Auditing Manipulation Benchmarks Before Training

Placement Radius, a Shipped-State Defect, and Embodiment Tolerance

Monish Allada
University of North Carolina at Charlotte

Avneh Bhatia*
Exea Labs

Lalith Samineni
University of Maryland

Joshua Selvaraj
Exea Labs

*Corresponding author: avneh@exealabs.org

Abstract

Manipulation benchmarks are often interpreted as tests of perception and language-conditioned control, but a score can also reflect regularities in the evaluation states. We introduce the *placement radius* R, the radius of the smallest disc containing an object's shipped positions. The condition $R \leq \delta$ is exactly the geometric condition for one task-indexed constant to lie within distance δ of every shipped position; behavioral equivalence additionally requires an embodiment-specific robustness assumption.

Applied to LIBERO, this audit reveals a shipped-state defect in LIBERO-Object. Its tasks declare $5 \times 5\,\text{cm}$ target-sampling regions, yet the shipped states span only 0.0–1.9 cm, and six of ten tasks place the target at one point to floating-point precision. In a paired intervention on one task, sampling the declared region reduces the lookup-goal condition from 25/30 to 10/30 successes ($p = 0.0007$). A reset-informed condition supplied the true target position and succeeded on all 30 repaired states. Thus the same motion controller can solve the repaired states when supplied the correct target.

A sparse displacement ladder shows a dose response but does not identify a uniform tolerance δ: outcomes change even at the smallest nonzero level, and the experiment varies direction and initial state together. Accordingly, placement radius is a pre-training diagnostic of geometric shortcut opportunity, not by itself a certificate of policy behavior. The causal intervention is limited to one task, and we do not establish that a trained policy exploits the defect.

1 Introduction

A manipulation benchmark score is commonly read as evidence that a policy identified the commanded object, localized it, and executed the requested action. That interpretation is not warranted when the evaluation states make a correct goal predictable from the task index alone. This paper audits that possibility before training a policy.

Our starting point is a mismatch in LIBERO [5]. Each task specifies a target-sampling region, but LIBERO-Object's shipped initial states do not exercise it. Six of ten tasks place the identity-bearing target at one point to floating-point precision, while sibling suites exercise their declared regions. We regenerate candidate states from the declared region and test whether the mismatch matters behaviorally. On one task, a task-indexed lookup-goal condition falls from 25/30 to 10/30 successes; a reset-informed condition that receives the true target succeeds on all 30 repaired states.

We use the *placement radius* R to summarize the geometric opportunity for such a shortcut. R is computed from shipped state files without training or simulation. It answers a narrow question: how accurately can one constant approximate an object's shipped positions? Turning that geometric fact into a behavioral claim requires a controller-, task-, object-, and phase-specific robustness condition. Our displacement experiment does not identify a uniform value for that tolerance, so we report radii and behavioral evidence separately.

Contributions.

1. We define placement radius and the exact geometric condition under which a task-indexed constant lies within a specified distance of every shipped position (§3).
2. We show that LIBERO-Object's shipped states do not sample the target regions declared by its task files, while three sibling suites do (§2).
3. We release a candidate repaired state set and a paired one-task intervention with a reset-informed condition (§2.2).
4. We separate geometric coverage from behavioral equivalence and give the correct confidence-set interpretation of a decision window (§4.1).

Relation to prior work. Dataset bias and shortcut learning can make benchmark scores overstate a model's intended capability [2, 3, 8]. In robotics, perturbation suites and recollection expose sensitivity after a policy has

been trained [1, 6, 7, 9]. Our diagnostic is complementary: it measures geometric dispersion directly from evaluation-state files. Recent work also shows that language-free probes can solve LIBERO tasks at high rates [4]; placement pinning is one possible shortcut opportunity, not a complete explanation of those results.

Scope and reproducibility. All experiments use simulation. The causal intervention is one task at $n = 30$ per primary condition. We do not show that a trained policy exploits the shipped-state defect or that the effect generalizes across LIBERO-Object. The released artifacts include per-trial outcomes, state hashes, figure-generation scripts, and scripts that recompute the reported statistics.

2 Shipped states do not match declared sampling regions

This comparison uses the benchmark's task declarations and shipped initial states. It does not require a policy, controller, tolerance, or statistical model.

Each LIBERO task declares, in its own BDDL, an axis-aligned region to sample its target from. For `pick_up_the_alphabet_soup` that region is 5×5 cm.

Write $x_j(s)$ for coordinate $j \in \{1, 2\}$ of a task's target in shipped state s. The **spread**, $\max_j(\max_s x_j(s) - \min_s x_j(s))$, is the longer side of the shipped placements' axis-aligned bounding box. The **exercised fraction** divides that spread by the corresponding side declared in the BDDL file. Both quantities compare the BDDL declaration with the shipped-state array and require no policy, simulator, or tolerance.

The fifty shipped states for task 0 place the soup at the exact center of its declared box (Figure 2). Figure 1 extends the same comparison to all twenty LIBERO-Object and LIBERO-Long tasks.

suite	declared	shipped spread	exercised?
Spatial	2.0 cm	2.9 cm	yes
Goal	2.0 cm	3.0 cm	yes
Long	5.0 cm	4.9 cm	yes
Object	**5.0** cm	**0.0–1.9** cm	**no**

Table 1: Mean declared box side and shipped spread by suite. Spatial, Goal, and Long exercise their declared regions; LIBERO-Object ships only 0.0–1.9 cm of its declared 5 cm side.

Table 1 summarizes all four suites. LIBERO-Object's shipped states conflict with its declared target sampler, whereas Spatial, Goal, and Long exercise their declared regions. We refer to this mismatch as a **state-generation defect**.

Candidate repaired states. We draw the target's position from the box the task file already declares. `scripts/resample_init_states.py` does this for all ten LIBERO-Object tasks (Table 2). It changes *only* the two state-array columns holding the target's x and y; every other object, joint velocity, and time field is bit-identical to the shipped file and verified by array comparison.

	shipped	repaired
spread, LIBERO-Object (10 tasks)	0.00–1.94 cm	4.44–4.96 cm
spread, LIBERO-Long (unmodified)	4.84–4.99 cm	
clearance to any other region	—	≥ 7.42 cm

Table 2: Shipped and candidate repaired spreads. Repaired positions satisfy the declared regions and a 7.42 cm center-to-center clearance constraint; resting height and post-settling table contact remain simulator-dependent.

2.1 Evaluation setup

All goal-supply conditions drive the same non-learned pick-and-place controller. The *lookup-goal* condition reads the task index and emits the mean shipped positions of the identity-bearing target and shared container; it receives no image or command at evaluation time. The *reset-informed* condition reads the true target and container positions once at reset and then uses static goals. The *learned-goal* condition uses MicroVLA's trained goal predictor and is reported only as an exploratory comparison.

Experiments use LIBERO commit `8f1084e3`, MuJoCo 2.3.7, robosuite 1.4.1, torch 2.8.0, ultralytics 8.4.115, numpy 2.2.6, Python 3.10, and `MUJOCO_GL=osmesa`. Trial indices select initial states deterministically, and paired comparisons use the same trial indices. Reported p-values are exact two-sided McNemar tests unless stated otherwise.

2.2 Repaired-state intervention

We evaluate task 0 on the shipped states and on states drawn from its declared region. The paired runs use the same controller, weights, seeds, trial indices, machine, and package versions; only the target-position columns change.

goal supply	shipped	repaired	paired
lookup	**25/30**	**10/30**	17–2; $\mathbf{p = 0.0007}$
learned	19/30	13/30	10–4; $p = 0.18$
reset-informed	10/10	**30/30**	vs lookup: 20–0; $\mathbf{p = 2 \times 10^{-6}}$

Table 3: Task 0 paired outcomes for lookup-goal, learned-goal, and reset-informed conditions. Repair changes lookup from 25/30 to 10/30 successes ($p = 0.0007$); reset-informed succeeds on all 30 repaired states. The learned-goal change is unresolved and is also compatible with distribution shift.

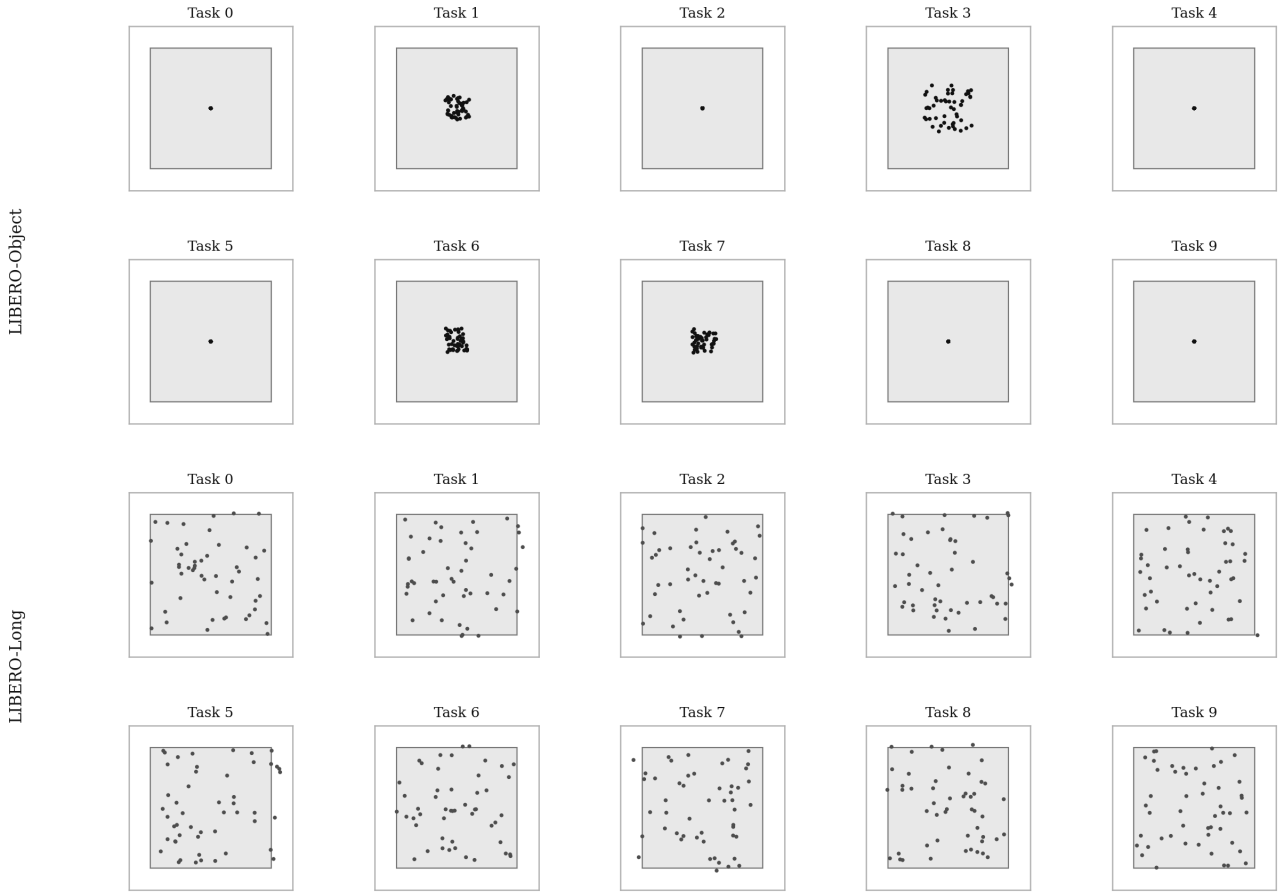

Figure 1: Declared target regions and fifty shipped states for every LIBERO-Object and LIBERO-Long task. Every Long task exercises its declared region; no Object task does, and six Object tasks collapse to one point.

Reset-informed condition. The repaired positions could be generally harder. To test that alternative, we hold the controller and repaired states fixed but read the true target position once at reset instead of using a constant compiled from the shipped files.

On the same thirty repaired states, the lookup-goal condition scores 10/30 and the reset-informed condition scores 30/30. All 20 discordant trials favor the reset-informed condition ($p = 2 \times 10^{-6}$). The result is limited to this task and controller.

3 Placement radius as a pre-training diagnostic

§2 establishes a mismatch between declared and shipped placements. This section measures the geometric concentration of those shipped placements and states the additional assumption needed for a behavioral claim.

3.1 Definitions

Fix a task t that ships a finite set S_t of initial states. Write O_t for the objects that the policy must localize; LIBERO declares this set as `obj_of_interest`. Let $x_t^o(s) \in \mathbb{R}^2$ denote object o's table position in state s. All radii below use planar Euclidean distance. A controller that consumes a three-dimensional or anisotropic goal requires the corresponding goal-space norm and radius.

Definition 1 (Placement radius)**.** *The* placement radius *of object o in task t is the radius of the smallest disc containing every shipped position,*

$$R_t(o) \;=\; \min_{c \in \mathbb{R}^2} \; \max_{s \in S_t} \|x_t^o(s) - c\|.$$

By construction, a constant within distance δ of every shipped position exists if and only if $R_t(o) \le \delta$; the minimizing center is one such constant. This is a geometric equivalence, not a behavioral one.

The distinction from the *diameter* $D_t(o) = \max_{s,s'} \|x_t^o(s) - x_t^o(s')\|$ matters. $D \le \delta$ implies $R \le \delta$ but not conversely: $R \ge D/2$ always, and Jung's theorem gives $R \le D/\sqrt{3}$ in the plane, so $D > \delta$ does not imply $R > \delta$.

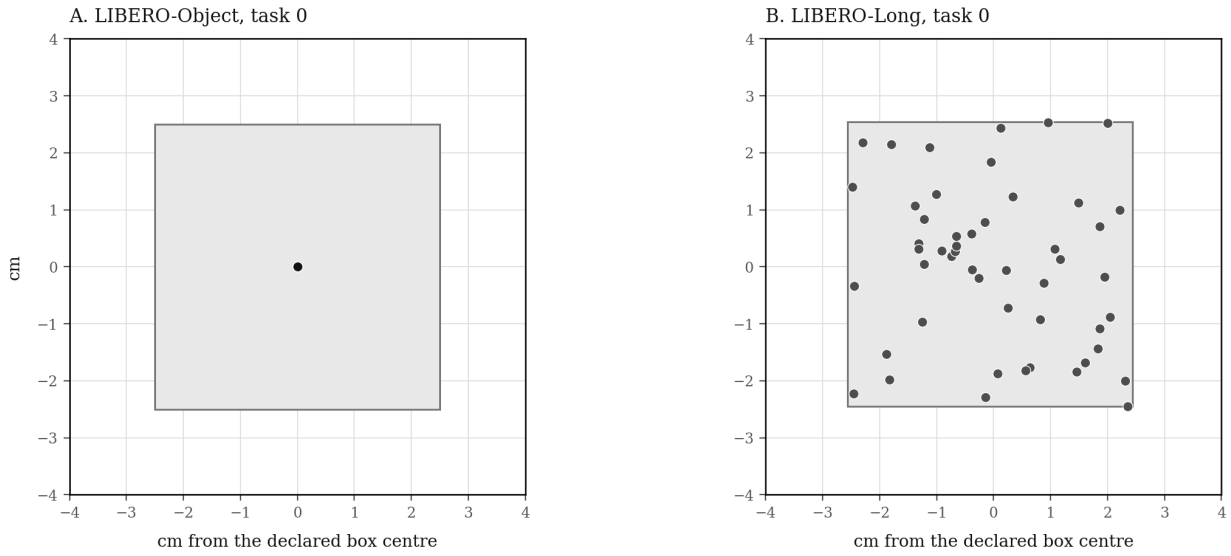

Figure 2: Declared target region and shipped states for task 0 of LIBERO-Object (**A**) and LIBERO-Long (**B**). The tasks declare the same 5×5 cm region, but only Long exercises it.

Definition 2 (Geometric δ-coverage)**.** *Object o of task t is* geometrically δ-covered *if $R_t(o) \leq \delta$. A task is covered only with respect to an explicitly selected object set, norm, and goal space.*

Which objects, and why it matters. Every LIBERO task specifies at least a manipulated object and a destination. Call *o identity-bearing* if it is what individuates the task from its siblings — the grocery in "pick up the *alphabet soup* and place it in the basket" — and *shared* otherwise. LIBERO-Object's ten tasks differ only in the identity-bearing object; all ten share one basket.

We report identity-bearing and shared objects separately. This distinction is essential: LIBERO-Object's identity-bearing targets have $R \in [0.00, 1.17]$ cm, whereas its shared baskets have $R \in [1.67, 1.98]$ cm. A task-level behavioral claim must specify which goals the controller receives and cannot be inferred from either column alone.

Task semantics and placement variation. LIBERO-Object's ten tasks are individuated by *which object*, so the suite need not vary that object's initial position to distinguish task identities. LIBERO-Spatial's tasks are individuated by *where the bowl is*, so that position must vary. Table 4's ordering therefore partly reflects the suites' task definitions.

The magnitude is not built in. Object-based task identity does not require a position to be frozen to one `float64` value when the task file declares a 5×5 cm sampling region. LIBERO-Object could randomize placement while retaining object identity, and LIBERO-Spatial demonstrates that a suite can carry both. **The finding is the gap between the declared region and the shipped states, not the ordering of the four suites.**

Assumption A1 (joint δ-robustness). For a specified task, controller, selected object set, norm, and coupled simulator randomness, replacing all true goals jointly by values within δ leaves each episode's outcome unchanged. A1 is an empirical assumption, not a consequence of placement radius. The displacement experiment in §4 does not establish A1 for any positive uniform δ.

Proposition 1 (Conditional lookup sufficiency)**.** *Suppose $R_t(o) \leq \delta$ for every selected object of every task, and suppose A1 holds at δ. A task-indexed goal supplier can then emit the minimum-enclosing-disc centers. Driven by those goals, the shared controller achieves the same outcomes as it does with the true positions.*

Proof. Take c_t^o to be the center of the minimum enclosing disc. By Definition 1, $\|x_t^o(s) - c_t^o\| \leq R_t(o) \leq \delta$ for every s, so $t \mapsto (c_t^o)_{o \in O_t}$ is a lookup table meeting Definition 2. A1 applies at each state, giving the same outcome as the true positions; averaging over S_t gives the claim. □

The proposition is conditional: geometric coverage supplies the constants, while A1 supplies the behavioral equivalence.

Corollary 1 (Outcome equivalence under A1)**.** *On a suite satisfying the hypotheses, episode outcomes cannot distinguish the lookup policy from the same controller supplied the true goals.*

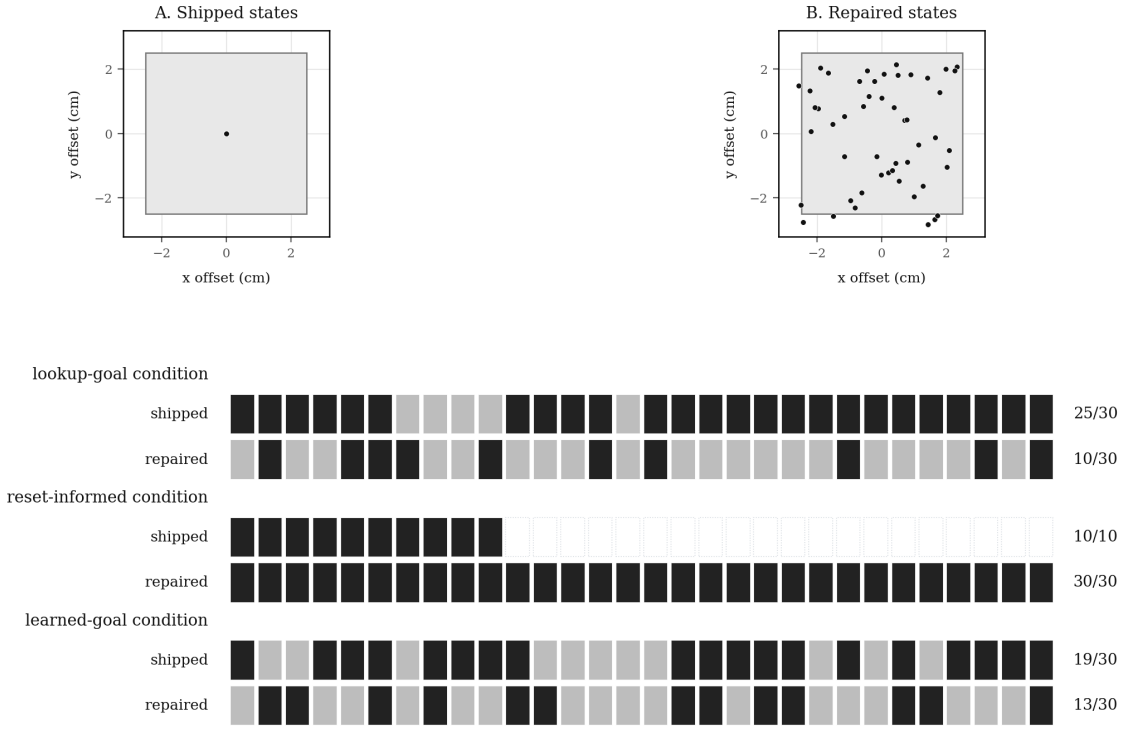

Figure 3: Target positions and paired outcomes under shipped and repaired states. The lookup-goal condition degrades under repair, while the reset-informed condition succeeds on all repaired trials. Dark cells denote success, light cells denote failure, and empty cells were not run.

This corollary does not assert that A1 holds or that a measured policy uses a lookup shortcut. It describes only the comparison class induced by the stated assumptions.

3.2 Measurement

A naive fixed-column layout fails for every suite containing an articulated fixture. We therefore compile each task's model and read `jnt_qposadr`. Appendix A gives the procedure and the cross-check.

Two LIBERO-Goal tasks name a fixture with no free joint — a cabinet drawer, a stove. A fixture that never moves has $R = 0$; these cases are identified separately from movable objects.

Placement concentration. In 6 of Object's 10 tasks the target's start position takes exactly *two* `float64` values, one unit in the last place apart on a single axis: constant to the precision the format can express. The other four have $R \in [0.60, 1.17]$ cm.

Across tasks, the ten identity-bearing targets occupy two table positions 22.1 cm apart, with five tasks at each position. Only two distinct constants are therefore needed to reproduce their centers.

	identity-bearing R (cm)		shared R
suite	mean	range	range
Object	**0.30**	0.00–**1.17**	1.67–1.98
Spatial	2.15	1.49–4.49	1.66–2.01
Goal	1.48	0.00–2.48	0.00–1.70
Long	3.10	2.85–3.28	0.00–3.17

Table 4: Placement radius of every LIBERO task, computed from the shipped initial states. LIBERO-Object has the smallest identity-bearing radii, while its shared baskets vary across episodes.

4 Embodiment tolerance and decision resolution

Placement radius is geometric. Relating it to behavior requires an embodiment-specific robustness region. We therefore displaced the lookup controller's grasp goal on two LIBERO-Object tasks. Each trial uses a fixed random direction across radii $r \in \{0, 1, 2, 4, 8\}$ cm, pairing direction and initial state across levels.

Observed success decreases monotonically on both tasks and reaches zero at 4 cm, demonstrating sensitivity to grasp-goal displacement. The experiment does *not* establish $\delta \geq 2$ cm under Assumption A1: task 3 changes

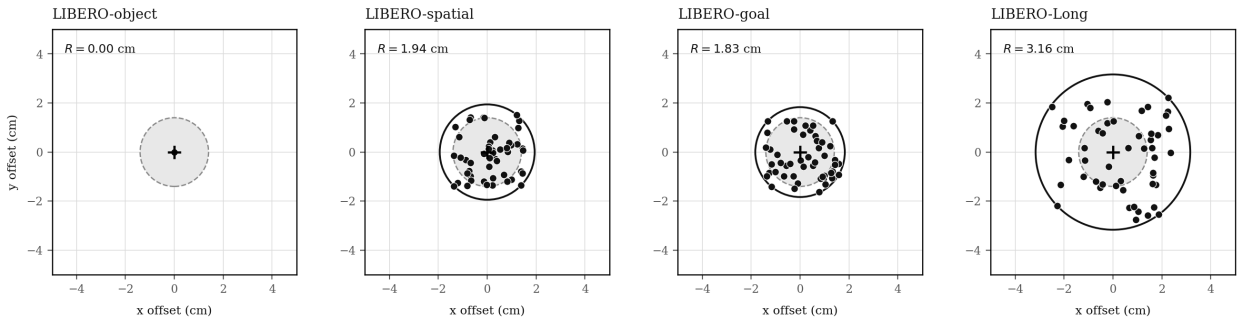

Figure 4: Placement radius for one representative task per suite. The solid circle is the minimum enclosing disc with radius R; the dashed gray disc shows an illustrative $\delta = 1.4\,\text{cm}$. A constant lies within δ of every shipped position exactly when $R \leq \delta$.

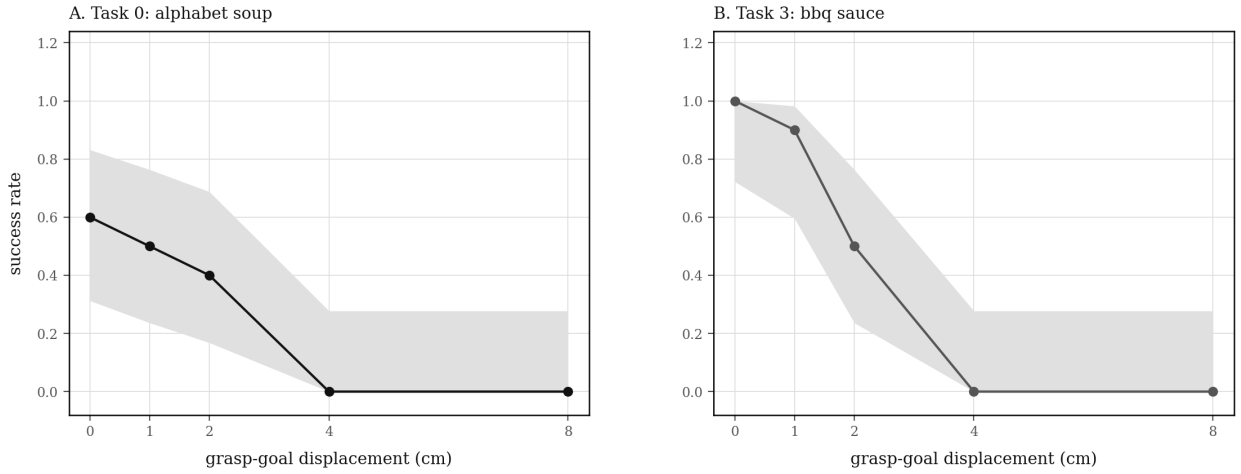

Figure 5: Success under radial displacement of the lookup-goal condition's grasp goal. Directions and initial states are paired across displacement levels; gray bands are Wilson 95% intervals.

r (cm)	0	1	2	4	8
task 3	10/10	9/10	5/10	0/10	0/10
Holm p	—	1.0	0.125	**0.008**	**0.008**
task 0	6/10	5/10	4/10	0/10	0/10
Holm p	—	1.0	1.0	0.125	0.125

Table 5: Paired success counts and exact McNemar tests against $r = 0$, with Holm correction within each task. The ladder shows degradation by 4 cm but does not estimate a uniform tolerance.

from 10/10 to 9/10 at 1 cm and to 5/10 at 2 cm. Failure to reject equality at $n = 10$ is not evidence of statewise equivalence. In addition, each trial couples one direction with one initial state, so the experiment does not identify a worst-case directional tolerance. We therefore make no positive estimate of a uniform δ.

4.1 Decision window

The radii still determine how precisely a *hypothetical shared scalar* δ would need to be known for a selected geometric comparison. Choose one object o_t per task and write $R_t = R_t(o_t)$. For a subset of tasks A, define

$$W(A) = \left[\max_{t \in A} R_t,\ \min_{t \notin A} R_t\right). \tag{1}$$

Proposition 2 (Geometric decision window)**.** *Under the selected objects, norm, and shared scalar δ, every task in A is geometrically δ-covered and every task outside A is not if and only if $\delta \in W(A)$. For a confidence set C_δ, the data support the comparison when $C_\delta \subseteq W(A)$, reject it when $C_\delta \cap W(A) = \varnothing$, and are otherwise inconclusive.*

Proof. The comparison requires $R_t \leq \delta$ for all $t \in A$ and $R_t > \delta$ for all $t \notin A$. These inequalities are equivalent to $\max_{t \in A} R_t \leq \delta < \min_{t \notin A} R_t$. The confidence-set cases follow by set inclusion and disjointness. □

For identity-bearing movable objects, the LIBERO-Object one-versus-rest comparison gives $W = [1.171, 1.492)$ cm using the unrounded radii. This interval is a property of that selected geometric comparison; it is not a suite-wide behavioral verdict. Our sparse ladder provides no confidence set for a uniform δ, so it neither supports nor rejects membership in this window. A future test would need to define an equivalence margin, vary direction independently of state, and estimate task- and object-specific robustness in the controller's actual goal space.

5 Discussion and limitations

The declared-versus-shipped mismatch does not depend on a policy or tolerance. LIBERO-Object declares target-sampling regions that its shipped evaluation states do not exercise. Placement radius summarizes the resulting geometric regularity and can be computed before training. It does not, on its own, show that a policy memorized a location or that a constant preserves behavior.

The repaired-state intervention adds one scoped causal result. On one task, the lookup-goal condition loses 15 of its 25 shipped-state successes when the target is sampled from the declared region. The reset-informed condition succeeds on all 30 repaired states, showing that this controller can solve those states when given the correct target. The learned-goal condition remains ambiguous because its repaired-state drop is also consistent with distribution shift.

Several limitations bound the interpretation. First, the causal intervention covers one task, one controller, and $n = 30$ paired trials. Second, the repaired states are author-constructed under a clearance constraint and are not an official LIBERO release. Third, the displacement ladder is small, couples state and direction, and does not estimate a uniform robustness tolerance. Fourth, the geometric analysis uses planar Euclidean positions; a controller operating in three-dimensional or anisotropic goal space requires the corresponding norm and radius. Finally, neither trained-policy exploitation nor transfer to other benchmarks is established.

Benchmark authors can make these checks inexpensive by publishing declared and shipped placement distributions, per-object radii in the policy's goal space, and evaluation tiers with controlled displacement. Behavioral certification requires a separately validated robustness region for the evaluated embodiment; geometric dispersion should not be presented as behavioral invariance.

6 Conclusion

LIBERO-Object declares 5 cm target-sampling regions but ships spreads of only 0.0–1.9 cm. Six tasks collapse to one point at floating-point precision. Placement radius exposes this regularity directly from the evaluation states.

On one task, replacing the shipped placement with samples from the declared region changes the lookup-goal condition from 25/30 to 10/30 successes, while the reset-informed condition succeeds on all 30 repaired states. This demonstrates that the shipped-state defect can support a location lookup for that controller and task. It does not establish that a trained policy uses the shortcut or that all LIBERO-Object tasks are behaviorally equivalent under a constant goal. The placement radius measures geometric shortcut opportunity before training; behavioral conclusions require task- and embodiment-specific validation.

Acknowledgements

All experiments reported in this paper were run on AMD Instinct MI300X accelerators. The large on-device memory kept the frozen detector, learned-goal head, and simulator tensors resident, so paired shipped-versus-repaired rollouts could be executed back to back in one invocation without host-device offloading. High HBM bandwidth reduced the cost of moving visual embeddings and batched trial tensors through that pipeline. We gratefully thank AMD for providing this compute.

Data and Code Availability

The package includes the manuscript source, bibliography, figures, and per-trial artifact references. The accompanying repository contains scripts for state extraction, repaired-state generation, validation, figure rendering, and statistical recomputation.

AI Use Statement

No generative AI tools were used in the creation, writing, analysis, or preparation of this work. All research, ideas, methodology, implementation, and written content were developed independently by the authors.

A Recovering which column owns which object

LIBERO stores each task's initial states as a MuJoCo flattened state $[\, t \mid q \mid \dot{q} \,]$; every shipped $\dot{q}$ is zero, so the columns that vary across states are exactly the position degrees of freedom the suite randomizes.

Recovering *which* object owns which column requires care. The fixed layout $[\, t \mid 9 \mid N \times 7 \mid \dot{q} \,]$ that works for LIBERO-Object predicts a width of $19+13N$ and fails for every suite containing an articulated fixture — LIBERO-Spatial has width 92 against 5 declared objects. We therefore run two passes: the fixed-layout one, and one that compiles each task's model and reads `jnt_qposadr` with

the joint names, mapping every column onto a named body. **The two agree wherever both apply**, and only the second is used for suites the first cannot parse. Artifacts: `results/suite_forensics_columns.json` (digests), `results/suite_forensics_joints.json` (per-object positions).

References

[1] Senyu Fei et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

[2] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2: 665–673, 2020.

[3] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, 2018.

[4] 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.

[5] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *NeurIPS Datasets and Benchmarks*, 2023.

[6] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021.

[7] 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.

[8] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] 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.