WHEN THE SIMULATOR FAILS, NOT THE POLICY: AUDITING CONTACT ARTIFACTS IN SIMULATED ROBOT BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation benchmarks score robot policies by success rate, and every failed episode counts against the policy. Testing a vision language action model and a world action model in simulation, we find that failures in which the policy has placed the object and the simulator then ejects it are frequent, and that the benchmark's own scripted demonstrations fail the same way. We present SIMUGUARD, an auditing plugin for LIBERO, RoboTwin, RoboCasa and RoboDojo. It leaves the evaluation unchanged. At every physics step, it records the state and flags an object that starts moving at a contact faster than the robot's contact or gravity account for. It replays a recorded episode bit for bit, and it reruns the episode with one solver parameter changed. With it we trace each such episode step by step to the contact solver's response at a resting contact: within one physics step the solver reports an overlap of about two centimeters between the placed object and the thin wall of its container, removes the whole overlap in that same step, and the separation speed throws the object out at several meters per second. In SIMUGUARD's controlled reruns, a limit on the separation speed slows the ejection and lets most of the flagged failures pass. The artifacts change the score, and every way of scoring them rests on an assumption. Reported rates assume every ejection is the policy's fault. Replaying the recorded actions with the separation speed limited on every body from before each artifact assumes those actions remain the policy's choice once the ejection is gone. Excluding the flagged episodes assumes they are as hard as the rest. Running the policy closed loop with the limit on for the whole rollout assumes the limit changes nothing the task needs, and our data refute this one: it lowers the score of a task the artifacts barely touch. SIMUGUARD lets a benchmark separate what the policy did from what the solver did.

1 INTRODUCTION

Simulation benchmarks give a repeatable way to compare robot policies, from vision language action (VLA) models to world action models (Brohan et al., 2023a;b; Octo Model Team et al., 2024; Kim et al., 2024; Black et al., 2024; Björck et al., 2025; Gemini Robotics Team, 2025; Li et al., 2026). LIBERO (Liu et al., 2023), RoboCasa (Nasiriany et al., 2024) and RoboTwin (Mu et al., 2025; Chen et al., 2025) fix the tasks and the protocol and treat the simulator as faithful physics. Their success rates record which episodes failed, not why. A failure caused by the simulator's treatment of contact enters the same count as a failure caused by the policy. Testing a world action model and a VLA in simulation, we find that such failures are frequent: the policy places the object, and the simulator throws it out.

In an episode of RoboTwin 2.0's Place Can Basket, LingBot-VA (Li et al., 2026) places a can inside the basket, resting against its wall. The gripper releases it. After 0.28 s, within a single 4 ms physics step, PhysX reports the can 2.75 cm inside the basket wall, where the step before it reported contact without overlap, and the can's speed rises from 0.2 to 22 m/s while no robot link touches it and the robot is nearly still; the basket ejects it, the can lands 1.65 m away, and the benchmark records a failure (Figure 1). Nothing in the recording supplies the 2.4 J the can gains: no robot contact in that step and no drop. Replayed with the same recorded actions and the overlap removed over several

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.

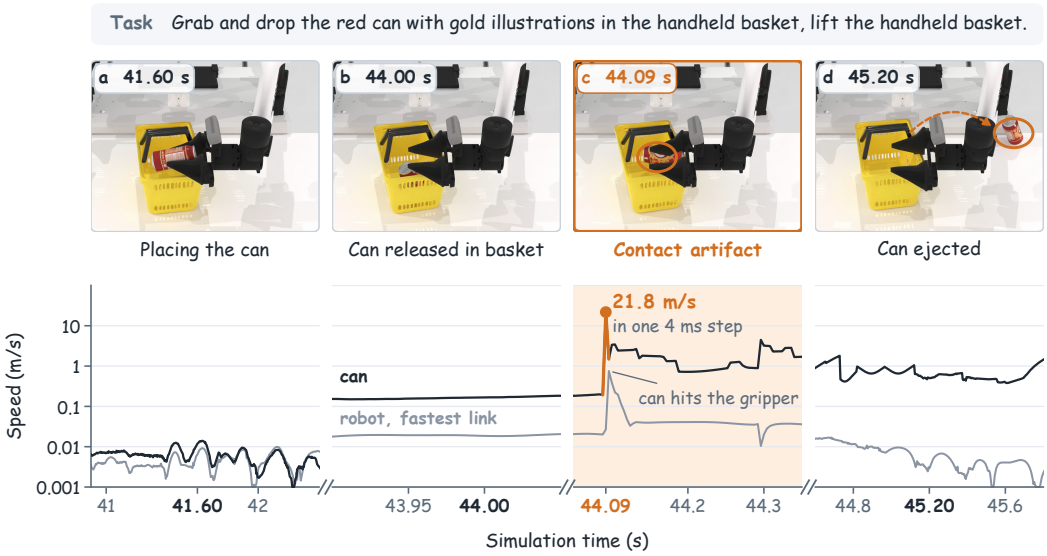

Figure 1: **A contact artifact turns a correct placement into a failure.** After LingBot-VA places the 10 g can in the basket on Place Can Basket (a, b), the can's speed rises from 0.2 to 22 m/s within one 4 ms physics step, a gain of 2.4 J of kinetic energy, while no robot link touches it and the fastest link moves at under 0.1 m/s (c); the can is ejected from the basket and the benchmark records a failure (d). The time axis has four linear segments, one under each frame.

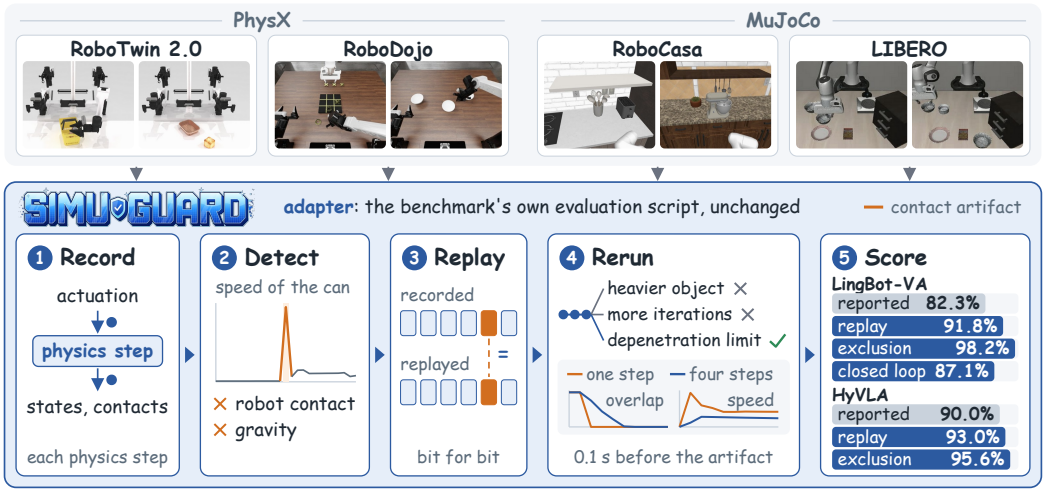

Figure 2: **SIMUGUARD audits a simulated evaluation without changing it.** Attached by an adapter to the benchmark's own evaluation script, SIMUGUARD records every physics step, detects contact artifacts, replays the episode exactly, reruns it from just before the artifact with one simulation parameter changed, and scores the rerun with the benchmark's own success check. Inset: PhysX removes a 20 mm overlap in one step and releases the can at 4.2 m/s, or over four steps at 1.2 m/s with the depenetration limit. Score, on the five container tasks: *reported* counts every ejection against the policy; *replay* keeps the recorded actions and limits the depenetration velocity from before the artifact; *exclusion* drops every flagged episode; *closed loop* keeps the limit on for whole new rollouts, on 170 of the 220 seeds, where the as run rate is 81.8%. Each rests on one assumption (Section 3.6). HyVLA's closed loop rate is in Table 1.

physics steps instead of one, the episode passes the benchmark's own success check (Figure 3). We call a contact response that sets an object moving faster than the robot's contact or gravity account for a contact artifact. The impulse that starts one lasts a single physics step and falls between two

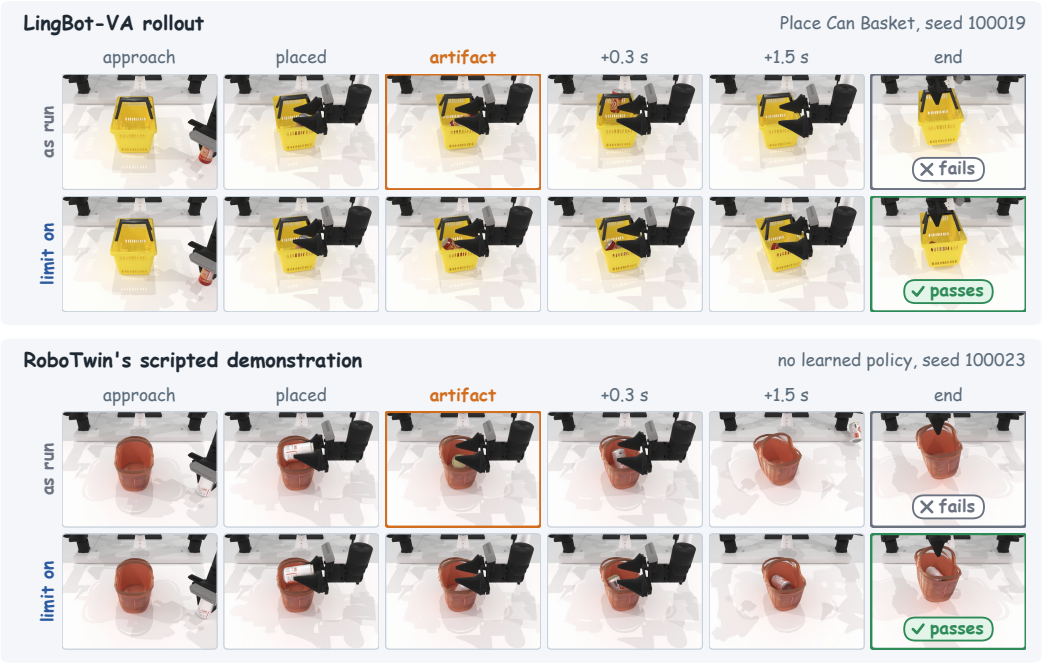

Figure 3: **Two episodes whose outcome the artifact window intervention flips.** Each is shown as run and replayed with the depenetration limit switched on 0.1 s before the first artifact; seed, actions and physics up to the switch are identical. *Top:* LingBot-VA on Place Can Basket. *Bottom:* one of RoboTwin's own scripted demonstrations, which the benchmark discards when it fails.

camera frames; a video shows the can flying, not what launched it, and the success check cannot tell whether the policy or the contact solver decided the episode.

SIMUGUARD is an auditing plugin that leaves the evaluation unchanged (Figure 2). An adapter attaches it to the benchmark's own evaluation script; we provide adapters for LIBERO, RoboTwin, RoboCasa and RoboDojo. At every physics step it records the actuation, the object states and the contacts, and flags an object that starts moving at a contact faster than the robot's contact or gravity account for. It replays a recorded episode bit for bit by rebuilding the seeded scene and applying the recorded actuation from the first step. A rerun is a replay with one solver parameter changed, with the recorded actuation and the success check fixed. Prior work adds metrics to the success rate (Wang et al., 2025; Snyder et al., 2026; Jiang et al., 2026) or measures how solvers depart from physics (Erez et al., 2015; Le Lidec et al., 2023); neither connects an anomalous contact in a rollout to the outcome of that rollout.

Our contributions are:

- **A finding.** Simulated benchmarks contain frequent episodes in which the policy has placed the object and the contact solver then ejects it, so the benchmark scores a failure of the policy. On five tasks that place objects into containers with thin walls, 36 of LingBot-VA's 39 failures and 6 of HyVLA's 10 contain such an ejection, and most of them pass once it is removed. The same ejection strikes the benchmark's own scripted demonstrations, which act on ground truth object poses, so it is not a property of the learned policy.
- **A tool.** SIMUGUARD monitors every physics step, replays a recorded episode bit for bit, and reruns it with one solver parameter changed, so that a single contact response can be traced to the outcome of its episode and the trace verified by anyone. It attaches to the benchmark's own evaluation script without changing it, on LIBERO, RoboTwin, RoboCasa and RoboDojo.
- **An evaluation method.** We compare four ways of scoring a benchmark with contact artifacts: ignoring them, which charges the policy for every ejection; excluding the flagged episodes; replaying each flagged episode with the depenetration velocity limited from

before the artifact, which we propose; and keeping the limit on for whole closed loop rollouts, which changes the task. Each rests on one stated assumption, and the experiments test them.

2 Related Work

Simulated evaluation of robot policies. VLA policies adapt large pretrained backbones to produce robot actions (Brohan et al., 2023a;b; Open X-Embodiment Collaboration et al., 2024; Octo Model Team et al., 2024; Kim et al., 2024; Black et al., 2024; Björck et al., 2025), and world action models predict future video together with the actions (Du et al., 2023; Wu et al., 2024; Cheang et al., 2024; Hu et al., 2025; Li et al., 2025; 2026). Both are compared on simulated suites built on two families of physics engines: LIBERO (Liu et al., 2023) and RoboCasa (Nasiriany et al., 2024) run on MuJoCo (Todorov et al., 2012), while ManiSkill (Mu et al., 2021; Gu et al., 2023; Tao et al., 2024) and RoboTwin (Mu et al., 2025; Chen et al., 2025) run on SAPIEN (Xiang et al., 2020), which uses PhysX, as does Isaac Lab (Mittal et al., 2023). SimplerEnv (Li et al., 2024b) tests whether their rankings carry over to real robots, and Jiang et al. (2026) audit five of these suites and find shortcut solutions, statistically insignificant gains and overfitting to the test distribution; they ask what a score says about the policy and take the simulator as given. SafeVLA-Bench (Fan et al., 2026) instruments existing benchmarks after the fact to report successes that violate a safety specification, and also takes the simulated physics as correct. All of these suites score each episode with a scripted check on the final state and do not examine the physics that produced it.

Contact simulation. Rigid body engines approximate contact differently. MuJoCo treats it as a soft constraint that removes an overlap over a time constant of several physics steps (Todorov et al., 2012); PhysX treats it as a hard constraint and pushes overlapping bodies apart within the step, up to a configurable maximum speed, and its documentation records that bodies created in overlap may be separated violently and names the depenetration limit as the remedy (NVIDIA Corporation, 2024). Compliant and pressure field models (Castro et al., 2023; Elandt et al., 2019) make other choices. Engines disagree on the same scenes (Erez et al., 2015) and on recorded real impacts (Acosta et al., 2022), contact models trade physical consistency for tractability (Le Lidec et al., 2023), and convex decomposition of assets adds its own approximation (Wei et al., 2022). These studies compare engines in isolated scenes or document a known behavior; none follows one contact response inside a fixed benchmark to the outcome of the episode, which is what SimuGuard does (Section A).

Reliability of benchmark numbers. Several studies ask what a success rate measures and how many trials it needs (Kress-Gazit et al., 2024; Vincent et al., 2024). RoboEval (Wang et al., 2025) adds behavioral metrics because equal success rates can hide different executions; Snyder et al. (2026) replace binary outcomes with graded progress; LIBERO-PRO (Zhou et al., 2025) and LIBERO-Plus (Fei et al., 2025) show that high scores often do not survive changes to objects, layouts and observations; and Henderson et al. (2018), Agarwal et al. (2021) and Bouthillier et al. (2021) show how variance and point estimates distort comparisons. These works improve how outcomes are aggregated and take the outcome of each episode as given. We test the outcome itself: SimuGuard traces one contact response to the outcome of the episode that contains it, with a replay anyone can verify, and compares four ways of scoring a benchmark once its artifacts are known.

3 SimuGuard

SimuGuard has three requirements: it must not change what it measures, it must reproduce an episode exactly so that a difference after an intervention belongs to the intervention, and it must run with the benchmark's own evaluator so that its findings apply to the reported numbers.

3.1 Setting

A benchmark evaluates a policy on a sequence of seeds. Each seed fixes the scene, and the evaluator runs the policy's actions through the simulator until a scripted success check passes or the action budget is spent. The simulator advances in physics steps of a few milliseconds (4 ms in RoboTwin,

2 ms in LIBERO and RoboCasa); one policy action spans tens of them. Of N scored episodes, S succeed, and the benchmark reports S/N.

A contact artifact is motion that starts at a contact and that neither the robot's contact nor gravity accounts for; in this paper the term names what the detector of Section 3.3 flags, and Section 4.5 measures its precision against human judgment. Let A be the set of scored episodes that contain one. Figure 2 shows the five steps that follow.

3.2 Recording without intrusion

SimuGuard wraps the physics step of the running scene, so the evaluator, the benchmark and the policy run unmodified in their own process. A thin adapter per engine provides the wrapper, around the step function of SAPIEN (RoboTwin), Isaac Sim (RoboDojo) or robosuite, which runs LIBERO and RoboCasa on MuJoCo. Before each step, SimuGuard records the actuation: joint drive targets and forces in PhysX; control inputs, applied forces and mocap poses in MuJoCo. It also records any state that benchmark code writes between two steps, which it must replay (Section 3.4). After each step, it records the pose and velocity of every body and the contacts, merged per pair of bodies, with impulse and penetration depth. A task specification names the objects a task manipulates and the containers they go into. The evaluator, the benchmark and the policy code are untouched, and with the monitor on or off, all 59 tracked object trajectories in our check are identical; the monitor adds 7.5 to 11.6 ms to each 4 ms step in SAPIEN (Section C.1).

3.3 Detecting contact artifacts

Detection has three stages, with the same thresholds for every task (Section B.5). The first finds motion that starts at a contact. An object becomes a candidate when it touched another body within the last 40 ms, moves faster than 0.5 m/s, and changes velocity by at least 0.35 m/s within one 4 ms step (the thresholds on velocity change scale with the step). The candidate is confirmed if within 240 ms it travels at least 4 cm and either flies free for 16 ms or stays in violent contact with its container.

The second stage removes motion that gravity explains. An object dropped into a bin also moves fast right after a contact. For each confirmed event, SimuGuard compares the object's peak speed in the 0.1 s after onset with the free fall speed $\sqrt{2g\,\Delta z}$ for the height it dropped before that peak, and keeps the event only if the ratio exceeds 1.5.

The third stage removes motion that the robot transmits through another body. A basket that the robot's base shoves across the floor carries the bread inside it at the basket's speed. SimuGuard finds every body that touched both the object and a robot link in the 40 ms before onset, and discards the event if the object's peak speed is less than 1.5 times the fastest such carrier's speed before onset. An event that passes all three stages is a candidate contact artifact. The stages compare speeds, not the work done on the object, so human review is the final standard: raters judged the flagged clips frame by frame (Section 4.5).

3.4 Exact replay

SimuGuard replays an episode by rebuilding the scene from its seed and applying the recorded actuation from the first physics step. Because the whole trajectory is regenerated from its initial conditions, the replay matches the recording bit for bit (Section 4.5). Three details make this hold. PhysX cannot restore a saved state exactly once bodies touch, since trajectories restored through its state interface diverge by millimeters to centimeters, so SimuGuard always replays from the start. Benchmark code changes the state between physics steps: a RoboCasa fixture sets its door joint, and robosuite 1.4, which LIBERO uses, recomputes the dynamics before each step and so rewrites the solver's warm start; SimuGuard records both and writes them back. Finally, outcomes depend on values outside the physics: RoboTwin's success checks read the gripper opening the robot was commanded to, which SimuGuard restores at each step from the recorded drive targets.

3.5 CONTROLLED RERUNS

A rerun is a replay with one simulation parameter changed: the mass of the manipulated object, the number of solver iterations, the maximum speed at which PhysX separates overlapping bodies (its maximum depenetration velocity), or the time constant of MuJoCo's contacts. A rerun either applies the change from the first physics step, or switches it on 0.1 s before the episode's first contact artifact. In the second form, everything before the switch is identical to the recording, which SIMUGUARD verifies bit for bit; from the switch on, the limit acts on every body until the episode ends. A different outcome therefore comes from the physics after the switch, of which the artifact is the first event the recording shows: an artifact window intervention, not an intervention on one contact. The benchmark's own success check scores every rerun, and the evaluator stops at the first success, as in the original evaluation. A third form keeps the change on for a new closed loop rollout on the same seed; the policy samples new actions, so the episode is no longer comparable one by one with the recording.

3.6 SCORING

Once the flagged episodes A of an evaluation are known, and if only they can change, the success rate under a physics that never ejects lies between $(S - S_A)/N$, every flagged episode failing, and $(S - S_A + |A|)/N$, every flagged episode succeeding, where S_A counts the successes in A; the one assumption, that no unflagged episode changes, fails where the detector misses an artifact. On the five RoboTwin tasks of Section 4.1 it is $[74.5\%, 98.6\%]$ for LingBot-VA and $[86.0\%, 96.0\%]$ for HyVLA. Each of the four rates we report is a point inside it under one further assumption (Figure 2, Score). The *reported* rate S/N assumes that every ejection is the policy's fault. The *replay* rate replaces the recorded outcome $y_e \in \{0, 1\}$ of each flagged episode by the outcome r_e of its recorded actions under the limit switched on before the artifact,

$$\hat{s} \;=\; \frac{1}{N}\Big(S \;+\; \sum_{e \in A} (r_e - y_e)\Big), \tag{1}$$

and assumes that those actions remain the policy's choice once the ejection is removed. The *exclusion* rate $(S - S_A)/(N - |A|)$ assumes that the flagged episodes are as hard as the others. The *closed loop* rate is the success rate of new rollouts with the limit kept on, and assumes that the limit changes nothing the task needs. We report the interval and the artifact count, and use the replay rate as the point estimate.

4 EXPERIMENTS

Setup. We run RoboTwin 2.0 (Chen et al., 2025) through its official evaluation script in its clean configuration (no background, table or lighting randomization, seen instructions), on SAPIEN (Xiang et al., 2020) with PhysX and the benchmark's 4 ms physics step. We audit two kinds of policy, each with its RoboTwin checkpoint: LingBot-VA (Li et al., 2026), a world action model that predicts future video together with its actions, on all 50 tasks, and HyVLA (Zhang et al., 2026), a vision language action model, on the five container tasks of Section 4.1 and six other tasks. Before the policy runs on a seed, RoboTwin runs a scripted demonstration and scores the policy only on seeds where the demonstration passes; every rate in this paper is over scored episodes, and skipped seeds do not enter N. On MuJoCo, with a 2 ms physics step, we run $\pi_{0.5}$ (Physical Intelligence et al., 2025) on the 40 tasks of LIBERO and XR-1 (Fan et al., 2025) on all 50 tasks of its RoboCasa evaluation set. RoboDojo (Chen et al., 2026) runs on Isaac Sim with PhysX and a 4 ms physics step; we run XR-1 on four of its tasks. SIMUGUARD monitors every episode, and all analysis uses the stored recordings. Episode counts, versions and checkpoints are listed in Section B.

4.1 WHERE CONTACT ARTIFACTS OCCUR

Failures with contact artifacts concentrate in five tasks, each of which places an object into a container with thin walls (Table 1). Of LingBot-VA's 39 failures on these tasks, 36 contain an artifact; in its other 45 tasks, 10 of 128 failures do, on six tasks (Table 17). HyVLA shows the same pattern: 6 of its 10 failures on the five tasks contain an artifact, and none of its 38 failures on the six other tasks.

Table 1: **Success rates on RoboTwin 2.0** under the four scores of Section 3.6, in percent. *Rep.*: as run. *Replay*: Equation (1), every episode with an artifact replayed with the limit switched on before it. *Excl.*: rate over the episodes without an artifact. *CL*: new closed loop rollouts with the limit on; as run rate on the shared seeds 81.8% (LingBot-VA) and 90.7% (HyVLA). Counts and the tasks outside the five: Table 16.

	LingBot-VA (world action model)				HyVLA (vision language action)			
Task	Rep.	Replay	Excl.	CL	Rep.	Replay	Excl.	CL
Place Can Basket	74.4	86.7	96.2	90.0	80.0	75.0	81.3	80.0
Stack Bowls Three	76.0	90.0	100.0	100.0	90.0	95.0	100.0	100.0
Place Object Basket	90.0	95.0	94.4	50.0	95.0	95.0	95.0	45.5
Place Bread Basket	97.5	100.0	100.0	85.0	100.0	100.0	100.0	70.0
Stack Bowls Two	95.0	100.0	100.0	100.0	85.0	100.0	100.0	100.0
Five tasks	82.3	91.8	98.2	87.1	90.0	93.0	95.6	78.4

Table 2: **The benchmark's own expert fails the same way.** Place Can Basket: failures that contain an ejection, and failures that pass the success check when replayed with the depenetration limit switched on before it. The scripted expert acts on ground truth object poses with a motion planner; RoboTwin runs it on every seed before scoring a policy and skips the seed when it fails.

Actor	Episodes	Failures	With an ejection	Pass with the limit
LingBot-VA (world action model)	90	23	21	18
HyVLA (vision language action)	20	4	1	1
Scripted expert (ground truth poses)	22	22	16	13

An artifact does not always end in failure: 17 successful LingBot-VA episodes of the five tasks contain one, 16 of them on Place Can Basket, where an ejected can may land back in the basket. Artifacts also act before the policy is scored: on Place Can Basket, 16 of the 22 seeds RoboTwin skipped because its own scripted demonstration failed contain one, and 13 pass with the limit (Table 2, Figure 3).

4.2 Cause: overlap removed within one physics step

At the onset of an artifact, PhysX's contact report places the object deep inside the collision geometry of its container. Over the 99 artifacts on Place Can Basket, the reported overlap between can and basket at the onset step is 21.5 mm in the median (interquartile range 11.6 to 27.9 mm); one step earlier the same contact pairs report contact without overlap in the five episodes we traced step by step. The can moves under 1 mm in that step, so the overlap is not motion but a change in the contact report: the basket wall is a thin convex piece, and once the contact feature switches from its inner face to its far face, a touching contact is reported as an overlap of the wall's thickness and beyond. PhysX removes it within the step: the can's velocity changes by 4.36 m/s, close to the 5.4 m/s at which 21.5 mm is removed in one step.

A stress test isolates how each engine removes such an overlap (Table 3a). We place RoboTwin's can in its basket without a robot, at 300 poses near the one recorded before the ejection of Figure 1 (Section D.1). With RoboTwin's settings, the overlap divided by the exit speed puts PhysX's removal at 2.6 ms, within one step, and throws the can in 91% of poses, a median 1.55 m away. Of the PhysX settings we tested, mass and solver iterations leave the rate at 90%, and only the maximum depenetration velocity, the parameter PhysX's documentation names for bodies that start in overlap (NVIDIA Corporation, 2024), matters: limited to 1 m/s, the removal stretches to 16.5 ms and the median distance falls to 10 cm; at 0.1 m/s no pose throws the can. MuJoCo removes the same overlaps over 38.6 ms and throws the can in 5% of poses.

Replays of recorded episodes need the second form of rerun, since a change from the first step alters outcomes that no artifact decided (Section D.2). Switched on 0.1 s before the first artifact, with everything before identical to the recording, only the depenetration limit keeps the object slow (Table 3b): the can's median peak speed after the switch is 1.39 m/s, against 3.72 m/s with a ten

Table 3: **Controlled reruns.** (a) RoboTwin's can at 300 resting poses inside its basket, in PhysX and in MuJoCo with identical assets and step (Section D.1). *Thrown*: at least 1 m/s and 4 cm away. *Removal*: initial overlap over exit speed. RoboTwin's values in parentheses. (b) LingBot-VA episodes with artifacts, replayed with one change from 0.1 s before the first artifact. *Rescued*: failures that pass, of the failures with an artifact. *Peak*: median object speed after the switch over all episodes with an artifact, successes included.

(a) Stress test, 300 poses

Engine and setting	Thrown (%)	Exit speed (m/s)	Distance (cm)	Removal (ms)
PhysX, RoboTwin settings	91	8.77	155	2.6
PhysX, can mass 100 g (10 g)	90	6.86	73	3.3
PhysX, solver iterations 32/8 (10/1)	90	16.78	200	1.4
PhysX, depenetration $\leq$ 1 m/s	65	1.36	10	16.5
PhysX, depenetration $\leq$ 0.1 m/s	0	0.14	3.7	169
MuJoCo, default (time constant 20 ms)	5	0.56	2.8	38.6
MuJoCo, time constant 8 ms	42	1.99	3.8	10.6

(b) Replays switched on 0.1 s before the first artifact

	Place Can Basket		Stack Bowls Three	
Change	Rescued	Peak (m/s)	Rescued	Peak (m/s)
Depenetration $\leq$ 1 m/s	18/21	1.39	7/12	0.65
Object mass 100 g	14/21	3.72	5/12	2.12
Solver iterations 32/8	12/21	11.20	4/12	7.00

times heavier can and 11.20 m/s with more solver iterations, and the two rescue fewer failures, 14 and 12 of 21 against 18.

4.3 What the artifacts do to the score

Every episode with an artifact is replayed with its recorded actions and the depenetration limit of 1 m/s switched on 0.1 s before the first artifact; the benchmark's success check scores the replay, and Equation (1) replaces the recorded outcome. All 53 replays on the five tasks match their recordings bit for bit up to the switch. Section F lists every replayed episode. Of LingBot-VA's 36 failures with an artifact on the five tasks, 28 pass; of its 17 successes with an artifact, 7 fail, all on Place Can Basket (Table 1). Its success rate on the five tasks moves from 82.3% to 91.8%. HyVLA gains 5 failures, loses 2 successes, and moves from 90.0% to 93.0% (difference $+3.0$ points, 95% interval $[-2.2, +8.2]$, sign test $p = 0.45$). Over all 50 tasks, counting the first pass of the other 40 (1080 episodes), LingBot-VA moves from 90.5% to 92.4% pooled over episodes ($+1.9$ points, $[+0.9, +3.0]$) and from 91.9% to 92.6% averaged over tasks; the five tasks, 220 of the 1080 episodes, were identified after the audit.

A benchmark could instead drop every confirmed flagged episode. That gives 98.2% for LingBot-VA and 95.6% for HyVLA on the five tasks (*Excl.*, Table 1). That the flagged episodes are as hard as the rest is untested. That the recorded actions remain the policy's choice is tested only indirectly: a policy that takes over from a replayed prefix avoids the artifact in 10 of 21 failures with the physics unchanged (Section D.2). That a limit kept on for whole rollouts changes nothing the task needs is refuted (*CL*): LingBot-VA reaches 87.1% on the five tasks against 81.8% as run on the same seeds, but Place Object Basket, with two artifacts in 20 episodes, falls from 90.0% to 50.0%. HyVLA falls from 90.7% on the shared seeds to 78.4%, to 70.0% on Place Bread Basket and 45.5% on Place Object Basket. We use the replay rate as the point estimate because it keeps the recorded actions and changes the physics only from 0.1 s before the artifact; it measures how sensitive the recorded execution is to that change.

4.4 Other platforms

SimuGuard also attaches to RoboDojo, LIBERO and RoboCasa, audited at a smaller scale (Table 4). RoboDojo, on PhysX, ejects objects as RoboTwin does: 7 of 40 XR-1 episodes contain

Table 4: **Four benchmarks, two engines.** Episodes audited, episodes with a contact artifact (failures and successes), the fastest object speed at an artifact, and the object's speed when the episode is replayed with the depenetration limit of 1 m/s switched on before the artifact (RoboTwin: median over Place Can Basket; RoboDojo: the tube of Figure 6). LIBERO and RoboCasa were not rerun.

Benchmark	Engine	Policy	Episodes	With artifact	Peak (m/s)	Limited (m/s)
RoboTwin 2.0	PhysX, SAPIEN	LingBot-VA	1880	70	21.8	1.39
		HyVLA	220	10	–	–
RoboDojo	PhysX, Isaac Sim	XR-1	40	7	6.30	1.88
LIBERO	MuJoCo	$\pi_{0.5}$	400	2	1.57	–
RoboCasa	MuJoCo	XR-1	300	5	1.19	–

artifacts, all failures, and the same limit slows a tube from 4.31 to 1.88 m/s. On MuJoCo, artifacts are few and slow: 2 of 400 LIBERO episodes and 5 of 300 RoboCasa episodes, none above 1.6 m/s.

4.5 Reliability of replay and detection

SimuGuard reproduces every recorded episode exactly, and only with the replay details of Section 3.4 (Table 8).

Detection agrees with human judgment: five robotics engineers rated each clip independently, without knowing where it came from, and settled disagreements by discussion (Table 8, Section G). On Place Can Basket, they judged 37 of 40 clips centered on detected events physically implausible (precision 0.925) and none of 20 hard negatives.

5 Discussion and Limitations

Discussion. The four scores of Table 1 answer four questions about the same episodes. The benchmark's own scripted expert, acting on ground truth poses, fails in the same way and is rescued by the same limit (Table 2), so the ejection belongs to the task and the solver, not to a learned policy. The reported overlap appears within one step at a resting contact with the thin wall; its one step removal is the artifact because of what it produces: a can drifting at 0.2 m/s gains 2.4 J while no robot link touches it (Figure 1). A limit kept on for whole rollouts is not a fix, since it changes the task. As reported, HyVLA leads LingBot-VA on the five tasks, 90.0 against 82.3; under exclusion the order reverses, 95.6 against 98.2, untested for significance.

Limitations. The main experiments are on RoboTwin; the three other benchmarks were audited at a smaller scale, without reruns on MuJoCo, and a full evaluation there is next. The mechanism, the intervention and the detector need tightening: the contact feature switch behind the reported overlap is not traced to a piece pair, the limit acts on every body after the switch, the negative control on episodes without an artifact and a limit kept only near the artifact are not run, and the detector's precision was measured on one task, with human review as the final standard, which we will extend to held out tasks. The four scores are references, with the replay rate as our point estimate; how close it comes to the policy's true ability is for extended experiments to decide.

6 Conclusion

Success rate in simulation is read as a policy's ability to act, but it also counts the failures the simulator causes. On tasks that place objects into containers with thin walls, most failures of two policies and of the benchmark's own scripted demonstrations contain a contact artifact. Controlled reruns trace the artifact to how fast the solver removes an overlap, and the smaller audits of three further benchmarks show the same pattern. Replaying the recorded actions with the removal slowed lets most of the flagged failures pass, moving LingBot-VA's rate on the five tasks from 82.3% to 91.8%, but it is not a fix: kept on for whole rollouts, the limit changes the task. We therefore report the interval the artifacts leave open, use the replay rate as the point estimate, and leave to extended experiments how close it comes to a policy's true ability. An audit of this kind lets a benchmark separate what the policy did from what the solver did.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic data, to develop theoretical models, to formulate or prove mathematical claims, or to propose hypotheses, and translation, dataset cleaning and qualitative analysis are not applicable to this work. We used generative AI tools to create and edit software code, to draft and edit parts of the paper, to identify relevant literature and to format references. We have reviewed all AI-assisted work: every number in the paper comes from runs the authors executed and can be regenerated by exact replay. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the validity of simulated robot benchmarks. It involves no personal data; the human judgments in Section G are annotations of simulation videos by five robotics engineers.

REPRODUCIBILITY STATEMENT

SIMUGUARD runs the benchmark's own evaluation script, and every episode in this paper is stored with its seed and the actuation of every physics step, so each result can be regenerated by exact replay. Section B lists tasks, episodes, upstream versions and checkpoints; Section B.5 lists every detector parameter; Section C.1 describes the adapters and the replay; Section D.1 specifies the stress test. The code, the detector, the replay and rerun tools, the episode manifest of Section F and the scripts that produce each table are at `https://anonymous.4open.science/r/simuguard-B109`; the recordings will be released with the code.

REFERENCES

Brian Acosta, William Yang, and Michael Posa. Validating robotics simulators on Real-World impacts. *IEEE Robotics and Automation Letters*, 2022.

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

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, et al. RoboArena: Distributed Real-World evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.

Johan Björck, Fernando Castañeda, Nikita Cherniadev, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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 Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

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. *arXiv preprint arXiv:2307.15818*, 2023a.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, et al. RT-1: Robotics transformer for Real-World control at scale. In *Robotics: Science and Systems (RSS)*, 2023b.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Alejandro Castro, Frank Permenter, and Xuchen Han. An unconstrained convex formulation of compliant contact. *IEEE Transactions on Robotics*, 2023.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, et al. GR-2: A generative Video-Language-Action model with Web-Scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, Honghao Su, Zhiyang Dou, Kaixuan Wang, Dandan Zhang, Yunze Liu, Yan Qin, Qiwei Liang, Qiwei Wu, Zijian Lin, Wenwei Lin, Yuran Wang, Minghua He, Tianshu Wu, Ruihai Wu, Jingquan Zhou, Kai-Chong Lei, Haibao Yu, Yuanfeng Ji, Weiyang Jin, Guanyu Lin, Xiaofan Li, Qi Xiong, Renjing Xu, Zhongyu Li, Wenhao Chai, Enze Xie, Ziwei Wang, Yao Mu, Hao Dong, Wojciech Matusik, Mingyu Ding, Wenbo Ding, Ping Luo, and Masayoshi Tomizuka. RoboDojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026.

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

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, et al. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, 2023.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via Text-Guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ryan Elandt, Evan Drumwright, Michael Sherman, and Andy Ruina. A pressure field model for fast, robust approximation of net contact force and moment between nominally rigid objects. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

Tom Erez, Yuval Tassa, and Emanuel Todorov. Simulation tools for model-based robotics: Comparison of Bullet, Havok, MuJoCo, ODE and PhysX. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

Jialiang Fan, Weizhe Xu, Oleg Sokolsky, Insup Lee, and Fanxin Kong. SafeVLA-Bench: A benchmark for the success-safety gap in vision-language-action models. *arXiv preprint arXiv:2606.00773*, 2026.

Shichao Fan, Kun Wu, Zhengping Che, Xinhua Wang, Di Wu, Fei Liao, Ning Liu, Yixue Zhang, Zhen Zhao, Zhiyuan Xu, Meng Li, Qingjie Liu, Shanghang Zhang, Min Wan, and Jian Tang. XR-1: Towards versatile vision-language-action models via learning unified vision-motion representations. *arXiv preprint arXiv:2511.02776*, 2025.

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.

Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable Vision-Language robotic manipulation: A benchmark and LLM-guided 3d policy. *arXiv preprint arXiv:2410.01345*, 2024.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. *arXiv preprint arXiv:2302.04659*, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *Nature*, 2025.

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

Taylor A. Howell, Simon Le Cleac'h, Jan Brüdigam, Qianzhong Chen, Jiankai Sun, J. Zico Kolter, Mac Schwager, and Zachary Manchester. Dojo: A differentiable physics engine for robotics. In *Robotics: Science and Systems (RSS)*, 2022.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, et al. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning (ICML)*, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

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

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, et al. VIMA: General robot manipulation with multimodal prompts. In *International Conference on Machine Learning (ICML)*, 2023.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *Conference on Robot Learning (CoRL)*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, et al. DROID: A Large-Scale In-The-Wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action models: Optimizing speed and success. In *Robotics: Science and Systems (RSS)*, 2025.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Quentin Le Lidec, Wilson Jallet, Louis Montaut, Ivan Laptev, Cordelia Schmid, and Justin Carpentier. Contact models in robotics: a comparative analysis. *arXiv preprint arXiv:2304.06372*, 2023.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, et al. igibson 2.0: Object-Centric simulation for robot learning of everyday household tasks. In *Conference on Robot Learning (CoRL)*, 2021.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, et al. BEHAVIOR-1K: A Human-Centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning (CoRL)*, 2022.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998*, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Robotics: Science and Systems (RSS)*, 2025.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, et al. Vision-Language foundation models as effective robot imitators. In *International Conference on Learning Representations (ICLR)*, 2024a.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating Real-World Robot Manipulation Policies in Simulation. *arXiv preprint arXiv:2405.05941*, 2024b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. *arXiv preprint arXiv:2306.03310*, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, et al. RDT-1B: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, et al. Isaac gym: High performance GPU-Based physics simulation for robot learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (NeurIPS)*, 2021.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for Language-Conditioned policy learning for Long-Horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, et al. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, et al. ManiSkill: Generalizable manipulation skill benchmark with Large-Scale demonstrations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (NeurIPS)*, 2021.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins. *arXiv preprint arXiv:2504.13059*, 2025.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

NVIDIA Corporation. PhysX SDK 5.4.1 documentation: Best practices guide. `https://nvidia-omniverse.github.io/PhysX/physx/5.4.1/docs/BestPractices.html`, 2024. Section "Overlapping objects explode".

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, et al. Octo: An Open-Source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, et al. FAST: Efficient action tokenization for Vision-Language-Action models. *arXiv preprint arXiv:2501.09747*, 2025.

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

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. THE COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation. In *Robotics: Science and Systems (RSS)*, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, et al. SpatialVLA: Exploring spatial representations for Visual-Language-Action model. In *Robotics: Science and Systems (RSS)*, 2025.

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

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. *arXiv preprint arXiv:2410.00425*, 2024.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Joseph A. Vincent, Haruki Nishimura, Masha Itkina, Paarth Shah, Mac Schwager, and Thomas Kollar. How generalizable is my behavior cloning policy? a statistical approach to trustworthy performance evaluation. *arXiv preprint arXiv:2405.05439*, 2024.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, et al. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Yi Ru Wang, Carter Ung, Christopher Tan, Grant Tannert, Jiafei Duan, Josephine Li, Anh Le, Rishabh Oswal, Markus Grotz, Wilbert Pumacay, Yuquan Deng, Ranjay Krishna, Dieter Fox, and Siddhartha Srinivasa. RoboEval: Where robotic manipulation meets structured and scalable evaluation. *arXiv preprint arXiv:2507.00435*, 2025.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *arXiv preprint arXiv:2205.02961*, 2022.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, et al. Unleashing Large-Scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning (CoRL)*, 2022.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-based Interactive ENvironment. *arXiv preprint arXiv:2003.08515*, 2020.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive Real-World simulators. In *International Conference on Learning Representations (ICLR)*, 2024.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, et al. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, Quanlu Zhang, Yongji Wu, Chunyang Zhu, Junhao Hu, Zixiao Huang, Mingjie Wei, Yuqing Xie, Ke Yang, Bo Dai, Zhexuan Xu, Jiakun Du, Xiangyuan Wang, Xu Fu, Letong Shi, Zhihao Liu, Kang Chen, Weilin Liu, Gang Liu, Boxun Li, Jianlei Yang, Zhi Yang, Guohao Dai, and Yu Wang. RLinf: Flexible and efficient large-scale reinforcement learning via macro-to-micro flow transformation. *arXiv preprint arXiv:2509.15965*, 2025.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, et al. Meta-World: A benchmark and evaluation for Multi-Task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.

He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, Wanjia He, Ling Chen, Kai Huang, Jiahao Chen, Sichang Su, Xumin Yu, Ziyi Wang, Chengwei Zhu, Xiao Teng, Yuchun Guo, Yufeng Zhang, Yuandong Liu, Rui Wang, Zisheng Lu, Han Hu, and Zhengyou Zhang. Hy-Embodied-0.5-VLA: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained bimanual manipulation with Low-Cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances Spatial-Temporal awareness for generalist robotic policies. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, et al. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

A Extended Related Work

Policies. Language conditioned manipulation moved from grounding language in affordances (Ahn et al., 2022) and multimodal prompts (Jiang et al., 2023) to transformers trained on robot data at scale (Brohan et al., 2023b; Open X-Embodiment Collaboration et al., 2024; Octo Model Team et al., 2024) and to VLAs that start from a vision language backbone (Driess et al., 2023; Brohan et al., 2023a; Li et al., 2024a; Kim et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025; Björck et al., 2025; Gemini Robotics Team, 2025). Later work changes the action head and its training (Chi et al., 2023; Zhao et al., 2023; Ke et al., 2024; Liu et al., 2025; Pertsch et al., 2025; Kim et al., 2025) or the spatial input (Zheng et al., 2025; Qu et al., 2025), and large real robot datasets support this training (Walke et al., 2023; Khazatsky et al., 2024). World action models add a generative model of future observations: video prediction as a planner (Du et al., 2023; Ko et al., 2024), video generative pretraining (Wu et al., 2024; Cheang et al., 2024; Hu et al., 2025), a single model for video and actions (Li et al., 2025; 2026), and latent actions learned from video (Ye et al., 2025). They draw on world models trained for control (Wu et al., 2022; Hafner et al., 2025) and on interactive video models (Yang et al., 2024; Bruce et al., 2024). Every one of these policies is scored in simulation by the outcome of an episode; SimuGuard audits that outcome.

Simulated benchmarks. Simulated suites for manipulation span RLBench (James et al., 2020), Meta-World (Yu et al., 2019), robosuite (Zhu et al., 2020), CALVIN (Mees et al., 2022), ManiSkill (Mu et al., 2021; Gu et al., 2023; Tao et al., 2024), LIBERO (Liu et al., 2023), RoboCasa (Nasiriany et al., 2024), RoboTwin (Mu et al., 2025; Chen et al., 2025) and RoboDojo (Chen et al., 2026), and household suites such as iGibson (Li et al., 2021), Habitat (Szot et al., 2021) and BEHAVIOR (Li et al., 2022); GPU physics (Makoviychuk et al., 2021; Mittal et al., 2023) makes large evaluations cheap. Generalization suites vary objects, layouts, lighting and instructions (Pumacay et al., 2024; Garcia et al., 2024; Zhou et al., 2025; Fei et al., 2025), SimplerEnv (Li et al., 2024b) and RoboArena (Atreya et al., 2025) ask whether simulated or crowd sourced rankings agree with real robots, and Acosta et al. (2022) validate engines on recorded impacts. All of them treat the physics engine as part of the fixed apparatus.

Evaluation practice. Guidance on evaluating robot policies concerns the number of trials, the choice of conditions and the reporting of uncertainty (Kress-Gazit et al., 2024; Vincent et al., 2024; Snyder et al., 2026); RoboEval (Wang et al., 2025) and Jiang et al. (2026) examine what the scores of manipulation benchmarks measure; and the machine learning literature documents how seeds and variance distort comparisons (Henderson et al., 2018; Agarwal et al., 2021; Bouthillier et al., 2021). These works improve how outcomes are aggregated. SimuGuard asks whether an individual outcome is the policy's.

Contact simulation. Rigid body engines differ in how they model contact: soft constraints with a time constant in MuJoCo (Todorov et al., 2012), hard constraints with penetration correction in PhysX, compliant and pressure field models (Castro et al., 2023; Elandt et al., 2019), and differentiable formulations (Howell et al., 2022). Comparisons across engines (Erez et al., 2015; Le Lidec et al., 2023) and against real impacts (Acosta et al., 2022) report that the same scene produces different forces and trajectories, and convex decomposition of assets adds another approximation (Wei et al., 2022). SimuGuard measures what one such difference, the speed at which PhysX removes an overlap, does to the score of a benchmark.

B Setup

B.1 RoboTwin

We use RoboTwin commit 6dde571, XPolicyLab commit fa431ec, LingBot-VA code 7c6ffa9, and the RoboTwin checkpoint of LingBot-VA from ModelScope, driven through its end effector action path. Inference runs on RTX 4090 or H800 GPUs; the copies of the weights are byte identical (sha256 over all 16 files), and launch settings are the same everywhere. RoboTwin runs its scripted demonstration on each seed in turn and skips the seed if the demonstration fails its check, until N seeds are scored. Every run uses RoboTwin's *demo_clean* task configuration, seen instructions, seed

base 0 and the expert check. The protocol is 40 episodes per task: the 40 tasks outside the ten of the cross task study were evaluated in two passes of 20 seeds from the same seed base, the second on H800 GPUs with RoboTwin commit 2a14726 and XPolicyLab fe71eb5, the same checkpoint and the same evaluator arguments; the two passes score 675 seeds in common and agree on the outcome of 629 of them. Only the first pass was replayed. Of the ten cross task tasks, Place Can Basket has 90 episodes and Stack Bowls Three 50; Place Bread Basket has 40, in two passes of 20 seeds from the same seed base; Place Object Basket, Stack Bowls Two and Dump Bin Bigbin have 20, and Place Cans Plasticbox, Put Bottles Dustbin, Put Object Cabinet and Stack Blocks Two have 10; their extension to 40 is in progress. Place Can Basket was evaluated three times with 30 scored seeds each, because a single repetition is not stable: 15 of the 37 seeds involved succeed in some repetitions and fail in others (23, 24 and 20 successes, $74.4 \pm 6.9\%$). Table 5 lists every task.

Table 5: **RoboTwin 2.0 runs** of LingBot-VA. *Artifact*: episodes with a contact artifact among failures and successes. *Skipped*: seeds discarded because the scripted demonstration failed. *Succ.*: successes; *Rep.*: reported rate. Bold: the five tasks of Table 1. The 40 tasks outside the ten of the cross task study were run twice with 20 seeds each, and both passes are pooled.

				Artifact in								Artifact in		
Task	N	Succ.	Rep. (%)	fail.	succ.	Skipped	Task	N	Succ.	Rep. (%)	fail.	succ.	Skipped
Adjust Bottle	40	40	100.0	0	0	11	**Place Can Basket**	3×30	67	74.4	21	16	22
Beat Block Hammer	40	39	97.5	0	0	19	Place Cans Plasticbox	10	10	100.0	0	0	0
Blocks Ranking RGB	40	39	97.5	0	0	0	Place Container Plate	40	39	97.5	0	0	9
Blocks Ranking Size	40	36	90.0	0	0	3	Place Dual Shoes	40	37	92.5	3	2	27
Click Alarmclock	40	40	100.0	0	0	8	Place Empty Cup	40	40	100.0	0	0	4
Click Bell	40	40	100.0	0	0	0	Place Fan	40	36	90.0	1	0	8
Dump Bin Bigbin	20	20	100.0	0	0	28	Place Mouse Pad	40	39	97.5	0	0	0
Grab Roller	40	40	100.0	0	0	2	**Place Object Basket**	20	18	90.0	1	1	8
Handover Block	40	39	97.5	0	0	12	Place Object Scale	40	36	90.0	0	0	16
Handover Mic	40	38	95.0	0	0	6	Place Object Stand	40	40	100.0	0	0	1
Hanging Mug	40	9	22.5	1	0	7	Place Phone Stand	40	39	97.5	0	0	20
Lift Pot	40	40	100.0	0	0	56	Place Shoe	40	39	97.5	0	0	3
Move Can Pot	40	37	92.5	0	0	4	Press Stapler	40	33	82.5	0	0	4
Move Pillbottle Pad	40	40	100.0	0	0	21	Put Bottles Dustbin	10	9	90.0	0	0	4
Move Playingcard Away	40	40	100.0	0	0	0	Put Object Cabinet	10	7	70.0	0	0	73
Move Stapler Pad	40	27	67.5	2	0	2	Rotate QRcode	40	37	92.5	0	0	3
Open Laptop	40	40	100.0	0	0	0	Scan Object	40	38	95.0	0	0	80
Open Microwave	40	31	77.5	0	0	32	Shake Bottle	40	40	100.0	0	0	16
Pick Diverse Bottles	40	38	95.0	0	0	46	Shake Bottle Horizontally	40	40	100.0	0	0	16
Pick Dual Bottles	40	40	100.0	0	0	2	Stack Blocks Three	40	35	87.5	2	0	3
Place A2B Left	40	40	100.0	0	0	9	Stack Blocks Two	10	10	100.0	0	0	0
Place A2B Right	40	37	92.5	0	0	9	**Stack Bowls Three**	50	38	76.0	12	0	43
Place Bread Basket	40	39	97.5	1	0	9	**Stack Bowls Two**	20	19	95.0	1	0	7
Place Bread Skillet	40	39	97.5	0	0	56	Stamp Seal	40	38	95.0	1	0	18
Place Burger Fries	40	40	100.0	0	5	0	Turn Switch	40	21	52.5	0	0	7

All 50 tasks: N = 1880, successes 1713 (91.1%), artifacts in 46 failures and 24 successes, 734 seeds skipped.

HyVLA. We use the RoboTwin checkpoint Hy-Embodied-0.5-VLA-RoboTwin through XPolicyLab, with the deployment settings of its own RoboTwin evaluation (end effector actions, seven actions executed per query), and the same evaluator arguments as for LingBot-VA (seed base, seen instructions, scripted demonstration check), so both policies are scored from the same seed sequence. Place Object Basket ran as five parallel shards that start at different seeds. Because RoboTwin's scripted demonstration can pass on a seed in one run and fail on it in another, two seeds were scored by two shards; we keep one run per seed, which leaves 20 distinct seeds (Table 6).

Table 6: **RoboTwin 2.0 runs** of HyVLA. *Artifact*: failures that contain a contact artifact. *Skipped*: seeds discarded because the scripted demonstration failed.

Task	N	Successes	Reported (%)	Artifact	Skipped
Place Can Basket	20	16	80.0	1/4	5
Stack Bowls Three	20	18	90.0	2/2	18
Place Object Basket	20	19	95.0	0/1	8
Place Bread Basket	20	20	100.0	0/0	5
Stack Bowls Two	20	17	85.0	3/3	7
Blocks Ranking Size	20	17	85.0	0/3	2
Open Microwave	20	9	45.0	0/11	2
Put Bottles Dustbin	20	15	75.0	0/5	10
Put Object Cabinet	20	16	80.0	0/4	156
Stamp Seal	20	15	75.0	0/5	9
Turn Switch	20	10	50.0	0/10	3

B.2 LIBERO

We use upstream LIBERO with robosuite 1.4.0 and MuJoCo 3.2.3, a 2 ms physics step and 25 physics steps per action. The policy is the $\pi_{0.5}$ checkpoint RLinf-Pi05-LIBERO-130-fullshot-SFT (Yu et al., 2025), evaluated with the protocol it was trained for: two camera images and the robot state as input, five actions per query, 15 steps with the arm at rest before the policy acts, at most 600 policy steps, and a stop at the first success. We run the initial states 0 to 9 of all 40 tasks, with the seed equal to the index of the initial state. The policy succeeds in 381 of 400 episodes (95.2%): 99 on LIBERO-Spatial, 97 on LIBERO-Object, 98 on LIBERO-Goal and 87 on LIBERO-10.

B.3 RoboCasa

We use RoboCasa (Nasiriany et al., 2024) on MuJoCo 3.3.1 with robosuite's split physics step, a 2 ms physics step and 25 physics steps per action, and XR-1 (Fan et al., 2025) with its RoboCasa checkpoint and evaluation protocol on ten object placement tasks, ten episodes each: Arrange Bread Basket, Coffee Setup Mug, Pick Place Counter To Cabinet, Pick Place Counter To Stove, Pick Place Drawer To Counter, Pick Place Sink To Counter, Pick Place Toaster To Counter, Stack Bowls Cabinet, Store Leftovers In Bowl and Wash Fruit Colander, and on the other 40 tasks of XR-1's evaluation set, five episodes each. The set holds 18 atomic tasks and 32 composite tasks, 16 of them unseen in XR-1's training; the ten placement tasks above are 4 atomic and 6 composite. XR-1 succeeds in 72 of the first 100 episodes and in 113 of the other 200.

B.4 RoboDojo

RoboDojo (Chen et al., 2026) runs Isaac Sim 5.1.0 and Isaac Lab 0.54.3 with two ARX X5 arms, a 4 ms physics step and 10 physics steps per action; RoboDojo switches PhysX to GPU dynamics. We run XR-1 with its official RoboDojo checkpoint on four tasks, Fill Pen Holder (three layouts), Insert Tubes (three), Put Bottles Into Dustbin (two) and Store Tools In Toolbox (two), under the three evaluation seeds that ship with the benchmark. Seed 0 was run twice, first with every episode replayed and then, together with seeds 1 and 2, without replay. XR-1 succeeds in 16 of the 40 episodes.

B.5 DETECTOR PARAMETERS

Table 7: **Detector parameters** (Section 3.3), identical for all tasks and engines. The thresholds on velocity change are given for a 4 ms step and scale with the physics step on other engines.

Stage	Parameter	Value
Contact ejection	recent contact window	40 ms
	minimum speed	0.5 m/s
	minimum velocity change in one 4 ms step	0.35 m/s
	confirmation window	240 ms
	minimum displacement	4 cm
	free flight	$\geq$ 16 ms
	or violent contact with the container: velocity change / speed	$\geq$ 0.75 / 1.0 m/s
Gravity	window before onset for the drop height Δz	0.6 s
	window after onset for the peak speed	0.1 s
	ratio of peak speed to $\sqrt{2g\,\Delta z}$	> 1.5
	speed floor in the denominator	0.05 m/s
Carrier	window before onset for carriers	40 ms
	window after onset for the peak speed	0.1 s
	ratio of peak speed to the fastest carrier's speed	≥ 1.5
	speed floor in the denominator	0.05 m/s

C RELIABILITY OF THE TOOL

C.1 RECORDING, REPLAY AND RERUNS

Recording. SIMUGUARD wraps the physics step of the running scene, which leaves the evaluator, the benchmark code and the policy untouched. In SAPIEN, it records the drive targets, drive velocity targets and joint forces of every articulation before each step, 85 to 105 bytes per step; monitoring costs 7.5 to 11.6 ms per step on top of about 0.5 ms of physics. In MuJoCo, it records the control inputs, applied forces and mocap poses, the solver's warm start whenever benchmark code changed it since the last step, and every write of positions, velocities or actuator states between two steps. For LIBERO, this takes 355 bytes per step for the actuation and 304 bytes for the object states, and monitoring adds 2.8 ms per step, which lengthens the wall time by 23% with cameras rendered on the CPU. In Isaac Sim, it records the drive position and velocity targets and joint forces of every articulation, reads states from PhysX tensor views and contacts from the PhysX contact report, which RoboDojo's object loaders are patched to enable; monitoring costs 6.9 to 18.4 ms per step, 3 to 7% of an XR-1 episode's wall time. Body velocities are those of the centers of mass.

Replay. A replay rebuilds the scene from its seed in a fresh process and applies the recorded actuation from the first physics step. In RoboTwin, it also restores the commanded gripper opening at each step from the gripper drive targets with RoboTwin's own mapping, because the success checks read it. In MuJoCo, it writes back the recorded warm start and the recorded writes between steps, and compares the full integration state with snapshots archived every 250 steps. On robosuite 1.4.0, LIBERO places fixtures with an unseeded random generator, so a scene is not reproducible from its seed; SIMUGUARD then writes the recorded body poses back before the replay. In Isaac Sim, restoring a saved state in place leaves trajectories 0.2 to 1 mm apart after 500 steps because PhysX's contact caches cannot be restored, while rebuilding the scene and replaying the recorded actuation is exact.

Reruns. The depenetration limit is set on every dynamic body and articulation link of the scene and stays on until the episode ends. A rerun from before the artifact switches it on at the physics step 0.1 s (25 steps) before the onset of the episode's first contact artifact and compares every recorded state before that step with the recording.

Replay checks, detection checks and recording cost. Table 8 collects every replay check we ran, the detection checks against human judgment, and the cost of recording on each engine. Restoring

a saved state in place is not exact on either PhysX platform; rebuilding the scene and replaying the recorded actuation is.

Table 8: **Reliability of SIMUGUARD and cost of recording.** *Exact*: every recorded float of every tracked body matches the recording.

Check	Platform	Result
Replay		
Fresh process, 96 failures (2.42 M steps); events and outcomes reproduced	RoboTwin	96/96 exact; 43/43, 96/96
Stack Bowls Three, 12 failures: outcomes with / without gripper restore	RoboTwin	12/12 / 7/12
Another machine, 25 episodes	RoboTwin	25/25 exact
Prefix before a rerun switch, 273 replays	RoboTwin	273/273 exact
Full episodes, 400 (1.88 M steps); without the recorded warm start	LIBERO	400/400 exact; 0/33
Rebuilt scene, 13 episodes	RoboDojo	13/13 exact
Monitor on versus off	RoboTwin, RoboDojo	59 and 26 bodies identical
State restored in place during contact	SAPIEN / Isaac Sim	mm to cm / 0.2–1 mm apart
Detection		
Detected events judged implausible (Place Can Basket)	RoboTwin	37/40
Hard negatives judged implausible (Place Can Basket)	RoboTwin	0/20
Agreement on 60 episodes of six tasks, with / without the gravity stage	RoboTwin	58/60 / 39/60
Recording cost per physics step	SAPIEN	7.5–11.6 ms, 85–105 bytes
	MuJoCo (LIBERO)	2.8 ms, 659 bytes, +23% wall time
	Isaac Sim	6.9–18.4 ms, 3–7% of episode wall time

C.2 WHAT EACH DETECTION STAGE REMOVES

Table 9 counts, per platform, the events that the first stage flags, those that survive the gravity stage, and those that survive the carrier stage. The gravity stage does most of the work where objects are dropped: on RoboTwin's Put Bottles Dustbin (10 episodes) it removes all 20 flagged drops, and on the 60 human-labeled episodes of six further RoboTwin tasks it raises agreement with the annotators from 39 to 58 (Table 8). The carrier stage changes no label on RoboTwin and RoboDojo and removes none of RoboCasa's remaining events.

Table 9: **Detection stages across platforms.** Events flagged by the contact ejection stage, and those left after the gravity and carrier stages; episodes in parentheses where counted.

Platform	Episodes	Flagged	After gravity	After carrier
RoboTwin, Put Bottles Dustbin (LingBot-VA)	10	20	0	0
RoboTwin, 40 tasks, second pass (LingBot-VA)	800	12 (12)	10 (10)	10 (10)
RoboCasa, ten placement tasks (XR-1)	100	7 (7)	3 (3)	3 (3)
RoboCasa, the other 40 tasks (XR-1)	200	15 (15)	2 (2)	2 (2)
RoboDojo, four tasks (XR-1)	40	25 (15)	12 (7)	12 (7)
LIBERO, stored initial states, no policy	2000	655	650	650

D ROBOTWIN: STRESS TEST AND RERUNS

D.1 STRESS TEST

The scene contains RoboTwin's can (model 1 of object 071, 22 convex pieces, 10 g) and basket (model 1 of object 110, 54 convex pieces, 0.5 kg) on a table, with the same collision pieces, masses, inertias and center of mass frames in SAPIEN 3.0.0b1 with PhysX and in MuJoCo 3.3.1, gravity on and a 4 ms step. The basket takes its pose from the episode of Figure 1 one physics step before the ejection. Each of the 300 can poses has a uniformly random orientation and a position within 3 cm of the recorded one horizontally and from 3 cm below to 1 cm above it vertically; the can starts at rest, and 288 poses start at least 5 mm deep in the basket. Each pose runs for 0.5 s.

The response depends on depth. With the can pushed straight down into the basket floor, PhysX with RoboTwin's settings releases it at 0.04 m/s from depths up to 10 mm and at 4.20 m/s from 20 mm (peak speed within 0.1 s); with the depenetration limit at 1 and 0.1 m/s, the 20 mm case leaves at 1.24 and 0.14 m/s, and with 32 position iterations at 11.92 m/s. MuJoCo with its default contact releases the can at 0.24 m/s from 10 mm and at 1.05 m/s from 20 mm. MuJoCo does not accept a time constant below two physics steps; with this safeguard disabled, time constants of 4 or 1 ms make the same scene diverge. The recorded poses of can and basket one step before the ejection, placed in this scene without the robot, show no overlap and do not reproduce the ejection (0.23 m/s), consistent with the recording, in which the overlap is reported only at the onset step; the stress test therefore starts from overlaps of the depth reported at onset rather than from the recorded state.

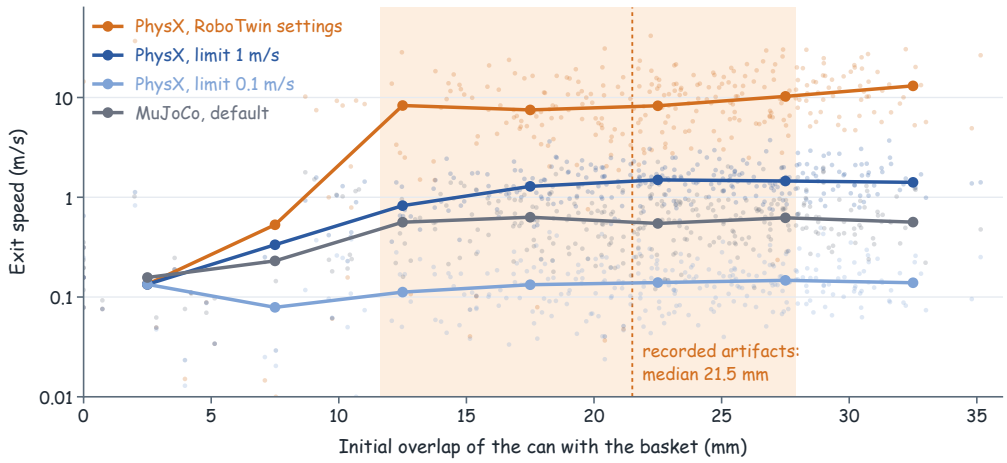

Figure 4: **Exit speed against initial overlap** in the stress test of Table 3a: RoboTwin's can at 300 resting poses inside its basket, without a robot, in PhysX with RoboTwin's settings, with the depenetration limit at 1 and 0.1 m/s, and in MuJoCo with its default contact. Dots are poses; lines are medians in 5 mm bins. With RoboTwin's settings the median exit speed is 0.13 m/s below 10 mm of overlap and 9.4 m/s at or above it; the limits and MuJoCo stay between 0.14 and 1.5 m/s at every depth. The band marks the overlap at the onset of the 99 recorded artifacts on Place Can Basket, median 21.5 mm and interquartile range 11.6 to 27.9 mm, above the depth at which PhysX starts to eject.

D.2 RERUNS

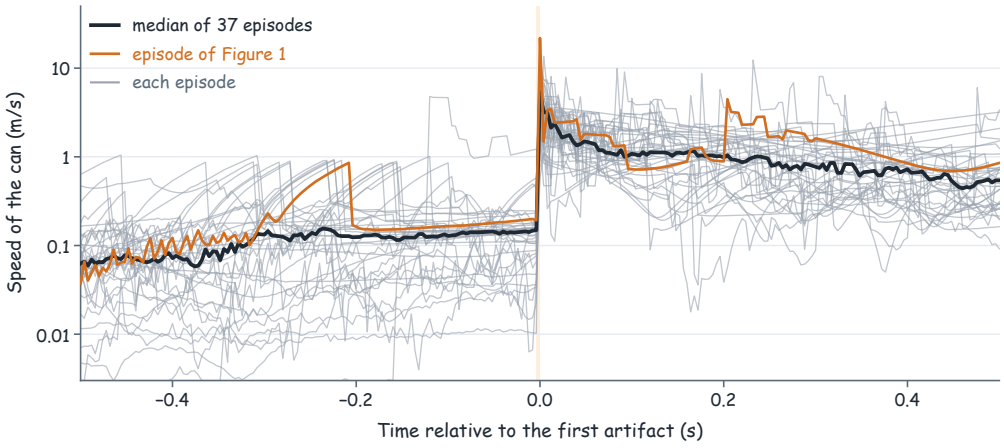

Figure 5: **The can's speed around the first artifact** in all 37 Place Can Basket episodes of LingBot-VA that contain one (21 failures, 16 successes), at every 4 ms step from 0.5 s before to 0.5 s after the onset. Gray: episodes; bold: median; orange: the episode of Figure 1. One step before the onset the can moves at a median 0.15 m/s. The largest one step rise in speed falls on the onset step in 33 episodes and within four steps in the other 4; it ranges from 0.53 to 21.6 m/s, median 5.7 m/s, more than an order of magnitude in the median.

Changes from the first physics step. Applied from the first step, each change of a simulation parameter also rescues Place Can Basket failures that contain no artifact (Table 10), so these replays do not separate the parameters. A clamp that caps the object's speed at 2 m/s after each step, which is not a simulation parameter, rescues none of them.

Table 10: **Replays from the first physics step** of the recorded failures: failures that pass the success check.

	Place Can Basket failures		Stack Bowls Three
Change	with artifact (21)	without artifact (2)	failures (12)
Depenetration $\leq$ 1 m/s	19	2	8
Depenetration $\leq$ 0.1 m/s	9	2	5
Object mass 100 g	17	2	6
Solver iterations 32/8	14	1	5
Speed clamp 2 m/s	14	0	0

Limit value. Switched on 0.1 s before the first artifact, limits from 0.1 to 1 m/s rescue nearly the same failures, and a limit of 2 m/s rescues fewer while the object leaves faster (Table 11). The peak speeds here are medians over the 21 failures; Table 3b takes the median over all 37 Place Can Basket episodes with an artifact, successes included, which is why it reads 1.39 m/s against 1.25 m/s here. All 108 replays of this sweep match their recordings up to the switch.

Table 11: **Limit value.** LingBot-VA failures with artifacts that pass when replayed with the depenetration limit switched on 0.1 s before the first artifact, and the median peak speed of the object after the switch over these failures. *Three tasks*: Place Object Basket, Place Bread Basket and Stack Bowls Two, one failure each. The successes with an artifact were replayed at 1 m/s only.

	Place Can Basket (21)		Stack Bowls Three (12)		Three tasks (3)	Five tasks
Limit	Rescued	Peak (m/s)	Rescued	Peak (m/s)	Rescued	Rescued
0.1 m/s	18	0.91	7	0.40	3	28/36
0.5 m/s	18	1.05	7	0.63	2	27/36
1 m/s	18	1.25	7	0.65	3	28/36
2 m/s	16	1.95	4	1.09	3	23/36

Successes with an artifact. The same replay is run for every success that contains an artifact; Table 16 lists how many keep their success per task and policy.

Closed loop evaluation with the limit on. We also evaluate LingBot-VA again with the depenetration limit of 1 m/s on for the whole policy rollout, with the scripted demonstration unchanged (Table 12). On Place Can Basket it succeeds in 81 of 90 episodes over three repetitions, against 67 reported; Stack Bowls Three and Stack Bowls Two rise to 20 of 20, Place Bread Basket falls from 19 to 17, and Place Object Basket from 18 to 10, with artifacts still present in 2 of its 20 episodes. The tasks without container artifacts move in both directions, from 15 to 10 of 20 on Open Microwave, from 7 to 5 of 10 on Put Object Cabinet, and from 17 to 18 on Blocks Ranking Size. The policy samples its actions without a fixed seed, so each closed loop episode is a new rollout, and 20 episodes resolve such differences only coarsely. Put Bottles Dustbin was not completed. The as run rate on the shared seeds counts each seed once with its recorded outcome. HyVLA, run the same way on the five container tasks, keeps 16 of 20 on Place Can Basket, rises to 20 of 20 on both bowl tasks, and falls from 20 to 14 of 20 on Place Bread Basket and from 19 of 20 to 10 of 22 on Place Object Basket (18 seeds, four of them run twice); 7 of its 102 episodes still contain an artifact.

Table 12: **Closed loop evaluation with the limit on.** Successes as reported and with the depenetration limit of 1 m/s on for the whole policy rollout. *The first 20 scored seeds of the reported run of 50.

Task	Reported	Limit on
LingBot-VA		
Place Can Basket	67/90	81/90
Stack Bowls Three	16/20*	20/20
Stack Bowls Two	19/20	20/20
Place Bread Basket	19/20	17/20
Place Object Basket	18/20	10/20
Stamp Seal	19/20	19/20
Blocks Ranking Size	17/20	18/20
Open Microwave	15/20	10/20
Turn Switch	9/20	9/20
Put Object Cabinet	7/10	5/10
Place Burger Fries	20/20	20/20
HyVLA		
Place Can Basket	16/20	16/20
Stack Bowls Three	18/20	20/20
Stack Bowls Two	17/20	20/20
Place Bread Basket	20/20	14/20
Place Object Basket	19/20	10/22

Closed loop evaluation with changed physics. Changing one parameter for the whole closed loop evaluation also changes the reported rate (Table 13). With a can of 100 g, LingBot-VA succeeds on Place Can Basket in 95.6% of episodes against 74.4% with RoboTwin's 10 g can (paired over seeds,

$p = 0.0015$; 95% interval of the difference $[+10.3, +32.8]$ points). More solver iterations give 86.7% (Wilcoxon signed rank test on success per seed, $p = 0.17$).

Table 13: **Closed loop success** of LingBot-VA on Place Can Basket under three physics configurations, 30 scored seeds per repetition. The solver configuration completed two repetitions.

Configuration	Rep. 1	Rep. 2	Rep. 3	Total	Mean $\pm$ std (%)
RoboTwin (can 10 g)	23/30	24/30	20/30	67/90	74.4 ± 6.9
Can mass 100 g	29/30	30/30	27/30	86/90	95.6 ± 5.1
Solver position iterations 32	26/30	26/30	–	52/60	86.7 ± 0.0

Closed loop takeover. We also let the policy take over from a replayed prefix, 1 s before the first artifact of each of the 21 Place Can Basket failures, with the speed clamp on or with the physics unchanged. All 42 prefixes match their recordings bit for bit. The unchanged physics rescues 10 of 21 failures and the clamp 15 (paired: 7 rescued only with the clamp, 2 only without; McNemar exact test $p = 0.18$). From the same state, the policy's own sampling often avoids the artifact, which is why the corrected rate of Equation (1) replays the recorded actions instead.

Skipped scripted demonstrations. RoboTwin scores a policy on a seed only if its own scripted demonstration passes there first. On Place Can Basket, 22 seeds were skipped for this reason over the three repetitions, and the demonstration of 16 of them contains a contact artifact. Replayed with the depenetration limit switched on 0.1 s before the first artifact, 13 of the 16 pass the success check, with all 16 prefixes identical to the recording up to the switch (Figure 3, bottom). The artifacts therefore also decide which seeds enter the evaluation, before any policy is scored.

E Other Platforms

The adapters for RoboDojo, LIBERO and RoboCasa are described in Section C.1; this section reports what the audit found on each.

E.1 RoboDojo

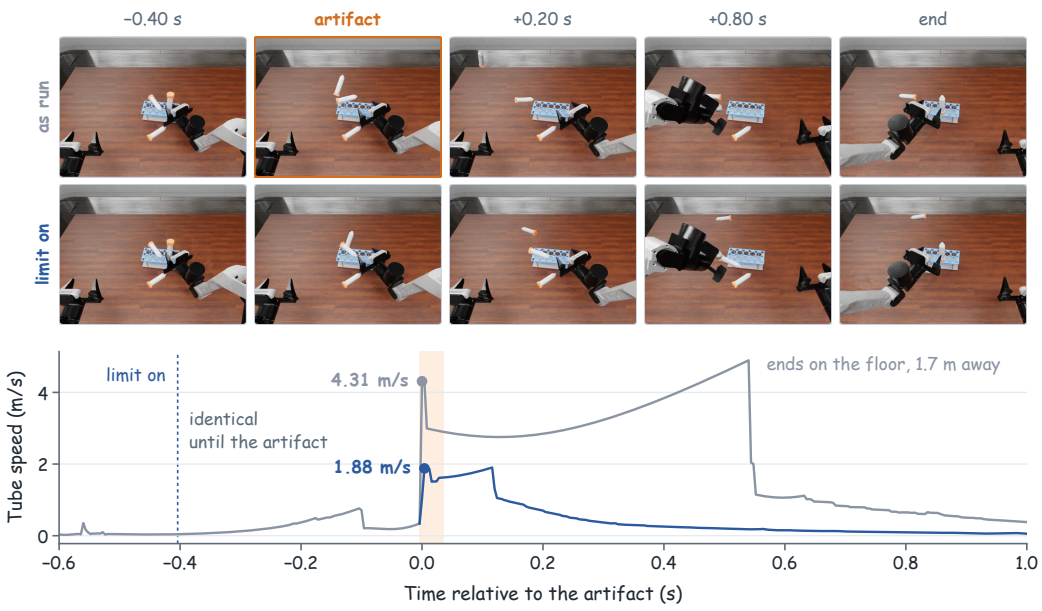

Figure 6: **RoboDojo, Insert Tubes (layout 2), XR-1.** Head camera frames at fixed times relative to the first artifact, as run (top) and replayed with the depenetration limit of 1 m/s switched on 0.4 s before it (bottom), and the speed of the ejected tube in both runs. As run, the tube leaves the rack at 4.31 m/s within one step, flies off the table and lands 1.7 m away; the later peak of the gray curve is its fall. With the limit it leaves at 1.88 m/s, slows below 0.4 m/s within 0.3 s and stays on the table. Both rows are replays of the recording; the as run replay matches it bit for bit, and the limited one first differs at the artifact step.

Seven of the 40 XR-1 episodes contain contact artifacts, all failures. In the first pass over seed 0: two events on Insert Tubes (layout 2) and three on Store Tools In Toolbox (layout 1). In the later passes: two on Insert Tubes (layout 2 under seed 0, at 6.30 m/s, and layout 0 under seed 2), four on Fill Pen Holder (layouts 1 and 2 under seed 0) and one on Store Tools In Toolbox (layout 1 under seed 1). The first detection stage confirmed 25 events in 15 episodes, and the gravity stage explained 13 of them as falls. Only the first pass was replayed. We replayed the Insert Tubes episode with the depenetration limit switched on 100 physics steps (0.4 s) before the ejection; the replay is identical to the recording up to the ejection itself (Table 14).

Table 14: **RoboDojo, Insert Tubes layout 2**: speed at which the tube leaves, replayed with the depenetration limit switched on 0.4 s before the ejection.

Depenetration limit	Exit speed (m/s)
None (replay identical to the recording)	4.31
1 m/s	1.88
0.3 m/s	0.62

E.2 LIBERO

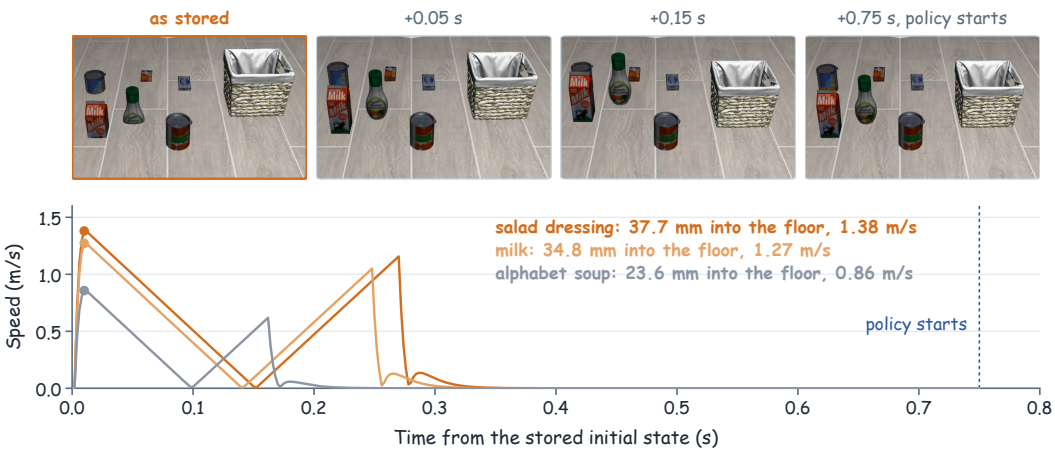

Figure 7: **LIBERO-Object task 0, stored initial state 0**, built as the evaluation builds it (official LIBERO, robosuite 1.4.0, MuJoCo 3.2.3) and run through the 15 settle steps that precede every LIBERO policy step. As stored, the salad dressing, the milk and the alphabet soup sit 37.7, 34.8 and 23.6 mm inside the floor. MuJoCo pushes them out within 10 ms at 1.38, 1.27 and 0.86 m/s; they hop, land within 0.3 s and are at rest before the policy acts at 0.75 s. Depths and peak speeds match the audit of all 2000 stored states, of which 650 start inside a support.

Table 15: **LIBERO's stored initial states**, built as the evaluation builds them and simulated for the 15 steps before the policy acts, without a policy.

Suite	Initial states	Contact ejection	After gravity	Contact artifacts	Tasks affected
LIBERO-Spatial	500	3	0	0	0/10
LIBERO-Object	500	500	500	500	10/10
LIBERO-Goal	500	0	0	0	0/10
LIBERO-10	500	152	150	150	3/10

Initial states. In the floor scene of LIBERO-Object and the living room table of LIBERO-10 tasks 0, 1 and 7, the stored initial states place objects inside their support (Table 15): the salad dressing 37.7 mm, the ketchup 37.4 mm and the milk 34.8 mm into the floor, and the ketchup 14.6 mm into the table. The depth is the same in every initial state of a scene, which differ only in horizontal positions. MuJoCo removes the overlap within the first 10 ms by launching the objects, at up to 1.38 m/s off the floor and 1.83 m/s off the table, with no robot contact. In the 130 policy episodes that start from these states, the policy succeeds in 96.2%, against 94.8% in the other 270.

Episodes with the policy. Two of the 400 episodes contain a contact artifact while the policy acts. In LIBERO-10 task 9 (a mug into the microwave), MuJoCo reports a contact 242.8 mm deep with 412 N between a collision box of the mug and a plate of the microwave, although its own distance query gives 1.4 mm between the two shapes; the mug leaves the gripper at 1.57 m/s and the episode fails. Whether this contact caused the failure is not tested. In LIBERO-Object task 4, a finger touches the ketchup again during the release, and its speed rises from 0.25 to 0.62 m/s for one step; the episode succeeds.

E.3 RoboCasa

Five of the 300 XR-1 episodes contain a contact artifact, all of them failures. On the ten placement tasks: bread against its basket at 0.61 m/s on Arrange Bread Basket, an object leaving a drawer at 1.19 m/s on Pick Place Drawer To Counter, and a bowl on the floor at 0.53 m/s on Stack Bowls Cabinet. On the other 40 tasks, which include doors, drawers, appliances and long horizon

composite tasks: a baguette against the door of a toaster oven at 0.61 m/s on Heat Kebab Sandwich, and a piece of meat on the floor, pushed by the mobile base, at up to 0.84 m/s on Pack Identical Lunches. The first detection stage flags 22 of the 300 episodes; the gravity stage explains the events in the other 17 as falls.

F Episode Accounting

Table 16 gives the counts behind Table 1: failures with an artifact out of all failures, the failures that pass when replayed with the limit, and the successes with an artifact that fail. It also lists the tasks outside the five, which Table 1 omits because their audit is incomplete: HyVLA's successes on its six other tasks were not checked for artifacts, HyVLA was not run on the remaining 39 tasks, and the closed loop runs cover the five tasks and, for LingBot-VA, the six others (Table 12), not the remaining 39.

Table 16: **Counts behind Table 1.** *Art.*: failures with an artifact, of all failures. *Resc.*: failures that pass under the limit. *Lost*: successes that fail under the limit, of all successes with an artifact. †Both passes of the 40 tasks and the single pass of the other five (Table 5); only the first pass, 4 failures and 3 successes with an artifact, was replayed.

	LingBot-VA			HyVLA		
Task	Art.	Resc.	Lost	Art.	Resc.	Lost
Place Can Basket	21/23	18	7/16	1/4	1	2/3
Stack Bowls Three	12/12	7	–	2/2	1	0/1
Place Object Basket	1/2	1	0/1	0/1	–	–
Place Bread Basket	1/1	1	–	0/0	–	–
Stack Bowls Two	1/1	1	–	3/3	3	–
Five tasks	36/39	28	7/17	6/10	5	2/4
Six other tasks	1/24	1	–	0/38	–	–
Other 45 tasks†	10/128	1	1/7	–	–	–
Rates outside the five tasks (as run / replay / exclusion), LingBot-VA then HyVLA						
Six other tasks	76.0 / 77.0 / 76.8			68.3 / 68.3 / –		
Other 45 tasks†	92.3 / – / –			– / – / –		

Table 17 lists every scored episode that contains a contact artifact, with its seed, recorded outcome and the outcome of its replay with the depenetration limit of 1 m/s switched on 0.1 s before the first artifact. Equation (1) sums $r_e - y_e$ over these rows. Place Can Basket was evaluated in three repetitions of 30 seeds (Section B), so a seed can appear in more than one repetition.

Table 17: **Every scored episode with a contact artifact** and its replay. S: success, F: failure; the arrow leads from the recorded outcome y_e to the replayed outcome r_e.

Task	Episodes	Seed, recorded → replayed
LingBot-VA		
Place Can Basket (rep. 1)	13	100000 F→S, 100002 S→S, 100004 S→F, 100006 F→S, 100007 F→S, 100008 F→S, 100014 S→S, 100017 S→F, 100021 F→S, 100022 S→S, 100024 S→F, 100029 F→S, 100031 F→F
Place Can Basket (rep. 2)	11	100004 S→S, 100011 F→S, 100012 F→S, 100017 S→F, 100019 F→S, 100021 S→S, 100022 F→S, 100025 S→F, 100026 S→S, 100027 S→F, 100029 F→F
Place Can Basket (rep. 3)	13	100000 S→S, 100003 F→S, 100009 F→S, 100016 F→F, 100019 F→S, 100021 F→S, 100022 F→S, 100023 F→S, 100027 F→S, 100028 F→S, 100029 S→F, 100030 S→S, 100032 S→S
Stack Bowls Three	12	100002 F→S, 100009 F→F, 100023 F→F, 100028 F→F, 100047 F→S, 100067 F→S, 100071 F→S, 100073 F→F, 100075 F→S, 100077 F→S, 100078 F→S, 100089 F→F
Place Object Basket	2	100013 F→S, 100015 S→S
Place Bread Basket	1	100003 F→S
Stack Bowls Two	1	100011 F→S
Place Fan	1	100015 F→F
Hanging Mug	1	100003 F→F
Stamp Seal	1	100026 F→S
Move Stapler Pad	1	100018 F→F
Place Burger Fries	2	100012 S→S, 100015 S→S
Place Dual Shoes	1	100015 S→F
HyVLA		
Place Can Basket	4	100004 S→F, 100021 S→F, 100022 F→S, 100023 S→S
Stack Bowls Three	3	100016 S→S, 100022 F→S, 100023 F→F
Stack Bowls Two	3	100020 F→S, 100021 F→S, 100026 F→S

G Human Verification Protocol

For Place Can Basket, 60 clips were rendered by replaying recorded episodes: 40 centered on a detected event and 20 hard negatives (the fastest object motion in episodes without an event, and detector flags that were never confirmed). Clips were shuffled and shown without their origin to five robotics engineers, who rated them independently and then resolved disagreements by discussion; the labels reported are the agreed ones. For each clip, they answered whether any object moves in a physically implausible way, and whether the robot was behaving sensibly for the task before that moment; the robot was judged on track in 38 of the 40 events. For six further tasks, 60 complete evaluation videos were shown in random order without seed or outcome, with three questions: implausible motion anywhere in the episode (objects falling after release excluded), sensible robot behavior, and whether the task was completed. Their completion judgments matched the benchmark's in 59 of 60 videos.