PER-TASK RETRAINING NOISE IN ROBOT IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Imitation learning trains robot policies from human demonstrations that a person collects one episode at a time by teleoperating the robot. Because gathering them is slow and expensive, curation methods try to make that effort go further by scoring each demonstration and retraining the policy on only the highest-scoring ones. Across the six curation papers we audit, reported success-rate improvements run from 5 to 32 percentage points in simulation. Yet retraining a policy on unchanged data also moves its success rate, and none of these improvements was measured against that retraining noise. We train 701 policies on controlled demonstration subsets, spanning diffusion policy and ACT from scratch and the vision-language-action models SmolVLA and $\pi_{0.5}$ fine-tuned. Each is evaluated on PushT, two bimanual ALOHA tasks and two ten-task LIBERO suites, every batch pre-registered. Retraining on identical data shifts one LIBERO task's success rate by a standard deviation of 20.4 percentage points and the ten-task average by 4.23, comparable to the reported gains. This per-task noise recurs in Meta-World, where one retrain of a small state-based policy solves a drawer-opening task 200 times out of 200 and another solves it 0 times. No score we built for a demonstration predicts whether keeping it improves the policy, so the safeguard is to measure the noise itself. Part of it comes from limited evaluation episodes and is captured by evaluating one policy many times. The rest comes from training itself and appears only when training is repeated, so eight retrainings of one fixed demonstration set are enough to capture it. This measurement tells a lab whether a new policy is a real improvement or a lucky retrain. We propose that every curation gain be quoted beside the retraining noise of its own setting, and we release the run records and pre-registrations that make that comparison possible.

1 INTRODUCTION

The field has built much machinery to choose which demonstrations to keep, and far less to tell whether a reported gain exceeds noise. Per-episode scoring and curation methods now report gains of 5–32 pp in simulation and up to $+88$ pp on hardware (Hejna et al., 2025; Chen et al., 2025; Agia et al., 2025; Kulkarni et al., 2026; Zhang et al., 2026; Hejna et al., 2024). Figure 1 sets each against the smallest gain its seed count could resolve (§2 reviews the methods); throughout, a gain inside its noise marker is *unverified, not refuted*, a claim about the reported evidence, not about whether the method works.

Retraining noise differs between settings that look alike, so it must be measured, not assumed. Three rarely tested assumptions underlie curation methods: that an episode has an intrinsic value a cheap score can estimate, that a measured gain exceeds retraining noise alone, and that a score under one policy class holds under another. This paper tests all three by intervention, retraining on controlled episode sets and counting closed-loop successes rather than correlating scores with outcomes, and states each answer's scope. A fourth measurement puts the answers in proportion: variance from *which* episodes are in the training set shrinks fast as the set grows, while retraining noise does not (§4.3).

This paper makes six contributions.

1. **Retraining noise is large and setting-specific.** Retraining on identical data moves success rate by an sd σ_{run} running from at or below the evaluation's sampling floor (finite-rollout binomial

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.

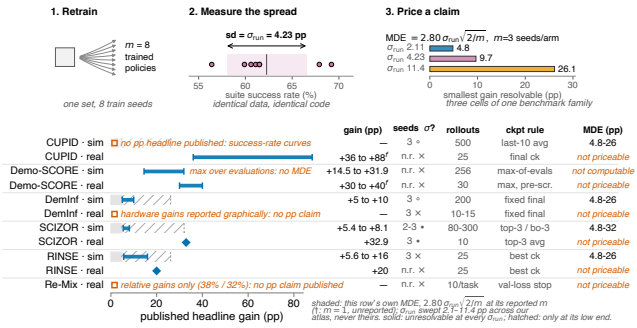

Figure 1: **The ruler (a), and the literature measured with it (b). (a)** Retraining one byte-identical episode set m times gives one cell's retrain lottery σ_{run} (here $\pi_{0.5}$-ft, `libero_object`, k=88), fixing the smallest gain the design can resolve (minimum detectable effect, MDE); across three of our cells that MDE runs 4.8 to 26.1 pp. **(b)** The same computation at each audited paper's *own* reported seeds, sim and hardware legs separate. Bars are each source's headline gain (^{f}: off a figure); ×/○/● mark σ_{run} as not/partly/weakly recoverable from that source, and shading uses *our* σ_{run} range under the lenient conventions of Appendix G; hardware legs get no marker, as our floors are simulation-only. **A gain inside its marker is unverified, not refuted** (ledger: Table A8).

noise) to 6.515 pp on bimanual ALOHA (Zhao et al., 2023) and 11.43 pp on small-data LIBERO. No scale-free transform we tested collapses that range, so a floor measured in one *cell* (one policy class, task suite and training-set size) cannot be assumed in another. The *atlas* of measured cells (Table 1) prices each of them, where to *price* a claim is to state the replication it requires; an unmeasured cell cannot be priced from it (§4).

2. **Per-task noise exceeds what the suite average shows and depends on the suite.** Net of the sampling floor, the sd of per-task success over retrains, σ_{pertask}, runs 3.29–11.0 pp across three vision–language–action (VLA) cells whose suite-level σ_{run} on the same runs is 2.04–4.23 pp. Pre-registered follow-ups rule out the truncated learning-rate schedule as the cause in the one cell tested and replicate the lottery on Meta-World, where choosing the retrain per task from a pool of eight retrains beats the best single retrain by 4–12 pp (§4.1, Appendix L).
3. **What makes an episode set valuable does not transfer across policy classes.** On identical episode sets, per-set success under diffusion policy and fine-tuned SmolVLA (Shukor et al., 2025) correlates at only $r \approx 0.29$. No selection principle transfers across classes once deconfounded, and at fixed budget even the *sign* of adding more data is suite-local (Liu et al., 2023). Exactly one rule survives, *within* one cell: coverage-maximizing (widest state-space spread) selection on PushT (Florence et al., 2021) (§5, §5.2).
4. **No per-episode score we could build predicts causal value.** Twelve estimators from three families, plus a thirteenth from the published method class, score $\rho \approx 0$ against causal ground truth, retrain-measured episode effects; all are our own implementations, and we re-ran no published curation code (§5.1, §2).
5. **Set value is not a sum of per-episode contributions.** A pre-registered paired design of 86 runs cut between-set variance by only 1.92× on PushT with diffusion policy (Chi et al., 2023) and 1.73× under fine-tuned SmolVLA, against pre-registered additive predictions of 11.5× and 4.9×. On the fine-tuned cell this rejects, by intervention, the additive premise beneath per-episode scoring, which explains why the $\rho \approx 0$ nulls are not merely an estimator problem. On PushT the rejection was withdrawn once the harness's own nondeterminism, run-to-run disagreement at fixed data and seed, was measured (§4.2).
6. **A released corpus and a fully pre-registered record.** The corpus is 701 trained policies with episode masks, configurations and, for 334 runs, per-episode success vectors, released with every pre-registration, verdict criteria and analysis kit. Every verdict was scored as registered, and the twelve contradicting our predictions are reported as such, among them the overturn of our own pretraining-scale claim (§6, Appendices D, H).

2 Related Work

Demonstration curation and scoring is the class under test. The class shares one recipe: score, filter, retrain, report the gain. **CUPID** (Agia et al., 2025), closest in goal, uses influence func-

Table 1: **The noise atlas: the cells this paper argues from, each against the sampling floor of its own evaluation.** Floors are binomial, $f = 100\sqrt{p(1-p)/n_{\text{eval}}}$ at the row's mean SR; floor share is $f^2/\sigma^2_{\text{run}}$. **Bold**: the readings the main text argues from. †At or below its own binomial floor, 1.292 against 1.682 and 3.31 against 3.530: unresolvable above sampling noise. Each is a measurement of its own cell, and treating the bound as a noise estimate *understates* the lottery. [L]LOW-SR: σ compressed near SR=0, never a class property. [H]Parenthesized: the *same* eight checkpoints at the pre-committed healthy horizon h=8, not an independent cell, ACT/PushT's one licensed figure. Intervals and superseded readings: Table A10, Appendices H, K.

cell (policy class / task / k)	mean SR	n_{eval}	σ_{run} raw	floor-sub.	floor share	replication (df)
ACT / PushT, chunk-100 open loop[L] (h=8)[H]	1.35 (16.35)%	500	0.784 (**4.849**)	0.590 (4.559)	43 (12)%	8 runs (7)
ACT / ALOHA `insertion`[L]	11.98%	500	2.261	1.734	41%	8 runs (7)
ACT / ALOHA `transfer_cube`	77.88%	500	**6.515**	6.245	8%	8 runs (7)
DP / PushT k=103, rebuilt stack†	17.05%	500	**1.292**	—	>100%	4 pairs (4)
DP / LIBERO-object k=22, single-task	66.56%	200	**11.43**	10.93	8.5%	8 pairs (7)
SmolVLA-ft / LIBERO-object k=22, single-task†	47.10%	200	≤**3.31**	—	>100%	6 pairs (5)
SmolVLA-ft / LIBERO-object multi-task, k=88	54.94%	2000	2.113	1.80	28%	8 runs (7)
$\pi_{0.5}$-ft / LIBERO-object multi-task, k=22	35.75%	200	6.02	4.97	32%	8 runs (7)
$\pi_{0.5}$-ft / LIBERO-object k=22, single-task	77.81%	200	**9.74**	9.29	9.1%	16 runs (10)
$\pi_{0.5}$-ft / LIBERO-object multi-task, k=88	62.26%	1000	4.228	3.94	13%	8 runs (7)

tions for per-demonstration causal effect on closed-loop return. **DemInf** (Hejna et al., 2025) ranks episodes by mutual information on RoboMimic (Mandlekar et al., 2021), and **Demo-SCORE** (Chen et al., 2025) classifies them by rollout success. **SCIZOR** (Zhang et al., 2026) assigns episodes self-supervised scores, **RINSE** (Kulkarni et al., 2026) ranks them by smoothness, and **Re-Mix** (Hejna et al., 2024) reweights whole data mixtures. None reports enough retrains to place its gain against the measured noise floor of its own regime, from-scratch or fine-tuned: at most three seeds ($\text{df} = 2$), too few to bound σ_{run} anywhere we measure. We do not re-run them but audit what they report (Figure 1, Appendix G) and interventionally test one representative (§5.1).

Data quality and attribution. Theory predicts the regime dependence we find: action divergence and transition diversity jointly bound compounding-error regret, and raw state diversity can *hurt* (Belkhale et al., 2023). Our causal ground truth is the datamodel of Ilyas et al. (2022): per-episode value fit to random-subset retrains. The attribution literature it anchors (Park et al., 2023; Koh & Liang, 2017; Dass et al., 2025; Ghorbani & Zou, 2019) targets held-out loss, cheap to compute and low in variance. Closed-loop success is neither: at $N = 200$ subsets its reliability of 0.29 caps any scorer's correlation at 0.54. Our constructive rule, textbook greedy facility-location maximization covering the pool's states (Wei et al., 2015; Mirzasoleiman et al., 2020), is not novel: we validate it causally and measure where validation stops. Appendix G covers observational metrics (Sakr et al., 2025; Sirigiri et al., 2026) and influence-function fragility (Basu et al., 2021; Bae et al., 2022); §5.2 the validated rule.

Seeds, noise, and statistical practice. Robot imitation learning lacks the noise accounting deep RL was forced into (Henderson et al., 2018; Agarwal et al., 2021) and supervised benchmarking has begun to price (Dodge et al., 2019; Picard, 2021). Bouthillier et al. (2021) decomposed variance by source, benchmark- and task-dependently. No cited work has our three additions: a floor *not transportable across cells of one benchmark*; a per-task σ larger than the suite σ averaging it; and an *interventional* rejection of the additive per-episode-value premise. Rigor has gone to evaluation rather than curation (Kress-Gazit et al., 2024; Snyder et al., 2025; TRI LBM Team, 2025), none of it measuring the cell-specific retraining noise floor curation claims face. LIBERO-Plus (Fei et al., 2025) makes the complementary *evaluation*-side point: ∼78 pp of spread for one model under camera perturbation, which, together with our training-side results, implicates the harness on both sides (§3, §4.1). Selection on noise is not robotics-specific (Szymanski et al., 2023; Shao et al., 2025).

3 SETUP: CELLS, ESTIMANDS, AND PROTOCOL

Notation. A *cell*, the unit every claim in this paper is made and priced in, is one policy class × task suite × training-set size k; single- and multi-task evaluation of a suite are distinct cells, and a LIBERO suite's ten tasks are learned jointly by one policy, which is what makes per-task and suite-average readings distinct quantities. k counts the *episodes* (training demonstrations) in a subset, and a *rollout* is one of the n_{eval} evaluation episodes per run. DP is diffusion policy and ACT the action-chunking transformer (Zhao et al., 2023), both from scratch; a VLA fine-tune (-ft) is SmolVLA or $\pi_{0.5}$. A *wave* is a pre-registered batch of runs, an *arm* is one condition in it, a *mask* is the episode subset a run trains on, and the *ladder* orders cells by pretraining scale: DP, then

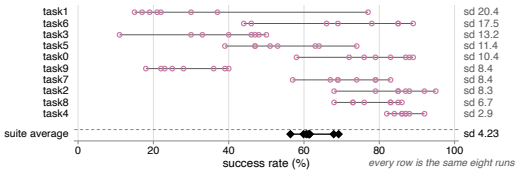

Figure 2: **Averaging over a suite hides the largest noise term we measure.** Eight retrainings of $\pi_{0.5}$-ft on byte-identical sets (`libero_object`, k=88, 100 episodes/task/retraining, truncated LR schedule), rows sorted by descending sd, bottom row the ten-task average *of those same eight runs*: $\sigma_{\mathrm{pertask}} = 11.00$ floor-subtracted / 11.82 raw against a suite average moving 4.23 pp. The top row reads sd 20.4 pp, a maximum over ten that one retrain carries: with that run deleted the task falls to 7.81, a different task becomes the worst at 17.8, and $\sigma_{\mathrm{pertask}}$ moves only to 10.51 (§4.1).

SmolVLA-ft, then $\pi_{0.5}$-ft. A reading is *licensed* when its wave's pre-registered scope covers its use. An all-caps *verdict token* such as INDETERMINATE names the numeric band, fixed before the data existed, that a hypothesis was scored into; each token is defined in the appendix its paragraph cites.

Estimands. Repeatedly retraining identical subsets under one evaluation protocol gives σ_{run}, a standard deviation (sd) in points of success rate SR; we call that distribution the cell's *retrain lottery*, of which a single retrain is one draw. It is $\mathrm{sd}(\Delta\mathrm{SR})/\sqrt{2}$ from train-seed pairs, or the plain sd across retrains; Table 1's replication column reads *pairs* or *runs* accordingly. σ_{set} is dispersion across episode *sets* at fixed k, and $\sigma_{\mathrm{pertask}}$ the sd across retrains of *per-task* SR. σ_0 is the harness's own nondeterminism, identical episodes and seed run twice, a floor no design cancels; $\hat{w}$ is the share of between-set variance that shared episodes do *not* cancel. The minimum detectable effect (MDE), the smallest true difference m retrainings per arm resolve at 80% power with $\alpha = 0.05$, is $\mathrm{MDE} = 2.80\,\sigma_{\mathrm{run}}\sqrt{2/m}$ (Figure 1a). Intervals are 95% χ^2 at the stated df, jackknife t_7 for $\sigma_{\mathrm{pertask}}$. Every σ is given raw and floor-subtracted, $\sqrt{\sigma^2 - f^2}$ with f Table 1's binomial floor.

Protocol. Training-step budgets are fixed per cell and evaluation uses the final checkpoint without best-checkpoint selection, whose bias is measured in §4. A shared evaluation seed gives common random numbers *within* a run but does not align start states across runs on LIBERO, so it waives no sampling floor (§4.2). Arms are size-matched, inference stays within a wave, and absolute success rates are never compared across software stacks. One provenance stamp covers every vision–language–action (VLA) value: all SmolVLA and $\pi_{0.5}$ fine-tunes use the ecosystem-default *truncated* LR schedule, 20k of a 30k-step cosine decay. Completing it measured null in the one cell tested (§4.1). **Four harness hazards.** First, LeRobot, the training and evaluation stack, counts evaluation episodes **per task**, not in total. Second, the schedule above ships truncated by default. Third, LIBERO's init-state pointer advances with episode outcome, so start states depend on the policy's own behavior, and Meta-World's per-reset seed selects no start state (Appendix L). Fourth, our own driver defect silently cut runs short of budget (Appendix C, §7).

4 Retraining noise is a measured, cell-specific quantity

The lottery falls with data but does not track pretraining scale. Under from-scratch DP, σ_{run} falls from 11.43 to 6.24 pp as k goes from 22 to 44. Under SmolVLA fine-tuning at identical k it falls to *at most* 3.31 pp, at or below the evaluation's own floor: unresolvable above sampling noise, and not read as zero. The variance budget flips with it: retraining noise is $\sim$75% of between-subset variance under DP (CI 37–100%) but $\sim$5% under fine-tuning, where composition (the episodes a set holds) dominates. Under $\pi_{0.5}$ fine-tuning, a pre-registered 16-run seed expansion overturned a preliminary reading of 2.02 pp at $\mathrm{df} = 2$, giving 9.74 pp, or 5.23 with one collapsed mask excluded. The frozen rule attributes that collapse to the training seed, not the subset, so it belongs to the retrain level: a single retrain can land $\sim$28 pp below the cell mean, invisible without a reference retraining. The ladder, 11.43 under DP, at most 3.31 under SmolVLA-ft, 9.74 under $\pi_{0.5}$-ft, is thus not monotone, and a floor-limited middle rung only sharpens that. That ladder compares variance structure only and is not sample-matched, and task breadth and then k take 9.74 back down 2.30$\times$ in an exploratory, not pre-registered decomposition (Appendices E, J).

σ_{run} is measurable within a cell, not transportable across cells. On the rebuilt PushT stack (§7) the pooled cell reads *below* its own binomial floor, superseding an earlier reading: a measurement,

not a prior. No scale-free transform we tested collapses the span of Table 1 under the floor-subtracted convention, and that span is not estimation noise in $\hat{\sigma}$: across the eight cells whose σ_{run} is measured rather than bounded, a common-σ_{run} null is rejected (Bartlett on raw σ_{run}, $p = 8.1\times10^{-5}$). We therefore quote no single cross-cell multiple; a pre-registered gate licenses a ratio only for cell pairs whose floor shares and evaluation scopes are close enough to compare. *That refusal is the result.* A third policy class does not restore a common law: the two ALOHA ACT rows are unpaired under the frozen rule, their $2.9\times$ difference confounded with success-rate regime within the wave, and the ACT/PushT row, at LeRobot's default open-loop horizon (the actions taken before re-observing), *understates* its own lottery $\sim6.2\times$ raw, so only its h=8 re-evaluation is in scope. A pre-registered horizon-artifact test required agreement at both horizons and found it at one, **INDETERMINATE**, so the row's LOW-SR stamp stands (Appendices K.6, K.4–K.5).

Two questions and two noise distributions. Comparing two fixed episode sets turns on σ_{run} alone, through the MDE of §3. Comparing two selection methods turns instead on set-identity dispersion, the spread between equal-size sets, which replicating seeds cannot reduce. In our replicate cells that spread is not a Gaussian σ_{set} at small k but a collapse *probability* and magnitude, so $\sqrt{\sigma^2_{\text{set}} + \sigma^2_{\text{run}}}$ propagation is invalid and we price method budgets under a two-component mixture. The audited designs distinguish neither question (Tables A1–A3, Appendices A, G).

A prospective replication corrected our own noise model in both directions. A replication registered in advance, in a cell where we had never trained (k=22, single-task), measured $\sigma_{\text{run}} = 3.51$ pp against our fit's 7.63 and $\sigma_{\text{set}} = 19.01$ pp against its 11.69: set identity outweighs run noise $5.4\times$ there. We **withdraw** the method-comparison rows, which propagated that σ_{set} as Gaussian; the restated, k-scoped prices are in Appendix F, and no audited evaluation meets even the cheaper k=88 price (Figure A2, Appendix G).

A gain reported without σ_{run} in its regime is uninterpretable. Of the six audited curation papers, exactly one prints per-condition seed spread, and one reports the *maximum* success over evaluations, a rule that on the $\pi_{0.5}$ cell's first 8 runs turns checkpoint noise into a $+4.00$ pp bias whose held-out value is indistinguishable from zero and, being within-run, adds to the retrain lottery instead of bounding it; Figure 1 draws the corollary (§5.1; Appendices G, J).

4.1 The lottery is task-level, and per-task noise is suite-specific

Figure 2 shows what the field's reporting unit, the suite average, hides: σ_{pertask}, the sd across retrains of *per-task* SR, measured from per-episode vectors in three VLA cells. Net of the floor it reads $\mathbf{11.00}$ pp for $\pi_{0.5}$-ft on `libero_object`, 6.69 pp for SmolVLA-ft on the same suite and 3.29 pp on `libero_spatial`, all at k=88; the raw readings are 11.82, 7.33 and 4.52. Suite σ_{run} on the same runs is $\mathbf{4.23}$, 2.11 and 2.040, a $2.60\times$ gap in the first cell (CIs: Appendix K.2). The worst single task moves with sd 20.4 pp, a maximum one retrain carries; every aggregate claim here is made from σ_{pertask}, never from that maximum. The pre-registered claim that per-task noise exceeds suite noise holds at its frozen scope of *2 classes, LIBERO,* k=88, neither widened nor narrowed, while a pre-registered widening to any LIBERO suite came back **INDETERMINATE**, underpowered rather than null. The mechanism is idiosyncratic rather than global: a seed effect shifting all ten tasks together would give cross-task correlation ≈0.46, against measured 0.011 for $\pi_{0.5}$ and -0.036 for SmolVLA on `libero_object`. That contrast restates the ratio rather than confirming it independently, since at m=8 a correlation cannot carry a claim alone; we argue from the ratio and its CI. A retrain thus redistributes competence across tasks more than it moves it up or down; a ten-task average shrinks the swing by $\approx\sqrt{10} = 3.16$ while each task keeps its own.

The everyday improvement loop pays out even with the fix removed. Development commonly targets a run's worst task, changes something, retrains, and reads the rate again; the corpus's retrains on byte-identical data let that loop run with its fix removed at zero compute. Over 56 ordered pairs a retrain that changes *nothing* lifts the target by $+12.96$ pp [3.94, 22.67], clearing 10 pp in 48% of pairs and 20 pp in 25%, up to $+62$ pp, while the suite average over those pairs is zero by antisymmetry and a rollouts-only null gives $+2.85$ (99th $+5.61$): the movement is in the retrained weights. Exposure is rank churn and cell-specific: two retrains disagree on which task is worst 68% of the time here, against 25% and 0% in the SmolVLA cells, which return $+1.22$, inside their null, and a structurally exact $+0.00$ (exploratory, Appendix K.10). Relaunching the same seed as

an independent wave gave statistically indistinguishable checkpoints in this cell, an exploratory sign that the seed, not run-to-run chance, fixes the draw (Appendix K.3). On `libero_spatial` the correlation is $+0.141$, so more of the draw is shared there, and the suite-agreement test returned **DIVERGENT**, the two suites' CIs not overlapping. σ_{pertask} is thus a property of the model–suite cell, so every value carries a per-suite stamp and none is pooled (Appendices K.2, J).

Three pre-registered follow-ups stress-test these values. The strongest objection, that every VLA σ above is read before the learning rate has fully decayed, meets a measured null: completing the schedule at matched steps moves σ_{pertask} by -0.34 pp, inside both CIs, so the per-task lottery survives. The evaluation stream is sub-binomial, making both arms *lower bounds*, and the scope is SmolVLA-ft on `libero_object` at k=88 with the schedule's *shape* varied at fixed compute. Training *length* we do not bound, our strongest open threat (Appendix K.1). Repeat evaluation attributes most single-rollout flips to rollout sampling on a lower-bound reading, aggregates remain seed-driven, and a model soup of the m=8 retrains is viable, scoped to these parents (Appendices K.1, K.7–K.8).

The lottery is not a LIBERO artifact, and a release step can turn it into a gain. Over the eight $\pi_{0.5}$ retrains above, the retrain with the best suite average on a selection half of the episodes (best-validation) beats a random draw by $+6.1$ pp yet, because the retrain that wins on the average is rarely the one that wins on each task, leaves 1.82 tasks at least 5 pp below the mean of the eight on held-out episodes, which we call damaged; choosing the retrain per task from the same eight adds $+5.6$ pp and leaves 0.22. The same pre-registered design on Meta-World MT10 and MT50 (Yu et al., 2019), a different simulator and policy class, gives $+4.2$ to $+11.7$ pp over best-validation across ten replicate pools, and at 50 tasks a best-validation release ships a task damaged beyond sampling noise with probability 1.000, as independent per-task hazards predict, while per-task selection with abstention holds it at 0.02 or below. Fixing the initialization or the batch order alone leaves the lottery intact (Appendix L).

4.2 The paired program: the harness floor and the locality of composition

The lottery prices one fixed set; a curation claim compares *different* sets, and at small k the spread between sets is the larger obstacle. A pre-registered paired program of 86 runs in both regimes, frozen before any run existed, measured how much between-set variance cancels when sets share episodes. Pairing cut it only $1.92\times$ on PushT with diffusion policy and $1.73\times$ under SmolVLA fine-tuning, against pre-registered additive predictions of $11.5\times$ and $4.9\times$. On the fine-tuned cell, whose measured harness floor σ_0 is exactly zero, set value is therefore *not* a sum of per-episode contributions; this is the interventional rejection of the premise beneath per-episode scoring (§5.1). On PushT the floor we later measured absorbs that cell's prediction, so we withdraw the rejection there (Appendix F.1).

The harness floor is measured on both sides. On PushT/DP, $\sigma_0 = 1.91$ pp at $\mathrm{df} = 8$, so 53% of between-set variance is harness nondeterminism, not composition. Under bit-deterministic SmolVLA fine-tuning $\sigma_0 = 0.00$ exactly, and the clean locality estimate $\hat{w} = 0.49$ [0.21, 0.91] says about half of between-set variance does not cancel on shared episodes. Only 10.06% of same-index cross-run pairs on LIBERO share an initial condition (LeRobot, 2026), so the shared evaluation seed is not episode-level pairing and waives no floor (§3). Pairing makes the increment question, the effect of swapping a few episodes, measurable, but paired whole-set comparison costs $6.8\times$ and $2.9\times$ the unpaired design's runs (Appendices E, F.1).

4.3 The k-curve: set identity is a small-data property, the lottery is not

Whether the set-identity wall persists as the training set grows is tested by a pre-registered 32-run k-curve at $k \in \{22, 44, 88, 176\}$ in one cell, SmolVLA fine-tuned on LIBERO-object with 32/32 accepted, drawing uniformly from the full 10-task pool P=454. σ_{set} falls fast with k, faster than additive value predicts: it drops 3.2–$5.1\times$ from k=22 to 176, from 12.47 down to 3.86 pooled or 2.43 original, and $\sigma \propto k^{-\alpha}$ fits $\alpha = 0.58$–0.79, 0.68 overlap-corrected, above the additive model's 0.5. The $5.4\times$ ratio and every price built on it are therefore k=22 measurements, not regime constants: in this multi-task pool a 10 pp effect costs ∼24–28 draws per arm at k=22, ∼6 at 88. The set-identity wall is small-data, the retrain lottery is not: along the same curve σ_{set} collapses while σ_{run} is flat-to-rising at each k (df 2 per point, Table A9; the k=176 point carries the

amendment below). At production k the lottery is not what recedes but what remains (Tables A9 and A3, Appendix D.1). A pre-registered re-measurement of the k=176 rung read $\sigma_{\text{run}} = 2.60$ pp at $\text{df} = 4$, an indeterminate band; pooled under its frozen guard it gives $\sigma_{\text{run}}(176) = 3.36$ pp at $\text{df} = 6$ and $\sigma_{\text{set}}(\text{pure}, 176) = 3.86$ pp net of retrain noise, where the original 2.43 had scored our prediction FAILED. Because this flip runs toward our thesis we report it as a flip with both readings carried, and the k=176 price and caveat stand.

Two pre-registered results test ways around retraining-based measurement. *Adaptive evaluation*, which spends episodes on the most discriminating items, returned a negative result: neither ranked nor two-stage selection beats random in any cell tested, and Appendix B measures its accept threshold's false-pass rate. A second wave asked whether curated sets retrain more noisily than random ones; its manipulation check *FAILED*, so pre-registration rules both co-primaries uninformative and the question stays open (Appendix B; evaluation-episode prices: Table A4).

5 VALUE, SELECTION RULES, AND DATA QUANTITY ARE REGIME-LOCAL

This section asks whether what makes an episode set valuable is the same property across policy classes: from-scratch DP and fine-tuned SmolVLA train on identical episode sets under pre-registered 200-episode evaluations, data and protocol held fixed while the policy class varies.

The two disagree about what a set is worth. Over twelve shared random k=22 sets, DP and VLA per-set success correlate at Pearson $+0.29$, rank correlation lower, neither significant.

A selection rule reverses sign. Subsets maximizing physical state-space coverage, a policy-agnostic criterion, average $+12.2$ pp under DP on three of three draws, hierarchical $P(>0) = 0.86$, yet the same sets lose 23.2 pp under the VLA. A pre-registered rule-versus-draw test then dissolved even the DP gain: four fresh draws per arm gave $+13.8$ pp, not significant, with same-rule sets differing by sd 9–12 pp, some three times the cell's training-noise floor. Coverage in the VLA's own embedding space is null too, -5.8 pp, not significant and underpowered at four draws per arm. Fixing coverage barely constrains success; the draw dominates. The bound $\hat{w}$ caps what is removable, since sets sharing twenty of twenty-two episodes leave half uncanceled (§4.2, §7). The single-set trap then reproduces prospectively: in the frozen-prediction test on held-out task 1 of `libero_object` (below), one coverage-max set reads as a large win but is uninterpretable, its gap about $1.4\,\sigma_{\text{set}}$ and nearly matched by a *random* draw, as single-set evaluations routinely are (Figure A2, Appendix F).

Doubling the data helps on one suite and hurts on another. On `libero_object`, doubling 22 to 44 demonstrations gains $+20.9$ pp under DP and $+28.3$ pp under the VLA, the latter against the cell's floor-limited $\sigma_{\text{run}} \leq 3.31$ pp. A pre-registered breadth test spent the same 20K-step budget on task 0 of three more suites, with suite-local sign: -15.4 pp on long-horizon `libero_10`; -2.9 pp on near-ceiling `libero_goal`, within noise; and $+3.8$ pp on `libero_spatial`. On held-out task 1 of `libero_object`, prediction frozen at $+30.6 \pm 11.7$ before any run on it, doubling gave $+26.5$ pp, correct in sign, inside tolerance, the corpus's only prospective transfer success. At fixed compute, more demonstrations mean fewer passes over each, a suite-local trade, so "collect more data" is a regime claim, not a law (floors, post-repair numbers: §4, Appendix E).

5.1 NOTHING SCORES VALUE

No per-episode score we tested predicts measured episode value in either regime. If set value is regime-local, a score could still work *within* one regime. Against PushT causal ground truth, a datamodel fit from a 200-model random-subset design at k=103 (§2), twelve kinematic, coverage and gradient/attribution estimators reach $\max|\rho| = 0.121$. The ground truth's own noise sets its 0.54 reliability ceiling, so this is genuine failure, not attenuation. A thirteenth, DemInf-style per-episode state–action mutual information on raw z-scored rather than learned features (a recorded deviation), agrees: $\rho = -0.06$ on that target, CI straddling zero, the estimator healthy at pooled $I(s;a) \approx 3.9$ nats. Prospectively, sets keeping a quality heuristic's cleanest episodes land below random, its messiest above.

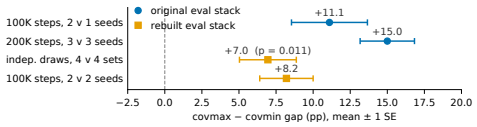

Figure 3: **The one principle that replicated, measured four independent times.** PushT gap between the coverage-maximizing (covmax) and coverage-minimizing (covmin) sets, mean $\pm$ 1 SE, across a $2\times$ budget change, independent re-draws and an environment rebuild, on two evaluation stacks.

These are our implementations. We re-ran no published curation code, so a reader may fairly blame our reimplementation rather than the method family, and our claim is narrower: these scorers, given ground truth and an interventional test, predict and move nothing. *Scope.* PushT demonstrations are scripted, so the pool is more uniform than the teleoperation datasets that curation methods target, RoboMimic among them. Range restriction, too little spread to score, is thus a live objection, yet the top-103 *by measured value* score 40.2% (§5.2): the spread is recoverable and the scorers fail where signal exists (§2, Appendices F and G).

The same quality heuristic is at chance on LIBERO. Its $+10.8$ pp "gain" for from-scratch DP on `libero_object` at k=22 sits inside $\pm 2\sigma_{\text{run}}$ of that setting's floor. Under the collapsed VLA floor, where the test has power (MDE ≈ 9 pp), the gap over eight pre-registered SmolVLA-ft runs is $+1.25$ pp, at chance even with power (floors: Table 1; scoping: Appendix F).

Early training loss predicts nothing either. Over the first 20–50% of its budget, a run's loss predicts final SR at $|\rho| \leq 0.19$ with negative cross-validated R^2 in the noisy regime, LIBERO-object under DP, and at $|\rho| \leq 0.12$ on the quiet PushT datamodel runs, where noise cannot be blamed (Appendix F). All fail because the paired program rejects the additive model interventionally (§4.2): the premise beneath per-episode scoring, that set value is a sum of per-episode contributions, fails a direct variance test, whatever the scorer, in the regime whose harness floor permits one, so the repeated $\rho \approx 0$ is not an estimator problem.

5.2 A REPLICATED CONSTRUCTIVE RESULT, AND WHY SCOPE MUST STILL BE STATED

One selection rule escapes the null: coverage-based selection replicates on PushT with a from-scratch diffusion policy. At k=103 and fixed budget, training-free greedy facility-location selection lifts success from 29.0% (covmin, the coverage-minimizing set) to 40.1% (covmax, coverage-maximizing), the 16th and 88th percentiles of 200 random subsets, matching the top 103 episodes by measured value at 40.2%. The low-coverage end rests on one surviving seed, so the swing is an upper bound. Unlike its LIBERO counterpart, it survives de-confounding: in four independent near-optimal draws per arm, each high-coverage draw beats each low-coverage one (Figure 3, row 3). Set-identity variance here is $\leq \sigma_{\text{run}}$, unlike under fine-tuning (§5). The facility-location *cost*, lower being better, predicts cross-set success at $r = -0.79$.

It also survived an accidental stress test. A rebuilt `gym_pusht` shifted these sets' absolute success by a directly measured -18.2 pp; a repair wave pre-registered before launch re-ran both deterministic sets on it, and the gap returned to the frozen $[+5, +20]$ band with complete arm separation. Figure 3 collects the four independent positive measurements, so in its regime the effect is method-level, not set luck or one build, and no audited curation paper reports comparable run-level replication (Appendix G; stack shift: §7; pre-registration and limits: Appendices D, F).

The result is regime-scoped, not a recommendation. Ported to LIBERO under the same policy class, it failed its frozen validation criteria, and under a fine-tuned VLA it does nothing reliable. With two seeds per arm and the covmax pair tied, the repair wave is a directional confirmation, not a powered estimate, and the rebuilt cell's σ_{run} sits below its own binomial floor in a lower-SR regime, a bound on that cell's noise and never a pricing prior (§4, Appendix E).

6 RELEASED ARTIFACTS, CORPUS, AND PRE-REGISTRATIONS

The record behind every claim above is released: a corpus of 701 trained policies in four policy classes, a frozen *census* (v0.7) of 621 runs from which our null distributions are drawn, plus 80

scored post-freeze runs. Each record carries its episode mask, configuration and SR (Appendix C), and 334 census runs add a per-episode success vector reproducing the recorded rate exactly, 0 mismatches in 334. Every rule is set against random subsets of matched size and budget, so the corpus is itself a null distribution and noise floor. **Seventeen frozen pre-registrations govern the census waves**, each later wave carrying its own, and **twelve outcomes contradicted our own predictions and declared interests**, four in the census and eight later, each reported as such where used (Appendices D, K). Per-model records, pre-registrations, analyzers and wave documents are collected in the anonymous supplementary material accompanying this submission, the reusable kit in Appendix H, and counts, budgets, the disclosed provenance defect, registry, hashes, verdict criteria and amendment chains in Appendices C, D and J–L.

7 Conclusions, recommendations, and limitations

Summary. Across 701 policies, retraining noise is large and cell-specific: eight retrainings on identical data move LIBERO per-task success by $\sigma_{\mathrm{pertask}}$ 11.82 pp raw, single tasks by up to 20.4, while the ten-task average moves 4.23 pp, and no scale-free transform we tested carries a floor between cells, so we quote no cross-cell multiple. The suite average understates per-task noise, which retraining redistributes across tasks almost independently, so a loop that retrains to fix one task and reports the average cannot tell a fix from a fresh draw. Where per-task noise is large, its structure pays: choosing the retrain per task from a fixed pool beats the best single retrain by 4 to 12 pp on Meta-World and `libero_object`, and all but removes the damaged-task tail, in proportion to the suite's own $\sigma_{\mathrm{pertask}}$, not as a constant (§4.1). Nothing we tested transferred unchanged between settings: which demonstrations help depends on policy class, no per-episode score we could build tracks causal value, and set value is not additive; one causally validated rule survives, in one cell. Against these floors the audited curation gains are unverified as reported, not refuted, and twelve pre-registered results went against us, all reported as such (§§4–5.2, §6).

A reporting standard for curation claims. *(A) Measure your own cell's floor.* **1.** Report no curation gain without σ_{run} in its regime and set-identity dispersion at the method claim's k, sized from Tables A1–A4. **2.** Assert the realized training budget before recording a result, and report it per run. **3.** Report each run's learning-rate schedule as executed, not as configured. *(B) Quote both readings and name your scope.* **4.** Quote every σ raw and floor-subtracted, with its floor share. **5.** Name the suite on every $\sigma_{\mathrm{pertask}}$, because the two we measured diverge. **6.** Stamp a near-floor σ LOW-SR, and never read it as a class property. **7.** Quote no cross-cell σ_{run} ratio unless both cells clear a stated comparability gate (§4). *(C) Design the test and pre-register the verdict.* **8.** Test a rule on sets built at the extremes of its own score, not on observational correlations. **9.** Scope any constructive claim to an explicit policy class. **10.** Pre-register the verdict criteria before training.

Standard 2 comes from our own defect: trusting checkpoint directories over the trainer's exit status, our driver corrupted 12 of 414 atlas runs and flipped one suite's quantity sign, and because truncation lowers scores it is *not* mean-zero noise but mimics whichever effect predicts a lower score; the resulting budget-assertion gate covered all 207 post-atlas and 80 post-freeze runs, none silently truncated (Appendix E).

Limitations. *Scope.* All measurements are in simulation: eight single-task settings, two of them bimanual ALOHA (Zhao et al., 2023), two ten-task LIBERO pools, four policy classes, plus two Meta-World suites under a small state-based policy outside the corpus (Appendix L). PushT demonstrations are scripted, not teleoperated; the three ACT cells allow no cross-class comparison; multi-task evaluation with pixels beyond LIBERO, a real-robot leg, and running the released curation codebases are future work. *Cross-stack comparability.* A pre-registered calibration put our two PushT stacks **18.2 pp** apart, wider than our frozen band predicted, and we apply no cross-stack correction (Appendices C, E). *Power, withdrawn rows, one open trend.* Several key results are **nulls at four sets per arm** ($p = 0.13$ set diversity, $p = 0.21$ mutual-information de-confounding); resolving them costs about 25 sets per arm. The Gaussian method-comparison rows are **withdrawn**, survivors claimed only at their k (Appendix F). The k-curve's df 2 per-k σ_{run} rise is unexplained, its k=176 **flip in the flattering direction** reported as such (§4.3), and retrain-level certification of a deployed policy is open (Appendices F.2, L). **The remedy is cheap**: repeat evaluations measure a setting's testing noise, eight retrainings of one unchanged demonstration set measure its training floor, and we propose that no curation gain be quoted without both beside it.

ETHICS STATEMENT

This work involves no human subjects, no personally identifying data and no real-robot deployment: every measurement is a simulation rollout. The released corpus consists of per-model records, evaluation outputs and analysis documents derived from public simulation benchmarks and public pre-trained checkpoints. A license audit of the source datasets is named in the Reproducibility statement as still in preparation, and we make no redistribution claim over material we did not produce until it completes.

This paper audits published work, so we state the scope of that audit precisely. Figure 1 and Appendix G place six papers' reported headline gains against the noise floors we measured, and *we did not run their code or reproduce their experiments*. The claim is about the reported evidence, not about whether the methods work: an effect inside its marker is **unverified as reported, not refuted**, and none of our results should be cited as showing that any audited method fails. Where we tested a method class interventionally, we tested *our own implementation* of it and say so at every use.

Compute for the experiments was provided by a third party, disclosed in the non-anonymous version of this paper. That party had no role in the research questions, the experimental design, the pre-registrations, the analyses or the conclusions, and no right of review over them. Our own prior expectations were recorded as declared interests inside the pre-registrations before each wave ran, which is why the predictions that went against us can be counted: §6.

REPRODUCIBILITY STATEMENT

All experiments, per-model records, pre-registrations, mask generators, wave drivers, analyzers and per-wave results documents are deposited in the anonymous supplementary material accompanying this submission. The pre-registrations were frozen and SHA-256-timestamped to cloud storage before each wave launched, and any in-flight amendments are hash-chained and disclosed. Appendix I gives the full artifact inventory and names what is still in preparation: a public mirror, a datasheet, Croissant metadata, a license audit of the source datasets, and one-command reproduction of the headline analyses. None of that is available at submission time.

Three properties of the bundle are worth stating here rather than leaving a reviewer to discover them. First, it carries its own `LICENSE` (MIT for code, CC BY 4.0 for the measurement records) and an `ENVIRONMENT.md` recording that every σ, interval and verdict in this paper recomputes from the shipped per-episode vectors with only `numpy` and `scipy`. Regenerating the corpus itself needs GPUs we no longer have, which is why the per-episode vectors and not merely the summary statistics are released. Second, anonymisation is a published bijection (`anonymise.py`), with its round trip asserted file by file when the bundle is built, so `verify_prereg_hashes.sh` hashes the *de-anonymised* bytes. Every pass it reports is a statement about the document as frozen, not about its redaction. It also now counts the pre-registrations and results documents that carry *no* frozen digest, rather than passing over them in silence; no priority is claimed for those. Third, one analysis does not reproduce from the bundle alone. The PushT causal ground truth is a ridge fit over 200 per-model records that are not included. The fitted per-episode weights ship, so every downstream claim can be checked against them, but the fit cannot be re-run from this bundle. We state this as a limitation rather than implying end-to-end reproducibility.

What the hash chain does *not* establish should also be plain. The digests were written to cloud storage we controlled. That fixes ordering and internal consistency; it is not a third-party timestamp, and we make no notarisation claim. An RFC-3161 or OpenTimestamps anchor is the right instrument and cannot be applied retroactively to waves already run. Every confirmatory quantitative claim traces to a frozen, hash-chained pre-registration and to a per-wave results document, with amendment chains and their SHA-256 hashes recorded per wave. Exploratory and descriptive analyses are labeled as such and carry no verdicts. §3 specifies the corpus protocol: fixed budgets, final-checkpoint evaluation, a shared evaluation seed, and in-wave inference. That shared seed does not establish episode-level common-random-number (CRN) pairing on LIBERO, and pairing claims are licensed per wave in §4. §3 also records the reporting conventions used throughout. Every σ quote carries both its raw and floor-subtracted readings. Every $\sigma_{\mathrm{pertask}}$ names its suite. Every VLA measurement carries its training-schedule provenance: truncated, 20k steps / 30k decay. All figures regenerate from the scripts in `figures/src/`, each with inline provenance comments mapping every plotted number to its source document.

AI USE STATEMENT

In this work, we used generative AI tools for drafting assistance on the manuscript and for codebase engineering: implementing analysis, scoring and figure code, and wave drivers. We did not use generative AI tools to generate the research questions or the experimental design, to write the pre-registrations or their verdict criteria, to produce or alter any measurement, or to interpret the results; those are our own. We have reviewed all AI-assisted work. Scoring and analysis code was verified against synthetic fixtures with known answers before any contact with wave data. Every confirmatory quantitative claim traces to a frozen, hash-chained pre-registration and to a per-wave results document, as recorded above. Each figure script carries inline provenance comments mapping every plotted number to its source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2108.13264.

Christopher Agia, Rohan Sinha, Jingyun Yang, Rika Antonova, Marco Pavone, Haruki Nishimura, Masha Itkina, and Jeannette Bohg. CUPID: Curating data your robot loves with influence functions. In *Conference on Robot Learning (CoRL)*, 2025. arXiv:2506.19121.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.05364.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2006.14651.

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Data quality in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.02437.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021. arXiv:2103.03098.

Annie S. Chen, Alec M. Lessing, Yuejiang Liu, and Chelsea Finn. Curating demonstrations using online experience. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2503.03707.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023. arXiv:2303.04137.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. arXiv:2011.03395.

Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Madry, Andrew Ilyas, and Roberto Martín-Martín. DataMIL: Selecting data for robot imitation learning with datamodels. *arXiv preprint arXiv:2505.09603*, 2025.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. arXiv:1909.03004.

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

Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning (CoRL)*, 2021. arXiv:2109.00137. Origin of the PushT task.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1802.10026.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1904.02868.

Joey Hejna, Chethan Bhateja, Yichen Jiang, Karl Pertsch, and Dorsa Sadigh. Re-Mix: Optimizing data mixtures for large scale imitation learning. In *Conference on Robot Learning (CoRL)*, 2024. arXiv:2408.14037.

Joey Hejna, Suvir Mirchandani, Ashwin Balakrishna, Annie Xie, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, Dhruv Shah, Coline Devin, and Dorsa Sadigh. Robot data curation with mutual information estimators. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2502.08623.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. arXiv:1709.06560.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*, 2024.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2202.00622.

Scott M. Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip S. Thomas. Evaluating the performance of reinforcement learning algorithms. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:2006.16958.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017. arXiv:1703.04730.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.13289.

Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot learning as an empirical science: Best practices for policy evaluation. *arXiv preprint arXiv:2409.09491*, 2024.

Soham Kulkarni, Raayan Dhar, and Yuchen Cui. Learning from the best: Smoothness-driven metrics for data quality in imitation learning. *arXiv preprint arXiv:2604.23000*, 2026.

LeRobot. LeRobot v0.6.0, 2026. Software release; the training and evaluation defaults referenced here (default evaluation seed; learning-rate schedule length) are properties of this version of the code, not of the library paper.

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 (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.03310.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021. arXiv:2108.03298.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (EMNLP)*, pp. 217–227, 2020. arXiv:1911.02969.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K. R. Zentner, Ryan Julian, J. K. Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-world+: An improved, standardized, RL benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv:2505.11289.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:1906.01827.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.14186.

David Picard. torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Maram Sakr, H.F. Machiel Van der Loos, Dana Kulic, and Elizabeth Croft. Consistency matters: Defining demonstration data quality metrics in robot learning from demonstration. *ACM Transactions on Human-Robot Interaction*, 2025. arXiv:2412.14309.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

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. *arXiv preprint arXiv:2506.01844*, 2025.

David Silver and Richard S. Sutton. Welcome to the era of experience. Technical report, Google DeepMind, 2025. Preprint of a chapter in *Designing an Intelligence*, MIT Press.

Sreevardhan Sirigiri, Nathan Samuel de Lara, Christopher Agia, Florian Shkurti, and Fabio Ramos. Diversity you can actually measure: A fast, model-free diversity metric for robotics datasets. *arXiv preprint arXiv:2603.11634*, 2026.

David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Is your imitation learning policy better than mine? policy comparison with near-optimal stopping. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2503.10966.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624:86–91, 2023. Author correction: *Nature*, 2026, following independent re-analysis.

TRI LBM Team. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

Jiachen T. Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. arXiv:2205.15466.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning (ICML)*, 2015.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2203.05482.

Shan Xie, Man Luo, Chadly Daniel Stern, Mengnan Du, and Lu Cheng. DemoShapley: Valuation of demonstrations for in-context learning. *arXiv preprint arXiv:2410.07523*, 2024.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:2011.09161.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019. arXiv:1910.10897.

Edwin Zhang, Vincent Zhu, Naomi Saphra, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham M. Kakade, and Eran Malach. Transcendence: Generative models can outperform the experts that train them. *arXiv preprint arXiv:2406.11741*, 2024.

Yu Zhang, Yuqi Xie, Huihan Liu, Rutav Shah, Michael Wan, Linxi Fan, and Yuke Zhu. SCIZOR: A self-supervised approach to data curation for large-scale imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. Oral. arXiv:2505.22626.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023. arXiv:2304.13705.

A THE COLLAPSE-MIXTURE MODEL AND ITS MONTE CARLO PRICING

A single independent set draw (one fresh episode set, one fresh training seed) is modelled as

$$\mathrm{SR} = (1 - Z) \cdot \mathcal{N}(\mu_h, \sigma_h^2) \; + \; Z \cdot \mathcal{N}(\mu_h - \Delta_c, \sigma_c^2), \qquad Z \sim \mathrm{Bernoulli}(p_c),$$

clipped to $[0, 100]$; a method effect δ shifts the healthy-mode mean of one arm, and the collapse process is common to both arms. With only 4–6 independent same-size draws per cell, estimation is interval-based, not point maximum-likelihood: p_c carries a Jeffreys Beta posterior on pooled collapse counts—2–4 of 16 sets at $k = 22$, where the count range reflects the structurally ambiguous duplicate 38.5 draws of the Addendum-A cell, carried as two readings through every table; zero observed collapses at each $k \geq 44$ yield per-k upper bounds only—σ_h carries its 95% χ^2 interval at its actual degrees of freedom (1–5), upper limits capped at 20 pp on boundedness grounds; Δ_c (healthy-mode mean minus collapse level; observed 21.3–29.1 pp) is swept over $[20, 30]$ pp; and σ_c is fixed at 4 pp (MDE shifts $\leq$ 1.5 pp over $[2, 8]$). Minimum detectable effects and draws per arm are computed by Monte Carlo simulation of two-arm Welch t-tests (two-sided $\alpha = .05$, sign-correct rejections, 80% power), reported as central values with min–max envelopes over this parameter box; p_c is scanned at all posterior quantiles because power is non-monotone in p_c. Envelopes cross marginal 95% limits—conservative outer bounds, not joint credible intervals. Realised type-I error under the null mixture is 0.02–0.05. Fitting and simulation: `experiments/influence/mde_mixture.py` (Monte Carlo seed 20260807—the session date; the K4B mask generator derives its distinct stream from the same date base, `random.Random(20260807 + 176)`, a different generator and offset); log `MDE_MIXTURE_LOG.txt`; every entry of Tables A1–A3 and the notes' auxiliary numbers mapped to the log in `MIXTURE_MDE_RESULTS.md`.

Table A1: **Mixture table R1—fitted collapse-mixture parameters per cell** (independent single-seed set draws; SR in pp). Model specification, priors, and Monte Carlo procedure: this appendix; fitting script `mde_mixture.py`, log `MDE_MIXTURE_LOG.txt`, entry-by-entry mapping for Tables R1–R3 in `MIXTURE_MDE_RESULTS.md` (released kit, Appendix I). All cells are VLA fine-tunes on the truncated LR schedule (schedule-provenance note below). **Each row is one (policy class, task scope, pool, k) cell and prices only itself**; the three k=22 cells are not replicates of one another.

cell	k	draws	collapses	p_c central [95%]	σ_h pp [95% CI] (df)	Δ_c pp
SmolVLA-ft / **single-task** `object` t1, task's own pool (Add. A)—(i) 38.5-pair healthy	22	4	0	0.13 [0.03, 0.34]	19.0 [10.8, 70.9 → cap 20.0] (3)	—
—(ii) 38.5-pair collapsed	22	4	2	0.26 [0.09, 0.49]	15.4 [6.9, 491.9 → cap 20.0] (1)	29.1
$\pi_{0.5}$-ft / **single-task** `object` t0, pool P=44¶ (collapse retrain-attributable—notes)	22	6	1	0.13–0.26 [0.03, 0.49]	4.5 [2.7, 12.9] (4)	24.5
SmolVLA-ft / **multi-task** 10-task `object`, pool P=454	22	6	1	0.13–0.26 [0.03, 0.49]	10.2 [6.1, 29.2 → cap 20.0] (4)	21.3
SmolVLA-ft / **multi-task** 10-task `object`, pool P=454	44	6	0	0.04 [0, 0.33]	5.8 [3.6, 14.2] (5)	—
SmolVLA-ft / **multi-task** 10-task `object`, pool P=454‡	88	6 (+12)	0 in 18 sets	0.01 [0, 0.13]	4.3 [2.7, 10.6] (5)	—
SmolVLA-ft / **multi-task** 10-task `object`, pool P=454†	176	6	0	0.04 [0, 0.33]	5.1 [3.2, 12.6] (5)	—

Notes to R1.

- *Pooling and the Addendum-A ambiguity.* p_c at $k = 22$ is the pooled Jeffreys posterior over the three $k = 22$ cells; the pooled count is 2/16 or 4/16 depending on whether Addendum A's two draws at exactly 38.5 are read as healthy or collapsed—unresolvable at $n = 4$, so both readings are carried through Tables R2–R3 (per-cell medians, $1/6 \to 0.19$ and $2/4 \to 0.50$, lie within or at the scanned envelope).
- *Zero-collapse rows are upper bounds.* $k \geq 44$ rows are per-k Jeffreys with zero observed collapses—upper bounds, not estimates; pooling the 0/30 sets across $k \geq 44$ (a constant-rate assumption) would tighten the bound to ≤ 0.080. The $k = 88$ row's zero count covers 18 sets (56 runs: 8 k-curve + 48 CURATED-88, the "+12" in the draws column).
- *Collapse-mode parameters.* Δ_c = healthy-mode mean − collapse level; envelope sweep [20, 30] pp; collapse-mode sd fixed at 4 pp ($\pi_{0.5}$ $m = 5$ MDE 20.3/20.7/21.8 at $\sigma_c = 2/4/8$). σ_h upper limits are capped at 20 pp on bounded-$[0, 100]$-scale grounds; the cap binds where marked.
- *Collapse attribution.* The multi-task $k = 22$ collapse (25.0) replicated at 28.0 under seed 1 (set-attributable). The $\pi_{0.5}$ collapse (50.0) resolved the opposite way: its mask retrained at a second seed scored 87.5, above the frozen 65.0 threshold, so per the pre-registered PI05-PAIRS rule it is retrain-attributable; the cell's 16-run re-measurement gives $\sigma_{\mathrm{run}} = 9.74$ pp

($\mathrm{df} = 10$) with $\sigma_{\mathrm{set}} = 0.00$ by method of moments—superseding the cell's earlier atlas reading of 2.02 pp ($\mathrm{df} = 2$, χ^2 CI [1.05, 12.67], from two replicate pairs); the supersession is reported in §4 and Table A10. This row's single-seed draws therefore mix retrain noise into what the fit reads as set-level dispersion (row retained as fitted; the R2–R3 prices remain correct for the independent-draw design—scope note below). The $k = 44$ draw at 47.0 is not classed as a collapse—if it is one, the $k = 44$ rows degrade toward the $k = 22$ regime.

- *Schedule provenance (every row here and wherever these cells appear in Tables R2–R3).* All cells are VLA fine-tunes trained on the ecosystem-default truncated LR schedule (schedule: truncated, 20k steps / 30k decay; final LR 28.4% of peak). ‡For the SmolVLA-ft k=88 cell the pre-registered ANNEAL-1 completion control measured the completion effect as null: completing the decay within the same 20,000 steps moves $\sigma_{\mathrm{pertask}}$ by -0.34 pp (the two arms' estimates lie deep inside each other's CIs)—LOTTERY-SURVIVES, scope SmolVLA-ft/`libero_object`/k=88 only; stamp carried with that verdict: episode-level CRN pairing between arms is NOT established; the comparator is matched-n; untransferred dispersion excess biases toward ANNEAL-EXPLAINS—the direction that did not win (Appendix K). ¶For $\pi_{0.5}$ rows the stamp is "truncated schedule; completion effect untested."
- †*K4B.* Scored (§4.3); row unaffected: mixture fits use set-level draws, not σ_{run}.

Table A2: **Mixture table R2—replacement method-comparison MDE** (pp of healthy-mode shift; 80% power, $\alpha = .05$, two-sided Welch t on m independent single-seed set draws per arm; Monte Carlo under the fitted mixture). Entries: central (range over the two Addendum-A readings) [min–max envelope over the parameter box]; >50 = not resolvable on a $[0, 100]$ SR scale. **Quoting rule: no entry may be quoted without its row**—every number here is a (policy class, task scope, pool, k, m) quintuple, and the three k=22 cells differ by up to 3.9× at fixed m because they differ in task scope and pool, not in precision.

cell (class / task scope)	pool	k	MDE (pp) at m independent set draws per arm				
			m=2	m=3	m=5	m=10	m=25
SmolVLA-ft / single-task `object` t1 (Add. A)	task's own	22	>50	>50 [32, >50]	40–41 [19, >50]	25–26 [12, 29]	15–16 [7, 18]
$\pi_{0.5}$-ft / single-task `object` t0	P=44	22	>50	46–>50 [12, >50]	21–24 [8, 40]	13–15 [6, 24]	8–9 [3, 15]
SmolVLA-ft / multi-task 10-task `object`	P=454	22	>50	41–43 [24, >50]	25–27 [14, 45]	16–17 [9, 29]	10–11 [6, 18]
SmolVLA-ft / multi-task 10-task `object`	P=454	44	>50	26 [13, >50]	15 [8, 38]	10 [5, 25]	6 [3, 15]
SmolVLA-ft / multi-task 10-task `object`	P=454	88	>50 [28, >50]	17 [10, >50]	10 [6, 30]	7 [4, 19]	4 [2, 12]
SmolVLA-ft / multi-task 10-task `object`†	P=454	176	>50	23 [11, >50]	14 [7, 37]	9 [4, 24]	6 [3, 15]

†K4B scored (§4.3); row unaffected: mixture fits use set-level draws, not σ_{run}.

Withdrawn Gaussian reference to R2. The withdrawn Gaussian propagation ($\mathrm{MDE} = 2.80 \cdot \mathrm{sd}_{\mathrm{total}} \cdot \sqrt{2/m}$, where a method comparison takes $\mathrm{sd}_{\mathrm{total}} = \sqrt{\sigma^2_{\mathrm{set}} + \sigma^2_{\mathrm{run}}}$, §4) is materially optimistic at m = 2–3 in every $k = 22$ mixture cell (a single collapse draw both shifts an arm mean and inflates the Welch variance at $\mathrm{df} \approx 2$) and at all m for $k \geq 44$ (it assumes known σ and no collapse tail); approximate agreement occurs only near $m = 10$ at $k = 22$. Its figures at $m = 2/3/5/10/25$: Add. A 54.1/44.2/34.2/24.2/15.3; $\pi_{0.5}$ 30.2/24.6/19.1/13.5/8.5; multi-task k=22 35.2/28.8/22.3/15.8/10.0; k=44 16.2/13.2/10.2/7.2/4.6; k=88 12.1/9.9/7.7/5.4/3.4; k=176 14.3/11.7/9.1/6.4/4.1.

One convention correction, recorded rather than applied silently. The Add. A figures above now propagate the same $\mathrm{sd}_{\mathrm{total}} = \sqrt{19.01^2 + 3.51^2} = 19.33$ that Table A11 and Figure A4 use for this cell, and that every other row here already used (12.58 at multi-task k=22, 5.78/4.34/5.12 at k=44/88/176). An earlier printing of this one row propagated $\sigma_{\mathrm{set}} = 19.01$ alone, giving 53.2/43.5/33.7/23.8/15.1 here and, in Table A3, 14/**57**/227 rather than 15/≈59/234 draws per arm at $\delta = 20/10/5$ pp. Both readings sit well inside this appendix's stated precision scope—direction and order of magnitude only—but only one of them is the paper's own method-comparison formula, so the quotable withdrawn-Gaussian price at this cell is ≈59 **draws per arm**, as §4 and Figure A4 state.

Scope (applies to Tables R2–R3).

- *Precision.* The 4–6 draws per cell license direction and order of magnitude only; no entry is a decimal-precise price.

Table A3: **Mixture table R3—independent set draws per arm to resolve a method effect δ (80% power),** Monte Carlo under the fitted collapse mixture (parameters: Table A1; model: this appendix). Entries: central (range over the two Addendum-A readings) [min–max envelope over the parameter box]; withdrawn Gaussian-propagation figure in parentheses (all rows on the $\mathrm{sd}_{\mathrm{total}} = \sqrt{\sigma^2_{\mathrm{set}} + \sigma^2_{\mathrm{run}}}$ convention—see the R2 correction note). **Quoting rule: a price is a (class, task scope, pool, k, δ) quintuple and may not be quoted without all five**—the three k=22 rows differ by up to 3.9× at $\delta = 10$ pp (63 vs 16 draws/arm) purely by task scope and pool.

cell (class / task scope)	pool	k	set draws per arm at effect size δ		
			$\delta = 20$ pp	$\delta = 10$ pp	$\delta = 5$ pp
SmolVLA-ft / single-task `object` t1 (Add. A)	task's own	22	15–17 [5, 21] (15)	**56–63 [14, 79]** (59)	220–244 [51, 302] (234)
$\pi_{0.5}$-ft / single-task `object` t0	P=44	22	6–7 [3, 14] (5)	**16–22 [4, 53]** (18)	57–84 [13, 200] (73)
SmolVLA-ft / multi-task 10-task `object`	P=454	22	7–8 [4, 20] (6)	**24–28 [9, 76]** (25)	93–105 [30, 301] (99)
SmolVLA-ft / multi-task 10-task `object`	P=454	44	4 [3, 15] (1)	10 [4, 54] (5)	36 [10, 216] (21)
SmolVLA-ft / multi-task 10-task `object`	P=454	88	3 [3, 10] (1)	**6 [3, 33]** (3)	18 [6, 130] (12)
SmolVLA-ft / multi-task 10-task `object`†	P=454	176	4 [3, 14] (1)	9 [4, 51] (4)	31 [8, 198] (16)

Notes. *The four bold entries are the prices quoted in §4, and three of them are k=22.* They are not alternative estimates of one price: 56–63 is the **Addendum-A single-task** cell (`libero_object` task 1, that task's own pool), 16–22 is the **other single-task** k=22 cell ($\pi_{0.5}$-ft on task 0, pool P=44), and 24–28 is the **multi-task** cell (10-task pool P=454)—so the 3.9× spread across them is task scope, pool and policy class, not disagreement, and any of the three quoted bare would misprice the other two. §4 states the single-task pair as a range for that reason. Upper envelope ends at $k \geq 44$ (33–54 draws at $\delta = 10$) are set by df=5 σ uncertainty plus unexcluded rare collapses ($0/6$ sets bounds $p_c \leq 0.33$ at k=44 and k=176). Prices of $\gtrsim$200 draws exceed the replication budgets reported in the robot-curation studies we audit (§2; the largest reported is three seeds). Realised type-I error of the Welch t under the null mixture is 0.02–0.05 across cells and m—conservative; no entry is anticonservative. A Wilcoxon rank-sum at the multi-task k=22 central parameters reaches 0.77 power at the Welch-t MDE (m=10): rank tests do not rescue the budget, because the collapse tail is rare rather than mild. †K4B scored (§4.3); row unaffected: mixture fits use set-level draws, not σ_{run}.

Table A4: **Mixture table R4—empirical anchor: eval episodes to rank two fixed checkpoints under a shared evaluation seed** (episode-level pairing scope: §4; Appendix B)—the complementary axis to Table A3's retraining draws; each R3 draw costs one retraining plus ≥100 eval episodes.

effect size	eval episodes needed
$\delta > 20$ pp	∼25 decides it
δ 10–20 pp	∼50 for >93%
δ 5–10 pp	∼100 for ∼93% (50 gets ∼80%)
$\delta < 2$ pp	42–63% accuracy at every budget tested—a coin flip

- *Estimand.* The model treats a method effect as a healthy-mode shift with the collapse process common to both arms; a method that changes p_c itself is a different estimand that none of these budgets resolves. σ_h is set-level dispersion under independent single-seed set draws, not σ_{run}; no cross-cell noise multiple may be read off these tables—cross-cell σ_{run} comparisons are licensed only where `atlas.compare_sigma_run()` returns one instead of refusing (§4).
- *Collapse attribution.* Whether collapse attaches to the training run or the episode set is now measured in both directions and is cell-specific: the pre-registered PI05-PAIRS seed expansion attributes the $\pi_{0.5}$ collapse to the training run (`null04`: 50.0 vs 87.5 across seeds on one mask—retrain-level, §4), while the multi-task $k = 22$ collapse replicated under a second seed (set-level). The pricing is correct for the independent-draw design either way; set-reusing designs must settle the attribution per cell—in a retrain-level cell like $\pi_{0.5}$, reusing a set does not avoid the collapse mode.
- *Monte Carlo.* 20,000 simulations per central entry, 8,000 per envelope corner (R3: 12,000/6,000), seed 20260807 (`mde_mixture.py`); resolution $\approx\pm1$ in the last digit.
- *Per-task scope.* Multi-task $n_{\text{eval}} = 200$ is 20 episodes $\times$ 10 tasks—the evaluation configuration's episode count is per-task, a reporting hazard noted in §3. These n=20/task evaluations are also out of band for per-task noise pricing: at n=20/task the per-task binomial floor ($\approx$10.9 pp) is the same size as the raw per-task SD, so the floor-subtracted σ_{pertask} estimator saturates; every per-task price in this paper comes from the $n\geq$100/task waves (Appendix K).
- *K4B.* The $k = 176$ rows were priced before K4B ran; the scored K4B wave (§4.3) reattributes variance between set identity and training noise without moving $\sigma_{\text{set}}(\text{total})$, leaving these rows unchanged.

B Adaptive evaluation and the curated-vs-random wave

B.1 Adaptive evaluation on the shared-seed corpus: real item structure, no exploitable item selection

Every run in the corpus is scored under a shared evaluation seed (§4.2), so the per-episode records form an item-response dataset: each evaluation episode index is an item every policy has answered. One scope caveat governs the LIBERO cells throughout this appendix: the shared seed pins the episode *index*, not the initial state—LIBERO's init-state pointer is outcome-dependent, and only $\sim$10% of same-index cross-run pairs share an initial condition (Appendix K.1)—so LIBERO cross-run item statistics mix state mismatch with policy variation, while PushT items are alignment-clean. We asked whether the item structure supports adaptive evaluation—spend rollouts on informative items, cut the episode budget—on the 246-run snapshot of the corpus that predates the 2026-08 waves (74,100 per-episode outcomes; `research/ADAPTIVE_EVAL_PROGRAMME.md`; the waves themselves are censused in Appendix C and reported in §4.3 and Appendices J–K). The result is negative with a usable residue, and it prices the one axis a reader can economize on: evaluation cost for a *fixed* checkpoint. It cannot touch σ_{run} or σ_{set}, which sit upstream in retraining.

The item structure is real: per-item success rates over-disperse relative to an equal-difficulty binomial by 0.95–3.09$\times$ across five cells (per-cell multipliers: Appendix E), and item difficulty even transfers partially across policy classes—Pearson $+0.484$ / Spearman $+0.407$ between DP and SmolVLA-ft on `libero_object`, $\sim$23% shared variance, far above the $\rho \approx 0$ every curation lever returns (§5.1). Exploiting it fails: neither discrimination-ranked nor two-stage adaptive item selection raises the discordant-item rate above random selection in any of the three cells tested, against oracles showing $\sim$2.5$\times$ headroom (full rate triplets and over-dispersion multipliers: Appendix E). The procedure's accept threshold falsely passes 16.1% of ordered no-change gates (62 of 386) on same-data pairs at stated $\alpha = 0.05$, or 4.9% net of σ_{run}. The mechanism is a **flip floor**: 33.1% (DP) / 28.2% (SmolVLA-ft) of item outcomes differ between runs trained on identical data, and only 0.3–0.5% of items never flip—which items a given pair disagrees on is idiosyncratic, not predictable from latent item traits. (The SmolVLA-ft figure is computed over same-index cross-run LIBERO pairs and therefore includes state mismatch, per the pointer caveat above; in the cell where state alignment is pinned, the between-checkpoint flip rate on matched states is 0.229, Ap-

pendix K.8. The same floor reappears in the curated-vs-random wave at k=88: flip rate 22.9% vs 20.8%, Appendix B.2.)

The flips do not simply "live in retraining"—an earlier version of this paragraph said they did, and the pre-registered REPEATEVAL-ZERO decomposition (Appendix K.8) reconciles the claim. A fixed checkpoint's responses are bit-stable only under the deterministic battery configuration that measured $\sigma_0 = 0.00$; under its stochastic flow-matching sampler the *same* checkpoint flips $f_{\text{within}} = 0.145$ of state-matched items against itself, versus $f_{\text{between}} = 0.229$ between identically-trained checkpoints on the same states. Leading with the difference: $\Delta = f_{\text{between}} - f_{\text{within}} = 0.0841$ (from the unrounded counts, $6{,}416/28{,}000 - 580/4{,}000$; $\text{sd}_{\text{seed}}(q_{\text{item}}) = \sqrt{\Delta/2} \approx 0.205$); the ratio $\Phi = 0.633$ is quoted second and is a lower bound (conventions 0.6328 same-pass / 0.6306 all-pass-pairs, both carried): **the majority of item-level flipping is the policy's own rollout sampling, not the training seed** (SAMPLING-MAJORITY; a descriptive decomposition with disclosed prior visibility, not a blind confirmatory test). Aggregate quantities are untouched: suite and per-task success rates average the sampling term over 100+ episodes, and the aggregate lottery (σ_{run} 2.11–4.23 pp at multi-task k=88, per-task σ 3.3–11.0 pp) remains seed-driven. The analysis also reproduced, on ourselves, the in-sample trap this paper keeps finding in the literature: selecting items by in-sample $|r|$ claims a $\sim 4\times$ budget saving; cross-validated, the saving deflates to $\sim 2\times$ on DP/PushT and to nothing on LIBERO. We report the deflated numbers (Appendix E).

What survives is a price list, not a method (Table A4, Appendix A): with random items under a shared evaluation seed, ranking two fixed checkpoints costs $\sim$25 episodes for $\Delta > 20$ pp; $\sim$50 for $\Delta = 10$–20 pp at $>93\%$ accuracy; $\sim$100 for $\Delta = 5$–10 pp at $\sim 93\%$ (50 buys $\sim 80\%$); and **checkpoints within $\sim$2 pp of each other cannot be separated at any rollout budget tested—a coin flip** (42–63% accuracy, including $m = 100$). Taking the max over near-tied checkpoints then converts checkpoint-level *selection noise* into inflation, at $\sim$7 pp in the programme's price-list analysis (`research/ADAPTIVE_EVAL_PROGRAMME.md`; the winner's-curse figures measured directly on the $\pi_{0.5}$ cell are $+4.00$ pp mean / $+12.5$ max, §4)—the quantitative case for §3's no-best-checkpoint rule. One diagnostic survives with no learning at all: saturation detection (Appendix E).

B.2 Are curated sets noisier to train on than random ones? A pre-registered test whose manipulation check failed

Every σ_{set} and set-draw price in §4 is computed over *random* draws—all 32 k-curve runs (§4.3) are uniform subsets by construction—while every arm of a real method comparison is a *curated* set. If curation itself destabilizes training, random-draw prices understate the replication a method comparison actually needs, and the penalty enters the budget roughly as the square of the sd ratio. A zero-GPU split of the existing k=22 replicate pairs by set-construction class said exactly that: $\sigma_{\text{run}} = \mathbf{3.66}$ pp across 6 random pairs ($\text{df} = 6$) vs $\mathbf{9.91}$ pp across 8 curated pairs ($\text{df} = 8$; coverage / heuristic / scoring extremes)—$F = 7.34$, sd ratio $2.71\times$, recorded as motivation, not measurement (`research/CURATED88_PREREG.md`, frozen 2026-08-05; the same heteroscedasticity §5 reports as "composition extremes destabilize training," now split by class). CURATED-88 tested whether that penalty survives at larger set size: **12 sets $\times$ 4 training seeds, fully crossed = 48 runs** in the k-curve's own cell ($k = 88$ from the 10-task pool $P = 454$; 20K steps, final checkpoint, 200 shared-seed eval episodes)—six curated sets from three rule families in both directions, six within-task uniform random, all twelve task-stratified to identical per-task counts and overlap-matched (16.1 vs 16.6 of 88). All 48 runs follow the ecosystem-default truncated LR schedule (20k steps of a 30k cosine decay); a later pre-registered completion experiment measured the completion effect as null for this class on `libero_object` (annealed per-task σ 6.936 vs control 7.275, $\Delta = -0.34$ pp, deep inside both CIs; both values are lower bounds under the measured sub-binomial dispersion $\hat{c} = 0.588$, and the arms are not episode-paired; Appendix K.1). Hypotheses, thresholds, and a manipulation check were frozen before launch; 48/48 runs reached their pre-registered budget.

The primary result is indeterminate, and the pre-registered gate says it must stay that way. H1: curated $\sigma_{\text{run}} = \mathbf{3.19}$ pp vs random $\mathbf{2.63}$ pp ($\text{df} = 18$ each)—$F = 1.467$, sd ratio $\mathbf{1.21}\times$, 95% CI $[\mathbf{0.75}, \mathbf{1.95}]$. Confirmation required $F \geq 2.217$ *and* ratio ≥ 1.5; refutation required a CI upper bound <1.6; neither fired, and no default ships. More important, the **manipulation check failed**: H6 required the curated arms' set-mean spread to exceed the random arms' by $\geq 1.5\times$, and it reached $\mathbf{1.28}\times$ (6.01 vs 4.71 pp). By pre-registration, an H1 null under a failed H6 is **uninformative**—this wave is not evidence that curation carries no noise penalty, and we do not report it as one. The pre-

registered distribution-free co-primary fails the same way: **H2 FAIL**—per-set Cochran's Q across the four crossed seeds gives curated median 3.73 against random 4.28; the frozen one-sided Mann–Whitney U (curated $>$ random, exact, 6 vs 6) gives $U = 23$, $p = 0.24 \geq 0.05$, the pre-registered falsification branch. H2 was frozen precisely to be immune to the mixture objection that voids Gaussian σ comparisons, and under the failed manipulation check its null carries the same gate as H1's. (A secondary decomposition also failed: curated systematic between-set component $1.61\times$ random against a required ≥ 2.0—**H3 FAIL**; Appendix E.) The seven frozen hypotheses, in one place:

	hypothesis (frozen)	threshold	scored verdict
H1	curated σ_{run} exceeds random	$F \geq 2.217$ and ratio ≥ 1.5	indeterminate ($1.21\times$ [0.75, 1.95]); gated by H6
H2	distribution-free co-primary (Cochran Q, MW-U)	one-sided $p < 0.05$	FAIL ($p = 0.24$); null gated by H6
H3	systematic between-set component (secondary)	$\geq 2.0\times$ random	FAIL ($1.61\times$)
H4	item flip floor class-invariant	ratio $\in [0.8, 1.25]$	PASS ($1.05\times$)
H5	random arm reproduces its cell	$\sigma_{\text{run}} \in [2.2, 5.0]$, mean $\in [48, 67]$	PASS (2.63; 63.1)
H6	manipulation check: curated set-mean spread	$\geq 1.5\times$ random	FAIL ($1.28\times$)
H7	no seed main effect	—	PASS ($F(3, 44) = 0.219$)

What did resolve. H4 PASS—the item-level flip floor is class-invariant: under the shared-seed evaluation, 22.9% of eval items flip outcome between identical retrainings of curated sets vs 20.8% for random, σ_{flip} ratio $1.05\times$, inside the frozen $[0.8, 1.25]$ band and replicating the k=22 cell. (Both rates are same-index cross-run LIBERO statistics and carry the pointer caveat of Appendix B.1; the curated-vs-random *comparison* is within-convention and unaffected.) Whatever excess *set-level* variance a curated set carries is therefore systematic at the suite level—attached to which items fail rather than to the rollout draw—and more evaluation episodes cannot buy it down; only retraining replication touches it. At *item* level the same reading does not hold: the majority of single-episode flipping is the policy's own sampler ($\Phi = 0.633$, a lower bound; Appendix K.8), so no individual flip should be read as a seed effect. H5 and H7 also PASS—the random arm reproduces the cell it copies, and the fully crossed design shows no seed main effect (details: Appendix E).

The manipulation failure is our own design flaw, and we record it as such. Stratifying selection *within* task held task coverage—the dominant success-rate driver in this pool—identical across all twelve sets by construction; the constraint that matched the classes also engineered away the extremity the curation rules were supposed to produce (full narrative: Appendix E). The fix is a stronger manipulation, not a larger sample (CURATED-STRONG; future work, not yet designed). Until that wave runs, whether curated arms pay a σ_{run} penalty at larger set sizes is an **open, pre-registered question**: the k=22 observation ($2.71\times$) stands as motivation, not measurement, and none of §4's random-draw prices should be assumed to cover curated arms.

C CORPUS COMPOSITION

The atlas—the frozen null-distribution corpus everything in §§4–5.2 is judged against—is given in Table A5.

The scope of §7, enumerated. The corpus covers **six single-task settings**—PushT, task 0 of all four LIBERO suites, and one held-out `libero_object` task—plus, since the 2026-08 waves, the **10-task `libero_object` pool** ($P = 454$) as a multi-task cell. Class-level scopes (for §7): the noise-floor and quantity breadth results are SmolVLA-fine-tune only, and DP-from-scratch claims stay scoped to PushT and LIBERO-object; PushT's demonstrations are scripted, not human teleoperated. Within the census these give nine task $\times$ policy-class cells over three policy classes: from-scratch DP, fine-tuned SmolVLA, and one 16-run fine-tuned $\pi_{0.5}$ cell (8 initial runs plus the 8-run PI05-PAIRS expansion). The post-freeze waves extend this scope with the **ACT class** (PushT plus two ALOHA bimanual tasks) and the ten-task `libero_spatial` pool (P=432), giving the **eight single-task settings** and four policy classes §7 enumerates (Appendices J and K).

The three addendum waves—run after the atlas was frozen, each judged against its own frozen pre-registration—add 103 runs: the **Phase-1c repair wave** (8; PushT/DP at full budget on the rebuilt eval stack—4 stack-calibration sets plus the 2+2 covmax/covmin replication), **Addendum A** (9; SmolVLA fine-tune on held-out `libero_object` task 1, the corpus's only fully prospective cell—4 random k=22 sets, one with 3 seeds, 2 full-pool runs, 1 coverage-max set; predictions frozen and hash-mirrored before launch, 0 crashes, all realised budgets verified), and the **paired-design**

Table A5: The frozen 414-run atlas, by task × policy-class cell.

cell	policies	waves
PushT, DP from scratch	251	200 datamodel random subsets; 22 designed arms + noise pairs (Phase 1c); 17 budget-doubling + driver A/B replicates; 8 set-diversity draws (Phase 1d); 4 MI-selection arms (Phase 4C)
LIBERO-object, DP from scratch	78	58 (Phase 2); 20 σ_{run}-scaling + full-data anchor (Tier 1)
LIBERO-object, SmolVLA fine-tune	52	24 (Tier 2); 8 embedding-coverage + heuristic (Tier 2b); 8 set-diversity draws (Tier 2c); 4 MI arms (Phase 4C); 8 MI de-confounding draws (Phase 4D)
LIBERO goal / spatial / 10, SmolVLA fine-tune	33	Tier 3 breadth

programme (86 = 46 anchor/rung/replicate + 6 floor-replicate runs on PushT/DP, and 34 SmolVLA task-1 runs: 10 bases × 3 rungs plus replicate and reseed arms). Addendum records carry the same per-run schema and are released alongside the atlas, flagged by wave; their absolute success rates obey the same in-wave-only rule—in particular the repair wave's rebuilt-stack PushT numbers are not poolable with the atlas's.

The **three 2026-08 session waves** add 88 runs: 8 $\pi_{0.5}$ fine-tune runs on `libero_object` task 0 (design frozen before results); 32 k-curve runs at $k \in \{22, 44, 88, 176\}$ drawn from the 10-task `libero_object` pool ($P = 454$), pre-registered (§4.3); and the 48 curated-vs-random runs at $k = 88$ in that same cell, pre-registered (Appendix B.2). All 88/88 reached their budget gate. Two **scored pre-registered follow-up waves** add 16 more, both with zero failures: the 8-run K4B $k = 176$ σ_{run} replication (§4.3), which takes the corpus $605 \to 613$, and the 8-run PI05-PAIRS $\pi_{0.5}$ seed expansion (§4), which brings it to 621. The corpus ladder is therefore $517 \to 605 \to 613 \to 621$: frozen atlas plus addenda, then the session waves, then each scored follow-up in turn.

Census boundary. The 621 count closes with PI05-PAIRS. Later pre-registered waves — reported in Appendices J and K, each against its own frozen pre-registration—add trained policies that are *not* counted in the 621: the 24 runs of Appendix J (PI05-CHAR 8, PI05-SCALE W1 8, W2 8) plus the 8 PERTASK-2 retrains reported there; then, in Appendix K, 8 (ACT on PushT, Leg B1), 8 (SmolVLA Leg A′), 16 (ACT on ALOHA, B2), 8 (ANNEAL-1), and 8 (PERTASK-4). Weight-averaged soups (Appendix K.7) are constructed from existing runs, not trained; ACT-HORIZON and REPEATEVAL-ZERO train nothing (re-evaluation and re-analysis of existing checkpoints). The 2026-08-19 to 09-06 waves of Appendix L train 112 small Meta-World networks and 8 LoRA adapters outside this corpus and its gates; none is counted in the 621, the 80 or the 701.

A harness parameter that is itself a reporting hazard (flagged in §3). LeRobot's `-eval.n_episodes` is **per task**, not total—`-env.task_ids=[0,1,2]` with `-eval.n_episodes=30` executes 90 episodes (measured 2026-08-04)—so our multi-task evaluations specify 20 episodes × 10 `task_ids` = 200 total, and any multi-task LIBERO result reported as "N evaluation episodes" under this harness in fact ran $N \times n_{\text{tasks}}$.

Each record carries the exact episode mask, the training configuration, the measured success rate, and—where the logs were retained—the full training-loss curve (used for the probe assays). In the per-task waves the 100–200 episodes *per task* amount to 1,000–2,000 evaluation episodes per policy. For **334 of the 621 census runs** —not of the 701-policy programme total of §6, against which 334 would be the wrong fraction—the record also carries the **per-episode success vector** harvested from the evaluation logs: **246 of the 414 atlas rows** (the 168 datamodel-PushT vectors were not retained) **plus all 88 2026-08 session runs**; the 103 addendum runs and the 16 K4B/PI05-PAIRS follow-up runs are outside the harvest. Every harvested vector reproduces its run's recorded success rate exactly (0 mismatches in 334), an independent integrity check on both the corpus and the item-level dataset behind Appendix B. (This 334 is a *harvest* count over the 621-run census. It is distinct in role from the gate-replay census of §6—334 runs replayed, 193 null pairs, 386 ordered gate calls—and neither count is the "382 controlled runs" programme figure.)

What that check does not reach, said plainly. The check is scoped to the 621-run census, and the 80 scored post-freeze runs sit outside that census by construction (*Census boundary*, above). So the **flagship per-task figure sits outside the 0-mismatch check**: $\sigma_{\text{pertask}} = 11.00$ pp for $\pi_{0.5}$-ft on multi-task `libero_object` at k=88 (PERTASK-2, §4.1) comes from 8 retrains that post-date the

harvest, as do every ACT row, Leg A′, ANNEAL-1 and PERTASK-4 value. Those waves are not unchecked—they carry their own integrity gates, quoted with their results rather than borrowed from this one: PERTASK recomputed 16/16 per-run success rates identical to the atlas and reproduced both waves' suite σ_{run} through the same grouping (Appendix J), ANNEAL-1 passed 8/8 gates with its one provenance defect disclosed rather than repaired, and REPEATEVAL-ZERO's alignment gate ($\geq$48/50) was passed by every one of SHIM-1's parent evaluation chunks (Appendix K). The honest reading of "334, zero mismatches" is therefore an integrity check on the *census corpus and the item-level dataset*, not on this paper's headline number, and we do not quote it as one. Two earlier legs on a prior harness are reported in the paper but held separate from this corpus and its absolute success rates: the drop-gate experiments (30 models, PushT/DP and ALOHA/ACT) and the leave-group-out decile study (99 cells).

D PRE-REGISTRATION REGISTRY

Seventeen pre-registration documents govern the atlas, addendum, and 2026-08 session waves, each frozen and timestamped to cloud storage *before* its wave launched, with numeric verdict criteria; the post-freeze waves add their own frozen documents, enumerated at the end of this section. *Thirteen for the atlas and addendum waves* (Phase 1c, 1d, 2, 2b, 2c, Tier 2b, 2c, 3, Phase 4, 4D, Phase-1c repair, Addendum A, and the paired programme—the last frozen with its analyzer hashed in `PAIRED.sha256` before any run existed, plus one declared amendment for the extra floor replicates, `AMENDMENT_FLOOR.md`): `PHASE1C_PREREG.md`, `PHASE1D_PREREG.md`, `PHASE2_PREREG_DRAFT.md`, `PHASE2B_PREREG.md`, `PHASE2C_PREREG.md`, `TIER2B_PREREG.md`, `TIER2C_PREREG.md`, `TIER3_PREREG.md`, `PHASE4_PREREG.md`, `PHASE4D_PREREG.md`, `forecast/PHASE1C_REPAIR_PREREG.md` (sha256 `767f7119...`), `forecast/ADDENDUM_A_PREREG.md`, `paired/PAIRED_PREREG.md`. *Four for the 2026-08 session and its follow-ups*: `research/KCURVE_PREREG.md`, `research/CURATED88_PREREG.md`, `research/K4B_PREREG.md` (frozen decision bands), and `research/PI05_PAIRS_PREREG.md` (sha256 `be2147c6...`, mirrored to cloud storage *before* mask generation).

Four census-era outcomes contradicted our own predictions and are reported as frozen: Phase 2's method validation FAIL; Tier-2's H1 missing its threshold by 0.11 pp; Tier-3's quantity prediction failing 0/3 (verdict frozen pre-repair; the repaired numbers change one suite's sign, Appendix E); and the repair wave's new-stack mean (14.8) landing 0.2 pp below its frozen $[15, 28]$ band.

Addendum A is the corpus's only genuinely blind *prediction* wave—all its quantitative predictions frozen and hash-mirrored to cloud storage *before any run existed*—and it cut both ways: its frozen calibration floor and quantity-transfer tolerance passed (7/9 within interval; $+26.5$ vs $+30.6 \pm 11.7$), while its fresh replicate measurements overturned our fitted variance components in both directions (§4).

The 2026-08 waves extend that ledger. The **k-curve** wave (§4.3) adds two scored failures: **K3 FAIL** on its committed pure reading ($\sigma_{\text{set}}(\text{pure}) = 2.43 \leq 3$ pp at $k = 176$)—subsequently **flipped to the PASS side** (3.86 pp) by the frozen K4B recompute, and reported as a flip with both readings kept—and **K4 partial fail** (σ_{run} rose with k where flat-to-falling was predicted). Both landed opposite our declared prior interest—K3's failing branch is the one on which the paper's method-comparison price recedes at production data sizes—which is why the follow-up wave settling $\sigma_{\text{run}}(k = 176)$ froze its decision bands before its data existed (`research/K4B_PREREG.md`, scored in `research/K4B_RESULTS.md`). **CURATED-88** (Appendix B.2) scores H4/H5/H7 PASS and **H2/H3/H6 FAIL**, with H1 indeterminate and gated *uninformative* by the failed manipulation check H6 (curated set-mean spread $1.28\times$ random against a required $1.5\times$); H2, the distribution-free co-primary, fell on its frozen falsification branch (one-sided $p = 0.24$) and its null carries the same gate. **PI05-PAIRS** (§4), frozen before mask generation, falsified this paper's own three-policy ladder claim—σ_{run} for the $\pi_{0.5}$ fine-tune re-measured at 9.74 pp ($\text{df} = 10$) against the 2.02 pp ($\text{df} = 2$) it replaces—and is reported as an overturn with both values.

The post-freeze waves extend the registry further. The pre-registrations for PI05-SCALE, PI05-CHAR, PERTASK, and the G3 pairing audit are described with their waves (Appendix J); PERTASK-2's was frozen 2026-08-13 (sha256 `b7163823...`) before any of its masks existed. Six further frozen documents govern the 2026-08-16 to 08-21

waves (Appendix K), quoted here by sha256 prefix: `research/PERTASK3_PREREG.md` (`22ff5c67...`, with the declared `PERTASK3_AMENDMENT_A.md`, `8088ab5b...`, authorized by the study lead and frozen before its training wave); `research/ACTHORIZON_PREREG.md` (`a4744452...`, frozen before any episode was scored); `research/ANNEAL1_PREREG.md`, whose amendments form a **hash chain**—v1 `be169823` → v2 `65871332` → v3 `6cc05f82` → v4 `8e103e86`, governing sha `7b22044e`, plus two corrections of record (§12.8/§12.9) written mid-battery with the primary endpoints unscored, every version retained; `research/PERTASK4_PREREG.md` v2 (`50adacdc...`; v1 `13abe2c2` retained, superseded before any mask existed); `research/SOUP1_PREREG.md` v2 (`4c23f9ef...`; v1 `8048aca7` retained, Amendment A frozen at 0/30 evaluation chunks banked); and `research/REPEATEVAL_ZERO_PREREG.md` (`647d1550...`, frozen before the pooled statistic existed, with one disclosed feasibility peek). The waves of 2026-08-19 to 2026-09-06 add sixteen further frozen documents, fifteen pre-registrations and one declaration, quoted by sha256 prefix in Appendix L.9.

D.1 THE k-CURVE'S SCORED VERDICTS, IN FULL

This subsection carries the full scored verdicts behind §4.3; the numbers are unchanged from the main text.

Why the draws come from the 10-task pool. Draws are uniform from the full `libero_object` pool $P = 454$, and the pool choice is load-bearing: task 0's own pool is $P = 44$, so drawing $k = 44$ from it forces 100% overlap and $\sigma_{\text{set}} = 0$ *by construction*—a single-task k-curve would rig its own answer.

K4B, in detail. The four fresh $k = 176$ sets gave pair differences $+4.0$, $+2.0$, -5.5, $+2.0$ pp; the seed-effect guard did not trigger because the signs are mixed, which is what licenses pooling to $\sigma_{\text{run}} = 3.36$ ($\text{df} = 6$). A 16-run unbalanced ANOVA over all ten $k = 176$ sets agrees (σ_{run} 3.36, $\sigma_{\text{set}}(\text{pure})$ 3.53). The monotone σ_{run}-rise reading of K4 is weakened—the new independent estimate sits below the $k = 44$ and $k = 88$ values—but, per band B3, not settled; the K4 caveat below stands.

The two floors change places along the curve. At $k = 22$, $\sigma_{\text{set}}(\text{pure})$ dominates σ_{run} by $7.4\times$ (12.47 vs 1.68 pp); by $k = 88$ run noise dominates. Which of the two instruments a design needs is therefore itself k-dependent.

- **K1 ($\sigma_{\text{set}}(\text{total})$ monotone non-increasing): PASS.** The $k = 88 \to 176$ uptick of $+0.78$ pp is inside 1 SE (1.62 pp at $\text{df} = 5$).
- **K2 (training collapses concentrate at small k): PASS, directional only.** Collapse frequency 1/6 at $k = 22$, 0/6 at each larger k; six draws cannot resolve a rate below $\sim 15\%$, so this is evidence of direction, not a rate estimate. It ties small-k σ_{set} to occasional collapse rather than a smooth spread.
- **K3 ($\sigma_{\text{set}} \geq 3$ pp at $k = 176$): FAIL on the committed reading**— $\sigma_{\text{set}}(\text{pure}) = 2.43 \leq 3$ pp. It passes on the overlap-corrected (3.08) and total (5.12) readings; an interpretation note committing to the pure reading was recorded 2026-08-04 15:30 UTC—after the $k = 176$ wave had run but before the other three—so the commitment could not be tuned to the full outcome, and the prereg's pure-vs-total ambiguity is a recorded wording failure. We score the committed reading, and it fails; the pre-registered K4B follow-up then flipped it (3.86, §4.3), and both readings are carried. K3 was pre-registered as the least-wanted branch: the one on which the method-comparison price this paper leads with recedes at production data sizes.
- **K4 (σ_{run} flat-to-falling with k, ≤ 5 pp): PARTIAL FAIL.** σ_{run} stays ≤ 5 pp (max 4.51) but *rises* monotonically—$1.68 \to 3.40 \to 3.35 \to 4.51$, each $\text{df} = 2$—opposite the prediction and opposite the from-scratch trend (DP: $11.43 \to 6.24$ across $k = 22 \to 44$). We have no explanation, and this is the weakest link in the decay claim: $\sigma_{\text{set}}(\text{pure})^2 = \sigma_{\text{set}}(\text{total})^2 - \sigma_{\text{run}}^2$, so an over-estimated σ_{run} at $k = 176$ manufactures part of the fall. The $k = 44$ pair's differences are same-signed (-5.50, -4.00; probability 0.25 under the null), so these data alone do not exclude a seed main effect.

Two measurement limits bound the small-σ_{set} end of the curve. Overlap rises with k by construction (expected $4.9\% \to 38.8\%$; measured $4.1\% \to 37.9\%$), suppressing σ_{set} exactly where the decay

appears; the correction applied is first-order, the prereg's fuller variance-reduction model was not fitted, and the $k = 176$ point should be read as a lower bound. And a naive binomial floor at $n_{\text{eval}} = 200$ would be $\sim$3.3–3.5 pp—*above* the measured $\sigma_{\text{set}}(\text{pure})$ at $k = 88$ and 176; this is coherent only because the shared-seed evaluation walks the same fixed init-state pool in every run, so there is no independent state draw between runs (§4.2; episode-*level* cross-run pairing is a separate, stronger property that LIBERO does not deliver, Appendix K.1), and it puts those two points near the resolution limit of this evaluation size. Within those limits, the overlap-corrected σ_{set} plateaus at 3.05 ($k = 88$) and 3.08 ($k = 176$): a $\sim$3 pp residual that more data does not remove—genuine floor or correction artifact, four points cannot say.

E AUDIT AND REPAIR NARRATIVES

E.1 THE BUDGET-TRUNCATION CONFOUND (REPORTING STANDARD 2)

Our training driver derived each run's step count from the highest-numbered checkpoint directory and evaluated it *without checking the trainer's exit status*, so a crashed run silently produced a complete-looking result that the skip-done guard then treated as finished. This corrupted 12 of the 414 atlas runs—including 8 of 33 in one wave, where it inflated a noise-floor estimate by an order of magnitude and manufactured an apparent effect in the same direction as our pre-registered prediction. It also inflated our headline PushT coverage gap: as first computed, $+16.9$ pp at 100K—but the covmin arm included a seed-0 run that had silently trained 25K of its pre-registered 100K steps and scored 17.4 as a result. Retiring it gives the $+11.1$ pp we report (§5.2). Budget truncation is *not* mean-zero noise: an undertrained model scores low, so truncation masquerades as whichever effect predicts a lower score. We caught it only because a later audit reconciled every run's realised step count against its design. We expect this failure mode is widespread and invisible in published work, since almost nobody reports realised budgets per run. The 103 addendum runs all trained under the budget-assertion gate: 103/103 reached their pre-registered step count, and the paired programme's gate—mandatory from launch—recovered five cold-start OOMs rather than silently truncating them (86/86 accepted, zero silent truncations).

The breadth-wave repair in detail. 8 of the 33 runs in our breadth wave were silently truncated by a shared-memory crash. Before repair we reported $\sigma_{\text{ft}} = 3.19$ pp and "quantity reverses on all three new suites" ($-3.4/-8.4/-14.1$); after re-running all 8 to the full budget, $\sigma_{\text{ft}} = 2.82$ pp (the noise-collapse claim survives) but the quantity result does not: `libero_spatial` **flips sign** from -8.4 to $+3.8$, because both of its full-pool runs had been 5K-step models. The contamination pushed every affected number in the direction of our own pre-registered prediction—an undertrained model scores low, and a low-scoring full-pool arm is exactly what "quantity reverses" predicts. We report the sequence rather than only the corrected number because the failure mode is systematic, not incidental.

E.2 THE EVAL-STACK SHIFT, MEASURED

Absolute success rates are not comparable across our two PushT evaluation stacks—and we have now measured the gap. A pre-registered calibration wave re-ran four datamodel sets with known old-stack scores on the rebuilt `gym_pusht`: the same episode sets lost a mean **18.2 pp** of absolute success (range -10.6 to -24.4, sd 7.05) purely from the environment rebuild. Our frozen prediction was that the new-stack mean would land in $[15, 28]$; it came in at 14.8—0.2 pp below the band. The shift is larger than we predicted, and we report the prediction as a miss. It is also not a uniform offset: higher-scoring sets lost more, so rank order is only partly preserved, and we accordingly apply no cross-stack correction anywhere in the paper—Phase-1c and Phase-1d PushT numbers remain within-stack comparisons, as frozen.

The repair wave's own numbers (behind §5.2): 2 seeds per arm at 100K steps; gap $+8.2$ pp inside the frozen $[+5, +20]$ band, with complete arm separation (min covmax $20.6 >$ max covmin 14.2); arm means 20.6 and 12.4 against 29–40% pre-rebuild, the lower-SR regime behind the never-compare rule.

The same wave closed an integrity flag: its covmax seed pair returned *identical* SRs (20.6/20.6), which an automated check raised as a possible seeding-independence bug. Investigated and refuted—the two runs have clearly different reward trajectories and merely tied in success count, and with SR

quantized to 0.2 pp at $n = 500$ eval episodes an exact tie is an ordinary coincidence; the matching concern previously raised against Phase-1d's pairs is downgraded likewise. The tie is also consistent with the cell's resolved noise floor: pooling all four post-rebuild replicate pairs gives $\sigma_{\text{run}} = 1.292$ pp ($\text{df} = 4$), *below* the cell's own binomial floor of 1.682 pp (mean SR 17.05%, $n = 500$)—no retrain variance is detectable above sampling in this regime, so an exact replicate tie is unremarkable (Table A10; Appendix K.6). The repair wave's own two-pair reading is 1.80 pp at $\text{df} = 2$ (one gap exactly 0.00; χ^2 95% CI $[0.94, 11.31]$); 1.292 pooled at $\text{df} = 4$ is the cell's canonical value.

E.3 The historical-reference defect in the first paired design

A pre-launch review found that the historical unpaired reference that our first paired design used would have manufactured a false PASS 62% of the time at the measured parameters. The defect is recorded, not quietly fixed, in the design document's §5.1; the launched design replaced it with the in-wave unpaired reference of §4.2 (anchors and rungs from different bases are marginally independent uniform k-subsets, so Leg 1's 46 runs supply 819 unpaired pairs at $\bar{d}$ from runs the wave was doing anyway).

E.4 LeRobot evaluation shares its seed; what that cancels, and where

Every run in the corpus is scored under the same evaluation seed, because `EvalPipelineConfig.seed` defaults to 1000 in lerobot 0.6.0 and no driver in our repository overrides it. What that shared seed *buys* depends on the environment, and the scope was measured, not assumed (Appendix K.1). On PushT, identically-ordered runs replay identical initial states, so the independent binomial draw between two runs cancels and their difference carries no binomial floor—which is why replicate pairs can legitimately differ by 0.32 pp when a naive binomial floor says 1.69 pp, a pattern we had earlier mis-flagged as an integrity failure. On LIBERO the shared seed does *not* establish cross-run episode pairing: the init-state pointer advances with the episode's *outcome* (Appendix K.1), so two runs desynchronise as soon as their outcomes diverge, and only 10.06% of same-index cross-run pairs share an initial condition. All LIBERO-side analyses therefore retain binomial floors and subtract them explicitly. Any analysis of LeRobot-derived success rates can verify the shared seed in one line, and should—before subtracting a binomial term that is not there on aligned PushT pairs, flagging sub-binomial agreement as fraud, or assuming episode-level pairing on LIBERO that the harness does not deliver. Two scope notes carried from the main text: on PushT, rollouts are a function of (init state, *position in the episode sequence*), so pairing holds between identically-ordered runs but not across differing orderings or episode counts; and PushT/ACT evaluations are not bit-reproducible across batch configurations (the training-side bit-determinism result of §4.2 is SmolVLA-ft training only).

E.5 A frozen forecasting exercise that did not validate

Before the repair wave we tested whether pre-run features alone (policy class, env, suite, task, k, steps, arm class) could forecast the distribution of a run's final SR, with acceptance criteria frozen first: they could not (80% intervals covered 20/57 held-out runs against an acceptance band of $[39, 51]$; NOT VALIDATED). Its single worst cluster was PushT: 0 of 12 runs inside their intervals. Applying the later-measured -18.2 pp stack shift post hoc puts 11 of 12 inside. The frozen verdict is unchanged—the shift was measured after the freeze, and this is a diagnostic, not a re-judgement—but the diagnosis relocates the failure: the PushT collapse was a measurement-stack artifact, not a forecasting failure, and the residual failure is forecasting new task cells on LIBERO.

E.6 The $\pi_{0.5}$ cell: scope notes and budget validation

The $\pi_{0.5}$ comparison of §4 (checkpoint `pi05_base`, 4B, full fine-tune; the cross-class ladder's other rungs are DP from scratch at 263M and the SmolVLA fine-tune at 450M) is **variance-structure only**—the protocol forbids cross-regime absolute-SR comparison (§3)—and three scope notes qualify it. *Sample matching:* the cell is not sample-matched, at 320K training samples against SmolVLA's 640K at equal steps; matching them would cost 47 h per run. *Normalization statistics:* $\pi_{0.5}$ requires q01/q99 quantile statistics that the public dataset does not ship, so we recomputed

them on a private copy; recomputation shifted image mean/std by $\sim$1–3% while the action and state statistics and the shared episode cache stayed *byte-identical*, which is what keeps the existing corpus comparable. *Budget:* the 20K-step budget was validated within-cell—mean $\Delta\mathrm{SR}$ was $+2.19$ pp from 10K$\rightarrow$20K and -0.50 pp from 15K$\rightarrow$20K, with 3 of 8 first-wave runs still marginally rising, so 20K sits inside $\pi_{0.5}$'s dynamic range and the pre-registered re-run trigger did not fire. All 16 runs of the expanded cell reached `final_step` $= 20000$.

Where the superseded 2.02 came from (provenance for §4). The 2.02 pp reading came from two seed pairs whose differences happened to be -0.5 and -4.0 pp—an optimistic low-df draw, and the second low-df σ in this corpus to move upward on re-measurement (Appendix H). Our earlier guidance for this cell—report the ladder and not the decimal—is therefore **retired**: the ladder itself was the artifact. In the expanded 16-run cell the seed main effect is not significant ($F(3,3) = 2.90$), which is what licenses pooling across the six masks; the 9.74 headline is computed by within-mask one-way ANOVA over the 16 runs.

The single-collapse structure of the withdrawn 10.76. The cell's four independent same-size draws scored $71.0/72.0/50.0/71.5$; the naive sd of 10.76 pp ($\mathrm{df} = 3$) is one collapse and nothing else—leave-one-out sds are 12.27–12.56 pp dropping any healthy draw versus $\mathbf{0.50}$ pp dropping the 50.0 draw. Checkpoint curves identify the mechanism: the collapsed run *learned and then degraded* ($62.5 \rightarrow 55.5 \rightarrow 48.0 \rightarrow 50.0$ over 5K/10K/15K/20K steps) while seven of eight rose to a plateau. The pre-registered seed expansion then attributed that collapse to the retrain (the same mask at a second seed scored 87.5), which is why $\sigma_{\mathrm{set}}(k{=}22, \pi_{0.5}) = 0.00$ by method of moments—between-mask variance does not exceed what $\sigma_{\mathrm{run}} = 9.74$ predicts, and at this noise level set identity is unresolvable from six masks. *Withdrawn:* the cross-class ratio an earlier version of §4 quoted ($10.76/2.02 = 5.3\times$ against Addendum A's $19.01/3.51 = 5.4\times$) is withdrawn—its $\pi_{0.5}$ numerator was retrain noise and its denominator an optimistic $\mathrm{df} = 2$ draw.

E.7 Adaptive evaluation: the full item-structure numbers

These are the measurements Appendix B summarizes (`research/ADAPTIVE_EVAL_PROGRAMME.md`).

Over-dispersion. Per-item success rates over-disperse relative to an equal-difficulty binomial by $3.09\times$ (DP/PushT: 500 items, 83 runs), $2.35\times$ (SmolVLA-ft/`libero_object`), $1.44\times$ (DP/`libero_object`), 1.45–$1.51\times$ (SmolVLA-ft on `libero_10` and `libero_spatial`), and $0.95\times$ on saturated `libero_goal` (mean $p = 0.925$). Between 2.0% and 22.6% of items are near-uninformative ($|\text{point-biserial } r| < 0.10$).

Discordant-item rates. For deciding which of two checkpoints is better, an item-selection scheme must raise the discordant-item rate—the fraction of selected items on which the pair actually disagrees. Realised rates for random / discrimination-ranked / two-stage-adaptive selection: $41.1/40.8/40.8\%$ on DP/`libero_object` (oracle 98.6%); $39.2/44.4/40.3\%$ on DP/PushT (oracle 100.0); $45.7/46.2/45.2\%$ on SmolVLA-ft/`libero_object` (oracle 100.0). The $\sim 2.5\times$ headroom is real and no implementable scheme captures any of it.

The in-sample trap, reproduced on ourselves. Selecting items by in-sample $|r|$ claims a $\sim 4\times$ budget saving (rank recovery $+0.943$ at $m = 50$ evaluation episodes vs $+0.767$ for random, DP/PushT); cross-validated, the saving deflates to $\sim 2\times$ on DP/PushT ($+0.807$) and to nothing on LIBERO—at or below random at every m on DP/`libero_object`, and $+0.956$ against random's $+0.976$ at $m = 100$ on SmolVLA-ft/`libero_object`. We report the deflated numbers.

Saturation detection, the one diagnostic that survives with no learning at all: on `libero_goal` ($p = 0.925$), 40.5% of items pass or fail identically in every run, against 0.0–0.5% dead items on `libero_object`.

Every item-level flip statistic above carries the Φ stamp. REPEATEVAL-ZERO (Appendix K.8) measures $\Delta = f_{\mathrm{between}} - f_{\mathrm{within}} = \mathbf{0.0841}$ (exactly $6{,}416/28{,}000 - 580/4{,}000 = 0.08414$ from the unrounded counts; the results document prints 0.0849, which no rounding of its own two components produces—corrected here, in Appendix B.2 and in Appendix K.8, and recorded rather than silently harmonised) and $\Phi = 0.633$ (a lower bound): the majority of item-outcome flipping between identically-trained checkpoints is the policy's own stochastic rollout sampling, not the train-

ing seed. Aggregate quantities (σ_{run}, per-task σ) average the sampling term over 100+ episodes and are unaffected.

E.8 CURATED-88: THE COMPONENTS BEHIND THE SCORED VERDICTS

Supporting detail for Appendix B.2 (results logged in `SESSION_LOG_2026-08-08.md` §5).

Design. Twelve sets, six curated from three rule families in both directions (state-coverage min/max, quality-heuristic clean/messy, episode-length short/long) and six within-task uniform random, all task-stratified to identical per-task counts (9 from each of tasks 0–7, 8 from tasks 8–9)—selection at 9 of $\sim$45 episodes $= 20\%$ per task—and overlap-matched at 16.1/88 within-class overlap for curated against 16.6/88 for random, so no overlap artifact.

H3 FAIL (secondary decomposition). The systematic between-set component is 2.11 for curated against 1.31 for random, a ratio of $1.61\times$ against a required ≥ 2.0.

H2 FAIL (distribution-free co-primary). Per-set Cochran's Q across the four crossed seeds on the 200 shared-seed items ranges over 2.14–11.14 for curated (median 3.73) against 1.39–8.10 for random (median 4.28); the frozen one-sided Mann–Whitney U (curated $>$ random, exact, 6 vs 6) gives $U = 23$, $p = 0.24$. Scored zero-GPU from the per-episode harvest (`experiments/forecast/score_curated88_h2.py`, log `CURATED88_H2_SCORING.txt`).

H4 PASS replicates the $k = 22$ cell's flip floor (2.80 vs 3.02 there). *H5 PASS:* the random arm reproduces the cell it copies—$\sigma_{\text{run}} = 2.63$ inside the frozen $[2.2, 5.0]$, grand mean 63.1 inside $[48, 67]$, against the k-curve's $k = 88$ reference of 3.35 pp at $\text{df} = 2$. *H7 PASS—no seed main effect:* seed means $62.75/63.12/63.67/64.54$, $F(3, 44) = 0.219$; the crossed-seed design is clean, and the same-signed $k = 44$ replicate pair of §4.3 is retired as chance.

Why the manipulation failed. Stratifying selection *within* task held task coverage—the dominant success-rate driver in this pool—identical across all twelve sets by construction, so the constraint that matched the classes also engineered away the extremity the curation rules were supposed to produce, while random $k = 88$ set means already spread 12.7 pp (56.88 to 69.62)—a moving target the within-task rules could not beat. The fix is a stronger manipulation, not a larger sample: unstratified selection over the whole pool, accepting task coverage as a confound to be modelled (CURATED-STRONG; future work, not yet designed).

F EXTENDED LIMITATIONS

F.1 THE PAIRED-DESIGN ESTIMATOR AND ITS IN-WAVE UNPAIRED REFERENCE

This subsection carries the full derivation behind §4.2. For two size-k sets at edit distance d, decompose

$$\mathbb{E}\big[(\text{SR}_1 - \text{SR}_2)^2 \,\big|\, d\big] \;=\; 2\sigma^2_{\text{set}}\left[(1-w)\,\frac{d}{\bar{d}} + w\right] + 2\sigma^2_0, \tag{1}$$

where w is the share of between-set variance not explained by episode overlap ($w = 0$ is the field's implicit additive per-episode-value model; $w = 1$ means the outcome re-rolls on any change), $\bar{d} = k(1 - k/N)$ is the expected distance between independent draws, and σ_0 is the harness's own nondeterminism—identical episodes, identical seed, run twice—the floor no design can cancel. Anchor/rung bases at exact distances m give $\text{VRF}(m) = \text{Var}_U/\text{Var}_P(m)$ against a **free in-wave unpaired reference**: anchors and rungs from different bases are marginally independent uniform k-subsets, so Leg 1's 46 runs supply **819 unpaired pairs** at $\bar{d}$ from runs the wave was doing anyway (the defective historical reference our first design used is recorded, not quietly fixed, Appendix E). Under Equation 1's primary null $H_0\colon w = 1$ every designed run is an independent draw from one distribution, so relabelling runs is an exact randomisation null. Pre-launch simulation put both legs at 80% power against $w \lesssim 0.3$.

Verdict detail. The two legs' cells (behind §4.2): Leg 1 is PushT/DP from scratch at $k = 103$ of a 206-episode pool, 14 bases; Leg 2 is the SmolVLA fine-tune on `libero_object` task 1 at $k = 22$ of 45, 10 bases. The primary VRFs carry CIs **1.92** [1.00, 3.94] (Leg 1, $p = 0.040$) and **1.73**

[1.06, 2.97] (Leg 2, $p = 0.032$), plotted against the additive-model predictions in Figure A1(a). Leg 2's second rung also rejects (VRF 3.05 [1.20, 7.10], $p = 0.016$), and Leg 1's leave-one-base-out fragility is worst at VRF 1.45, $p = 0.148$. The most precise contrast in either leg is Leg 1's within-base $\mathrm{Var}(m{=}16)/\mathrm{Var}(m{=}4) = \mathbf{2.15}$ [1.35, 3.44] on the same 14 anchors, which **excludes 1.0** and **excludes 4.0**—sub-linear growth of between-set variance with edit distance. On the floor side, PushT/DP's $\sigma_0 = 1.91$ pp ($\mathrm{df} = 8$) supersedes an initial $\mathrm{df} = 2$ estimate of 1.17 that was unstable; the declared six-run amendment changed the conclusion. The $m = 4$ rung's mean squared gap of 7.17 pp^2 is indistinguishable from the floor (ratio 0.98 [0.39, 3.38]; Figure A1(b) decomposes that variance). The bit-determinism test is that both same-seed SmolVLA replicate pairs are bit-identical at all 21 logged loss steps, while the different-seed control diverges by step 500 and PushT's from-scratch same-seed control at step 14,000.

The additive predictions are pre-registered, and one of them is conditional on a floor the experiment then measured differently. 11.5 and 4.9 are the $w{=}0$ row of the power table frozen in `paired/PAIRED_PREREG.md` before any run existed, computed from Equation 1 by `paired/power_paired.py`. That code includes the $2\sigma_0^2$ term on both sides, exactly as the equation states; what it could not know were the *inputs*. Leg 1's prediction assumed $\sigma_0 = 0.32$ pp and $\sigma_{\mathrm{set}} = 3.2$ pp; the wave then measured $\sigma_0 = 1.91$ pp ($\mathrm{df} = 8$, itself superseding an unstable $\mathrm{df} = 2$ reading of 1.17) against a composition term of $2\sigma_{\mathrm{set}}^2 = 6.48$ pp^2. Re-evaluating the same expression at the measured inputs gives an additive prediction of $(6.48{+}7.30)/(6.48{\cdot}4/51.5{+}7.30) = \mathbf{1.77}\times$, against an observed 1.92 [1.00, 3.94]. **On Leg 1 the rejection of additivity therefore does not survive substituting the measured harness floor**: at $m{=}4$ that cell's paired variance has reached σ_0, which is the same fact the $m{=}4$ rung already reports (mean gap^2 7.17 pp^2, ratio to the floor 0.98 [0.39, 3.38]). Leg 2 moves the other way. Its frozen 4.9 assumed $\sigma_0 = 3.51$ pp; bit-deterministic fine-tuning measured $\sigma_0 = \mathbf{0.00}$ exactly, which *raises* the additive prediction to $\bar{d}/m = 5.62\times$, and the observed 1.73 [1.06, 2.97] excludes both readings. **The interventional rejection of the additive per-episode-value premise stands on Leg 2, at $k{=}22$ on `libero_object` task 1 under SmolVLA fine-tuning, and we withdraw it on Leg 1.** We report the frozen predictions as frozen, and this recomputation beside them, rather than re-deriving the pre-registration after the fact. The $p = 0.040$ and $p = 0.032$ attached to the two VRFs are randomisation tests of $H_0{:}\, w = 1$—no pairing benefit—not tests against the additive prediction, and Leg 1's is fragile to dropping one base (worst VRF 1.45, $p = 0.148$).

What pairing buys, priced. Measured MDEs at the primary rungs: **2.0 pp at 14 bases on PushT** (2.7 pp at 8); **13.2 pp at 14 bases on LIBERO** (17.5 pp at 8), where $\mathrm{Var}_P = 313$ pp^2 means a 10 pp increment effect needs 25 bases (50 runs) and a 5 pp effect 99. For full-set method comparison the ratio is $n_{\mathrm{paired}}/n_{\mathrm{unpaired}} = k^2/(m\,\bar{d}\,r^2) \approx 6.8\times$ (Leg 1, $m{=}4$) and $2.9\times$ (Leg 2, $m{=}2$), minimised by not pairing at all; in the from-scratch regime even the increment question is floor-limited.

F.2 Caveats carried from the main text

The head-to-head tests our implementation, not the published methods. We ran a DemInf-style MI ranker with raw features standing in for learned embeddings, and no code from CUPID, DemoSCORE, SCIZOR, RINSE, or Re-Mix. Interventionally, that ranker's top-k-vs-bottom-k gap is $+0.6$ pp on PushT (bar ± 4.1), and its LIBERO gap of $+19.5$ pp **fails set-diversity de-confounding** (Welch $p = 0.21$ against a 20-pp set-identity sd)—the exact failure mode single-set, few-seed evaluations cannot see. A reader is entitled to the reading that our reimplementation, not the method family, is what failed (the twelve-estimator sweep behind the scoring null is named estimator by estimator in Appendix F.3); the claim we defend is narrower—this scorer, given ground truth and an interventional test, predicts and moves nothing—plus the separate, purely bibliographic finding that the published evidence for the family cannot be checked against a noise floor. Running the original implementations is the obvious next experiment.

Scoping the quality heuristic's powered null (§5.1). Its $+1.25$ pp gap under the collapsed VLA noise floor is powered *as a comparison of the fixed sets tested*; a method-level reading would be bounded by set-identity dispersion, at this k a collapse mixture, which makes it an underpowered null like those below and cuts against any rescue of the heuristic, not for it.

The two tables erred in both directions, and we record it twice over (§4). Pre-correction they were optimistic for the method question; the first correction then modelled as a Gaussian sd what two inde-

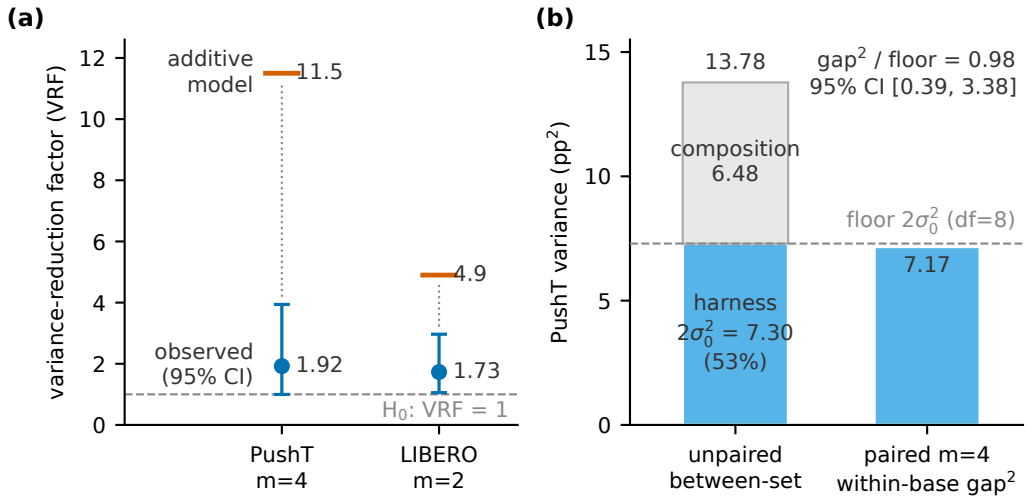

Figure A1: **The paired-design programme.** (a) Observed VRF (95% CI) per cell vs the *pre-registered* additive-model predictions (11.5 and 4.9) and H_0: VRF $= 1$. The additive per-episode-value model is rejected under fine-tuning, whose measured harness floor is exactly zero; on PushT the prediction is conditional on an assumed $\sigma_0 = 0.32$ pp that the wave then measured at 1.91, and the rejection there is withdrawn (Appendix F.1). (b) PushT variance decomposition: the harness floor $2\sigma_0^2 = 7.30\,\text{pp}^2$ (df 8) is 53% of the $13.78\,\text{pp}^2$ unpaired between-set variance; the m=4 within-base mean gap^2 ($7.17\,\text{pp}^2$) is indistinguishable from the floor (ratio 0.98 [0.39, 3.38]).

pendent cells show is a collapse process. Our earlier belief that these tables erred only conservatively was itself wrong.

An early small-n transfer read (§5). Across the first five shared sets, DP-SR and VLA-SR correlated at $+0.73$; at the full 12 shared sets the correlation is $+0.29$ (Spearman $+0.20$, n.s.). The interim value is recorded because it is itself a cautionary tale about small-n transfer claims.

The single-set trap, in the numbers (§5). In the frozen-prediction task-1 wave the coverage-max set that scored 81.5 was the *second-highest* result in the wave, on an arm whose fitted effect from task 0 was -5.4 pp. With $n = 1$ set a method cannot be separated from a draw: resolving a 10 pp method effect takes 56–63 independent draws per arm in this cell, and this is the Tier-2c lesson reproducing under predictions frozen before the runs existed. The accompanying heteroscedasticity—designed/extreme sets showing seed-pair spreads of 10–21 pp against a 4.7 pp null-pair expectation—is seen under from-scratch DP as well as under fine-tuning.

Anchors from directly measured σ_{run} (behind §4): at $\sigma_{\text{run}} = 11.4$, a 3-seed/arm comparison detects only $\geq$26 pp effects (retro-power against our own $+12$ pp observed gap: 26%), and percentile-vs-null verdicts carry an irreducible $0.84 \cdot \sigma_{\text{null}} \approx 13$ pp floor; neither anchor is touched by the table corrections, because both come from directly measured rather than fitted σ_{run}.

Several of our key results are nulls at four sets per arm (Tier-2c $p = 0.13$, Phase-4D $p = 0.21$). Those are underpowered nulls, not demonstrations of zero effect, and our own design tables say so: against the set-identity dispersion of 9–20 pp measured in those k=22 cells, Table A3 brackets a true 16 pp method effect between its $\delta = 20$ price (15–17 draws per arm) and its $\delta = 10$ price (56–63)—order 25 sets per arm, k=22-scoped and mixture-based, not Gaussian. We report them as "not resolvable at a realistic budget," which is the defensible reading and is also precisely the paper's point about this literature—but the asymmetry cuts both ways and we do not claim otherwise.

The Addendum-A variance corrections rest on minimal degrees of freedom—and the replication they were pending has now arrived, partly against them. The prospective cell's $\sigma_{\text{run}} = 3.51$ pp is a single 3-seed replicate (df $= 2$) and its $\sigma_{\text{set}} = 19.01$ pp comes from four set draws (df $= 3$) whose 40 pp spread is the rare-large-drop pattern Table A1 prices, and the other single-task k=22 cell prices the same 10 pp effect at 16–22 draws per arm (both moved here from §4 in the readability pass); against the fitted values they replace (7.63 and 11.69), the fit over-predicted σ_{run} 2.17$\times$ and under-predicted σ_{set} 1.63$\times$ (§4). The *directions* of the corrections remain well supported—the

σ_{run} over-prediction independently confirms the 1.55–$2.33\times$ estimate from our audit of the fitted components, and the σ_{set} under-prediction matches the Tier-2c set-identity pattern—and the fixed-set magnitude (8.0 pp at 3 seeds per arm) stands as a σ_{run}-based Gaussian estimate. The method-comparison magnitudes did not survive as first stated: the Gaussian method rows tabled from this cell (44.2 pp at 3 per arm; **59** draws per arm for a 10 pp effect) rested on propagating σ_{set} as a Gaussian sd—which two independent k=22 cells show it is not (a collapse mixture, Table A1; in the $\pi_{0.5}$ cell the pre-registered seed expansion attributes the collapse to the retrain rather than the set)—and on treating a k=22 single-task measurement as k-general, which the k-curve refutes. They are **withdrawn**; the surviving statements are k-scoped: **56–63 draws per arm** $[14, 79]$ **for a 10 pp effect at** k=22 **single-task,** $\sim$6 $[3, 33]$ **at** k=88 (Table A3), with the $k = 176$ σ_{run} attribution settled by the scored pre-registered K4B wave (§4.3; both readings carried, σ_{set}(total) and the mixture prices unmoved). The mixture parameters themselves rest on 4–6 draws per cell and license direction and order of magnitude only (Appendix A). The Addendum-A wave remains one cell, nine runs, four sets, `libero_object`/SmolVLA only.

Which withdrawn number, and why two of them are printed. The withdrawn Gaussian price for a 10 pp method effect in this cell is quoted here, in Table A3 and in Figure A4 as **59** draws per arm: that is $2\,(2.80 \cdot 19.33/10)^2$ evaluated at $\sqrt{\sigma_{\text{set}}^2 + \sigma_{\text{run}}^2} = 19.33$ pp, the formula exactly as this paper states it. An earlier printing of that one Addendum-A row propagated σ_{set} alone (19.01 pp) and gave **57**—a 1.7% difference in the sd and two draws in m. We standardise on **59**, the value the printed formula produces, and Appendix A keeps the superseded 57 on the page rather than harmonising the two silently—both figures are withdrawn in any case, and neither is in force anywhere in the paper.

The k-curve's σ_{run} estimates are minimal-df and their rise is unexplained. Each per-k σ_{run} in §4.3 rests on two seed pairs ($\text{df} = 2$), and the monotone rise ($1.68 \to 3.40 \to 3.35 \to 4.51$ pp) contradicts both our frozen prediction and the from-scratch trend; because $\sigma_{\text{set}}(\text{pure})^2 = \sigma_{\text{set}}(\text{total})^2 - \sigma_{\text{run}}^2$, an over-estimated σ_{run} at $k = 176$ manufactures part of the measured fall. $\sigma_{\text{run}}(k{=}176)$, the K4 verdict, and the K3 restatement were settled by a follow-up wave whose decision bands were frozen before its data existed (`research/K4B_PREREG.md`): band B3 (indeterminate, $\hat{\sigma_{\text{run}}}(\text{new}) = 2.60$, $\text{df} = 4$, CI $[1.56, 7.48]$), pooled $\sigma_{\text{run}} = 3.36$ ($\text{df} = 6$, CI $[2.17, 7.40]$), K3 flipped to the PASS side under the frozen recompute ($\sigma_{\text{set}}(\text{pure}, 176) = 3.86$ vs 2.43 originally)—**a flip in the flattering direction, reported as a flip** with both readings; the decay figure is quoted as 3.2–5.1$\times$ accordingly, and the K4 rise remains unsettled.

Table A6: The two regimes are limited by opposite things (§4.2).

	PushT / DP from scratch	SmolVLA fine-tune
harness floor σ_0	1.91 pp—$2\sigma_0^2$ is 53% of unpaired var.	**0.00 pp—bit-deterministic**
locality $\hat{w}$	unidentified [0.00, 1.00]	**0.49** [0.21, 0.91]
what caps the paired gain	the harness	the data
VRF observed	1.92 (at its floor-implied ceiling of 1.89)	1.73–3.05

The paired programme leaves w unidentified in the from-scratch regime and the variance–distance curve unresolved in both. The two regimes are limited by opposite things (Table A6): on PushT the harness floor consumes the $m = 4$ signal, so $\hat{w}$'s CI is $[0.00, 1.00]$ and the clean locality estimate ($\hat{w} = 0.49$ [0.21, 0.91]) rests on the fine-tuning cell alone; Leg 1's primary pass is one base from non-significance (leave-one-base-out $p > 0.05$ in 6 of 14 refits); and neither cell resolves the shape of $\text{Var}(\Delta)$ versus d—Leg 1's $m = 16$ rung (15.42 pp^2) lands above the in-wave unpaired arm (13.78 pp^2), which no distance model allows, and Leg 2's rungs invert within noise ($\text{Var}(m{=}4)/\text{Var}(m{=}2) = 0.57$ [0.23, 1.32]). The additive-model rejection and the σ_0 measurements do not depend on any of these; the shape of the locality curve does, and it is open. (Both paired waves also reproduce the atlas where they overlap it: Leg 1's reseed arm gives $\sigma_{\text{run}} = 2.10$ pp against §4's 2.09; Leg 2's gives 2.37 pp against the prospective cell's 3.51—both minimal-df estimates, consistent within their wide intervals.)

The PushT coverage effect changed size across stacks. Phase 1c measured $+11.1$ pp at 100K and $+15.0$ pp at 200K with deterministic covmax/covmin sets; Phase 1d's de-confounded replication measured $+7.0$ pp with independent draws; the repair wave's re-run of the deterministic sets on the rebuilt stack measured $+8.2$ pp. The Phase-1d draws sit 4–6% off the deterministic coverage optimum, and the rebuilt stack compresses the absolute success range (which is why we restrict

inference to within-wave comparisons). Direction holds in all four measurements, and every frozen criterion attached to any of them passed, but the magnitude spans $+7$ to $+15$ pp: on the rebuilt stack the deterministic sets give $+8.2$ against the draws' $+7.0$, so most of the drop from the original $+11.1/+15.0$ tracks the stack, not the loss of determinism. The one-line summary this licenses is given in §5.2.

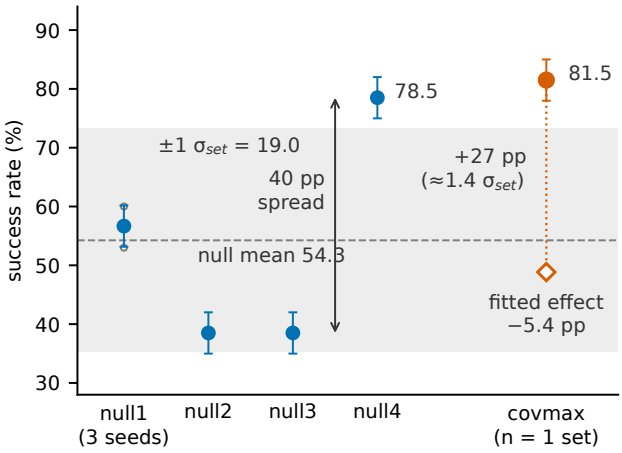

Figure A2: **The single-set trap** (§5). `libero_object` task 1, SmolVLA fine-tune: four same-size random draws span 40 pp (56.7/38.5/38.5/78.5; band $\pm 1\,\sigma_{\text{set}} = 19.0$ about the null mean 54.3), while the single coverage-max draw scores 81.5 ($+27$ pp, $\sim 1.4\,\sigma_{\text{set}}$) against a fitted effect of -5.4 pp. Read alone it is a large win; read against the null draws it is set identity.

Per-episode value reliability at $N = 200$ subsets is 0.29—per-*set* claims are the defensible unit at this scale, and the $N = 500$ extension that would raise the ceiling was cut. *Mechanism is exploratory:* across the 12 shared sets, the only statistic that tracks DP value is coverage in the policy-relevant state space ($\rho = -0.52$, $p = 0.08$; matching Phase-1d's objective↔SR $r = -0.79$ on PushT—the objective is a facility-location *cost*, so the negative sign is better coverage, §5.2), while under fine-tuning no summary statistic in any tested representation tracks value at all—not even the VLA's own embedding coverage ($\rho = -0.29$, n.s.; also null interventionally). With the set-conditional training instability, the working hypothesis is that from-scratch value is partly data-geometry while fine-tuning value is dominated by optimization-path effects invisible to per-set statistics; confirmatory probes are future work.

Deployment-side verification (moved here from §7, 2026-08-27). Theory bounds self-improving systems by their external verifier (Zhang et al., 2024; Huang et al., 2024; Silver & Sutton, 2025): that machinery certifies *within* a run, and retrain-level certification is the open problem. A first retrain-level release rule, per-task selection with a pre-registered abstention that defects from the incumbent only at a stated one-sided z, holds the probability of shipping a damaged task at 0.02 or below in the Meta-World cells of Appendix L.3; it certifies a choice among retrains the budget already produced, in simulation, and certification of a policy updated on its own experience remains open.

F.3 The twelve estimators of §5.1, named

§5.1's central null cites twelve estimators against PushT causal ground truth and does not list them; Table A7 does, with each one's definition, family, Spearman ρ and bootstrap CI, so the null can be checked rather than taken on trust. The early-loss probe of §5.1 reads, in full, $|\rho| \leq 0.19$ with negative cross-validated R^2 on LIBERO-object under DP, and $|\rho| \leq 0.12$ with CV-$R^2 = -0.02$ on the quiet PushT datamodel runs under from-scratch DP (`experiments/influence/PHASE1B_RESULTS.md`; scripts `phase1b_estimators.py`, `phase1b_model_estimators.py`, `phase1b_reliability.py`). All twelve are scored against the same target—the $N = 200$ datamodel's per-episode value at k=103—on the same $n = 206$ episodes.

Table A7: **The twelve estimators, named.** Target: measured per-episode causal value from the 200-model PushT datamodel (k=103, from-scratch DP); $n = 206$ episodes for every row; Spearman ρ with bootstrap 95% CI. Families: *kin* kinematic/heuristic, *cov* coverage, *grad* gradient/attribution from the full-data 200K checkpoint. Every CI straddles zero. The thirteenth estimator— DemInf-style per-episode state–action mutual information, our implementation substituting raw z-scored features for learned embeddings—scores $\rho = -0.06$ $[-0.19, +0.07]$ on the same target and is reported in §5.1.

estimator	definition	family	ρ	95% CI
`episode_length`	frames (trivial size baseline)	kin	$+0.084$	$[-0.06, +0.21]$
`sparc_badness`	$-$SPARC (spectral arc length); the drop-gate's smoothness heuristic	kin	$+0.057$	$[-0.08, +0.19]$
`action_entropy`	differential entropy of the episode's actions	kin	$+0.050$	$[-0.09, +0.18]$
`state_span`	log-det of the episode's state covariance	kin	$+0.043$	$[-0.10, +0.17]$
`coverage_knn_novelty`	mean k-NN distance of the episode's states to states of *other* episodes	cov	$+0.031$	$[-0.10, +0.16]$
`quality_combined`	the shipped drop-gate heuristic: the larger of z-scored `sparc_badness` and z-scored `divergence`	kin	$+0.017$	$[-0.12, +0.16]$
`divergence`	action-statistic outlierness; the drop-gate's divergence heuristic	kin	$\mathbf{-0.121}$	$[-0.25, +0.02]$
`trak_rowsum`	Fisher-whitened influence row-sum via SVD of the episode-gradient matrix (TRAK-style)	grad	$+0.044$	$[-0.09, +0.17]$
`grad_align`	cosine between the episode gradient and the mean training gradient	grad	$+0.033$	$[-0.11, +0.17]$
`grad_norm`	norm of the mean episode gradient over the UNet-head parameters (TracIn self-influence)	grad	$+0.019$	$[-0.12, +0.16]$
`influence_rowsum`	$\sum_{j \neq i} \langle g_i, g_j \rangle$; TracIn-CP influence of the episode on the rest of the data	grad	$+0.011$	$[-0.13, +0.15]$
`loss_mean`	per-episode mean diffusion loss at the full-data checkpoint (EL2N / hard-example score)	grad	-0.078	$[-0.21, +0.06]$

The exact bound. The sweep's bound, quoted as $\max|\rho| = \mathbf{0.121}$ in §5.1, is `divergence`; an earlier draft rounded it to 0.12, one thousandth too tight. Nothing turns on the thousandth; we print the source-precision value in both places rather than leave two forms in the paper.

Disattenuated against the reliability ceiling. The target is itself noisy: split-half $\rho = +0.172$ gives a Spearman–Brown reliability of $\mathbf{0.29}$ for the $N = 200$ per-episode weights, hence an attenuation ceiling of $\sqrt{0.29} = \mathbf{0.539}$—the largest correlation any perfectly reliable estimator could show against this target. The estimators are deterministic functions of the episode, so taking their own reliability at 1.0 (the assumption most favourable to them), the disattenuated bound is $0.121/\sqrt{0.29} = \mathbf{0.225}$: under 6% of the variance in measured value, and the largest *positive* estimator ($+0.084$) disattenuates to $+0.156$. That is why we report this as genuine failure rather than attenuation—the ceiling is 0.54 and nothing gets within a factor of two of it—and why the honest fix is a larger datamodel (the $N = 500$ extension that would raise the ceiling was cut), not a better scorer.

Six further scores, from the theory rather than the toolbox. If value were coverage, the right cheap score under $\alpha = 0.5$ random subsets is expected unique coverage, $\mathrm{EUC}_i = \sum_{b \in i} 2^{-(c_b - 1)}$. It fails in every space and resolution tried, and so do the alternatives: EUC state 16^2 -0.041, EUC state 8^2 $+0.023$, EUC action 16^2 -0.031, EUC state+action 6^4 $+0.116$, rarity-sum state 16^2 $+0.023$, Monte Carlo marginal facility-location gain $+0.029$ (all CIs straddle zero; `combinatorial_value.py`, part D). These six are counted separately from the twelve. The subset-level analysis behind them is the one that reframes the null: a linear datamodel's held-out CV-R^2 is 0.080 against the 0.83 that 500-episode binomial evaluation noise permits, so $\approx$75% of between-subset success variance is systematic-looking but *not* attributable to additive data composition—which the paired programme then rejects interventionally (§4.2, Appendix F.1).

G PRIMARY-SOURCE AUDIT OF THE CURATION LITERATURE

Seed-only variance outside robotics. The construction this paper rests on — retrain an identical pipeline on identical data, vary only the seed, and compare — has prior art in other fields, and §2's claim that no cited work has our three additions is made against it, not in ignorance of it. D'Amour et al. (2022) retrain 50 ImageNet pipelines differing only in seed: they agree to ± 0.11 pp on the iid test set yet vary an order of magnitude more on individual stress tests. McCoy et al. (2020) fine-tune 100 BERTs on one MNLI training set, obtaining a 1.2 pp span in-distribution against a 66 pp span on one syntactic subcase. Dehghani et al. (2021) show that task *selection* alone reorders methods. Jordan et al. (2020) make the matched argument inside RL evaluation. Each is an aggregate concealing disagreement, as ours is. The difference, and the reason the three additions stand: their disaggregations are onto a *separate* stress set or a different choice of tasks, whereas $\sigma_{\mathrm{pertask}}$ is measured across the very tasks whose arithmetic mean is the reported suite number, so the concealment is within-aggregate cancellation rather than out-of-distribution fragility. None of them measures a noise floor's transportability across cells of one benchmark, and none tests additive per-datum value interventionally.

We audited the reported evidence of six prominent curation papers against the question: does the paper report the run-level replication needed to compare its gain to the noise floor of the regime it was measured in? Table A8 summarizes the facts established in §4; the full audit table, with per-paper protocol details, ships with the corpus (`research/PHASE4_AUDIT.md`).

Figure 1b conventions. Shading uses *our* measured σ_{run} range, as we did not measure the audited papers' own, and it is lenient in their favor, excluding their rollout floors. Hardware legs get no marker: every σ_{run} we measured is in simulation, and pricing a hardware claim from a simulation floor is the cross-cell transfer our own first result forbids, so those rows read *not priceable* and rest on their replication columns alone.

Best-checkpoint selection, measured per run. One audited paper (Demo-SCORE) reports the *maximum* success over evaluations, a rule that converts run and checkpoint noise into positive bias. §4 quantifies what it buys on a 4B-parameter VLA at checkpoint counts in common use. Across the first 8 runs of the $\pi_{0.5}$ cell, max-over-4-checkpoints exceeds the final checkpoint by $+\mathbf{4.00}$ pp on average ($+12.5$ pp at most), the per-run deltas being $+5.0$, $+4.0$, 0.0, 0.0, $+4.0$, $+1.5$, $+12.5$, $+5.0$. Pooled over the full 16-run cell once the pre-registered pairs wave landed, the inflation falls to $+\mathbf{2.34}$ pp

Table A8: **Audit summary, one row per leg, matching Figure 1 leg for leg.** Every value is transcribed from the summary row of `research/PHASE4_AUDIT.md`, which is also what the figure is drawn from; the full audit document carries the complete per-paper protocol table. [f]: read from the source's figures rather than a printed number. **Simulation and hardware legs are split for every method that reports both**: judging one merged sim+hardware bar against a simulation-measured resolution floor would be exactly the cross-regime conflation this paper objects to. The "—" this table previously carried in the gain column is gone: it stood where a published number exists, which left the abstract's "5–32 pp in simulation and up to +88 pp on hardware" traceable to no printed row in the paper. It now reads off this table—simulation floor DemInf $+5.0$, simulation ceiling Demo-SCORE $+31.9$, hardware ceiling CUPID $+88$. CUPID's hardware leg is reproducible from its source's printed bars, as CUPID minus base policy: Figure-8 $72-36$ and $72-32$, TuckBox $84-0$ and $88-0$, Bookshelf $84-44$, giving $+36$ to $+88$; the source's own sentence, that CUPID "improves over the base policy's success rate by 38% (averaged over filtering and selection)", is exactly $(36+40)/2$ for Figure-8 and fixes the units as points. Its *simulation* leg publishes no such headline: those results are success-rate curves, and the source states that the Lift MH base policy already succeeds 100% of the time, so we record the leg as none-published rather than convert a curve into a gain. No audited paper reports a measured retraining noise floor for its evaluation regime; where our own audit finds a gain *likely to clear* any plausible floor, the last column says so. Two published statements these measured ranges replace are carried here so that no number leaves the paper: CUPID's own headline claim is that *under 33%* of curated data can yield state-of-the-art diffusion policies, and Demo-SCORE's abstract states "15–35% higher absolute" where the transcribed per-task figures are $+14.5$ to $+31.9$ (simulation) and $+30$ to $+40$ (hardware).

method (leg)	gain (pp)	replication reported, and what it licenses
CUPID (Agia et al., 2025), *sim*	none published (curves)	3 seeds, 500 rollouts per condition (50 × last-10 checkpoints, averaged, not best-of); per-seed success rates not printed, so no floor is recoverable. **Our audit records this gain as probably above any plausible noise floor.**
CUPID, *hardware*	$+36$ to $+88$[f]	single unreplicated run, 25 trials, final checkpoint; no per-condition spread
Demo-SCORE (Chen et al., 2025), *sim*	$+14.5$ to $+31.9$	no seeds reported (one run per condition); 256 rollouts; reports the *maximum* success over evaluations—a rule that converts run/checkpoint noise into positive bias
Demo-SCORE, *hardware*	$+30$ to $+40$[f]	no seeds reported; 30 trials; same max-over-evaluations rule, with checkpoints pre-screened on 1–2 rollouts
DemInf (Hejna et al., 2025), *sim*	$+5.0$ to $+10.0$	cleanest protocol in this set: 3 seeds, 200 trials each at a fixed final checkpoint (600 rollouts pooled). **Our audit records this gain as probably above any plausible noise floor.**
DemInf, *hardware*	none published (graphical)	3 seeds, 10–15 trials, fixed final checkpoint; no pp headline printed
SCIZOR (Zhang et al., 2026), *sim*	$+5.4$ to $+8.1$	the only audited paper printing per-condition seed spread ($n = 2$–3); 80 trials/task with top-3-checkpoint averaging, 240–300 OXE trials best-of-last-3
SCIZOR, *hardware*	$+32.9$	3 seeds, 10 trials, top-3-checkpoint averaging
RINSE (Kulkarni et al., 2026), *sim*	$+5.6$ to $+16.0$	3-seed means with no spread reported; 25 evaluation episodes, best checkpoint
RINSE, *hardware*	$+20.0$	single unreplicated run, 25 trials, best checkpoint
Re-Mix (Hejna et al., 2024), *hardware*	none published (38%/32% *relative*)	one training run per mixture; 10 real trials per task; validation-loss early stop

mean (+12.5 pp max): the pairs wave on its own contributes +0.69 pp, and 7 of the 16 runs peak at the final checkpoint.

Set draws per arm. No audited robot-curation evaluation meets even the $k \geq 88$ method-comparison price of §4: every design in this audit uses a *single* set draw per arm, so its arms differ in set identity as well as in method, and the k-scoped mixture tables (Appendix A) price that gap at $\sim$6 draws per arm at k=88 and 56–63 at k=22.

And the audit exonerates two of the six on their simulation legs. Our own primary-source pass records, in the same paragraph as the criticism, that **CUPID's and DemInf's simulation gains are probably above any plausible noise floor** (3 seeds, 500–600 rollouts each), even though their per-seed success rates are unpublished and so no floor is recoverable from what they printed; CUPID's *real-robot* claim is a single unreplicated run at 25 trials, while DemInf's reports 3 seeds at 10–15 trials but publishes no pp headline; both are a separate matter. SCIZOR is the only one of the six whose gains can be compared against a seed-level spread from its own tables, and only coarsely (n = 2–3). We state this as prominently as the criticism because it is our own finding, because two of the six audited groups are plausible reviewers of this paper, and because the claim we defend is regime by regime and method by method—not a blanket charge against a literature.

The literature's largest real-robot gains (+30–88 pp) rest on 10–30 eval trials and 0–3 replicates. Our σ_{run}-based fixed-set table puts the resolvable-effect floor of a 1-seed, 25-episode comparison at $\sim$20–45 pp in the σ_{run} regimes we measured—several headline numbers live inside that band—and Table A4's shared-seed pricing independently finds that $\sim$25 eval episodes decides only effects above $\sim$20 pp. An effect inside that band is *unverified*, not refuted: the claim is about the reported evidence, not about whether the methods work; only replication at measured noise floors can settle that, and the papers as published do not permit it.

Extended related-work notes (for §2). RoboMimic (Mandlekar et al., 2021) showed that data quality alone moves success $\sim$30 pp at fixed algorithm, and its validation-loss-fails finding is the ancestor of our probe-assay null (§5.1); that $\sim$30 pp quality effect is 72% vs 42% success, Proficient- vs Multi-Human Transport. On observational quality metrics: Sakr et al. (2025) report consistency/jerk/duration metrics predicting success at $R^2 \approx 0.70$–0.89, and FAKTUAL (Sirigiri et al., 2026) a diversity metric correlating with success—but across random subsets the summary statistics concentrate, so those correlations are estimated over a range that designed extremes leave far behind. On the deep-RL precedent: Henderson et al. (2018) showed seed variance swamped many reported RL improvements, and Agarwal et al. (2021) turned that into reporting practice; TRI LBM Team (2025) is blind randomized evaluation of *large behavior models*. Data Shapley (Ghorbani & Zou, 2019) was extended to in-context demonstrations by DemoShapley (Xie et al., 2024). And the cross-field selection-on-noise pattern (Szymanski et al., 2023; Shao et al., 2025) is exactly the structure this atlas prices: results selected on noisy evaluations that did not survive re-measurement against adequate nulls.

Attribution, valuation, and subset selection (relocated from §2 in the 2026-08-24 pass; verbatim). Our ground truth is the datamodel construction of Ilyas et al. (2022); TRAK (Park et al., 2023) makes attribution cheap (we include a TRAK-style whitened influence among our estimators); influence functions (Koh & Liang, 2017) underpin CUPID, with well-documented fragility in deep networks (Basu et al., 2021; Bae et al., 2022); DataMIL (Dass et al., 2025) applies datamodels to robot data *selection*, where we instead use them as causal ground truth; and Data Shapley (Ghorbani & Zou, 2019; Xie et al., 2024) is typically approximated by the same machinery. The difference is the target: held-out loss is cheap and low-variance, whereas closed-loop success has reliability 0.29 at $N = 200$ subsets (ceiling 0.54; Appendix F), which bounds what "per-episode value" can mean. Two results bracket the additive premise our paired programme tests. Koh et al. (2019) ran the closest prior interventional version of it, removing whole *groups* and retraining rather than trusting the sum of individual influences, and found the additive prediction tracks the truth well on held-out loss; our §4.2 finds it does not on closed-loop success, so the disagreement is about the target, not the method. Wang & Jia (2023) then shows the supervised case is not safe either once training is stochastic: semivalue rankings of individual data points move under retraining noise alone, which is the same failure mode we measure downstream of a much larger floor. On observational quality metrics (Sakr et al., 2025; Sirigiri et al., 2026): these do not contradict our null against *causal* value, because across random subsets the summary statistics concentrate, so their correlations are estimated over a range that designed extremes leave far behind.

H THE NOISE-FLOOR MEASUREMENT KIT

The reusable instruments, in the order a new cell would apply them:

1. **Duplicate-retraining seed pairs for σ_{run}:** identical subset, identical eval protocol, $\sigma_{\text{run}} = \text{sd}(\Delta\text{SR})/\sqrt{2}$.
2. **Independent same-size set draws for σ_{set}:** the set-to-set sd of independent draws of the same rule, which seed replication cannot reduce.
3. **Design tables—two questions, two instruments.** For *fixed-set* claims, $\text{MDE} = 2.80 \cdot \sigma_{\text{run}} \cdot \sqrt{2/m}$ over m seed retrainings per arm (80% power, $\alpha = .05$), corrected by the prospective replicates of §4; this rule stands, and the correction makes fixed-set comparisons $\sim 2.2\times$ easier than the pre-correction tables said (**a 10 pp effect between two fixed sets needs ~ 2 seeds per arm**, Table A11; Appendix H.1 gives this and the new-cell price together, with a measured GPU-hour line—those are the numbers a reader should leave with). For *method-level* claims the Gaussian propagation $\text{MDE} = 2.80 \cdot \sqrt{\sigma_{\text{set}}^2 + \sigma_{\text{run}}^2} \cdot \sqrt{2/m}$ is **withdrawn**—at small k the distribution of independent set draws is a collapse mixture, not a Gaussian spread—and replication is sized instead from the k-scoped mixture tables (Tables A1–A4, mixture tables R1–R4; §4, Appendix A).
4. **The designed-subset percentile test** against a random-subset null of the same size and budget.
5. **The set-diversity de-confounding protocol:** re-draw a selection method several times independently and compare arm variance to within-arm set variance. This protocol overturned two of our own headline results (Tier-2's -23.2 pp reversal, Phase-4's $+19.5$ pp win) and confirmed a third (Phase-1d), which is the case for it; Addendum A then reproduced its lesson prospectively (a single coverage-max draw at $+27$ pp over the null mean, uninterpretable against $\sigma_{\text{set}} = 19$ pp).
6. **The adaptive-evaluation instruments and the fixed-checkpoint price list** (Appendix B, Table A4): the item-response view of a shared-seed corpus, the discordant-item rate as the quantity any item-selection scheme must raise, the flip floor that bounds it, and the resulting episode budgets for ranking two fixed checkpoints.
7. **The paired programme's instruments:** the anchor/rung construction whose across-base pairs supply a free in-wave unpaired reference; the exact randomisation test of $H_0: w = 1$; and the σ_0 replicate arm—identical data, identical seed—that separates harness nondeterminism from composition.
8. **The 2026-08 additions** (inventoried for §6): the per-task decomposition estimator with the both-readings rule (`pertask_decomp.py`: raw and floor-subtracted σ_{pertask} with jackknife CI, suite named on every quote); the cross-cell refusal gate `compare_sigma_run()` with pre-check 11b (SR-regime scope enforced through declared `sr_range`) and pre-check 11c (an anti-conservative-prior warning whenever a floor-dominated cell is used to price); the sub-binomial dispersion diagnostic $\hat{c}$; and the σ-transform convention as runnable code (`sigma_transferability.py`, Appendix H.2 and Appendix K.6).

Table A9: **The pre-registered k-curve** (SmolVLA fine-tune, multi-task pool P=454, 20K-step budget, 200-episode shared-seed evaluation; six independent draws plus two replicate seed pairs per k; SR and σ in pp; schedule stamp: truncated, 20k steps / 30k decay—completion effect measured null on `libero_object`, §4). Scored verdicts K1–K4 are in Appendix D.1.

k	mean SR	$\sigma_{\text{set}}(\text{tot})$	σ_{run} (df 2)	$\sigma_{\text{set}}(\text{pure})$	overlap	$\sigma_{\text{set}}(\text{pure})$, corr.	draws/arm, 10 pp[G]
22	42.75	12.58	1.68	12.47	4.1%	12.74	24.8
44	58.50	5.78	3.40	4.67	8.0%	4.87	5.2
88	57.50	4.34	3.35	2.75	19.0%	3.05	2.9
176	69.50	5.12	4.51[K]	2.43[K]	37.9%	3.08[K]	4.1

[G]Gaussian reference only—the mixture prices govern (Table A3). [K]As originally measured ($\text{df} = 2$); the pre-registered K4B re-measurement revises the pooled σ_{run} to 3.36 ($\text{df} = 6$) and hence $\sigma_{\text{set}}(\text{pure})$ to 3.86, overlap-corrected 4.90; both readings are carried wherever these quantities are used.

Table A10 gives the full noise atlas behind Figure A3. The $\pi_{0.5}$ entry is reported at both readings deliberately. The cell-1 value (2.02 pp, df=2) was superseded by PI05-PAIRS, whose pre-registered

Table A10: **The noise atlas, every row against the sampling floor of its own evaluation.** σ_{run} across eight measured regimes (the first also re-measured on a rebuilt eval stack), displayed graphically as Figure A3; the body prints the same convention over the cells it prices in Table 1. Every σ_{run} is a duplicate-retraining measurement, not a fit, and **no σ_{run} here is changed** by this table—what is added is the second reading every σ in this paper is required to carry (reporting standard (6)), which this table previously withheld. Intervals are 90% χ^2 at the row's stated replication unless marked otherwise (the $\pi_{0.5}$ row reports its 95% χ^2 CI); n_{eval} is evaluation episodes per run. Floors are binomial, $f = 100\sqrt{p(1-p)/n_{\mathrm{eval}}}$ at the row's realised mean SR; floor-sub. is $\sqrt{\sigma_{\mathrm{run}}^2 - f^2}$; share is $f^2/\sigma_{\mathrm{run}}^2$. [F]**At or below its own floor**: unresolvable above sampling—a measurement, never a prior, and pricing from it *understates* the lottery. [S]Floor share above 75%: the row sizes an evaluation, it does not price a retrain lottery. [C]PushT only: identically-ordered PushT runs replay identical initial states, so a replicate *difference* carries no binomial term at all (Appendix E); the naive floor printed here therefore *overstates* what binds on those two rows, which is why a PushT σ_{run} can legitimately sit beneath it. LIBERO delivers no such cancellation—only 10.06% of same-index cross-run pairs share an initial condition—so its floors bind as printed. Mean-SR provenance: note $*$.

regime	mean SR	n_{eval}	σ_{run} (90% CI)	floor f	floor-sub.	share	replication
PushT k=103, DP from scratch[C,F]	33.38%	500	2.09 (1.47–3.76)	2.109	—	102%	8 pairs (df=7)
— **rebuilt eval stack** (pooled)[C,F]	17.05%	500	**1.292**	1.682	—	170%	4 pairs (df=4)
LIBERO-object k=22, DP	66.56%	200	11.43 (8.06–20.54)	3.336	10.93	8.5%	8 pairs, df=7 (+10 repl.)
— hierarchical posterior	66.56%	200	13.7 (10.4–17.2)	3.336	13.29	5.9%	
LIBERO-object k=31, DP	80.67%	200	7.48	2.792	6.94	13.9%	6 pairs
LIBERO-object k=44, DP	88.38%	200	6.24	2.266	5.81	13.2%	8 seeds, 1 set
LIBERO-object k=22, **SmolVLA fine-tune**[T,F]	47.10%	200	$\leq$**3.31** (2.23–6.92)	3.530	—	114%	6 pairs (df=5)
LIBERO goal/spatial/long10, SmolVLA-ft (pooled)[T,S]	68.92%	200	**2.82**	2.487	1.33	78%	9 pairs (3 suites)
LIBERO-object task 1, SmolVLA-ft (**prospective**)[T,S]	58.14%	200	**3.51**	3.488	0.39	99%	3 seeds, 1 set
LIBERO-object k=22, $\pi_{0.5}$ **fine-tune**[P]	77.81%	200	**9.74** (χ^2 **CI 6.81–17.10**)	2.938	9.29	9.1%	16 runs, 6 masks (df=10)
— excluding the collapsed mask null04	79.11%	200	5.23 (3.60–9.55)	2.875	4.37	30%	14 runs (df=9)
— *superseded* cell-1 reading[F]	73.13%	200	*2.02 (95% CI 1.05–12.67)*	3.135	—	241%	2 pairs (df=2)

[F]**At or below its own binomial floor—four rows, two of them newly stamped in this pass.** The rebuilt-stack row (1.292 against 1.682) is the case §4 leads with, pooled over all four post-rebuild replicate pairs (two repair-wave, two prior-calibration); the two half-evidence readings it supersedes are 1.80 (repair pairs, df=2, one gap exactly 0.00, 95% χ^2 CI [0.94, 11.31]) and 0.32 (prior-calibration pairs, df=2). **The SmolVLA-ft k=22 row, 3.31 against a 3.530 pp floor, now carries the same stamp and did not before**—it is the middle rung of the three-policy ladder our own pre-registered PI05-PAIRS wave falsified, and stamping it makes that falsification *stronger, not weaker*: the ladder is not merely non-monotone ($11.43 \to \leq 3.31 \to 9.74$), its middle rung is an **upper bound**, so "the noise floor falls monotonically with pretraining scale" fails at the third rung whatever the second rung's true value, and the fine-tuning collapse is a bound on σ_{run} rather than a measurement of a smaller one. The *superseded* $\pi_{0.5}$ cell-1 reading (2.02 against 3.135) was floor-dominated too: **two of the original ladder's three rungs were unresolvable at** n_{eval}=200, which is the honest reason it should never have been read as a ladder. The PushT k=103 row (2.09 against 2.109) is stamped arithmetically and read under note C. [S]Floor share above 75%: sampling dominates the variance, so the row sizes an evaluation rather than pricing a retrain lottery, and each is quoted only in the direction its stamp allows. The prospective row's 3.51 bounds σ_{run} *from above*, which strengthens rather than weakens the 2.17× over-prediction of the fitted component it corrects (§4); the breadth row's 2.82 likewise bounds the fine-tuning collapse from above and is never read as a measured floor. [T]Schedule stamp: truncated LR schedule (20k steps of a 30k cosine decay—the ecosystem default recipe); a pre-registered completion experiment measured the completion effect as null for SmolVLA-ft on `libero_object` (Appendix K.1); untested on other suites. [P]Truncated schedule; completion effect untested for $\pi_{0.5}$. $*$*Mean-*SR *provenance, and the one convention choice it rests on.* Realised mean SR is recomputed for this table, cell by cell, with `python3` over `experiments/forecast/atlas.csv` (the frozen 414-run atlas), averaging *every* corpus run in the row's task × policy-class × k cell—so the SmolVLA-ft k=22 row averages all 50 `smolvla_ft/libero_object/`k=22 runs to 47.10%. Rows whose runs are not in that file use their own wave's records: the rebuilt-stack row its four post-rebuild replicate pairs (17.05%), the prospective row the seven k=22 runs of Addendum A (58.14%), the three $\pi_{0.5}$ rows `atlas_pi05.csv` + `atlas_pi05_pairs.csv` (16 runs, 77.81%; 14 excluding `null04`; 8 for the superseded cell-1 reading). Restricting a row's mean instead to just the replicate runs its σ_{run} is computed from moves no floor by more than 0.1 pp and changes no stamp (SmolVLA-ft k=22: 53.29% gives 3.528 against 3.530; PushT k=103: 35.71% gives 2.143 against 2.109). The breadth row's floor is pooled across its three suites at their own repaired means—`libero_goal` 94.32%/1.637, `libero_spatial` 85.77%/2.470, `libero_10` 26.68%/3.128—df-weighted, because one floor at the pooled 68.92% mean (3.273) would be wrong for all three; on that reading *each* per-suite σ_{run} (goal 1.08, spatial 1.84, long10 2.94) also sits below its own suite floor, and the pooled 2.82 exceeds all three only because pooling centres nine pair gaps on a single grand mean. Finally, the DP/LIBERO-object k=22 row's mean SR is not carried with its σ_{run} record: the 66.56% Table 1 prints is this recomputation over that cell's 58 runs, and restricting it to the 8 replicate runs the σ_{run} is built from (63.50%) moves the floor 3.336→3.404 and the share 8.5%→8.9%, changing no stamp.

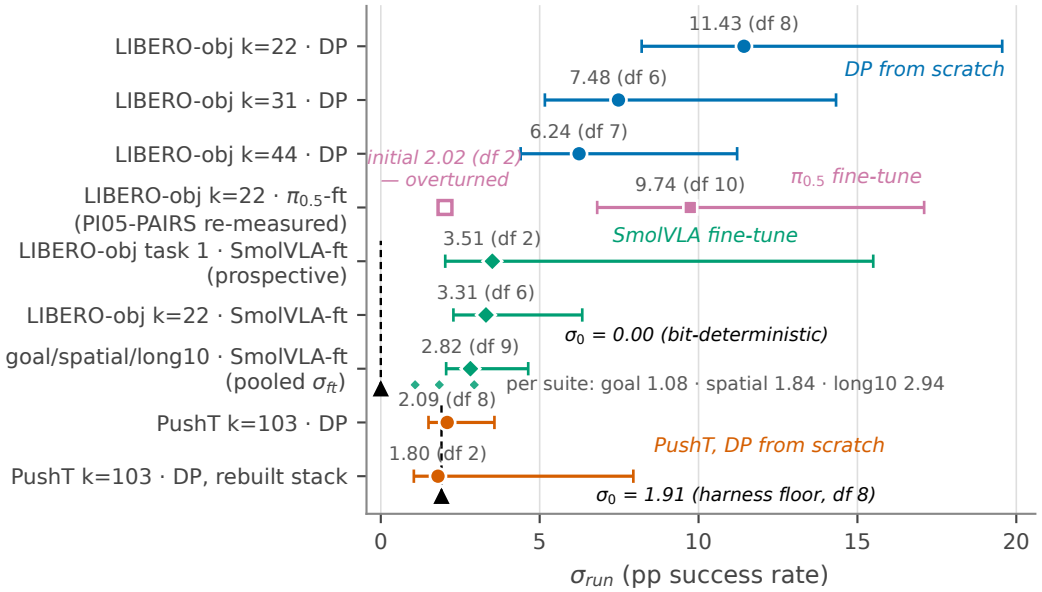

Figure A3: **The noise floor is a measured, cell-specific variable, not a constant.** σ_{run} across the eight measured regimes of Table A10, plus the rebuilt-stack re-measurement of the first (open square: the **overturned** initial $\pi_{0.5}$ reading of 2.02, re-measured at 9.74 by its pre-registered 16-run seed expansion; dashed: the measured harness floors σ_0). Intervals are 90% χ^2 except the $\pi_{0.5}$ row, which carries its reported 95% CI [6.81, 17.10] at df 10. The eight cells that record *both* a realised success rate and a σ_{run}—a different set from the eight regimes drawn here, and the one the cross-cell transform analysis runs over—are enumerated in Table A12; its endpoints are ACT cells not drawn here. Values: Table A10.

null04 rule resolved on its SEED-ATTRIBUTABLE branch: the same episode set scored 50.0 at seed 0 and 87.5 at seed 1, a 37.5 pp gap on byte-identical data and the largest null gap in the corpus. That single mask contributes 703 of the 950 pooled within-mask sum of squares, which is why the null04-excluded robustness reading is given beside the headline rather than in its place. Note that the superseded point estimate lies inside its own interval—the df=2 confidence statement was honest; only the point was misleading, which is the case for quoting intervals rather than point estimates at low df. Low-df σ estimates in this corpus have now moved in *both* directions on re-measurement: upward twice (the $\pi_{0.5}$ 2.02 $\rightarrow$ 9.74 overturn above, and a prior-calibration re-measurement of the PushT/DP entry finding 3.58 pp at df $= 24$ against the atlas's 2.09 at df $= 8$) and downward once (the rebuilt-stack reading, 1.80 at df $= 2 \rightarrow$ 1.292 pooled at df $= 4$, which lands below its own eval floor—Table A10, note F). A low df licenses an interval, not a direction. Table A11 gives the corrected design rules for the fine-tuned-VLA LIBERO regime; Figure A4 displays the correction graphically. The σ_{run} over-prediction (7.63 fitted vs 3.51 measured) is traceable to pooling heteroscedastic designed-set replicate pairs into the fitted component.

Table A11: Design rules for the fine-tuned-VLA LIBERO regime, **scoped to** k=22 **single-task**: MDE (pp) at 80% power, $\alpha = .05$. The fixed-set rows stand (m = seeds per arm). The method rows (m = independent set draws per arm) are **withdrawn**—they propagated $\sigma_{\text{set}} = 19.01$ through the Gaussian formula, and the generating process is a collapse mixture—and are reproduced in italics only as the withdrawn reference; the replacement prices are Tables A2 and A3 (Appendix A).

MDE	m=2	m=3	m=5	m=10
two fixed sets — pre-correction	21.4	17.5	13.5	9.6
two fixed sets — **corrected**	**9.8**	**8.0**	**6.2**	**4.4**
two selection methods — pre-correction (withdrawn)	*39.1*	*31.9*	*24.7*	*17.5*
two selection methods — Gaussian corrected (withdrawn)	*54.1*	*44.2*	*34.2*	*24.2*

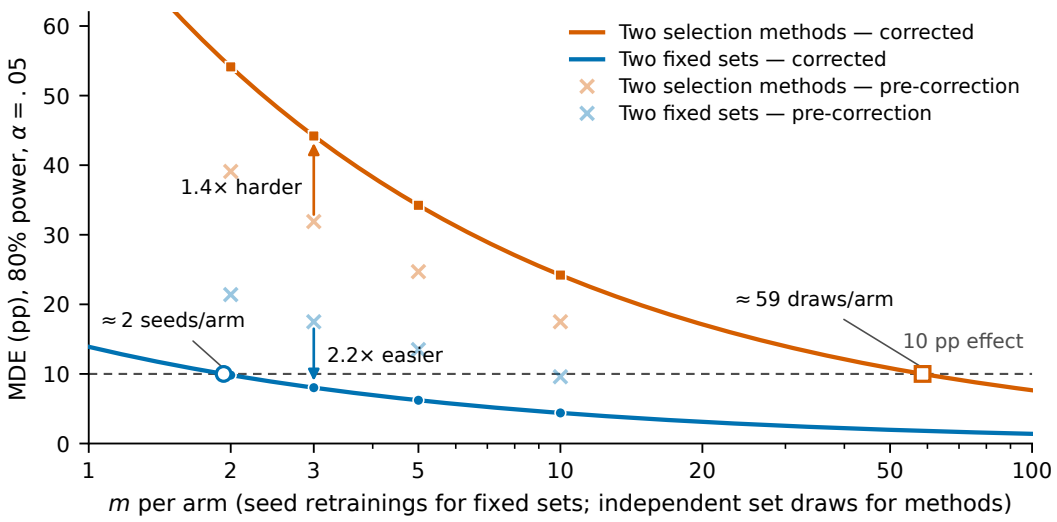

Figure A4: **Two questions, two prices—fixed-set correction standing, method curve withdrawn.** MDE $= 2.80\,\sigma\sqrt{2/m}$ (80% power, $\alpha = .05$), fine-tuned-VLA LIBERO regime, **scoped to** k=22 **single-task**. Corrected curves use the prospectively measured $\sigma_{\text{run}} = 3.51$ and $\sigma_{\text{set}} = 19.01$ (method comparisons take $\sqrt{\sigma_{\text{set}}^2 + \sigma_{\text{run}}^2}$); faded ×: pre-correction values (Table A11). The fixed-set reading stands: a 10 pp effect needs ≈2 seeds per arm between two fixed sets. The method-comparison curve—≈59 independent set draws per arm for a 10 pp effect—is the **withdrawn Gaussian reference**, drawn here only for comparison: at this k the set-draw distribution is a collapse mixture, and the replacement prices are 56–63 draws per arm $[14, 79]$ at k=22 single-task and ∼6 $[3, 33]$ at k=88 (Tables A2–A3, Appendix A).

H.1 The affordable prices

The price a reader is most likely to carry away from this paper is **56–63 independent set draws per arm**, and on its own that number is misleading: it is the *method-comparison* price in the single hardest cell we measured (k=22, single-task `libero_object`, SmolVLA-ft, where set identity is a collapse mixture), and it is a reason to stop running that experiment, not a bar on doing careful work. The two prices that actually govern day-to-day practice are cheap, and both are measured here rather than assumed.

Comparing two fixed sets: ∼2 seeds per arm for a 10 pp effect. With the prospectively measured $\sigma_{\text{run}} = 3.51$ pp in the fine-tuned-VLA LIBERO regime, $\text{MDE} = 2.80 \cdot \sigma_{\text{run}}\sqrt{2/m}$ gives 9.8 pp at m=2 and 8.0 at m=3 (Table A11, corrected row). Two retrainings per arm resolve a 10 pp fixed-set difference. That is the correction landing in the *helpful* direction, and it is the one result of this programme that makes an experiment cheaper rather than more expensive. Scope: fixed sets only—it prices "does this set beat that set," never "does this selection rule beat that one," and the row is k=22-scoped and carries the note-S caveat that 3.51 bounds σ_{run} from above.

Pricing a new cell: 8 retrains of one fixed set, plus an evaluation large enough to be beaten. Every post-freeze wave in this paper priced its cell the same way and we recommend it as the minimum: **8 retrainings of one byte-identical episode set, differing only in training seed** ($\text{df} = 7$), evaluated under one fixed protocol, with n_{eval} chosen so the binomial floor $100\sqrt{p(1-p)/n_{\text{eval}}}$ sits clearly below the σ the cell is expected to have—otherwise the wave buys an upper bound, as four rows of Table A10 did. Report both readings, raw and floor-subtracted, and the floor share. Fewer than 8 is not a saving but a different measurement: this corpus contains two df=2 estimates that moved on re-measurement, one upward by 4.8× (2.02 → 9.74) and one downward (1.80 → 1.292), and **a low df licenses an interval, not a direction.**

One worked cost line. In the cell where we have a measured bill—SmolVLA fine-tune, 10-task `libero_object` pool, k=88, 20K steps, 200 evaluation episodes per run (20/task)—the CURATED-88 wave ran 48 such runs for ≈53 **GPU-hours** on an 8-GPU node, i.e. ≈1.1 **GPU-hours per run including its evaluation** (budgeted and realised; `research/CURATED88_PREREG.md`, `CURATED88_RESULTS.md`). The recommended 8-retrain cell price is therefore **8 runs** × **200**

episodes ≈ 9 GPU-hours, and a fixed-set comparison at 2 seeds per arm is **4 runs ≈ 4.5 GPU-hours**. Against those, the withdrawn Gaussian method price (59 draws per arm, i.e. 118 runs ≈ 130 GPU-hours) and the surviving mixture price at k=22 (56–63 draws per arm) are what make method-level claims in that regime unaffordable—not the instrument, and not the noise floor. Where the floor is the binding constraint the cost moves to evaluation rather than training: taking a multi-task LIBERO cell from 20 to 100 episodes/task (n_{eval} 200 → 1000) is what dropped PERTASK's per-task floor from ≈10.9 to ≈4.9 pp and turned an INDETERMINATE verdict into a resolved one (Appendix J), and it buys resolution no number of extra retrains can.

H.2 Cross-cell transfer: the eight dual-recording cells, enumerated

The σ-transform analysis of Appendix K.6 is computed over the eight cells in this corpus that record *both* a realised mean SR and a σ_{run} measured by duplicate retraining. Those eight cells are not the eight regimes of Table A10 and not the nine rows of Figure A3; until this pass they were enumerated **nowhere in this paper**, so a reader could not check the dispersion figure they feed, nor see that three of them are one training wave. Table A12 prints them, transcribed row by row from the `CELLS` list of the convention script that is the published definition of the convention (`experiments/influence/audit_2026_08_18/sigma_transferability.py`).

Table A12: **The eight cells behind the cross-cell dispersion figure**, transcribed from `sigma_transferability.py`. Floors are binomial at each cell's own mean SR and n_{eval}; floor-sub. is $\sqrt{\sigma_{\text{run}}^2 - f^2}$; share is $f^2/\sigma_{\text{run}}^2$. [H]`pusht-act`, `-h8` and `-h16` are the *same eight ACT checkpoints re-evaluated at three inference horizons* (h = 100, 8, 16; no training anywhere, Appendix K.5)—**not three independent cells**, so the dispersion counts one training wave three times. [R]Enters at its two-pair 1.800 reading (df=2), not the canonical pooled **1.292** (df=4) that Table A10 prints for the same cell. [L]OW-SR: σ compressed near SR=0, never a class property. • marks the **six independently measured cells** that survive this paper's own quoting rules—drop the two duplicate horizon rows, and take h=8's **4.849** as the licensed ACT/PushT figure, chosen in advance (Appendix K.4) precisely because the 1.35%-SR `pusht-act` row *understates* that cell's lottery by ∼6.2× raw. Over those six the max/min ratio is **3.6**× (rebuilt PushT at 1.800), **5.0**× (at its canonical 1.292) or **3.1**× (admitting only rows under the 75% floor-share bar), against the 8.31× the eight-cell set gives.

	cell	mean SR	n_{eval}	σ_{run}	floor f	floor-sub.	share
	`pusht-act`[L,H]	1.350%	500	0.784	0.516	0.590	43%
•	`aloha-insertion-act`[L]	11.975%	500	2.261	1.452	1.733	41%
•	`pusht-dp-rebuilt`[R]	16.500%	500	1.800	1.660	0.696	85%
•	`pusht-act-h8`[H]	16.350%	500	**4.849**	1.654	4.558	12%
	`pusht-act-h16`[H]	22.100%	500	4.839	1.856	4.469	15%
•	`smolvla-ft-k88`	54.944%	2000	2.113	1.113	1.797	28%
•	`pi05-ft-k88`	62.263%	1000	4.228	1.533	3.940	13%
•	`aloha-transfer-cube-act`	77.875%	500	**6.515**	1.856	6.245	8%

The earlier framing overstated, and this is the disclosure. An earlier version of this paper quoted a single 8.31× cross-cell multiple, and led the abstract with it. Three properties of the set it is computed over, all visible only now that the set is printed, say it should not have carried that weight: three of the eight rows are one ACT training wave at three inference horizons; the low endpoint 0.784 is the row our own pre-committed rule declines to quote alone; and the `pusht-dp-rebuilt` row enters at 1.800 rather than the 1.292 this paper elsewhere calls canonical for that cell. Honouring our own rules leaves six independently measured cells and a ratio of **3.1**×–**5.0**×. The multiple was ours, we printed it, and it was inflated by roughly a factor of two; recording that here is cheaper than defending it, and it is why the body now quotes a range and a refusal instead.

What the enumeration licenses instead. Across the cells we price, σ_{run} runs from at or below a cell's own binomial floor (rebuilt PushT/DP, 1.292 against a 1.682 floor: a measurement, never a prior) to **6.515** pp on bimanual ALOHA and **11.43** pp in a small-data LIBERO cell, and no scale-free transform we tested collapses that spread under the floor-subtracted convention (Appendix K.6 gives the CVs, including the one transform that beats pp on raw σ and therefore falsifies our own strong form). **We deliberately quote no single cross-cell multiple.** Our own refusal gate, `atlas.compare_sigma_run()`, declines these pairs on floor share and evaluation scope—

three of the eight rows sit above 40% floor share and one above 85%, and n_{eval} ranges 500–2000 across them—and *that refusal is the result*: it is this paper's thesis applied to itself. The atlas sizes an experiment; it cannot price a cell nobody measured. $8.31\times$ survives only as the dispersion input to the transform CVs, never as a price.

I REPRODUCIBILITY INVENTORY

All artifacts below live in the anonymous supplementary material accompanying this submission. (Paths are given relative to the artifact root; the public repository URL is withheld for double-blind review and is given in the non-anonymous version of this paper.) The non-anonymous version of this paper names the public repository directly. What is deposited is the full record listed below; still to build before the corpus is turn-key for others are a public mirror of the per-model records, a datasheet, Croissant metadata, a license audit of the source datasets, and one-command reproduction of the headline analyses.

Experiments: `experiments/influence/`, `experiments/within_task_sweep/`, `experiments/forecast/` (Addendum A, Phase-1c repair, and the forecasting exercise), `experiments/paired/`; per-model records in the supplementary data archive (repair-wave, Addendum-A, and paired-programme raw records mirrored there; artefact hashes in `PAIRED.sha256`).

Pre-registrations (all frozen and timestamped to cloud storage before their wave launched): `PHASE1C_PREREG.md`, `PHASE1D_PREREG.md`, `PHASE2_PREREG_DRAFT.md`, `PHASE2B_PREREG.md`, `PHASE2C_PREREG.md`, `TIER2B_PREREG.md`, `TIER2C_PREREG.md`, `TIER3_PREREG.md`, `PHASE4_PREREG.md`, `PHASE4D_PREREG.md`, `forecast/PHASE1C_REPAIR_PREREG.md`, `forecast/ADDENDUM_A_PREREG.md`, `paired/PAIRED_PREREG.md` (with its declared amendment `AMENDMENT_FLOOR.md`); and for the 2026-08 session and its follow-ups, `research/KCURVE_PREREG.md`, `research/CURATED88_PREREG.md`, `research/K4B_PREREG.md` (frozen decision bands), and `research/PI05_PAIRS_PREREG.md` (sha256 `be2147c6...`, mirrored to cloud storage before mask generation).

Analyzers: `analyze_phase2.py`, `analyze_tier2.py`, `hierarchical_phase2.py`, `power_analysis_phase2.py`, `probe_assay_phase2.py`, `probe_assay_pusht.py`, `noise_atlas.py`, `phase3_repr_stats.py`, `deminf_lite.py`, `combinatorial_value.py`; paired: `analyze_paired.py`, `power_paired.py`, `pilot_overlap_variance.py`, `check_historical_reference.py`.

Selection and wave drivers: `gen_{datamodel,phase1c,phase1d,phase2,tier2,tier2c,tier3,phase4_libero}_masks.py`; `gen_paired_masks.py`, `gen_floor_amendment.py`; `run_{phase2,tier2,tier3,phase1d}.sh`.

Results documents (one per wave, each carrying its frozen verdicts): `DATAMODEL_RESULTS.md`, `PHASE1B_RESULTS.md`, `PHASE1C_RESULTS.md`, `PHASE1D_RESULTS.md`, `PHASE2_RESULTS.md`, `PHASE2C_RESULTS.md`, `TIER2BC_RESULTS.md`, `TIER3_RESULTS.md`, `PHASE3_NOTES.md`, `PHASE4_RESULTS.md`, `forecast/PHASE1C_REPAIR_RESULTS.md`, `forecast/ADDENDUM_A_RESULTS.md`, `forecast/FORECAST_RESULTS.md`, `paired/LEG1_RESULTS.md`, `paired/PROGRAMME_RESULTS.md` (design and pre-launch derivations in `paired/DESIGN.md`), `POWER_ANALYSIS.md`, plus the literature audit `research/PHASE4_AUDIT.md`.

2026-08 session artifacts: `experiments/influence/PI05_CELL1_RESULTS.md` (design freeze: `experiments/influence/pi05_pilot/PI05_PILOT_PREREG.md`, with its three timestamped amendments); `research/KCURVE_PREREG.md` and `research/KCURVE_RESULTS.md`; `research/CURATED88_PREREG.md` (results logged in `SESSION_LOG_2026-08-08.md`; the H2 co-primary scored in `experiments/forecast/score_curated88_h2.py`, log `CURATED88_H2_SCORING.txt`; item-level analyses `forecast/item_analysis.py`);

research/K4B_PREREG.md (frozen decision bands) with research/K4B_RESULTS.md; research/PI05_PAIRS_PREREG.md (sha256 be2147c6..., mirrored to cloud storage *before* mask generation) with research/PI05_PAIRS_RESULTS.md, atlas rows experiments/forecast/atlas_pi05_pairs.csv, and per-run results plus checkpoint-curve evaluations mirrored to the cloud archive; research/ADAPTIVE_EVAL_PROGRAMME.md; the mixture-pricing artifacts experiments/influence/mde_mixture.py, MDE_MIXTURE_LOG.txt, and MIXTURE_MDE_RESULTS.md (Tables A1–A3, Appendix A); and the per-episode success vectors with harvest manifest experiments/forecast/PER_EPISODE_MANIFEST.json. All session waves are mirrored to the supplementary data archive.

2026-08-19 to 2026-09-06 artifacts (Appendix L): research/ROUTE1_PREREG_2026-08-29.md and research/ROUTE1_RESULTS_2026-08-29.md with scorers experiments/influence/score_route1.py, score_route1_secondary.py, route1_crosseval_check.py, route1_budget_curves.py and route1_within_run.py; research/SEEDSEL_PREREG.md, SEEDSEL_RESULTS.md and their two follow-ups with score_seedsel.py; research/SEGMENT1_PREREG.md and SEGMENT1_RESULTS.md; the directories research/kr1_metaworld/, research/cleanrel50/, research/steer1/ and research/kr2_rung1/, each holding its pre-registrations, freeze manifests, results documents, scorers and scored replicate files; and research/FLOOR0_*, DISCRETE1_*, PIN1_*, PIN_DISCRETE_*, STREAM1_*, HARNESS_SHARE_1_*, CKPT_SELECTION_LOTTERY_* and CKPT_COLLISION_SCAN_* with experiments/influence/score_floor0.py, score_hs1.py and verify_phase_e.py. Raw evaluation outputs of the harness waves, the LoRA checkpoints and the Meta-World demonstration files are mirrored to the supplementary data archive.

J POST-FREEZE PRE-REGISTERED RESULTS (2026-08-10 TO 2026-08-14)

The main text was frozen and verified against DRAFT v0.6 on 2026-08-09. Six scored analyses landed in the following five days; this appendix reports them in full, under the same rules as everything else in the paper: every verdict is scored against bands frozen before analysis, nothing here renegotiates a frozen band, and both readings are carried wherever a frozen rule requires it. All $\pi_{0.5}$ values in this appendix carry the schedule stamp truncated, 20k steps / 30k decay; the completion effect is untested for this class (Appendix K.1). Three waves added **24 training runs** (8 characterization, 8 production-scale k=88, 8 multi-task k=22), and the pre-committed PERTASK-2 follow-up (scored 2026-08-14, reported below) 8 more; these post-date the corpus census in §6 and are not counted in the 621. The remaining two analyses are zero-GPU: one observational re-analysis of existing runs and one instrument audit. The waves of 2026-08-16 to 08-21 follow in Appendix K.

J.1 PI05-SCALE: THE FRONTIER CELL AT PRODUCTION SCALE

Two eight-seed waves on byte-identical episode sets drawn from the full 10-task LIBERO-object pool (P=454), pre-registered jointly (prereg frozen and sha256-mirrored before any mask of either wave existed; 8/8 runs accepted in both waves, 0 gate failures). W1 (multi-task k=88): $\sigma_{\text{run}} = 4.05$ pp (df = 7, χ^2 95% CI [2.68, 8.24]), mean SR 61.38. W2 (multi-task k=22): $\sigma_{\text{run}} = 6.02$ pp (df = 7, CI [3.98, 12.25]), mean 35.75. W2's four disk-exhaustion casualties were caught by the realised-budget gate of reporting standard (2) and re-run clean from scratch—the gate functioning as designed.

The three-cell ladder for the $\pi_{0.5}$ class is therefore: single-task k=22 (pool 44) **9.74** (df = 10) → multi-task k=22 (pool 454) **6.02** → multi-task k=88 (pool 454) **4.23** pp. *Supersession note on the k=88 endpoint:* W1's own reading is 4.05 at n=20/task; the pre-registered PERTASK-2 wave (below) re-measured the same cell at n=100/task ($n_{\text{eval}} = 1000$, suite binomial floor 1.53 pp), giving $\sigma_{\text{run}} = 4.23$, and 4.23 is the cell's comparator everywhere in this paper—4.05 is quoted only as the superseded n=20/task reading. Derived quantities move with the supersession: the $\pi_{0.5}$ per-task/suite ratio reads $11.00/4.23 = 2.60$, not 2.72, against the $\sqrt{10} \approx 3.16$ baseline that independent per-task effects predict by construction (§4.1). The 2.30× total drop (2.40× as first computed on the superseded endpoint) decomposes as 1.62× (single-task → multi-task at k=22) ×

$1.42\times$ (k=22 $\rightarrow$ 88 within pool 454): a $\sim$58/42 log-scale split between task breadth and k—neither change alone explains the drop, and they contribute comparably. Reading W1 alone as a pure k-effect would have attributed roughly half of a task-breadth effect to data scale; W1 was not safely interpretable without its anchor wave.

The frozen production-k verdict is **S3, INDETERMINATE**: S1 ("the lottery survives") required $\hat{\sigma} \geq 6.0$ with CI-lo above the binomial floor (no: $4.05 < 6.0$; $2.68 < 3.44$), and S2 ("noise collapses") required $\hat{\sigma} \leq 3.5$ with CI-hi < 6.81 (no: $4.05 > 3.5$; $8.24 > 6.81$). Both readings are carried and no claim is strengthened on this wave. Nor is "σ_{run} exceeds the sampling floor at k=88" a resolvable claim: $\hat{\sigma}$ minus floor is 0.61 pp at n=200, below the programme's measured 2 pp resolvability floor—asserting it would repeat the error PI05-CHAR's H1 made (below). At multi-task k=22 the same claim *is* resolvable: excess 4.97 pp $= 1.47\times$ the floor, $\hat{\sigma}$ minus floor $= 2.63$ pp. The drop ratios ($2.30\times$ on the superseding 4.23 endpoint, $2.40\times$ as first computed; against the single-mask 12.64, $\mathbf{2.99}\times$ on the superseding 4.23 endpoint and $3.12\times$ as first computed on 4.05—both readings given, because the neighbouring ratios in this sentence honour the supersession and this one had not) are **exploratory, not pre-registered**, and labeled so wherever quoted. Pre-registered secondaries: the k-slope $6.02 > 4.05$ (4.23 at the superseding n=100/task convention) is directional only (CIs overlap); the bimodal collapse mode behind the k=22 cell was **falsified at** k=88 (P1: 0/8 runs $\geq$ 15 pp below the median of the others; largest gap 4.50 pp); the $\pi_{0.5}$-vs-SmolVLA noise premium shrinks with k (SmolVLA references $\text{df} = 2$—no quotable multiple; direction only). One attribution question remains open, stated as a hypothesis: part of the $1.62\times$ task-breadth term may be pure arithmetic averaging (a 10-task mean is less variable than a 1-task mean), in which case per-task retrain instability is unchanged and suite-average reporting hides it. The per-task decomposition below is the zero-GPU test of exactly this.

J.2 PI05-CHAR: CHARACTERIZING THE COLLAPSED MASK

Eight further seeds on `null04` alone (the mask whose first seed scored 50.0) give $\sigma_{\text{run}}(\texttt{null04}) = \mathbf{12.64}$ pp ($\text{df} = 9$, χ^2 CI [8.69, 23.08]), mean 77.50—the largest retraining variance in the corpus at this df, against a sampling floor of 2.95 pp ($4.3\times$ excess). The atlas keeps $\sigma_{\text{run}}(\pi_{0.5}\text{-ft}) = 9.74$ ($\text{df} = 10$) as headline; **this wave corroborates it rather than undermining it**, and the frozen robustness reading (5.23 pp at $\text{df} = 9$, χ^2 CI [3.60, 9.55], excluding `null04`) remains published as the robustness arm only—adopting it as headline would exclude the corpus's highest-variance mask on the grounds that its variance was a fluke, which this wave refutes (H4: 12.64 on one mask exceeds the six-mask pooled value).

The wave's confirmatory prediction was scored against itself: H1 (a new seed below 60.0) is **FALSIFIED**—0/8 below threshold, minimum 60.5, missed by 0.5 pp (one episode on a 200-episode grid)—recorded without renegotiation, and simultaneously ledgered as a design fault of the *prediction*: the 0.5 pp margin sits inside the programme's measured $<$2 pp resolvability floor, so the verdict was never resolvable at n=200. H1's frozen consequence (demote 9.74) did not execute because its stated premise was refuted by H4: the rule fired; its stated reason did not hold; recorded, not silently resolved. The cell's collapse count carries both frozen readings, neither promoted after the fact: **1/10** under H1's 60.0 threshold (H2 PASS, at the boundary of its [1/10, 4/10] band) and **2/10** under H3's $\geq$ 15 pp gap criterion (a 17.0 pp gap places both 50.0 and 60.5 in the low cluster).

The wave also surfaced and repaired a silent corpus defect: two `set_hash` conventions coexisted (the atlas builder's `sha1(",".join(sorted(eps)))` vs the corpus's `sha1(str(sorted(eps)))`), which prevents same-episode-set runs from grouping as replicates *with no error raised*. 28 values across four $\pi_{0.5}$ files were corrected (6 also un-truncated); post-fix, 350/350 atlas runs carry 16-character hashes and `null04` groups as 10 replicates across three waves, mechanically reproducing 12.64. One pre-existing instance remains open: `atlas_k4b.csv` carries 9-character hashes on all 8 rows, so K4B runs do not yet participate in hash-grouped replicate pooling (the K4B numbers in §4.3 come from that wave's own frozen scoring scripts, not from hash grouping); which convention is renormalised is an open decision.

J.3 PERTASK: THE PER-TASK VARIANCE DECOMPOSITION

A zero-GPU decomposition of the existing multi-task per-episode shared-seed vectors (prereg frozen and mirrored *before the analysis script existed*; integrity: 16/16 per-run SRs recomputed identical to

the atlas, and both waves' suite σ_{run} reproduced exactly through the same grouping). The question is the averaging hypothesis above: does the multi-task suite average hide single-task-sized per-task instability?

At k=88 (W1) the instrument saturates and the frozen primary is **INDETERMINATE**: floor-subtracted per-task $\sigma = 5.03$ pp with jackknife 95% CI **[0.00, 16.41]**. WEDGE-LIVES required $\sigma \geq 8.0$ with CI-lo > 5.0 (no); WEDGE-DIES required $\sigma \leq 5.0$ with CI-hi < 8.0 (no). The point estimate sits 0.03 pp above the DIES edge and no reading of proximity is permitted—the CI conditions fail decisively in both directions. The cause is arithmetic: at n=20 episodes/task the per-task binomial floor (≈ 10.9 pp) equals the raw per-task SD (10.92), so the floor subtraction differences two nearly equal numbers. What *is* directly measured, true under either hypothesis: on the fixed 20-episode battery, **a per-task success number moves $\sim$11 pp across retrains while the suite average moves 4**—a per-task dashboard entry swings $\sim 2.7\times$ more than the headline average—with how much is episode sampling vs true per-task instability undecided (the raw 10.92 is carried alongside the binding 5.03 wherever quoted). The mechanism secondaries ($R = 7.26$; mean cross-task correlation 0.075) are recorded to prevent over-reading: the pure-global rival predicts $R \approx 7.5$ and $\rho \approx 0.04$ from the floor alone, so they do not discriminate at n=20.

At multi-task k=22 (W2; $\pi_{0.5}$, `libero_object` pool) the pre-registered secondary **resolved**: floor-subtracted per-task $\sigma = \mathbf{12.01}$ pp (jackknife CI [8.70, 15.31]), raw 15.33—and the mechanism check is decisive there, because raw 15.33 exceeds the pure-global ceiling (≈ 11.8 given suite 6.02), so **a task-idiosyncratic component is required at** k=22. Two tasks sit at 0.0 for all 8 seeds while the worst swings 22.0 pp (mean 57.5); the composition yields a comfortable 6.02 suite average. The licensed statement: at k=22 multi-task on `libero_object`, the per-task retrain lottery is real, single-task-sized (the 9.74–12.64 class), and suite-average reporting halves it. Directional secondaries, never headline: per-task σ falls with k ($12.01 > 5.03$); $\pi_{0.5}$ exceeds SmolVLA's per-task cells at k=88 (SmolVLA references at 2 pairs/df=1, several truncated at zero). The pre-committed follow-up (PERTASK-2, frozen 2026-08-13, sha256 `b7163823...`) re-runs the k=88 cell at 100 episodes/task, dropping the per-task floor to ≈ 4.9 pp so the two hypotheses separate, under identical bands and a frozen escalation cap: if it lands INDETERMINATE again, no third wave is licensed.

PERTASK-2 resolved the cell (scored 2026-08-14, 8 fresh retrains, 8/8 accepted; these postdate the 24-run census above): per-task $\sigma = \mathbf{11.00}$ pp, jackknife CI $[6.86, 15.13]$, raw 11.82 — **WEDGE-LIVES**, on the $\pi_{0.5}$/`libero_object` multi-task suite at n=100/task. The same runs give the cell's superseding suite $\sigma_{\text{run}} = 4.23$ pp ($n_{\text{eval}} = 1000$; suite floor 1.53 pp; χ^2 CI [2.80, 8.61] at $\text{df} = 7$; per-task jackknife CI [6.86, 15.13] on the 11.00), so the per-task/suite ratio is $11.00/4.23 = 2.60$ floor-subtracted (raw $11.82/4.23 = 2.79$; under fully task-idiosyncratic noise the raw ratio would be $\sqrt{10} \approx 3.16$ by construction, since a 10-task average shrinks idiosyncratic noise by $\sqrt{10}$). At n=100/task the mechanism statistics discriminate: a purely global seed effect (suite σ 4.23 plus the mean per-task floor $\approx$4.4 pp) predicts raw per-task SD $\approx$5.9 pp and cross-task correlation $\approx$0.46; observed are 11.82 pp and 0.011 (`research/PERTASK2_RESULTS.md` §3), so the lottery is task-idiosyncratic. The k-slope secondary did not survive ($12.01 \to 11.00$: per-task instability barely moves with k while the suite average falls $6.02 \to 4.23$). The resolved secondary landed in the programme's favour—the first pre-registered result to do so since the K4B flip—and is recorded as such. See `research/PERTASK2_RESULTS.md`. Two later waves priced the SmolVLA class directly (Appendix K.2): per-task retrain σ is suite-dependent in magnitude (6.69/7.33 on `libero_object` vs 3.29/4.52 on `libero_spatial`, SmolVLA-ft k=88, floor-subtracted/raw), so every quoted per-task σ names its suite; the honest pricing sentence is that per-task retrain σ is a property of the (model, suite) cell—measured 3.3–11.0 pp floor-subtracted across the three VLA cells priced so far—and suite averages hide it wherever it is large.

J.4 Checkpoint promotion, held out: an observational companion

Not pre-registered. An observational split-half analysis of all 28 existing $\pi_{0.5}$ runs (4 checkpoints each, 200 shared-seed episodes; deterministic stdlib script, no RNG—halves are defined by episode *index*, which on LIBERO does not pin the initial state across separate checkpoint evaluations, Appendix K.1), reported as such; the confirmatory version—a frozen prediction on a future wave, and a second policy class—has not been run. Its implementation validation reproduces the published cell-1 winner's-curse figures (+4.00 pp mean / +12.50 pp max, §4) exactly.

Selecting the best-of-4-checkpoints on one half of the episodes and scoring on the other: the pooled held-out advantage of the selected checkpoint over the final one is $-\mathbf{0.38}$ pp (run-weighted, sem of order 0.6)—indistinguishable from zero, and if anything negative—while the same runs report a naive max-over-checkpoints gain of $+1.73$ pp. The honest bias measured under split-half selection is $+2.82$ pp, scoped to selection on n=100 (selection on the full 200 episodes would incur *less* bias); it is not comparable to the naive figure and the two are never quoted as if they were. Per cell, the curse ordering tracks measured σ_{run}: cell 1—the cell that *carries* the `null04` collapse, and whose own σ_{run} readings are 2.02 ($\text{df} = 2$, superseded) and **9.74** ($\text{df} = 10$) on the expanded 16-run cell it belongs to—shows $+4.00$ naive / -0.94 held-out; (the 12.64 at $\text{df} = 9$ that the source document attaches to "cell 1" is `null04`'s *single-mask* σ from PI05-CHAR, a different quantity, and is corrected here;) the pairs cell $+0.69$ / -0.62; the k=88 multi-task cell (σ_{run} 4.05 at this wave's n=20/task evals; 4.23 at the superseding n=100/task convention) $+0.94$ / $+0.12$—selection bias is a function of the regime's noise, not a constant (observational, three headline-eligible cells, one policy class). The four-run multi-task k=22 row is excluded from any headline: those runs are the survivors of the W2 disk failure, selected by surviving. Argmax concordance across halves is 14/28 (50%): above the 25% four-way-chance baseline, but the excess is largely the genuine monotone 5K→20K improvement, not discrimination among late checkpoints. Scope: $\pi_{0.5}$-ft only; four fixed checkpoints, so the measured curse is a lower bound for teams selecting over denser sweeps; halves are interleaved (even/odd episode index within task) because a contiguous split on a 10-task eval would confound selection with task difficulty.

J.5 G3: THE PAIRING SCOPE AUDIT

A pre-registered bit-level replay audit of the evaluation instrument. Frozen verdict: **FAIL under the driver's own regime; PASS with the env rebuilt per rollout.** Replaying one episode five times on a long-lived env (the SHARED regime—which is how `lerobot_eval` evaluates every run in this corpus) produced 0/10 bit-identical trace pairs; rebuilding the env per rollout produced 10/10. The mechanism is the render path, not the physics: MuJoCo state after reset is bit-identical ($0.000\text{e}{+00}$), all 34 proprioceptive elements identical, physics identical for 60/60 steps under fixed actions—but rendered images differ by up to 207/255 over $\sim 6\%$ of pixels and stay divergent for 60/60 frames. The power check passed (the comparator fires on genuinely different initial states); the root cause inside MuJoCo/robosuite is unidentified (two candidates, not distinguished); the repair (env rebuilt per rollout, now the default in the handoff driver) costs $+3.1$ s/episode, $+13.4\%$ wall clock in its one measurement.

Scope for every pairing claim in this paper: *rollouts are a function of (init state, position in the episode sequence), not of the init state alone*—and, on LIBERO, the position-to-state mapping is itself *outcome-dependent*. A later instrumented probe (ANNEAL-1 §12.8, Appendix K.1; validated 23/23 against directly logged `init_state_id` and per-reset MuJoCo state hashes) established that `LiberoEnv` advances its init-state pointer on every reset and that a successful episode fires one additional seedless autoreset, so $\text{pointer}(k) = (k + \#\text{successes strictly before } k) \bmod 50$: two runs desynchronise as soon as their outcomes diverge. Measured on the ANNEAL-1 control, only **10.06%** of same-index cross-run pairs share an initial condition (56,000 pairs; flip rate 0.1723 on aligned pairs vs 0.3742 on misaligned, 0.3539 overall). An earlier version of this paragraph claimed that identically-ordered runs remain CRN-paired at episode level and that episode j of task t has a bit-identical initial state in every run; on LIBERO both claims are **withdrawn**. What survives: pairing at *run and suite* level (matched protocol, matched episode sets, matched n-prefix); and *within-arm* prefix identity—the pointer mapping is independent of `n_episodes`, so a run's first 100 episodes are byte-identical to the first 100 it would have run at n=100, which is what licenses matched-n prefix comparators. Whether the render-path contamination *biases* success rate, rather than merely perturbing individual rollouts, is not measured and is asserted in neither direction (the bounding measurement—one checkpoint, SHARED vs FRESH at n=200—is identified but unrun); it is not, on this evidence, a reason to distrust the corpus's headline numbers. Audit scope: one task, one policy class, one checkpoint, one GPU; cross-suite generalisation is an inference, not a measurement.

K POST-FREEZE PRE-REGISTERED RESULTS (2026-08-16 TO 2026-08-21)

Eight scored, pre-registered analyses landed between 2026-08-16 and 2026-08-21, reported here—together with one convention audit and one clearly-labeled exploratory note—under the same rules as Appendix J: every verdict is scored against bands frozen before analysis (pre-registration sha prefixes in Appendix D), nothing renegotiates a frozen band, and every σ quote carries both readings (raw and floor-subtracted) wherever a frozen rule requires it. Five waves added **48 training runs**—8 ACT on PushT (Leg B1), 8 SmolVLA Leg A′, 16 ACT on ALOHA (B2), 8 ANNEAL-1, 8 PERTASK-4—none counted in the 621-run census (Appendix C). The rest train nothing: ACT-HORIZON re-evaluates the eight B1 checkpoints at two horizons; SOUP-1 evaluates weight averages of existing runs; REPEATEVAL-ZERO and the σ-transferability audit are zero-GPU reanalyses of banked data.

K.1 THE TRUNCATED-SCHEDULE DISCOVERY AND ANNEAL-1: THE LOTTERY SURVIVES SCHEDULE COMPLETION

Every VLA σ in this paper was measured on a mid-anneal iterate, because the ecosystem's default recipe stops training before its own learning-rate schedule finishes—and a pre-registered completion experiment finds the per-task lottery unchanged. LeRobot's official SmolVLA fine-tuning documentation prescribes `-steps=20000` while `SmolVLAConfig` defaults `scheduler_decay_steps` $= 30\,000$ (verified against upstream source): cosine progress at stop is 0.6552 and the learning rate at step 20,000 is still **28.4% of peak**. The string `decay_steps` appears zero times in our repository, so every SmolVLA and $\pi_{0.5}$ fine-tune in this corpus follows the truncated schedule—as does, by default, every recipe-following fine-tune in the ecosystem. Banked two-checkpoint data showed the concern was live: on the eight $\pi_{0.5}$ k=88 runs ($n_{\text{eval}} = 1000$ at both checkpoints), mean suite SR moved $55.05 \to 62.26$ ($+7.21$ pp) from step 10k to 20k with per-run deltas $-1.8 \ldots +18.3$, and suite σ_{run} fell $8.85 \to 4.23$—still halving at the stopping point. An unknown fraction of the per-task lottery could therefore have been incomplete-annealing dispersion. The documentation defect stands regardless of the result below.

Design. ANNEAL-1 ran 8 SmolVLA-ft retrains byte-identical to Leg A′ (`libero_object`, k=88, set `2d4a570b7025a761`, seeds 100–107, eval seed 4242, n=100/task) in every respect except one flag: `-policy.scheduler_decay_steps=20000`, which completes the cosine anneal inside the same 20,000 steps (final LR 2.5% of peak against the control's 28.4%)— isolating schedule shape, not training length or compute. The intervention is provenance-enforced, not assumed: `scheduler_decay_steps` $=$ 20000 was verified per model from its banked `train_config.json`. All integrity gates passed 8/8. One provenance defect is disclosed rather than repaired (noted in §6): the wave's `model_result.json` files inherit training-gate `arm` and `eval_seed` labels from the verbatim parent driver; the banked bytes are unmodified, and the per-chunk sidecars record true provenance (eval seed 4242), which the scorer establishes per model.

Result (frozen order). A1, the binding endpoint: annealed per-task $\sigma = \mathbf{6.936}$ pp, jackknife CI $[5.562, 8.309]$, against the frozen n=100 control 7.275 $[6.076, 8.473]$—**LOTTERY-SURVIVES**. Raw readings: 8.149 vs 8.439. A2, mean suite SR: 53.775% vs 54.650% ($\Delta = -0.875$ pp, no flag). A3, suite σ_{run}: 2.017 vs 2.371 (directional only). Completing the schedule changes per-task σ by -0.34 pp—deep inside both CIs—and the mean by -0.9 pp: the cross-policy atlas comparisons are not confounded by anneal position at this cell, and the 10k→20k mean gain had already plateaued by the stopping point.

Mandatory stamps, carried on every quote of this result. **(i) CONCORDANT:** the comparator set was CLOSED at three before scoring (first-100 7.275 $[6.076, 8.473]$, primary; second-100 5.457 $[4.408, 6.506]$; full-200 6.686 $[5.602, 7.770]$), and the full-200 comparator also reads LOTTERY-SURVIVES; the disclosed selection texture stands (the first-100 realisation sits at the 92nd percentile of random 100-episode subsets of the control). **(ii)** SURVIVES-region coverage for a fresh 8-run arm is **∼70–74%, not 95%**. **(iii) The §12.8 stamp, verbatim:** "episode-level CRN pairing between arms is NOT established (§12.8); the comparator is matched-n, and under §12.4's frozen residual-honesty branch the untransferred dispersion excess biases toward ANNEAL-EXPLAINS"—the direction that did not win. (§12.8 is the correction of record, written mid-battery with the endpoints unscored, that discovered LIBERO's outcome-dependent init-state pointer: $\text{pointer}(k) = (k + \#\text{successes before } k) \bmod 50$, validated 23/23 against instru-

mented ground truth; only 10.06% of same-index cross-run pairs share an initial condition, with aligned-pair flip rate 0.1723 against 0.3742 misaligned. This finding scopes every pairing claim in the paper—Appendix E.) **(iv) Dispersion lower-bound stamp:** the binomial floor assumes iid episodes, and the eval stream is measurably sub-binomial (dispersion factor $\hat{c} = 0.588$ of binomial variance, $[0.41, 0.77]$, exact test $p = 0.0012$), so **6.936 and 7.275 are lower bounds**, never absolute σ quotes; correctly-floored control values are 7.78 (n=100) and 6.96 (n=200), the A1 *contrast* is common-mode safe, and $\hat{c}$ does not transfer across policy classes or SR levels. **(v) Operating characteristics:** false-EXPLAINS 0.112 under the frozen-vector convention (0.081 only under a random-control convention, quoted with that label or not at all); power against true σ reductions of 20/30/40% is 0.58/0.84/0.96, so this null is informative against the strong form of the artifact hypothesis and reductions below $\sim$20% remain unresolved by design.

Consequences and scope. Every VLA atlas row now carries the schedule scope stamp (*schedule: truncated, 20k steps / 30k decay*); the SmolVLA-ft k=88 row additionally records this null; $\pi_{0.5}$ rows say *completion effect untested*. Per the frozen kill-switch K-D, **no second anneal wave is licensed, whatever anyone wishes this had said**. No transfer to $\pi_{0.5}$, ACT, DP, other suites, other k, or real robots; no claim about training to 30k steps (only about decay shape within 20k); no episode-paired contrast between arms; no selection among comparators. Prereg chain and corrections of record: Appendix D.

Named limitation: the undertraining threat. ANNEAL-1 answers one form of this objection, and the strongest form is not the one it answers, so we state that one ourselves rather than leave it to a reader. *The threat, at full strength.* Every VLA σ in this paper is measured at a stopping point the ecosystem's recipe chose, not at convergence—and the only two-checkpoint evidence in the corpus shows the lottery still shrinking fast at that point. On the eight $\pi_{0.5}$ k=88 runs, suite σ_{run} falls $8.85 \rightarrow 4.23$ between step 10k and step 20k: **a 52% fall over the last doubling, still halving at the exact step where training stops**. Taken as a trend that is a serious artifact hypothesis—a converged σ could be materially smaller than every VLA number we print—and it is not aimed at a peripheral value: 4.23 is the $\pi_{0.5}$ atlas row and one of the σ_{run} figures the body quotes. The mechanism is visible in the same data and is entirely ordinary: over that interval the mean gains 7.21 pp with per-run deltas $-1.8 \ldots + 18.3$, so runs that were behind catch up, and catch-up compresses spread. *What ANNEAL-1 bounds.* Schedule *shape* at fixed compute, in one cell. Completing the cosine decay inside the same 20,000 steps moves per-task σ by -0.34 pp (6.936 vs 7.275, deep inside both CIs), suite σ_{run} by -0.354 (2.017 vs 2.371, directional only), and the mean by -0.875 pp—so the residual 28.4%-of-peak learning rate is not what keeps the lottery large on SmolVLA-ft / `libero_object` / k=88, and the 10k $\rightarrow$ 20k mean gain had already plateaued there. Power against true σ reductions of 20/30/40% is 0.58/0.84/0.96. *What is not bounded, plainly.* **(a) Training length.** ANNEAL-1 holds steps fixed at 20k by design: it is a decay-shape experiment, not a convergence experiment, while the $8.85 \rightarrow 4.23$ fall is a *length* effect. Nothing in this corpus measures σ_{run} at 30k or 40k steps in any cell, so "the lottery keeps shrinking with training length" is **untested here, not refuted**. **(b) Class and quantity.** The falling reading is $\pi_{0.5}$ suite σ_{run}; the null is SmolVLA-ft per-task σ. Under this paper's own transfer rules neither licenses the other. **(c) Resolution.** Two checkpoints is one segment, at $n = 8$ runs, with no CI quoted on 8.85; and reductions below $\sim$20% are unresolved by ANNEAL-1's design. **(d) The non-VLA half of the atlas.** No ACT or DP cell has a two-checkpoint reading at all, so undertraining is untested for `pusht-act`, for `pusht-dp-rebuilt`, and for both ALOHA rows—including the 6.515 endpoint that carries the widest σ_{run} we report. *What we therefore claim, and what we do not.* The atlas prices **the lottery a practitioner following the ecosystem's default recipe actually faces**, which is the deployment-relevant quantity and is exactly what the schedule stamp records (§3). It does *not* establish an intrinsic lottery for a converged policy, and no sentence in this paper should be read as claiming one. Settling it needs a *length* ladder—σ_{run} measured at several step budgets with the decay completed at each—which is a different pre-registration from ANNEAL-1's: kill-switch K-D bars a second anneal wave, and no length ladder has been pre-registered here.

K.2 Per-task σ on a second class and a second suite: Leg A′ and PERTASK-4

Leg A′ (scored 2026-08-18): the per-task lottery is in both measured classes. Eight fresh SmolVLA-ft retrains under the frozen Amendment-A protocol (`libero_object`, k=88, n=200/task) give per-task $\sigma = \mathbf{6.69}$ pp floor-subtracted / **7.33** raw, jackknife CI $[5.60, 7.77]$ (df = 7), against same-

wave suite $\sigma_{\text{run}} = 2.11$ ($n_{\text{eval}} = 2000$). The frozen bands pass on both conditions ($\hat{\sigma} \geq 5.0$; CI-lo $5.60 \geq 2\sigma_{\text{run}} = 4.23$): **T1 WEDGE-TRANSFERS**. An independent audit re-derived the result from raw per-episode vectors exactly and found it robust to leave-one-task-out (6.00–7.00, TRANSFERS in all 10) and leave-one-seed-out (6.44–7.02). Cross-task correlation is -0.036 against ≈ 0.46 expected under a global (suite-wide) seed effect: the lottery is task-idiosyncratic. The licensed headline becomes "2 classes, LIBERO, k=88"; no third SmolVLA wave is licensed.

PERTASK-4 (scored 2026-08-21): the suite question did not settle, and per-task σ is suite-dependent in magnitude. Eight further SmolVLA-ft retrains, Leg A′ verbatim except the suite (`libero_spatial`, pool 432, k=88, n=200/task, seeds 200–207), give per-task $\sigma = \mathbf{3.29}$ pp floor-subtracted (BINDING), CI $[1.18, 5.41]$ (df = 7) / **4.52** raw—by the both-readings rule, 3.29 is never quoted without 4.52—against same-wave suite $\sigma_{\text{run}} = 2.040$ (df = 7). The bands were carried verbatim from Amendment A §4, deliberately not re-tuned: HOLDS required $\hat{\sigma} \geq 5.0$ with CI-lo $\geq 2\sigma_{\text{run}} = 4.08$; SUITE-SCOPED required CI-hi < 4.08. $\hat{\sigma} = 3.29$ fails the HOLDS floor and CI-hi $5.41 > 4.08$ refuses SUITE-SCOPED: **T4-1 VERDICT: INDETERMINATE**. The prebound consequence is honoured exactly: both readings reported, the suite stamp stays as it is, and no generalisation is claimed in either direction—the headline remains "2 classes, LIBERO, k=88," neither widened to two suites nor narrowed to `libero_object` only. The secondary is decisive: $[1.18, 5.41]$ and $[5.60, 7.77]$ do not overlap, **T4-2 DIVERGENT**—though the interval gap is knife-edge (0.19 pp), so the practice, not the divergence's size, carries the weight—and per-task σ is suite-dependent in magnitude, so **every quoted per-task σ carries a per-suite stamp** (SmolVLA-ft k=88: 6.69/7.33 on `libero_object`, 3.29/4.52 on `libero_spatial`; never a pooled or unqualified number).

INDETERMINATE here means underpowered, not "no effect": the $\sigma \geq 5.0$ band was calibrated against 11.00 and 6.69, and against $\sim$3.3 the CI straddles the boundary by construction; task 1's 9.7 pp SD against a 3.4 pp floor is not noise. The *shape* of the lottery also differs, not just its size: on `libero_spatial` task 1 carries 67.4% of the excess variance and 4/10 tasks carry zero (dropping task 1 collapses $\hat{\sigma}$ to 1.70), while on `libero_object` the largest task carries 27.6% and 0/10 are zero. The raw per-task/suite ratio is 2.22 against the $\sqrt{10} \approx 3.16$ independence baseline (`libero_object`'s is 3.47), and cross-task correlation turned positive ($+0.141$): on this suite more of the retrain draw is shared across tasks. The pricing sentence all of this licenses: **per-task retrain σ is a property of the (model, suite) cell—measured 3.3–11.0 pp floor-subtracted across the three VLA cells priced so far—and suite averages hide it wherever it is large.** No third suite is licensed by this prereg. Schedule stamp: truncated schedule, inherited ecosystem default; ANNEAL-1 measured the completion effect as null on `libero_object`; untested on this suite.

K.3 Cross-wave seed reproducibility (exploratory)

Exploratory; not pre-registered; banked data only; no verdict. Two independently launched training waves three days apart share one cell: $\pi_{0.5}$-ft on `libero_object` at k=88, byte-identical episode lists, matched train seeds, separate invocations (the first wave's on-box artifacts were reclaimed before the second launched, and the checkpoint cadences differ). At both saved steps the paired cross-wave suite differences are *smaller* than evaluation sampling alone predicts under identical weights: observed sd 1.89/1.98 pp against an expected 3.36/3.26 pp (χ^2 2.47/3.23 on 8 df), so the excess attributable to the weights is 0.00 pp; matched-seed suite correlations are $+0.980$ (step 10,000) and $+0.886$ (step 20,000). Same-wave σ_{run} across seeds is 4.23 pp. In this cell, then, the lottery is drawn by the seed: change the seed and the effect is large; relaunch the same seed and the checkpoint is statistically indistinguishable. Indistinguishable is not bitwise-identical, since shared evaluation randomness makes the expected-sd model conservative, and determinism here may be a property of this software stack. Determinism is also not predictability: which seed yields good per-task behavior is unknown without training and evaluating it, arms are retrained with free-running seeds in practice, and every sizing result in this paper therefore stands, with σ_{run} still the price of a comparison. An attempted second-cell check on banked PushT/DP driver replicates was indeterminate for want of a same-checkpoint re-evaluation floor in that cell, so nothing beyond this cell is claimed. Analysis and provenance: `research/CROSSWAVE_DETERMINISM_2026-08-27.md` in the supplementary data archive.

K.4 THE ACT DOSSIER: B1 ON PUSHT, B2 ON ALOHA

Leg B1 (scored 2026-08-16) delivers the first ACT σ_{run} row—at a degenerate operating point. Eight ACT retrains on PushT at pure LeRobot defaults (`chunk_size = n_action_steps = 100`: three open-loop decisions per 300-step reactive episode), $n_{\text{eval}} = 500$, 8/8 accepted: $\sigma_{\text{run}} = \mathbf{0.784}$ pp raw, χ^2 CI $[0.52, 1.60]$ (df $= 7$) / **0.590** floor-subtracted (floor 0.516 pp, **43% of variance**), mean SR 1.35% (per-run SRs 0.2–3.0%). **LOW-SR scope stamp, mandatory on every quote of this row:** ACT at chunk-100 open-loop is a near-floor policy on PushT; σ in pp is mechanically compressed near $\text{SR} = 0$; this cell must never be read as "ACT is a low-noise class"; and the row *understates* ACT's PushT retrain lottery by $\sim$6.2$\times$ raw / 7.7$\times$ floor-subtracted (ACT-HORIZON, below). It remains the correct row for anyone running LeRobot ACT defaults, which is what the Hub contains.

Leg B2 (scored 2026-08-19) completes the dossier with two ALOHA rows—unpaired by demotion. The frozen §7 rule demoted the wave before scoring: the G2 observation-hash check licensed pairing in 0/6 checks, so the harness makes *no* `gate()` call and computes σ_{run} only, and every `model_result.json` carries the stamp `pairing: UNLICENSED-G2FAIL-20260818`. 16/16 runs accepted (df $= 7$ per task):

	`transfer_cube`	`insertion`
mean SR (n=500)	77.875%	11.975%
σ_{run} raw (χ^2 95% CI)	**6.515** pp [4.31, 13.26]	**2.261** pp [1.50, 4.60]
binomial floor (share of variance)	1.856 pp (**8%**)	1.452 pp (**41%**)
floor-subtracted	6.245 pp	1.734 pp
σ as % of mean SR	8.4%	18.9%

`transfer_cube` is the healthy-regime ACT measurement B1 could not provide: 18.8 pp separates two byte-identical-data retrains at a full 50-episode set, with only 8% of the variance attributable to sampling. `insertion` carries its own LOW-SR stamp (41% floor share: much of its 2.261 is sampling, not lottery). No cross-cell multiple is quotable from either row: the ratio against `pusht-dp-rebuilt` has a 95% F interval of $[0.58, 9.26]$—it **contains 1** (df 7 vs 2, and the denominator sits below its own eval floor, Table A10)—so `compare_sigma_run()` refuses it. What the pair *does* establish is the SR-regime confound within a single wave: same class, same k, same protocol, and σ_{run} in pp differs 2.9$\times$ purely from where the SR sits—quoted only as the within-wave SR-regime confound, never as a pricing multiple—while in *relative* terms the ordering reverses (8.4% vs 18.9% of mean). Floor shares use the iid-binomial convention; $\hat{c}$ is unmeasured on ALOHA streams, so the 8%/41% splits are convention-dependent. No paired statistic, no per-task decomposition (single-task envs: per-task $\sigma \equiv \sigma_{\text{run}}$), and no cross-wave ACT comparison without checking the recorded FFmpeg major—decode is part of the data path.

K.5 ACT-HORIZON: THE OPERATING POINT IS PART OF THE STORY, AND THE LOTTERY IS NOT AN ARTIFACT OF IT

A pre-registered re-evaluation of the same eight B1 checkpoints at two decision horizons ($n_{\text{eval}} = 500$ each, no training anywhere; only `-policy.n_action_steps` differs) returns **T-H1 INDETERMINATE by construction**: the frozen bands required agreement at *both* horizons, and h=8 (pre-committed, non-argmax) lands INDETERMINATE (CV 29.7%) while h=16 (the sweep argmax) lands ARTIFACT-CONFIRMED (CV 21.9%). This is precisely the failure mode the prereg was written to catch: had h=16 been run alone, the verdict would have read ARTIFACT-CONFIRMED and would not have been trustworthy. The pre-bound consequence is honoured: INDETERMINATE licenses **no re-wording of the `pusht-act` scope stamp in either direction**.

Both readings, every horizon: h=100 (frozen B1) $\sigma_{\text{run}} = 0.784$ raw / 0.590 floor-subtracted; h=8 **4.849** $[3.21, 9.87]$ / 4.559 (floor share 12%); h=16 **4.839** $[3.20, 9.85]$ / 4.469 (15%). The measurement refutes the audit hypothesis that motivated it ("the eight retrains are behaviourally interchangeable and B1's σ_{run} is threshold-crossing noise"): at healthy operating points the same eight checkpoints spread 12.2 pp (h=8) and 15.4 pp (h=16), and σ_{run} does not collapse under correction—it rises **6.2$\times$ raw / 7.7$\times$ floor-subtracted**. What survives of the audit's claim: B1 measured in a place where the lottery could not show up (floor share 43% $\rightarrow$ 12–15%), so `pusht-act` understates ACT's PushT lottery by $\sim$6$\times$. σ_{run} is horizon-stable in pp (4.849 vs 4.839) while the

mean moves $16.35 \rightarrow 22.10\%$, so CV falls through its denominator, not through any change in spread. Seed rankings reshuffle across horizons only directionally: tie-corrected Spearman $+0.195$ (h=100 vs 8), $+0.317$ (h=100 vs 16), $+0.595$ (h=8 vs 16)—none significant at $n = 8$. Atlas rows `pusht-act-h8`/`-h16` are written with `sr_range` and an EVAL-HORIZON stamp; `pusht-act` is left exactly as it was. **Selection disclosure on `-h16`:** h=16 was the sweep argmax, so its SR level is optimistically biased and must never be quoted as "ACT achieves 22% on PushT"; when a single σ_{run} figure is needed, it is h=8's 4.849, chosen in advance. Pre-check 11b confirms the three ACT/PushT cells refuse to price one another.

K.6 No transferable σ_{run} law: the atlas is a sizing tool, not a pricing oracle

Across the eight atlas cells that record both a realised SR and a σ_{run}—enumerated row by row, with each cell's n_{eval} and binomial floor, in Table A12 (Appendix H.2)—σ_{run} in pp spans $8.31\times$ ($0.784 \ldots 6.515$). **That multiple is the dispersion input to the transform analysis below. It is not a price**—the body quotes no cross-cell multiple at all (§4)—and three properties of this eight-cell set, disclosed here, are why.

(i) Three of the eight cells are one cell. `pusht-act`, `pusht-act-h8` and `pusht-act-h16` are the *same eight ACT checkpoints* re-evaluated at three inference horizons ($h = 100$, 8, 16; no training anywhere, Appendix K.5), not three independent measurements. The CV across cells therefore counts one training wave three times, and the span's low endpoint plus two of its interior points come from that one wave.

(ii) The low endpoint is a figure we pre-committed not to quote alone. 0.784 is the `pusht-act` row at 1.35% mean SR, which Appendix K.4 records as *understating* ACT's PushT retrain lottery by $\sim 6.2\times$ raw / $7.7\times$ floor-subtracted, and whose pre-committed replacement—"when a single σ_{run} figure is needed, it is h=8's 4.849, chosen in advance"—is fixed in Appendix K.5. A span whose low endpoint is a number we have bound ourselves not to quote as *the* ACT/PushT σ_{run} is not a span we may quote either.

(iii) The remaining low end is convention-split. The analysis reads `pusht-dp-rebuilt` at 1.800 (repair pairs, $\text{df} = 2$) while Table A10 and §4 print $\mathbf{1.292}$ ($\text{df} = 4$, pooled over all four post-rebuild pairs) for the same cell. Both readings are published and neither is wrong. The *eight*-cell endpoints do not depend on the choice; the six-cell endpoints below do.

The arithmetic, done under our own rules. Drop the two duplicate horizon rows and take the licensed 4.849 for the ACT/PushT checkpoint set, and six independently measured cells remain: 2.261 (`aloha-insertion-act`), 1.800 or 1.292 (`pusht-dp-rebuilt`), 4.849 (`pusht-act-h8`), 2.113 (`smolvla-ft-k88`), 4.228 (`pi05-ft-k88`), 6.515 (`aloha-transfer-cube-act`). The span is then $6.515/1.800 = \mathbf{3.62}\times$ on the analysis's DP reading, $6.515/1.292 = \mathbf{5.04}\times$ on Table A10's, and $6.515/2.113 = \mathbf{3.08}\times$ if that DP row is not admitted at all—which `compare_sigma_run()` would enforce, since its floor share is 85% at 1.800 and exceeds 100% at 1.292, either way past the gate's 75% ceiling. **Six independent cells give 3.1–$5.0\times$, not $8.31\times$**, and that is before the refusals the gate additionally raises on evaluation scope. **This is why §4 quotes a range and a refusal in place of a multiple**: what we can say is that σ_{run} runs from at or below a cell's own binomial floor (`pusht-dp-rebuilt`) to 6.515 pp on bimanual ALOHA and 11.43 pp in a small-data LIBERO cell (Table A10), and that we decline to price any one cell against another.

The transform analysis, and the convention audit that caught our own over-claim. No scale-free transform we tested collapses that spread under the floor-subtracted convention. The claim is convention-dependent, and an audit caught our own strong form over-claiming: three choices decide the answer and all three must be stated—(1) *which* σ is transformed (raw or floor-subtracted), (2) *which* p (the row's measured mean SR), and (3) *which* delta-method factor (logit $1/(p(1-p))$; arcsine $1/(2\sqrt{p(1-p)})$). The convention is published as runnable code (`experiments/influence/audit_2026_08_18/sigma_transferability.py`), with dispersion measured as CV across cells. We print it twice—over the eight cells as the published script computes it, and over the six that survive (i)–(iii), with the horizon rows collapsed to h=8 alone—so that the cell-set choice is visible rather than ours to make silently:

	eight cells (convention of record)		six independent cells (h=8 only)	
transform	CV on RAW σ	CV on floor-sub.	CV on RAW σ	CV on floor-sub.
identity (pp)	57.3%	69.1%	51.9%	**66.4%**
CV (σ/mean)	89.5%	87.9%	73.4%	87.0%
logit	58.8%	61.0%	53.2%	67.8%
arcsine	**45.6%**	**59.1%**	**51.5%**	66.5%

The six-cell columns drop `pusht-act` and `pusht-act-h16` and keep `pusht-dp-rebuilt` at the script's 1.800; bold marks the tightest entry in each column. Lower is more transferable, and every column is a CV over 6–8 numbers, so gaps under a few points decide nothing.

On raw σ over the eight cells the arcsine transform *beats* pp (45.6% against 57.3%), and it beats pp under floor subtraction too (59.1% against 69.1%), so the strong form ("every scale-free transform is worse than pp") is **falsified**—it was our claim, and it was wrong until the convention was written down. **The six-cell columns do not un-falsify it and are not offered as though they did**: the eight-cell set is the convention of record, frozen in the published script, and a cell set re-picked after seeing the answer cannot retract an error. What those columns do show is where the arcsine margin came from—collapse the duplicated horizon rows and drop the disowned low-SR endpoint and it narrows from 11.7 CV points to 0.4 (51.5% against pp's 51.9%), and under floor subtraction pp becomes the tightest of the four by 0.1 (66.4% against 66.5%). Both of those are coin flips at six cells and we read neither as vindication. The weak form is what everything rests on, and it stands on every column of the table: σ_{run} is cell-specific, differs several-fold between settings that look alike, and no transform we tested makes it transferable under the floor-subtracted convention. Consequences, enforced in code: the atlas is a *sizing* tool and an existence proof per measured cell, not a pricing oracle for unmeasured cells; a cross-cell σ_{run} multiple may be quoted only when `atlas.compare_sigma_run()` returns one (it gates on recorded n_{eval}, a resolved σ convention, floors on the same scale, both dfs, floor share below 75%, and a ratio CI excluding 1); and a σ_{run} below its own binomial floor is a *measurement, never a prior*—pre-check 11c warns on the record that pricing from a floor-dominated regime understates the lottery (`pusht-dp-rebuilt`, 1.292 vs floor 1.682, is the live example).

K.7 SOUP-1: AVERAGING ALL EIGHT RETRAINS INTO ONE POLICY IS VIABLE AT THIS CELL

A uniform weight average of all eight Leg A′-protocol parents is a deployable policy at this cell: the frozen primary (m=8) scores paired $\Delta = \mathbf{-1.375}$ pp against its *own* parents' mean, 95% CI $[-3.33, +0.58]$, on matched init states (SHIM-1 P-arm)—**SOUP-VIABLE**, with no stratum reaching DEGRADES. The mandatory null calibration is quoted beside every Δ: a matched-norm *random* weight step of the 8-soup's length landed at $+1.8$ pp on its single draw, and the measured step-cost spread at this step length is ± 11.6 pp, so the soup's -1.4 sits comfortably inside what step-length effects produce in either direction and no cost is attributed to averaging itself. The dispersion secondary decides nothing and is reported once with its frozen stamp: per-task σ of the four disjoint 2-soups is 4.503 (raw 6.176), S=4 CI $[0, 11.94]$. Scope: **"this soup of these parents"**—m=4/8 carries no population degrees of freedom; the four m=2 rows are a disclosed re-slice of SEGMENT-1's banked M-arm, not independent evidence; SmolVLA-ft, `libero_object`, k=88, truncated-schedule recipe, n=100/task, simulation. One look; **no SOUP-2 is licensed**, and nothing here revises any frozen upstream value. What it licenses: the cure hypothesis is a funded question rather than a speculation—a single soup, as one artifact, carries no retrain-selection lottery downstream, published as a falsifiable, properly-powered future design with this pilot as its evidence.

K.8 REPEATEVAL-ZERO: AT ITEM LEVEL, THE MAJORITY OF THE "RETRAIN LOTTERY" IS ROLLOUT SAMPLING

A zero-GPU decomposition of item-level flipping, riding SEGMENT-1's banked SHIM-1 harvest, in which every chunk contains a perfect within-checkpoint repeat: episodes k and k+50 share a byte-identical initial state (validated by sim-state hash; alignment gate ≥48/50 passed by all 80 parent chunks). Led, per its independent-review addendum, with the difference rather than the ratio: f_{within} (same checkpoint, same state, pass 1 vs pass 2) = $\mathbf{0.1450}$ (per-model 0.124–0.184; 4,000 pairs); f_{between} (different checkpoints, same state and pass) = $\mathbf{0.2291}$ (28 pairs × 1,000 items); $\Delta = f_{\text{between}} - f_{\text{within}} = \mathbf{0.0841}$ (exactly $6{,}416/28{,}000 - 580/4{,}000 = 0.08414$; both components

are printed rounded, so the difference is quoted from the unrounded counts), an unbiased estimate of $2\,\mathbb{E}[\mathrm{Var}_{\mathrm{seed}}(q_{\mathrm{item}})]$, giving $\mathrm{sd}_{\mathrm{seed}}(q_{\mathrm{item}}) = \sqrt{\Delta/2} \approx 0.205$; and $\Phi = f_{\mathrm{within}}/f_{\mathrm{between}} = \mathbf{0.633}$, quoted second (conventions 0.6328 same-pass / 0.6306 all-pass-pairs, both carried), against the frozen band $\Phi \geq 0.50$: **SAMPLING-MAJORITY**. Φ is a *lower bound* (within-process pairs can only be independent or positively correlated, which deflates f_{within}; observed McNemar $z = -0.83$ shows no measurable dependence). The honest label, adopted from the same review, is **descriptive decomposition with disclosed prior visibility**—the unfrozen draft had quoted the f_{within} range, and one feasibility cell ($f_{\mathrm{within}} = 0.120$) was disclosed in the frozen prereg—never a blind confirmatory test; the pre-bound band consequence stands.

Consequences. The Φ stamp is mandatory on every item-level claim in this corpus: $\sim$63% of outcome flipping between identically-trained SmolVLA checkpoints at a single (task, initial-state) episode is the policy's own stochastic sampler (flow-matching noise drawn from the global CUDA RNG, verified in source), not the training seed, so nobody should read a single episode flip as a seed effect. **Aggregate-level claims are untouched**: suite and per-task success rates average the sampling term over 100+ episodes, and the retrain lottery measured at those levels (σ_{run} 2.11–4.23 at multi-task k=88, per-task σ 3.3–11.0 pp) already survives its binomial floor and remains seed-driven. Souping does not reduce the sampler's stochasticity (midpoint models 0.156 vs parents 0.145): the sampler's noise is a property of the sampler, not of the weights' position. Scope: SmolVLA-ft stochastic flow-matching sampler, `libero_object`, k=88, 8 checkpoints, simulation; one look, no follow-up licensed; nothing here revises any frozen aggregate-level value.

K.9 Exploratory: cross-cell structure of the per-task lottery (no verdicts)

Exploratory, $n = 3$ cells, computed from banked data after the confirmatory waves scored; no bands, no verdicts, no consequences. Its only legitimate use is as the hypothesis source for a future frozen pre-registration. On the same ten `libero_object` tasks measured by two model classes: which tasks are *hard* transfers almost perfectly (per-task SR correlation $+0.904$ between SmolVLA-ft and $\pi_{0.5}$), while which tasks are *unstable* under retraining barely transfers at all (per-task excess-variance correlation $+0.273$; rank $+0.224$). Read as a hypothesis: task difficulty is model-portable, task instability is not—the per-task lottery would then be dominated by a (model $\times$ task) interaction rather than a task property, consistent with the near-zero cross-task correlations within models. Were it to survive a frozen test, no atlas could price a new model's per-task risk from task identity alone; the cell would have to be measured. The honest negatives are recorded with it: mid-SR concentration of excess variance is directionally present but inconsistent across cells; no cell-level law relates per-task σ to mean SR or task spread; and the harder-task-bigger-lottery correlation flips sign across cells. The candidate pre-registrations it generates (a scale axis: per-task σ 6.69 for the 450M SmolVLA vs 11.00 for the larger $\pi_{0.5}$ at fixed suite and k, two points only; a held-out profile-transfer band; a mid-SR band) run only under a new frozen prereg and do not touch K-D or PERTASK-4's no-third-suite clause.

K.10 DEVLOOP: running the collect–retrain–report loop with the fix removed (exploratory)

Exploratory. Not pre-registered, no verdict token, no consequence. A re-analysis of three already-scored waves (PERTASK-2, PERTASK-3 Leg A′, PERTASK-4) asking a question none of their pre-registrations posed. Script `experiments/influence/devloop_redraw.py`; machine output `DEVLOOP_RESULTS.json`; narrative `research/DEVLOOP_RESULTS.md`.

Design. The main text argued in four places that a loop collecting data for a failing task, retraining and reporting the suite average cannot separate a fix from a fresh draw, and each time called it untested. It is testable at zero compute, because the corpus holds repeated retrains on *byte-identical* data: run the loop with the intervention deleted. Take run A, target its worst task, retrain changing nothing — same episode set, budget and schedule — and read that task in run B. Every movement is the lottery by construction. All $8 \times 7 = 56$ ordered pairs per cell.

Three arms and one null, because the targeted number alone would mislead. *Targeted* is the real loop and contains regression to the mean by construction: the worst task of a noisy draw is expected to improve on re-measurement even with no lottery. That is not a confound to be stripped out — it is half of why the loop is fooled — but it is named. *Random task* removes it. *Suite average* is what

gets reported, and is exactly antisymmetric over ordered pairs. The *binomial-only null* holds each task's ability fixed at its across-run mean and varies only rollouts, answering "is this just evaluation noise?"

Table A13: **DEVLOOP, three cells.** Targeted arm with 2,000-fold bootstrap CI over retrains. *Churn* is how often two retrains disagree about which task is worst. Suite-average mean is $+0.00$ in all three cells by antisymmetry.

cell (k=88)	σ_{pertask}	ep/task	churn	targeted mean	$\geq$10 pp	max	binomial null (99th)
$\pi_{0.5}$-ft / `libero_object`	11.00/11.82	100	**68%**	**+12.96** [3.94, 22.67]	**48%** [22, 75]	$+62.0$	$+2.85$ ($+5.61$)
SmolVLA-ft / `libero_object`	6.69/7.33	200	25%	$+1.22$ [0.00, 3.21]	7% [0, 25]	$+13.5$	$+0.58$ ($+2.13$)
SmolVLA-ft / `libero_spatial`	3.29/4.52	200	0%	$+0.00$	0%	$+6.5$	$+0.00$

The mechanism is rank churn, and it is the finding. Exposure is not set by how large the lottery is but by whether it is large enough to *reorder* the tasks. If the same task is bottom of every retrain, a developer targeting it and redrawing sees only that task's own noise, symmetric about zero — annoying, not misleading. The loop is misled only when retraining reshuffles which task is worst, because then the target sat at the bottom partly by luck, regression to the mean pays out on the next draw, and the suite average stays flat while it does. Churn is ordered with σ_{pertask} across the three cells and the targeted arm with churn—three points, a consistency observation rather than a fitted relationship. On `libero_spatial` task 5 is bottom in all eight retrains, so the statistic is exactly antisymmetric and returns $+0.00$: a structural zero, not a measured null.

Robustness, and why this reading is preferred to the 20.4 pp one. The same cell's worst-task sd of 20.4 pp is a maximum over ten tasks that one retrain carries: deleting it drops that task to 7.81 and promotes a different task to worst at 17.77, against 10.51 for σ_{pertask} under the same deletion. The DEVLOOP statistic survives every single-run deletion: targeted mean $+9.10$ to $+14.95$, $\geq$10 pp 40–57%, above the null's 99th percentile throughout. It is the same phenomenon measured in a form that does not rest on one run, which is why Figure 2 now argues from the aggregate.

Scope. One of three cells, and absent in both SmolVLA cells at the same k, one at the same suite: never a property of LIBERO, of VLAs, or of imitation learning, but of the cell, as every σ here is. The $\pi_{0.5}$ cell evaluates 100 episodes/task against the others' 200, so its per-task floor is higher; the binomial-only null prices exactly that and the observation clears it four-fold, but the cells are not floor-matched and no cross-cell multiple is quoted from them. It does not show that curation loops fail — only that in an exposed cell the evidence such a loop collects cannot separate a fix from a redraw, which leaves the audit's *unverified as reported, not refuted* framing unchanged.

L Post-freeze results II (2026-08-19 to 2026-09-08): the per-task lottery outside LIBERO, and what a release step can do with it

Sixteen further analyses scored against numeric bands frozen before their data existed, under fifteen pre-registration documents (DISCRETE-1 and PIN-1 share one), together with one declaration-form analysis and four clearly labeled exploratory notes, landed between 2026-08-19 and 2026-09-08 under the rules of Appendices J and K: no band renegotiated, every verdict quoted in its frozen language, and every σ carried raw wherever no floor convention was frozen for it. They answer the question that §4.1 raises and answers only in summary: whether the per-task lottery is a property of LIBERO, and whether anything can be done with it beyond pricing it. The short answer is that the lottery replicates in a second simulator with a fifth policy class at zero GPU cost, that a per-task choice among a fixed pool of retrains converts it into a gain wherever the suite offers per-task complementarity to harvest, and all but removes the damaged-task tail that a suite-average release ships, and that its origin is chaotic: fixing the initialization or the batch order alone leaves it intact.

What trained, and what did not. These waves sit **outside the 701-policy corpus** and outside its budget-assertion gate (§7); each carries its own integrity gates, listed in its results document, and none of its runs is counted anywhere in §6. They trained 112 small multi-task networks on Meta-World (Yu et al., 2019; McLean et al., 2025) on a laptop CPU (24 for KR-1, 24 for CLEANREL-50, 16 for CLEANREL-50B, 16 for SOUP-PT-1, 32 for STEER-1; 1.4 seconds per retrain) and 8 rank-64 LoRA fine-tunes of SmolVLA on one rented A100. Everything else re-scores banked per-episode

vectors or re-evaluates banked checkpoints. Because five variables change at once between LIBERO and the Meta-World cells (simulator, task family, observation space, policy class, demonstration source), the transfer claims below are disjunctions, and no LIBERO magnitude is claimed to carry. The no-transferable-σ_{run} rule of §4 applies to every constant here: each is a property of its cell. Wave names such as SEEDSEL, ROUTE-1 and KR-1 are the internal labels of their pre-registrations and carry no meaning beyond identification.

The ledger. Ten of the sixteen scored outcomes landed against the hypothesis that motivated them or in an INDETERMINATE band, and each is reported as it fell: suite-average selection damages tasks (SEEDSEL S3); the payoff forecaster is MARGINAL (SEEDSEL-3); the basin structure is UNRESOLVED (SEGMENT-1); a post-hoc trend is a FLUKE (FLOOR-0); determinism flags FAIL to remove the harness floor (PIN-1); the scheduling hypothesis is REFUTED (STREAM-1); rank-64 compression FAILS to carry the routed policy (KR2-RUNG1) and trained rank-64 adapters are INDETERMINATE (LORA-LOTTERY-1); naive weight averaging FAILS as a single artifact (SOUP-PT-1); and steering the lottery by batch order is INDETERMINATE (STEER-1).

L.1 Selecting among retrains: the suite winner damages tasks, the per-task winner rarely does (SEEDSEL, ROUTE-1)

SEEDSEL (scored 2026-08-19; prereg sha `e73f5766`..., frozen before any selection statistic existed). On the eight $\pi_{0.5}$ retrains of Figure 2 (k=88, n=1000 per model, banked CRN vectors, no new episodes), a split-half design selects on one half of each task's episodes and reads the other. Shipping the retrain with the best suite average beats shipping one at random by $+6.09$ **pp** held out (S1, SELECTION-PAYS; stamp CONFIRMATORY-WEAK, because the eight suite rates were visible before the freeze and bounded this quantity at $\max - \text{mean} = 6.94$). Selecting and reading on the same episodes overstates it by $+1.08$ pp, inside the frozen $[0.3, 1.5]$ winner's-curse band (S2). The blind primary is S3: the suite winner is at least 5 pp below the pool mean, on held-out episodes, on **1.82 of 10 tasks**, its worst task sits -10.76 pp below, and it damages at least one task in 97.5% of draws: **SUITE-SELECTION-DAMAGES-TASKS**. A task at least 5 pp below the retrain pool's mean on held-out episodes is what "damaged" means in this appendix; Appendix L.3 adds a significance requirement and says so where it applies. The mechanism is §4.1's: with cross-task correlation near zero, the retrain that wins on the average is the one whose per-task damage canceled best in the mean, not the one that is good on each task. Four retrains capture 73.4% of the eight-retrain gain (S4, prediction HELD). Exploratory and unclaimed: the obvious per-task repairs make it worse, maximin over tasks paying $+0.30$ and damaging 2.38 tasks, a constrained argmax $+0.81$ and 2.34, because a per-task rate on 50 episodes carries a binomial SE near 6.9 pp and the minimum of ten such estimates is mostly noise. A follow-up (SEEDSEL-3, sha `d1dc5ca6`...) tested whether an atlas row forecasts the payoff and damage in three further cells and returned **FORECASTER-MARGINAL** (calibrated on 1 of 3, every miss an over-estimate of the tail); no forecaster is claimed.

ROUTE-1 (scored 2026-08-29; prereg sha `fffe7dda`..., frozen before any routing statistic). The same eight retrains, R=4000 stratified split-half draws: select on 50 episodes per task, evaluate on a disjoint 50. Task-conditioned selection ("routing") ships, for each task, the retrain that won that task on the selection half; a margin δ defects from the suite winner only when the per-task advantage exceeds δ pp.

Table A14: **ROUTE-1 primary cell.** $\pi_{0.5}$-ft, `libero_object`, k=88, 20k steps, J=8 candidates; held-out SR on the disjoint evaluation half, 4000 draws. *Damaged* counts tasks at least 5 pp below the pool mean; ORACLE picks per task on the evaluation half and is an upper bound, never a gain.

arm	held-out SR	gain vs RANDOM	damaged tasks	worst task
RANDOM (ship one retrain)	62.27	—	—	—
SUITE (ship the best-validation retrain)	68.34	+6.08	1.82	−10.76
ROUTE, δ=0 (per-task winner)	**73.94**	**+11.68**	**0.22**	**−1.95**
ROUTE, δ=10 (the frozen band arm)	72.76	+10.50	0.73	−5.54
ORACLE (upper bound)	78.76	+16.49	—	—

The frozen bands (ROUTE-10 gain $\geq +8.0$: $+10.50$; ROUTE-10 minus SUITE $\geq +2.0$: $+4.42$; damaged $\leq$ 1.82: 0.73) all pass: **ROUTE-DOMINATES**. The result recomputes SEEDSEL's $+6.09$ and 1.82 as $+6.08$ and 1.82 on a different RNG seed, so the new arms sit beside a reproduced old one. The δ grid ran monotone against the prereg's guess: δ=0 is both the highest-gain and

the lowest-damage arm, so the margin written to protect the tail cost it; the grid was frozen and is reported entire. Against the best-validation retrain, routing adds $+5.60$ **pp** and cuts damaged tasks from 1.82 to 0.22.

Secondary cells, texture only, no verdicts. Adding each seed's banked 10k-step checkpoint (16 candidates) gives $+17.90$ pp over RANDOM against SUITE's $+9.69$; the independent $\pi_{0.5}$ k=22 cell (n=200) gives ROUTE minus SUITE $+4.39$; the same primary policies read on their n=200 evaluation, 10 selection episodes per task, give $+0.13$**, the honest negative**: below about 20 selection episodes per task the advantage is unreliable. Forty-eight SmolVLA runs spanning twelve *data* curations give SUITE $+12.00$ and ROUTE $+12.87$: routing pays where candidates are comparable on the average and differ in which tasks they solve, which is the lottery, and adds little where one candidate is simply better. *Robustness against LIBERO's pointer defect.* Selecting on one evaluation run and deploying on a physically separate evaluation run of the same eight policies (n=200 against n=1000, verified disjoint) gives RANDOM 62.26, SUITE 67.90 with one damaged task, ROUTE 70.00 with none, ORACLE 77.50; the (seed×task) competence profile correlates $+0.595$ across the two runs after task and seed main effects are removed. One draw, a consistency check, not an estimate. *Budget curves, post-hoc and exploratory.* At J=8 the advantage over SUITE rises from $+4.24$ at 5 selection episodes per task to $+5.78$ at 50; at 50 episodes per task it rises from $+2.96$ at J=2 to $+5.72$ at J=8 (separate exploratory draws, which is why the J=8, 50-episode corner reads $+5.78$ and $+5.72$ against the primary's $+5.60$); SUITE's damage count stays near 1.8–2.0 at every budget while ROUTE's falls to 0.2. Within a single training run, routing between its two banked checkpoints gains $+1.42$ pp over the final checkpoint (post-hoc; a real run saves many more). *Scope.* One suite family and two policy classes in simulation; task identity must be known at inference; J checkpoints are held and served (Appendix L.5); the comparison is against a pool the training budget already paid for, not a claim that training J seeds beats spending that compute otherwise. Policy routing over pools of *heterogeneous* policies is an established technique; routing among identical-recipe retrains, and the measurement that it is the per-task lottery being harvested, is what is new here.

Exploratory scope check, post hoc and not pre-registered: the gain does not replicate on a second LIBERO suite. PERTASK-4's eight SmolVLA-ft retrains on `libero_spatial` (Appendix K.2; k=88, seeds 200–207, n=200 per task per retrain) had never been routed. Running ROUTE-1's design on them afterwards, block split-half with 100 selection and 100 held-out episodes per task and 4000 bootstrap draws, gives ROUTE minus best-validation $+1.7$ **pp, 95% interval** $[-2.1, +5.4]$, zero inside, and $+1.8$ $[-2.5, +6.0]$ with the halves exchanged: twice ROUTE-1's selection budget, against $+5.60$ on `libero_object`. The *damage* result does replicate here, best-validation leaving 1.51 tasks at least 5 pp below the pool mean against routing's 1.02. This is what §4.1 predicts rather than an anomaly: routing harvests per-task complementarity, and on the same wave this suite carries σ_{pertask}=3.29 pp against `libero_object`'s 6.69. We therefore quote the routing gain only in the cells where it was pre-registered and measured, and Standard 5 of §7 binds it exactly as it binds σ_{pertask} itself. The per-episode vectors and this scorer (`experiments/influence/route1_spatial_scope_check.py`) are in the supplementary data archive, so the check reruns from the released records.

L.2 KR-1: THE PER-TASK LOTTERY AND ITS ROUTING GAIN REPLICATE ON META-WORLD MT10, FOR $0

Design (prereg sha `1e49f926...`, including Amendment A; all four scorers hashed before any wave ran, re-verified identical at scoring). Three independent replicates, each a fresh scripted-expert demonstration set (3 demonstrations per task) and J=8 identical-recipe retrains of a 23k-parameter multi-task MLP with a task one-hot on 39-dimensional state, 4,000 steps, differing only in training seed; 200 evaluation episodes per task per retrain under common random numbers (identical episode list, `eval_seed0`, `num_goals`); frozen bands ROUTE minus SUITE $\geq +2.0$ pp TRANSFERS, $\leq +1.0$ DOES-NOT-TRANSFER, median over replicates with all three on one side. Gates in the frozen order: G2 init-state diversity (≥ 150 distinct states in 200), G3 CRN, G4 recipe identity, G1 pool SR in $[35, 85]$, G5 a noise-only null, G6 the mid-band subset, G7 replicate stability. Apple M1 Pro laptop, 1.4 s per retrain, 48,000 evaluation episodes, no GPU.

Result. Median $D = +\mathbf{6.86}$ pp, all three replicates above $+2.0$ on the same side: **TRANSFERS**. Every gate passed; no interval straddles both band edges; no hashed file changed after freezing. G5

Table A15: **KR-1 (Meta-World MT10) and the CLEANREL-50 law (MT50).** Top: held-out SR by arm and the routing advantage D with its split-half 95% interval, the noise-only null (G5: every retrain given its task's pool rate, real binomial noise, no seed×task structure) and the damaged-task count under each rule. Bottom: MT50, the probability that the best-validation release ships at least one truly damaged task (priced above measurement noise) as suite size T grows, against the same checkpoints routed per task with abstention.

MT10 replicate	pool SR	best-val.	routed	D	95% CI	null	damaged, best-val. / routed
r0	70.73	76.50	80.65	$+4.16$	$[+2.30, +5.50]$	-0.09	0.61 / 0.00
r1	64.44	66.66	73.52	$+6.86$	$[+5.10, +8.70]$	$+0.08$	1.40 / 0.00
r2	70.51	72.65	80.09	$+7.44$	$[+5.30, +9.50]$	$+0.04$	1.54 / 0.03

MT50, $P(\geq 1$ damaged)	T=2	5	10	20	30	50	σ_{suite}, T=2 → 50
best-validation, r0	0.012	0.172	0.567	0.919	0.992	1.000	$10.7 \rightarrow 3.3$ pp
best-validation, r1	0.009	0.122	0.443	0.850	0.981	1.000	
best-validation, r2	0.010	0.102	0.340	0.767	0.951	1.000	
routed + abstain, all replicates	≤ 0.005 at every T; expected damaged tasks 0.00–0.01						

is the load-bearing gate: on a simulated corpus in which every retrain shares its task's pool rate the same estimator returns -0.09, $+0.08$ and $+0.04$ pp, so the estimator does not manufacture a gain. The per-task lottery is the same object as in LIBERO: one retrain solves `drawer-open-v3` 200 times in 200 and another, from the same demonstrations and recipe, 0 times, with per-task spreads of 100, 78, 49, 42 and 34.5 pp on the five widest tasks while the suite σ_{run} of the same pools reads 2.80, 2.12 and 4.44 pp (raw; no floor convention was frozen for this cell and none is quoted). Both exploratory curves reproduce ROUTE-1's shape: D rises monotonically with J ($+1.7$ to $+2.5$ at J=2, $+4.2$ to $+7.4$ at J=8) and with selection budget ($+1.9$ to $+4.5$ at 5 episodes per task, $+4.2$ to $+7.4$ at 100). *Licensed, verbatim:* per-task complementarity between identical-recipe retrains is not a LIBERO artifact; it reproduces in a different simulator (MuJoCo/Meta-World against robosuite/LIBERO), task family (ten heterogeneous Sawyer skills against ten object-manipulation variants), observation modality (state against pixels), policy class (a 23k-parameter MLP against a 3B VLA) and demonstration source (scripted experts against human teleoperation). *Not licensed:* anything about pixels, VLA scale, hardware or the LIBERO magnitudes specifically, since the five variables move together. Two weaknesses were stated in the prereg: the informative middle band is reached by a restricted budget, and the task one-hot could manufacture complementarity architecturally. Both are discharged in Appendix L.3.

A second init-state defect, in a second benchmark. Measured before freezing and recorded as Amendment A: in Meta-World v3, `env.reset(seed=k)` does not select an initial state; the state is a function of the environment's construction seed and its position in the reset sequence. At the package default of 50 goals, 200 distinct evaluation seeds yield **49 distinct initial states**; at 500 goals, 168. It was caught only because G2 was a gate rather than an assumption, and it had silently made three "independent" demonstration sets byte-identical until the knob was moved to the construction seed. It does not break CRN here (one explicit reset per episode, so the reset count is outcome-independent), unlike LIBERO's pointer (§4.2). Two benchmarks examined, two init-state defects found: §3's harness hazards name both.

L.3 CLEANREL-50 AND CLEANREL-50B: THE SUITE-MEAN RELEASE GATE GOES BLIND GEOMETRICALLY IN SUITE SIZE

CLEANREL-50 (scored 2026-09-01; prereg sha `e4a86877`…, all four scripts and the r0 demonstration file hashed before any MT50 statistic existed). Meta-World MT50, the KR-1 recipe, three replicates of J=8, 200 episodes per task per retrain, 240,000 evaluation episodes in 1 h 26 m on the laptop. Damage is *priced*: a task counts as damaged only when it sits at least 5 pp below the pool mean by more than two standard errors on held-out half-samples, because under the unpriced "5 pp below" call the structureless null itself reaches P=0.99 at T=50 (the RAW5 reading, reported so nobody quotes the wrong formula). Four frozen clauses, all passing in all three replicates: **LAW**, the measured $P(\geq 1$ damaged) over the suite-size grid tracks the one-line series-reliability prediction $1 - \prod_t (1 - \hat{q}_t)$ from per-task hazards to within 0.045–0.050 (band ≤ 0.15), so per-task damage hazards compound near-independently (residual cross-task $r = +0.03$ to $+0.06$); **BLIND**, the replicate sd of the reported suite mean falls to 0.31–0.35 of its T=2 value by T=50 (band ≤ 0.70); **BITE**, a

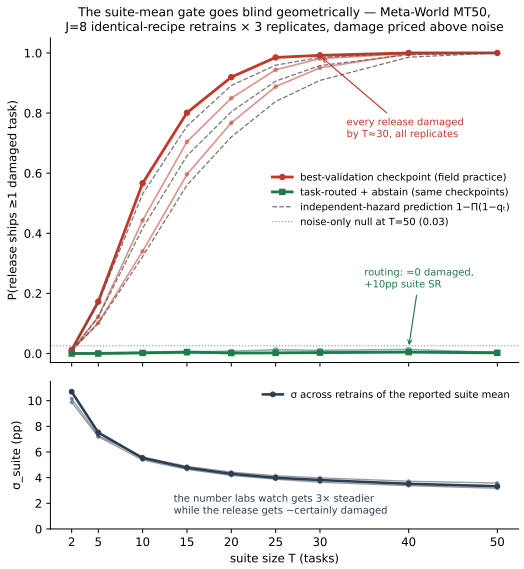

Figure A5: **The suite-mean gate goes blind as the suite grows (CLEANREL-50, Meta-World MT50).** Top: probability that a release ships at least one truly damaged task against suite size T, for the best-validation checkpoint (field practice) and for the same checkpoints routed per task with abstention; the dashed overlay is the independent-hazard prediction $1 - \prod_t (1 - \hat{q}_t)$ and the dotted line the noise-only null. Three replicates; the heavy lines are r0. Bottom: sd across retrains of the reported suite mean, which shrinks as $1/\sqrt{T}$ while the release becomes certainly damaged. Source figure from `research/cleanrel50/make_figure.py` over the scored replicate files.

best-validation release ships at least one truly damaged task with probability **1.000** at $T{=}50$ (band ≥ 0.50), carrying 4.1–6.1 damaged tasks; **ROUTE**, the same checkpoints routed per task with a pre-registered abstention rule (defect from the incumbent only at one-sided $z \geq 1.645$ on the selection half) hold $P \leq 0.005$ with 0.00–0.01 expected damaged tasks (band ≤ 0.25 and ≤ 0.5). **VERDICT: LAW-CONFIRMED + GATE-BLIND.** The abstaining router keeps the incumbent on 53–59% of tasks, retains about 91% of the routing gain, and pays $+9.9$ **to** $+10.9$ **pp** of suite SR over best-validation ($+10.7$ to $+11.2$ without abstention). Table A15 (bottom) and Figure A5 show the shape: at the ten-task scale where the field's reporting habits formed, damage is an even-odds event (0.34–0.57), easy to miss; by $T \approx 30$ every replicate ships damaged with probability at least 0.95, while the number used to certify the release becomes three times steadier. Leave-one-replicate-out, per-task hazards from two replicates predict the third's P at $T{=}50$ to within 0.001. MT50's own lottery, for the record (r0): `door-close` 100 and 0, `sweep` 100 and 4.5, `plate-slide-back` 93.5 and 0, `hammer` 83.5 and 0 across retrains differing only in seed, at suite σ_{run} 3.2–3.6 pp raw.

CLEANREL-50B (scored 2026-09-01; prereg sha `b4c4d8ec…`; single replicates, disclosed). Two owed legs discharge KR-1's stated weaknesses. **B1, no one-hot**: with the task one-hot removed end to end (state input only; the one-hot is worth about 10 pp of suite SR, pool 51.03%), $D = +\mathbf{10.33}$ pp (band $\geq +2.0$, SURVIVES), the law deviation is 0.026, suite $P(\geq 1) = 1.000$ against 0.004 routed, and the residual cross-task $r = +0.013$ is *lower* than with the one-hot: the one-hot was not manufacturing the independence. **B2, capacity**: at about three times the data and five times the compute (10 demonstrations per task, 20,000 steps, pool 78.89%), $D = +\mathbf{8.93}$ pp (SURVIVES), suite $P = 1.000$ against 0.000, and routing lifts the suite from 83.61 to 92.54%; the hazards shrink slightly with capacity but the compounding and the gain are intact at the higher operating point. *Licensed, verbatim:* in this cell the probability that a best-validation release ships at least one damaged task compounds with suite size as independent per-task hazards predict while the suite mean's replicate σ shrinks, and task-conditioned selection with abstention removes the compounding. *Not licensed:* VLA-scale policies, pixels, human demonstrations, hardware, or any other cell's constants; what is claimed transferable is the *mechanism*, measured in two Meta-World suites atop ROUTE-1's LIBERO complementarity. *Positioning.* $1 - \prod(1 - q_t)$ is textbook series reliability, and the regression-free-update literature owns the negative-flip framing (Yan et al., 2021); what exists nowhere else are the measured constants, per-task hazards, σ_{pertask} and cross-task independence, for identical-recipe retrain pools of robot policies, because they require $J \times T$ replicate matrices.

L.4 WHERE THE LOTTERY COMES FROM: INITIALIZATION OR BATCH ORDER ALONE SUFFICES, AND IT IS NOT DESIGNABLE (SOUP-PT-1, STEER-1, SEGMENT-1)

SOUP-PT-1 (scored 2026-09-01; prereg sha `b6fffa1…`). Pool B holds the initialization fixed (`init_seed`= 1000) and varies only the batch-order seed. Suite σ_{run} *halves* (3.33 to 1.80 pp raw) while the per-task lottery survives largely intact: maximum spread still 100 pp, mean spread 30.9 against 37.1, eleven tasks with spread ≥ 50 pp against fourteen, $D = +10.96$ pp against pool A's $+11.21$ under the same scorer, suite $P(\geq 1) = 1.000$ against 0.020 routed with abstention. **BATCH-ORDER-BORNE** by the frozen band. *STEER-1 (scored 2026-09-06; prereg sha `3e174f6a…`, frozen before any Leg-1 statistic and before any Leg-2/3 evaluation; gates G-FID 8/8 bitwise, G-STREAM 24/24 pure permutations, G-CRN 32/32 batteries).* Pool C completes the 2×2: fixed order, varied initialization, $D = +11.7$ pp (Leg 3, **INIT-ALSO-SUFFICIENT**), suite σ_{run} 2.11 pp.

pool (MT50, J=8)	init	order	D = route − suite	mean per-task spread	σ_{run} raw
A (CLEANREL-50 r0)	varies	varies	+11.4	37.1	3.33
B (SOUP-PT-1)	fixed	varies	+11.2	30.9	1.80
C (STEER-1)	varies	fixed	+11.7	29.4	2.11

The table's D values are STEER-1's Leg-3 re-score of all three pools with one split-half draw, which is why pools A and B read $+11.4$ and $+11.2$ here against $+11.21$ and $+10.96$ in their own waves. Either perturbation alone produces most of the lottery and all of the routing gain. Pinning one source makes the *certifying number* steadier and leaves the per-task lottery, and the routing gain, where they were: a reproducibility program that pins initialization or data order buys suite-level cosmetics, not release safety. Pool C's members share one batch stream and their residual profiles correlate $+0.23$ with the pool-A seed that used it and -0.01 with the others: the order leaves a fingerprint but does not decide the outcome. *Is the lottery a knob?* Leg 1a asked whether natural recency of a task's draws in a random stream predicts which retrain wins it: Spearman -0.02, -0.09, $+0.00$ across the three banked pools, exposure CV about 2%, **NO-PREDICTION**. Leg 1b asked whether same-stream, different-init retrains share a per-task profile: $r = 0.294$ against a band edge of 0.30, permutation $p = 0.0008$, **INDETERMINATE**, reported as it fell. Leg 2 designed the dose: for eight declared (seed, task) pairs, every draw of the target task placed in the final 10% of steps against the first 10%. The mean late-minus-early effect is $+22.8$ pp, bootstrap 95% CI $[-2.3, +54.0]$, positive in 5 of 8: **INDETERMINATE** (the $\geq +15$ clause met; the CI-lower-bound and 6-of-8 clauses not). Two pairs swung 80–100 pp (`sweep` 3 to 83, `plate-slide-back` 0 to 100), four did not move or moved the wrong way, and the *chaos reference*, the same reordering aimed at a different task, moved the target by 16.6 pp on average (maximum 29.5): an order perturbation scrambles tasks it was never aimed at with the same magnitude as the aimed effect. Per the frozen partition Leg 2 stays INDETERMINATE; no pairs were added and no band edited. *Consequences, in this cell:* the trainer's batch log is not a free pre-ranking signal, so a release step cannot skip the selection episodes; the lottery is not a data-ordering artifact that a fixed order removes; and designed complementarity is not licensed, so a measurement-based release step remains the way to harvest the gain.

Averaging does not replace the choice. A uniform soup of eight different-init retrains collapses to 4.39% suite SR (-56.07 pp, **COLLAPSES**); a soup of the eight shared-init retrains runs but loses 7.71 pp against its own parents (**PARTIAL**) and sits **21.04 pp below the routed policy**; in both pools the per-task argmax selects all eight members, so no small routed subset exists to average. Against the frozen single-artifact clause (any one artifact within 2 pp of the routed policy) this **FAILS** for the from-scratch cell. It is consistent with SOUP-1's parity (Appendix K.7), which averaged same-base SmolVLA *fine-tunes*: souping tolerates weight diversity born of fine-tuning a shared base (Wortsman et al., 2022), not of training from scratch. *SEGMENT-1 (scored 2026-08-21; prereg sha `b8a46d0e…`; SmolVLA-ft, `libero_object`, k=88, 200 evaluation chunks under SHIM-1 positional init-state enumeration, zero training)* places the fine-tuned lottery inside one basin: the straight-line midpoint between identically trained siblings costs **0.588 pp** (delete-one-parent CI $[-2.109, 3.284]$, **CONNECTED**), against 4.750 ± 11.604 pp for a random weight step of the same length, and the eight fine-tune deltas share about 76% of their direction (cosine 0.753–0.762). The dispersion ratio at the midpoint, 0.757 with CI $[0.209, 2.743]$, did not resolve (INDETERMINATE-E0), so the verdict is **ONE-BASIN-STRUCTURE-UNRESOLVED** and no second wave was licensed (Garipov et al., 2018). The reading across the three waves: the lottery is

a task-idiosyncratic wobble inside one connected basin, produced by either source of training chaos, seed-specific in its per-task expression, and not steerable at the precision a release would need.

L.5 THE GAIN IS HIGH-RANK: SERVING THE ROUTED POOL (KR2-RUNG1, LORA-LOTTERY-1)

KR2-RUNG1 (scored 2026-09-03; prereg sha `a5374f91`... with Amendment A, two mechanical scorer fixes frozen after the battery completed and before any success rate was read; 170 of 170 chunks; about $30 on one rented A100, zero training). The harness was rebuilt from the wave's environment manifest on a freshly rented machine two weeks after the retrains were trained, with upstream libraries having moved on; fresh originals reproduce the banked per-seed rates at a median $|\Delta|$ of 1.23 pp (H1, band ≤ 5.0), and the SHIM-1 battery pairs initial states exactly across all 17 models (H2, 100 of 100 identical everywhere). Routing on the *originals* replicates in this fresh harness, the fifth independent harness or cell to do so: the banked table pays $+7.4$ pp over best-validation and a fully fresh split-half estimate gives $D = +\mathbf{5.34}$ pp, 95% CI $[+1.60, +9.20]$. The question was whether base plus a rank-64 SVD truncation of each retrain's delta is a faithful serving form. It preserves the average policy (median paired suite drop $+2.20$ pp against a 3.0 band, SUITE-PRESERVED; median per-task profile $r = 0.962$) and erases the lottery: the routed compressed pool reads -0.60 pp against the best-validation *original* (clause: $\geq +2.0$), so **RUNG-1 FAILS**. Compression costs the routed policy 8.0 pp while costing the pool average about 1.6, and within the compressed pool fresh split-half routing yields $D = -0.60$ $[-4.80, +3.20]$: nothing harvestable remains even with a perfect table; 8 of 10 per-task argmaxes flip. The spectra say why: rank 64 holds 32.3% of the delta energy and rank 256 only 70.7%, so the routing signal lives in the high-rank remainder that truncation discards, while the surviving low-rank third carries the seed's average behavior. *The constructive finding needs no approximation:* 378 of 500 tensors are bitwise identical across all eight retrains, because the recipe never moves the SmolVLM backbone; the routed release therefore ships *exactly* as one shared backbone plus eight action-expert modules, about 2.5 GB against 7.2 GB for eight full checkpoints, with zero behavioral change. Recipes that unfreeze the backbone forfeit this.

LORA-LOTTERY-1 (scored 2026-09-06; prereg sha `a9c31937`...; about $85). Eight rank-64 LoRA fine-tunes (13.7M adapter parameters, the Leg A′ recipe otherwise verbatim, seeds 100–107) ask whether the lottery expresses itself inside the subspace when training is constrained to it. The pool reaches 41.88% suite SR (gate G-CAP ≥ 40, passed barely) against 53.68 for the full pool at identical data and steps, so the constraint costs about 12 pp of competence. The primary reads $D = +2.09$ pp, 95% CI $[-2.0, +8.2]$, spanning both band edges: **INDETERMINATE**, reported as such, with no seeds or episodes added to move it. The claim-free readings worth keeping: the per-task lottery *regrows* inside the subspace, larger, with mean per-task spread 32.1 pp (maximum 54) against the full pool's 20.8 and σ_{run} 5.04 pp against about 3.3, so whatever generates the lottery does not need high-rank freedom to express variance; and at that competence best-validation selection captures more of the pool's quality axis ($+7.6$ pp), leaving a noisier residual for per-task routing at 50-episode selection halves. Ladder state for a single deployable artifact: soups FAIL (SOUP-PT-1), SVD truncation FAILS (here), trained adapters INDETERMINATE; the exact modular bundle above is the shippable form today.

L.6 EXPLORATORY: EPISODES CARRY INTRINSIC IDENTITY, AND A PER-STATE CEILING SITS ABOVE PER-TASK ROUTING (KR6-PRECURSOR)

Declaration form, not a blind pre-registration: the three point estimates were seen in-session on 2026-09-03 before the nulls, intervals and wording bands were frozen (sha `e98ecb97`...); the per-state readings that follow are not pre-registered at all. The KR2-RUNG1 battery is the corpus's first *paired* replicate-pool outcome tensor (J=8 × T=10 × n=100, qpos-hash-verified identical initial states across models), which LIBERO's pointer defect had made unavailable (§4.2). E1: 447 of 1,000 episodes are unanimous across the eight retrains (203 all-fail, 244 all-pass) against 170.5 expected if episodes were coin flips at their task's pool rate, a $2.62\times$ ratio, $p \approx 0$: *episodes carry intrinsic identity*. E2: whether the other seven retrains solve a state predicts whether the held-out retrain does, a leave-one-out lift of $+71.9$ pp over 64 cells against $+0.05$ under permutation. E3: the ladder random ship 53.7, per-task routing oracle 63.6, per-episode oracle 79.7, so $+16.1$ pp $[13.1, 17.0]$ is reachable only by per-state selection; the independence null would cover 91.3, and

E1 is exactly what pulls real coverage below it, the structure that creates the per-state signal also bounding it. The rank-64 compressed pool keeps all three (E1 $2.82\times$, E2 $+70$ pp, headroom $+16.1$) while its per-task complementarity is dead: episode difficulty is robust and largely shared, and which retrain wins which task is the fragile high-rank signal. Decomposed with the battery's paired repeats (each of the 50 states per task rolled out twice per model): shared state difficulty, model$\times$state interaction and model main effect carry 0.52, 0.44 and 0.03 of the variance; the interaction reproduces across repeats at $r = 0.35$ against 0.81 for the per-task profile; an exact-state lookup router scores 65.5 against 62.0 for per-task routing ($+3.5$ pp, a fifth of the band); a nearest-neighbour router on the 110-dim initial-state vector for *unseen* states sits at the edge of noise ($+0.4$ to $+1.1$; null -1.2 ± 1.2) while the same vector predicts a state's shared difficulty at $r = 0.72$. The initial state knows which episodes are hard, not which retrain wins them. Two consequences are worth stating even unverified: every binomial interval in the field models an object that is not iid at the episode level, and difficulty-stratified initial-state sampling is available for free in this cell. One pool, one cell; no runtime signal is claimed.

L.7 The evaluation harness on a different stack: a discrete floor beneath the retrain floor (FLOOR-0, DISCRETE-1, PIN-1, STREAM-1, HARNESS-SHARE-1)

Stamp on every number in this subsection: DP on PushT at k=103, banked checkpoints re-evaluated on LeRobot 0.4.4, torch 2.9.1, one NVIDIA L4, 2026-08-29, a *different* software stack from the corpus. Nothing here is imported into any atlas value, §4.2's $\sigma_0 = 1.91$ pp is a different estimand (a full retrain at fixed seed, n=500, corpus stack), and the cross-wave check of Appendix K.3 stays INDETERMINATE: its verdict inverts inside this floor's own interval, and importing a floor across stacks is exactly what §4 forbids.

FLOOR-0 (prereg sha `db457a69...`). One checkpoint, J=10 re-evaluations, n=200, eval seed 1000 identical, weights and initial states fixed: success rates 37.5, 38.0, 39.0, 40.0, 42.0, 39.5, 44.0, 38.5, 35.0, 39.5, a range of **9.0 pp**, $\sigma_0 = \mathbf{2.452}$ pp (df 9, 90% χ^2 CI $[1.69, 4.48]$), within-checkpoint flip rate $f_{\text{within}} = 0.2184$. The first five repeats arrived strictly monotone ($p = 1/120$); the prereg froze a test on repeats 6–10 before they existed and it returned **TREND-FLUKE** (Spearman 0.176; second-half minus first-half mean $+0.00$). *DISCRETE-1 (prereg `32fd14d9...`, Amendment A `11891f92...`).* At J=20: $\sigma_0 = 2.241$ pp (df 19), $f_{\text{within}} = 0.2083$, and **6 of 190 pairs agree on all 200 labels**, forming exactly two closed triples, $\{6, 10, 12\}$ and $\{8, 14, 19\}$, whose members share 182–189 float-exact trajectories against a median of 6 for other pairs: **MODES-CONFIRMED**. Agreement is transitive, which is what an equivalence class looks like and coincidence does not. An earlier significance argument for the first colliding pair ($0.782^{200} \approx 10^{-21}$) was **withdrawn**: it conditioned on nothing, and given the 182 shared trajectories the figure is $0.782^{18} = 0.0105$; the combinatorial triple argument replaces it and needs no independence assumption. A uniform-mode estimate ($\hat{k} = 32$) is internally inconsistent with two triples in 20 draws, so no mode count is claimed, only existence and repeatability. *PIN-1 (same prereg):* with `cudnn.deterministic`, `use_deterministic_algorithms` and a fixed cuBLAS workspace verified per run, $f_{\text{within}} = 0.2623$ against the default 0.2184, band ≥ 0.15: **PIN-FAILS**; a first attempt whose flags silently failed to load was caught by the per-run verification line and discarded. *STREAM-1 (prereg `9265d16f...`):* removing rollout vectorization (`batch_size` 1 against 50) gave f_{within} 0.3283 against 0.2700, band ≥ 0.15: **STREAM-REFUTED**. Kernel autotuning, RNG stream alignment and continuous per-episode noise are excluded; simulator state, policy sampling outside `-seed` and GPU nondeterminism outside torch's determinism contract remain live. *HARNESS-SHARE-1 (prereg `0b2986fe...`, Amendment A `1ea0bbfc...` correcting two identifiers before data).* Five banked checkpoints forming four same-set different-train-seed pairs, each re-evaluated J=5 times: $f_{\text{within}} = 0.2530$, $f_{\text{retrain}} = 0.3860$, $H = f_{\text{within}}/f_{\text{retrain}} = \mathbf{0.6554}$ (band ≥ 0.50): **HARNESS-DOMINANT**, *item-level*. Two-thirds of the item-level flipping between independently retrained checkpoints is reproduced by re-running one of them. The *aggregate* share from the same 25 evaluations is 38.0% ($\sigma_0 = 2.146$ at df 20 against $\sigma_{\text{seed}} = 2.742$ at df 3), with a 95% interval of $[4.2\%, 65.6\%]$ that makes it unquotable alone; the mandatory stamp is that no sentence of the form "N% of the retrain lottery is the harness" appears without *item-level* or *aggregate*, and the earlier claim that REPEATEVAL-ZERO's $\Phi = 0.633$ corroborates H was withdrawn (within-process against cross-process estimands). The one operational reading, exploratory: averaging J evaluations per arm divides the harness term by J and leaves σ_{seed} untouched, so the

comparison sd for two retrained checkpoints falls at most 21.2%, two-thirds of it by $J{=}3$; to compare retrains, buy training seeds, not evaluation episodes. Aggregate-level results stand untouched everywhere: suite and per-task rates average this term over 100 or more episodes, and the retrain lottery at those levels survives its binomial floor. σ_0 on this stack is not a Gaussian noise parameter but the sd of a mixture over discrete execution modes, and sizing from it as a Gaussian would misprice.

L.8 Exploratory, zero GPU: the best checkpoint is itself a lottery, and no two runs share weights

Post-hoc, no pre-registration, no verdicts. *Checkpoint selection.* Nine PushT/DP data conditions and three ALOHA and PushT panel conditions, each with three training seeds, an in-training checkpoint sweep on $n{=}100$ evaluations and a selected best checkpoint: in **0 of 12 conditions** did the three seeds agree on which checkpoint was best (selected steps span the whole grid, and 5 of the 27 PushT sweep runs selected the final checkpoint), including ALOHA `transfer_cube` under ACT at about 80% success, so the pathology is not a near-floor artifact. The mechanism is the winner's curse on the checkpoint axis: at $p \approx 0.4$ one $n{=}100$ evaluation carries a binomial sd of 4.9 pp, and the argmax of eight such draws is noise unless true checkpoint differences exceed it, which late in training they do not. It is an independent leg for the final-checkpoint protocol of §3 that does not rest on the held-out promotion argument (Appendix J): even a reader who disputes that promoted gains fail to replicate must accept that the object being promoted is not identifiable. A tempting "selection inflates σ_{run} by 29%" comparison is **withdrawn** as a success-rate-regime violation; the licensed claim is the count. *Bitwise uniqueness.* Every `model.safetensors` under the archive carries a content hash: 1,578 checkpoints, 1,461 distinct values, 117 duplicate groups all accounted for by 109 mirror copies and 8 reuses of one run under two experiment prefixes, and **0 independently trained runs sharing a checkpoint**, across 263 run directories, three policy classes and five months. With Appendix K.3's controlled result (matched seed, data order and config, weights differing 8 of 8 at every step while behavior is statistically indistinguishable), bitwise and behavioral reproducibility are separate properties, and only the second is under the experimenter's control. One stamp for the corpus index: the drop-gate "full" ALOHA arm is a byte copy of the panel run, never two observations.

L.9 Pre-registrations, hashes and cost for this appendix

Frozen documents, by sha256 prefix, each hashed and mirrored to cloud storage before its data existed unless stated: `research/SEEDSEL_PREREG.md` (`e73f5766`), `SEEDSEL3_PREREG.md` (`d1dc5ca6`), `SEGMENT1_PREREG.md` (`b8a46d0e`), `ROUTE1_PREREG_2026-08-29.md` (`fffe7dda`), `kr1_metaworld/PREREG_KR1_2026-08-29.md` (`1e49f926`, with Amendment A), `cleanrel50/PREREG_CLEANREL50_2026-09-01.md` (`e4a86877`), `PREREG_CLEANREL50B_2026-09-01.md` (`b4c4d8ec`), `PREREG_SOUPPT1_2026-09-01.md` (`b6ffffa1`), `steer1/PREREG_STEER1_2026-09-06.md` (`3e174f6a`), `kr2_rung1/PREREG_KR2RUNG1_2026-09-02.md` (`a5374f91`, Amendment A frozen after the battery and before any rate was read), `PREREG_LORALOTTERY1_2026-09-03.md` (`a9c31937`), `KR6P_DECLARATION_2026-09-03.md` (`e98ecb97`, declaration form), `FLOOR0_TREND_PREREG_2026-08-29.md` (`db457a69`), `PIN_DISCRETE_PREREG_2026-08-29.md` (`32fd14d9`, Amendment A `11891f92`), `STREAM1_PREREG_2026-08-29.md` (`9265d16f`) and `HARNESS_SHARE_1_PREREG_2026-08-29.md` (`0b2986fe`, Amendment A `1ea0bbfc`). Each wave's freeze manifest lists its scorer hashes, re-verified at scoring. Two waves pre-registered on 2026-08-29 (PIN-2, FLOOR-N) were never run and no data for them exists; they are listed so that their absence is not read as selective reporting. Compute: the Meta-World waves and every re-scoring ran on one laptop for $0; KR2-RUNG1 and LORA-LOTTERY-1 cost about $30 and $85 of rented GPU time; the harness waves ran on one rented L4. Results documents and scorers are in the anonymous supplementary material accompanying this submission under the paths named in the source of this appendix.