spatial is largest for

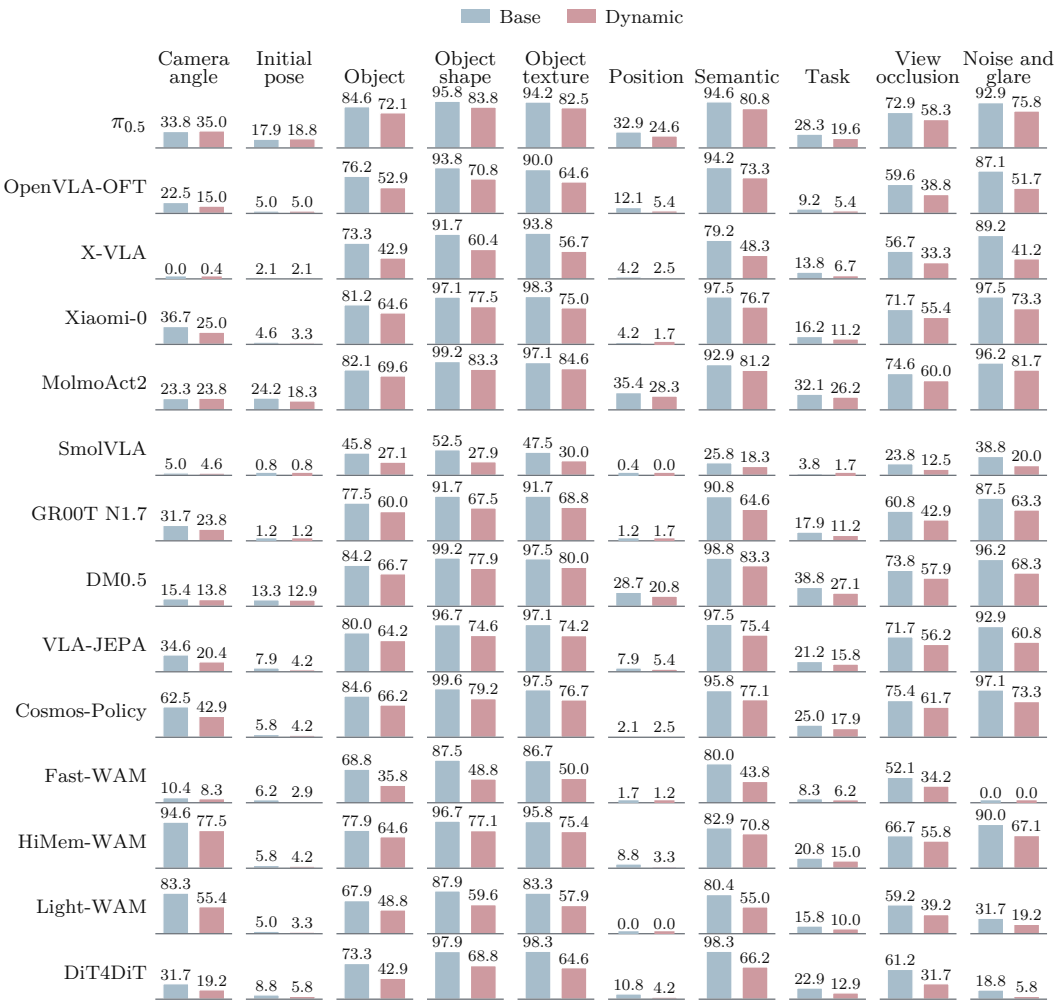


Figure 13: **Base and Dynamic success across the ten PRO derived source categories.** Each category contains 240 matched pairs. Low Base success explains the small or positive differences for several pose, position, and task settings.

Fast-WAM, HiMem-WAM, Light-WAM, and SmolVLA; DiT4DiT is most affected on LIBERO-Goal. The 17-category macro deltas remain negative for all fourteen evaluated checkpoints, so the overall loss persists when each source category receives equal weight. Individual PRO categories with Base success near zero still require care: reporting Base and Dynamic results together makes these floor effects visible.

F Camera Controls: Viewpoint and Execution History

F.1 Design and paired comparisons

A new camera view may be difficult even when the policy sees it from the start. We compare that setting with introducing the same view during an ongoing task, asking how viewpoint difficulty and the preceding trajectory affect task completion. X-VLA, $\pi_{0.5}$, and HiMem-WAM share all 1,000 camera cases from the frozen Max benchmark: 700 Plus-derived and 300 PRO-derived cases, selected without policy outcomes. Base, Reset, and Mid-task give 9,000 complete rollouts across the three models. This diagnostic preserves the original 8,000-pair benchmark and its primary results.

Three conditions. Base retains each case's original source configuration without an additional camera event. Reset restores that source configuration and initial state, then applies the complete

Table 14: **Losses persist across task suites, with sensitivity to event parameters.** All gaps are Dynamic minus Base success in percentage points. Category macro weights all 17 source categories equally. Medium and high contain 1,000 and 7,000 cases with different event compositions. Split is the largest within-event Dynamic success difference between two variant cells.

Task-suite and category gaps

Model	Macro	LIBERO-10	Goal	Object	Spatial
Cosmos-Policy	−16.6	−22.9	−19.8	−12.9	−18.5
$\pi_{0.5}$	−12.2	−21.2	−14.7	−10.4	−10.1
OpenVLA-OFT	−19.3	−22.3	−21.4	−21.8	−18.6
X-VLA	−23.3	−30.1	−24.6	−17.2	−29.7
Xiaomi-Robotics-0	−16.6	−23.5	−17.0	−13.8	−19.8
MolmoAct2	−12.0	−16.0	−15.0	−9.6	−14.1
SmolVLA	−10.7	−10.9	−5.4	−12.5	−14.1
GR00T N1.7	−17.4	−25.6	−16.8	−14.5	−20.8
DM0.5	−16.1	−25.1	−17.1	−13.2	−16.2
VLA-JEPA	−17.7	−24.1	−18.7	−14.1	−21.3
Fast-WAM	−17.6	−16.0	−17.0	−16.6	−22.8
HiMem-WAM	−14.3	−17.1	−13.5	−11.8	−19.5
Light-WAM	−17.2	−12.1	−17.0	−16.4	−25.3
DiT4DiT	−23.4	−26.3	−30.7	−18.3	−30.5

Severity groups and parameter sensitivity

Model	Medium	High	Largest-split event	Split (pp)
Cosmos-Policy	−20.6	−17.8	Receptacle	31.0
$\pi_{0.5}$	−13.5	−14.0	Camera	35.6
OpenVLA-OFT	−20.0	−21.3	Target	27.6
X-VLA	−41.1	−22.6	Receptacle	28.4
Xiaomi-Robotics-0	−20.2	−18.1	Target	35.6
MolmoAct2	−14.4	−13.3	Sensor	31.6
SmolVLA	−16.7	−10.2	Receptacle	12.2
GR00T N1.7	−32.2	−17.4	Receptacle	35.6
DM0.5	−16.1	−18.0	Target	34.8
VLA-JEPA	−18.2	−19.4	Target	34.6
Fast-WAM	−17.4	−18.1	Receptacle	20.6
HiMem-WAM	−23.1	−14.2	Receptacle	36.4
Light-WAM	−31.2	−15.6	Receptacle	23.0
DiT4DiT	−16.9	−27.0	Target	25.8

camera transform before the first policy input; the changed camera remains active until the episode ends. Mid-task uses the original Dynamic protocol: replay the matched Base action prefix up to the original proximity trigger, apply the same transform, and continue with the native action queue. Reset generates its own actions from the changed observations. Within each model, the conditions share the case, initial state, policy seed 195, source runtime, step budget, and success criterion.

X-VLA retains $H = Q = 30$, $\pi_{0.5}$ retains $H = 10$, $Q = 5$, and HiMem-WAM retains $H = 32$, $Q = 10$ with ten inference steps. Checkpoints and action interfaces are fixed across conditions. All 3,000 model–case triples pass the initial-state, executed-prefix, and camera checks. Executed Base/Mid prefixes match exactly before the event, and Reset's persistent camera position, orientation, and field of view match Mid-task in every triggered case. Trigger and subsequent-query counts are 916 and 915 for X-VLA, 951 and 951 for $\pi_{0.5}$, and 941 and 941 for HiMem-WAM. Every model retains all 1,000 cases in its primary denominator, including unreached events.

F.2 Timing effects and source consistency

The changed view reduces success even when present from the first input (Table 3). X-VLA achieves 68.2% in Base, 1.3% in Reset, and 12.4% in Mid-task; $\pi_{0.5}$ achieves 79.5%, 51.3%, and 55.3%, respectively. HiMem-WAM achieves 72.5%, 68.8%, and 72.6%. Mid-task exceeds Reset by 11.1, 4.0, and 3.8 points (paired 95% intervals in Fig. 14). All three policies therefore succeed more often when they first make progress in the original view, although X-VLA and $\pi_{0.5}$ still struggle with the changed view.

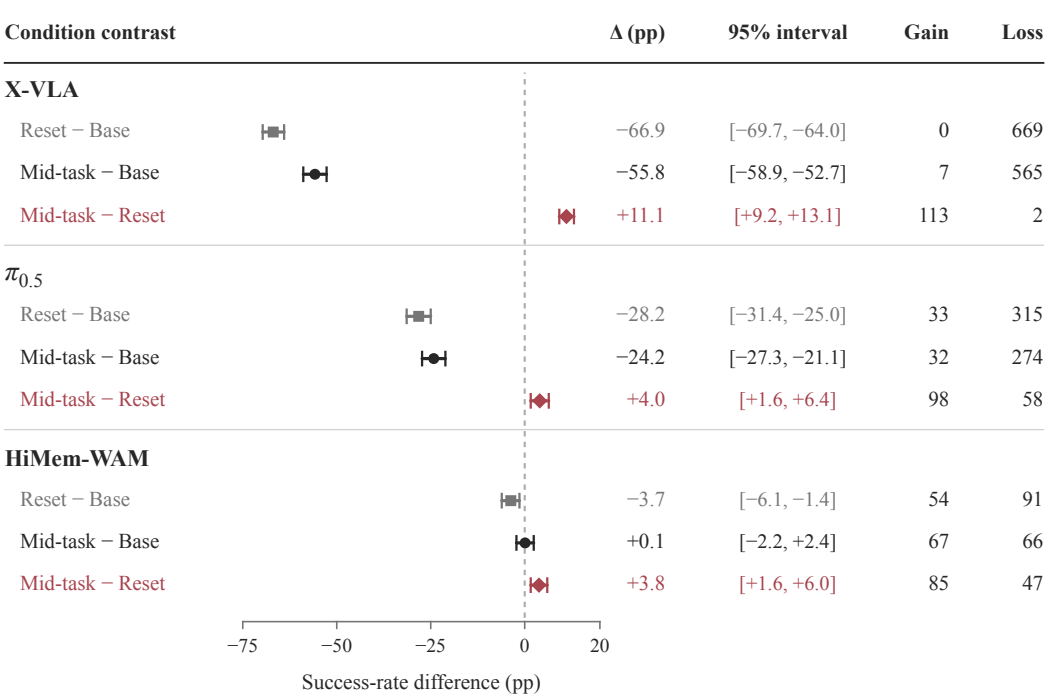


Figure 14: **Mid-task success exceeds Reset for all three policies.** Points show paired success-rate differences, with pointwise 95% empirical bootstrap intervals, over the same 1,000 camera cases per model; resampling retains the 700 Plus / 300 PRO composition. Gray squares denote Reset minus Base, black circles Mid-task minus Base, and red diamonds Mid-task minus Reset. Gain and Loss count cases where the first condition succeeds and the second fails, or vice versa.

For each contrast, we average within-case differences in binary success. Resampling retains the 700/300 source composition and all three paired conditions. We compute the 2.5th and 97.5th percentiles of the empirical bootstrap distribution by discrete convolution.

Source and task sensitivity. Mid-task success exceeds Reset success on both sources: by $+11.6/+10.0$ points for X-VLA, $+4.1/+3.7$ for $\pi_{0.5}$, and $+4.7/+1.7$ for HiMem-WAM on Plus/PRO, respectively. The PRO-only intervals for $\pi_{0.5}$ and HiMem-WAM include zero. Table 15 reports all source counts.

The 1,000 cases cover 995 source/task/initial-configuration groups and 40 reference tasks. We also resample the 40 reference tasks jointly across sources, include all cases belonging to each sampled task, and combine source-wise pooled success rates with fixed 0.7/0.3 weights. Across 50,000 replicates, the paired Mid-task minus Reset intervals are $[6.4, 17.0]$, $[0.8, 7.2]$, and $[0.8, 7.2]$ points for X-VLA, $\pi_{0.5}$, and HiMem-WAM. Resampling the 995 configuration groups or 80 source-specific reference-task groups independently within source supports the same Mid-task advantage. These checks keep the original case-weighted point estimates and test whether the interval estimates depend on how cases are grouped. They cover the observed cohort; additional policy seeds and training runs remain outside this analysis.

F.3 Paired outcomes and trajectory diagnostics

HiMem-WAM's near-zero Base-to-Mid difference combines 66 cases that succeed in Base but fail in Mid-task with 67 that fail in Base but succeed in Mid-task. Among cases where both Base and Reset succeed, Mid-task fails in 2/13 X-VLA cases, 48/480 $\pi_{0.5}$ cases, and 34/634 HiMem-WAM cases. Thus, each policy has individual Mid-task failures despite its higher aggregate success than Reset. These groups are selected from each policy's observed successes; they do not independently establish that the policy can reliably handle the new view.

Table 15: **Camera-control outcomes vary across source partitions.** Cells are success counts, with the source-specific denominator in the third column. The same cases appear in all three conditions.

Model	Source	Cases	Base	Reset	Mid-task
X-VLA	Plus	700	524	12	93
	PRO	300	158	1	31
$\pi_{0.5}$	Plus	700	606	382	411
	PRO	300	189	131	142
HiMem-WAM	Plus	700	531	506	539
	PRO	300	194	182	187

The two comparisons answer different questions. Base versus Mid-task measures the effect of adding a camera event after the same executed history. Reset versus Mid-task compares seeing the changed view from the start with seeing it after making progress in the original view. This second comparison also changes how long the policy sees the new view, which states it visits, and how many steps remain. The results measure whether the policy completes the task under these conditions; identifying internal detection or deliberate replanning would require additional instrumentation.

Approach and post-event behavior. We use the trigger's 18 cm Euclidean end-effector distance to the designated entity as an approach diagnostic. Reconstructing the first threshold crossing from all 9,000 trajectories agrees with the logged crossing step in every case. Among Reset failures, X-VLA has 516 cases that never enter this neighborhood and 471 that enter before failing; the corresponding counts are 103 and 384 for $\pi_{0.5}$, and 69 and 243 for HiMem-WAM. This threshold describes approach rather than a verified grasp or necessary task stage: one successful $\pi_{0.5}$ Reset rollout never crosses it.

Among cases where Base and Mid-task succeed but Reset fails, the Reset trajectory never enters the neighborhood in 74/106 X-VLA cases, 21/89 $\pi_{0.5}$ cases, and 2/59 HiMem-WAM cases. This pattern suggests that progress made before the camera change helps the policy approach the object, particularly for X-VLA. Conversely, of the 84 Base/Reset-success cases that fail in Mid-task, all reach a new model query and 13 execute zero old queued actions after the event (eight $\pi_{0.5}$ and five HiMem-WAM cases). Old-action exposure therefore cannot explain every such failure.

F.4 Milder camera shifts

To examine whether X-VLA's near-zero Reset success at full strength drives the timing comparison, we evaluate $\alpha \in \{0.25, 0.5, 1\}$ on the same 1,000 camera cases (700 Plus and 300 PRO). We multiply each case's camera-position, yaw, and field-of-view deltas by α, preserving the shift direction and original trigger. These runs retain the checkpoint, initial states, seed, source runtimes, and $H = Q = 30$. The design has seven outcomes per case—one shared Base and Reset/Mid-task at each strength—for 7,000 trajectories. Every Mid-task branch retains the exact Base action prefix and recorded pre-event end-effector and trigger-entity positions. The same 916 cases reach the event at every strength, and 915 reach a subsequent model query; all 1,000 remain in the primary denominator.

As strength decreases from 100% to 50% to 25%, Reset success rises from 1.3% to 7.0% to 36.0%, and Mid-task success from 12.4% to 29.1% to 51.7% (Fig. 15); shared Base success is 68.2%. At 25%, 50%, and 100% strength, the paired Mid-task minus Reset differences are $+15.7$ points ($[12.8, 18.6]$), $+22.1$ ($[19.3, 24.9]$), and $+11.1$ ($[9.2, 13.1]$), respectively. Mid-task therefore still outperforms Reset at quarter strength, where Reset success is well above its full-strength floor. Both changed-view conditions still fall below Base: Reset by 32.2 points ($[-35.4, -29.0]$) and Mid-task by 16.5 ($[-19.1, -13.9]$). Even this milder camera shift makes the task harder.

The trajectories show whether Reset failures occur before or after approach. At 25% strength, 119 Reset rollouts fail without entering the 18 cm neighborhood, 521 enter but fail, and 360 succeed; at 50%, these counts are 305, 625, and 70. Among cases that succeed in Base and Mid-task but fail in Reset, the Reset trajectory never enters the neighborhood in 18/189 cases at 25%, 55/221 at 50%,

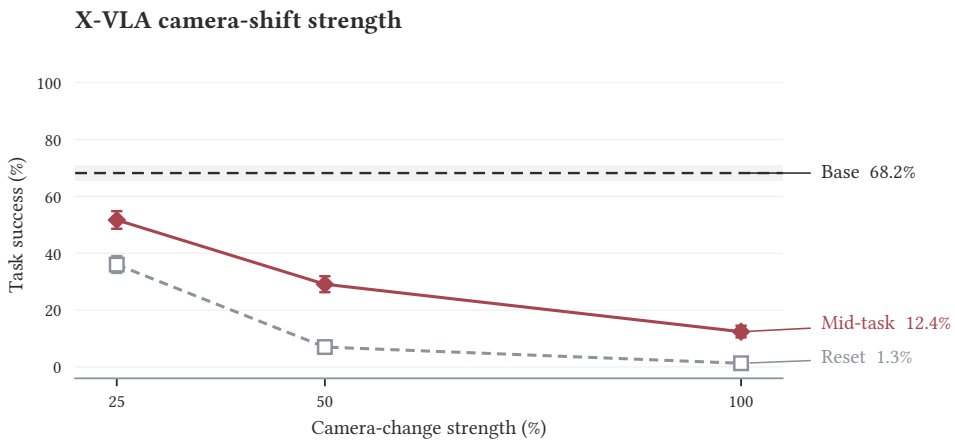


Figure 15: **Milder camera changes improve X-VLA in both timing conditions.** Reset and Mid-task use 25%, 50%, or 100% of each case's original camera transform. The dashed Base line and pale band show the shared no-event success rate and its interval. Error bars are pointwise 95% source-stratified empirical bootstrap intervals over the same 1,000 cases. Paired Mid-task minus Reset differences (95% intervals) are $+15.7$ ($[12.8, 18.6]$), $+22.1$ ($[19.3, 24.9]$), and $+11.1$ ($[9.2, 13.1]$) percentage points at 25%, 50%, and 100% strength, respectively.

and 74/106 at full strength. At the milder settings, most of these Reset failures occur after reaching the approach threshold.

Bootstrap resampling keeps all seven outcomes paired within each case and retains the 700/300 source composition. The Mid-task minus Reset gap is largest at 50% strength: it exceeds the 25% gap by 6.4 points ($[2.8, 10.0]$) and the full-strength gap by 11.0 ($[8.4, 13.7]$). The 25% gap also exceeds the full-strength gap by 4.6 points ($[1.4, 7.8]$). These are pointwise paired intervals, rather than simultaneous intervals across strengths and contrasts. The reference-task resampling described above also retains positive Mid-task minus Reset intervals at 25% and 50% strength: $[10.9, 20.8]$ and $[15.8, 28.5]$ points, respectively.

G Query Cadence and Action Commitment

G.1 Sweep design and complete results

The sweep evaluates 18 policy/cadence settings on the fixed 800-pair Lite pool, totaling 14,400 Base/Dynamic pairs. Each setting includes 100 cases per event with a 70:30 Plus/PRO split. Case IDs, event parameters, and policy seeds are fixed across query intervals. Within each setting, Base and Dynamic share an identical executed pre-event prefix. Changing the query interval can change that prefix across settings.

The returned action horizon H and executed prefix Q satisfy $Q \leq H$. The $\pi_{0.5}$ sweep uses $H = 10$, distinct from the $H = 50$ primary protocol, and tests $Q \in \{2, 4, 8\}$. The longer intervals are unavailable under that sweep interface. Fast-WAM fixes $H = 32$ and tests $Q \in \{2, 4, 8, 12, 16\}$. X-VLA and GR00T N1.7 test the same five intervals while decoding $H = Q$ actions. No checkpoint is retrained. For the latter two policies, the decoded horizon and executed prefix vary together, so the sweep compares practical serving settings rather than isolating Q alone.

X-VLA rises from 21.9% to 60.4% Base success and from 10.8% to 39.2% Dynamic success as Q increases from 2 to 16. Over its valid range, $\pi_{0.5}$ also improves, with a nearly constant 12.8–13.1-point paired loss. GR00T N1.7 peaks at $Q = 8$ in Base and $Q = 12$ in Dynamic; Fast-WAM peaks at $Q = 8$ in both conditions. Across all settings, Dynamic remains 11.1–23.0 points below Base. These policy-specific optima agree with evidence that fixed chunk optima vary across tasks (Liang et al., 2026).

Table 16: **Complete results for the fixed 800-pair action cadence sweep.** Base and Dynamic are success rates in percent. Gap is Dynamic minus Base in percentage points. H is the decoded action horizon and Q is the number of actions executed before the next observation. Only valid protocols with $Q \leq H$ are listed; $Q \in \{12, 16\}$ is therefore undefined for $\pi_{0.5}$ at $H = 10$.

Model	Q	H	Base	Dynamic	Gap
X-VLA	2	2	21.9	10.8	−11.1
X-VLA	4	4	40.0	19.0	−21.0
X-VLA	8	8	52.4	29.5	−22.9
X-VLA	12	12	57.6	34.6	−23.0
X-VLA	16	16	60.4	39.2	−21.1
$\pi_{0.5}$	2	10	74.1	61.1	−13.0
$\pi_{0.5}$	4	10	77.2	64.1	−13.1
$\pi_{0.5}$	8	10	80.0	67.2	−12.8
GR00T N1.7	2	2	61.2	41.9	−19.4
GR00T N1.7	4	4	66.2	48.2	−18.0
GR00T N1.7	8	8	69.2	51.4	−17.9
GR00T N1.7	12	12	65.5	53.2	−12.2
GR00T N1.7	16	16	66.1	50.8	−15.4
Fast-WAM	2	32	38.2	18.2	−20.0
Fast-WAM	4	32	47.4	25.0	−22.4
Fast-WAM	8	32	47.6	28.0	−19.6
Fast-WAM	12	32	46.8	26.4	−20.4
Fast-WAM	16	32	43.9	25.9	−18.0

G.2 Native stale-action exposure

For six policies with step-level traces, median stale-action exposure is two actions for $\pi_{0.5}$, four for OpenVLA-OFT and GR00T N1.7, three for VLA-JEPA, and seven for Cosmos-Policy and Fast-WAM. The corresponding means are 2.02, 3.59, 3.60, 2.96, 7.47, and 7.24. The policy can therefore act on stale observations before receiving fresh feedback, even though every response-evaluable case eventually produces a post-event action chunk.

Exposure measures how long feedback is delayed under the native serving protocol. It does not measure internal change detection or physical safety. The sweep shows that query timing changes task success, but changing cadence alone leaves a substantial Dynamic deficit.

H Targeted Recovery from Sensor Corruption

H.1 Study design and restoration module

We test whether a simple observation-processing module can recover part of the measured performance gap. The study uses X-VLA and 300 fixed `sensor_noise_onset` cases: 210 from Plus and 90 from PRO, stratified by source category and corruption setting. Each setting contains 150 cases: noise standard deviation 24 with 8% image occlusion, or 36 with 16% occlusion. We compare the native policy, always-on restoration, and quality-gated restoration. Each strategy receives a fresh Base/Dynamic pair for every case, giving 1,800 rollouts. The released X-VLA checkpoint, policy seed 195, and native action horizon and query interval $H = Q = 30$ are fixed. Within each strategy, Dynamic replays that strategy's Base actions until the assigned event; strategies share case IDs and initial physical states, but can have different executed prefixes. All 300 cases remain in the denominator, including those that do not reach the trigger. An independent 16-case development set checks execution and instrumentation. The formal cohort and restoration parameters were fixed before formal evaluation, without using its outcomes to select or adjust them. All 900 pairs completed; recorded pre-event actions and end-effector/trigger-object positions matched within each pair.

Observation-only restoration. At each policy query, the module independently examines the received agent-view and wrist RGB images. A high-pass median absolute response on the green

Table 17: **Restoration improves Dynamic success with a small aggregate Base change.** Paired differences use all 300 X-VLA cases and are in percentage points. Gap closure equals Dynamic gain minus Base gain. Intervals are pointwise 95% bootstrap intervals.

Comparison	Quantity	Estimate	95% CI
Gated vs. native	Dynamic gain	+7.0	*[3.0, 11.3]*
	Base gain	+0.3	*[−1.0, 1.7]*
	Gap closure	+6.7	*[2.3, 11.3]*
Always vs. native	Dynamic gain	+6.7	*[2.3, 11.0]*
	Base gain	+1.7	*[−0.7, 4.0]*
	Gap closure	+5.0	*[0.0, 10.0]*
Gated vs. always	Dynamic gain	+0.3	*[−3.3, 4.0]*

channel estimates noise, excluding inferred occlusions and their boundaries. The module identifies occlusions as near-black connected components (all RGB values at most 2) occupying at least 0.2% of the image, with both bounding-box sides at least 8 pixels and rectangular fill at least 85%. Estimated masks covering more than 30% of an image are discarded. The gate activates when estimated noise is at least 12 on the 0–255 intensity scale, or the inferred mask covers at least 3.5% of the image. Restoration applies Telea inpainting with radius 3 pixels, followed by a bilateral filter with diameter 5, color scale 45, and spatial scale 2 pixels. Always-on restoration uses the same operator at every query; the gated version passes unflagged images through unchanged. The module is stateless and requires no training, clean image, simulator state, event flag, or ground-truth occlusion mask. It responds to current image quality rather than estimating an event-onset time.

H.2 Paired effects

We compare gated restoration with the native policy and use always-on restoration as an ablation. We report paired 95% percentile bootstrap intervals from 10,000 resamples, stratified by source category and clustered by task/initial-state group. The cohort contains 300 distinct groups. For strategy s, define gap closure as $(B_{\text{native}} - D_{\text{native}}) - (B_s - D_s)$, where B and D are success rates. We report Base and Dynamic changes separately so that a drop in Base success cannot be mistaken for recovery. Table 17 shows a 7.0-point Dynamic gain and a 6.7-point gap reduction for the gated strategy, with a 0.3-point Base change. It converts 33 native Dynamic failures to successes while losing 12 native successes. Always-on restoration also improves Dynamic success. The 0.3-point difference between the two strategies is too uncertain to establish whether gating improves success over always-on processing. A 19.3-point Base/Dynamic gap remains under gated restoration.

H.3 Coverage and processing cost

Gated restoration increases Dynamic success on both sources and at both corruption settings (Table 18). The intervals are pointwise; the two corruption settings vary both noise and occlusion. Gated restoration activates on 1.43% of measured Base queries and all fresh Dynamic queries. Replayed prefixes are excluded from these activation denominators, and Base may already contain source-level visual corruption. Measured preprocessing of both images takes 9.1 ms on average in Base and 48.8 ms in Dynamic (95th percentiles: 9.9 and 58.8 ms); always-on means are 14.6 and 48.9 ms. These query-level measurements include quality estimation and restoration, exclude policy inference and simulation, and describe the trajectories actually visited. Native diagnostic measurement alone takes approximately 9–10 ms, so these are not end-to-end speedup estimates.

I Discussion and Future Directions

I.1 Interpreting progress on the benchmark

The paired gap measures how adding an online event changes task success after the same executed history. A policy must handle both the changed conditions and the need to continue from its current trajectory. Reset controls help distinguish these demands, while paired transitions show whether a

Table 18: **Dynamic gains persist across both sources and corruption settings.** Success rates are percentages; gains are percentage points for gated versus native on the same cases. Intervals are pointwise 95% bootstrap intervals. Noise denotes its standard deviation; mask denotes image occlusion.

Group	N	Native	Gated	Gain	95% CI
Plus	210	44.3	51.0	+6.7	*[1.4, 12.4]*
PRO	90	32.2	40.0	+7.8	*[2.2, 13.3]*
Noise 24, mask 8%	150	53.3	60.7	+7.3	*[2.0, 12.7]*
Noise 36, mask 16%	150	28.0	34.7	+6.7	*[1.3, 12.7]*

higher Dynamic rate preserves existing successes or solves different cases. An effective response should improve post-change completion while retaining Base success. Released policies also differ in training, scale, and action interfaces, so family rankings alone do not identify which design causes greater robustness.

I.2 Scope and limitations

The benchmark evaluates simulated manipulation using LIBERO's visual, contact, and task abstractions. Its main track introduces one persistent event from a pre-grasp trigger family. This design supports controlled comparisons across event types; transfer to physical robots, later task stages, and repeated or reversible changes requires further evaluation. Geometry screening checks support, workspace, visibility, and collisions, but the human feasibility audit remains incomplete. Some interventions also occur abruptly in the simulator, so their naturalness and physical realization need separate assessment.

The camera controls cover three policies and 1,000 fixed camera cases per policy. The restoration study covers one policy and one sensor-noise event. These studies show how viewpoint difficulty and execution history affect success, and how one restoration module can help. Testing more events, policies, and policy seeds would show how far these findings generalize. Endpoint success and post-event queries measure observable behavior; they do not directly establish internal change detection or deliberate replanning. Safety outcomes were not measured, so task completion cannot establish safe deployment. Measuring contacts and constraint violations would allow future evaluations to assess how safely a policy completes the task.

I.3 Learning and evaluating responses to change

Policies could be trained to handle changes during execution. Separately generated dynamic-event trajectories could expose them to altered observations, displaced targets, and recovery behaviors while reserving the frozen benchmark cases for evaluation. Adaptive VLA and WAM policies could learn when to preserve an action sequence, acquire fresh evidence, or revise the plan. Event-aware rewards could favor restored task progress and successful completion while accounting for unnecessary interruptions, additional queries, and unsafe contacts. The benchmark's diagnostics could then guide what these policies are trained to improve.

One proposed mechanism is to compare the next observed representation with a latent state predicted from the preceding observation and action. A task-conditioned gate could use this disagreement to continue execution, cancel the pending action suffix and query again, or stop before replanning. Task relevance matters: a visual change need not invalidate the current action. This remains a future method, to be evaluated against fixed-cadence execution and simpler change triggers under matched observation and computation budgets, including its effects on Base success.

Extensions could also introduce transient, recurrent, and compound event sequences in separate supplemental tracks. Object changes offer a complementary direction to the novel-instance generalization studied by ORION (Zhu et al., 2026): a cover could restrict a target's grasp regions, or a detached handle could remove a usable part during execution. Such cases would require stable contact behavior and demonstrated feasibility after the change. Base, reset-change, and mid-task controls could then distinguish changed-object difficulty from execution history, while recovery la-

tency, action cancellation, and contact measurements characterize the response beyond endpoint success.