EVALUATOR VALIDITY IS A PROPERTY OF THE (EVALUATOR, TASK) PAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Cheap stand-ins for expensive robot policy evaluation, whether simulators, learned world models, or vision-language judges, rest on one load-bearing argument: a high rank correlation between the cheap and the expensive ordering on a handful of policies. Careful papers check more; none of those checks replaces it. We show this evidence does not support the conclusion drawn from it, and say what to report instead. Using a deterministic ground-truth evaluator, so that every disagreement is judge error and nothing else, we score 12 (judge, task suite) pairs built from three vision-language judges and four LIBERO suites, holding the prompt, the policy family and the construction of the policy set fixed. Validity turns out to belong to the pair: within a single suite the probability of choosing the truly best policy varies by 0.835 depending only on the judge, and no judge is best on every suite. Rank correlation tracks none of it. It exceeds 0.82 in 10 of the 12 cells while the probability of picking the best policy inside that group runs from 0.212 to 1.000, and its sign changes with how the policy set was assembled and with the wording of the prompt. In the practitioner's own units, following one of these certified evaluators gives up as much as 14.7 points of true success rate, and the worst evaluator we measure is beaten by a coin. The same failure appears between two independently released checkpoints, where one judge is significantly *backwards* on a true gap of 12.6 points. Our coverage is narrow, three judges and two public checkpoints, so we map this failure rather than bound it. We give the perfect-evaluator ceiling that makes any reported correlation readable; on a published leaderboard a flawless evaluator would score only $\rho = 0.500$ on LIBERO-Long at the budget the field uses.

1 INTRODUCTION

Whenever a measurement is expensive, the field that depends on it eventually replaces it with a cheap stand-in, and then needs an argument that the stand-in can be trusted. The argument is almost always the same one. Run a handful of systems through both the expensive procedure and the cheap one, correlate the two rankings, and report that the correlation is high. Large language models scoring other models' outputs are validated this way against human preference (Zheng et al., 2023), and so are the substitutes that have grown up around robot policy evaluation, where the expensive procedure costs hundreds of physical rollouts: a simulator standing in for the real robot, a learned video model standing in for the simulator, or a vision-language model standing in for the success predicate (Li et al., 2024; Jeon et al., 2026; Wang et al., 2026). None of these papers is naive about what a rank statistic can carry, and Section 7 records what each already says and already checks.

This paper asks what that argument establishes, and answers it in a setting where the question can be asked cleanly. Work on model judges usually compares the judge against human annotation, which is itself noisy, so a disagreement has to be apportioned between two imperfect measurements. Robot policy evaluation has a property that setting lacks: a simulator's success predicate is *deterministic*. Re-running an identical configuration of 500 episodes end to end flips 0 outcomes. Every disagreement between the cheap evaluator and the ground truth is therefore attributable to the cheap evaluator and to nothing else, which is what lets us measure how far the usual argument can be trusted rather than argue about it.

Two findings follow, one about the thing being certified and one about the statistic that certifies it.

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

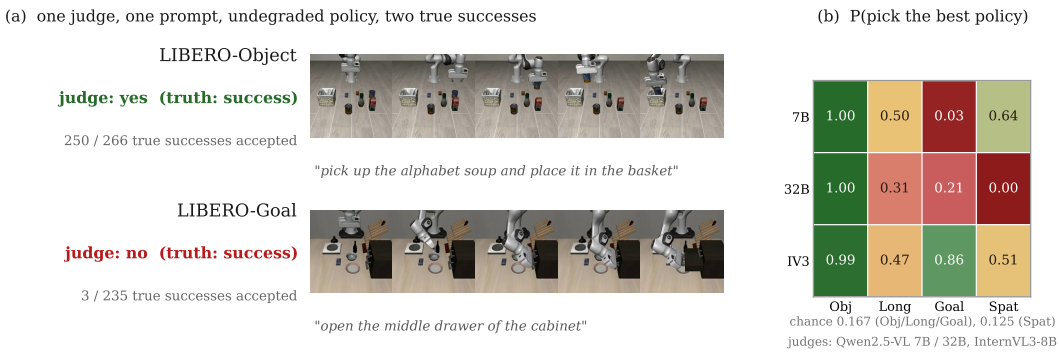


Figure 1: Evaluator validity is a property of the (evaluator, task) pair. *Left:* two episodes of the same *undegraded* policy, scored by one judge (Qwen2.5-VL 7B) under one prompt; the frames shown are the judge's actual input. Both are ground-truth successes, yet the judge accepts the LIBERO-Object episode and rejects the LIBERO-Goal one. Over that policy it recognises 250 of 266 true successes on Object and 3 of 235 on Goal. *Right:* the probability of picking the truly best policy, for three judge families across four suites. **No judge is uniformly best**: Qwen-7B is the strongest of the three on Spatial and the weakest on Goal. Rank correlation tracks none of this, reaching at least 0.82 in 10 of the 12 cells while $P_{\text{top-1}}$ inside that group runs from 0.212 to 1.000. Details and per-cell counts are in Table 1.

Validity does not attach to the evaluator. Across three judges and four task suites, holding the prompt, the policy family and the construction of the policy set fixed, the probability of choosing the best policy varies by 0.835 inside a single suite depending only on which judge is used, and no judge is best everywhere. Figure 1 shows the extreme case: one judge, one prompt, two true successes, accepted on one suite and rejected on the other. Tasks do differ in difficulty, but difficulty interacts with the judge, so validity belongs to the pair and cannot be inferred from either component alone. A judge validated on somebody else's benchmark carries no guarantee on yours.

Rank correlation tracks none of it. Over those same 12 cells, the correlation is at least 0.82 in 10 of them, and inside that group the probability of picking the best policy runs from 0.212, barely above chance, to 1.000. The statistic meant to certify the evaluator takes nearly the same value whether the evaluator is excellent or useless. Put in the units a practitioner actually cares about, following one of these certified evaluators costs up to 14.7 points of true success rate, and the worst evaluator we measure is beaten by a coin.

Contributions. We measure 12 (judge, suite) pairs against a deterministic ground truth and show that validity belongs to the pair, with the ordering of judges reversing across suites (Section 4.3) and the same reversal appearing between two independently released checkpoints (Section 4.4). We show the reported statistic says nothing about the decision it is used for, since its sign turns over with the construction of the policy set (Section 4.2) and with the wording of the prompt. We give the perfect-evaluator ceiling, transplanted from the classical correction for attenuation, without which no reported correlation can be read, and measure it on a published leaderboard (Section 3). And we give a reporting protocol whose verdict has two cells, one for choosing a policy and one for reporting its number, with a tool that computes it from a generic table of judgements (Section 5).

Every experiment here would pass the criterion we are criticising. Our strongest correlations, $|\rho|$ above 0.9, sit on settings whose $P_{\text{top-1}}$ runs from 0.000 to 0.866. That is the point. The criterion is easy to satisfy, which is why satisfying it certifies so little.

2 What a rank correlation can and cannot certify

Read this section as bookkeeping, not as a contribution. Proposition 1 is the definition of a rank statistic restated and Proposition 2 is a textbook best-arm bound; we say so before a referee does. Their role is to turn the measurements of Section 4 into instances of something already known, and to name the quantity the field should have been reporting. Throughout, let $p_1, \ldots, p_k$ be the true success rates of k policies under a ground-truth evaluator and $q_1, \ldots, q_k$ the rates a surrogate reports

for the same policies, each estimated from n episodes, and write ρ for Spearman's rank correlation between p and q.

2.1 Rank correlation is blind to magnitude

Proposition 1 (Monotone invariance)**.** *Let g be strictly increasing on $[0, 1]$ and let a surrogate report $q_i = g(p_i)$ exactly. Then $\rho = 1$ regardless of how far g departs from the identity. Consequently ρ places no upper bound on calibration error and none on sensitivity.*

The proof is immediate: g strictly increasing preserves ranks, and ρ depends on p and q only through ranks. The content is the corollary. An evaluator can be perfect by the criterion the field reports while being arbitrarily wrong about every number it produces. This is not a hypothetical family: on 6 distortions we constructed, ρ stays at 1.000 while calibration error reaches 0.326 and sensitivity falls to 0.080, and Appendix A shows that the two extremes of that family sit close to the measured profiles of the 32B and 7B judges we go on to study.

2.2 When the signal is small, the sign is arbitrary

Write the surrogate's rate as

$$q_i \;=\; a\,p_i + b + \varepsilon_i, \tag{1}$$

where a is the sensitivity, b an offset, and ε_i collects everything else the evaluator responds to. Because ρ is invariant to the scale of q, it does not compare $|a|\,\mathrm{span}(p)$ against $\mathrm{spread}(\varepsilon)$; it simply ranks. When the first is small relative to the second, the ordering of q, and therefore the sign of ρ, is determined by ε rather than by a, and nothing in the reported value of ρ reveals which regime produced it.

This makes a the diagnostic the field is missing, and it is estimable from data already collected. Two facts constrain how it may be used, and Appendix B gives both with intervals. It is not a property of the evaluator alone, since holding the 32B judge and the prompt fixed and changing only how the policy set was built puts a entirely below zero on one ladder and entirely above it on the other two. Nor is it a property of the task alone, the subject of Section 4.3, since across our four suites the ranges of $|a|$ for the 7B and 32B judges overlap heavily. Only the pair pins it down.

Both facts are post hoc, so we also wrote the reading rule and its predictions to a timestamped file before looking at the ladders it had not seen. The rule follows from Equation 1 and nothing in it was fitted: the sign of ρ is trustworthy when the interval for a excludes zero, and the full ordering is trustworthy when $|a|\,\Delta_{\min} > 1.96\sqrt{2}\,s_\varepsilon$. Two of the three registered predictions held and one failed, and it failed because we wrote an undiscriminating prediction inside the regime our own rule calls stable. Appendix C gives the registration, the outcomes and that diagnosis in full.

Remark 1 (A trap we fell into)**.** When bootstrapping a, the resampling unit must be the episode, not the ladder level. Resampling levels treats six or eight points as the sample and inverts the conclusion. Our first pass had the sign backwards for this reason.

2.3 How many rollouts a comparison actually needs

Proposition 2 (Sufficient sample size)**.** *Let $\Delta_i = p_{(1)} - p_{(i)}$ for $i \neq 1$. Under a ground truth evaluator with independent episodes, $\Pr[\text{best policy not identified}] \leq \sum_{i\neq 1} \exp(-n\Delta_i^2)$, so $n \geq \ln((k-1)/\delta)\,/\,\Delta_{\min}^2$ suffices for confidence $1-\delta$.*

This is Hoeffding plus a union bound and we claim no novelty for it. We include it because it converts an observed leaderboard gap into a rollout budget, which Section 3.1 uses to show that published budgets are short by factors of about two and two hundred. Appendix D records two cautions on quoting it, including how conservative it is against Monte-Carlo simulation of the same setting.

3 How high could ρ have been?

A reported $\rho = 0.970$ invites the reading that the evaluator is almost perfect. That reading needs a reference point the literature does not supply: the value a *perfect* evaluator would obtain on the same

policy set at the same budget. Both sides of the comparison are finite samples, so even an evaluator that measures exactly the right quantity ranks noisily. Fix true rates $p_1, \ldots, p_k$ and a per-policy budget n; a perfect evaluator draws its own independent $\mathrm{Binomial}(n, p_i)$ counts, and the ceiling is the sampling distribution of ρ between the two draws. It depends only on the policy set and the budget, never on the evaluator, and no real evaluator can beat it except by luck. Appendix E illustrates the dependence on a controlled ladder, where holding the evaluator perfect and only clustering the policies drops the ceiling from 1.000 to 0.738. Most of the variation in a bare ρ is attributable to how far apart the policies were.

Remark 2 (This is a transplant, not a new statistic)**.** In psychometrics this is attenuation due to measurement error, and the correction for attenuation goes back to Spearman (1904). Our contribution is not the estimator; it is the observation that a literature which reports ρ as evidence of evaluator quality never computes it, and that without it the reported numbers cannot be read at all.

3.1 The same line, measured on a real leaderboard

The construction above uses our own ladder, which invites the objection that we built the difficult regime ourselves. So we computed the same ceiling from numbers we did not produce: the 3 methods reported side by side in the OpenVLA release, each averaged over three seeds of $n = 500$ rollouts (Kim et al., 2024). On LIBERO-Long those methods lie within 0.032 of one another, and a perfect evaluator at the field-standard budget would obtain $\rho = 0.500$ $[-1.000, 1.000]$; tripling the budget to the full three-seed total leaves it at 0.500. Any ρ reported on that leaderboard is being read off a statistic whose noiseless value is a coin flip. Appendix F gives the ceiling for all four suites together with the sample size Proposition 2 requires, and works through the two sample-size questions, which are easy to conflate and which we conflated ourselves at first. Both answers exceed the 500 rollouts the field reports, by factors of about two and two hundred.

This answers half of the objection that we manufactured a low-signal regime. Real leaderboards are already inside it. It is only half, because the mechanism differs: here the policies are too close together, whereas in Section 4.3 the judge carries too little signal. The consequence is the same, that ρ can be large and uninformative at once, but the two causes should not be written as though they were one.

4 Experiments

Our setting isolates one link in the chain. A learned world model that scores policies makes two kinds of error at once: its dynamics drift, and its scoring head is wrong. We remove the first by running the policies in the real simulator, keeping the recorded video, and handing that video to a vision-language judge. Whatever the judge and the simulator then disagree about is judge error and nothing else.

4.1 A trustworthy ruler and a controllable ladder

LIBERO's own success predicate is our ground truth. Re-running an identical configuration of 500 episodes end to end produced 0 flipped outcomes, leaving only the finite-sample term, about ± 1.6 points at $n = 500$, so no disagreement ever has to be split between two noisy measurements. The policy is a fine-tuned OpenVLA (Kim et al., 2024), and no suite entered the study until it had reproduced the published number for that suite, with $|z| > 2$ aborting the run. All 4 passed, combined $z = -0.81$ (Appendix G).

We need policies whose true ranking we know exactly. Public checkpoints have spacing we cannot control, so instead we degrade one policy along three axes: Gaussian noise on the action, a resolution bottleneck on the observation, and action repetition. The noise axis yields 8 strictly ordered levels from 85.6% down to 0.2% at $n = 500$ each. Two properties carry weight later: the undegraded level reproduces the official number episode for episode, and the three axes can be trimmed to a common true span, which lets us change how the policy set was built while holding its spread fixed. Section 6 says what this construction costs us. Our judges are Qwen2.5-VL at 7B and 32B (Bai et al., 2025) and InternVL3-8B (Zhu et al., 2025); each sees nine frames sampled from the recorded episode plus the task instruction, decodes greedily, and answers a binary question. Over 29,800 paired judgements under this prompt they flipped their own verdict 0 times and returned 0 unparseable answers, so judge-side run-to-run noise is not a confound below.

Table 1: $P_{\text{top-1}}$ with ρ underneath, three judges across four suites. Object, Long and Goal each aggregate 1,800 episodes over 6 strictly ordered levels, so chance is 0.167; Spatial aggregates 4,000 over 8 levels, so its chance is 0.125, which puts its $P_{\text{top-1}}$ on a different scale from the other three columns. The perfect evaluator's ceiling is 1.000 in every cell, so any ρ below one is charged to the judge rather than to the design. Read down a column and the task looks easy or hard; read across a row and the within-column ranking changes.

Judge	LIBERO-Object	LIBERO-Long	LIBERO-Goal	LIBERO-Spatial
Qwen2.5-VL 7B	0.997 $\rho = 1.000$	0.498 $\rho = 0.986$	0.029 $\rho = -0.247$	0.643 $\rho = 0.976$
Qwen2.5-VL 32B	1.000 $\rho = 1.000$	0.306 $\rho = 0.943$	0.212 $\rho = 0.829$	0.000 $\rho = -0.970$
InternVL3-8B	0.986 $\rho = 1.000$	0.468 $\rho = 0.943$	0.865 $\rho = 0.986$	0.512 $\rho = 0.976$
spread within suite	0.014	0.192	0.835	0.643

4.2 The sign is a property of the ladder, and not of it alone

Were ρ measuring the judge, its sign would survive a change in how the policy set was assembled. It does not. Holding the judge (32B), the prompt and the true span fixed, the three ladders differing in span by at most 0.021, the noise ladder gives $\rho = -0.943$ $[-1.000, -0.314]$, the observation ladder 0.600 $[0.086, 0.943]$ and the action-repeat ladder 0.696 $[0.029, 0.943]$. No interval covers zero, so the sign difference is not sampling noise. Over the same judgements ρ swings by 1.639 while $P_{\text{top-1}}$ stays at or below 0.111 on all three. Nor does a high $|\rho|$ rescue the decision: the 7B judge reaches 1.000, 0.943 and 1.000 on those same ladders, at or near the perfect evaluator's ceiling, yet its $P_{\text{top-1}}$ is 0.866, 0.336 and 0.621. The raw rates show the mechanism: the 32B judge's accept rate is nearly flat across levels, so whichever policy happens to sit highest on that flat line fixes the whole ordering, and rank correlation turns order noise into what looks like strong agreement.

Running the same three ladders on a second suite, this time with all three judges, shows the sign is not a property of the ladder alone either. The ladders are the same three constructions, so whatever changes has to come from the other side of the pair. On Goal the swing in ρ sorts itself by sensitivity, and the reversal is the point (Appendix H has the table).

The 7B judge, whose sign was the stable one above, is here the one that turns over, from $\rho = -0.247$ on the noise ladder to 0.618 on the observation ladder, with $|a|$ never above 0.013 and significant on only 1 of 3 ladders. InternVL3, whose $|a|$ runs from 0.221 to 0.398 and is significant on all 3, stays between 0.943 and 1.000. Which of these happens is predicted by neither the judge nor the ladder, but by the sensitivity measured on those same episodes. Since the perfect evaluator's ceiling is 1.000 in all nine cells, none of this is a ceiling artifact.

Two further controls are reported in full in Appendices I and J: the sign also turns over across four wordings of the prompt for the 32B judge while the 7B judge stays stable, and the failure is not a visibility problem, since holding ground truth fixed leaves nothing in the image that separates the 32B judge's accepted episodes from its rejected ones.

4.3 Main result: validity belongs to the pair

Table 1 is the paper's central measurement: four suites, three judges, one prompt, one policy family, one ladder construction, with only the pair changing. The starkest cell is LIBERO-Spatial, where the two Qwen judges score the same footage under the same prompt and reach $\rho = 0.976$ and -0.970. Those two magnitudes are indistinguishable, one ranking is right and the other exactly reversed, and the decision follows the sign rather than the magnitude: $P_{\text{top-1}}$ is 0.643 against 0.000, so the 32B judge almost never picks the best policy under resampling. The effect size explains how a rank statistic gets into this position at all. Ground truth spans 0.854 on that ladder while the 32B judge's own accept rate moves 0.070, an order of magnitude less than the thing it is tracking.

Tasks do differ in difficulty. On LIBERO-Object all three judges come close to perfect and the spread across judges is 0.014. Were difficulty the whole story the spread would stay that small everywhere, and it does not. On LIBERO-Goal it is 0.835, from 0.029 for Qwen-7B, below chance, to 0.865 for InternVL3.

Table 2: Two real policies, three judges. Of the 6 (judge, suite) cells only 2 order the pair correctly at $|z| > 1.96$, 3 cannot separate them, and 1 is significantly *reversed*; correcting for 6 tests leaves 1. On LIBERO-Goal the OpenVLA side is $n = 300$ and the OFT side $n = 500$; elsewhere both are $n = 500$.

Suite	Judge	OpenVLA	OFT	z
Spatial	*ground truth*	85.6	98.2	$+7.50$
	Qwen2.5-VL 7B	36.2	39.0	$+0.91$
	Qwen2.5-VL 32B	14.2	3.0	-6.45
	InternVL3-8B	45.0	45.4	$+0.13$
Goal	*ground truth*	77.4	98.6	$+10.91$
	Qwen2.5-VL 7B	1.0	12.8	$+7.37$
	Qwen2.5-VL 32B	20.7	22.6	$+0.65$
	InternVL3-8B	67.7	75.8	$+2.46$

Three consequences follow, none of them readable from a single number. First, **no judge is uniformly best**. Qwen-7B is the strongest of the three on Spatial and the weakest on Goal, so validating a judge on one suite and carrying it to another buys no protection. This is the sense in which validity belongs to the pair: not recoverable from the task alone, because the judge reorders it, and not from the judge alone, because the task reorders it. Second, **rank correlation tracks none of it**. Across the 12 cells $\rho \geq 0.82$ in 10, and inside that group $P_{\text{top-1}}$ runs from 0.212 to 1.000, barely above chance at one end and essentially perfect at the other; the remaining 2 cells have negative ρ. Third, **even the direction of the bias is not a judge property**. InternVL3 over-reports on 3 of its 4 cells while both Qwen judges under-report on 3 of theirs, and on LIBERO-Goal the signed calibration of the three is 0.131, -0.446 and -0.266, so anyone correcting for a judge's known optimism on one suite would apply that correction backwards on another.

Stated as a cost rather than as a statistic, Table 1 reads as follows. Following a judge's choice instead of the truth gives up 0.0 to 14.7 points of true success rate among the 10 cells that pass the field's own criterion of $\rho \geq 0.82$. In the worst cell it gives up 73.7 points, against 40.6 for choosing at random on that suite, so the practitioner would have done better with a coin.

Figure 3 in Appendix K shows why ρ is blind here. Plot each judge's accept rate against ground truth level by level and the slope is the sensitivity a of Equation 1: on Object all three slopes sit near the diagonal, on Goal the 7B trace is flat at the floor ($a = 0.001$), and on Spatial the 32B trace descends ($a = -0.072$, CI $[-0.112, -0.032]$). Rank correlation is invariant to exactly this axis.

4.4 The same thing happens on two real, independently released policies

Everything above uses a policy set we built by degrading one policy, which invites the objection that we manufactured the low-signal regime. Here we remove the part that can be removed; Section 6 keeps the rest. We take two policies we did not make: the fine-tuned OpenVLA checkpoint used throughout (Kim et al., 2024), and OpenVLA-OFT (Kim et al., 2025), which differs structurally by decoding an action chunk in parallel and regressing continuous actions instead of emitting discrete tokens. Each passed the same reproduction gate against its own published per-suite number. Their true gap is not subtle (Table 2): 12.6 points on LIBERO-Spatial ($z = +7.50$) and 21.2 on LIBERO-Goal ($z = +10.91$), against the $n \geq 189$ and $n \geq 67$ episodes Proposition 2 asks for. We ran 500.

On LIBERO-Spatial no judge separates the two policies meaningfully, and the 32B judge is significantly *backwards*: it reports OFT as -11.2 points worse ($z = -6.45$) when OFT is in fact 12.6 points better. A coin flip landing badly would not look like this. It is a confident statement in the wrong direction, on two published checkpoints, at a true separation of $z = +7.50$. We do not overstate it, because on LIBERO-Goal two of the three judges order the pair correctly and significantly. What the two suites together show is again the pattern this paper is about. Read down a suite and Spatial defeats all three judges; read across a judge and the same model swings between failing and succeeding with the suite. One observation survives the sign question. The effect size collapses even where the direction is right: against a true gap of 21.2 points on Goal, the three judges move $+11.8$, $+1.9$ and $+8.1$ points, sensitivities 0.56, 0.09 and 0.38, so reading those as success rates would badly understate the improvement. This pair is also the $k = 2$ case that Section 5.2 builds on.

5 WHAT TO REPORT INSTEAD

Nothing above argues that rank correlation is meaningless. It argues that a single ρ, reported without a ceiling and without an effect size, cannot support the conclusion it is routinely asked to support. Five quantities should be reported together, and none is recoverable from the others; Table 5 reports all five for every cell of Table 1.

- **Sensitivity** a, the slope of the evaluator's rate against the true rate: does the evaluator move as far as reality moves? Proposition 1 says ρ cannot see it, since $a = 1$ and $a = 0.01$ both give $\rho = 1$. It also carries the sign information ρ garbles, because a negative slope whose interval excludes zero is an unambiguous statement that the evaluator is anti-correlated with reality in magnitude and not merely in rank.
- **Calibration**, absolute as well as signed. Signed calibration is actively deceptive when the evaluator errs in both directions: on LIBERO-Long the 32B judge has signed calibration 0.003, which reads as almost unbiased, while its absolute calibration is 0.131 and its sensitivity 0.288. The errors cancel in the average and not in any individual number a practitioner would quote.
- ρ **next to its own ceiling** for your k and n (Section 3). Without it the number is uninterpretable, and comparing ρ across suites with different k or n compares two different scales.
- $P_{\text{top-1}}$, the probability of picking the best policy. This is what a practitioner actually consumes and it separates cases ρ merges. It is a bootstrap estimate whose Monte-Carlo error across the 12 cells of Table 1 runs from 0.000 to 0.008, so the third decimal should not be read.
- **Regret**, the true success rate given up by following the evaluator's choice instead of the truth. It is $P_{\text{top-1}}$ in the practitioner's own units, measuring the cost of a wrong choice rather than its frequency. Its reference line is the regret of choosing at random on the same policy set, and an evaluator can be worse than that line (Section 4.3).

To these we add the sample size implied by Proposition 2, which converts an observed gap between the top two policies into the number of rollouts needed to separate them.

5.1 THE VERDICT HAS TWO CELLS, NOT ONE

Splitting the conclusion is not a way of softening it. It is what Proposition 1 forces: ranking validity and reading validity are separable properties, so collapsing them into one boolean is exactly the move this paper argues against. Cell **A**, *can I use it to choose between policies?*, is governed by $P_{\text{top-1}}$ against chance and by the sign and interval of a. Cell **B**, *can I report its numbers as success rates?*, is governed by absolute calibration and by whether a is near one.

The two come apart in our measurements. Under the occlusion-aware prompt the 7B judge reaches $P_{\text{top-1}} = 0.834$ against chance 0.125 with $\rho = 1.000$ sitting at the ceiling, and simultaneously underreports by 0.129. Cell A passes and cell B fails; one number for this judge would have to be wrong in one of the two directions. Any thresholds are conventions and we say so: we use $P_{\text{top-1}} > 0.80$, absolute calibration below 0.05, and $a \in [0.80, 1.25]$, and we always print the continuous values next to the verdict so a reader who prefers different cut-offs can apply them.

5.2 THESE QUANTITIES SURVIVE WHERE ρ DOES NOT

A referee may reasonably ask what happens on real leaderboards rather than constructed ladders. The answer is uncomfortable for ρ and comfortable for everything else. At $k = 2$, the case of Section 4.4, Spearman's ρ can only be ± 1 and so separates nothing, while sensitivity, calibration, $P_{\text{top-1}}$, regret and the sample-size bound all stay informative; Appendix L says what each becomes. The regime where the field most wants an answer is precisely the regime where the statistic it reports is fixed in advance by the number of policies.

`surrocheck`, included in the supplementary material, is a dependency-light utility that computes all of the above, plus the two-cell verdict, from a generic table of `(policy_id, ground_truth,`

`evaluator_pred`) rows. It imports none of our experiment code and knows nothing about robotics. Appendix M says what it checks and where the two implementations agree only within error.

6 LIMITATIONS

Most of our policies are degraded copies of one policy. We build 10 of the policy sets here by degrading a single checkpoint rather than by collecting independent methods, and for the ladder experiments the objection that we manufactured the low-signal regime stands. It no longer stands for the paper's conclusion: Section 4.4 repeats the comparison on two policies we did not build, each reproducing its own published per-suite number and separated by 12.6 and 21.2 points, and of 6 cells only 2 order the pair correctly while 1 is significantly reversed. Independently, the published LIBERO leaderboard already sits inside the low-signal region (Section 3.1). What remains genuinely limited is coverage rather than validity: our real pair is $k = 2$ from one architecture lineage, so it demonstrates the failure without mapping its extent. A larger set of structurally distinct public checkpoints would let one ask how often rather than whether this happens, and is the most valuable next experiment.

We can diagnose after the fact and cannot predict beforehand. Two pre-registered attempts to predict from the task alone which pairs would fail were both refuted (Appendix N), and the reason is structural, not a missing feature: changing the judge moves sensitivity further than changing the task category does (0.421 against 0.309), so a rule whose inputs are task properties is trying to predict something that is not a function of those inputs. The comparison also says what a predictor would need as input: the judge's own measured behaviour, not the task's properties. Our recommendation is consequently a measurement protocol and not a screening rule, and we can quote its price. Appendix D gives it in episodes per policy: the smallest true gap a perfect evaluator reverses less than five percent of the time is 11.5 points at $n = 100$ and 5.5 at $n = 500$. Anyone who wants to know whether a judge works on their task has to run it on their task, and can budget for it first.

One simulator family, one policy architecture, three judges. All four suites come from LIBERO, so rendering, physics and success predicates share a lineage and a different simulator could behave differently. The base policy is a single architecture; we added a third judge instead, having judged the dependency risk high and the marginal evidence low. Two judge families and two model sizes are enough to show the effect is not a Qwen artefact, and not enough to describe how judging quality scales with judge capability.

We measure a judge, not a world model. Isolating the judge is what makes every disagreement attributable, and it is also the boundary of the result. A learned world model used as an evaluator compounds dynamics error with judging error and we have characterised only the second. We expect the compound case to be worse rather than better, but we have not measured it and do not claim it. Appendix O records three statistical caveats we hold ourselves to.

7 RELATED WORK

Cheap stand-ins for expensive evaluation. Hardware evaluation is slow, so the field has built substitutes. SimplerEnv put simulation-based evaluation of real-world policies on a systematic footing and introduced a rank-based discrepancy measure (Li et al., 2024). Learned video models have since been proposed in place of hand-built simulators: RoboWorld pairs an autoregressive world model with a progress-aware scorer and reports agreement with a real-robot leaderboard (Jeon et al., 2026); an interactive world simulator reports strong sim-to-real correlation (Wang et al., 2026); SC3-Eval adapts a video foundation model into a policy evaluator through self-consistency objectives (Tseng et al., 2026); and Genie Envisioner folds a neural simulator and a benchmark into one platform (Liao et al., 2025). Hand-built simulation has moved the same way, with REALM designed so that simulated scores correlate with real ones (Sedlacek et al., 2026). World models built for other purposes (Shang et al., 2025; Gao et al., 2026) are context here rather than instances of the pattern.

The concern is voiced already; the number is not. The field is not unaware. SimplerEnv states that a rank statistic discards the margins between policies and can swing on differences attributable to

evaluation noise, and its Mean Maximum Rank Violation weights each violation by the real margin to address exactly that (Li et al., 2024). The interactive world simulator reports Clopper-Pearson intervals on its per-policy rates and notes that its own fit is optimistic (Wang et al., 2026). RoboWorld rescores with a second frontier model and finds the ordering largely preserved (Jeon et al., 2026). Three things none of these checks supplies. Nobody computes what a *perfect* evaluator would have scored on the same policy set at the same budget, without which a reported correlation cannot be read. The intervals that are reported attach to the individual success rates rather than to the correlation carrying the argument. And a two-judge check on a single task distribution is precisely the evidence our results say is insufficient, since reading down one column of Table 1 would also suggest the judges agree, and it is reading across the rows that reverses them. One further RoboWorld detail is the cleanest existing evidence that the scoring convention alone moves the statistic: switching the rubric from graded progress to binary success takes ρ from 0.970 to 0.922, with no uncertainty attached to the difference. Their pipeline also compounds a world model with a judge, so neither error source is separately identified, whereas ours holds the simulator fixed and varies only the judge.

Statistically grounded evaluation. Our closest methodological relative is SureSim (Badithela et al., 2025), which casts a large imperfect simulation plus a small real evaluation as prediction-powered inference (Angelopoulos et al., 2023) and returns non-asymptotic intervals on true policy performance. That is the right way to *use* an imperfect evaluator once paired real data exists; we address the prior question of whether one is usable at all, where such data does not exist and the evaluator is a neural judge. Our sample-size statement is the standard fixed-budget best-arm argument, Hoeffding (Hoeffding, 1963) with a union bound, whose sharp characterisation is well known (Kaufmann et al., 2016); we claim no novelty for it.

Models as judges. Success detection as visual question answering for a pretrained vision-language model was proposed by Du et al. (2023) and is now standard in neural-simulator pipelines and in reward labelling. Judge reliability has been studied directly in the language domain (Zheng et al., 2023), and miscalibration of confident neural predictors is a long-standing finding (Guo et al., 2017). What has not been done, to our knowledge, is to isolate the judge against a *deterministic* ground truth. Those literatures compare judges with human annotations, themselves noisy, so a disagreement must be split between two imperfect measurements. Here it need not be, which is what lets us say that a larger model from the same family can be much worse (Section 4.3) and that a judge can respond strongly to the visual evidence and still be badly calibrated (Appendix J).

Attenuation, and auditing the benchmark itself. The ceiling of Section 3 is attenuation due to measurement error, whose correction dates to Spearman (1904); we transplant it rather than invent it. A parallel line audits the ground truth rather than the surrogate, showing that LIBERO (Liu et al., 2023) scores are sensitive to perturbations that should not matter (Fei et al., 2025) and proposing cheaper or more diagnostic real-world benchmarks (Huang et al., 2026; Wu et al., 2026; Yu & Qiu, 2026). We instead take the ground truth as given and exploit its determinism. The two concerns compose: a surrogate validated against a fragile ground truth inherits both problems.

8 Conclusion

The question a practitioner needs answered is not whether vision-language judges or neural simulators work in general, but whether the specific evaluator they are about to use works on the specific tasks they are about to run. Our measurements say this question has no shortcut. The answer is not a property of the evaluator, so a judge validated elsewhere carries no guarantee; nor of the task, so a suite other judges handled well carries none either; and the statistic the field reports cannot tell an evaluator that will choose correctly from one that will choose at chance.

What is left is unglamorous and cheap. Run the evaluator on your own tasks against whatever ground truth you have, and report how far it moves when reality moves, how far its absolute numbers are off, the rank correlation next to the value a perfect evaluator would reach at your k and n, and what following it costs you in true success rate. Then answer separately whether it can be used to choose and whether its numbers can be reported, because those can have different answers. None of these quantities is new. What our results argue is that reporting one of them, chosen because it is the one that looks best, is not evidence of validity.

AI USE STATEMENT

We used generative AI tools throughout this work, in the form of an AI coding assistant used interactively by the authors, and we disclose that use here in full. What this paper claims, which results it reports, what would have counted as falsification, and how the argument is organised were decided by the authors, and everything the assistant produced was checked before it was used.

The authors designed the experiments; the assistant proposed candidate designs and drafted the protocols, which the authors then accepted, altered or rejected. The same division holds for the readings of the results, for the wording of the pre-registered predictions reported in Appendix C, and for how the propositions of Section 2 are stated: the assistant drafted and the authors adjudicated. The assistant also wrote code, namely the evaluation harness, the analysis scripts, the independent second implementation used to cross-check every statistic, the scripts that draw every figure in this paper, and the build tooling. It drafted and edited the manuscript prose. The tool released with this paper was written first in the authors' working language and translated into English by the assistant, and the authors read the translation before release. Part of the related work was surfaced by AI-assisted search and the rest identified by the authors directly.

The checking was not a formality. The code was checked by running it, by reproducing published baseline numbers under a gate declared in advance that was able to abort the study, and by implementing every statistic twice. The numbers were checked by generating each one from result files with a script, so that the document does not build if a number appears that the script did not produce. Each cited paper was read by the authors before being described, so the related work section reflects those readings and not any summary the assistant produced.

We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, its claims and its released artifacts. Separately, and not as a form of AI assistance: the vision-language models studied in this paper are the object of the investigation rather than tools used to produce it, and their role is described in Section 4.

REPRODUCIBILITY STATEMENT

Every number in this paper is generated from result files by a script rather than transcribed, and the text will not build if a number appears that the script did not produce. The five quantities Section 5 asks for are tabulated for every cell in Appendix P. The ground-truth evaluator, the policy ladders, the judges and the prompts are specified in Section 4.1 and Appendix I. The reproduction gates each suite had to pass before entering the study, together with their published reference numbers, are in Appendix G. Three pre-registrations are reported: the two task-level attempts we report as refuted, in Appendix N, and the sign criterion of Section 2.2, two of whose predictions held and one of which failed. All were written down before the data they concern existed. The statistics are computed twice, once in the analysis pipeline and once in an independent utility that we include in the supplementary material, and the two agree on every quantity reported here except the bootstrap estimate of $P_{\text{top-1}}$, where they agree within Monte-Carlo error.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. arXiv:2301.09633.

Apurva Badithela, David Snyder, Lihan Zha, et al. Reliable and scalable robot policy evaluation with imperfect simulators. *arXiv preprint arXiv:2510.04354*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yuqing Du, Ksenia Konyushkova, Misha Denil, et al. Vision-language models as success detectors. In *Conference on Lifelong Learning Agents*, 2023. arXiv:2303.07280.

Senyu Fei, Siyin Wang, Junhao Shi, et al. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, et al. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. arXiv:1706.04599.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Alex S. Huang, Jiahui Zhang, Shiqing Tang, and Yu Xiang. VLA-REPLICA: A low-cost, reproducible benchmark for real-world evaluation of vision-language-action models. *arXiv preprint arXiv:2605.20774*, 2026.

Byeongguk Jeon, Seonghyeon Ye, JaeHyeok Doo, et al. RoboWorld: Fast and reliable neural simulators for generalist robot policy evaluation. *arXiv preprint arXiv:2607.01060*, 2026.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016. arXiv:1407.4443.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024. arXiv:2406.09246.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems*, 2025. arXiv:2502.19645.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, 2024. arXiv:2405.05941.

Yue Liao, Pengfei Zhou, Siyuan Huang, et al. Genie Envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023. arXiv:2306.03310.

Martin Sedlacek, Pavlo Yefanov, Georgy Ponimatkin, et al. REALM: A real-to-sim validated benchmark for generalization in robotic manipulation. *IEEE Robotics and Automation Letters*, 11 (7):8315–8322, 2026. arXiv:2512.19562.

Yu Shang, Xin Zhang, Yinzhou Tang, et al. RoboScape: Physics-informed embodied world model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 71035–71059, 2025. arXiv:2506.23135.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, et al. SC3-Eval: Evaluating robot foundation models via self-consistent video generation. *arXiv preprint arXiv:2606.18610*, 2026.

Yixuan Wang, Rhythm Syed, Fangyu Wu, et al. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

Zenan Wu, Bingqing Wei, Lu Liu, et al. ATOM-Bench: A real-world benchmark for atomic skills and compositional generalization in manipulation policies. *arXiv preprint arXiv:2606.16826*, 2026.

Yi Yu and Xinchuan Qiu. Benchmarking vision-language-action models on SO-101: Failure and recovery analysis. *arXiv preprint arXiv:2606.08881*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. arXiv:2306.05685.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A NUMERICAL VERIFICATION OF PROPOSITION 1

We verified this numerically on 6 distortions, ranging from a pure compression to a logistic squeeze. In every case $\rho = 1.000$, while calibration error reaches 0.326 and sensitivity falls to 0.080.

The construction is not hypothetical. The worst distortion we wrote down, a compression to sensitivity 0.080, is close to the measured profile of the 32B judge, whose sensitivity on our ladders runs from -0.072 to 0.089; the milder compression at sensitivity 0.350, with calibration error 0.293, is close to the 7B judge. We did not stumble on an unusually bad evaluator. We measured evaluators that sit inside the family the proposition constructs.

B SENSITIVITY IS A PROPERTY OF NEITHER SIDE ALONE

This makes a the diagnostic the field is missing, and it is estimable from data already collected. Two facts constrain how it may be used. It is not a property of the evaluator alone: holding the 32B judge and the prompt fixed and changing only how the policy set was built gives $a = -0.072$ $[-0.112, -0.032]$ on the noise ladder against 0.089 $[0.037, 0.142]$ and 0.082 $[0.032, 0.133]$ on the other two, the first interval lying entirely below zero and the other two entirely above it. Nor is it a property of the task alone, which is the subject of Section 4.3: across our four suites $|a|$ runs from 0.001 to 0.861 for the 7B judge and from 0.072 to 0.807 for the 32B judge, and the two ranges overlap heavily. Only the pair pins it down.

C THE REGISTERED SIGN CRITERION, AND THE PREDICTION THAT COULD NOT BE TESTED

Both facts above are post hoc. To check that a does work rather than merely describe data we had already seen, we wrote the reading rule and its predictions to a timestamped file and only then looked at the two ladders it had not seen, on a suite that played no part in shaping the rule. The registered inputs were the noise ladder of that suite and nothing else; the observation and action-repeat ladders were held back. The rule follows from Equation 1 and nothing was fitted: the sign of ρ is trustworthy when the interval for a excludes zero, and the full ordering is trustworthy when $|a|\,\Delta_{\min} > 1.96\sqrt{2}\,s_\varepsilon$, with $\Delta_{\min}$ the smallest true gap between adjacent policies and s_ε the residual spread. Read from that noise ladder alone, all three judges came out sign-trustworthy. Both transfer predictions held: the reading was unchanged on both held-out ladders (yes), and no judge changed sign across the three, with ρ staying inside $[0.941, 1.000]$ (yes).

A third registered prediction failed (no), and the fault is in the prediction rather than in the rule. We had asked that the swing in ρ order inversely with $|a|$. On that suite every one of the nine cells has $|a|$ between 0.277 and 0.845, all of them deep inside the regime the rule itself calls stable, and the three swings land between 0.057 and 0.059. We had asked for an ordering among three quantities our own model says should be indistinguishable. The ordering does appear where sensitivity actually varies, which is Section 4.2, but that is a different test and we do not get to claim this one.

D SAMPLE SIZE FOR A STABLE PAIRWISE ORDERING

Proposition 2 asks a question anyone designing an evaluation has to answer, and for a single pair the same question can be answered by simulation instead of by a bound. Simulating a perfect evaluator on two policies, keeping the probability of a reversed pair below five percent needs a true gap of 11.5 points at $n = 100$, 7.5 points at $n = 250$ and 5.5 points at $n = 500$.

Two cautions. The bound is sufficient, not tight: at a given n it over-states the probability of a reversed pair by a factor of 9 to 49 against Monte-Carlo simulation of the same setting, so it should be quoted as a budget that suffices and never as the budget that is required. And $\Delta_{\min}$ must be chosen to match the question being asked: the gap to the runner-up if only the winner matters, the smallest adjacent gap if the whole ordering matters.

E THE CEILING ON A CONTROLLED LADDER

At $k = 8$ and $n = 500$, a perfect evaluator scoring policies spread from 0.2% to 85.6% reaches $\rho = 1.000$ with a 95% interval of $[0.976, 1.000]$. The same perfect evaluator on policies clustered near one another reaches only 0.738 $[0.285, 0.964]$. Nothing about the evaluator changed.

F THE TWO SAMPLE-SIZE QUESTIONS ON A PUBLISHED LEADERBOARD

Table 3: Perfect-evaluator ceilings on a published leaderboard of 3 methods at the field-standard budget $n = 500$, with the sample size Proposition 2 requires for two different questions. The two differ by up to two orders of magnitude, so they must be labelled separately.

Suite	span	ceiling ρ	n for best	n for order
Spatial	0.064	0.866	1,097	102,469
Object	0.068	1.000	2,194	5,060
Goal	0.163	1.000	1,265	1,265
Long	0.032	0.500	5,457	102,469

The sample-size column of Table 3 deserves care, because the two questions it answers are easy to conflate: identifying only the top method on Spatial means separating a gap of 5.8 points and needs $n \geq 1{,}097$, whereas recovering the full ordering means separating the closest adjacent pair, 0.6 points, and needs $n \geq 102{,}469$. Both exceed the 500 rollouts the field reports, by factors of about two and two hundred. A further detail belongs to the finding rather than to the fine print: at $k = 3$, Spearman's ρ can take only the values ± 1 and ± 0.5, so the statistic the field reports has four possible values on the leaderboard the field actually publishes.

G REPRODUCTION GATES

Every suite had to reproduce the published number for that suite before it entered the study, with $|z| > 2$ aborting the run. Ours is a single seed of $n = 500$ against a published three-seed average, so the honest reading is *within error*, not *replicated*. Three of the four sit slightly low, which is worth stating even though it does not reach significance.

Table 4: Reproduction gates against the numbers published with OpenVLA.

Suite	Ours (%)	Published (%)	z
Spatial	85.6	84.7 ± 0.9	$+0.50$
Object	86.4	88.4 ± 0.8	-1.16
Goal	77.4	79.2 ± 1.0	-0.85
Long	53.4	53.7 ± 1.3	-0.12

H THE THREE LADDERS REPEATED ON A SECOND SUITE

Sensitivity, its significance, and the resulting swing in ρ across the three ladders, for all three judges on LIBERO-Goal.

Judge	$\lvert a \rvert$ range	a significant	swing in ρ
7B	0.001 to 0.013	1 of 3	0.865
32B	0.069 to 0.095	2 of 3	0.365
InternVL3	0.221 to 0.398	3 of 3	0.057

I PROMPT SENSITIVITY

Four prompts, from a bare binary question to one that names occlusion as an allowed reason to answer yes, move the 32B judge from $\rho = -0.970$ to $\rho = 0.647$. Over the same four prompts the 7B judge

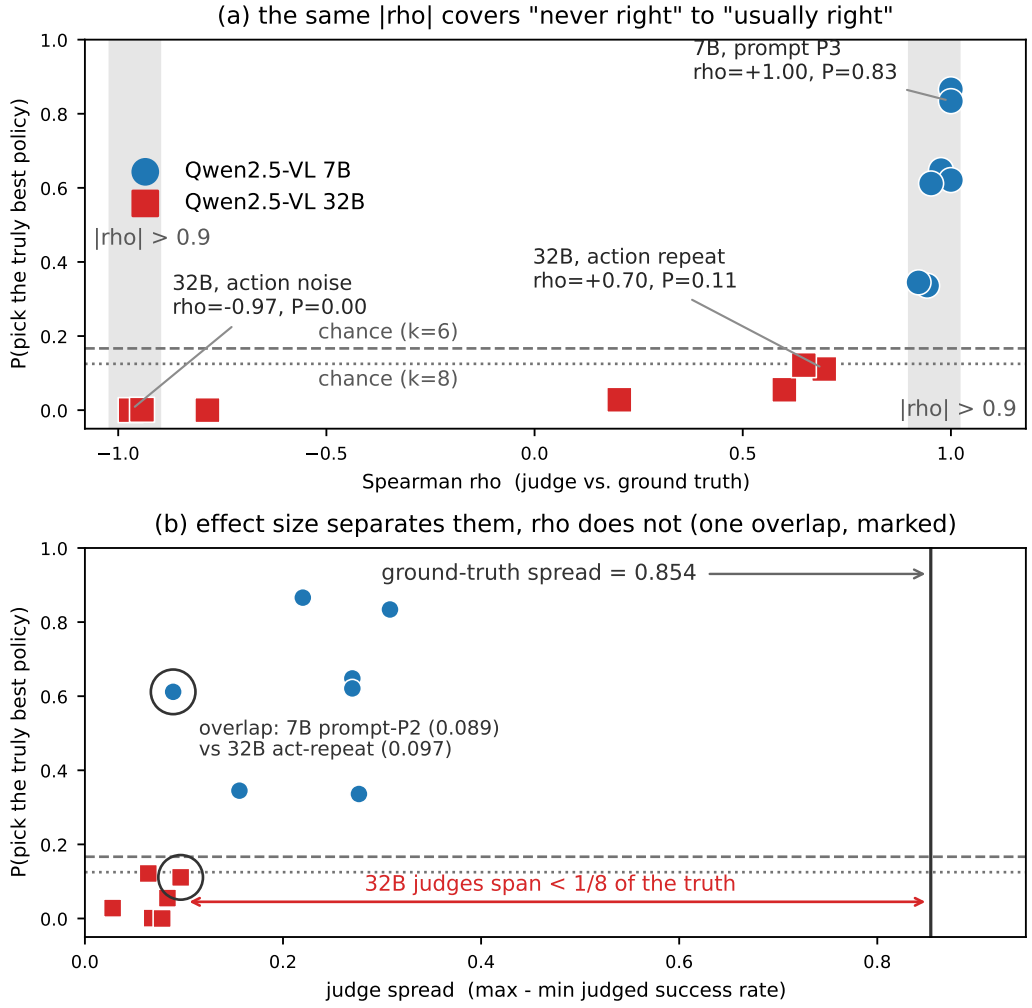


Figure 2: All settings from Sections 4.2 to 4.3, which vary the judge, the ladder, and the prompt on LIBERO-Spatial. *Top:* $P_{\text{top-1}}$ against ρ. The shaded bands are $|\rho| > 0.9$, and they contain settings whose $P_{\text{top-1}}$ ranges from never correct to usually correct. *Bottom:* the same points against the judge's own spread, that is its maximum minus minimum accepted rate. Effect size separates the two judges where ρ does not, with a single overlapping pair marked; note that every 32B setting has a spread under an eighth of the ground-truth spread, which is drawn as the vertical line.

never drops below $\rho = 0.922$. This blocks the obvious dismissal that we simply failed to tune the prompt: for one judge the finding is prompt-stable, and for the other the prompt alone decides the sign.

Figure 2 collects every setting from Sections 4.2 to 4.3 in one view. Restricting attention to settings with $|\rho| > 0.9$ still leaves $P_{\text{top-1}}$ scattered from 0.000 to 0.866, whereas plotting the same points against the judge's own spread separates the two judges almost cleanly. One pair overlaps, the 7B judge under the describe-then-decide prompt at 0.089 against the 32B judge on the action-repeat ladder at 0.097, and we mark it in the figure rather than rounding it away.

One prompt is worth isolating because it is the paper's clearest positive result. Telling the 7B judge that a partially occluded success still counts raises $P_{\text{top-1}}$ from 0.643 to 0.834 and ρ to 1.000, while signed calibration stays at -0.129. The judge becomes good at choosing and remains unusable as a success-rate estimate. Section 5 makes that split explicit.

A protocol defect in one prompt variant. The prompt that asks for a description before the verdict made the 7B judge emit 297 answers with no verdict in them. Scoring those as *no* silently inflated accuracy, because most of them were true failures anyway. Our tooling now reports both conventions and a gate refuses to aggregate a run whose parse failures were never inspected. We read the excluded convention throughout.

J THE FAILURE IS NOT A VISIBILITY PROBLEM

The natural explanation for a judge that rejects true successes is occlusion: the arm covers the object, or the final frame is ambiguous. We tested it by computing three image statistics per episode, namely how far the final frame moved from the first, the fraction of the frame occupied by the arm, and the distance from the final frame to the average successful final frame for that task, and asking whether they separate the judge's accepted from its rejected episodes. To hold ground truth fixed we restrict the comparison to ground-truth failures. With 3 tests the corrected threshold is 0.0167.

For the 32B judge nothing clears it: the smallest p is 0.0221. Its verdicts are close to independent of what is visible in the frame. The 7B judge is the opposite, separating strongly on goal appearance ($p = 1.2 \times 10^{-15}$) and on arm coverage ($p = 3.4 \times 10^{-11}$), yet it is the judge that under-reports by 0.198 on the same data. Using the visual evidence and being calibrated are different properties, and the first does not imply the second.

K SENSITIVITY TRACES BEHIND TABLE 1

Figure 3 plots the accept rate of each judge against the ground-truth success rate, one point per degradation level, for all four suites. Reading it against the diagonal is what sensitivity means: an evaluator that tracked reality exactly would lie on it. Two departures matter and neither is visible in ρ. A trace can be flat, so the judge barely moves when reality moves, and a trace can descend, so it moves the wrong way; both are compatible with $|\rho|$ close to one. Appendix P tabulates the sensitivity of every cell beside the other four quantities.

L WHAT EACH QUANTITY DOES AT $k = 2$

Comparable policies with both public weights and their own published per-suite numbers are scarce; at the time of writing we could identify two on LIBERO. At $k = 2$ Spearman's ρ can only take the values ± 1, whatever the evaluator does. Sensitivity is still estimable, with a wide interval that should be reported as wide; calibration is unaffected; $P_{\text{top-1}}$ degenerates gracefully into whether the evaluator ranks the stronger of two policies correctly, against a chance level of one half; regret degenerates into the true gap whenever the evaluator picks the weaker policy; and the sample-size bound applies unchanged. There is therefore no need to manufacture a third checkpoint simply to give ρ room to move.

M THE RELEASED TOOL

`surrocheck` is included in the supplementary material. It takes a generic table of `(policy_id, ground_truth, evaluator_pred)` rows and emits every quantity above, the two-cell verdict, and a LaTeX table. It imports none of our experiment code and knows nothing about robotics, so it applies to any surrogate evaluator with a ground truth to compare against. It was written as an independent second implementation and reproduces our primary pipeline on the same judgements, which is the strongest check we have on the numbers in this paper. The one quantity where the two agree only within error, rather than digit for digit, is $P_{\text{top-1}}$, because it is a bootstrap mean.

The package ships three worked examples drawn from this paper's own judgements: the 7B judge under the baseline prompt, the same judge under the occlusion-aware prompt, and the 32B judge. Their $|\rho|$ values are 0.976, 1.000 and 0.970, close enough to be mistaken for one another, while the tool separates them into *marginal*, *usable for choosing but not for reporting*, and *ordering reversed*.

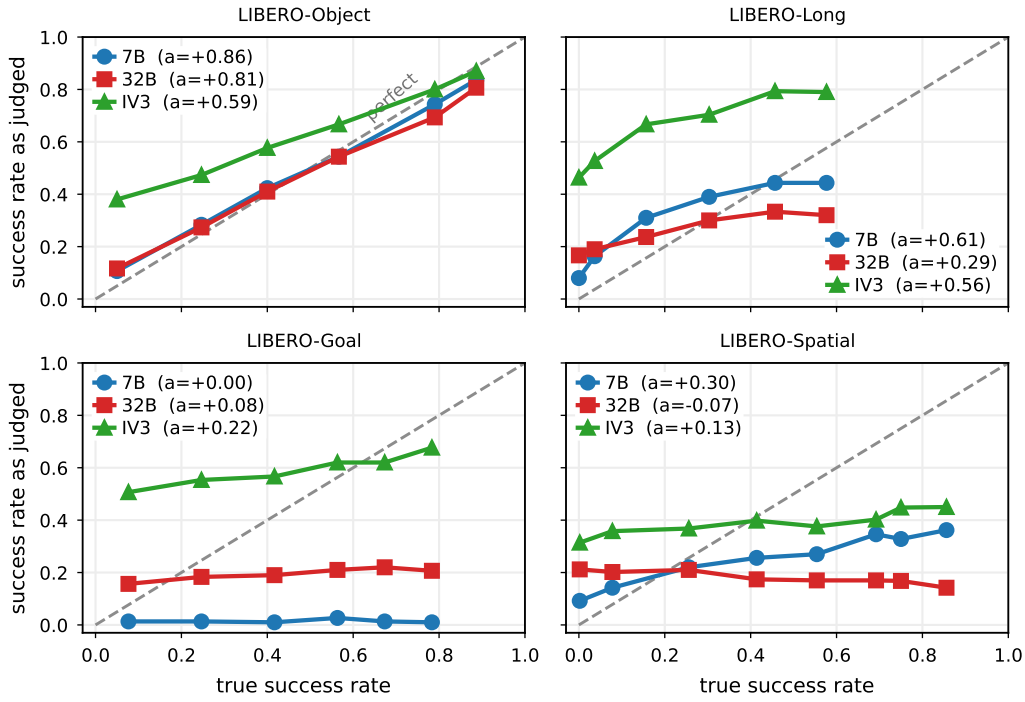


Figure 3: Judge accept rate against ground-truth success rate, one panel per suite, one trace per judge, with the perfect evaluator shown as the diagonal. The slope is the sensitivity a. A flat trace and a descending trace can both sit at $|\rho|$ near one, which is the axis rank correlation cannot see. Judges are Qwen2.5-VL 7B (7B), Qwen2.5-VL 32B (32B) and InternVL3-8B (IV3).

N TWO PRE-REGISTERED ATTEMPTS TO PREDICT FAILURE

Since validity belongs to the pair, the natural next question is whether a practitioner can predict, from the task alone, which pairs will break. We tried twice and pre-registered both attempts before looking at the relevant results.

The first attempt classified all 40 tasks from their text alone into four families, predicting that tasks whose success is a single object entering a container would have high sensitivity and that tasks whose success is a small articulated state change, or that require telling identical objects apart, would have low sensitivity. The predicted ordering does not hold: the class medians are 0.790, 0.160, 0.371 and 0.107, so the class we predicted to be third is second. The second attempt registered a containment hypothesis before the InternVL3 results existed. Pooled across suites the gap looks convincing at 0.445, but conditioning on suite splits it into 0.546 on Goal and -0.151 on Long. The pooled effect was a suite effect wearing a task-property costume.

The diagnosis is in the same data. Take a single task and change only the judge family: the median swing in sensitivity is 0.421 and the largest is 1.479, against a within-class standard deviation of 0.309. Switching judge moves sensitivity more than the task category does. Any rule written over task properties alone is therefore predicting a quantity that is not a function of its inputs. We report this as a limitation rather than as a result, and Section 6 keeps it.

O THREE STATISTICAL CAVEATS WE HOLD OURSELVES TO

The confidence interval on $P_{\text{top-1}}$ is conditional on this set of episodes. It is not the interval one would get by repeating the entire experiment. Separately, our reproduction runs use a single seed against published figures that average three, so Table 4 supports the statement that our stack is within error of the published one and not the stronger statement that we replicated it.

Third, the per-cell tests behind Table 2 are 6 tests, not one. Read at $|z| > 1.96$ they give 2 cells that order the pair correctly and 1 that is significantly reversed. Under a Bonferroni threshold of 0.0083 the reversed cell survives and only 1 of the correctly ordered ones does, so the corrected reading is the less flattering one for the judges and we report it here rather than leaving the uncorrected count to stand alone. The same correction is applied in Appendix J, where it is the reason nothing clears the threshold for the 32B judge.

P EVERY QUANTITY WE ASK FOR, ON EVERY CELL

Section 5 asks for five quantities rather than one. Table 5 reports all of them for the 12 cells of Table 1, so that the recommendation can be checked against the same data the rest of the paper uses. Calibration is given both ways because the signed value cancels errors in opposite directions; regret is in points of true success rate, against the reference line of choosing at random on the same policy set.

Table 5: All five reported quantities for every (judge, suite) cell, with the perfect-evaluator ceiling beside ρ. Sensitivity runs from 0.001 to 0.861 in absolute value while ρ is at least 0.82 in 10 of the 12 cells, which is the point of Section 5. Regret is in points; the random-choice line is 39.7 on Object, 32.2 on Long, 32.3 on Goal and 40.6 on Spatial; 1 of the 12 cells is worse than its own line.

Judge	Suite	a	|cal|	cal	ρ	ceiling	$P_{\text{top-1}}$	regret
Qwen2.5-VL 7B	Object	0.861	0.039	−0.001	1.000	1.000	0.997	0.0
Qwen2.5-VL 7B	Long	0.611	0.099	0.050	0.986	1.000	0.498	5.7
Qwen2.5-VL 7B	Goal	0.001	0.446	−0.446	−0.247	1.000	0.029	27.0
Qwen2.5-VL 7B	Spatial	0.300	0.237	−0.198	0.976	1.000	0.643	5.3
Qwen2.5-VL 32B	Object	0.807	0.051	−0.016	1.000	1.000	1.000	0.0
Qwen2.5-VL 32B	Long	0.288	0.131	0.003	0.943	1.000	0.306	9.6
Qwen2.5-VL 32B	Goal	0.079	0.292	−0.266	0.829	1.000	0.212	14.7
Qwen2.5-VL 32B	Spatial	−0.072	0.353	−0.269	−0.970	1.000	0.000	73.7
InternVL3-8B	Object	0.587	0.143	0.138	1.000	1.000	0.986	0.1
InternVL3-8B	Long	0.556	0.402	0.402	0.943	1.000	0.468	5.8
InternVL3-8B	Goal	0.221	0.184	0.131	0.986	1.000	0.865	2.1
InternVL3-8B	Spatial	0.133	0.237	−0.061	0.976	1.000	0.512	5.6