WHAT SUCCESS RATE HIDES: MAPPING DISTURBANCE TOLERANCE IN VISION-LANGUAGE-ACTION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies are ranked by task success, but success is an endpoint, not a description of behavior. Two policies that both finish a task from a nominal state can respond very differently when the same in-progress state is physically displaced. We measure this directly. A calibrated, policy-free push creates one mid-task release state that is restored bit-identically for every policy, and a disturbance-tolerance map records next-milestone success over push direction and severity. A scalar score Q_μ summarizes the map under a declared disturbance distribution; the map shows where policies differ. Across three released LIBERO policies, clean success leaves tolerance unresolved: 18 of 20 fresh anchors at which every policy is clean-competent separate under at least one matched disturbance, and all 28 selected decisive cells retain their gap after expansion to 10 stochastic draws, including local reversals of the aggregate ranking. The differences persist when execution cadence is equalized. VLA evaluation should report clean competence, the tolerance score, and the map, together with the execution setting that produced them.

1 INTRODUCTION

Figure 1 shows two VLA policies carrying a bowl toward a plate. Both complete this step every time from the undisturbed state. We push the gripper 29 mm sideways, release it, and hand the identical release state to each policy. GR00T N1.7 places the bowl in all 10 draws; SmolVLA misses the plate in all 10. A success-rate table assigns the two policies the same label at this state.

Task success is the right headline for a benchmark, but it compresses an entire closed-loop episode into one bit. It does not say how a policy behaves once execution leaves the path it was trained to follow, and in manipulation that happens constantly: contacts slip, objects shift, and controllers lag. Whether a policy recovers from a displaced state is a property of its behavior off the nominal path, and success on that path does not reveal it.

We treat off-nominal behavior as a hierarchy of measurements. *Clean competence* p_0 asks whether a policy completes the next milestone from a state. *Local response* asks how its short-horizon progress changes after a small displacement. *Disturbance tolerance* asks which finite displacements it can still recover from. The last level is summarized by a tolerance map $M_\pi(s, d, r)$, the probability of reaching the next milestone after a push of direction d and realized magnitude r from state s, and by its average Q_μ under a declared distribution of pushes. No level determines the next, and the score does not determine the map.

Measured this way, clean ties hide large and reproducible differences. On 18 of 20 fresh anchors at which all three policies are clean-competent, some matched push separates two policies by at least 0.6 in success rate. All 28 selected decisive cells keep their gap at 10 draws, and six of them favor the policy that is weakest on aggregate. Equalizing how often the policies re-plan does not remove the phenomenon.

Contributions.

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

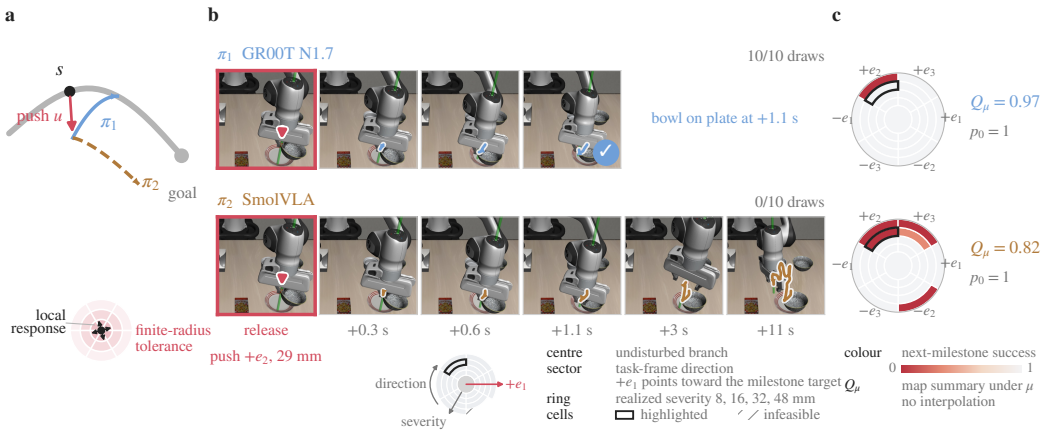


Figure 1: Same clean success, different disturbance tolerance. **a** A push moves the state off the nominal path; local response probes its immediate neighborhood, tolerance asks which finite pushes are recoverable. **b** One LIBERO transport state. A policy-free 29 mm push along $+e_2$ produces a single release state (red outline) shared by both policies; lines trace the end effector. GR00T N1.7 places the bowl 1.1 s later; SmolVLA does not. **c** Tolerance maps of the same state, with the encoding used by every map in the paper (key, bottom): the centre is the undisturbed branch, each sector a task-frame direction ($+e_1$ points toward the milestone target), each ring a realized severity (8, 16, 32, 48 mm, inside out), and color next-milestone success. The outlined cell is the push shown in **b**; hatched cells are infeasible; cells are never interpolated. Q_μ summarizes the map under the declared distribution μ.

- **A hierarchy for off-nominal VLA evaluation** that separates clean competence, local response, and finite disturbance tolerance, with results showing that none of these levels determines the next.
- **Matched disturbance-tolerance maps**: calibrated policy-free pushes from identical release states, summarized by a score Q_μ and localized by a direction $\times$ severity map.
- **Prospective evidence** that policies tied on clean execution differ in tolerance, that the differences are local and replicate, and that they survive a common execution cadence.

2 Success Does Not Determine Off-Nominal Behavior

Clean competence. At a task state s, clean competence $p_0(\pi, s)$ is the probability that policy π reaches the next declared milestone without disturbance. It is a prerequisite for interpreting any disturbed failure: a policy that fails from s undisturbed is not fragile to a push, it is simply not competent there.

Local response. Let u be a small probe command and ϕ a task-progress potential that decreases toward completion (Appendix B). The local response over h controls is measured relative to a matched zero branch,

$$z_h(u) = \big[\phi(s_h(u)) - \phi(s_0(u))\big] - \big[\phi(s_h(0)) - \phi(s_0(0))\big], \tag{1}$$

so negative values mean the probed branch makes more progress. To first order $z_h(u) \approx g^\top u$, and measuring response means estimating the slope g.

Imitation learning constrains a policy mainly on the states its demonstrations visit (Ross et al., 2011; Simchowitz et al., 2025). The first result makes the consequence exact: nominal behavior leaves the off-nominal feedback free.

Proposition 1 (Nominal agreement does not determine local response)**.** *Two policies can produce identical rollouts on a nominal state manifold and still have opposite local closed-loop behavior off it: one stabilizing and one destabilizing the transverse error.*

Proposition 2 (Probe rank)**.** *With noiseless responses $y = Xg$ to probe commands $X \in \mathbb{R}^{B\times 3}$, the slope g is identified if and only if* $\operatorname{rank} X = 3$*. With fewer than three probes, the unobserved direction is set only by the regularizer.*

Proposition 2 makes probe geometry matter at small budgets. On held-out tasks, four A-optimal probes estimate held-probe response with 0.69 [0.63, 0.76] of the error of a random four-probe design, and weighting the design toward task-relevant directions helps when only two probes are affordable (0.76 [0.69, 0.87] of the A-optimal error). Local response is therefore cheap to measure when the probes are geometrically informative (Figure 2d). It is still a local quantity, and the next section asks the finite question directly.

3 Mapping Finite Disturbance Tolerance

Matched disturbances. An anchor s is a saved in-progress simulator state, either approaching an object (pick) or carrying it (transport). A task frame fixes six directions $\pm e_1, \pm e_2, \pm e_3$: e_1 points from the end effector toward the current milestone target, e_2 is orthogonal to it in the gravity plane, and $e_3 = e_1 \times e_2$. From the anchor, a bounded controller translates the end effector for 0.4 s with the gripper held and no policy in the loop. The realized severity r is the end-effector offset from the matched zero branch at release. Each release state is generated once and restored bit-identically for every policy, draw, and execution setting (Appendix C).

Tolerance map and score. The map is the success probability from the matched release $R(s, d, r)$,

$$M_\pi(s, d, r) = \Pr\{\pi \text{ reaches the next milestone within } 6\,\text{s} \mid R(s, d, r)\}, \tag{2}$$

and the tolerance score is its average under a declared distribution μ over directions and severities,

$$Q_\mu(\pi, s) = \mathbb{E}_{(d,r)\sim\mu}\big[M_\pi(s, d, r)\big]. \tag{3}$$

We use uniform weights over feasible directions and a trapezoidal average over severity. A push that contact prevents from reaching its target is declared infeasible by the release program before any policy runs.

What the map adds. Local agreement does not settle finite outcomes.

Proposition 3 (Local agreement, distant disagreement)**.** *Two policies can agree on the clean rollout and on every sufficiently small probe, yet have opposite outcomes after a finite matched disturbance.*

The construction changes the policy only near the disturbed trajectory, far from the clean and probe branches (Appendix A). Conversely, a finite tolerance grid can also leave local feedback behavior underdetermined (Appendix A). The score, in turn, is a linear summary of the map:

Lemma 1 (Score and map)**.** *For maps M_A, M_B on a common support, $Q_\mu(A) \geq Q_\mu(B)$ for every distribution μ if and only if $M_A \geq M_B$ in every cell.*

Unless one map dominates the other everywhere, some choice of μ reverses the ranking. Q_μ answers how much of the declared disturbance distribution a policy tolerates; M shows where.

The deployed policy. Tolerance belongs to the deployed stack $\Pi_{\text{deploy}} = (\pi_\theta, e, \text{controller})$, where e is the execution cadence: how many actions of each predicted chunk are executed before the policy is queried again. Changing e changes the closed loop even with identical weights, so maps at different cadences are different estimands.

4 Experimental Setting

Policies. We evaluate three released LIBERO policies (Liu et al., 2023): OpenVLA-OFT (Kim et al., 2024; 2025), SmolVLA (Shukor et al., 2025), and GR00T N1.7 (NVIDIA, 2025), each with its native preprocessing and the same end-effector controller. SmolVLA and GR00T sample actions stochastically; OpenVLA-OFT is deterministic. At native cadence SmolVLA executes 50 controls per query, while OpenVLA-OFT and GR00T execute 8.

Table 1: Levels of off-nominal evaluation.

Level	Question	Measurement
Clean competence p_0	Does it finish the step undisturbed?	zero branch
Local response z_h	How does it react to a small displacement?	5 mm probes
Tolerance map M	Which finite pushes can it recover from?	matched pushes, direction $\times$ severity
Score Q_μ	How much of a declared push distribution?	average of M under μ

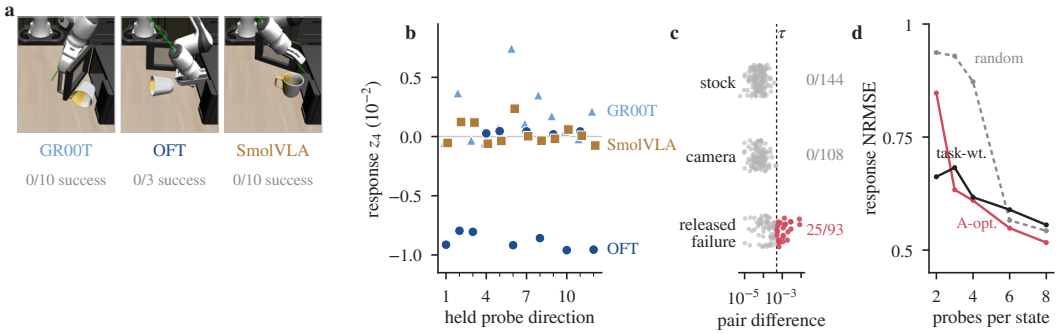


Figure 2: Same outcome, different local response. **a** Final frames from one released failure state; all repeated rollouts of all three policies fail. **b** Zero-relative response z_4 to 12 held 5 mm probes at the same state. **c** Pairwise response differences by state type; red points exceed the materiality threshold τ. **d** Held-probe response error against probe budget on clean and camera-shifted states (median over 36 cells).

Response study. The local-response panel covers 12 LIBERO tasks held out from design development and 115 states: stock initial states, camera-shifted states, and released failure states. Design selection and scoring use disjoint probe banks.

Tolerance study. A reference panel of 18 anchors over eight tasks provides full maps (8–48 mm). The prospective panel is built for the central question. Of 31 fresh candidate anchors, screened on zero branches only, 20 are clean-competent for all three policies ($p_0 = 1$ in every clean draw), spanning seven tasks, 7 pick and 13 transport states. Each is pushed along all six directions at 16, 32, and 48 mm. SmolVLA and GR00T receive five draws per cell and OpenVLA-OFT one. A *hidden difference* is a cell in which two clean-competent policies differ in success rate by at least 0.6. Aggregate scores average anchors within a task and then tasks; intervals are 95% task-cluster bootstrap intervals.

5 Results

5.1 Same Outcome, Different Local Response

Outcome ties can hide different responses. Response differences above the materiality threshold never occur at stock or camera-shifted states (0/252 policy pairs) but do occur at released failure states (25/93). Most of these are outcome ties: 19 of the 25 material pairs have the same ordinary rollout outcome, usually failure. Figure 2 shows one such state. Every repeated rollout of every policy fails, yet OpenVLA-OFT responds to almost every held probe with a progress shift near 10^{-2}, an order of magnitude larger than GR00T or SmolVLA. Two policies can share the same failure and still behave differently near it. A fresh released-failure panel reproduces the pattern: 11 of 33 policy pairs differ materially in local response while their repeated outcomes tie (Appendix E).

5.2 Clean Ties Hide Different Tolerance

Clean success leaves disturbance tolerance unresolved. On 18 of the 20 fresh clean-competent anchors, at least one matched push separates two policies by 0.6 or more (Figure 3a,b). The anchors

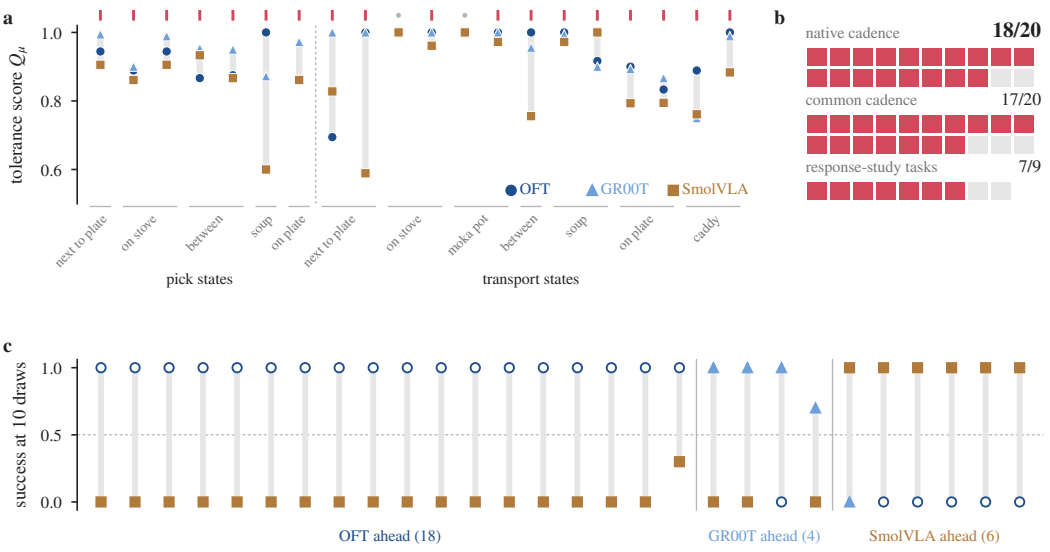


Figure 3: Clean ties hide different tolerance. **a** Tolerance score of each policy at the 20 fresh anchors where all three are clean-competent, grouped by stage and task; red ticks mark anchors with a hidden difference (a cell with a success gap ≥ 0.6). **b** Raw counts of anchors (or tasks) with a hidden difference, one square each. **c** The 28 decisive cells at 10 draws, showing the two policies of each decisive pair and grouped by which policy succeeds. Open circles: deterministic OpenVLA-OFT.

were chosen by clean competence alone, before any disturbed outcome was observed, so this is the rate at which clean ties conceal tolerance differences on these tasks rather than a curated set of examples. The phenomenon also transfers to the tasks of the response study: 9 of those 12 tasks yield a clean-competent anchor, and 7 of the 9 show a hidden difference.

The decisive cells are not sampling noise. For each hit anchor, one decisive cell was fixed by a predeclared rule and expanded to 10 draws for the stochastic policies. All 28 cells retain their sign and a gap of at least 0.5, most of them at 10/10 against 0/10 (Figure 3c).

The differences are local. In six of the 28 cells SmolVLA, the policy with the lowest aggregate score, is the one that succeeds. Figure 5 shows the pattern on reference anchors: SmolVLA loses an entire lateral sector at one transport state, yet tolerates pushes that defeat OpenVLA-OFT at a pick state. Two anchors from the task-breadth panel contain both directions of the same axis, one where GR00T fails ($+e_3$, 48 mm) and the opposite one where SmolVLA fails ($-e_3$, 48 mm).

The local reversals are behavioral, not only numerical. Figure 4 replays two of them from their logged release states. At an alphabet-soup transport state, a 46 mm push along $-e_3$ leaves SmolVLA dropping the can into the basket in all 10 draws, while GR00T carries it to the basket and keeps holding it above the rim in all 10. At a bowl-on-stove pick state, SmolVLA grasps the bowl after the same lateral push in 10 of 10 draws; OpenVLA-OFT, which is deterministic, does not complete the grasp. Both policies were clean-competent at each state, and the policy that loses on aggregate wins the cell. A lower score therefore does not mean weaker everywhere, and the map says where the exception lies. More matched rollouts, balanced over policies, stages, and cadence arms, are collected in Appendix G.

5.3 The Score Summarizes; the Map Localizes

Q_μ provides a stable aggregate summary. On the fresh panel, OpenVLA-OFT and GR00T score 0.933 [0.899, 0.967] and 0.947 [0.913, 0.978], and SmolVLA 0.863 [0.816, 0.916]. In the reference panel the same ordering holds under small-disturbance-heavy, stress-heavy, and lateral-only weightings (Appendix I).

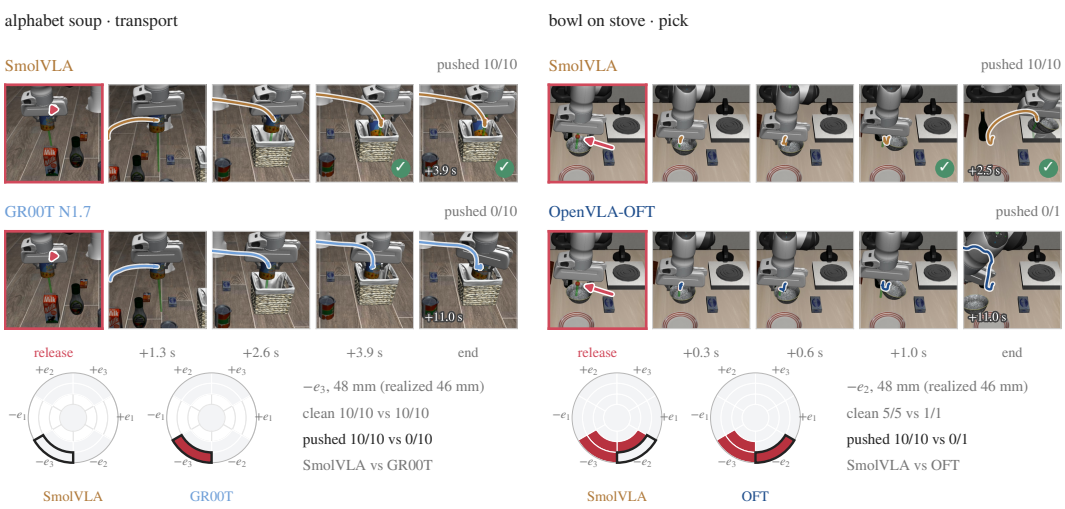


Figure 4: Local reversals in real rollouts. Each column is one clean-competent anchor and one matched push (red outline: the shared release frame; arrow: push direction). Lines trace the end effector; a check marks the next milestone. Maps below use the Fig. 1 encoding at the anchor's five-draw panel and outline the decisive cell; counts use all draws of that cell (10 for SmolVLA and GR00T, one deterministic run for OpenVLA-OFT). Frames are re-rendered from logged actions without re-running any policy.

The score cannot replace the map. OpenVLA-OFT and GR00T are within a few hundredths of each other, and maps cross: in 8 of the 54 anchor–pair comparisons of the reference panel, each policy wins at least one cell by 2/3 or more, and five pairs with equal scores fail in different cells. By Lemma 1, such pairs have no ranking that holds for every disturbance distribution. Figure 5 (right) is an example: OpenVLA-OFT and SmolVLA have nearly the same score at a pick state, but OpenVLA-OFT fails the large pushes away from the object ($-e_1$) that SmolVLA tolerates, while SmolVLA fails several small pushes. The map exposes which disturbances drive a score and whether a lower score reflects a broad deficit or one severe weakness.

5.4 Tolerance Differences Persist at a Common Cadence

SmolVLA executes 50 actions between queries, the other policies eight, so its lower tolerance could reflect slower feedback rather than learned behavior. We test this with a common cadence (C8): every policy is re-queried after eight executed controls, which leaves OpenVLA-OFT and GR00T unchanged and only increases SmolVLA's feedback rate. All policies see the same release states as in the native arm.

The gap remains (Figure 6a). Under C8, SmolVLA scores 0.810 [0.661, 0.940] and the mean gap to the two other policies is 0.130 [0.012, 0.263], against 0.077 [0.046, 0.111] at native cadence; 17 of the 20 anchors still show a hidden difference. Excluding the two anchors at which C8 breaks SmolVLA's clean competence leaves a gap of 0.064 [0.012, 0.124].

Cadence does change behavior, and the change depends on the stage. Faster re-querying helps SmolVLA before the grasp ($\Delta Q_\mu = +0.057$ [0.021, 0.092]), where its failures are mostly missed grasps. During transport it halves the number of dropped objects but multiplies placement failures, and the transport score does not improve (-0.065 [-0.246, 0.096]). Over a dose-response range from 50 to 4 controls per query, next-milestone success on 32 disturbed cells rises from 0.30 to 0.47 while policy queries per rollout rise from 4.0 to 33.6 (Figure 6b). At C8, SmolVLA uses 10.3 queries per disturbed rollout instead of 2.1. A cadence change is therefore a new deployed policy: it must be re-checked for clean competence and reported with its query cost.

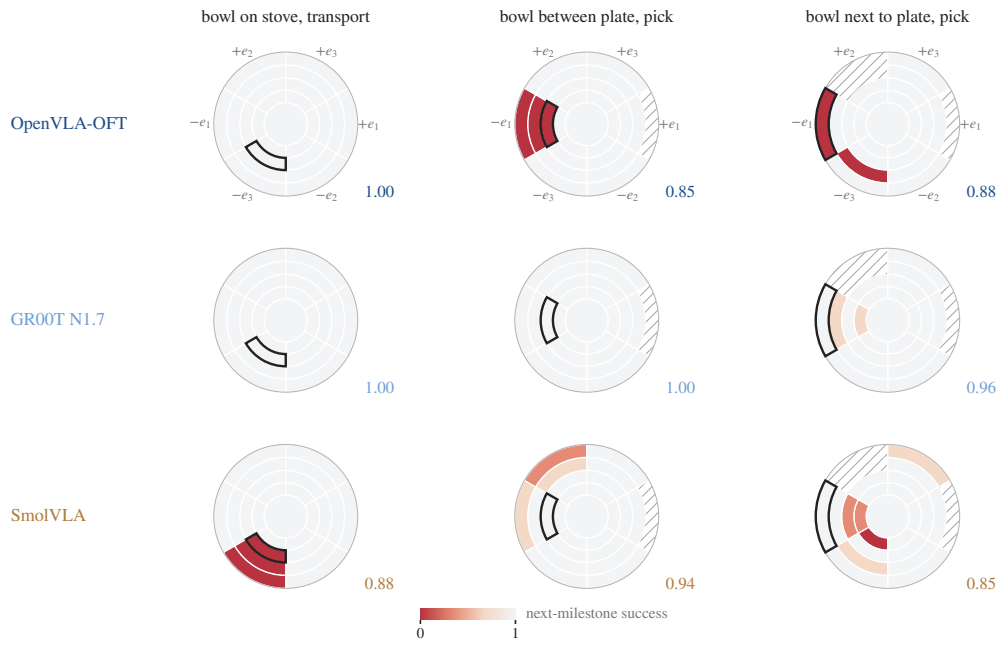


Figure 5: Tolerance maps at three reference anchors at which all policies are clean-competent; each column shares one set of release states. Left: a transport weakness specific to SmolVLA. Middle: a local reversal, where SmolVLA tolerates pushes that OpenVLA-OFT does not. Right: nearly equal scores with different failing cells. Angular position is task-frame direction and radius is disturbance severity, using the encoding in Fig. 1. Numbers are Q_μ; outlined cells mark the largest gap.

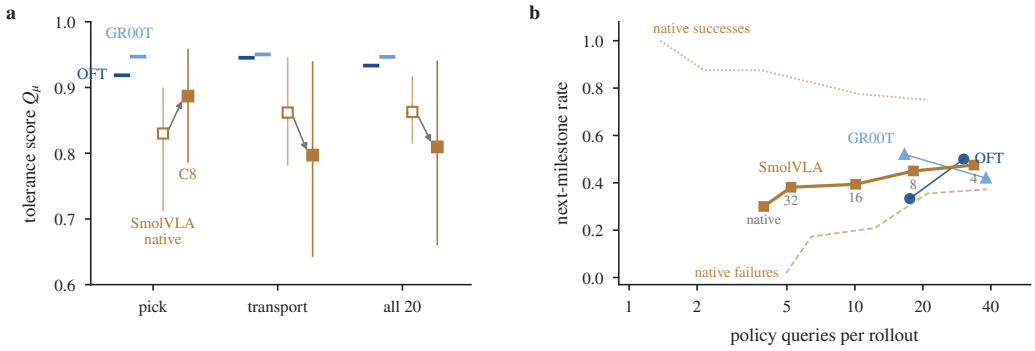


Figure 6: Tolerance differences persist at a common cadence. **a** Q_μ on the 20 fresh anchors at native cadence and with SmolVLA re-queried every eight controls (C8); OpenVLA-OFT and GR00T already run at this cadence. Bars are task-cluster bootstrap intervals. **b** SmolVLA dose response on 32 disturbed cells as controls per query fall from 50 (native) to 4, against query cost; thin lines split cells by native outcome.

6 Related Work

Evaluating VLA policies. LIBERO provides the manipulation suites used here (Liu et al., 2023). LIBERO-Plus measures how success degrades under structured perturbations of camera, lighting, layout, and language (Fei et al., 2025), and LIBERO-Recover evaluates recovery from execution-born failure states (Liu et al., 2026). IndustrialVLA-Bench reports capability, robustness, and deployment cost on separate axes (Wang et al., 2026). Active evaluation methods choose informative trials to estimate policy performance efficiently (Anwar et al., 2025; Liao et al., 2026). We add a complementary object: matched, policy-free physical disturbances applied mid-task from identical

release states, compared among policies that are tied on clean execution and resolved over direction and severity. The same concern extends beyond robotics: recent work studies conditional behavior hidden by aggregate scores in chest X-ray detection (Anonymous, 2026a), while a broader evaluation perspective asks when aggregate evaluation warrants richer measurement (Anonymous, 2026b).

Local response and experiment design. System identification and optimal experimental design formalize when finite measurements identify a local model (Ljung, 1999; Pukelsheim, 2006; Chaloner & Verdinelli, 1995), and task-optimal exploration tailors measurements to a downstream objective (Wagenmaker et al., 2021). We use these tools for the first off-nominal level and then show why local response does not settle finite tolerance.

Execution cadence and progress within episodes. Adaptive and reactive action chunking study the trade-off between responsiveness and consistency in chunked execution (Liang et al., 2026; Chen et al., 2026). Our cadence experiment is not a chunking method; it shows that tolerance must be attributed to a deployed setting. Stage-aware reward models and timestamp-level failure detectors look inside long-horizon episodes (Chen et al., 2025; Ma et al., 2026); our milestone endpoint is a coarser, fixed readout of progress after a controlled disturbance.

7 Implications for VLA Evaluation

What to report. A policy's behavior at a state is described by three quantities: whether it completes the step (p_0), how much of a declared disturbance distribution it tolerates (Q_μ), and where it fails (M). Each answers a question the others leave open.

What "better" means. Under this measurement, one policy is better than another when it tolerates more of a declared disturbance distribution at matched clean competence. When maps cross, the answer depends on which disturbances matter for the application, and the map says which ones do. In our panels SmolVLA is less tolerant on aggregate, while OpenVLA-OFT and GR00T are close and trade cells.

Cadence is part of the policy. Re-querying more often changes which failures occur and how many queries a rollout costs. Tolerance should be reported with the execution setting and query cost that produced it, and a changed setting should be re-evaluated from clean competence onward.

Scope. The study covers three released policies in LIBERO simulation, translational pushes of one duration, and next-milestone endpoints. The map is an evaluation object: it identifies operating conditions that deserve closer diagnosis or data collection. Extending it to rotations, contact-rich disturbances, and physical robots is the natural next step.

8 Conclusion

Task success tells us whether a VLA policy finishes, not how it behaves once execution is displaced. Matched disturbance-tolerance maps measure that behavior directly: policies tied on clean execution frequently separate, the separations replicate, and they persist at a common cadence. Q_μ summarizes a declared disturbance distribution and the map shows where the difference lives; both belong next to success rate.

Reproducibility Statement

Appendix A contains the proofs, the remaining appendices specify the response and disturbance-tolerance protocols and supporting analyses, and the supplementary package provides code, paper configurations, frozen result tables, and scripts for reproducing the reported figures and tables. Public policy checkpoints and LIBERO assets are referenced rather than redistributed.

AI-USE STATEMENT

We used generative AI tools to assist with research ideation and experiment design, implementation, mathematical derivations and proof drafting, literature search and synthesis, analysis, and the preparation and editing of figures and text. The authors reviewed the AI-assisted work and take responsibility for the final content of the submission.

REFERENCES

Anonymous. Beyond aggregate scores: Revealing conditional model behavior in chest x-ray nodule detection. In *Submitted to The Fifteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=OCOVgCj6RD. under review.

Anonymous. When should ai evaluation measure more? In *Submitted to The Fifteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=SXWnGa2DDw. under review.

Abrar Anwar, Rohan Gupta, Zain Merchant, Sayan Ghosh, Willie Neiswanger, and Jesse Thomason. Efficient evaluation of multi-task robot policies with active experiment selection. In *Proceedings of the Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 1636–1653. PMLR, 2025.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Qianzhong Chen, Justin Yu, Mac Schwager, Pieter Abbeel, Yide Shentu, and Philipp Wu. SARM: Stage-aware reward modeling for long horizon robot manipulation, 2025. URL https://arxiv.org/abs/2509.25358.

Wenxi Chen, Kaidi Zhang, Chi Lin, Zhiyuan Zhang, Yu She, Yuejiang Liu, Raymond A. Yeh, Shaoshuai Mou, and Yan Gu. DREAM-Chunk: Reactive action chunking with latent world model, 2026. URL https://arxiv.org/abs/2606.18589.

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, 2025. URL https://arxiv.org/abs/2510.13626.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2024.

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

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*, 2026. URL https://arxiv.org/abs/2604.04161.

Andrew Liao, Hanchen Cui, Karthik Desingh, and Aryan Deshwal. Active real-world factor-based evaluation for generalist robot policies, 2026. URL https://arxiv.org/abs/2607.14439.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Lin Liu, Zhicheng Bao, Lu Zhang, Ziying Song, Wu Yang, Yuzheng Zhuang, Shuai Tao, Wulong Liu, Caiyan Jia, and Huchuan Lu. LIBERO-RECOVER: Beyond task success towards failure recovery in robotic manipulation models, 2026. URL https://arxiv.org/abs/2609.05178.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, second edition, 1999.

Jie Ma, Zongxi Liu, and Yi Zhu. FailureSpot: Label-efficient timestamp-level failure detection for vision-language-action models, 2026. URL https://arxiv.org/abs/2609.04277.

NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. SIAM, 2006.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics, 2025. URL https://arxiv.org/abs/2506.01844.

Max Simchowitz, Daniel Pfrommer, and Ali Jadbabaie. The pitfalls of imitation learning when actions are continuous, 2025. URL https://arxiv.org/abs/2503.09722.

Andrew J. Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10641–10652, 2021. URL https://proceedings.mlr.press/v139/wagenmaker21a.html.

Yiqi Wang, Zhifeng Rao, Jiaqi Zhang, Xiaoyang Li, Zhangkai Wu, Yiqun Duan, Mingkai Zheng, Fei Wang, Shan You, and Taotao Cai. IndustrialVLA-Bench: A traceable multi-axis evaluation of open robot policy models, 2026. URL https://arxiv.org/abs/2609.25562.

A PROOFS

We state the propositions of the main text in full and prove them. Policies are maps from states (or time and state) to actions, and the plant is a discrete-time controlled system $x_{t+1} = f(x_t, a_t)$.

Proposition 1, full statement. Let $\mathcal{N}$ be an embedded C^2 nominal state manifold, π a C^1 policy, and L_z a C^1 field of linear maps on the normal spaces $N_z\mathcal{N}$. There is a C^1 policy π_L with $\pi_L = \pi$ on $\mathcal{N}$ and normal derivative L_z. If the transverse linearization $e_{t+1} = A_\perp e_t + B_\perp \delta u_t$ admits L_s, L_u with $\rho(A_\perp + B_\perp L_s) < 1 < \rho(A_\perp + B_\perp L_u)$, then π_{L_s} and π_{L_u} produce identical rollouts on $\mathcal{N}$ but have locally stable and unstable transverse error dynamics, respectively.

Proof. In a tubular neighborhood write $x = z + e$ uniquely with $z = \Pi_{\mathcal{N}}(x)$ and $e \in N_z\mathcal{N}$; the projection is C^1 there. With a smooth cutoff χ equal to one near zero, set $\pi_L(x) = \pi(z) + \chi(\|e\|^2) L_z e$. This map is C^1, equals π at $e = 0$, and has normal derivative L_z. Normal feedback $\delta u_t = L e_t$ gives the transverse Jacobian $A_\perp + B_\perp L$, whose spectral radius sets local stability. The nominal restrictions coincide, so every nominal experiment has the same observation law under both policies. The result is local, in the unrestricted C^1 class; it does not assert a particular architecture and cannot make an unactuated direction controllable. □

Proposition 2. A slope g' matches noiseless data iff $X(g' - g) = 0$. If $\operatorname{rank} X = 3$, $X^\top X$ is invertible and $g = (X^\top X)^{-1} X^\top y$. Otherwise there is a unit $n \in \ker X$ and $g + tn$ fits for every t. For $B < 3$, $X^\top X$ has a zero eigenvalue, so $X^\top X + \lambda I$ has smallest eigenvalue λ for every design, and the ridge estimate along n is fixed by the regularizer alone. □

Proposition 3, full statement. Let $\mathcal{X}$ be a smooth paracompact manifold or an open subset of $\mathbb{R}^n$, and let policies be time-indexed maps $\pi(t, x)$, smooth in x. Fix π_1, anchor s, horizon T, and an open clean/probe tube $U \subset [0, T] \times \mathcal{X}$ containing the clean trajectory graph and all sufficiently small probe branches through horizon h. Let $x^* = R(s, d, r)$. Suppose there are action sequences $a^{\pm}_{0:T-1}$ whose trajectories $\tau^{\pm}$ from x^* satisfy: (i) τ^+ reaches the milestone within T and τ^- does not; (ii) their time graphs $V_{\pm} = \{(t, \tau^{\pm}_t)\}$ are compact and disjoint from a closed core $K \subset U$; (iii) $\pi_1(t, x)$ and $a^{\pm}_t$ lie in a convex subset of the action set on the construction neighborhoods. Then there are policies $\pi_{\pm}$, smooth in x, equal to π_1 on a neighborhood of K and to $a^{\pm}_t$ near $V_{\pm}$, such that the clean rollout, clean success, all finite-difference probe responses, and all derivatives of the clean/probe branch map coincide, while $M_{\pi_+}(s, d, r) = 1$ and $M_{\pi_-}(s, d, r) = 0$.

Proof. Choose smooth bumps $\varphi_{\pm}$ on $[0, T] \times \mathcal{X}$ that vanish near K and equal one near $V_{\pm}$ (smooth Urysohn lemma), and set $\pi_{\pm} = (1 - \varphi_{\pm})\pi_1 + \varphi_{\pm} a^{\pm}_t$, which is admissible by convexity. Equality with π_1 on an open neighborhood of K gives equality of every clean and probe quantity and of all their derivatives. Near $V_{\pm}$ the policy reproduces $a^{\pm}$, so the disturbed rollout follows $\tau^{\pm}$ and realizes the chosen outcome. Stochastic policies use the same construction on mixtures of action kernels. This is a non-determination result for an unrestricted policy class, not a statement about a fixed architecture. □

The map does not determine local response. Conversely, let the tolerance grid be a finite set of releases with $r \geq r_{\min} > 0$. If the grid rollouts avoid an open time-indexed tube around the clean trajectory, a bump supported in that tube changes the normal derivative along the clean trajectory, and hence the local response, while leaving the clean actions and every grid rollout unchanged.

Lemma 1. Pointwise dominance integrates to dominance under every μ. Conversely, if $M_B(c) > M_A(c)$ in some cell c, the point mass on c reverses the scores. With severity weights fixed and only direction weights free, the same statement holds for severity-integrated direction scores. □

B Local-Response Protocol

Probe. A probe with unit direction d is one control step commanding a translation of $r\,d$ with $r = 5$ mm (normalized command $d\,r/0.05$ for the controller's 0.05 m action scale), with zero rotation and zero gripper command. Every branch restarts from the same simulator snapshot, including the controller target and the policy's action queue, and the policy then runs for $h = 4$ controls.

Task-progress potential. For a task with native goal atoms $k = 1, \ldots, K$,

$$\phi = \frac{1}{K}\Big[\sum_{k \text{ unsatisfied}} \big(1 + \rho(\delta_k, \ell_k)\big) + 0.25 \min_{k \text{ unsatisfied}} \rho(\alpha_k, 0.2\,\text{m})\Big], \qquad \rho(\delta, \ell) = \frac{\delta}{\ell + \delta}. \quad (4)$$

Here δ_k is the object-to-target distance for placement atoms (scale $\ell_k = 0.25$ m) or the joint-interval gap for articulated atoms (0.2 m for slides, 1.5 rad for hinges), and α_k is the end-effector distance to the atom's source object, zero while it is grasped. Satisfied atoms contribute zero, so $\phi = 0$ exactly when all native goal predicates hold. ϕ is a declared task cost computed from simulator geometry, not a learned value function, and reading it never changes the simulator state.

Probe banks. The 24 candidate directions are 12 antipodal pairs on a golden-spiral hemisphere. The 12 held directions are six antipodal pairs of the same construction after a fixed rotation; every held direction is at least 13° from every candidate (median 22°). Designs select probes from the candidate bank only; held directions are used only for scoring. Both banks were fixed on development tasks.

Designs. Random uses a preregistered order of candidates. A-optimal greedily minimizes $\mathrm{tr}(X^\top X + \lambda I)^{-1}$. Task-weighted minimizes $\mathrm{tr}\, W(X^\top X + \lambda I)^{-1}$, where $W = \sum_i w_i c_i c_i^\top$ and the weights come from geometry alone: $s_i = |\phi(x_i^+) - \phi(x_i^-)|/(2r)$ with $x_i^{\pm}$ displacing the end effector (or grasped object) by $\pm r c_i$, and $w_i \propto s_i + 0.05\,\bar{s}$.

Fitting and normalization. Each design fits the slope by ridge regression with $\lambda = 0.05$. For each state–policy unit and design, the regression is made unit-free by scaling with the root-mean-square of the $h = 4$ responses that design has itself observed at its budget (selected-only normalization), so λ has the same meaning across units and unexecuted candidates never set the scale. Held-response error for a task–policy cell is $\big(\sum \text{squared error}/\sigma_t^2\big)^{1/2}/\sqrt{n}$ over its states and held directions, with σ_t the root-mean-square held response of task t in the same state group. Design comparisons take per-cell ratios, their median, and a task-cluster bootstrap (2,000 resamples).

Materiality. For policies p, q at a state, D_{ref} is the median over the 12 held directions of $|z_4^p - z_4^q|$. A pair is material if $D_{\text{ref}} > \tau = 5 \times 10^{-4}$ (ϕ units). The threshold was fixed before the held-out phase as the default of the development grid $\{1, 2, 5, 10\} \times 10^{-4}$ and never tuned on held-out responses.

Input validity. A state is excluded if the image of its zero branch or of any held branch is invalid under an outcome-blind rendering check (grey-level mean absolute difference between adjacent pixels; valid frames score at most 14.4, invalid frames at least 30.2). The rule removes five released states and leaves 115.

Cost. One interaction is one executed branch: restore, probe, and h controls. A B-probe measurement costs $B + 1$ interactions, including the shared zero branch.

C Disturbance-Tolerance Protocol

Clock and endpoints. Controls run at 20 Hz. The disturbance lasts 8 controls (0.4 s). The stage endpoint is met if the milestone predicate holds for five consecutive controls starting within 120 controls (6 s) of release, or if the native task succeeds by then. The task endpoint is native LIBERO success within 220 controls (11 s). Drops, collisions, displaced objects, and timeouts are failures.

Anchors and milestones. Anchors are cut from donor rollouts of one native policy per task, rotating among the three policies. A pick anchor is the first control at which the gripper is within 8 cm of the active object without a grasp; its milestone is a secure grasp or the satisfied sub-goal. A transport anchor is taken 10 controls after a grasp has been held for five controls; its milestone is the object's native placement predicate. Two released LIBERO-Recover states form a separate reference stratum. Anchors are reproduced by open-loop replay of the donor commands with a bit-exact state-hash check.

Task frame. e_1 is the unit vector from the end effector to the geometric target of the milestone atom (the object before a grasp, the goal region after it). e_2 is gravity projected orthogonally to e_1; if $|g^\top e_1| > 0.95$ (near-vertical e_1), world $+x$ is projected instead. $e_3 = e_1 \times e_2$.

Disturbance program and calibration. The program commands a translation $a\,d\,[0.5, 1, 1, 1, 1, 1, 1, 0.5]$ with zero rotation and the gripper held. The realized offset r is the end-effector distance between the disturbed and zero branches at release. The amplitude $a \le 1$ is set per anchor, direction, and target by bracketing and bisection with up to two direction corrections. A cell is accepted if $|r - r^*| \le \max(0.25\,\text{mm},\, 0.1\,r^*)$ and the direction error is at most $15°$; otherwise it is infeasible. A direction enters Q_μ only if all of its severities are feasible. Feasibility depends only on the policy-free release, never on policy outcomes. In the reference panel, 421 of 432 disturbed releases were accepted (median magnitude error 4.7%, median angle error $6.6°$); the 11 infeasible cells are $+e_1$ pushes into the object at pick anchors.

Rollout and cadence. At release the policy's action queue is cleared and the policy is queried on the release observation. At native cadence SmolVLA executes 50-step chunks, OpenVLA-OFT 8, and GR00T 8 of 16. The common cadence C8 re-queries every policy after eight executed controls and discards the rest of the chunk. Sampling seeds are fixed per draw, query, and cell.

Clean gate and hidden differences. An anchor is clean-competent for a policy if every zero-branch draw reaches the milestone (five draws for SmolVLA and GR00T, one for OpenVLA-OFT).

The prospective cohort contains anchors that pass for all three policies; screening used zero branches only. A hidden difference is a cell in which two policies differ in stage success rate by at least 0.6 (for example 3/5 against 0/5); the task-breadth panel, with three draws, uses 2/3. Under C8 the clean gate is re-evaluated because the deployed stack has changed.

Scores and intervals. Q_μ is the trapezoidal integral of the stage success rate over the severity grid divided by the largest severity, averaged over feasible directions. Aggregates average anchors within a task and then tasks. Intervals are task-cluster bootstrap intervals (10,000 resamples).

D Full Tolerance Maps

Figures 7 and 8 show every reference anchor. States where a policy is not clean-competent are marked with its p_0 and are excluded from all hidden-difference counts. Figure 9 shows the aggregate maps of the fresh panel at native and common cadence.

E Extended Local-Response Evidence

Probe budget. Table 2 lists the curve of Figure 2d. On stock and camera-shifted states, four A-optimal probes reach 0.69 [0.63, 0.76] of the held-response error of the preregistered random order. At two probes, where the slope is underidentified, task-weighted probes reach 0.76 [0.69, 0.87] of the A-optimal error, consistent with Proposition 2: weighting chooses which direction remains unobserved. For pairwise screening on released failure states, a two-probe audit costs 23% of the full held reference; the A-optimal order recovers 76% of the 25 material pairs at a false-flag rate of 0.059, against 68% and 0.118 for the random order. These are single design draws; over equally valid A-optimal tie-breaks, two-probe recall ranges from 0.52 to 0.80, so we treat sparse screening as a cost result rather than a ranking of designs.

Table 2: Median held-probe response error (NRMSE) against probe budget, 36 stock and camera-shifted cells.

Design	2	3	4	6	8
Random	0.937	0.929	0.873	0.566	0.543
A-optimal	0.847	0.633	0.610	0.549	0.517
Task-weighted	0.662	0.682	0.617	0.589	0.556

Fresh failure states. A fresh released-failure panel, rendered and scored independently of the design panel, reproduces the outcome-tie finding: 11 of 33 policy pairs are material (95% interval 0.09–0.61 of pairs), and none of the 11 differs clearly in repeated outcomes (10 both fail consistently; one is 0/5 against 1/5).

Re-query versus replay. On 12 development states, replaying each policy's clean action chunk at the perturbed state, instead of re-querying the policy, reduces the median between-policy response difference from 5.8×10^{-5} to 0.55×10^{-5}. Separation therefore decreases under clean-action replay: most of the measured difference comes from how each policy reacts to the perturbed observation rather than from the shared physical displacement. This analysis was done after the design panel was fixed and is descriptive.

Physical and task response. Across 345 pairs, end-effector response differences and task-progress response differences are correlated (Spearman 0.72). Every material pair also differs by more than 0.5 mm in end-effector response, but so do 254 of 320 non-material pairs, so physical displacement alone does not identify which differences matter for the task.

F Prospective and 10-Draw Confirmations

Of 31 fresh candidate anchors, 20 pass the clean gate for all three policies. They cover seven of the eight reference tasks (13 transport and 7 pick anchors). Hidden differences occur at 18 of 20

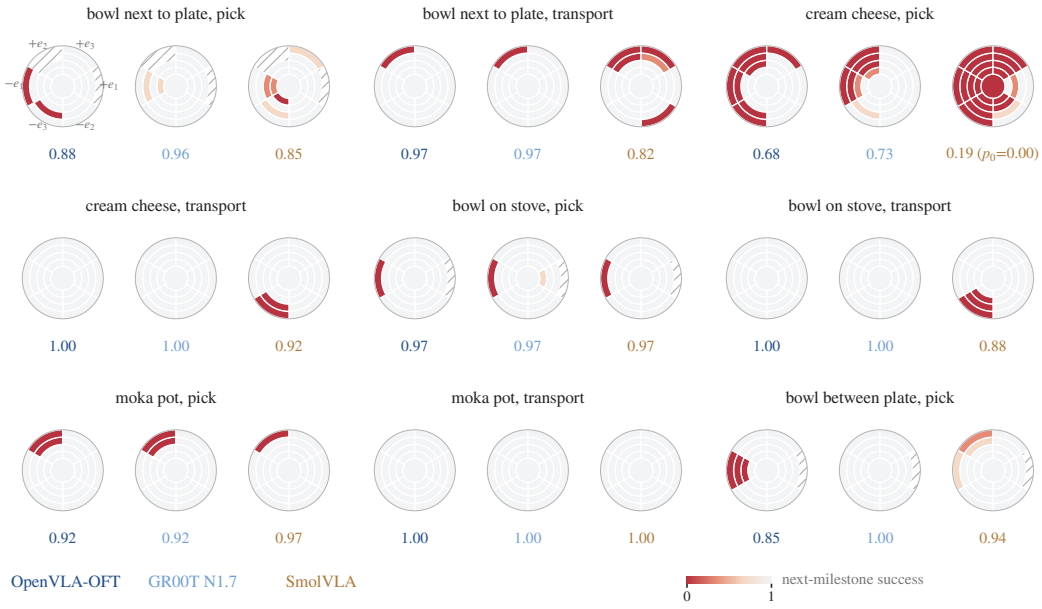


Figure 7: Reference-panel tolerance maps, part 1. Each group shows one anchor for OpenVLA-OFT, GR00T N1.7, and SmolVLA (left to right); layout as in Figure 1, rings 8–48 mm, hatching marks infeasible cells. Numbers are Q_μ.

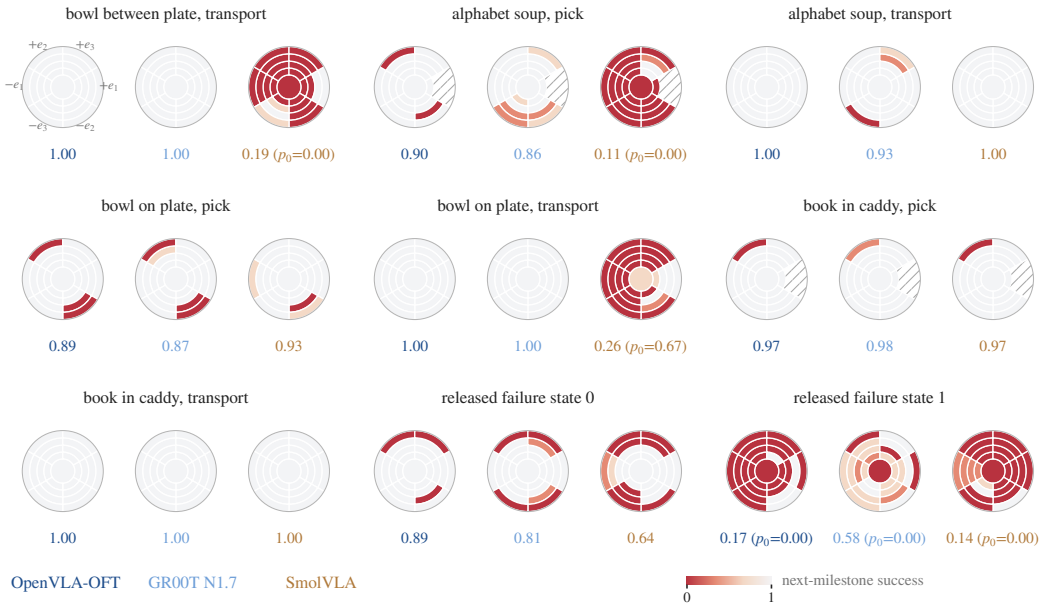


Figure 8: Reference-panel tolerance maps, part 2, including the two released failure states.

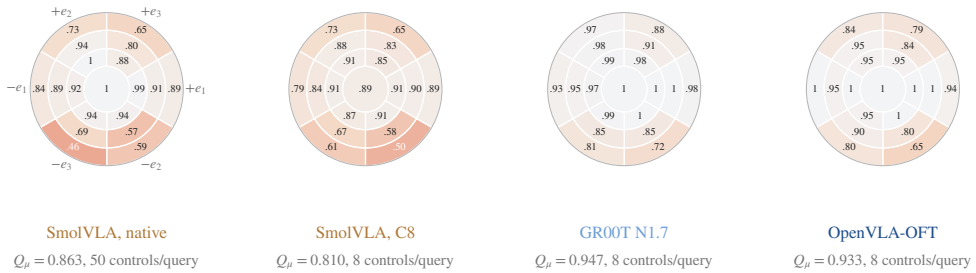


Figure 9: Aggregate maps over the 20 fresh clean-competent anchors (rings 16, 32, 48 mm; centre: undisturbed). Numbers are mean cell success rates.

anchors at native cadence and 17 of 20 under C8. In the task-breadth panel (one fresh anchor per response-study task; $\pm e_2, \pm e_3$ at 32 and 48 mm; three draws), 9 of 12 tasks yield a clean-competent anchor and 7 of 9 show a hidden difference.

For each hit anchor, one decisive cell was selected by a fixed rule before new draws and expanded to 10 draws for SmolVLA and GR00T; OpenVLA-OFT keeps its deterministic rollout. Ten cells come from an earlier fresh panel and 18 from the prospective panel. All 28 keep the sign of the original gap with $|\text{gap}| \geq 0.5$, and 25 keep a gap of 1.0. Every release state in the expansion matched its original hash. In six cells SmolVLA succeeds where the other policy of the pair fails.

G Representative Matched Rollouts

Figures 10–12 show ten further cases, chosen by fixed rules from the logged cohort: three clean ties with a large split, two further local reversals, two anchors where two policies have close scores but fail in different cells, and three cells where a common 8-control cadence (C8) either leaves SmolVLA's failure in place, introduces one, or removes one. Every row starts from the same bit-identical release state (red outline). Frames are re-rendered from logged actions, and each replay reproduced the logged end-effector trajectory to 0.0 mm and the logged outcome. Row labels give clean and pushed counts over all logged draws of that cell; maps use the Fig. 1 encoding at the anchor's five-draw panel with the row's cell outlined. Time labels are listed under each case; the last frame is the end of that rollout.

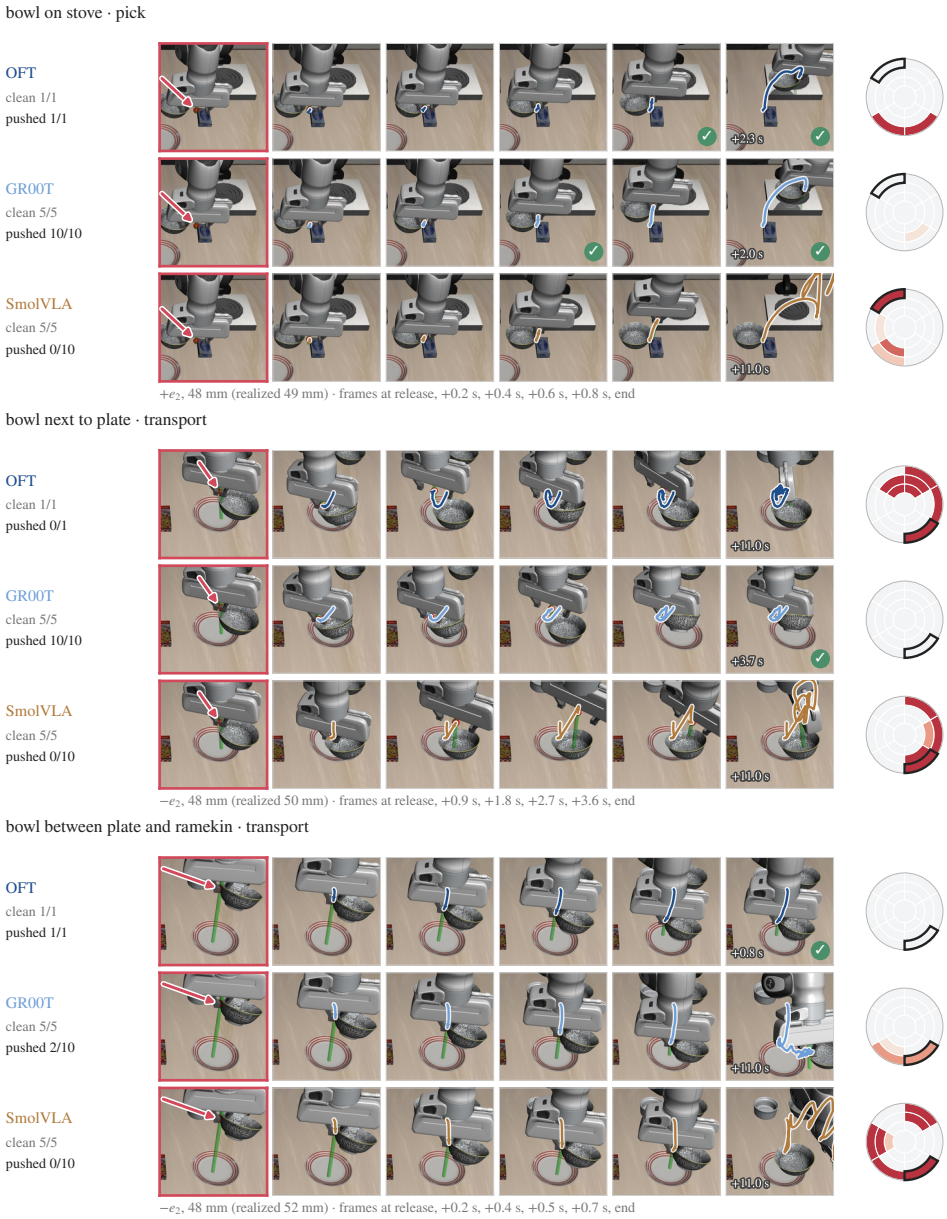


Figure 10: Clean ties with a large split. Top: GR00T and OpenVLA-OFT grasp after a lateral push that SmolVLA never recovers from. Middle: only GR00T completes the transport. Bottom: OpenVLA-OFT places the bowl; GR00T succeeds in 2 of 10 draws and SmolVLA in none.

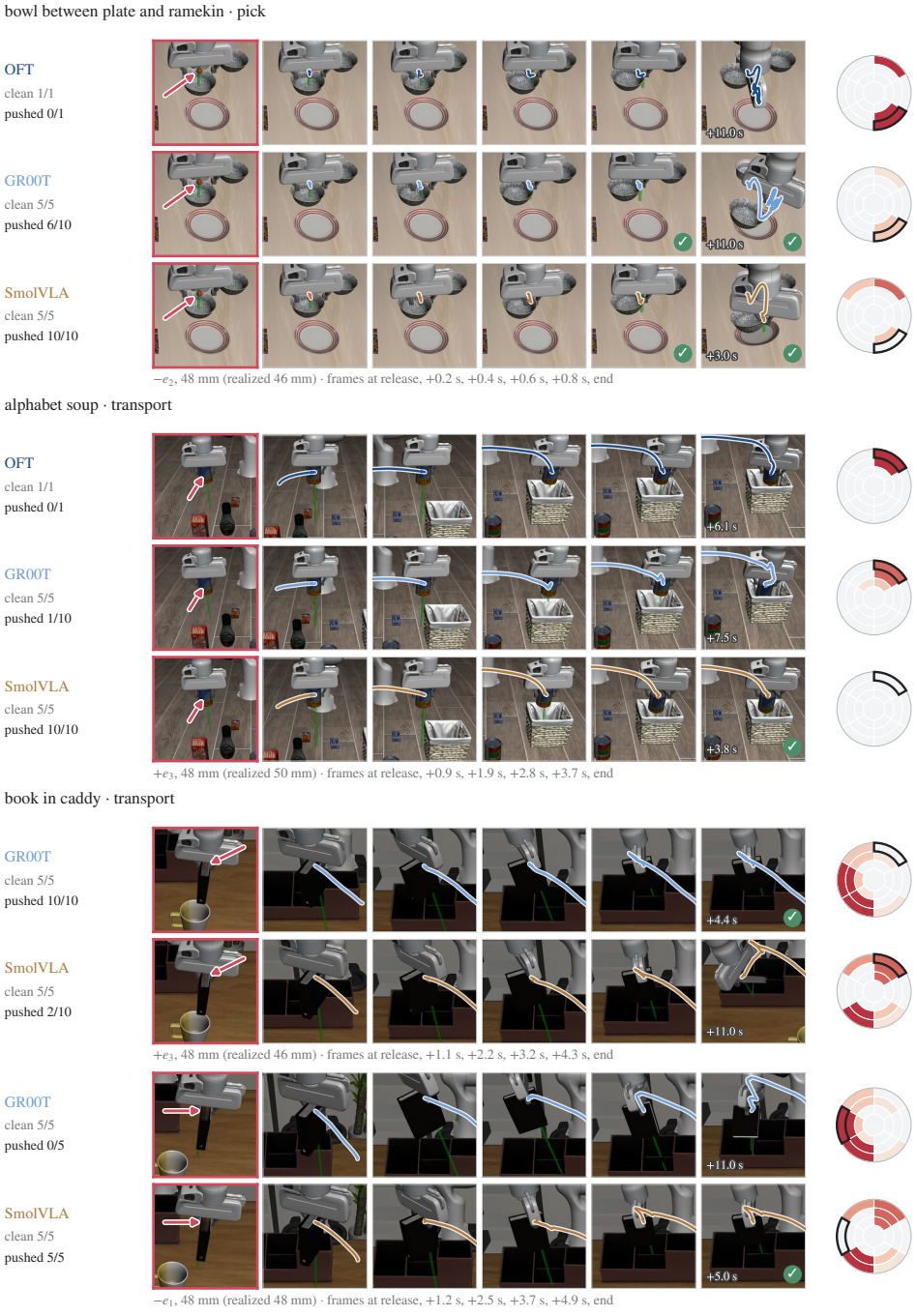


Figure 11: Top and middle: local reversals in which SmolVLA succeeds in 10 of 10 draws while OpenVLA-OFT fails and GR00T succeeds in 6 and 1 of 10. Bottom: GR00T and SmolVLA have nearly equal scores (0.750 and 0.761) at one transport state but fail in opposite cells: GR00T at $-e_1$, 48 mm, SmolVLA at $+e_3$, 48 mm.

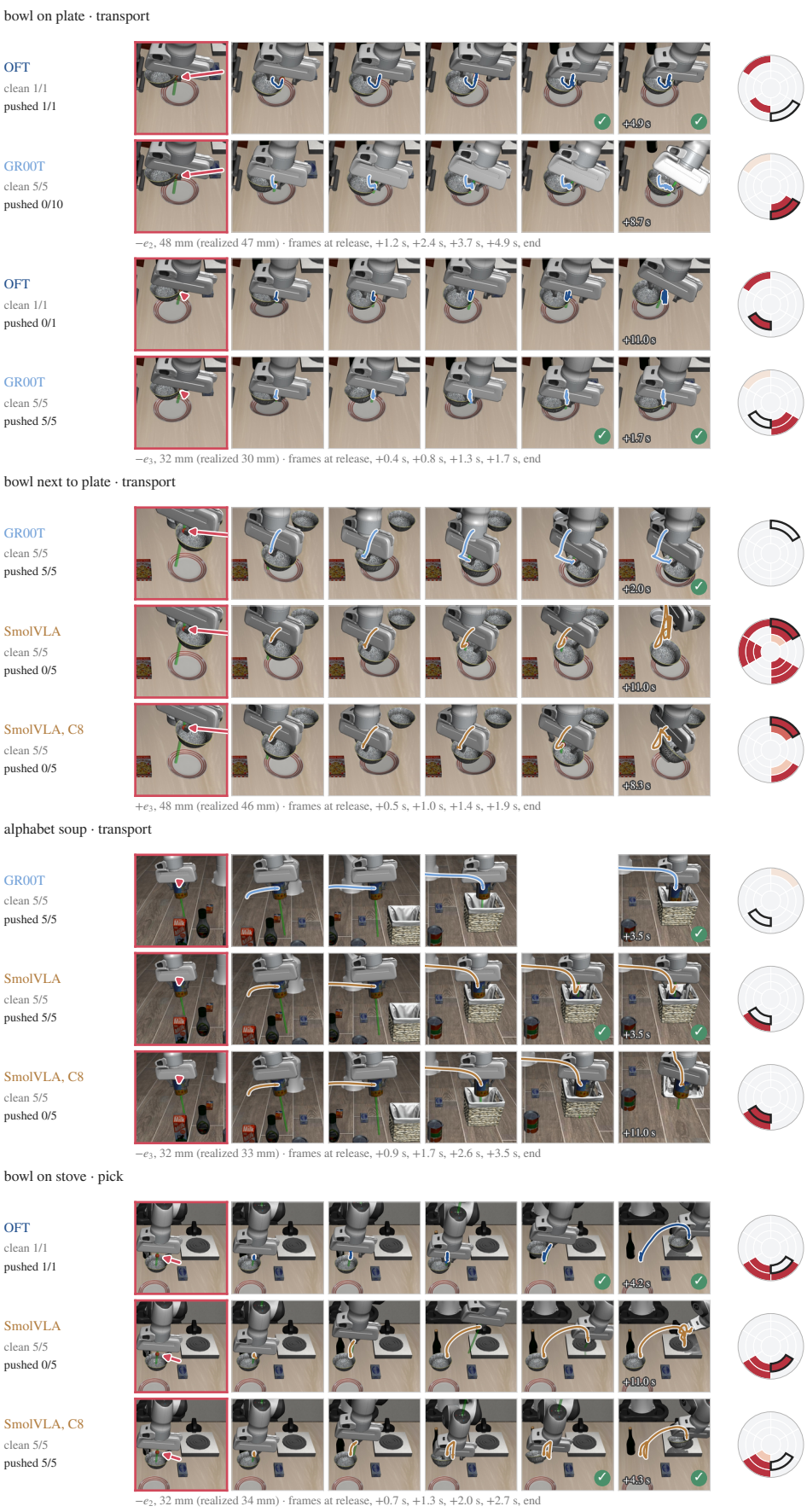


Figure 12: Top: OpenVLA-OFT and GR00T have nearly equal scores (0.900 and 0.893) but each fails a cell the other passes. Remaining cases: common cadence. SmolVLA's transport failure persists at C8 (second), a success at native cadence becomes a failure at C8 (third), and a pick

H Common Cadence and Dose Response

Table 3 gives the common-cadence comparison. The post-hoc row keeps the 18 anchors at which SmolVLA remains clean-competent under C8. Under C8, SmolVLA's missed grasps at pick anchors fall from 118 to 83 of 630 rollouts; at transport anchors, dropped objects fall from 168 to 83 of 1,170 while placement failures rise from 42 to 191. Before the grasp, all three policies fail mainly by missing the grasp; during transport SmolVLA mostly drops the object, whereas OpenVLA-OFT and GR00T more often keep it and stall near placement. All policy arms share the same release states (no release or onset hash mismatch).

Table 3: Native versus common cadence on the 20 fresh anchors. Gap: mean of OpenVLA-OFT and GR00T minus SmolVLA. Queries: SmolVLA policy queries per disturbed rollout.

	SmolVLA Q_μ	Gap	Hidden differences	Queries
Native cadence	0.863 [0.816, 0.916]	0.077 [0.046, 0.111]	18/20	2.1
Common cadence (C8)	0.810 [0.661, 0.940]	0.130 [0.012, 0.263]	17/20	10.3
C8, 18 clean-competent anchors	$\Delta = +0.030$ [-0.018, 0.075]	$0.094 \to 0.064$ [0.012, 0.124]	–	10.3
C8, pick anchors	$\Delta = +0.057$ [0.021, 0.092]	$0.103 \to 0.046$	–	–
C8, transport anchors	$\Delta = -0.065$ [-0.246, 0.096]	$0.086 \to 0.151$	–	–

Table 4: SmolVLA cadence dose response on 32 disturbed cells (25 transport, 7 pick), five draws per cell.

Controls per query	50 (native)	32	16	8	4
Next-milestone success	0.30	0.38	0.39	0.45	0.47
Policy queries per rollout	4.0	5.2	10.1	18.2	33.6

I Supporting Analyses

Score weighting. Table 5 recomputes the reference-panel scores (all 18 anchors) under three alternative disturbance distributions written down before recomputation. SmolVLA is lower under every weighting; OpenVLA-OFT and GR00T remain unresolved. Among anchor-level contrasts with $|\Delta Q_\mu| \geq 0.05$, at most one of 29 changes sign.

Table 5: Reference-panel Q_μ (task-macro, stage endpoint) under alternative disturbance distributions. Severity weights are on $\{0, 8, 16, 32, 48\}$ mm.

μ	OFT	SmolVLA	GR00T	OFT$-$SmolVLA	GR00T$-$SmolVLA
Uniform (trapezoid)	0.856	0.677	0.898	0.18 [0.05, 0.34]	0.22 [0.09, 0.37]
Small-disturbance-heavy	0.880	0.704	0.917	0.18 [0.04, 0.35]	0.21 [0.08, 0.38]
Stress-heavy	0.819	0.642	0.863	0.18 [0.05, 0.33]	0.22 [0.09, 0.36]
Lateral only ($\pm e_2, \pm e_3$)	0.852	0.645	0.885	0.21 [0.07, 0.37]	0.24 [0.09, 0.40]

Donor anchors. A policy is not favored on anchors cut from its own rollouts: every own-donor effect on the stage score is negative (-0.06 to -0.03; task-permutation $p \geq 0.60$), and the ordering is unchanged when any donor's anchors are removed.

Object contact. About half of the releases leave every object untouched. On those releases the pick-anchor gap between GR00T and SmolVLA is $+0.23$ [0.05, 0.42], and transport gaps are about 0.24 regardless of how far the carried object moves.

Draw stability. OpenVLA-OFT repeats are identical in every repeated cell (36/36 in the reference panel). Only 5.8% (SmolVLA) and 7.3% (GR00T) of reference cells have mixed outcomes across draws, anchor-level score differences correlate at 0.98–0.99 between the first two draws and a fresh third draw, and the decisive-cell expansion of Appendix F shows that the largest gaps are stable at 10 draws.

Input validity. In the tolerance study the same image check runs at every policy query; a rollout with an invalid frame is excluded from rates rather than counted as a failure. No release observation was invalid. The rule used for the response study is given in Appendix B.