FROM ASSOCIATION TO DECISION VALUE: TESTING PAIRED INSTRUCTION DEPENDENCE FOR VLA CORPUS SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing among robot demonstration corpora is a ranking problem: an upstream diagnostic must preserve which candidates produce better downstream policies, not merely correlate with average performance. We study *paired instruction dependence*, the rate at which a demonstrator changes its first decision when only the goal changes. Across 136 independently seeded quadruped-navigation corpora, predictions developed on 56 corpora are evaluated prospectively on 80 held-out corpora. The audit is reliable and associated with downstream paired correctness, and lowers mean absolute prediction error when added to a correctness-aware baseline, but the primary proper-log-score increment is $+0.059$ nats (95% corpus-bootstrap CI $[-0.027, 0.133]$) and the prespecified top-20% comparison does not establish an incremental selection gain. An exploratory response decomposition suggests a structural source of this predictive–ranking mismatch: under the tested predictor specification, separating both-correct and both-wrong responses yields larger held-out gains than their unsigned switch statistic. Seed-to-seed agreement indicates that the held-out endpoint retains recoverable rank signal. These results show that counterfactual response structure can be reliable and predictive while providing limited incremental decision value in this setting when compressed into an unsigned statistic, motivating diagnostics that preserve outcome-relevant structure and are validated against the decisions they are intended to support.

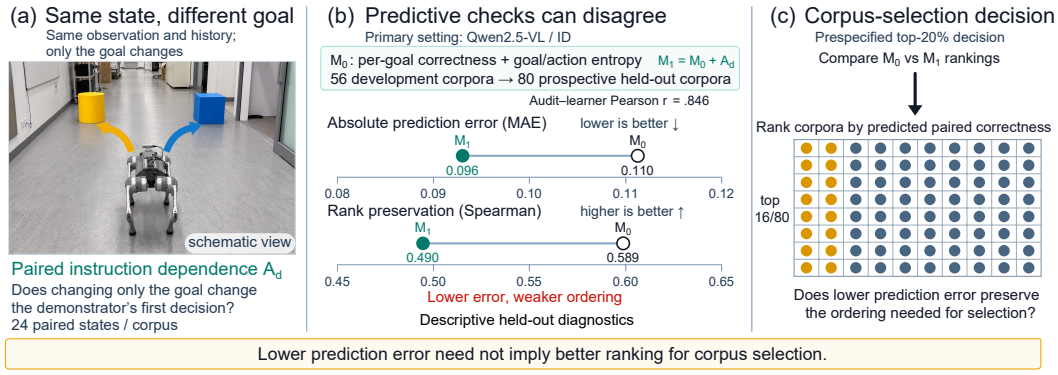

Figure 1: **Why prediction is not yet selection;** (a) A schematic third-person view illustrates the same-observation goal swap used to test whether the demonstrator changes its first response. (b) In the primary held-out test, adding A_d lowers MAE $.110 \rightarrow .096$ while Spearman falls $.589 \rightarrow .490$. (c) The prespecified top-20% rule ranks 80 corpora and selects 16, so ordering determines utility.

1 INTRODUCTION

The downstream value of a robot demonstration corpus is usually known only after a vision–language–action (VLA) policy has been fine-tuned on it. Evaluating every candidate therefore pays much

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.

of the cost that an upstream diagnostic is meant to avoid. The practical question is consequently decision-focused: can a measurement rank candidate corpora well enough to select the ones that will produce better policies?

Language-conditioned behavior creates a tempting diagnostic. If two goals require opposite turns from the same observation, a demonstrator that follows language should usually change its decision when the goal changes. We call the resulting switch rate *paired instruction dependence*, A_d. Unlike ordinary correctness, it uses a controlled same-state intervention and tests whether the response depends on the instruction. This distinction is relevant to generalist VLAs, whose success can otherwise conceal shortcut policies that weakly use language (Geirhos et al., 2020; Brohan et al., 2023; Kim et al., 2025; Ghosh et al., 2024; Black et al., 2025).

Dependence, however, is unsigned. A demonstrator can switch in the correct direction for both goals or in the wrong direction. With binary left/right responses, A_d is exactly the sum of these both-correct and both-wrong cases; the observed HALT action makes that identity approximate rather than exact in our data. The relevant question is therefore narrower than whether language sensitivity matters in general: does unsigned same-state response structure improve prediction and selection after marginal demonstrator correctness is known?

This question also separates four criteria that are often conflated:

$$\text{reliable measurement} \longrightarrow \text{association} \longrightarrow \text{incremental prediction} \longrightarrow \text{decision value}.$$

A signal may track broad outcome differences or lower mean error while failing to preserve the ordering near a selection cutoff. Figure 1 illustrates this *decision-value gap*.

We test it prospectively on 136 independently seeded quadruped-navigation corpora. Nested probabilistic predictors are developed on 56 corpora, while 80 disjoint corpora form the held-out test population. The correctness-aware baseline M_0 uses per-goal demonstrator correctness and corpus entropies; M_1 adds only A_d. We train 544 adapters across Qwen2.5-VL-3B and Qwen3.5-VL-0.8B and evaluate in-distribution and under a controlled scene shift. The primary selection endpoint is paired first-decision correctness; closed-loop evaluation separately tests how this local endpoint relates to full navigation. Predictors, held-out corpus identifiers, and the top-20% operating point were prespecified before held-out outcomes were revealed.

The study makes three main contributions:

1. **A prospective decision-value evaluation:** we separate reliability, association, proper probabilistic prediction, and top-k selection for a same-state robot-data diagnostic, using the corpus as the unit of inference.
2. **Correctness-resolved response structure locates a candidate source of the predictive–ranking mismatch:** the unsigned switch statistic mixes both-correct and both-wrong behavior; separating them yields larger exploratory held-out gains under the tested predictor specification.
3. **Decision resolution and practical alternatives:** a cross-seed resolution benchmark establishes recoverable corpus ordering, while demonstrator and proxy baselines locate the audit relative to realistic selectors.

2 Related Work

Vision–language–action learning: Language-conditioned robot learning spans task-specific imitation (Lynch & Sermanet, 2021; Mees et al., 2022; Nair et al., 2022), large pretrained embodied models (Driess et al., 2023; Brohan et al., 2023; Kim et al., 2025; Ghosh et al., 2024; Black et al., 2025), and heterogeneous robot corpora (O'Neill et al., 2024). Physical VLM-to-VLA transfer is increasingly practical (Hancock et al., 2026); NaVILA separates language-level navigation decisions from low-level legged control (Cheng et al., 2025), as we do. We study the preceding data decision: which fixed demonstration corpus should train a policy?

Language grounding and counterfactual evaluation: Vision-and-language navigation established grounded instruction following (Anderson et al., 2018), but success can hide instruction-insensitive shortcuts. Controlled tests expose memorized action sequences and semantic failures (Fei et al., 2025; Zhou et al., 2025; Hou & Zhao, 2026; Yu et al., 2026), while mismatched instructions can supply corrective supervision (Kondoh et al., 2026). Our intervention similarly holds the state fixed while

changing the goal, but it measures a data-generating demonstrator and predicts a separately trained learner. Crucially, we distinguish unsigned response change from correctness-resolved response structure.

Robot-data selection and valuation: Robot policies are sensitive to demonstration composition and scale (Mandlekar et al., 2022; Saxena et al., 2025; Lin et al., 2025). Data valuation uses marginal contribution, training dynamics, gradients, or standardized retraining (Ghorbani & Zou, 2019; Swayamdipta et al., 2020; Xia et al., 2024; Gadre et al., 2023). DemInf scores demonstrations through state–action mutual information (Hejna et al., 2025); DataMIL learns policy-aware datamodels (Dass et al., 2026). Unlike end-to-end selectors such as DataMIL, we test whether an interpretable pre-training corpus diagnostic retains incremental value at the downstream selection decision. Our unit is an entire candidate corpus, and the nested comparison conditions on marginal demonstrator correctness.

Proxy selection and decision-aware evaluation: Selection via Proxy uses a smaller model to rank data efficiently (Coleman et al., 2020); compute-aware selection makes the appropriate proxy depend on its cost (Yin & Rush, 2025). Small training runs can preserve data-recipe rankings inconsistently across proxy configurations (Wang et al., 2026). Upstream probes can correlate with control (Dong et al., 2026), and proxy environments can be judged by absolute prediction and rank preservation (Quevedo et al., 2026). Prior work therefore distinguishes absolute prediction quality from rank preservation and selection utility. We extend this distinction to prospective robot-corpus selection with prespecified decisions, a proper-score test, a seed-based resolution benchmark, and direct comparison with practical selectors.

3 METHOD

Figure 2 summarizes the prospective comparison and the disjoint roles of audit, training, and held-out evaluation data.

3.1 SAME-STATE GOAL-SWAP AUDIT

For each audit state, we query the demonstrator twice with an identical visual observation and action history. Only the object goal changes, and the two goals require opposite branches. Let d index a corpus, u a state, x^-, x^+ the goals, and $B_d(u, x)$ the first response. For a deterministic demonstrator, paired instruction dependence is

$$A_d = \frac{1}{n_A} \sum_{u=1}^{n_A} \mathbf{1}\left[B_d(u, x^-) \neq B_d(u, x^+)\right]. \tag{1}$$

It is an unsigned response statistic: changing correctly for both goals and changing incorrectly for both goals can each increase A_d. Under the balanced deterministic intervention, each unequal pair identifies the goal and contributes one bit, so Eq. 1 equals state-averaged conditional mutual information (Appendix A). This equivalence does not make A_d a correctness or downstream-value measure.

The audit requires counterfactual access to a queryable demonstrator and is unavailable from a fixed corpus containing only one instruction per state. It is therefore applicable in settings such as automated simulation or online data generation, where the same state can be replayed under alternate goals before committing to a large collection. The study therefore concerns this access model rather than arbitrary human demonstrations.

3.2 WHAT DOES THE SWITCH RATE CONTAIN?

Let a_d, b_d, and c_d be the fractions of audit states for which both goal responses are correct, both are wrong, or exactly one is correct. Marginal per-goal demonstrator correctness is

$$p_d = a_d + \tfrac{1}{2} c_d = \tfrac{1}{2} + \tfrac{1}{2}(a_d - b_d). \tag{2}$$

If responses are restricted to left/right, opposite required branches imply $A_d = a_d + b_d$. Thus, in the binary case, (p_d, A_d) is an invertible reparameterization of (a_d, b_d): p_d contains their signed

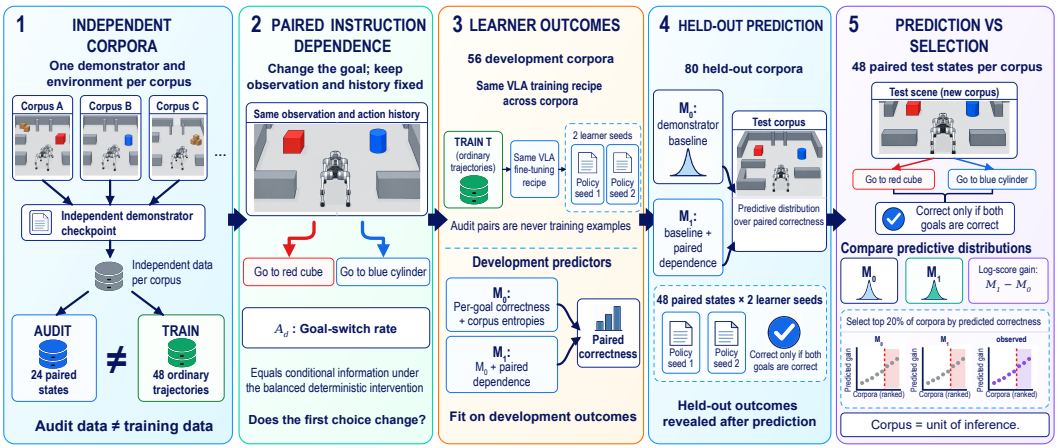

Figure 2: **Prospective test of paired response structure;** Audit states and training trajectories are disjoint, and only the goal changes within an audit pair. Baseline M_0 uses marginal per-goal correctness and corpus entropies; M_1 adds the same-state switch rate A_d. Both models predict paired learner correctness and rank the same held-out corpora.

difference, whereas A_d contains their unsigned sum. Our demonstrators may also emit HALT; this breaks the exact $A_d = a_d + b_d$ identity in our data, although Eq. 2 remains exact. The decomposition makes the primary question precise: does unsigned joint response structure improve prediction beyond marginal correctness and corpus composition?

3.3 Nested prediction test

After fine-tuning with learner seed s, let $C_{d,s} \in \{0, \ldots, m\}$ count evaluation states where the VLA predicts the correct branch for *both* goals. This conjunction rejects constant-side policies. We compare nested beta–binomial predictors:

$$C_{d,s} \sim \text{BetaBinomial}(m, \mu_{d,s}, \rho), \tag{3}$$

$$M_0: \quad \text{logit}(\mu_{d,s}) = \alpha + \beta^\top W_d + \gamma_s, \tag{4}$$

$$M_1: \quad \text{logit}(\mu_{d,s}) = \alpha + \beta^\top W_d + \eta A_d + \gamma_s. \tag{5}$$

W_d contains marginal per-goal demonstrator correctness, training-goal entropy, and coarse training-action entropy. Both predictors therefore use the same 48 demonstrator responses; $M_1 - M_0$ isolates pairing structure rather than extra queries. The crossed seed effect γ_s is a nuisance term and ρ captures extra-binomial variation.

Feature scaling and the ridge penalty are learned on development corpora. The primary M_1 shares M_0's penalty so that adding A_d does not also add a hyperparameter search. Test-audit uncertainty is marginalized by resampling the 24 paired states. Exploratory checks separately tune the ridge penalty for each response parameterization, decompose A_d into a_d and b_d, and add existing rollout features. None changes the prespecified endpoint or primary result.

3.4 Prediction, ranking, and selection

The primary estimand is the equal-corpus improvement in proper log score,

$$\bar{\delta} = \frac{1}{|\mathcal{B}|} \sum_{d \in \mathcal{B}} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left[\log p_{1,d,s}(C_{d,s}) - \log p_{0,d,s}(C_{d,s}) \right], \tag{6}$$

where $\mathcal{B}$ is the held-out corpus population and $\mathcal{S}$ the two learner seeds. A corpus-cluster bootstrap supplies the 95% interval. The prespecified primary rule requires its lower bound to exceed zero. The $0.105 \approx -\log(0.9)$-nat magnitude reference corresponds to multiplying the geometric-mean assigned probability by approximately 1.11; it is not a universal utility threshold. Log score (Gneiting & Raftery, 2007) evaluates the full predictive distribution and penalizes confident errors.

For selection, model M_j ranks each corpus by predicted mean correctness,

$$\widehat{q}_j(d) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{E}_{M_j}[C_{d,s}/m], \tag{7}$$

and selects $K = 16$ of 80 corpora. We report selected correctness, regret relative to the realized oracle set, and oracle overlap. The operating point was prespecified before outcome reveal; its confidence interval is exploratory. No held-out outcome enters fitting, tuning, or ranking.

3.5 Resolution and comparator analyses

Finite evaluation states and learner seeds limit any ranking method. We use cross-seed ranking as an empirical resolution benchmark and compare against demonstrator and proxy selectors. We also repeat M_0/M_1 comparisons after excluding the two least mature checkpoints. These exploratory analyses do not alter the prespecified primary test.

4 Experimental Setup

4.1 Corpus population

We study Unitree Go2 object-goal navigation in Isaac Lab (Mittal et al., 2023). Given forward RGB, instruction text, and action history, the VLA emits a discrete navigation command for a fixed low-level controller; only that controller sees proprioception. Every corpus uses a separately trained tabular Q-learning demonstrator with unique training, layout, physics, and rendering seeds. Checkpoints span 10, 50, 250, 1,000, and 50,000 episodes. The held-out population has 16 corpora per level; development has 12 at 10 episodes and 11 at each remaining level. Checkpoint maturity is excluded from the deployable baseline because it is an experimental assignment, not a property available for an arbitrary corpus.

Each corpus contains 24 paired audit states, 48 unpaired multi-step training trajectories, and 48 paired in-distribution (ID) evaluation states. Audit, training, and evaluation packets are disjoint at the state, image, and trajectory levels. All 6,528 attempted training rollouts passed prespecified structural-validity checks, so acceptance is constant across corpora and cannot explain their ranking; first-branch correctness instead ranges from .375 to .813. Appendix I gives the complete validity checklist.

Across 136 corpora, the study contains 6,528 training trajectories (43,084 steps), 3,264 audit states, and 6,528 ID evaluation states. The 56 development and 80 held-out corpora share no demonstrator seeds or scene realizations. Figure 3 summarizes the data roles and evaluations.

The audit requires 48 demonstrator queries and costs a median 4.18 GPU-min per corpus. A Qwen2.5 fit and ID evaluation average 11.46 GPU-min for one seed, or $2.74\times$ the audit cost. Appendix Table 16 reports the complete measured cost comparison.

4.2 VLA training and evaluation

The primary learner is Qwen2.5-VL-3B-Instruct (Bai et al., 2025); Qwen3.5-VL-0.8B is both a supporting backbone and an exploratory proxy selector (Qwen Team, 2026). Within each backbone, only corpus and learner seed vary. We hold the vision tower fixed and train rank-8 LoRA adapters (Hu et al., 2022) on language-model projections. Outputs are left/right turns, forward, halt, or done; invalid generations are errors. The full factorial gives 544 adapters. A Qwen2.5 fit averages 10.68 GPU-min and Qwen3.5 averages 13.41 GPU-min over the measured runs.

ID and controlled shift: The ID endpoint has 48 paired states per corpus; both goals must receive the required branch. The 24-state shifted set crosses two spawn/target geometries with two lighting settings, six states per cell. Camera, warehouse, objects, prompt, action space, and controller remain fixed, so this is controlled rather than open-world shift.

Closed-loop simulation: For 16 held-out corpora at evenly spaced A_d ranks, the first Qwen2.5 seed runs 24 scenes and both goals (768 episodes). We record first-command correctness, target reach,

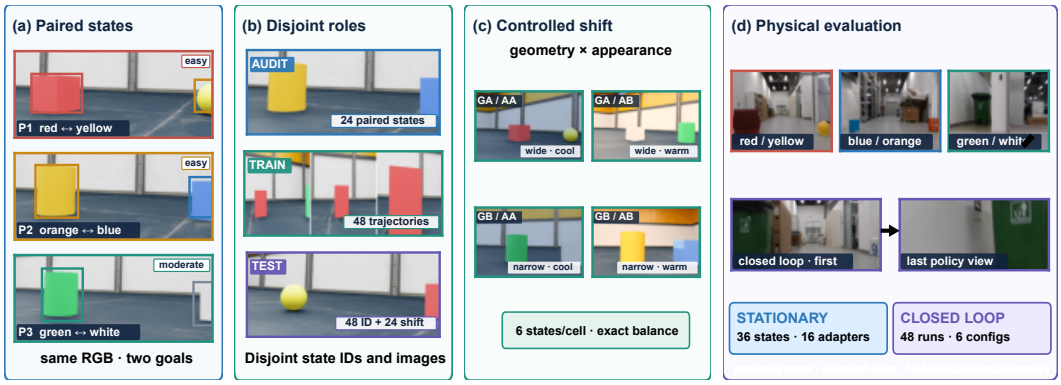

Figure 3: **Experimental design across simulation and hardware;** (a) Paired states span three semantic families with RGB fixed across goals. (b) Audit, training, and evaluation observations are disjoint. (c) Controlled scene shift balances a 2×2 geometry/appearance design. (d) Stationary real images evaluate 16 adapters; 48 physical runs describe two deployed adapters.

collision/fall, path efficiency, and recovery. These aggregate rates and exploratory rank correlations describe the stratified subset, not all 80 held-out corpora.

Physical evidence: Stationary evaluation uses 36 human-screened Go2 images across three layouts, three surrogate-object pairs, two sides, and two difficulties (72 queries per adapter). Closed-loop deployment uses one adapter per backbone, chosen before physical evaluation by simulation outcomes from the 16 committed roots. The VLA acts from onboard RGB and history; a depth safety mask can veto forward motion. The 48 physical runs are descriptive because terminal state was not externally adjudicated and parts of the recorded protocol departed from the initial design (Appendix G.1).

4.3 Inference protocol

The primary metric is Eq. 6. Secondary diagnostics include absolute error, Pearson/Spearman association, top-20% correctness, regret, and oracle overlap. Predictors, held-out corpus identifiers, and the top-20% operating point were prespecified before held-out outcomes were revealed. Inferential intervals resample whole corpora, preserving nested states, paired goals, and crossed seeds. The main endpoint and selection rule were unchanged after reveal; decomposition, resolution, proxy, calibration, and closed-loop correlation analyses are explicitly exploratory. Appendix I records the content-addressed prospective procedure.

5 Results

5.1 Reliability, association, and incremental value diverge

Figures 4–5 and Table 1 summarize the held-out association, predictive, and selection results.

Across all 136 corpora, A_d has Spearman $\rho = .467$ (95% CI $[.300, .616]$) with Qwen2.5 paired correctness; on the 80 held-out corpora, $\rho = .356$ ($[.111, .565]$). Audit reliability is .802 overall and .766 on held-out corpora. The audit therefore measures a demonstrator property with substantial estimated reliability and tracks broad learner differences.

The incremental comparison is more uncertain. Adding A_d to M_0 gives a Qwen2.5-ID mean log-score gain of $+.059$ nats (95% corpus-bootstrap CI $[-.027, .133]$). The interval crosses zero, and separately tuning M_1's ridge penalty changes the gain only to $+.046$ ($[-.102, .171]$). The same pattern appears under controlled shift and with Qwen3.5: mean gains are positive but small, and all intervals include zero (Table 1). Central predictive intervals are conservative rather than globally overconfident: 80% coverage is .925 for M_0 and .938 for M_1 (Appendix Table 9).

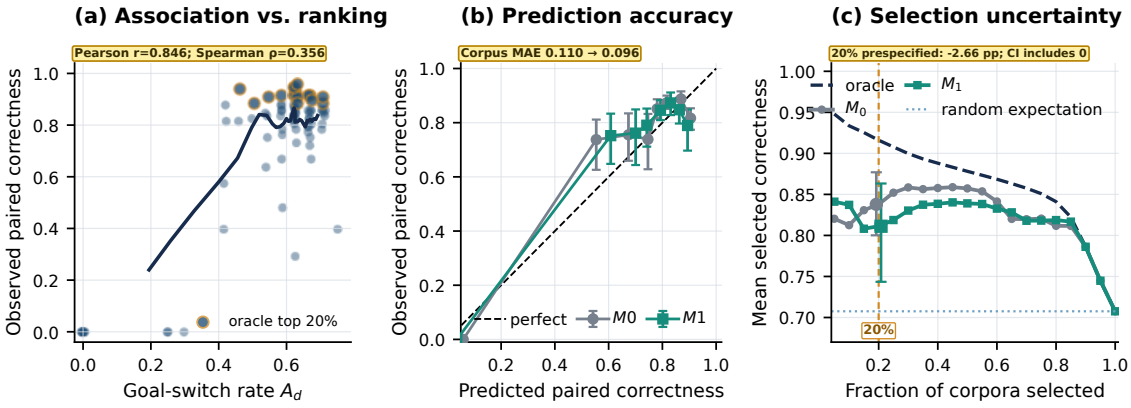

Figure 4: **Association, prediction, and selection answer different questions;** (a) Held-out corpora, a 15-corpus moving average, and the realized top 20%. (b) Equal-count bins with corpus-bootstrap intervals and corpus MAE. (c) The 20% operating point was prespecified before outcome reveal; its displayed confidence interval and the remaining curve are exploratory.

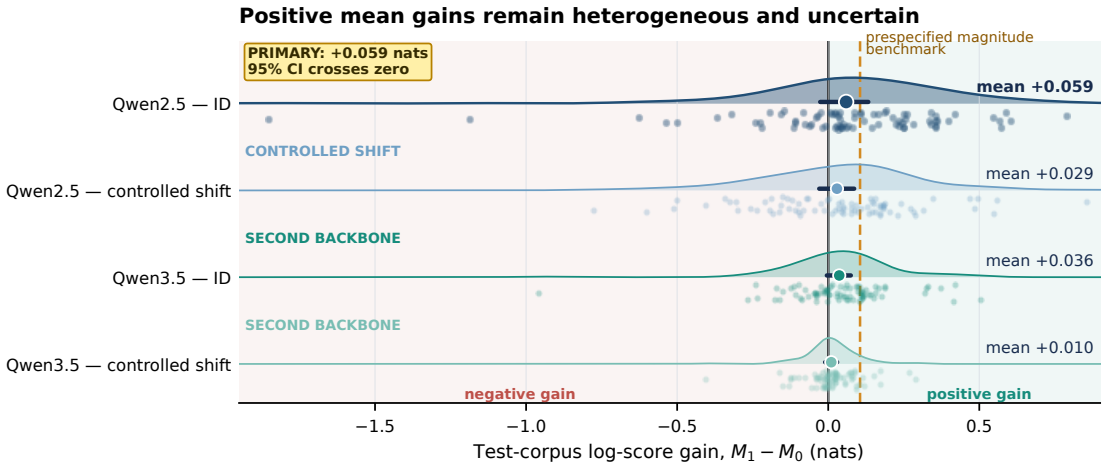

Figure 5: **Incremental gains are heterogeneous and uncertain;** Light bands distinguish negative from positive gain. Points are held-out corpora, half-densities show gain distributions, markers are equal-corpus means, and bars are corpus-bootstrap 95% intervals. Every interval crosses zero and every mean lies below the prespecified magnitude reference.

Table 1: **Cross-setting summary;** gain is $M_1 - M_0$ log score. Arrows show $M_0 \to M_1$; lower MAE and higher Spearman/selected correctness are better.

Setting	Gain [95% CI]	MAE	Spearman	Selected
Qwen2.5 ID	$+.059\ [-.027, .133]$	$.110 \to .096$	$.589 \to .490$	$.838 \to .811$
Qwen2.5 shift	$+.029\ [-.030, .086]$	$.106 \to .097$	$.460 \to .391$	$.945 \to .917$
Qwen3.5 ID	$+.036\ [-.005, .075]$	$.077 \to .073$	$.752 \to .745$	$.965 \to .971$
Qwen3.5 shift	$+.010\ [-.010, .030]$	$.080 \to .078$	$.729 \to .723$	$.969 \to .964$

5.2 Unsigned switching hides correctness-resolved structure

The demonstrator emits HALT in 10.9% of audit queries, and 107/136 corpora contain at least one nonbinary pair. Thus the binary identity $A_d = a_d + b_d$ is not exact: $A_d - (a_d + b_d)$ averages .100

and reaches .250. The marginal-correctness identity in Eq. 2 is exact. At the 10-episode checkpoint, mean $(a_d, b_d, c_d) = (.156, .016, .827)$; at 50 episodes it is $(.389, .074, .537)$. Both-wrong behavior is present but does not dominate these checkpoint averages.

The decomposition identifies a candidate source of the mismatch. In exploratory held-out tests, $M_0 + a_d$ gains $+.104$ nats ($[.049, .153]$) and $M_0 + b_d$ gains $+.110$ ($[.022, .188]$), each exceeding the $+.059$ gain from their unsigned, HALT-affected mixture. Their selected correctness is .846 and .842, compared with .811 for M_1. Under the same exploratory modeling protocol, these correctness-resolved parameterizations yield larger held-out gains than the unsigned switch rate. This ordering persists when the ridge penalty is selected separately on development corpora: the selected penalty is 1 for all three alternatives, with gains $+.126$ for a_d, $+.123$ for b_d, and $+.046$ for A_d (Appendix E). The comparison remains post-hoc and can reflect the nonbinary HALT cases, finite-sample estimation, or regularized model geometry; it does not establish an information-theoretic advantage.

5.3 The endpoint contains resolvable rank signal

The two Qwen2.5 learner seeds have rank agreement $\rho = .567$ across 80 corpora. Selecting 16 corpora by seed 1 and evaluating with seed 2 gives correctness .874, oracle overlap .50, and regret .057; reversing the seeds gives the same selected correctness, .50 overlap, and .052 regret. This empirical resolution benchmark is imperfect, but it is materially above both M_0 (.838 selected) and M_1 (.811), so finite-state and learner-seed variation do not fully explain the selection result.

The ranking gap persists after excluding the 10- and 50-episode checkpoints that create the clearest low-performance cluster. Among the remaining 48 corpora, M_0 has Spearman .349 and selected correctness .821, whereas M_1 has Spearman .172 and selected correctness .791. The ordering problem is therefore not solely an artifact of immature demonstrators.

5.4 Practical baselines locate the audit's decision value

Table 2 compares all methods on the same Qwen2.5 held-out outcomes. Demonstrator closed-loop success measured on existing simulator rollouts is inexpensive but weak ($\rho = .281$). Training-trajectory first-branch correctness, also available from the existing simulator rollouts, improves proper score by $+.156$ nats ($[.078, .225]$), yet does not improve top-20% selection over M_0. Structural rollout acceptance is 100% for every corpus and adds no information.

Qwen3.5 provides the highest observed rank correlation among the tested selectors: pooled outcomes correlate $\rho = .715$ with Qwen2.5 and select corpora with .866 Qwen2.5 correctness. Even one Qwen3.5 seed reaches $\rho = .644$–$.662$ and .846–.874 selected correctness. In this regime, the proxy therefore preserves the target-model ordering better than either audit-based predictor. It does not establish a cost win: a measured Qwen3.5 fit averages 13.41 GPU-min, versus 4.18 for the audit and 11.46 for one Qwen2.5 fit plus evaluation. At these measured costs, an economic advantage for the audit would require amortizing its queries across multiple downstream learners or collection decisions rather than replacing a single inexpensive adapter fit.

Table 2: **Exploratory selection baselines on Qwen2.5-ID;** all methods rank the same 80 corpora; $K = 16$. Gain is proper log score relative to the prespecified M_0 when defined. The seed-to-seed row is an empirical resolution benchmark rather than a deployable selector.

Selector	Gain [95% CI]	Spearman	Selected	Regret	Overlap
M_0	—	.589	.838	.078	.312
$M_1 = M_0 + A_d$	$+.059$ $[-.027, .133]$	.490	.811	.105	.250
$M_0 + a_d$	$+.104$ $[.049, .153]$	.560	.846	.070	.312
$M_0 + b_d$	$+.110$ $[.022, .188]$	.531	.842	.074	.250
M_0+ training first-correct	$+.156$ $[.078, .225]$	.584	.831	.085	.312
Demonstrator rollout success	—	.281	.838	.078	.312
Qwen3.5 outcomes	—	.715	.866	.050	.438
Qwen2.5 seed 1 $\rightarrow$ seed 2	—	.567	.874	.057	.500

5.5 Local response and navigation success diverge

Across 768 simulated episodes, first-command correctness is .829 (637/768), while target reach is .655 (503/768), a 17.4-point gap ($[10.4, 24.9]$). Of 637 correct-first episodes, 134 fail later; none of 131 wrong-first episodes recovers. Across the 16 corpora, target-reach rank correlations are .368 for A_d, .361 for M_0, and .423 for M_1. These small-sample descriptive values do not establish a corpus selector for embodied success. They show that the local endpoint and full navigation are distinct.

On 48 physical Go2 runs, the first choice is correct in 47, but strict terminal success is not estimable under the recorded protocol. We therefore retain physical results as qualitative transfer evidence and separate objective logs, operator reports, interventions, and protocol departures in Appendix G.1.

6 Scope and Limitations

The audit requires counterfactual demonstrator access and is tested on one simulator, one tabular demonstrator family, short corpora, and a largely maturity-driven population. The resulting claim is specific to this controlled corpus population. Generalization to different demonstrator families, corpus-generation mechanisms, and larger real-world corpora remains untested.

A_d tests one unsigned additive summary. Correctness-resolved parameterizations yield larger exploratory gains under the tested predictor specification, including after separate development-set penalty tuning, but these alternatives were not prespecified and do not by themselves identify the cause of that difference. Corpus value is defined relative to the fixed learner architecture and fine-tuning recipe used in each comparison. Qwen3.5 is a related backbone rather than an independent architecture family, and the proxy comparison uses outcomes collected for this study. Finally, closed-loop and physical evaluations support endpoint interpretation but do not enlarge the evidence for held-out corpus selection.

7 Conclusion

Paired instruction dependence provides a controlled case study of how proxy structure can be lost through compression. The switch rate is measured with substantial reliability and associated with learner behavior, but its unsigned form mixes both-correct and both-wrong responses; correctness-resolved parameterizations yield larger exploratory held-out gains under the tested predictor specification. Cross-seed agreement shows that corpus ordering remains recoverable. Together, these results motivate preserving outcome-relevant structure in counterfactual data diagnostics and validating them against the decisions they are intended to support.

Reproducibility Statement

The supplement specifies corpus construction, prompts, seeds, partitions, runtime, and statistical decisions. The artifact provides code, ordered split identifiers, predictions.

Ethics Statement

The corpus study and 768-episode closed-loop evaluation use simulation. The 48 physical Go2 runs used operator supervision and a depth-based safety mask; no in-run intervention was logged. These descriptive trials do not establish safety beyond the recorded environments. Deployment requires oversight, emergency stopping, and site-specific validation.

AI Use Statement

Language models assisted with code refinement and grammar review. The authors verified all suggested changes and remain fully responsible for the manuscript content.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018. doi: 10.1109/CVPR.2018.00387.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.010.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://robotics-transformer2.github.io/.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.018.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJg2b0VYDr.

Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Madry, Andrew Ilyas, and Roberto Martín-Martín. DataMIL: Selecting data for robot imitation learning with datamodels. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AcTsKglDdh.

Jiahua Dong, Yunze Man, Pavel Tokmakov, and Yu-Xiong Wang. Capturing visual environment structure correlates with control performance. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AmczI1k3Yk.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

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. doi: 10.48550/arXiv.2510.13626. URL https://arxiv.org/abs/2510.13626.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112, 2023. doi: 10.52202/075280-1179.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Tan, Lawrence Chen, Quan Vuong, Ted Xiao, Pannag Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.090.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Asher J. Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning VLMs into VLAs without catastrophic forgetting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sFO9d6XSlf.

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 XXI*, 2025. doi: 10.15607/RSS.2025.XXI.023.

Yuchen Hou and Lin Zhao. LangGap: Diagnosing and closing the language gap in vision-language-action models. *arXiv preprint arXiv:2603.00592*, 2026. doi: 10.48550/arXiv.2603.00592. URL https://arxiv.org/abs/2603.00592.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Haru Kondoh, Kei Ota, Asako Kanezaki, and Yueh-Hua Wu. CounterAlign: Counterfactual supervision for vision-language-action models. *arXiv preprint arXiv:2608.21740*, 2026. doi: 10.48550/arXiv.2608.21740. URL https://arxiv.org/abs/2608.21740.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pISLZG7ktL.

Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. In *Robotics: Science and Systems XVII*, 2021. doi: 10.15607/RSS.2021.XVII.047.

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 *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1678–1690. PMLR, 2022. URL https://proceedings.mlr.press/v164/mandlekar22a.html.

Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning over unstructured data. *IEEE Robotics and Automation Letters*, 7(4):11205–11212, 2022. doi: 10.1109/LRA.2022.3196123.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

Suraj Nair, Eric Mitchell, Kevin Chen, Brian Ichter, Silvio Savarese, and Chelsea Finn. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1303–1315. PMLR, 2022. URL https://proceedings.mlr.press/v164/nair22a.html.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, et al. Open x-embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://robotics-transformer-x.github.io/.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hidBHy1CAw.

Qwen Team. Qwen3.5: Towards native multimodal agents. Technical report and model release, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LqhorpRLIm.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746.

Jiachen T. Wang, Tong Wu, Kaifeng Lyu, James Zou, Dawn Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2FZC0c06jP.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Junjie Oscar Yin and Alexander Rush. Compute-constrained data selection. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4es2oO9tw1.

Bin Yu, Yao Zhang, Haishan Liu, Shijie Lian, Yuliang Wei, Xiaopeng Lin, Zhaolong Shen, Changti Wu, Ruina Hu, Bailing Wang, Cong Huang, and Kai Chen. RoboSemanticBench: Diagnosing semantic grounding in action prediction for VLA models. *arXiv preprint arXiv:2606.02277*, 2026. doi: 10.48550/arXiv.2606.02277. URL https://arxiv.org/abs/2606.02277.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025. doi: 10.48550/arXiv.2510.03827. URL https://arxiv.org/abs/2510.03827.

A Information-Theoretic Motivation

For completeness, let $q(U, X, B)$ be the demonstration distribution and define the best achievable log losses with and without the goal:

$$\mathcal{R}^*_U = \inf_f \mathbb{E}_q[-\log_2 f(B \mid U)], \qquad \mathcal{R}^*_{U,X} = \inf_g \mathbb{E}_q[-\log_2 g(B \mid U, X)]. \tag{8}$$

Expected log loss conditioned on the inputs equals the entropy of the true conditional distribution plus a nonnegative KL divergence. It is minimized by $f^*(b \mid u) = q(b \mid u)$ and $g^*(b \mid u, x) = q(b \mid u, x)$, giving

$$\mathcal{R}^*_U - \mathcal{R}^*_{U,X} = H_q(B \mid U) - H_q(B \mid U, X) = I_q(B; X \mid U). \tag{9}$$

Thus $I_q(B; X \mid U)$ measures the ideal log-loss value of the goal for predicting demonstration actions. Under the balanced deterministic intervention in the main study, each state contributes exactly zero or one bit, so their empirical average is A_d. Outside that setting, switch rate and conditional mutual information need not coincide. This is a statement about paired demonstrator responses, not a theorem about optimization, finite-model learning, action correctness, or navigation success.

B Endpoint and Predictor Details

For evaluation state u, let B^-_u and B^+_u be the required branches and $\widehat{B}_{d,s}$ the VLA prediction. The count modeled in Eq. 3 is

$$C_{d,s} = \sum_{u=1}^{m} \mathbf{1}\left[\widehat{B}_{d,s}(u, x^-) = B^-_u \ \wedge \ \widehat{B}_{d,s}(u, x^+) = B^+_u\right]. \tag{10}$$

This conjunction rejects constant-side policies that can attain deceptively high per-goal accuracy. It also keeps the local decision endpoint separate from closed-loop target reach.

For the top-K decision, let $\bar{Y}_d$ be realized correctness for test corpus d averaged over the two learner seeds, and let $q_M(d)$ be its score under predictor M. Define the selected and oracle sets as

$$\widehat{\mathcal{T}}_M = \mathrm{TopK}_{d \in \mathcal{B}}\, q_M(d), \qquad \mathcal{T}^* = \mathrm{TopK}_{d \in \mathcal{B}}\, \bar{Y}_d. \tag{11}$$

The reported selected correctness is $K^{-1} \sum_{d \in \widehat{\mathcal{T}}_M} \bar{Y}_d$, regret is the oracle mean minus this quantity, and oracle overlap is $|\widehat{\mathcal{T}}_M \cap \mathcal{T}^*|/K$. Score ties are broken deterministically by corpus identifier. These definitions apply only to source-level selection among the 80 fixed test corpora.

We use the mean–intraclass-correlation parameterization

$$\alpha^{\mathrm{BB}}_{d,s} = \mu_{d,s}\frac{1-\rho}{\rho}, \qquad \beta^{\mathrm{BB}}_{d,s} = (1-\mu_{d,s})\frac{1-\rho}{\rho}, \qquad C_{d,s} \sim \mathrm{BetaBinomial}\big(m, \alpha^{\mathrm{BB}}_{d,s}, \beta^{\mathrm{BB}}_{d,s}\big)\,. \tag{12}$$

Each fitted predictor has one unpenalized dispersion scalar ρ, shared over all development corpus–seed rows within that fit; M_0 and M_1, backbones, and evaluation settings are fitted separately. We estimate regression coefficients and ρ by penalized maximum likelihood on the 56 development

corpora using L-BFGS-B. The ridge penalty applies only to standardized scientific features: the intercept, learner-seed nuisance effect, and ρ are unpenalized.

The varying columns of W_d are marginal per-goal demonstrator correctness, instruction entropy, and action entropy. The ridge value $\lambda = 10$ is selected by whole-corpus cross-validation of M_0 and then shared with M_1 to isolate the added feature rather than an added hyperparameter search. Coefficients in Table 4 are on standardized development features. Their magnitudes are descriptive in the presence of correlated predictors; they should not be interpreted causally. An exploratory five-fold whole-corpus search for M_1 selects $\lambda = 1$ and yields a held-out gain of $+.046$ nats ($[-.102, .171]$), so separate regularization does not materially change the conclusion.

Table 3: Measurements entering M_0 and M_1. Marginal correctness and A_d reuse the same 48 demonstrator responses but summarize them differently.

Feature	Definition	Same-state pairs?	Correctness labels?	Range
Per-goal correctness	Fraction of goal-conditioned queries with the correct demonstrator branch	No	Yes	.438–.792
Instruction entropy	Entropy of training instruction labels	No	No	1.564–1.585
Action entropy	Entropy of coarse training actions	No	No	.902–1.524
A_d	Fraction of paired states whose first choice changes	Yes	No	0–.750

Table 4: Baseline feature ranges, correlation with A_d over 136 corpora, and standardized coefficients. Checkpoint maturity is included only in the exploratory sensitivity.

Feature	Range	Pearson r	Spearman ρ	M_0 coef.	M_1 coef.
Per-goal correctness	.438–.792	.718	.462	.842	.541
Instruction entropy	1.564–1.585	$-.470$	$-.454$	$-.001$	.113
Action entropy	.902–1.524	.815	.564	.966	.549
A_d	0–.750	1.000	1.000	—	1.089
$\log_{10}$ checkpoint episodes	1.000–4.699	.570	.566	—	—

The primary baseline excludes checkpoint episode count because it is a simulator-specific experimental assignment rather than a measurement available for an arbitrary collected corpus. In an exploratory sensitivity, adding it to both nested predictors changes the test-corpus log-score gain to $+0.030$ nats (95% corpus-bootstrap CI $[-0.036, 0.090]$). This sensitivity does not strengthen the incremental case.

For each test corpus, the recorded 24 binary switch indicators are sampled with replacement 256 times. Each draw supplies one audit value to M_1; the 256 resulting beta–binomial mass functions are averaged before scoring. Development audit values enter this primary fit as observed features. The primary log score therefore integrates test-corpus finite-audit variability but does not propagate development-audit error through predictor fitting. The nested sensitivity in Appendix D performs that additional resampling and refit. The outer 10,000-draw bootstrap resamples whole corpora and serves a different purpose: uncertainty over the population of corpora.

C Prospective Corpus-Population and Precision Design

The 56-development/80-test allocation, ordered corpus IDs, fitted predictors, and predictions were stored in a content-addressed internal commit before held-out outcomes were revealed. The accompanying certificate records the commit hash and forbids subsequent refitting. This is an auditable prospective record, not a public preregistration; the anonymized commit and certificate are included with the artifacts. The top-20% operating point was prespecified, whereas its confidence interval was computed exploratorily.

Candidate allocations varied development corpora, test corpora, and learner seeds under a generative model fitted to preliminary development results. The effect coefficient was the smallest value on the prespecified grid whose calibration mean exceeded the 0.105-nat magnitude benchmark; the discrete grid produced a simulated mean gain of 0.174 nats. Among the three highest-power cells, the rule required estimated power of at least .82 and zero failed fits, then selected minimum Qwen

compute, more test corpora, and higher power in that order. The chosen 56/80/two-seed cell had screening power .898 over 2,000 simulations. An independent 5,000-simulation confirmation gave .8934 (Monte Carlo 95% interval $[.8848, .9020]$), with zero failed fits. This documents the design rule, not 89% power for an effect of exactly 0.105 nats; only 12.5% of the realized corpus-bootstrap replicates exceed that magnitude reference.

Table 5: Target corpus population. All varying seeds are disjoint across development and test sets; nested states and goals do not increase the independent-corpus count.

Factor	Levels or sampling	Status
Corpus population	independently trained RL demonstrators	varying
Split	56 development / 80 test	disjoint
Demonstrator checkpoint	10, 50, 250, 1,000, 50,000 episodes	test: 16 each; dev.: 12/11/11/11/11
Demonstrator and scene seeds	unique training, layout, physics, rendering seeds	varying
Goal families	three semantic object pairs, balanced sides	fixed design
Audit states	24 paired states / 48 counterfactual queries	fixed quota
Training data	48 unpaired multi-step trajectories	fixed quota
Evaluation data	48 paired states per setting	fixed quota
Learner seeds	two shared seeds per corpus	crossed

D Robustness of the Primary Conclusion

For corpus d, estimate finite-state sampling variance as $A_d(1 - A_d)/(n_A - 1)$. If s_A^2 is the observed between-corpus variance, the classical variance-decomposition reliability is

$$R_A = \max\left(0, 1 - \frac{|\mathcal{D}|^{-1} \sum_{d \in \mathcal{D}} A_d(1 - A_d)/(n_A - 1)}{s_A^2}\right). \tag{13}$$

$R_A = 1$ means estimated state-sampling noise explains none of the observed between-corpus variance; $R_A = 0$ means it explains all of it. It is not a calibration curve or intraclass coefficient.

This estimate treats the 24 audit states within a corpus as exchangeable draws. It therefore does not capture additional dependence among states or uncertainty from unmodeled state strata.

Table 6 separates the prespecified estimate from exploratory sensitivity checks. Reliability intervals resample corpus roots, and seed rows retain one crossed learner seed at a time. For the nested sensitivity, each replicate samples the 24 observed switch indicators with replacement within every development and test corpus, recomputes A_d, refits the unchanged M_1 specification on development outcomes, and evaluates the same test outcomes while holding M_0 fixed. A whole-corpus resample at evaluation propagates test-population uncertainty.

Table 6: Sensitivity of the primary Qwen2.5-ID conclusion. The prespecified row remains primary; all other rows are exploratory checks. Intervals resample whole corpora unless explicitly described as audit-measurement resampling.

Check	Result	Interpretation
Audit reliability, all 136	$.802\ [.688, .859]$	stable corpus property
Audit reliability, test 80	$.766\ [.523, .853]$	stability persists on test corpora
Primary gain, pooled seeds	$+.059\ [-.027, .133]$	positive, uncertain
Bootstrap replicates > 0	92.3%	positive direction
Bootstrap replicates $> .105$	12.5%	12.5% exceed benchmark
Learner seed 20262001	$+.067\ [-.051, .164]$	positive, uncertain
Learner seed 20262002	$+.052\ [-.025, .127]$	positive, uncertain
Audit resampling + refit	$+.025\ [-.084, .111]$	interval spans zero
Selected correctness, $M_1 - M_0$	$-.0266\ [-.0990, +.0234]$	interval spans zero
Resampled selection, $M_1 - M_0$	$+.0019\ [-.0605, +.0430]$	centered near zero

The 92.3% and 12.5% values are proportions of bootstrap replicates, not posterior probabilities. The $-.0266$ primary selection difference evaluates the single prespecified pair of selected sets on observed

outcomes. The $+.0019$ sensitivity repeatedly resamples audit states, refits M_1, and reselects corpora; it is therefore a different, exploratory measurement-error estimand. Its sign change reinforces that the selection effect is centered near zero rather than reversing the primary result. The audit coefficient remains positive across perturbations, but propagated log-score and selection effects remain uncertain.

E Exploratory Structural and Decision Diagnostics

These exploratory analyses examine identifiability, ranking resolution, and practical alternatives. They were conducted after held-out outcome release and do not modify the primary endpoint or prespecified predictions.

Audit decomposition: The audit responses are LEFT-30, RIGHT-30, or HALT. HALT constitutes 10.9% of queries, and 107/136 corpora contain a nonbinary pair. Consequently, $p_d = .5 + (a_d - b_d)/2$ holds exactly, while the binary-response identity $A_d = a_d + b_d$ does not: $A_d - (a_d + b_d)$ averages $.100$ and has maximum absolute value $.250$. Table 7 reports the two least mature checkpoints.

Table 7: Joint audit correctness at the two least mature checkpoints. a, b, and c denote both correct, both wrong, and exactly one correct.

Episodes	Corpora	a	b	c	Halt/query
10	28	.156	.016	.827	.015
50	27	.389	.074	.537	.098

Alternative features and selectors: Table 8 uses the same Qwen2.5-ID outcomes. The a_d and b_d rows add exact audit-derived values to M_0 at the common $\lambda = 10$. Separate five-fold tuning selects $\lambda = 1$ for A_d, a_d, and b_d, yielding $+.046$ ($[-.102, .171]$), $+.126$ ($[.037, .206]$), and $+.123$ ($[.026, .211]$) nats. This persistent ordering is post-hoc and does not isolate nonbinary responses from finite-sample or model-geometry effects. Training first-correct comes from accepted trajectories; acceptance is constant at 1.0 (6,528/6,528) and adds numerically zero. Qwen3.5 rows use collected learner outcomes rather than a refitted predictor.

Table 8: Exploratory Qwen2.5-ID diagnostics. Selected correctness, regret, and overlap use $K = 16$; gain is relative to the prespecified M_0 where available.

Method	Gain [95% CI]	Spearman	Selected	Regret	Overlap
$M_0 + a_d$	$+.104$ $[.049, .153]$	.560	.846	.070	.312
$M_0 + b_d$	$+.110$ $[.022, .188]$	.531	.842	.074	.250
M_0+ train first-correct	$+.156$ $[.078, .225]$	.584	.831	.085	.312
M_0+ acceptance	$+.000$ $[.000, .000]$	.589	.838	.078	.312
Demonstrator rollout success	—	.281	.838	.078	.312
Qwen3.5 pooled outcomes	—	.715	.866	.050	.438
Qwen3.5 seed 1	—	.644	.846	.070	.500
Qwen3.5 seed 2	—	.662	.874	.042	.500

Selection resolution: Qwen2.5 learner-seed outcomes have Spearman $\rho = .567$. Selecting by seed 1 and evaluating on seed 2 gives .874 correctness, .057 regret, and .50 oracle overlap; reversing seeds gives .874, .052, and .50. After excluding 10- and 50-episode corpora, 48 roots remain. M_0 then has Spearman .349, selected correctness .821, and overlap .20; M_1 has .172, .791, and .10. Thus seed variation limits empirical resolution but does not account for the full M_1 ranking deficit.

Predictive calibration: Because $C_{d,s}$ is discrete, we report mid-PIT moments and central predictive-interval coverage over 160 corpus–seed rows. Both predictors are conservative (Table 9); a few severe log-score misses are not evidence of global undercoverage.

Embodied target reach: Across the 16 rank-stratified closed-loop corpora, Spearman correlations with target reach are .368 for A_d, .361 for M_0, and .423 for M_1. These descriptive estimates use too few, deliberately stratified corpora to establish a target-reach selector.

Table 9: Held-out Qwen2.5-ID predictive calibration.

Model	Mid-PIT mean	Mid-PIT variance	50% cov.	80% cov.	90% cov.
M_0	.486	.061	.725	.925	.956
M_1	.487	.054	.763	.938	.963

F Complete Predictive and Selection Results

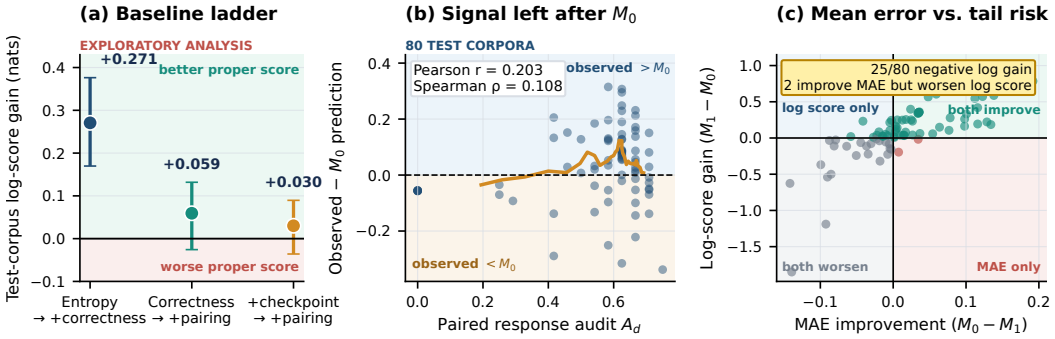

Figure 6: **Exploratory diagnostics of the remaining paired signal;** Light bands label the interpretation of each region. (a) Marginal per-goal correctness contributes substantially beyond corpus entropies, whereas the paired increment is smaller and uncertain. (b) A_d has a weak relationship with the signed residual left by M_0. (c) The four regions distinguish improvements in MAE, proper log score, both, or neither; confident misses can lower MAE while worsening proper score.

Table 10: Test-corpus predictive results. Correct is mean paired correctness across test corpora and learner seeds. Gain is $M_1 - M_0$ in nats; intervals resample whole corpora.

Backbone	Setting	Correct	Gain	95% CI
Qwen2.5	ID	.708	+.059	$[-.027, .133]$
Qwen2.5	Shift	.801	+.029	$[-.030, .086]$
Qwen3.5	ID	.761	+.036	$[-.005, .075]$
Qwen3.5	Shift	.776	+.010	$[-.010, .030]$

F.1 Development interface checks

Figure 7(c) shows the side-specific interface checks available for Gemma 3 4B and Qwen3.5. Gemma attains .99 left-goal but .55 right-goal correctness and fails the .60 per-side criterion; Qwen3.5 attains 1.00/.63 and advances to the corpus study. SmolVLM2 was evaluated in an earlier development screen rather than this side-specific check: 18 adapters achieve 31/432 paired decisions (.0718), compared with Qwen2.5's .6933 in the same screen. It was therefore not advanced to the corpus-level evaluation. These development diagnostics choose viable interfaces; they are not evidence for the primary predictive claim.

G Embodied and Physical Results

Figure 8 decomposes closed-loop outcomes at corpus and episode levels, while Figure 9 shows representative trajectories for the resulting success and failure modes.

The closed-loop decomposition is: 637/768 correct first commands, 503/768 target-reaching successes, 134 failures after a correct first command, and 131 failures after a wrong first command. A paired corpus bootstrap gives a 95% interval of $[10.4, 24.9]$ percentage points for the first-command minus task-success gap. Recovery is 0/131 and collision/fall is 0/768, with both zeros observed in every

Table 11: Absolute prediction error and rank diagnostics. Lower absolute error is better; higher correlation is better.

		Abs. error		Pearson		Spearman	
Backbone	Setting	M_0	M_1	M_0	M_1	M_0	M_1
Qwen2.5	ID	.110	.096	.887	.900	.589	.490
Qwen2.5	Shift	.106	.097	.908	.921	.460	.391
Qwen3.5	ID	.077	.073	.927	.934	.752	.745
Qwen3.5	Shift	.080	.078	.923	.925	.729	.723

Table 12: Top-20% corpus-selection diagnostics. Selected is realized mean correctness; regret is relative to the oracle top set. Regret and overlap are descriptive point estimates.

		Selected		Regret		Oracle overlap	
Backbone	Setting	M_0	M_1	M_0	M_1	M_0	M_1
Qwen2.5	ID	.838	.811	.078	.105	.312	.250
Qwen2.5	Shift	.945	.917	.055	.083	.250	.188
Qwen3.5	ID	.965	.971	.031	.026	.562	.625
Qwen3.5	Shift	.969	.964	.031	.036	.438	.438

one of the 16 corpus groups; we report these as descriptive counts without episode-level binomial intervals. Mean path efficiency across the 16 corpora is .995 under the recorded metric, so it does not localize the 134 downstream failures.

G.1 Physical closed-loop deployment

Figure 10 summarizes the hierarchy from objective recorder outputs to operator judgment and unavailable terminal verification.

Table 14 separates objective recorder outputs, operator judgment, and strict terminal evidence. The first two rows share the same 47/48 count but not the same evidential status: first-choice correctness is reconstructed from the expected and executed branch commands, whereas completion is the contemporaneous operator report. No run has the external target-zone measurement and independent adjudication required to verify strict success. The resulting 0/48 verified count denotes missing verification, not 48 observed failures.

The 47/48 hardware first-choice rate should not be compared directly with the 82.9% simulation-subset mean. Simulation averages 16 rank-stratified corpora, whereas hardware uses one median-selected corpus per backbone on six different physical configurations.

The seven known strict failures comprise four runs that reach the six-decision horizon without a final HALT or DONE, two communication/system aborts after motion, and the single run with a wrong first branch. The other 41 are indeterminate under the strict rule. For both backbones, the deployed corpus is the joint median of simulation-ID paired correctness over the 16 roots committed for closed-loop evaluation; the learner seed is fixed, and corpus-ID order resolves ties. No physical-image or physical-navigation outcome entered this choice; the exact corpus and seed remain in the reproducibility manifest. The selected Qwen2.5 adapter makes 12/12 correct first choices on each side; the selected Qwen3.5 adapter makes 12/12 on left goals and 11/12 on right goals. These nearly symmetric selected-policy counts do not contradict the aggregate stationary real-image result below, which averages 16 heterogeneous adapters.

Figure 11 reports the recorded motion, inference latency, and depth-safety mediation for the deployed policy stack.

G.2 Physical-Go2 image strata

Across the 36 stationary physical states, the same rank-stratified 16-adapter subset gives Qwen2.5 per-goal correctness .743, paired correctness .517, and instruction-switch rate .545; Qwen3.5 gives .811, .623, and .634. Exact parse rate is 1.0 for both. The left/right gaps (.958/.528 and .993/.628) motivate paired reporting: high aggregate per-goal correctness can conceal a side prior across

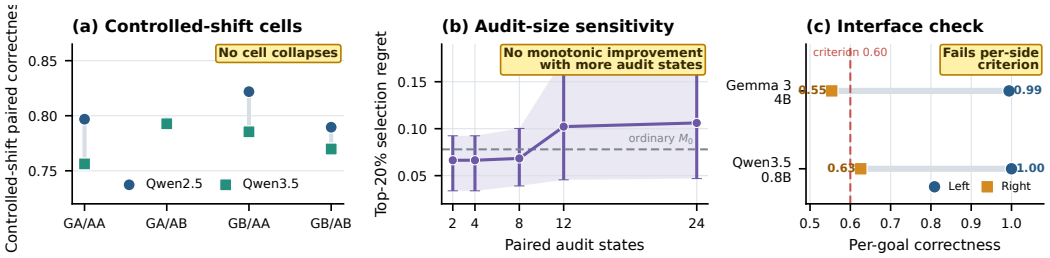

Figure 7: **Supporting analyses;** Controlled-shift performance remains substantial in every geometry/appearance cell; audit-size prefixes do not improve selection monotonically; and only models with an adequate instruction-conditioned action interface enter the test-corpus comparison.

Table 13: Paired correctness in the balanced geometry × appearance shift.

Cell	Qwen2.5	Qwen3.5
Geometry A / appearance A	.797	.756
Geometry A / appearance B	.794	.793
Geometry B / appearance A	.822	.785
Geometry B / appearance B	.790	.770

adapters. Before inference, human review required a readable camera frame, both intended surrogate targets identifiable on the recorded sides, sufficient visibility to assign the first branch, and a valid identical-image instruction pair. All 36 prescreened images were retained, including documented surrogate-prop and incidental-scene variation; no model output entered screening. The props are semantic surrogates rather than replicas of simulated objects, so this setting tests external-image robustness only.

For the same 16 corpus adapters and learner seed, matched ID simulation states yield left/right first-command accuracies of .973/.677 for Qwen2.5 and .999/.749 for Qwen3.5. The asymmetry therefore predates physical transfer, while the physical images amplify its right-goal deficit. This comparison is descriptive for the rank-stratified subset and does not estimate prevalence over all 80 test corpora.

Figure 12 summarizes stationary-image transfer across layouts, surrogate objects, and goal sides.

H Measured Cost

The full paired query set costs 48 demonstrator queries and a median 4.18 GPU-min per corpus. Once those responses have already been collected for the correctness-aware baseline, arranging them into same-state pairs and computing A_d requires zero additional demonstrator queries and negligible arithmetic. Thus "zero marginal query cost" is conditional on existing correctness-labeled responses; a pipeline without them incurs the full acquisition cost above. In this small-data regime, one Qwen3.5 proxy fit is not cheaper than one Qwen2.5 target fit plus evaluation (13.41 versus 11.46 mean GPU-min). Costs exclude one-time model download, rendering, and environment setup.

I Reproducibility Details

Within each backbone, the revision, prompt, processor, optimizer, update budget, decoder, grammar, and controller are fixed; Table 17 gives their exact values. We preserve all 136 corpus definitions, 544 adapter configurations, seeds, ordered partitions, asset hashes, 256 audit draws per corpus, the 10,000-draw primary bootstrap, $\lambda = 10$, and the 0.105-nat benchmark. Audit, training, and evaluation observations are disjoint. Predictions and rules were prespecified before test outcomes, with no outcome-dependent exclusion. Higher-draw robustness artifacts and all exploratory or descriptive analyses are separately labeled and do not alter the primary test.

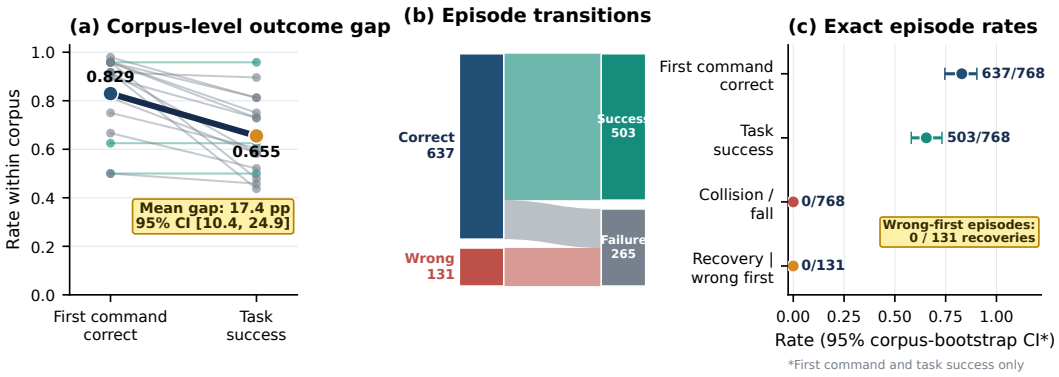

Figure 8: **Local response and embodied completion diverge;** (a) Lines pair first-command correctness and target success within each of 16 rank-stratified corpora. (b) Exact episode flows expose 134 later failures after a correct first command. (c) First-command and task-success intervals resample corpus groups; zero collision/fall and recovery counts are descriptive.

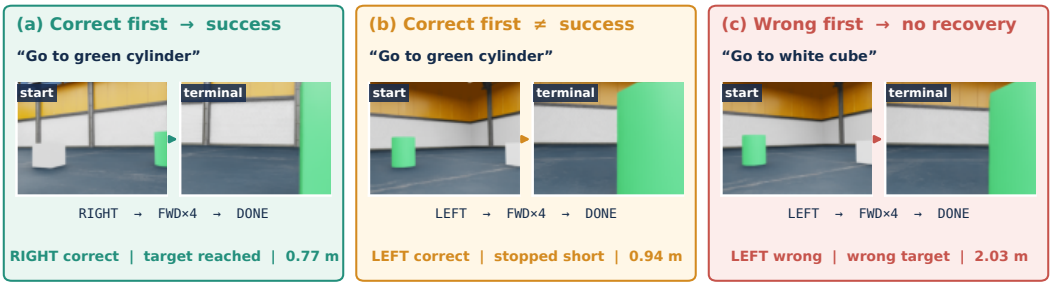

Figure 9: **A correct local decision can still fail downstream;** Representative simulated trajectories show success after a correct first command, failure after a correct first command, and failure after a wrong first command.

The exact prompt is version `go2_vla_prompt/v1`: *"You are the navigation policy of a Unitree Go2 quadruped robot. Instruction: '{instruction}'. The image is the current view from the robot's forward camera. Your recent actions, oldest first: {history}. Reply with exactly one next action and nothing else"*, followed by the five canonical forms for forward 25/50/75 cm, left or right 15/30/45 degrees, hold position, and instruction completion. The same builder and parser are used for training and deployment. During corpus rollout, a valid initial turn is followed by FORWARD-50, then DONE after the reach condition.

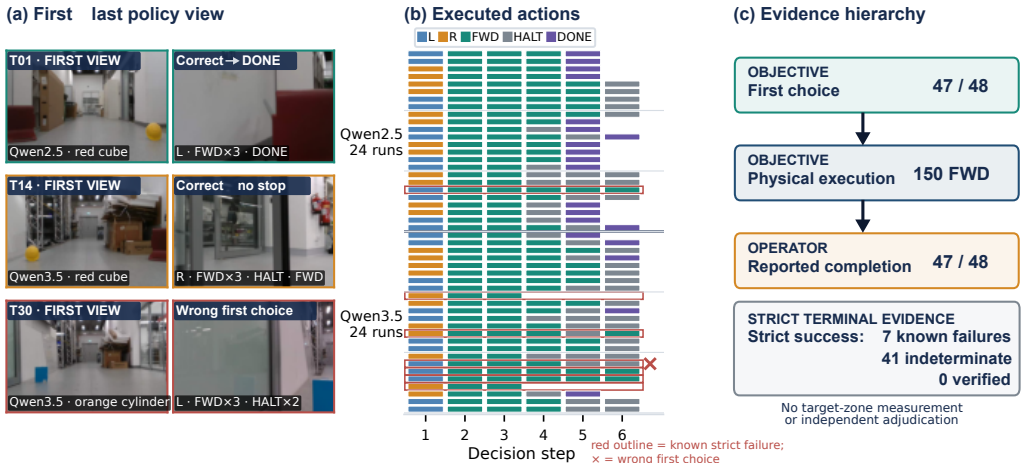

Figure 10: **Physical deployment transfers the local decision but does not establish strict terminal success;** (a) Representative first/last policy views. (b) Executed actions; red outlines mark seven known strict failures. (c) Recorder output, operator judgment, and terminal verification distinguish three evidential levels.

Table 14: Physical Go2 results for one fixed adapter per backbone. All counts are descriptive; the study was not designed to compare backbones.

Endpoint	Qwen2.5	Qwen3.5	Evidence
Correct first choice	24/24	23/24	recorder log
Operator-reported completion	24/24	23/24	operator judgment
Strict success	not verifiable		terminal evidence unavailable
Known strict failure	1/24	6/24	objective violation/report
Executed forward actions	72	78	controller log
Safety FWD→HALT vetoes	20	29	depth-mask log
Mean summed forward motion	1.54 m	1.63 m	odometry
Communication aborts	0	2	recorder event

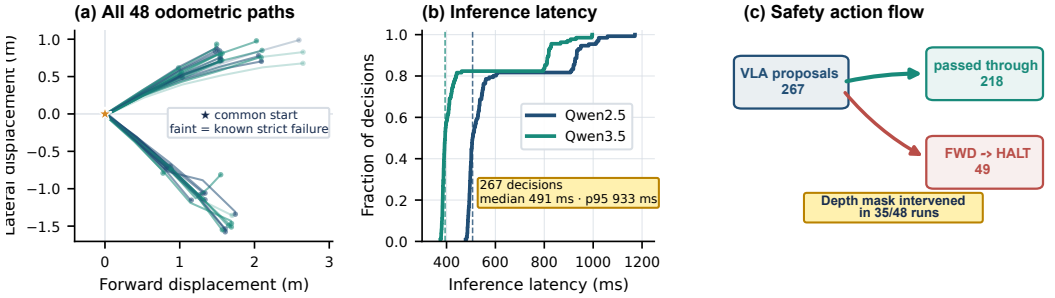

Figure 11: **The physical logs expose motion, latency, and safety mediation;** (a) All odometric paths after translating and rotating their starting poses to a common frame. (b) Empirical inference-latency distributions over 267 decisions; dashed lines mark model medians. (c) The depth mask changes 49 proposed forward actions to HALT, affecting 35/48 runs. The study therefore evaluates a VLA–controller–safety stack rather than unconstrained VLA stopping.

Table 15: Departures from the planned physical evaluation protocol. They restrict the deployment study to descriptive embodied-transfer evidence.

Element	Recorded study	Consequence
Decision horizon	six rather than eight	terminal rule differs from plan
Trial status	40 pilot; 8 unspecified	all 48 marked nonconforming
Instruction string	17/48 non-exact	wording is not fully standardized
Terminal evidence	no external video, target coordinates, distance, or adjudication	strict success is not estimable
Configuration metadata	required pretrial files absent	geometric conditions cannot be fully reconstructed
Run order	grouped rather than counterbalanced	order effects are uncontrolled
Visual review	person in one start frame; tracked robot or duplicate prop in five runs	deployment scenes are not uniformly clean
System availability	two response-timeout aborts	two known strict failures

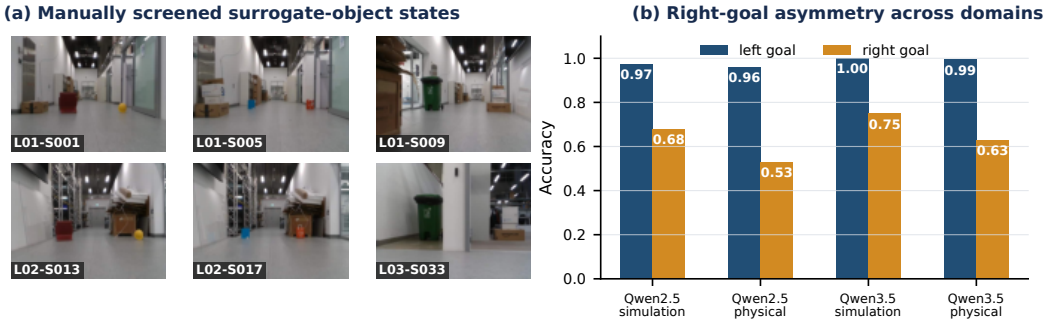

Figure 12: **Stationary physical images provide breadth across adapters;** (a) Six of 36 physical-Go2 states across three layouts and surrogate object pairs. (b) Matched first-command accuracy for the same 16 Qwen2.5 and Qwen3.5 corpus adapters in simulation and on physical images. The aggregate right-goal weakness does not imply that every selected adapter has the same asymmetry.

Table 16: Measured per-corpus cost. Rendering used one simulator GPU; adapter time is per fit.

Operation	Median	Mean	Output
Paired audit rendering	4.18 min	4.78 min	24 states / 48 queries
Training-trajectory rendering	10.61 min	11.35 min	48 trajectories
Qwen2.5 adapter fit	—	10.68 min	one learner seed
Qwen2.5 evaluation generation	—	0.78 min	48 goal-conditioned queries
Qwen3.5 adapter fit	13.08 min	13.41 min	one learner seed (121 measured fits)

Table 17: Learner and demonstrator configuration held fixed within each backbone. Qwen3.5 is a supporting replication under the same learner recipe.

Item	Configuration
Base artifacts	Qwen2.5-VL-3B-Instruct, Qwen3.5-VL-0.8B.
Input processor	One current RGB frame per decision; Qwen pixel cap 102,400; architecture-native chat template; action history serialized oldest first. No explicit token truncation beyond the base processor/model limit.
Quantization	NF4 4-bit weights, double quantization, bfloat16 compute; vision tower held fixed.
LoRA	Rank 8, alpha 16, dropout .05, no bias; language modules `q/k/v/o_proj`, `gate/up/down_proj`; all visual modules excluded and asserted absent.
Optimization	AdamW (PyTorch), learning rate 2×10^{-4}, cosine schedule, zero warmup, weight decay 0, batch size 1, gradient accumulation 8, three epochs (105–135 optimizer updates as corpus steps vary from 273–358).
Decoding	Greedy (no sampling), at most 16 new tokens; strict five-class grammar {forward, left, right, halt, done}; invalid generations count as errors.
Seeds	Learner seeds `20262001` and `20262002` cross every corpus; Python, NumPy, and all CUDA generators receive the learner seed.
Demonstrator	Tabular Q-learning with deterministic argmax; state is RGB-derived target bearing $\{-1, 0, +1\}$ and three binary depth-clearance sectors; actions are LEFT-30, RIGHT-30, and HALT; no target coordinates or branch labels are inputs.
Reward / maturity	Branch reward 2 if desired and -1 otherwise, plus 0.5 for a safe sector and -2 otherwise; checkpoints at 10, 50, 250, 1,000, and 50,000 episodes; held-out allocation 16 each and development allocation 12/11/11/11/11.
Corpus rollout	Unique teacher, layout, physics, rendering, neighborhood, and outcome seeds; at most eight decisions, 0.8 m reach radius, termination on DONE or collision/fall, timeout otherwise. Structural validity requires nonempty fresh observations, consistent histories, permitted actions, and finite translations; correctness and success are not acceptance gates. All 6,528 attempts were accepted.