Same Scene, Different Answer: Variance and Rollout Budgets for Flow-Policy Evaluation

Nihal Gunukula
Purdue University
ngunukul@purdue.edu

***Abstract*—A reported success rate for a flow-matching vision-language-action policy is a sample from a distribution nobody has decomposed. We decompose it. Holding the initial state fixed leaves 61–84% of outcome variance in play across three policies on LIBERO: the scene accounts for the smaller share. A replicate block splits that remainder on the informative task into the flow sampler (38.5% of total) and system nondeterminism (46.0%), both invisible in a single reported rate; the partition replicates at three environment seeds. The consequence is a budget: rerunning the field's standard 500-episode protocol while changing only the flow-sampling seed spans 2.2 points, a band already containing 27% of the recomputable improvements highlighted across 50 recent VLA papers, of whose highlighted comparisons 34.4% state no usable rollout count at all. From the measured components we derive rollout budgets and the paired-design efficiency gain they imply, then spend them: a powered 4×3 factorial of denoising steps × execution horizon on *released* checkpoints shows the shipped default is dominated on $\pi_{0.5}$, where a *single* Euler step at horizon 10 is statistically indistinguishable from the 10-step default at 2.7× lower measured latency; π_0 does not tolerate the same cut. The cut's one confirmed price on $\pi_{0.5}$ is 4.7 points under camera-viewpoint shift, while replanning every step costs 9–41 points. Across four perturbation axes a 16.5-point clean gap stands for anything from 9.6 to 67 points of deployed margin, and rankings invert only where an axis drives the trailing policies toward the floor. Every analysis is pre-registered or labeled exploratory in a public deviations log; per-episode logs and RoboRigor, the harness, with an exact rollout-budget calculator accompany the release.**

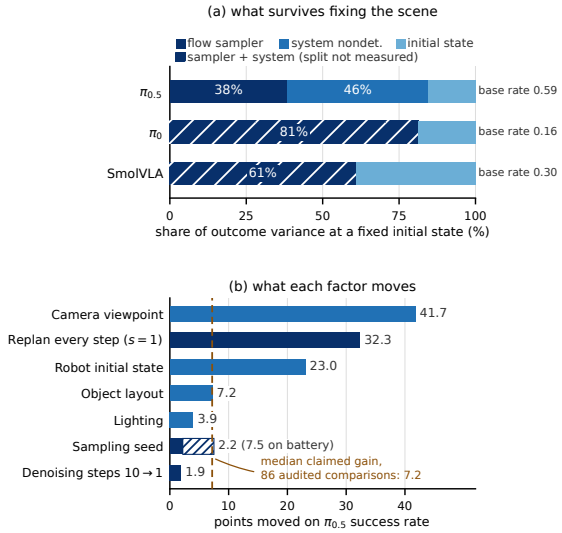

Fig. 1. (a) What survives fixing the scene, on the informative task across three policies (base rates 0.16 to 0.59). The replicate block splits that share into flow sampling and system nondeterminism on $\pi_{0.5}$; the other two have no replicate block, so only their combined share is measured (hatched). (b) What each factor moves on the reported $\pi_{0.5}$ rate: inference dials the evaluator sets (dark) and deployment shifts (mid). The seed moves the 500-episode aggregate by only 2.2 points because averaging suppresses the per-episode variance in (a); that is the budget the protocol must buy. Dashed line: the median claimed improvement across the audited comparisons.

I. Introduction

Flow-matching action heads are now the dominant recipe for vision-language-action (VLA) policies, a line advancing rapidly [1], [2], [3], [4], and their inference behavior is governed by three dials every deployment sets implicitly: the number of ODE integration steps, the number of actions executed before replanning, and the random seed feeding the flow sampler. Published evaluations rarely vary or cost these dials, and seed reporting is nearly absent: only 31 of 131 audited highlighted comparisons state an evaluation-seed count. We measure what each one buys, what each one costs, and what each one does to the statistical meaning of a reported success rate (Fig. 1).[1]

Auditing 50 recent VLA papers under frozen intake rules, over a third of their highlighted comparisons carry no usable rollout count; at the audited median of 50 rollouts per task, the minimum detectable effect at LIBERO-typical base rates is 18 points; nearly a third of the recomputable highlighted comparisons fail an exact test at their own rollout counts, and nearly half fall below their own protocol's MDE (Sec. VIII). The dials and the statistics are not separate problems: two dials move the number, and the third alone spans more than a quarter of the field's claimed advances.

Contributions. (1) A variance decomposition for robot policy evaluation: 61–84% of outcome variance survives fixing the initial state across three policies, and a replicate block separates it into flow sampling and system nondeterminism rather than bounding the latter away (Sec. IV); (2) the budget machinery those components imply: an exact rollout calculator and the measured paired-design efficiency gain (Sec. V); (3) a powered 4×3 factorial of denoising steps × execution horizon

[1]Per-episode logs for every reported number, the frozen pre-registration with its deviations log, the audit extraction, and the harness: https://github.com/nihalgunu/roborigor

on *released* checkpoints, showing $\pi_{0.5}$'s shipped default is dominated (Sec. VI); (4) a perturbation study across four LIBERO-Plus axes and three policies, in which one 16.5-point clean gap stands for 9.6 to 67 points of deployed margin and rankings invert only in the floor regime (Sec. VII).

II. Related Work

Inference dials as improvement levers. Recent work treats denoising steps and initial noise as levers for making flow policies faster [5], and real-time chunking [6] redesigns execution scheduling. Both of our first two dials are individually established: extra Euler steps are known to degrade flow policies [7], a two-step policy can match a full sampler [8], execution horizon has an interior optimum [9], and their coupling is already reported [10]. What is missing is a measurement: a balanced factorial, powered and significance-tested, on *released* checkpoints rather than models retrained for the purpose, which is what lets us say a shipped default is dominated rather than that a knob matters. We measure these dials as *evaluation variables*: our single-step result needs no retraining and is policy-conditional, a different claim from making few-step policies *trainable*, and our horizon finding is the benchmark-side complement of RTC, not a contradiction of it.

Evaluation methodology. Deep RL's reproducibility literature is the closest precedent: seed variation alone can flip conclusions [11], and few-run protocols report intervals tighter than their evidence supports [12]; our third dial is the VLA analogue with an architectural rather than training-time noise source. The same program is prescribed for language-model evals [13]; we supply the components that make its power analysis computable for a flow policy. In robotics the strongest existing practice is the LBM study [14], which pairs by initial condition, reports full posterior uncertainty, and states that many robotics papers may be measuring noise; we do not claim to introduce that discipline but supply what it lacks at simulation scale, namely the measured variance components that size a run. Kress-Gazit et al. [15] set out best practices; Jiang et al. [16] audit LIBERO significance and find most SOTA claims unprovable, and we extend that audit with per-comparison MDEs, unit-tagged counts, and non-reporting as a first-class outcome. STEP [17] and its successor [18] give sequential comparison with near-optimal stopping and anytime-valid inference; we add no sequential theory, only the components such tests need to size themselves. PhAIL [19] is a real-robot distributional protocol; our simulation scope is one SIMPLER [20] validates against hardware, and active experiment selection [21] complements our calculator.

Perturbation suites. LIBERO-Plus [22] supplies the battery; we use frozen variant sets over four of its seven axes (camera viewpoint, object layout, lighting, robot initial state). LIBERO-PRO [23] adds semantics-breaking probes whose floor effects are degenerate for rank analysis, so we cite rather than run it.

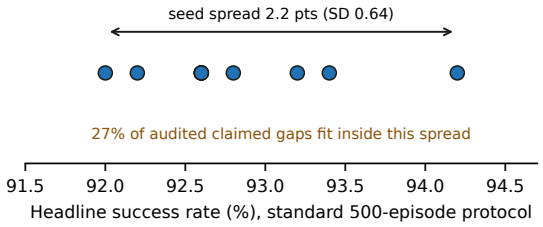

Fig. 2. Ten runs of the standard 500-episode LIBERO-10 protocol differing only in flow-sampling seed (5,000 fresh episodes). The 2.2-point spread contains 27% of the 86 recomputable improvements highlighted across the 50 audited papers.

III. Experimental Protocol

Every reported aggregate is regenerated from per-episode JSONL by released scripts, and a consistency check pins the paper's numerals to those records; no aggregate is carried by hand. **Design.** LIBERO [24] exposes deterministic initial states indexed per task, so all comparisons are paired episode-for-episode and use the exact McNemar test [25]; confirmatory families use Holm [26], exploratory readouts are uncorrected and labeled. Intervals are Clopper–Pearson [27] (Wilson [28] intervals ship alongside; conclusions are interval-choice-invariant), differences use Newcombe's score interval [29] (computed unpaired, which is conservative under the positive correlation that pairing induces), unpaired literature recomputations use Boschloo's exact test [30], falling back to Fisher's [31] for the ten comparisons above n=2000 where Boschloo does not terminate in reasonable time; Fisher is the conservative direction there (deviations log, Entries 1 and 16). **Seed semantics.** A "sampling seed" names a noise *stream*, not a draw: each action chunk receives a seed derived from (episode seed, chunk index), so a replicate reproduces the whole trajectory's noise sequence and consecutive chunks never reuse noise. The residual system nondeterminism this leaves is measured in Sec. IV. **Equivalence margin.** The ±5-point margin used throughout was fixed at the design reframe, anchored to the audit's median per-comparison MDE (4.35 points), an external quantity: it postdates the 4×3 grid itself but predates the distribution-shift (Sec. VI-A), replication, and camera data (deviations log, Entry 5(a)). **Pre-registration.** Analysis choices, audit intake rules, and go/no-go thresholds were frozen before data collection; the released deviations log records every post-freeze change, including this paper's direction after both registered provocations failed.

IV. Where the Variance Lives

A success rate is an average over two sources of randomness that the field reports as one: which initial state the episode drew, and which noise the sampler drew. Separating them is the measurement this paper is built on. Fixing everything but the seed exposes the second source at two scales. At the field's standard protocol (the full LIBERO-10 suite, 50 initial states per task, 5,000 fresh episodes across ten flow-sampling seeds)

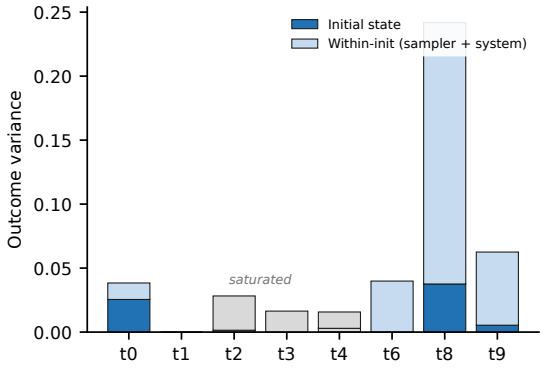

Fig. 3. Outcome-variance components per battery task: initial-state share (dark) vs. the within-init component, sampler and system nondeterminism combined (light). On saturated tasks (grey, observed rate $\geq$ 0.96) both components collapse: their total outcome variance is at most 0.028, below every non-saturated task.

the headline rate spans 2.2 points (SD 0.64, mean pairwise shift 0.71; Fig. 2). That band is not small relative to what the field reports: 27% of the 86 recomputable highlighted improvements are $\leq$ 2.2 points (34% among LIBERO-family comparisons, where the band was measured). The band is measured on one policy and one suite; lower base rates give wider ones (Sec. VI's variance coupling). The band also grows with benchmark informativeness: on a battery weighted toward mid-difficulty tasks (the V1 grid marginalized by seed, 160 episodes per replicate), the same reseeding spans 7.5 points (SD 2.4), the scale of the field's median claimed gap (7.2). Part of that growth is mechanical (the smaller replicate implies $\sqrt{500/160} \approx 1.8\times$ the SD); the measured 3.8× leaves roughly a 2× factor to informativeness. Saturated protocols do not remove seed sensitivity; they hide it in tasks whose outcome no longer varies.

Where the variance lives. With initial states fixed and seeds varied (8 tasks × up to 40 inits × 10 seeds), a nested beta-binomial decomposition attributes outcome variance to the initial state (ICC) versus everything within it, with a method-of-moments cross-check and cluster bootstrap over inits (Fig. 3). Variance-component analysis of benchmark scores is standard [32], and that the sampler can decide an episode is established [33]; what has not been done for robot policies is the partition itself. **Splitting the residual.** What survives fixing the scene is not all sampler. Fixed-request inference is bitwise deterministic on our A100, but closed-loop replicates at fully fixed factors still disagree, and the disagreement sits exactly where the decomposition is measured: 0% on five of the eight battery tasks and 38.9% on the informative one, so the pooled 7.6% understates it there fivefold. We therefore estimate the component from the replicate block rather than bounding it. The three-way split of the informative task's variance is initial state 15.5%, flow sampler 38.5% [12, 65], system nondeterminism 46.0% [19, 73] (the last measured directly at 0.111 [0.046, 0.176]). The sampler is the larger of the two components an evaluator might control, by 2.5×, but the machine is not negligible against it and at the top of its interval the two cannot be ordered. The partition is within task. On the informative task (base rate 0.59), the within-init component is 0.204 of a 0.242 total (84% of outcome variance; init ICC 0.16 [0.06, 0.24]): at a fixed initial state, something other than the scene decides most outcomes. On saturated tasks (observed rate $\geq$ 0.96) every component degenerates: four of the eight battery tasks sit there in our own measurement, and their total outcome variance is at most 0.028 against 0.242 on the informative task, the quantitative form of the observation that the benchmark can no longer separate frontier policies. A 20-init scan of the other three suites found no task below 0.90. **The decomposition replicates across environment seeds.** Rerunning the four qualifying tasks at two further environment seeds (24 inits × 10 sampling seeds each, 1,920 fresh episodes) reproduces the structure: on the informative task the within-init component is 0.200 and 0.214 (vs. 0.204), and the init-dominated task (ICC 0.66, the one task where the scene decides most outcomes) keeps its high init share (0.57 at both new seeds). The replicated set holds one seed-7 saturated task, which drops just below threshold at seed 8 (0.954), so the replication speaks to the informative and init-dominated tasks. The *level*, by contrast, moves: the informative task's base rate shifts from 0.59 to 0.53 and 0.54. Where the variance lives is stable for the policy–benchmark pair; the headline number is not. **And across policies.** The same instrument cell rerun on π_0 and SmolVLA (40 inits × 10 sampling seeds each) gives the same verdict at very different competence levels: with the initial state fixed, within-init noise accounts for 84%, 81%, and 61% of outcome variance for $\pi_{0.5}$ (base rate 0.59), π_0 (0.16), and SmolVLA (0.30) respectively, with the method-of-moments cross-check agreeing in every case. Averaging over sampling seeds is not an optional refinement for some particular policy; it is where most of the per-init variance lives for all three flow policies we test.

What this costs. At the field's default of 50 rollouts per task, the minimum detectable effect at base rate 0.95 is 18.3 points (80% power, exactly enumerated); 500 episodes buy 4.6 points and 1,500 buy 2.5. Our released calculator inverts these exactly for paired and unpaired designs, and the measured components supply the variance inputs a sequential test needs to size itself [17].

V. The Budget Those Components Imply

The components in Sec. IV are the inputs a rollout budget needs: a design dominated by within-init noise is sized differently from one dominated by initial-state spread.

Report the dials. Steps, horizon, and seed policy move the number by up to 41 points; a success rate without them is not a measurement. **Size the run before launching it.** Table I is the cost of honesty: the field's median protocol (50 rollouts) resolves only 18-point gaps, and a 2-point claim costs 2,200 episodes per arm. **Pair everything.** LIBERO's indexed initial states make episode-level pairing free, and the measured relative efficiency over the unpaired design is 1.4–2.3×. **Average over sampling seeds** on informative tasks,

TABLE I

EPISODES PER ARM REQUIRED TO DETECT A GAP AT BASELINE 0.95 (UNPAIRED SCORE TEST; POWER COMPUTED BY EXACT BINOMIAL ENUMERATION; 80% POWER, α=0.05). PAIRING BY INITIAL STATE REDUCES THESE BY THE MEASURED RELATIVE EFFICIENCY BELOW; THE RELEASED CALCULATOR INVERTS BOTH DIRECTIONS EXACTLY.

Gap to detect (points)	Episodes per arm
18	54
10	133
5	424
3	1,048
2	2,200

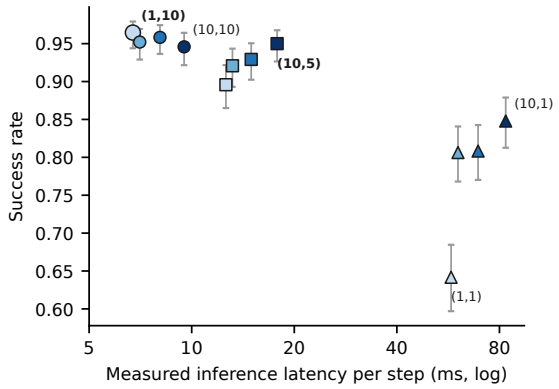

Fig. 4. Success rate vs. measured inference latency for all 12 (*num_steps*, *exec_horizon*) cells (n=480 each, Clopper–Pearson 95% intervals; latency measured server-side from the episode records). For $\pi_{0.5}$ the published default $(10, 5)$ is dominated; the frontier is a single integration step at horizon 10 (Sec. VI-B for what replicates on π_0).

where within-init noise carries 61–84% of per-init outcome variance, and report the seed count (Sec. IV). **Report a margin band, not a clean gap**: a clean gap predicts deployed separation only up to a factor of seven (Sec. VII). **Stop early honestly** with sequential tests [17], [18], sized with the measured components.

The harness (RoboRigor, released at the repository above) runs every policy as a websocket server speaking the openpi client protocol, so a new policy is one serve script of about a hundred lines, and episodes land as append-only JSONL whose factor-complete key gives exactly-once resume and the pairing used throughout. That discipline is not decorative: a wrong hardcoded Clopper–Pearson bound in our interval code survived until the suite cross-validated it against a direct inversion of the exact binomial test.

VI. SPENDING THE BUDGET: STEPS AND HORIZON

We evaluate $\pi_{0.5}$ on a 4×3 grid of denoising steps $s \in \{1, 2, 5, 10\}$ and execution horizon $h \in \{1, 5, 10\}$ over the registered eight-task LIBERO-10 battery (20 paired inits, 3 sampling seeds; n=480 per cell; 5,760 episodes, exactly-once coverage verified against the manifest; Table II). This is a different draw from Sec. IV's arm (up to 40 inits, 10 seeds), so the informative task reads 0.59 there and 0.72 at the default here. The battery is the V1 selection rule (the four tasks then inside $p \in [0.5, 0.95]$ plus the four least-saturated others) applied to a 50-episode-per-task development scan predating this protocol. Our own budget table applies to that step: at 50 episodes per task the scan could not separate 0.95 from 0.98.

The default is dominated. The published default (s=10, h=5) reaches 0.950 at 17.8 ms of amortized inference latency per executed step, measured server-side across the released episodes. The frontier cell (s=1, h=10) reaches 0.965 $[0.944, 0.979]$ at 6.7 ms (Fig. 4); nothing on the grid dominates it, and per-chunk cost falls from 88 to 66 ms. The uncontended cost model predicts a 2.9× speedup, within 10% of the measured one, though its absolute latencies run up to 17% low at h=10 since they omit the per-step overhead worker contention exposes; we report the measured pair as the conservative of the two. Held to this paper's own standard, the right claim is equivalence, not superiority. Cells are paired on (initial state, sampling seed), so a pair shares its noise stream and differs only in the dials; the 29 of 480 pairs that disagree are what the dials themselves move, which is consistent with Sec. IV, where the seed varies and the dials are held fixed. The exact McNemar test does not separate the two cells (18 vs 11 discordant wins, p=0.26), and the difference interval, $[-1.2, +4.1]$ points by Newcombe and $[-1.0, +4.0]$ under an init-cluster bootstrap of the paired difference, establishes equivalence within ±5 points. A tenth of the denoising budget and a 2.7× measured latency reduction buy a statistically indistinguishable success rate. Mechanistically one Euler step executes approximately the conditional-mean chunk, which we do not quantify from the per-cell components for the reason below. Intervals in Fig. 4 are Clopper–Pearson; init-cluster bootstrap intervals are wider in 11 of 12 cells (median +36%, up to +54%; the saturated $(10, 10)$ cell is 7% narrower) and change no conclusion.

What the certificate rests on. The ±5 verdict is a property of this battery, and this battery is near ceiling: at the default seven of eight tasks exceed 0.96 and two sit at 1.000. Those two contribute 120 concordant pairs and no discordant ones, tightening the interval through n alone. Restricted post hoc to the six non-ceiling tasks the interval widens to $[-1.1, +5.6]$ and no longer certifies ±5; on the one task with more than five points of headroom it is $[-10, +23]$ at n=60. The direction is stable across all three subsets (the frontier leads by 1.5, 1.9, 6.7 points); the certificate is not. Equivalence at ±5 holds on the battery as registered; establishing it where headroom exists needs a benchmark that has some, the saturation of Sec. IV returning as a bound on what any comparison here can prove. Both restrictions are exploratory (deviations log, Entry 13).

The coupling, quantified. At h=1 success climbs from 0.642 to 0.848 as s goes $1 \rightarrow 10$; at h=10 the ordering inverts and s=1 is nominally best. Replanning cadence dominates denoising budget throughout, every h=1 cell losing 10–32 points to its h=10 counterpart. That the dials couple is known [10]; what the factorial adds is its size on released weights. We decline to read a variance mechanism off these cells: the beta-binomial within-init component is $\mu(1-\mu)(1-\mathrm{ICC})$, so across cells running 0.642 to 0.965 it largely tracks the base rate.

TABLE II

$\pi_{0.5}$ SUCCESS ON THE REGISTERED BATTERY FOR ALL (s, h) CELLS (n=480 EACH; CLOPPER–PEARSON 95% INTERVALS). THE PUBLISHED DEFAULT IS $(10, 5)$; THE FRONTIER IS $(1, 10)$.

Steps s	h=1	h=5	h=10
1	.642 [.597, .685]	.896 [.865, .922]	.965 [.944, .979]
2	.806 [.768, .841]	.921 [.893, .943]	.952 [.929, .969]
5	.808 [.770, .843]	.929 [.902, .951]	.958 [.936, .974]
10	.848 [.813, .879]	.950 [.926, .968]	.946 [.922, .964]

The base-rate-free share stays within 0.45–0.55 across the h=1 column with no monotone trend in s.

Scope. These results are conditional on LIBERO's quasi-static manipulation: nothing demands mid-chunk reactivity, and where that is false real-time chunking [6] addresses the tension our frontier sidesteps.

A. Does the cheap frontier survive distribution shift?

The natural objection to a mean-chunk shortcut is that its slack appears only off-distribution. We re-ran $s \in \{1, 2, 10\}$ at h=10 on identical LIBERO-Plus variant sets (688 variants, one paired episode per variant per cell). Under object-layout perturbation the cheap cells hold: 0.920 vs 0.913 for s=1 vs s=10 (difference CI $[-3.8, +5.1]$ points by the same unpaired Newcombe estimator used throughout, non-inferior at the 5-point margin, though the upper bound precludes formal equivalence). Under camera-viewpoint perturbation the cost is real, by the strongest evidence in this paper: a pre-declared confirmatory sample (all 419 camera variants, two draws per variant at each step count, 838 pairs, single registered analysis; Entries 7 and 9) measures 0.4057 vs 0.4523, a 4.7-point deficit, discordants 45–84, exact McNemar p=0.0008 (unpaired Newcombe interval $[-9.4, +0.1]$ points, conservative under pairing). The exploratory draws that motivated it (632 pairs, pooled p=0.017 raw and p=0.033 two-look corrected, the value Entry 5(b) registers) carried this paper's caution in miniature: a screen pooling camera with layout read an ambiguous p=0.076 where the camera-only paired test was already p=0.015. Wrong aggregation, like wrong sample size, manufactures apparent equivalence. A second step does not recover the cost: on the exploratory sets s=2 scores 0.468 against 0.505 under camera (376 variants, a different sample from the confirmatory one, hence the different levels) and 0.907 against 0.913 under layout (312), so it is a many-steps effect. The frontier is not separated from the default on-distribution, non-inferior under layout shift, and carries a confirmed 4.7-point viewpoint cost.

B. Which parts replicate on a second policy?

Rerunning the key cells on π_0 (same battery, paired inits, n=480) separates the universal from the policy-specific. Horizon replicates emphatically: replanning every step costs π_0 40.8 points at s=1 and 18.1 at s=10. Its single-step cell is again not separated from the default (0.773 vs 0.787, p=0.56; CI $[-6.7, +3.8]$, too wide to establish equivalence), but unlike $\pi_{0.5}$ it pays for dropping steps at h=10: $(10, 10)$ beats $(1, 10)$ by 3.0–12.8 points (p=0.00025). The scoped conclusion: horizon dominates denoising budget for both policies, while single-step sufficiency is a property of the policy, not the benchmark. At a second environment seed $\pi_{0.5}$'s single-step cell reproduces to within one episode in 480 (0.9625 vs 0.9646) and never trails the default there (paired difference $[+0.1, +5.9]$, p=0.02). A third policy widens the pattern: SmolVLA [3], served through the same harness, shows the same horizon ordering (h=10 0.648 vs. h=5 0.596, discordants 63–38, p=0.017) and drops to 0.554 at h=1, 9.4 points below (CI $[3.2, 15.5]$, $p<10^{-4}$). That cell carries its own caution: an interim look truncated at 26 episodes, all from the battery's hardest task, read 2/26 and suggested outright collapse, which the pre-declared completion to n=480 (deviations log, Entries 4(c), 6, 8, 10) shows to be a task-composition artifact. Partial looks at non-random subsets manufacture effects. Its single-step cell is not separated from its default (0.625 vs. 0.648, paired difference −2.3 points $[-8.3, +3.8]$, p=0.29), an inconclusive result at our ±5-point margin rather than established equivalence. Horizon dominance is thus three-for-three across two model families while its magnitude is policy-dependent (10–32 points for $\pi_{0.5}$, 18–41 for π_0, 9 for SmolVLA), and single-step sufficiency is established for $\pi_{0.5}$, refuted for π_0, and inconclusive for SmolVLA.

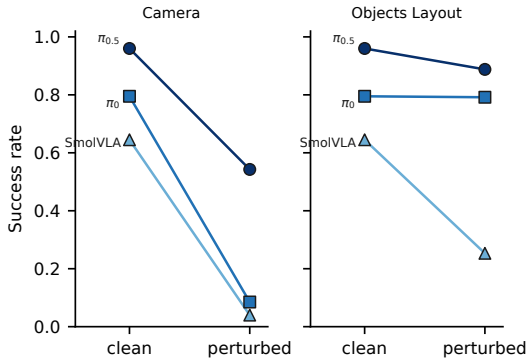

Fig. 5. Clean vs. perturbed success for three policies on identical LIBERO-Plus variant sets. Lines never cross (ordering stable) while separations change drastically.

VII. RANKINGS SURVIVE OFF THE FLOOR; MARGINS DO NOT

No inversions on the registered axes. Across three policies ($\pi_{0.5}$, π_0, SmolVLA) evaluated on an identical set of 688 LIBERO-Plus variants (Fig. 5), not one pairwise ordering inverted at the axis level (Kendall τ=1.0; a null simulation redrawing observed rates under binomial noise yields p=1.0; the single nominal level-scan inversion has p=0.26). Clean anchors are our own paired 200-episode runs ($\pi_{0.5}$ 0.960, π_0 0.795, SmolVLA 0.645), not published numbers. Anchors and perturbed runs alike are single draws at an uncontrolled sampling seed: the $\pi_{0.5}$ anchor across ten seeds spans 92.5 to 96.0 (mean 93.6, SD 1.1) and sits at that band's top, putting the clean gap at 14.1 on the seed mean rather than 16.5. The two-regime conclusion turns on tens of points and is unaffected.

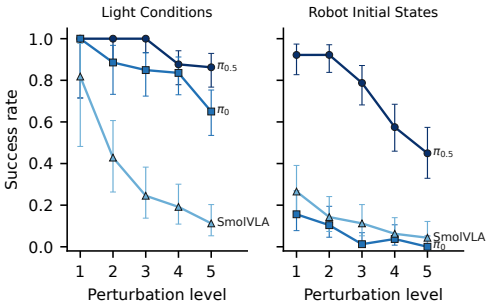

Fig. 6. Success by perturbation level on two further axes (frozen variant sets, identical across all three policies). Lighting degrades without reordering; robot-initial-state shift floors π_0, where the clean π_0–SmolVLA ordering inverts (p=0.0022).

Meanwhile the *gaps* moved wildly in both directions: the 16.5-point $\pi_{0.5}$–π_0 gap nearly tripled to 45.7 under camera and fell to 9.6 under layout; the 15.0-point π_0–SmolVLA gap collapsed to 4.5 under camera and tripled to 53.8 under layout. Two caveats: this analysis covers two of the seven axes, and with separations this large the design has little power to detect inversions between closely matched policies, so the finding is the magnitude instability, not the absence of flips. Ordering is more trustworthy than our premise assumed; a clean gap still says almost nothing about the margin behind it.

Two further axes find the regime where ordering fails. On frozen variant sets for two additional axes run on all three policies (252 lighting and 370 robot-state variants, Fig. 6), the axes behave qualitatively differently. Under Light Conditions every clean ordering is preserved ($\pi_{0.5}$ 0.921, π_0 0.794, SmolVLA 0.238; π_0–SmolVLA discordants 146–6). Under Robot Initial States the top of the order holds: $\pi_{0.5}$ at 0.730 degrades from 0.92 to 0.45 while π_0 is floored at every level (0.059, discordants 250–2, a 67.0-point gap). At the floor, however, the clean ordering breaks: SmolVLA (0.122) significantly outscores π_0 (38–15 discordants, exact p=0.0022, unchanged under Holm across the six pairwise tests of this section; exploratory, as this axis postdates the freeze), inverting a 15-point clean-benchmark gap. The collapse is not a serving artifact: through the same fork wrapper, server process, and session π_0 scores 0.794 under Light Conditions, matching its clean 0.795, so what remains is its own sensitivity to shifted initial joint states. The pattern is a two-regime rule. While every policy retains headroom, perturbation rescales gaps without reordering (the same 16.5-point clean gap stands for anything from 9.6 to 67.0 points), the robotics form of accuracy-on-the-line [34] with the margin moving and not the ordering; once an axis floors the competitors, even the ordering carries no information.

VIII. The Pre-Registered Gates and the Audit

This paper began as a provocation with two pre-registered go conditions, and both failed. A pre-registration that only survives success is theater. The first required rank inversions under perturbation at the registered scale; on the two registered

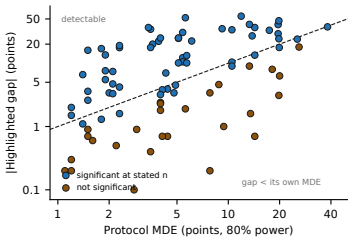

Fig. 7. Each audited highlighted comparison: its claimed gap vs. the minimum detectable effect of its own protocol. Points below the diagonal claim less than their protocol could detect.

axes the ordering held everywhere while pairwise gaps ranged from 4.5 to 63.5 points around clean gaps of 15.0 to 31.5, and the gate failed. A post-gate extension to two further axes (Entry 4(b)) found one significant inversion in the floor regime, outside the frozen scope, so we report it as exploratory. The second required at least one third of audited comparisons to fail exact tests at their stated rollout counts, or the audited median MDE to exceed the median claimed gap. The statistic logged at gate time was the uncorrected failure rate: it fell short at 31.4%, and the median MDE sits below the median gap, so both disjuncts failed. The registered secondary reading (Holm within paper families, 33.7%) would cross; we report it but do not adopt it retroactively as the gate statistic (deviations log, Entry 12).

The audit. Across 50 papers sampled under frozen intake rules, 34.4% of highlighted comparisons carry no usable rollout count (45/131), so the claim cannot be evaluated in principle; 13 of 50 papers [15%, 40%] report none at all, and eval seed counts appear for 31 of 131. Among recomputable comparisons, 31.4% fail an exact test at the papers' own rollout counts (33.7% under Holm), the median claimed gap (7.2 points) exceeds the median per-comparison MDE (4.35), and 47.7% of gaps (41/86) fall below their own protocol's MDE (Fig. 7). That failure rate is the same genus as the LIBERO audit of Jiang et al. [16] under different intake rules; we report it as concordance, not a finding.

IX. Limitations

All results are simulation-only and LIBERO-family; dials one and two are specific to flow-matching heads, since policies that decode actions directly [35], [36] have no denoising budget, though the seed and margin analyses apply unchanged. Latency is inference time on one A100, excluding simulator cost. SmolVLA's per-task success spans 0.25 to 0.98, so suite-pooled numbers hide task-level differences. The findings differ in reach: the variance partition should appear wherever a flow policy is scored at small n, whereas the horizon result rests on open-loop-tolerant scenes and the frontier's robustness price is measured on $\pi_{0.5}$ alone. The audit was single-rater against pre-frozen rules; the extraction ships so it can be rechecked.

X. Conclusion

Most of what moves a reported success rate is invisible at a fixed initial state. Measuring it is what sizes a rollout budget.

References

[1] K. Black *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv:2410.24164*, 2024.

[2] Physical Intelligence, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," *arXiv:2504.16054*, 2025.

[3] M. Shukor *et al.*, "SmolVLA: A vision-language-action model for affordable and efficient robotics," *arXiv:2506.01844*, 2025.

[4] Physical Intelligence, "$\pi_{0.7}$: A steerable model with emergent capabilities," Whitepaper, https://www.pi.website/blog/pi07, 2026.

[5] W. Luan, J. Li, W. Zhao, W. Zhang, T. Wu, and R. Ma, "SnapFlow: One-step action generation for flow-matching VLAs via progressive self-distillation," *arXiv:2604.05656*, 2026.

[6] K. Black, M. Galliker, and S. Levine, "Real-time execution of action chunking flow policies," *arXiv:2506.07339*, 2025.

[7] Z. Chen, Z. Guo, P. Wang, T. I. Egbe, Y. Lyu, and C. Qian, "Dense-jump flow matching with non-uniform time scheduling for robotic policies: Mitigating multi-step inference degradation," *arXiv:2509.13574*, 2025.

[8] C. Pan *et al.*, "Much ado about noising: Dispelling the myths of generative robotic control," *International Conference on Learning Representations (ICLR)*, 2026.

[9] H. Wang *et al.*, "VLA knows its limits: Adaptive execution horizons for robot policies," *arXiv:2602.21445*, 2026.

[10] Y. Chen *et al.*, "Let it be simple: One-step action generation for vision-language-action models," *arXiv:2606.05737*, 2026.

[11] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *AAAI Conference on Artificial Intelligence*, 2018.

[12] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[13] E. Miller, "Adding error bars to evals: A statistical approach to language model evaluations," *arXiv:2411.00640*, 2024.

[14] TRI LBM Team, "A careful examination of large behavior models for multitask dexterous manipulation," *Science Robotics*, 2026.

[15] H. Kress-Gazit *et al.*, "Robot learning as an empirical science: Best practices for policy evaluation," *arXiv:2409.09491*, 2024.

[16] T. Jiang, X. Tan, S. Wheeler, L. Sun, T. W. Ayalew, and M. Walter, "What are we actually benchmarking in robot manipulation?" *arXiv:2606.04233*, 2026.

[17] D. Snyder, A. J. Hancock, A. Badithela, E. Dixon, P. Miller, R. A. Ambrus, A. Majumdar, M. Itkina, and H. Nishimura, "Is your imitation learning policy better than mine? policy comparison with near-optimal stopping," *arXiv:2503.10966*, 2025.

[18] D. Snyder, A. Badithela, N. Matni, G. Pappas, A. Majumdar, M. Itkina, and H. Nishimura, "Beyond binary success: Sample-efficient and statistically rigorous robot policy comparison," *arXiv:2603.13616*, 2026.

[19] S. Arkhangelskiy, "PhAIL: A real-robot VLA benchmark and distributional methodology," *arXiv:2605.29710*, 2026.

[20] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating real-world robot manipulation policies in simulation," in *Conference on Robot Learning (CoRL)*, 2024.

[21] A. Anwar, R. Gupta, Z. Merchant, S. Ghosh, W. Neiswanger, and J. Thomason, "Efficient evaluation of multi-task robot policies with active experiment selection," *arXiv:2502.09829*, 2025.

[22] S. Fei *et al.*, "LIBERO-Plus: In-depth robustness analysis of vision-language-action models," *arXiv:2510.13626*, 2025.

[23] X. Zhou *et al.*, "LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models," *arXiv:2510.03827*, 2025.

[24] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "LIBERO: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.

[25] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.

[26] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979.

[27] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[28] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *JASA*, vol. 22, pp. 209–212, 1927.

[29] R. G. Newcombe, "Interval estimation for the difference between independent proportions: Comparison of eleven methods," *Statistics in Medicine*, vol. 17, pp. 873–890, 1998.

[30] R. D. Boschloo, "Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities," *Statistica Neerlandica*, vol. 24, pp. 1–9, 1970.

[31] R. A. Fisher, *The Design of Experiments*. Oliver and Boyd, 1935.

[32] X. Bouthillier *et al.*, "Accounting for variance in machine learning benchmarks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

[33] O. Patil *et al.*, "You've got a golden ticket: Improving generative robot policies with a single noise vector," *arXiv:2603.15757*, 2026.

[34] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," in *International Conference on Machine Learning (ICML)*, 2021.

[35] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," *Conference on Robot Learning (CoRL)*, 2024.

[36] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv:2502.19645*, 2025.