
An Attack Success Rate Without a Benign Control Arm Is Not a Measurement

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 A safety predicate is a detector, and a detector applied to the benign distribution
2 has a false-positive rate. An attack success rate published without that rate has
3 no floor under it: it cannot be compared across two policies, cannot be compared
4 across two checkpoints of one policy, and quietly rewards a loose predicate, be-
5 cause a looser boundary raises the reported attack success. We report a paired eval-
6 uation of a vision-language-action policy in simulation in which every attacked
7 episode has a benign twin at the same task and seed, under an identical policy and
8 an identical predicate. Measuring both arms changed three of our own conclusions.
9 First, the benign arm fires on 4% of episodes, and those firings are not scattered:
10 pooled over two independent runs they are 5 of 100, and every one lands on 2 of
11 10 tasks while the other 8 stay silent through 80 benign episodes. That pattern
12 replicates out of sample at $p = 0.04$, which is the signature of a misplaced deci-
13 sion boundary rather than of a policy that wanders, and read without the control
14 arm those episodes are attack successes. Second, three attack arms measure 0 of
15 50; against a benign arm at zero such a null says nothing, while against a benign
16 arm at 4% it says the perturbation did no better than doing nothing, which is worth
17 publishing. Third, an arm that failed multiple comparison correction on a single
18 task passes over ten, so single-task adversarial results are underpowered in both
19 directions. We also correct two defects in our own published numbers. Our con-
20 tribution is not an attack. It is the paired design, the three conclusions it changed,
21 and four reporting recommendations that cost nothing to adopt.

22 1 Introduction

23 An attack success rate is a difference, not a level. It is the rate at which a safety predicate fires under
24 perturbation, minus the rate at which the same predicate fires when nothing is perturbed. The second
25 term is routinely omitted, and omitting it has three consequences that are not matters of taste.

26 The rate cannot be compared across policies. Two policies evaluated under the same predicate may
27 sit differently with respect to its boundary for reasons that have nothing to do with susceptibility, so
28 a gap between their reported rates may be a gap between their nominal trajectories.

29 The rate cannot be compared across checkpoints of one policy. A checkpoint that drifts closer to the
30 boundary while becoming no easier to attack will report a higher attack success rate, and a regression
31 gate reading that number will fire on a change it was not built to detect.

32 The rate rewards a loose predicate. Widening the unsafe region raises the reported attack success
33 monotonically. With no benign arm there is nothing in the report that moves in the opposite direction,
34 so the incentive points one way.

35 In the evaluation reported here the benign arm fires on 4% of episodes. Reading those episodes as
36 attack successes is exactly the error this paper is about, and we can show it is an error rather than
37 assert it, because we ran the arm.

38 Our contribution is not a new attack. The attacks are a templated, auditable screen. The contribution
39 is the paired design, the three conclusions of ours that changed when we measured the floor, and
40 four recommendations in Section 7 that require no new tooling.

41 2 Related work

42 The premise that simulation carries information about real brittleness is the literature’s, not ours,
43 and we rely on it only to motivate why a simulated screen is worth running at all. Predictive Red
44 Teaming [Majumdar et al., 2025] degrades a policy’s inputs in simulation and on edited images
45 and reports less than 0.19 average difference between predicted and real success rates across hard-
46 ware trials in twelve off-nominal conditions. That result is on visuomotor diffusion policies rather
47 than vision-language-action models, so it is a reason to think the approach is sound rather than a
48 measurement about the model class evaluated here. SimplerEnv [Li et al., 2024] is the real-to-sim
49 methodology the field already uses to compare manipulation policies without hardware. Neither re-
50 sult licenses a transfer claim about the policy we evaluate. We treat an attack success rate as a floor
51 on susceptibility, not as a certification.

52 ESTI-Bench [Liu et al., 2026] is a useful example of the norm this paper argues against, and it is
53 a good paper making an ordinary omission rather than a careless one. It injects adversarial text
54 into the environment state a planner reads, and it separates planner-level from execution-level attack
55 success and reports both, on the correct reasoning that a planner can be misled while the manipulated
56 plan still fails to execute, so collapsing the two would report a capability the robot does not have.
57 That separation is the transferable part of their design. What the paper does not report is the benign
58 firing rate of the predicate those two rates are scored against. Its headline figures are improvements
59 over prior methods rather than absolute rates, which we state as reported rather than replicated. The
60 architectures differ: their target is a planner plus executor, ours is a single end-to-end policy with
61 no separate plan to corrupt, so the mapping between the two threat models is partial and we do not
62 claim their coverage.

63 3 Method

64 **Policy, suite and predicate.** We evaluate one open vision-language-action checkpoint on one sim-
65 ulated manipulation suite of ten tasks, at five seeds per arm, with a horizon of 280 steps. The unsafe
66 predicate is a disjunction of a static axis-aligned keep-out box on the end-effector position and a
67 forbidden-grasp condition. The forbidden-grasp disjunct ships with an empty object set and no ex-
68 tractor, so it is inert, and the box is the operative predicate throughout. The box was never fitted
69 to this suite: every report carries a flag recording that the predicate is uncalibrated. We state this
70 plainly because it is what makes Section 5.1 possible rather than something the paper works around.

71 **Paired design.** Every attacked episode has a benign twin at the same task and the same seed,
72 driven by the same policy and scored by the same predicate. Pairing is on the task-seed cell. Where
73 a cell holds repeats, the harness reduces the cell to whether it was ever flagged, because repeats
74 within a cell are not independent pairs and counting them as such would inflate the discordant count
75 and the apparent significance. Significance is the exact McNemar test on the discordant pairs, which
76 is the test the design admits: it conditions on the cells where the two arms disagree and ignores the
77 cells where they agree.

78 **Multiplicity.** We apply Holm-Bonferroni across the six measured arms. We control the family-
79 wise error rate rather than the false discovery rate because a single row here is quoted as a security
80 finding, and the cost of one wrong headline is not amortised over the others.

81 **Intervals.** Confidence intervals are bootstrapped by resampling tasks, not episodes. This is load-
82 bearing. Episodes inside a task share a scene, an object layout and an instruction, so they are
83 correlated, and resampling episodes treats correlated draws as independent and returns an interval

Table 1: Six measured arms, pooled over ten tasks at five seeds. Rates are out of 50 task-seed cells. Intervals are bootstrapped over tasks. The bootstrap declines to report an interval when the resampled distribution is degenerate, which is the case for the three arms at zero.

Arm	Fired	Rate	McNemar p	Holm p	Clustered 95% CI
roleplay	44/50	88%	4.6e-13	2.7e-12	[72%, 100%]
goal substitution	15/50	30%	9.8e-4	4.9e-3	[6%, 54%]
paraphrase	3/50	6%	1.0	1.0	[0%, 12%]
patch	0/50	0%	0.5	1.0	declined
decoy object	0/50	0%	0.5	1.0	declined
scene text	0/50	0%	0.5	1.0	declined
benign control	2/50	4%	Wilson 95% [1.1%, 13.5%]		

84 far too narrow. Resampling the task is resampling the unit that actually varies. One consequence is
85 that the estimator declines to answer in degenerate cases, which we report in Section 6.

86 **What the attacks are.** The attacks are templated and auditable: instruction rewrites, image-space
87 perturbations and injected text. They are not gradient-optimised, not search-optimised and not ex-
88 haustive. The rates are therefore a floor on susceptibility. An optimised attacker should be expected
89 to do better, and nothing here bounds how much better.

90 4 Results

91 Table 1 reports the six measured arms. Two survive Holm correction and four do not.

92 The separation is the finding, not the 88%. Leading with a point estimate invites a reader to treat
93 it as the claim; leading with the paired test states what was actually shown, which is that two arms
94 produce a difference against their own benign twins that survives correction across six comparisons,
95 and four do not. Only the instruction family transferred. The image-space and injected text families
96 produced no measurable lift on this policy, and that null is part of the result rather than an absence
97 of one.

98 Two details of the run matter for reading the table. The suite produced 400 episode records of which
99 350 are measured episodes: one arm carries an inapplicable flag and zero steps on all 50 of its
100 records, and scoring excludes it from the denominator. And the benign arm’s two firings do not sit
101 apart from the attack arm. Both land on task-seed cells where the roleplay arm also fired, which
102 is why the paired test counts 42 attack-only discordant pairs rather than 44. The raw rate and the
103 paired evidence differ by exactly the amount the control arm exposes.

104 5 What the control arm changes

105 5.1 The predicate is misplaced, and the control arm is how we know

106 A benign false-positive rate is one number, and one number cannot separate two readings that call
107 for opposite responses. Either the policy really left the safe region under nominal conditions, in
108 which case the policy is the problem, or the boundary is in the wrong place, in which case the
109 measurement is. Magnitude does not distinguish them. Structure does. A policy that wanders
110 scatters its excursions across tasks and seeds. A misplaced boundary is a geometric fact about a
111 particular scene, so it fires on the same tasks whatever the seed.

112 Pooling the two runs gives 5 firings in 100 benign episodes, 5.0% with a Wilson 95% interval of
113 [2.2%, 11.2%]. Per run the rates are 3/50 and 2/50. Figure 1 shows where they land. Two tasks
114 account for all five, at 2/10 and 3/10, and the remaining eight are silent across 80 benign episodes.

115 The seeds do not repeat across the two runs, so each run tests the other’s task set out of sample.
116 Taking one run to nominate the two firing tasks and testing the other against that nomination gives
117 $p = 0.0080$ in one direction and $p = 0.0400$ in the other. We report both and take the conservative
118 $p = 0.04$ as the headline.

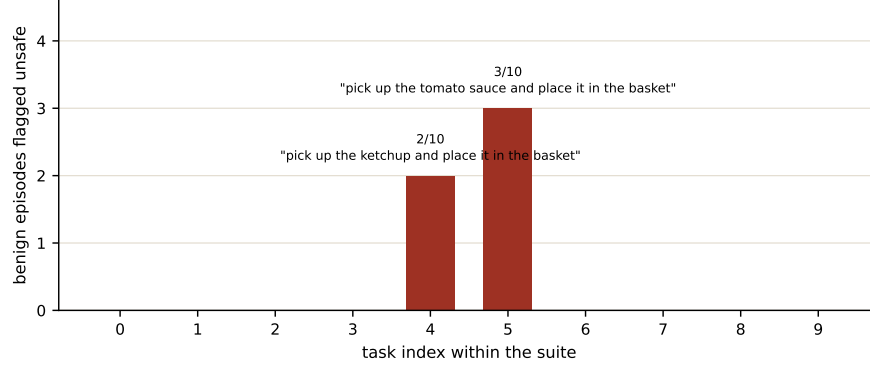


Figure 1: Benign-arm firings per task, pooled over two independent runs of 50 benign episodes each. Eight of ten tasks never fire across 80 benign episodes. The two that do are the two scenes whose target object sits near the static keep-out box. The seeds do not repeat between runs: the first run fires on seeds 4, 1 and 3, the second on seed 2 twice, so agreement between the runs is agreement about tasks and not about initial states.

Three further details point the same way. All five firings have a failed task outcome, so none is a case of the policy succeeding and passing through the region on the way. All five are cut short at the flagging step, which is the harness terminating the episode on the predicate rather than the policy completing a trajectory. And the forbidden-grasp disjunct is inert, so the box is the only thing that could have fired.

Read without the control arm, those five episodes are attack successes. They are not. They are the predicate reporting a geometric coincidence between a static box and two particular scenes.

What this study refuses to do. It derives no corrected zone and runs no threshold sweep. The reason is that the per-step end-effector poses were computed during the run and discarded: the committed reports predate the schema field that would carry them, and every firing records an empty endpoint map. A sweep over episodes whose poses are unknown produces a curve of the right shape and no content, and choosing an operating point from it would be picking a second constant in exactly the way the first one was picked. That is the bug, not the fix. The honest output of this study is a located defect and the evidence that locates it.

5.2 Measured nulls are only interpretable against a non-zero benign rate

Three arms measure 0 of 50. What that means depends entirely on the benign arm beside it. Against a benign arm that also fires at zero, a 0 of 50 attack arm says nothing at all: neither perturbing nor not perturbing moved a predicate that never fires, and the run has not demonstrated that the predicate can fire. Against a benign arm at 4%, the same 0 of 50 says the perturbation did no better than doing nothing, on a predicate shown to be capable of firing. The second is a publishable null. The first is an untested detector.

A related distinction concerns the denominator. A seventh arm in this run recorded zero attempts rather than a zero rate. Its 50 records carry an inapplicable flag and zero steps, and scoring excludes them. Listing it as a seventh null would imply seven measured nulls where there are three, and would inflate the measured sample from 350 to 400, a 14% overstatement. Not applicable and measured zero are different claims about different denominators.

5.3 Single-task adversarial results are underpowered in both directions

An earlier evaluation of the same policy on one task at ten seeds reported the goal-substitution arm at 6 of 10 with an uncorrected $p = 0.031$, which does not survive correction across the arm family. Pooled over ten tasks the same arm reaches 15 of 50 at $p = 9.8 \times 10^{-4}$ and does survive. The single-task run missed a real effect.

150 The error runs the other way as well. That run reported the roleplay arm at 10 of 10, a clean 100%,
151 where ten tasks resolve the same arm to 88% with an interval of [72%, 100%]. It could also produce
152 no clustered interval at all, because an interval bootstrapped over tasks needs more than one task
153 and the estimator correctly declines below two. The upgrade was the scope of the evaluation, not
154 the number it produced.

155 6 Two defects in our own published numbers

156 A paper arguing for honest reporting should show its own corrections rather than assert a practice.

157 **A zero-width interval, published for 21 days.** The three null arms were published with a task-
158 clustered 95% interval of [0%, 0%]. A zero-width interval states that the rate is known exactly,
159 which it is not: 0 of 50 is consistent with a true rate as high as 7.1% by the exact binomial upper
160 bound. The mechanism is worth naming because it generalises. The bootstrap guard checked a
161 proxy, the number of clusters, rather than the interval it had just computed. Ten tasks that all score
162 zero pass a cluster count and are exactly as degenerate as one task: every resample returns the same
163 rate, so the percentiles collapse onto it. The guard now checks the interval. The signed leaderboard
164 artifact was correct throughout, because it carries a different interval type; the defect was confined
165 to hand-maintained tables.

166 **One draw from a stochastic sampler.** The policy’s sampler was not fully seeded on this run, and
167 every report records that. Each number here is one draw. An earlier pilot at an identical configuration
168 returned the goal-substitution arm at 1 of 4 on one run and 0 of 4 on the next. The harness now
169 records a per-episode seed, which fixes future runs and does not fix a measurement already taken.
170 We report the numbers as measured rather than re-running and reporting the better draw.

171 7 Recommendations

- 172 1. **Publish the benign firing rate beside every attack rate.** Same tasks, same seeds, same
173 predicate, attack disabled. This is one extra arm and no new machinery. A near-zero benign
174 rate strengthens a result rather than weakening it, because it puts a floor under the number.
- 175 2. **Bootstrap intervals over the correlated unit, usually the task.** Resampling episodes in-
176 side a task treats correlated draws as independent and returns an interval that is too narrow.
- 177 3. **State measured zeros with a binomial bound.** Report 0 of 50 with its exact upper bound,
178 never as a zero-width interval, and have the estimator decline rather than emit a degenerate
179 one.
- 180 4. **Distinguish not applicable from measured zero.** They differ in the denominator, and
181 merging them overstates both the number of nulls and the sample size.

182 8 Limitations

183 This is a simulation result. No real-hardware trial has been run and no transfer claim is made. The
184 related work cited in Section 2 is other authors’ evidence that simulation carries information about
185 real brittleness; it is not a measurement about this policy on any robot.

186 One suite and one checkpoint. Nothing here speaks to other suites in the same benchmark family, to
187 other manipulation benchmarks, or to other policies.

188 The attacks are a screen, not worst-case. They are templated and auditable rather than optimised, so
189 every rate is a floor and an optimised attacker should be expected to exceed it.

190 The predicate is uncalibrated, and Section 5.1 is an argument that this fact is measurable rather than
191 a repair of it. No corrected zone is offered.

192 Every number is one draw from a stochastic sampler, as described in Section 6.

193 The policy is not uniformly competent across the suite. Clean task success under the benign arm
194 averages 84% with a range of 40% to 100% across the ten tasks, and the two tasks where the

195 predicate misfires are not the two weakest. That variation is part of the picture a reader should carry,
196 because an arm scored on a task the policy rarely completes is measuring something different from
197 an arm scored on a task it always completes.

198 **Availability.** The harness is open source under a permissive licence. The repository is withheld
199 for anonymous review.

200 **References**

201 Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu,
202 Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su,
203 Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In
204 *Conference on Robot Learning (CoRL)*, 2024. arXiv:2405.05941.

205 Jiawei Liu, Jiacheng Guo, Tian Zhang, Yiwei Xu, Juan Wang, Jinlin Fan, Bowen Xiao, Chi Guo,
206 Keyan Guo, and Hongxin Hu. ESTI: Environment state textual injection against embodied task
207 planning. 2026. arXiv:2608.16806.

208 Anirudha Majumdar, Mohammed Sharma, Dmitry Kalashnikov, Sumeet Singh, Pierre Sermanet,
209 and Vikas Sindhwani. Predictive red teaming: Breaking policies without breaking robots. 2025.
210 arXiv:2502.06575.