# MEASURE, DON'T AVERAGE: SMALL-COHORT MEASUREMENT OF VLA POLICIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language-action (VLA) studies often compare only four to six policies, yet report benchmark means without testing whether items, ranks, and uncertainty are identified. We audit a 3,155-episode matrix over six variants and a 14,412-episode LIBERO-Plus spatial bank. Standard 2PL estimation converges while item discrimination is prior-sensitive at $N = 4$. A hierarchical Bernoulli model grounded in observed rollout structure predicts unseen initializations (log-loss .1865 versus .1978) and has the best unseen-task point loss (.2039 versus a predeclared .2596 baseline, but not separated from a .2084 suite mean). Under a common observed-compatible data-generating process, nominal-90% measurement-replication coverage is .912 for the hierarchy, .898 for 1PL, and .931 for 2PL; the hierarchy has the best interval score but is not uniformly better calibrated. An equal-12,800-observation control reverses our earlier policy-count explanation: discrimination recovery is .764 with $N = 4 \times 8$ repeats and .731 with $N = 32 \times 1$. For acquisition, we implement outcome-enumerated joint weak-rank information gain and compare matched atomic task queries on independent future matrices. It beats uniform, but does not robustly beat task EIG or Fisher: 25% rank-difference intervals include zero on both pre-specified data-generating processes, no method has a median sustained-$\tau \geq .90$ budget within 50%, and the joint method triggers no model-based stops. Architecture-DIF tests also remain negative and partly unidentified. The contribution is validated small-cohort measurement with transparent negative boundaries, not generic adaptive or certified robot evaluation.

## 1 INTRODUCTION

Embodied evaluation is expensive: an episode consumes simulator and policy time, real-robot trials cost more, and every policy multiplies the bill. Consequently, vision-language-action (VLA) papers often report benchmark means for a small cohort, sometimes with one trial per task and initial state. Yet the questions readers ask are comparative: is policy A better than B, will that order survive a new task composition, and does an architecture respond differently to a type of item? Benchmark audits already document shortcuts, weak power, overfitting, and data-source dependence (Jiang et al., 2026b). We study the complementary measurement question: what can be inferred when the cohort itself contains only four to six policies?

On a response matrix built for this purpose, three failures are concrete. First, the realized top pair at $N = 6$ is an exact tie, 195/200 successes each, despite separated latent-ability point estimates. Second, a 2PL optimizer and Bayesian sampler converge while item discrimination remains prior-sensitive. Third, conditional model probabilities exclude task-composition uncertainty that materially changes policy comparisons. These facts motivate a measurement model, but not every initial diagnosis survives control: at fixed observation count, repeat depth compensates for respondent breadth in our discrimination-recovery simulation.

We therefore treat evaluation as measurement rather than score averaging. The instrument should represent verified rollout structure, predict genuinely withheld items and tasks, name the target behind each rank probability, and expose rather than hide a failed acquisition hypothesis. Our three contributions are:



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.

1. We assemble and audit a 3,155-episode LIBERO bank and a 14,412-episode LIBERO-PLUS spatial bank over six variants. Standard IRT diagnostics expose small-cohort pathologies, while an equal-observation control rejects the simpler claim that policy count alone is binding.

2. We validate a hierarchical measurement model whose family-conditional trial layer and pooling resolution follow measured rollout structure. It improves pre-registered unseen-initialization and unseen-task prediction, attains near-nominal same-DGP coverage, and makes rank probabilities estimand-specific.

3. We implement the corrected outcome-averaged joint weak-rank objective and evaluate it with matched task queries and independent future matrices. Its failure to robustly beat task EIG or Fisher is a negative boundary on adaptive-evaluation claims, not a hidden optimization result.

Architecture-dependent item functioning (DIF) is a design diagnosis rather than a fourth contribution: P1/P2 were prospectively specified, P3 was selected after the OpenVLA marginal collapse was observed, and all protected tests are negative.

**Problem formulation.** Let $v \in \{1, \ldots, N\}$ index policy variants, let $i = (s, t, j)$ denote an item—suite $s$, task $t$, and initialization $j$—and let $y_{vir} \in \{0, 1\}$ be trial $r$ for variant $v$ on item $i$. The observed bank $D = \{(v, i, r, y_{vir})\}$ is an evaluation instrument, not by itself the population to which a rank refers. We therefore pair $D$ with a declared target $Q$ and report posterior pairwise orders and best-policy probabilities. Three targets recur below: the fixed observed items with trial noise averaged away; a replication of the recorded task-and-initialization design with new outcomes; and a modeled super-population of items. They answer different questions and need not produce the same order. The acquisition target is the joint weak ordering of all six policies in an independent future task, initialization, and outcome replication. A four-policy legacy bank supplies item structure before two new policies reveal complete atomic task queries. Acquisition and evaluation matrices are independent.

We label model development, pre-specified scale replication, registered frozen-bank diagnostics, and the one-shot VLA-015 repair separately. The repair plan, equations, action spaces, data-generating processes (DGPs), gates, and seeds were committed before its result; all new analyses ran regardless of intermediate outcome.

## 2 AUDITED RESPONSE BANKS

### 2.1 COHORT AND REPEAT STRUCTURE

The core bank covers six variants from five checkpoint lineages and three broad families: OpenVLA (Kim et al., 2024); MolmoAct2 with continuous flow and discrete greedy-AR decoding (Fang et al., 2026); $\pi_{0.5}$ (Physical Intelligence et al., 2025); GR00T-N1.7 (NVIDIA et al., 2025); and VLA-JEPA (Sun et al., 2026). Items are task-by-initial-state pairs from LIBERO spatial, object, goal, and long-horizon suites (Liu et al., 2023), with task and initialization indices, seeds, revisions, rendering, maximum steps, and success predicates pinned. Tier A collected four variants under a pre-registered gate and cell plan; Tier B added $\pi_{0.5}$ and GR00T under the same conventions. All 3,155 rows are schema-valid with zero duplicate keys or integrity-error cells. The 36 warnings are documented partial database episodes superseded by two kill-resume cycles; none were silently dropped (Table 1).

A designed repeat slice exposes a structural split. Greedy-AR variants reproduce success and step count bitwise on 45/45 independently launched repeats, including five failure scenes. Flow decoders show step-count spread on 9–10 of 10 repeat items and success flips on up to 4/10. Thus an additional recorded repeat is informative for a flow decoder but not for a deterministic decoder under this serving stack. We encode this fact in both the likelihood and the acquisition candidate set.

Only 68/220 (30.9%) variant-by-task cells lie strictly inside the pre-stated $(.05, .95)$ mid band; 152 are at ceiling and none at floor. This population cannot identify many item or architecture interactions, regardless of repeat count.

Table 1: Response-bank provenance and audit. Cell and item definitions differ by bank because the first supports repeat-aware fitting and the second de-censors category contrasts.

| Bank | Variants | Episodes | Population | Mid band | Integrity |
|---|---|---|---|---|---|
| Tier-A/B LIBERO | 6 | 3,155 | 4 suites; 220 variant×task cells | 68/220 (30.9%) | 0 errors; 36 documented warnings |
| LIBERO-PLUS spatial | 6 | 14,412 | 2,402 items; 42 variant×category cells | 27/42 (64.3%) | 0 problems; 0 warnings |

### 2.2 A DE-CENSORING POPULATION

LIBERO-PLUS perturbs scenes along seven labeled axes, including camera, language, lighting, sensor noise, and object layout (Fei et al., 2026). A pilot validated the harness and a monotone difficulty signal, then a pre-registered sweep collected one official outcome for each of 2,402 spatial perturbations and six variants. The resulting 14,412-row bank has zero missing items, duplicates, or integrity warnings. Its mid-band census rises to 27/42 (64.3%) variant-by-category cells. On this population, $\pi_{0.5}$ exceeds VLA-JEPA, while OpenVLA drops from .78 on base spatial to .2494 [.2325, .2671]. These are population-specific robustness outcomes, not causal architecture effects.

## 3 WHAT SMALL-$N$ DIAGNOSTICS ESTABLISH—AND DO NOT

We fit means, Rasch/1PL, 2PL, logistic matrix factorization, and beta-binomial shrinkage to the $N = 4$ bank. The original cell-masked CV, which leaves other initializations from the same policy–task cell in training, is a known-task interpolation diagnostic rather than an unseen-item test. Under that split the cheap beta-binomial baseline is strongest among standard models (log-loss .23731 versus .2632 for 2PL and .2647 for 1PL). Rasch and 2PL are serious baselines, not straw men (Rasch, 1960; Birnbaum, 1968).

Clean convergence still does not imply item recovery. Maximum-likelihood 2PL pushes 98/400 real-data discriminations to its cap. Under matched Rasch truth, 24.6% hit the cap despite 100% optimizer convergence; Bayesian discrimination posteriors retain 90–99% of prior width. The earlier simulation also found recovery correlation .26 at $N = 4$ and .67 at $N = 32$, but it changed total observations together with cohort size. A registered equal-observation control fixes 12,800 Bernoulli observations and trades policies against repeats: correlation is .7642 for $N = 4 \times 8$, .7484 for $N = 8 \times 4$, .7381 for $N = 16 \times 2$, and .7314 for $N = 32 \times 1$. The paired $N = 32$ minus $N = 4$ difference is $-.03275$ $[-.03447, -.03105]$, opposite the frozen gate. Respondent breadth and repeat depth trade off in this simulation; policy count alone is not the binding explanation. The ML cap fraction nevertheless falls from .0503 to .0162, so the two diagnostics expose different aspects of instability.

Rank uncertainty also depends on target. Standard fixed-item models give $P(\text{VLA-JEPA} > \text{MolmoAct2-disc}) = .92$–$.96$, while a task-plus-initialization bootstrap preserves that order only .67 of the time (flip probability .28–.34; item-level McNemar $p = .75$). This is an estimand mismatch, not proof that every standard interval under-covers.

## 4 HIERARCHICAL MEASUREMENT BEYOND SEEN CELLS

For variant $v$, item $i$ in task $t$ and suite $s$, the model is

$$\eta_{vi} = a_s\,(\theta_v - b_t - d_i) + u_{vt} + \mathrm{DIF}_{vt}, \tag{1}$$

$$k_{vi} \sim \mathrm{Binomial}\big(n_{vi}^{\mathrm{eff}},\ \mathrm{logit}^{-1}(\eta_{vi})\big)\,. \tag{2}$$

Here $\theta$ is partially pooled ability; $b$ combines a suite anchor, instruction metadata, and a learned-scale task deviation; $d$ is an initialization-within-task deviation; and $u$ is a variant-by-task random effect. Discrimination $a$ is estimated at suite resolution. All pooling scales are learned with non-centered NUTS. The primary prediction model omits DIF; the contrast sensitivity is evaluated separately in Section 7.

The trial layer follows observed stochasticity. Independent flow rollouts enter the binomial count; verified deterministic greedy-AR repeats are deduplicated to $n^{\mathrm{eff}} = 1$. The random effect $u$ absorbs policy-specific task behavior that might otherwise be misread as an architecture effect.

Table 2: Blocked prediction and same-DGP interval validation. Log-loss is variant-macro; paired log-loss differences use 95% task-clustered bootstrap CIs. Coverage and interval score use the task+initialization+outcome replication estimand.

| Test | Quantity | HB | Standard comparator(s) | Paired result / boundary |
|---|---|---|---|---|
| Unseen initialization | log-loss | .18648 | task BB .19782 | $-.01134$ $[-.02157, -.00317]$ |
| Unseen task | log-loss | .20391 | suite BB .25959 | $-.05568$ $[-.10696, -.01445]$ |
| Unseen task sensitivity | log-loss | .20391 | suite mean .20843 | $-.00451$ $[-.01218, +.00223]$ |
| HB DGP | 90% coverage | .9122 | 1PL .8979; 2PL .9310 | no uniform calibration win |
| HB DGP | interval score | .1524 | 1PL .1614; 2PL .1603 | HB lower is better |
| Rasch DGP | 90% coverage | .9096 | 1PL .9180; 2PL .9255 | all near nominal |

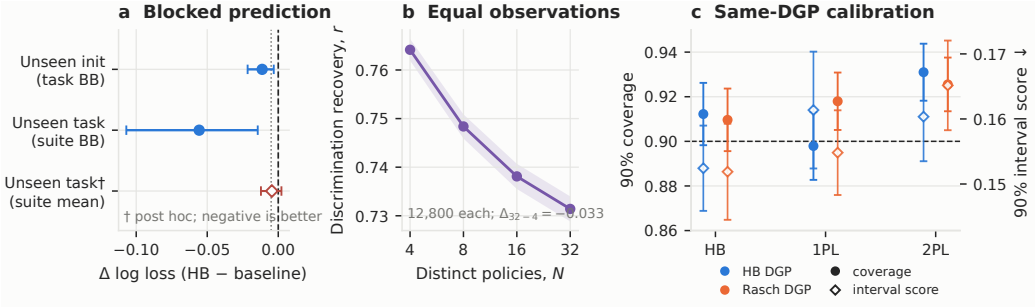



Figure 1: Measurement validation and corrected diagnosis. Blocked prediction tests genuinely unseen initializations and tasks; common-DGP intervals separate coverage from proper score; and the equal-observation control rejects a policy-count-only explanation of discrimination recovery.

VLA-015 replaces the permissive cell split with two five-fold blocked tests on all 2,200 base cells. In the item-blocked split, the same task–initialization item is held out across every covering policy; in the task-blocked split, eight complete tasks per fold are absent for all policies. The no-DIF hierarchy beats the predeclared local baseline on both (Table 2): unseen-initialization log-loss is .18648 versus .19782, and unseen-task log-loss is .20391 versus .25959. Paired suite-stratified task-bootstrap intervals exclude zero. On unseen tasks HB also has the best point estimate, but its post-hoc difference from the stronger variant–suite mean (.20843) is only $-.00451$ $[-.01218, +.00223]$; it does not significantly beat the strongest simple predictor.

The interval comparison likewise uses common DGPs and common pair-score estimands. On 256 observed-compatible HB matrices, nominal-90% measurement-replication coverage is .9122 for HB, .8979 for 1PL-Wald, and .9310 for 2PL-Wald. HB has the best proper interval score (.1524 versus .1614 and .1603), but 1PL is closest to nominal coverage. Under Rasch truth, coverage is .9096, .9180, and .9255. Thus HB is self-consistent and competitive across both DGPs, not uniformly better calibrated than standard methods. Prior-predictive SBC remains a model-conditional implementation check.

## 5 Estimand-aware rank uncertainty

A rank probability is incomplete without its target. *Fixed-item conditional ability* asks who wins on exactly the observed items with infinite trials. *Measurement replication* asks who wins when task composition and outcomes are redrawn according to the evaluation design. A *super-population* estimand asks about the modeled item population. At $N = 4$, $P(\text{VLA-JEPA} > \text{MolmoAct2-disc})$ is .978, .781–.824, and .843, respectively. The measurement-replication posterior contains the observed cluster-bootstrap flip rate in its predictive interval, reconciling rather than dismissing the design-based calculation.

Formally, let $Z_v^{(Q)}$ be the score induced for variant $v$ by target $Q$. We report

$$\pi_{vw}^{(Q)} = P\left(Z_v^{(Q)} > Z_w^{(Q)} \mid D, Q\right), \tag{3}$$

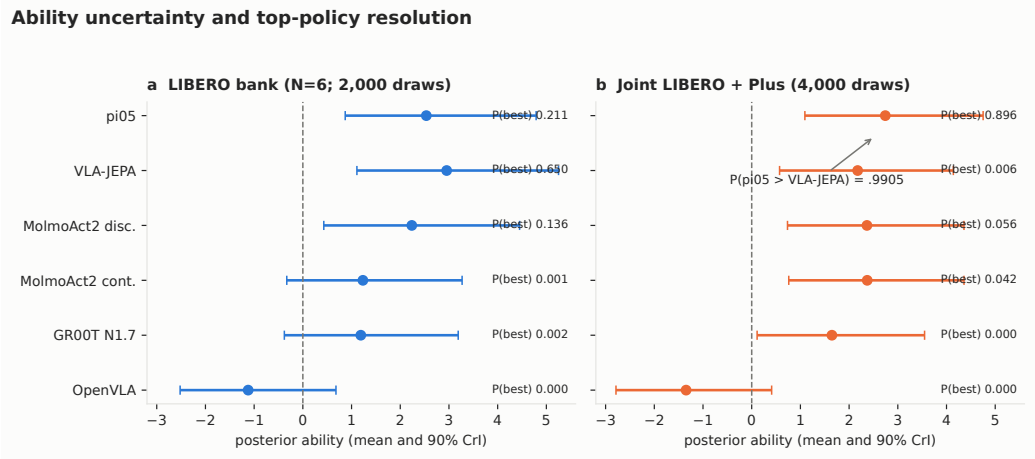



Figure 2: Ability uncertainty and top-policy resolution. Central-90% intervals are shown for LIBERO and the joint fit. The targeted top-pair probability becomes .9905 only after adding the de-censoring population.

with equality retained as a distinct event. Changing $Q$ changes the random quantities integrated in Equation 3: fixed-item draws retain parameter uncertainty, measurement-replication draws additionally resample the evaluation design and outcomes, and super-population draws generate new item effects from the fitted hierarchy. Thus a confident conditional order is not evidence that the order will survive a new task composition.

At $N = 6$, latent means place VLA-JEPA above $\pi_{0.5}$, but both achieve 195/200 on shared LIBERO items. Their measurement-replication best probabilities are only .53–.56 and .33–.34, respectively, rather than evidence for a strict order. This is a model-based uncertainty statement, not an anytime-valid decision guarantee.

The joint LIBERO plus LIBERO-PLUS fit changes the measured population and resolves the comparison: $P(\pi_{0.5} > \text{VLA-JEPA}) = .9905$ and $P(\pi_{0.5} \text{ best}) = .8962$. This does not prove a universally better policy. It shows that explicitly changing the target population can change the comparison; it does not transfer the order beyond the specified simulator population.

## 6 ADAPTIVE EVALUATION: A FALSIFIED HEADLINE

The VLA-014 replay completed one recorded matrix and used unequal action spaces; its procedure-seed intervals were conditional Monte Carlo uncertainty on that bank. We replace it with the target readers actually need. Let $O_{\text{rep}}$ be the joint weak ordering of six policies in an independent future suite-stratified task, initialization, and outcome replication. It is represented by all 15 pair relations in $\{-1, 0, +1\}$. For a complete atomic task query $q$ with outcome vector $\mathbf{y}$, the acquisition score is

$$\Delta_H(q; D) = H(O_{\text{rep}} \mid D) - \sum_{\mathbf{y}} p(\mathbf{y} \mid q, D) H(O_{\text{rep}} \mid D, q, \mathbf{y}). \tag{4}$$

All at most 64 outcomes of a two-policy query are enumerated. Each branch performs an outcome-specific sequential Gaussian/Laplace update of posterior mean and covariance, then common random numbers generate the future replication. Joint entropy is computed from the empirical weak-order distribution; a sum of marginal pair entropies is a separate ablation.

For tractability, every eligible task is first scored by matched-query Fisher information; exact Equation 4 is evaluated on the top four using eight structure and 32 replication draws. This pre-result computational amendment is part of the method name: *Fisher-screened joint-rank EIG*.

Table 3: Fresh-matrix acquisition with matched atomic task queries. Values are means over 64 independent acquisition/evaluation matrix pairs per DGP. AUC averages Kendall $\tau$ over 5–50% of the 800-rollout budget at complete-query checkpoints; stops use a model-based .95 threshold.

| DGP | Acquisition | $\tau$ at $\approx 25\%$ | $\tau$ AUC | model-based stops |
|---|---|---|---|---|
| HB | Joint-rank EIG | .7520 | .7472 | 0/64 |
| HB | Pair-marginal EIG | .7458 | .7477 | 0/64 |
| HB | Task EIG | .7414 | .7336 | 0/64 |
| HB | Fisher | .7563 | .7505 | 0/64 |
| HB | Uniform | .6993 | .7010 | 0/64 |
| Task-Beta | Joint-rank EIG | .6161 | .6044 | 0/64 |
| Task-Beta | Pair-marginal EIG | .6031 | .6017 | 0/64 |
| Task-Beta | Task EIG | .5888 | .5940 | 1/64, wrong |
| Task-Beta | Fisher | .6115 | .6068 | 0/64 |
| Task-Beta | Uniform | .5088 | .5250 | 0/64 |

**Algorithm 1: Outcome-enumerated joint weak-rank acquisition**
**Require:** legacy item bank, active policies, future-replication design, complete task queries, common random draws, model-based threshold .95.
At each query: (1) condition the active-policy posterior on acquired outcomes; (2) Fisher-screen every eligible complete task query and retain four; (3) for every retained query, enumerate its outcome vectors, update the fantasy posterior mean and covariance, simulate future measurement replications, and average the resulting joint weak-rank entropy; (4) acquire the query with largest expected reduction and refit; (5) record a model-based stop only if one weak rank has posterior probability at least .95 for two consecutive queries. Return ranks, uncertainty, and cost. The stop is not a statistical certificate.

**Registered fresh-matrix protocol.** The four-policy legacy bank predates $\pi_{0.5}$ and GR00T; neither new policy contributes outcomes at initialization, and incumbent uncertainty is propagated rather than replaced by full-bank means. Every method selects the same atomic task query—up to three initializations for each active policy, with the final remainder kept complete. Acquisition and evaluation matrices are independent. We run 64 paired matrices under each of two frozen DGPs: posterior-predictive HB truth and an empirical per-policy–task Jeffreys-Beta truth. Methods are joint-rank EIG, its pair-marginal ablation, the task-EIG transplant of Anwar et al. (2025), matched-task Fisher, and uniform task selection. The main readout is approximately 200 of 800 rollouts; curves cover complete-query checkpoints through at least 50% of the 800-rollout budget.

The adaptive headline fails its frozen gates. Under HB truth, joint minus task-EIG $\tau$ is $+.01064$ $[-.01638, +.03787]$ at the readout; under task-Beta truth it is $+.02723$ $[-.00654, +.06301]$. The 5–50% AUC difference is positive under HB ($+.01368$ $[+.00090, +.02726]$) but not task-Beta ($+.01040$ $[-.00533, +.02514]$). Joint EIG also does not separate from Fisher. Every method's median sustained-$\tau \geq .90$ budget lies beyond the observed 50% rollout-budget window, and joint EIG fires no model-based stop.

There is nevertheless a narrower positive result: joint EIG beats uniform at the readout under both HB ($+.05274$ $[+.01345, +.09291]$) and task-Beta ($+.10727$ $[+.05906, +.15592]$) truth, with positive AUC differences as well. This supports adaptive selection over random tasks, not a new superiority claim over targeted EIG or Fisher.

## 7 DIF AS A DESIGN DIAGNOSIS

DIF is an ability-adjusted architecture-by-item interaction, not a raw family gap. With only six variants, the model uses five pre-stated contrasts, including a clean same-checkpoint decoding contrast

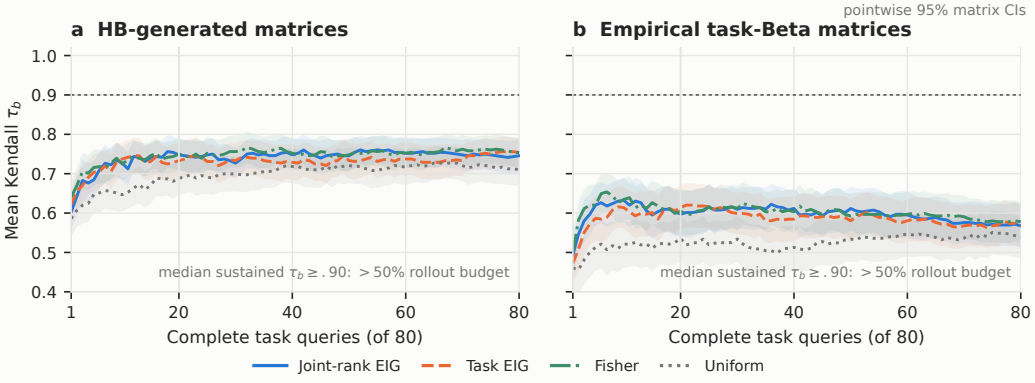



Figure 3: Fresh-matrix rank recovery under matched task queries. Curves report mean Kendall $\tau$ across independent acquisition/evaluation matrices for both frozen DGPs. Joint-rank EIG improves over uniform but does not robustly separate from task EIG or Fisher; no method has median sustained $\tau \geq .90$ within 50%.

Table 4: DIF coordinates. P1/P2 were prospectively specified; P3 is data-motivated exploratory. All intervals include zero under the stated decision rule.

| Coordinate | Mean [90% credible interval] | Verdict |
|---|---|---|
| P1: world-model contrast × camera | $+.026$ $[-.762, +.809]$ | not detected |
| P2: AR-decoding contrast × language | $+.183$ $[-.609, +.950]$ | not detected |
| P3: OpenVLA-excess contrast × camera | $-.916$ $[-1.995, +.084]$ | exploratory; not detected |

and four explicitly confounded lineage/family contrasts. The bundles include backbone, checkpoint scheme, and, for one lineage, a 50-versus-10 control-settling convention.

On base LIBERO, all adjusted posteriors include zero. Injected $-1$-logit effects are detected 0% of the time; even $-2$ logits yield only 0–16.7% detection across contrast axes. This is the first negative mechanism: 152/220 cells are at ceiling. The DIF hypothesis was removed from the headline and LIBERO-PLUS was pre-registered as a de-censoring instrument.

The adjusted joint refit passes fit and predictive checks, but none of its three reported coordinates is detected (Table 4). P1 and P2 were protected before the adjusted fit; P3 was chosen after the OpenVLA marginal collapse had been observed and is data-motivated exploratory. The prospective rule requires the 90% credible interval to exclude zero. $P(\text{P3} < 0) = .935$ is only an exploratory lean.

De-censoring exposes a second, deeper problem. In eight simulations, injected $-1$ at P1 and P3 and $-2$ at P1 are each detected 0/8 times; null calibration yields 0/240 coordinate detections. The injection demonstrably changes outcomes, but the fitted coordinate returns near zero while its credit is partly absorbed by variant-by-task effects and otherwise by shrinkage. Full versus no-DIF held-out log-loss is $+.00061$ $[-.00013, +.00135]$, so DIF adds no predictive value. At a single suite, a category interaction and a repeated pattern of task random effects are structurally unidentified. These nulls therefore do not establish absence.

Seven of 45 exploratory coordinates exclude zero under the same rule, including an OpenVLA sensor-noise deficit and relative object-layout robustness. They are **exploratory leads**, not findings: specificity under the fitted null does not undo the measured sensitivity failure. The identifying follow-up is cross-suite replication of the same perturbation category, which a task-within-suite random effect cannot mimic. The held all-suite LIBERO-PLUS collection would answer that question, but it is not required for this paper's negative conclusion and has not been run.

## 8 Related work and claim boundary

Robot benchmark audits and high-throughput matrices establish reliability problems and collection infrastructure (Jiang et al., 2026b; Selvaraj et al., 2026); PhAIL adds distributional time-to-success intervals and pairwise tests for real VLAs (Arkhangelskiy, 2026). Our narrow target is item-level measurement with $N = 4$–$6$. Standard IRT and recent AI work motivate both the baseline and its small-cohort failure mode (Rasch, 1960; Birnbaum, 1968; Jiang et al., 2026a; Rivera et al., 2026). ATLAS and CAT theory study adaptive item selection (Li et al., 2026; Lee & Li, 2026). Balkır et al. already target uncertainty-aware ranking and adaptive stopping for continuous LLM scores (Balkır et al., 2026); Arviv et al. use sequential tests and confidence-width/effect-size stops for model evaluation (Arviv et al., 2026). Formal cost-efficient and anytime-valid comparison also predates this work (Wang et al., 2026a; Snyder et al., 2026). We claim neither adaptive-ranking nor stopping priority, and our empirical threshold is not a certificate.

The closest robot active-evaluation methods optimize information over real-world task factors (Liao et al., 2026) or cost-aware task-by-policy experiments (Anwar et al., 2025); ROBOGATE adaptively searches VLA failure boundaries (Kim, 2026). These establish generic active robot evaluation. RoboArena aggregates distributed preferences (Atreya et al., 2025); vla-eval and XPolicyLab standardize serving and cross-platform evaluation (Choi et al., 2026; XPolicyLab Community et al., 2026). Sim/real studies directly warn that simulator rankings need not transfer (Wang et al., 2026b). Our remaining positive claim is the validated small-cohort stochastic measurement model. The corrected joint rank objective is reported as a negative experiment, not an acquisition novelty claim.

## 9 Limitations and release status

The study is simulator-only and covers six policy variants from five checkpoint lineages on one benchmark family: four LIBERO suites and LIBERO-Plus spatial. Ability and robustness claims apply only to those checkpoints and protocols. Architecture contrasts retain named lineage, backbone, checkpoint, wait-step, and seed confounds. The single-suite design is structurally unidentified for family-by-category DIF, and the exploratory interactions await cross-suite replication. Sim/real rank validity is untested.

Blocked prediction stays within the same suites. On unseen tasks HB beats the predeclared beta-binomial baseline, but not the strongest simple suite-mean predictor with statistical separation. Predictive margins may depend on the ten-initialization design. Greedy-AR determinism must be retested after any serving or seeding change.

Fresh acquisition uses synthetic matrices from two observed-grounded DGPs, not new simulator or robot rollouts. Sixty-four matrices per DGP limit precision, and the Fisher shortlist plus $8\times32$ Monte Carlo panel approximates exhaustive joint-rank EIG. The task-EIG baseline is a paper-faithful transplant, not a line-by-line reproduction: unavailable author code and a shared hierarchical surrogate prevent an end-to-end surrogate comparison. The model-based stop has no optional-stopping or frequentist error guarantee. Same-DGP coverage spans two fitted DGPs, not arbitrary misspecification. The equal-observation control is one simulation design and supports a trade-off, not a universal benefit of repeats.

The reader artifact is complete locally, but public release is **withheld**. The LIBERO-Plus code fork has no license file, one evaluated checkpoint has no license tag, and the project has no approved release license. Model weights, gated files, third-party source, and raw media are excluded from any prospective reader artifact.

## Reproducibility statement

A private reproducibility record retains the frozen configurations, item-level outcomes, paired tests and uncertainty summaries, model and benchmark revisions, exact commands, environment records, experiment cards, and SHA-256 manifests used for every reported headline. The VLA-015 plan was frozen before results; literal tests cover outcome enumeration, fantasy mean/covariance updates, joint versus marginal entropy, target replication, and atomic queries. Raw blocked predictions, same-DGP interval rows, equal-observation replicates, and fresh-acquisition curves are re-

tained with final manifests. The repair used zero GPU rollouts. Its authoritative v3 result stage took 700.6 CPU-seconds and reused hash-pinned validation outputs completed before a recorded summary-compatibility failure. Appendix F maps each claim class to its record. This is a paper-only submission: no supplemental package or public artifact is included.

## STATEMENT ON THE USE OF GENERATIVE AI

Generative AI tools, including OpenAI Codex, assisted under human direction with methodology and experiment feedback, implementation and testing, data cleaning and formatting, experiment orchestration and documentation, analysis and interpretation, literature search and summarization, figure and table generation, and drafting and editing. The author remains responsible for the research questions, locked protocols, interpretation, final verification, and any submission decision. AI-assisted work is checked against frozen configurations, item-level outputs, paired tests with uncertainty, independent headline reproduction, and SHA-256 manifests. The author takes full responsibility for the paper's contents.

## REFERENCES

Abrar Anwar, Rohan Gupta, Zain Merchant, Sayan Ghosh, Willie Neiswanger, and Jesse Thomason. Efficient evaluation of multi-task robot policies with active experiment selection. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 1636–1653, 2025. URL https://proceedings.mlr.press/v305/anwar25a.html.

Sergey Arkhangelskiy. PhAIL: A real-robot VLA benchmark and distributional methodology. *arXiv preprint arXiv:2605.29710*, 2026. URL https://arxiv.org/abs/2605.29710.

Ofir Arviv, Kristjan Greenewald, Yotam Perlitz, Hadar Mulian, Michal Shmueli-Scheuer, and Leshem Choshen. Stop guessing when to stop testing: Efficient model evaluation with just enough data. *arXiv preprint arXiv:2607.08522*, 2026. URL https://arxiv.org/abs/2607.08522.

Pranav Atreya et al. Roboarena: Distributed real-world evaluation of generalist robot policies. In *Proceedings of the 9th Conference on Robot Learning*, 2025. URL https://arxiv.org/abs/2506.18123.

Esma Balkır, Alice Pernthaller, Marco Basaldella, José Hernández-Orallo, and Nigel Collier. Confident rankings with fewer items: Adaptive LLM evaluation with continuous scores. *arXiv preprint arXiv:2601.13885*, 2026. URL https://arxiv.org/abs/2601.13885.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

Suhwan Choi, Yunsung Lee, Yubeen Park, Chris Dongjoo Kim, Ranjay Krishna, Dieter Fox, and Youngjae Yu. vla-eval: A unified evaluation harness for vision-language-action models. *arXiv preprint arXiv:2603.13966*, 2026. URL https://arxiv.org/abs/2603.13966.

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

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2510.13626.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for ai evaluation? *arXiv preprint arXiv:2607.15190*, 2026a. URL https://arxiv.org/abs/2607.15190.

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

Azuki Kim. ROBOGATE: Adaptive failure discovery for safe robot policy deployment via two-stage boundary-focused sampling. *arXiv preprint arXiv:2603.22126*, 2026. URL https://arxiv.org/abs/2603.22126.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Seungwon Lee and Xiaoou Li. On the asymptotics of item selection in multidimensional computerized adaptive testing. *arXiv preprint arXiv:2608.00201*, 2026. URL https://arxiv.org/abs/2608.00201.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V. Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.04689. Spotlight.

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

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. URL https://arxiv.org/abs/2306.03310.

NVIDIA, Johan Bjorck, Fernando Castaneda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

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

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, 1960.

Joshua Fonseca Rivera, Neil Shah, David Demitri Africa, and Konstantinos Voudouris. Item response theory for ai safety. *arXiv preprint arXiv:2608.05086*, 2026. URL https://arxiv.org/abs/2608.05086.

Praveen Selvaraj, Lorenzo Uttini, and Ville Kuosmanen. Armnetbench v0.1: Parallel real-world evaluation of manipulation policies on a low-cost arm farm. *arXiv preprint arXiv:2607.24481*, 2026. URL https://arxiv.org/abs/2607.24481.

David Snyder, Apurva Badithela, Nikolai Matni, George Pappas, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Beyond binary success: Sample-efficient and statistically rigorous robot policy comparison. *arXiv preprint arXiv:2603.13616*, 2026. URL https://arxiv.org/abs/2603.13616.

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

Ganghua Wang, Zhaorun Chen, Bo Li, and Haifeng Xu. Cer-eval: Certifiable and cost-efficient evaluation framework for llms. *OpenReview, ICLR 2026 submission*, 2026a. URL https://openreview.net/forum?id=eCx0fOWiSA.

Shuo Wang, Hanyuan Xu, Yingdong Hu, Fanqi Lin, and Yang Gao. A practical recipe towards improving sim-and-real correlation for VLA evaluation. *arXiv preprint arXiv:2606.10366*, 2026b. URL https://arxiv.org/abs/2606.10366.

XPolicyLab Community, Tianxing Chen, et al. XPolicyLab: A unified standard and open ecosystem for robot policy evaluation and deployment. *arXiv preprint arXiv:2608.09892*, 2026. URL https://arxiv.org/abs/2608.09892.

## A Response-bank provenance and integrity

The core bank is organized by policy variant, suite, task, initialization, and trial. Every row retains model and benchmark revisions, task and initialization indices, environment and policy seeds, rendering and horizon settings, success predicate, elapsed time, and collection stage. The immutable audit checks schema validity, duplicate composite keys, missing cells, superseded partial episodes, and consistency between database and exported outcome records. The 36 warnings in Table 1 refer only to partial database episodes superseded by two documented kill-resume cycles. They are retained in the audit trail and excluded by the same pre-stated completeness rule for all variants.

The repeat study launched each trial independently rather than replaying a cached response. For greedy-AR policies, all 45 paired repeats match in success and step count bitwise, including five failure scenes. For flow policies, nine or ten of ten repeat items show nonzero step-count spread and as many as four of ten flip success. The paper therefore uses a structural deterministic flag only for the verified serving configuration; it is not inferred from checkpoint family labels alone.

The LIBERO-Plus bank is a complete Cartesian sweep over six variants and 2,402 spatial perturbations. The manifest checks 14,412 unique keys, zero missing combinations, and zero integrity warnings. Perturbation category is part of item provenance, not a post-hoc label generated from outcomes.

## B Model specification and validation

Task difficulty in Equation 2 combines a suite anchor, a regression on instruction length, spatial-term count, two-part structure, and articulated-object status, and a learned-scale task deviation. Initialization deviations are nested within task and shared across variants. Ability, task, initialization, and variant-by-task blocks use non-centered parameterizations; discrimination is positive and partially pooled at suite resolution. The family-conditional trial layer replaces $n_{vi}$ by one for verified deterministic policies and preserves the number of independent trials for flow policies. DIF uses a fixed contrast matrix and a shared hierarchical scale; Appendix E describes its confounds.

All posterior fits report divergences, $\hat{R}$, and effective sample size. Prior-predictive SBC at $N = 4$ and $N = 6$ produced near-nominal coverage and rank-uniformity, but this is a model-conditional implementation check, not a comparison to standard estimators. Posterior-predictive checks cover task dispersion, deterministic repeats, held-out means, suite means, and rank-flip statistics. The joint fit over 110 tasks, 2,802 items, and 16,612 cells has zero divergences, $\hat{R} \leq 1.0056$, passes effective-sample-size gates, and flags 0/42 variant-by-category posterior-predictive cells.

The original cell-masked CV stratified individual cells by variant and suite; it left other initializations from the same task in training and is reported only as known-task interpolation. VLA-015 uses common blocks across policies. The item split holds out 80 task–initialization items (440 cells) per fold while retaining their tasks; the task split holds out eight complete tasks (440 cells) per fold.

Both have five folds. The no-DIF HB prediction gates pass against the predeclared split-specific beta-binomial baselines. The unseen-task comparison to the post-hoc strongest suite mean is $-.00451$ $[-.01218, +.00223]$, so point performance is best without a significant simple-baseline advantage.

Same-DGP coverage draws eight fixed truths and 32 matrices per truth (256 matrices per DGP). HB, 1PL-Wald, and 2PL-Wald estimate the same finite-expected-score and measurement-replication pair differences for all 15 policy pairs. Intervals are central 90%; paired interval-score comparisons bootstrap whole synthetic matrices. On HB truth, HB interval score improves on 1PL by $-.00894$ $[-.01428, -.00408]$ and on 2PL by $-.00792$ $[-.01201, -.00393]$, while its absolute coverage error does not improve on 1PL. Under Rasch truth HB remains near nominal.

The equal-observation control uses 200 paired replicates, 400 items, ability support $[-2, 2]$, and 32 trials per item. Cohorts $(N, \text{repeats})$ are $(4, 8)$, $(8, 4)$, $(16, 2)$, and $(32, 1)$, each totaling 12,800 observations. Common item truths and paired random uniforms isolate the breadth–depth trade-off.

## C Estimands and rank uncertainty

For fixed observed items, posterior draws average away trial noise while retaining parameter uncertainty. Measurement-replication draws resample tasks (and, in the most design-faithful version, initializations) and generate outcomes through the fitted trial layer. Super-population draws sample modeled task and initialization effects. Each reported pair probability names which of these operations it uses.

At $N = 4$, the contested VLA-JEPA/MolmoAct2-discrete comparison is .978 under the fixed-item estimand, .781–.824 under measurement replication, and .843 under the super-population estimand. The realized cluster-bootstrap flip probability is .319 and lies in the model's predictive interval. At $N = 6$, VLA-JEPA and $\pi_{0.5}$ each succeed on 195/200 shared base items; measurement-level best probabilities are .53–.56 and .33–.34, respectively, rather than a resolved strict order. Under the joint population, $P(\pi_{0.5} > \text{VLA-JEPA}) = .9905$ and $P(\pi_{0.5} \text{ best}) = .8962$.

## D Fresh-matrix acquisition protocol

The VLA-014 procedure is retained only as a historical fixed-bank diagnostic: each policy's so-called held-out bank was fit on the other policies over the same item universe, its score summed marginal pair entropies after outcome-independent covariance shrinkage, and its 250-seed intervals measured procedure randomness on one matrix. None of its adaptive numbers support the corrected claim.

VLA-015 freezes a future-replication estimand over the 20 shared goal/long-horizon tasks. A legacy posterior contains four incumbents; $\pi_{0.5}$ and GR00T are active and initially contribute no outcomes. A task query purchases up to three unseen initializations for each active policy, using chunks $(3, 3, 3, 1)$ and never truncating a query at a reported budget. All methods see the same eligible queries and evaluation matrix. Incumbent measurement-replication uncertainty remains in every rank draw.

For joint/pair EIG, the scorer enumerates all $2^{m_q} \leq 64$ query outcomes in fixed bit order. A sequential Laplace fantasy update changes both posterior mean and covariance for each branch; normalized predictive branch probabilities weight the entropy. Common structure/replication panels eliminate candidate-level Monte Carlo noise. The primary joint entropy counts empirical frequencies of the complete 15-component weak-order vector; the marginal ablation sums ternary entropies. A literal unit test checks the paper equation against code. The Fisher screen scores all tasks and retains four before exact rank scoring.

Each DGP supplies 64 paired acquisition and independent evaluation matrices. HB truth draws from the base-only hierarchical posterior. Empirical task-Beta truth draws each policy–task rate from $\text{Beta}(k + .5, n - k + .5)$ and then draws fresh initializations. The task-EIG comparator follows the archival Bernoulli BALD rule with ten predictive samples, $\epsilon = .1$, and unit task-switch cost, but uses the common surrogate because the paper-linked code is unavailable. Curves retain every complete query through at least 50% of the 800-rollout budget. The primary readout is the last complete

query at or below 200 actual rollouts; AUC integrates 5–50% of rollout budget at complete-query checkpoints. Paired intervals bootstrap fresh matrices.

The frozen headline requires joint EIG to beat task EIG at the readout and over AUC, with a 25% median sustained-budget reduction, on both DGPs. It fails. Against Fisher, readout differences are $-.00424$ $[-.02924, +.02076]$ on HB and $+.00461$ $[-.01436, +.02372]$ on task-Beta. Against uniform, joint EIG improves AUC by $+.04623$ $[+.02003, +.07252]$ and $+.07937$ $[+.04803, +.11165]$, respectively.

The model-based stop requires maximum posterior weak-rank probability at least .95 at two consecutive queries. Joint EIG stops 0/64 times on each DGP; the one task-EIG stop under task-Beta is wrong. For 0/64 events, the one-sided 95% Clopper–Pearson upper bound is .0457. These are empirical summaries, not optional-stopping guarantees.

## E DIF CONTRASTS, POWER, AND IDENTIFICATION

The five contrast axes are: (1) discrete minus continuous MolmoAct2 decoding, the clean same-checkpoint comparison; (2) continuous MolmoAct2 minus $\pi_{0.5}$, bundled with lineage and a 50-versus-10 wait convention; (3) GR00T minus $\pi_{0.5}$, bundled with lineage and checkpoint scheme; (4) VLA-JEPA minus $\pi_{0.5}$, a world-model family contrast; and (5) OpenVLA excess beyond the clean AR-versus-flow contrast, bundled with backbone and collection conventions. These are predictive contrasts, not causal architecture interventions.

On base LIBERO, all adjusted posteriors straddle zero. At $N = 4$, injected $-1$ and $-2$ AR-decoder effects are detected 0% of the time. At $N = 6$, detection for a $-2$ effect is 8.3%, 16.7%, and 0% on the AR, GR00T, and VLA-JEPA axes. The adjusted de-censoring refit prospectively protects P1 world model by camera and P2 AR decoding by language. P3 OpenVLA excess by camera was selected after its marginal collapse was observed and is data-motivated exploratory; Table 4 gives all null verdicts.

The Plus validation suite simulates eight null datasets and eight datasets for each reported injection. Null calibration yields zero detections across 240 coordinate fits. Injecting $-1$ at P1 moves affected-cell model probability from .6105 to .4721 and realized rate from .610 to .473, yet the coordinate refits to mean $-.044$ and is never detected. A $-2$ P1 injection refits to $-.181$ and is detected 0/8 times; a $-1$ P3 injection refits to $-.018$ and is detected 0/8 times. Diagnostic fits show partial absorption by variant-by-task effects and further shrinkage not uniquely assignable to a reported block. Cross-suite replication of the same perturbation category is the identifying follow-up because task-within-suite effects cannot mimic a repeated suite-level pattern.

## F REPRODUCIBILITY AND ARTIFACT BOUNDARY

The private reader package is organized into five immutable layers:

1. collection manifests and item-level response records for both banks;
2. frozen model, CV, rank-estimand, replay, and DIF configurations;
3. raw posterior summaries, replay tensors, paired losses, and item-level audit outputs;
4. deterministic scripts that regenerate the three figures and four tables and independently reproduce blocked-prediction, coverage, cohort-control, and fresh-acquisition values; and
5. experiment cards, hardware and elapsed-time records, environment versions, exact commands, and input/output SHA-256 manifests.

The submission build itself is one-shot and CPU-only. It pins the unmodified official ICLR 2027 style archive, TeX engine, Python PDF-audit environment, bibliography, figure hashes, source commit, and build command before producing the PDF. Automated checks cover compilation, the main-text page boundary, anonymity, author and local path leakage, placeholders, unresolved citations, frozen figure hashes, required headline values, and source/PDF manifests.