RETHINKING DEMONSTRATION UNLEARNING IN IMITATION LEARNING FOR ROBOTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot policies learn from demonstrations, some of which may later need to be withdrawn, for example when consent is revoked or an episode proves faulty. Demonstration unlearning edits the trained policy to remove their influence; ideally, the result matches a policy trained without them. Unlearning is often judged by isolated proxies such as restored task success, higher loss on withdrawn data, or membership-inference scores. When these proxies disagree with retraining, an edit can be credited with removing demonstrations whose influence remains detectable, or rewarded for overshooting what retraining would produce. We propose a retrain-calibrated audit that makes removal claims testable and shows whether an edited policy departs from retraining in behavior, data fit, or both. It compares the edited policy with independently retrained policies through actions executed at the same observations and prediction error (loss) on withdrawn episodes, including whether that loss separates them from never-used episodes; a symmetric rank test calibrates these differences against retraining variation. In physical cup insertion, an ACT edit raises success from 5/20 to 18/20 trials, yet its loss still perfectly separates audited withdrawn from never-used episodes and its actions remain far from retraining; a second policy class shows the same success-versus-evidence disagreement. In simulation, action and episode measurements detect complementary discrepancies; tests on fresh retain-only models and controlled data reintroduction reveal false rejections and sensitivity to checkpoint selection and queried states. Useful robot repair is thus not evidence of demonstration removal: unlearning claims should be checked against retraining in behavior and data fit.

1 INTRODUCTION

Robot imitation learning acquires policies from demonstrations: observations are paired with the actions a demonstrator took, and the learned policy uses these examples to act in new situations. Recent visuomotor policies learn dexterous manipulation from teleoperation and represent complex action sequences through chunked or generative predictions (Mandlekar et al., 2021; Zhao et al., 2023; Chi et al., 2023). As demonstration collections grow, so does the need to revise what a policy has learned when particular episodes are withdrawn or identified for removal.

Machine unlearning addresses this problem after training: it updates a model so that designated data no longer influence it, while preserving what it learned from the retained data (Cao & Yang, 2015; Bourtoule et al., 2021; Xu et al., 2023). Retraining without the designated data achieves this exactly but is costly, motivating approximate edits such as influence-based updates, selective weight scrubbing, and teacher–student objectives (Guo et al., 2020; Golatkar et al., 2020; Kurmanji et al., 2023).

The standard target of unlearning is defined by the training data, not by behavior: the ideal unlearned model is one trained without the removed data. Exact unlearning realizes this target by retraining; approximate unlearning is defined or evaluated by closeness to it (Ginart et al., 2019; Guo et al., 2020; Sekhari et al., 2021; Hayes et al., 2025; Triantafillou et al., 2024). For a robot policy, the ideal result is thus a policy trained on the retained demonstrations alone. Avoiding the characteristic failure of withdrawn episodes is a different goal, closer to policy correction. In robotics, the distinction between an edit and deletion is easily obscured. Suppressing an unwanted motion can repair a task,

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.

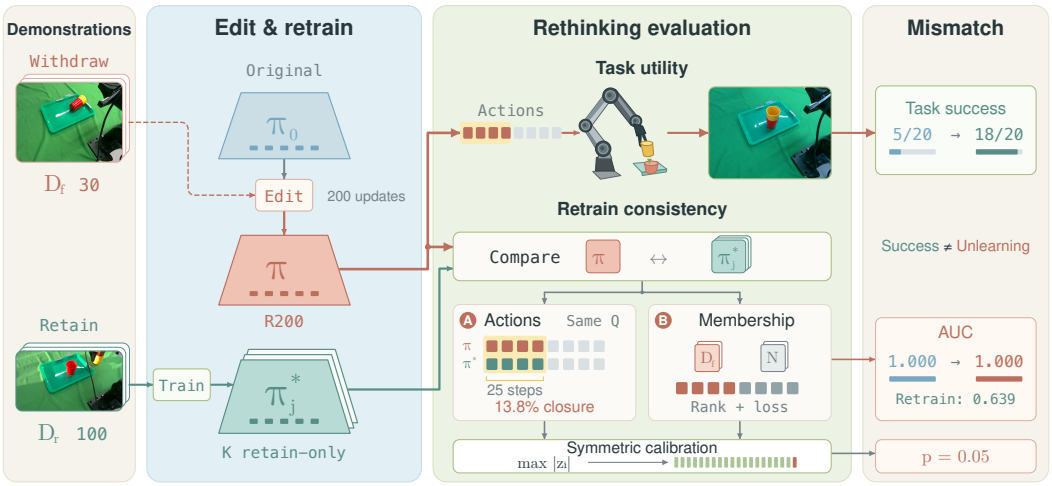


Figure 1: An unlearning attempt and its retrain-calibrated audit. A withdrawal request for $\mathcal{D}_f$ (30 withdrawn demonstrations) drives a policy edit; the 100 retained demonstrations $\mathcal{D}_r$ independently train references. The same R200 candidate improves ACT success from 5/20 to 18/20, yet membership AUC remains 1.000 versus a retrain mean of 0.639, with only 13.8% arm-action gap closure. Task success alone does not establish removal. The audit compares executed actions at matched observations and episode rank plus calibrated loss. Symmetric calibration yields joint $p = 0.05$ ($K = 19$, $\alpha = 0.05$) under an exchangeable null. The calibration-strip z_i denotes standardized coordinates (t_{ik} in Eq. 4). Closure uses three retrains and the first 25 of 50 predicted steps on 71 BRANCH observations overlapping the edit window. AUC averages five retrains and three frame draws on fixed cohorts (30 withdrawn, ten non-members). Photos show withdrawn episode 110, a retained demonstration, and R200 trial 0 from its first block; success counts aggregate 20 blind-scored trials per condition. Bars span $[0, 1]$; policy, robot, and audit diagrams are schematic.

yet the resulting policy may differ from one trained without the demonstrations that taught that motion. Conversely, withdrawing useful demonstrations can make the appropriate retrained policy less successful. We therefore compare edited policies with retain-only retraining and report task success separately.

Existing evaluation measures expose different aspects of this problem, and none of the usual proxies is sufficient alone. Task success measures a rollout outcome; forgetting loss measures fit to selected examples; membership AUC measures how scores rank audited members against non-members. Higher success does not establish removal, and a favorable ranking can conceal a loss level that has moved beyond the retraining reference. Multidimensional benchmarks already recognize related limitations (Shi et al., 2025; Pan et al., 2026). Sequential control adds a further requirement: a policy comparison should concern actions the robot actually executes, not only its endpoint or an unused part of its prediction.

We develop a *retrain-calibrated audit* that brings these comparisons under one removal target (Fig. 1). Independent retain-only policies define reference variation for matched-observation executed actions and episode-level evidence. Membership rank and calibrated loss provide complementary evidence readings; a symmetric conformal rule tests their joint consistency. Our contributions are:

- A robot-policy audit combining executed-action and episode-evidence comparisons with symmetric candidate-inclusive calibration and explicit reference assumptions (Sec. 3).
- Physical and simulated counterexamples showing that task recovery, action agreement, and membership evidence can give conflicting accounts of the same edited policy (Sec. 4).
- Independent-candidate validation using fresh retain-only models and controlled data reintroduction, together with tests of dependence on selection, execution horizon, and observation coverage (Secs. 4.4–4.5).

The central question is what an edit establishes about withdrawn demonstrations, beyond whether the robot now succeeds. Section 2 distinguishes the preprint formulation from the validation and protocol analyses added here.

2 Related Work

Visuomotor imitation policies learn from human demonstrations, using action chunks or generative action models to represent temporally extended behavior (Zhao et al., 2023; Chi et al., 2023; Physical Intelligence et al., 2025). Corrective methods such as MoRE (behavior uncloning) and RedFlow redirect policy behavior (Wang et al., 2026; Yan et al., 2026). Their objective differs from removal relative to a specified training history. VLA-Forget explicitly studies unlearning in OpenVLA-style policies, and TrajDeleter studies trajectory removal in offline reinforcement learning (Ranjan & Polyzou, 2026; Gong et al., 2025). These systems are not evaluated here; our fixed editing interventions examine what commonly reported outcomes establish about demonstration withdrawal.

Machine unlearning compares a model with training that excludes designated data (Bourtoule et al., 2021; Guo et al., 2020). MUSE evaluates utility, memorization, membership privacy, and over-unlearning; TOUR benchmarks trajectory deletion in offline reinforcement learning with matched non-members, retraining references, and multiple membership attacks (Shi et al., 2025; Pan et al., 2026). TOUR's action-error membership score compares predictions with recorded targets, whereas our action instrument compares policies with independent retrains at matched observations. Privacy and utility can also fail in different ways: Zhang et al. (2026b) recover removed graph edges from unlearned models through inversion attacks, while Zhang et al. (2026a) inject selected graph nodes and later request their deletion to induce utility degradation. These graph results motivate separate privacy and utility measurements; they do not establish demonstration removal in sequential control.

Example difficulty affects membership scores (Carlini et al., 2022); we hold episode cohorts fixed across candidates and retrains. Statistical audits target different notions: Sarmasarkar et al. (2026) lower-bound an unlearning parameter, while Ribero et al. (2026) show that distributional tests can reject retain-only models trained with different procedures. Our target is consistency with a specified retained-data recipe in action and evidence space; rank calibration uses established exchangeability arguments (Vovk et al., 2005; Shafer & Vovk, 2008).

The preprint of Li et al. (2026) introduces the behavior/evidence formulation and physical counterexamples, including executed-slice comparisons, retrain floors, membership rank and loss, and a joint conformal test. This manuscript develops that study with fully symmetric candidate-inclusive calibration, independent retain-only nulls and controlled data reintroduction. Checkpoint-selection and observation-bank comparisons, coordinate decomposition, and a corrected Can edit cohort characterize when the audit detects discrepancies and when its conclusions change. These additions validate and qualify the formulation; they do not introduce a new unlearning algorithm or a new conformal principle.

3 Retrain-Calibrated Audit

Let $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ contain retained and withdrawn demonstration episodes. An original policy π_{θ_0} trains on $\mathcal{D}$; an edit produces candidate π. Independent runs of a specified procedure $\mathcal{T}(\mathcal{D}_r)$ define retain-only references π_j^*. This procedure includes initialization, optimization, and checkpoint selection. The removal request defines the counterfactual; task utility is reported separately.

Before evaluating candidates, fix the reference procedure, observation bank, executed horizon, coordinate aggregation, episode cohorts, score interface, and significance level. References can be reused for the same retained set and training procedure. The workflow in Fig. 1 separates matched-state action measurements, episode evidence, and their joint calibration.

3.1 Executed-action consistency

For a bank Q_r of observations in trajectory region r, define

$$d_r(\pi, \pi') = \frac{1}{|Q_r|\,H\,|\mathcal{G}|} \sum_{o \in Q_r} \sum_{h=1}^{H} \sum_{g \in \mathcal{G}} \mathbb{E}_\xi \big| a_{\pi,h,g}(o;\xi) - a_{\pi',h,g}(o;\xi) \big|. \tag{1}$$

Here $a_{\pi,h,g}(o;\xi)$ is coordinate g, from the specified group $\mathcal{G}$, of the h-th of H executed actions that π outputs at observation o; stochastic policies use paired sampling randomness ξ. Every policy receives the same observations. For chunked policies, the executed actions are the deployed prefix of the predicted chunk. Coordinate groups and scales are part of the metric: pooling unlike units can change closure even on identical predictions. Groups should be reported separately when a common physical scale is unavailable.

This comparison concerns the supplied observations, not agreement of closed-loop trajectory distributions. A bank overlapping the editing window gives an in-sample diagnostic, while retain-only rollouts need not cover states visited by the edited policy.

For J descriptive retrains, let $b_r(\pi) = J^{-1}\sum_j d_r(\pi, \pi_j^*)$ and let f_r be their mean pairwise divergence. Report these distances and gap closure

$$c_r(\pi) = \frac{b_r(\pi_{\theta_0}) - b_r(\pi)}{b_r(\pi_{\theta_0}) - f_r}. \tag{2}$$

Zero means no change from the original; one reaches the retrain-pair mean. Negative closure means divergence increased; closure above one passes the retrain-pair mean. These are effect-size anchors, not acceptance thresholds or degrees of deletion. Closure requires a resolvably positive denominator; pairs sharing a model are dependent.

3.2 Episode membership rank and calibrated loss

Let $\ell_\pi(e)$ be the mean frame loss for episode e under a fixed score interface and frame-sampling rule. Define $m_\pi = |\mathcal{D}_f|^{-1}\sum_{e\in\mathcal{D}_f}\ell_\pi(e)$. An audit non-member cohort $\mathcal{N}$ is excluded from original training and editing. Membership AUC u_π is the fraction of withdrawn/non-member episode pairs for which the withdrawn episode has lower loss, with half credit for ties. Episodes, rather than correlated frames, are the ranking units.

Every reference scores the same cohorts, accounting for difficulty differences that can put retain-only AUC above 0.5. Rank discards loss level, so we also report `mem/null`: m_π divided by mean member loss across descriptive retrains. When model normalization differs, first divide by that model's non-member loss. This choice is fixed by the score interface. Low ratios and high-loss overshoot reveal opposite departures from retraining. Training loss and action-residual loss can have different attack strength; episode resampling and independent retrains quantify different sources of variation. Across all three readings, the target is the corresponding retrain reference: zero action distance, AUC 0.5, and arbitrarily high withdrawn loss are not universal removal targets.

3.3 Joint and marginal decisions

Index the candidate by 0 and its K retain-only references by $1,\dots,K$. Each model is compared with its *complement*, the other K models. With complement-mean member loss $\bar{m}_{-i}$ and AUC $\bar{u}_{-i}$, define

$$z_i = \left(\frac{1}{K}\sum_{j\neq i} d_r(\pi_i,\pi_j),\ \left|\log\frac{m_i+\varepsilon}{\bar{m}_{-i}+\varepsilon}\right|,\ |u_i-\bar{u}_{-i}|\right), \qquad \varepsilon = 10^{-12}. \tag{3}$$

These coordinates measure action, loss-level, and rank discrepancies against the same retained-data counterfactual. The candidate participates in each reference's complement; a descriptive reference subset cannot replace it.

After computing all z_i, standardize coordinate k using the mean $\mu_{-i,k}$ and floored sample SD $\widetilde{\sigma}_{-i,k}$ of $\{z_{jk} : j \neq i\}$, then take the maximum:

$$t_{ik} = \frac{|z_{ik}-\mu_{-i,k}|}{\widetilde{\sigma}_{-i,k}}, \qquad s_i = \max_{k\in\{1,2,3\}} t_{ik}, \qquad p = \frac{1+\sum_{j=1}^{K}\mathbf{1}\{s_j \geq s_0\}}{K+1}. \tag{4}$$

The same numerical-floor rule is used for every model (Appendix B). Computing all coordinates before standardizing them preserves the candidate/reference symmetry: every model's score depends on the other models in the augmented set.

Under a null in which the augmented candidate/reference outputs are exchangeable, symmetric scoring gives a conservative finite-sample rank p-value (Vovk et al., 2005; Shafer & Vovk, 2008). This guarantee is marginal over exchangeable augmented samples, not conditional on every realized reference panel. Matching the retained dataset alone does not establish exchangeability; training and checkpoint selection also matter.

Reject at a prespecified α. A rejection detects a measured discrepancy; non-rejection establishes neither equivalence nor demonstration removal. With $K = 19$, p-values lie on a 0.05 grid: at $\alpha = 0.05$, the candidate must strictly exceed all 19 reference scores. A tied maximum does not reject. The test cannot distinguish stronger tail evidence below 0.05; more references improve resolution, while fresh nulls across independent panels are needed to assess operating calibration.

Replacing s_i by t_{ik} in Eq. 4 yields an unadjusted marginal p-value for coordinate k. Joint and marginal decisions can disagree, so report both with raw effects. Three-marginal Bonferroni cannot reject at 0.05 with this panel (minimum adjusted p-value 0.15). Keep α fixed across reference sizes; the checkpoint-wise tests do not control multiplicity over an edit sweep.

4 EXPERIMENTS

We ask whether task recovery establishes removal, whether actions and episode evidence identify different discrepancies, and how the audit behaves on independent candidates.

4.1 EXPERIMENTAL SETUP

A PiPER follower arm places a yellow cup inside a red cup. The dataset contains front and wrist RGB observations and six joint coordinates plus a gripper coordinate, recorded at 30 Hz. Thirty of 130 teleoperated demonstrations exhibit a displaced-release mode and are withdrawn. ACT, Diffusion Policy (DP), and $\pi_{0.5}$ share this split. ACT and DP retrain from scratch; $\pi_{0.5}$ repeats LoRA fitting from the same pretrained base, so its removal target concerns adapter-training data. Table 1 summarizes populations and reference roles.

Table 1: Experimental populations and reference roles. J: descriptive action references; E: descriptive evidence references; K: joint-test fleet; a dash marks an unmeasured component. Withdrawn and non-member columns count audited episodes. References, original-policy lineages, and rollout trials are distinct replication units.

Setting	J	E	K	Withdrawn	Non-member	Evaluation
ACT cup	3	5	19	30	10	100 physical trials (5 policies)
$\pi_{0.5}$ cup	–	3	–	30	10	100 physical trials (5 policies)
DP cup	4	3	–	30	10	Offline only
PushT harmful	–	5	–	80	80	250 rollouts per checkpoint
PushT benign	–	3	–	40	40	250 rollouts per checkpoint
Can validation	19	19	19	60	25	20 fresh nulls, 12 reintroduced
Can edits	19	19	19	60	25	86 edits, 500 rollouts each

The five conditions are the original policy, retain-only retraining, localized redirect, calibrated loss ascent, and budget-matched retain fine-tuning (FT). Redirect learns retrieved clean targets in a detected temporal window with retained replay; ascent raises withdrawn-data loss with retained-data anchors. Recorded budgets and monitoring rules are specified in Appendix A; R and B identify redirect and ascent checkpoints by update count. These are fixed interventions for evaluating unlearning measurements.

ACT and $\pi_{0.5}$ each have 20 blind-scored trials per condition, totaling 200 physical trials with matched initial cup positions. Success requires sustained seating of the yellow cup inside the red cup through the clip's end. DP is evaluated offline. The ACT action audit uses 71 BRANCH observations from the 30 withdrawn episodes and the first 25 of 50 predicted actions. BRANCH is the redirect training window, making this an in-sample action diagnostic. Main ACT results use six arm joints, separating the gripper whose physical unit is undocumented. This is a retrospective decomposition of saved predictions; the original pooled metric remains a sensitivity comparison. All

nine ACT joint and marginal tests are recomputed with arm-only distances and retain their original p-values.

PushT contains harmful mirrored demonstrations and a separate request withdrawing useful demonstrations; each evaluated checkpoint has 250 rollouts. Our Can study uses the robomimic multi-human dataset (Mandlekar et al., 2021), with our split of 180 retained, 60 withdrawn, and 25 audit non-member episodes. It includes 20 fresh retain-only nulls, three reintroduction seeds per dose (3, 15, 30, or 60 withdrawn episodes in one nested draw), and 86 corrected edits spanning 14 base objectives, tuned variants, and seed expansions. The prospective null/dose recipe trains for 2000 epochs and selects the most recent best-success checkpoint from 30-rollout evaluations every 100 epochs. A manifest fixed before training specifies a panel of 19 references (13 existing and six new). Corrected edits use the same request and 500 rollouts; the superseded request-mismatched cohort is excluded (Appendix F).

ACT's descriptive AUC averages five retrains and three frame draws. ACT and Can use action-residual evidence; DP and $\pi_{0.5}$ use recorded training-loss interfaces. For $\pi_{0.5}$, member loss is divided by non-member loss within each model before reference calibration. Joint tests use $K = 19$ and $\alpha = 0.05$. Models, frame draws, and rollouts are distinct replication units; full selection rules and cohort identities appear in Appendix A.

4.2 Task recovery does not establish demonstration removal

ACT redirect R200 increases physical success from 5/20 to 18/20, against 20/20 for retain-only retraining. Yet its membership AUC remains 1.000, compared with the retrain mean of 0.639 (Fig. 2): every audited withdrawn episode has lower loss than every non-member episode. Executed arm-action divergence decreases from 4.21 to 3.90 raw joint units against a retrain mean of 2.00: only 13.8% gap closure, versus 0.1% for matched FT. Across three independently trained ACT original policies, R200 closes 13.8%, 23.7%, and 17.9% while AUC stays 1.000; physical trials cover the primary lineage.

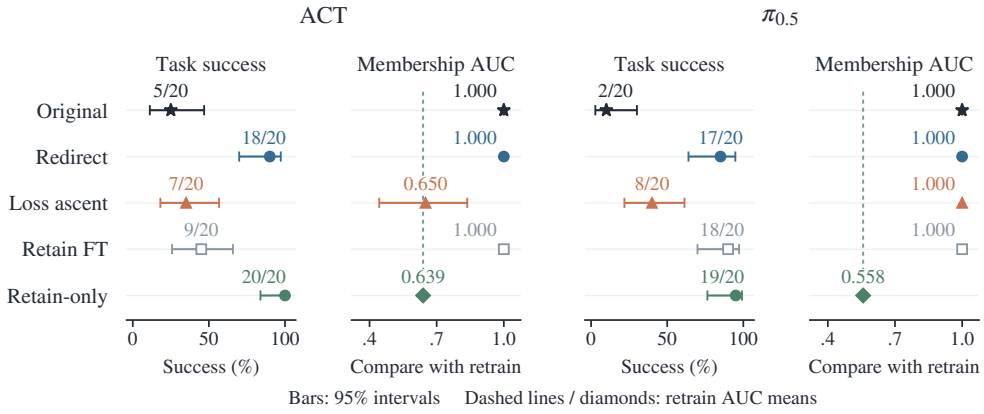


Figure 2: Physical recovery can leave membership evidence unchanged. Success bars are Wilson 95% intervals over 20 blind-scored trials per condition; candidate AUC bars are episode-bootstrap intervals from 30 withdrawn and ten non-member episodes. Green AUC diamonds/dashes show reference means (ACT: five retrains, three frame draws; $\pi_{0.5}$: three retrains); retrain success uses one deployed model. ACT edits are R200, B200, and FT200; $\pi_{0.5}$ uses its fixed operating points.

The disagreement recurs for $\pi_{0.5}$: redirect and FT recover 17/20 and 18/20 successes from 2/20, against 19/20 for retraining, but both retain AUC 1.000. This arm extends the task/evidence finding to a second policy class; it has no matched-state action-floor audit. Similar observed success counts also do not establish statistical equivalence to retraining.

The withdrawal request determines how success should change. In harmful PushT withdrawal, original policies average 57.9% success across three seeds, against 72.6% across five retain-only retrains. In benign withdrawal, removing 40 useful demonstrations lowers success from 80.8% to a mean of 62.5% across three retrains. Frozen-dose ascent retains harmful-request AUCs of

0.979–0.997 versus a retrain mean of 0.534; benign ascent and FT retain AUCs 0.961 and 0.968 versus 0.447 (Appendix F). Thus neither increased success nor suppression of a harmful behavior defines the removal target.

4.3 Actions and episode evidence expose complementary discrepancies

ACT ascent B200 brings AUC to 0.650, near the retrain mean of 0.639, while action divergence increases. At B400, AUC reaches 0.533, but member loss is 1.83 times its reference and arm-action divergence is 8.04 raw joint units (Fig. 3). Favorable rank can therefore coexist with action divergence and loss overshoot. High loss indicates a retrain discrepancy; it need not indicate preserved membership information.

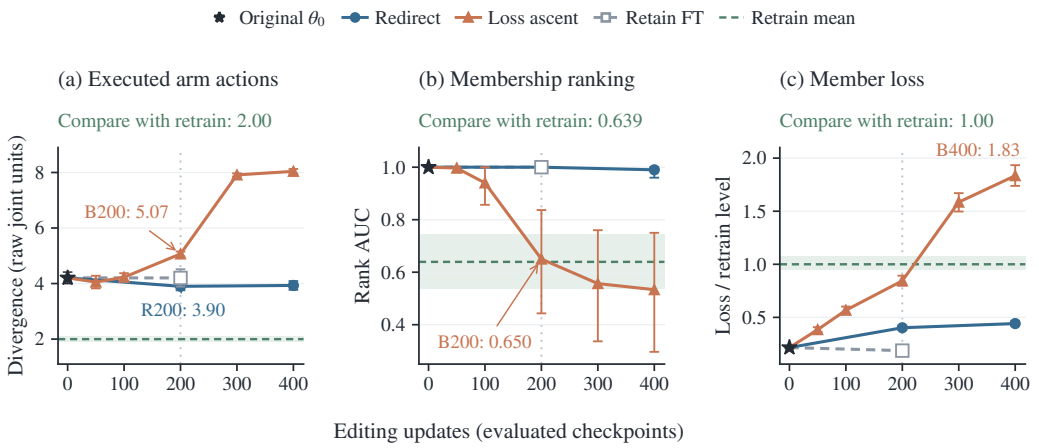


Figure 3: ACT measurements diverge along the same edit sequence. Vertical guides align 200 updates: B200 approaches reference AUC while its actions depart; B400 overshoots reference loss. Lines join evaluated checkpoints. Bars show action sample SD across three retrains and episode-bootstrap 95% intervals for rank/loss (loss conditional on five references). Green shading describes retrain variation: action mean $\pm$ sample SD across three pairs, AUC mean $\pm$ two SD across five seeds and three frame draws, and the range of five reference loss ratios. These bands are not decision boundaries.

The strong overshoot is a primary-lineage counterexample. Final loss ratios across three ACT lineages are 1.83, 0.91, and 1.19, with AUCs 0.533, 0.913, and 0.863. DP provides another rank/level disagreement: ascent AUC is 0.560 versus 0.534 for retraining, with a loss ratio of 1.52. Its original action gap is poorly resolved, so closure is not interpreted.

The complementary instruments detect both directions of disagreement. ACT B200 rejects on actions ($p = 0.05$) but not rank ($p = 0.50$) or loss ($p = 0.10$). A corrected Can loss-ascent edit (seed 9) has 58.4% success and a loss ratio of 0.60: loss rejects ($p = 0.05$), while actions ($p = 0.10$) and rank ($p = 0.20$) do not. Stronger fit to withdrawn episodes can coexist with similar member/non-member ordering. Both joint p-values are 0.05. These tests use separate 19-reference panels; marginals are unadjusted. The Can example is selected retrospectively from 86 edits to illustrate complementarity, not compare power. Marginal non-rejection does not establish agreement.

Across the corrected Can edits, the joint rule rejects 82/86; action, loss, and rank reject 69, 68, and 60. Relative to action alone, the joint rule rejects 14 additional edits but misses one action rejection. All 15 edits with success at least 30% reject jointly, including the two within a descriptive ± 0.10 AUC band around the reference. The four joint non-rejections have 0.2–26.0% success and each has a rejecting unadjusted marginal. These are shared-lineage checkpoint counts, not independent tasks or a ranking of published methods. The joint rule is not an "any marginal rejects" rule; on the nine primary ACT checkpoints, it adds no detections to action alone.

Regional evidence explains another distinction. ACT R200 moves BRANCH loss from 0.26 to 1.07 times the regional reference while TAIL remains at 0.20. BRANCH accounts for only 8.6% of

withdrawn frames. A local repair can leave episode-wide separation; the profiles do not identify a causal location of stored information.

4.4 Validation with independent candidates

Fresh retain-only candidates test the audit on new models generated without the withdrawn data. Controlled reintroduction tests known changes in training exposure independently of an editing algorithm. Its fraction measures exposure, not memorization; both comparisons use the prospectively fixed recipe in Sec. 4.1. Table 13 in Appendix E reports the complete marginal and joint counts.

At $\alpha = 0.05$, the joint rule rejects 2/20 nulls: an observed 10%, with a one-sided 95% binomial upper bound of 28.3% conditional on this reference panel. It detects 0/3, 2/3, 3/3, and 3/3 candidates at 5%, 25%, 50%, and 100% reintroduction, respectively. At 50%, rank and loss detect all three while actions detect none. At 25%, rank detects all three but the joint rule only two; marginal tests are unadjusted. Across the 12 exposed candidates, the joint rule detects eight, versus one for actions, nine for rank, and six for loss. It gains complementary detections without uniformly preserving marginal sensitivity. This pilot is too small to establish a 5% operating error rate: larger null cohorts address precision, additional references address p-value resolution, and independent panels address reference-panel variability. The three seeds per dose and one nested episode draw likewise do not establish a general power curve.

The two rejected nulls have joint scores only 1.012 and 1.061 times their respective largest reference scores. For every dose, the audit scores all 60 withdrawn episodes: exposure changes while the measurement cohort stays fixed. The three seeds replicate training randomness, not independent withdrawal requests.

Both null detections arise from actions, and selected epoch correlates with divergence at $r = 0.95$. A retrospective comparison evaluates the same references, nulls, and alternatives at epoch 2000. Joint null rejection becomes 0/20 and dose detection becomes 2/3, 3/3, 3/3, and 3/3. Pairwise reference divergence changes from 0.079 to 0.132. This demonstrates selection-rule sensitivity on reused runs; it does not replace the prospective result or establish improved calibration (Appendix E).

The pilot's 44 new training runs took approximately 18 workstation hours with three concurrent jobs. Auditing took under a minute using 21 cached and 44 newly evaluated banks, excluding cold evaluation of all 65 models. Training dominates this measured cost. References are reusable for a fixed retained set and training recipe; a new request generally requires new references.

4.5 Dependence on the audit protocol

ACT predicts 50 actions but executes only the first 25 before replanning. The same R200 checkpoint has 43.8% arm-action closure on the full prediction and 13.8% on the executed prefix (Fig. 4). Extending from 25 to 50 steps raises original divergence from 4.21 to 5.95, but edited divergence only from 3.90 to 4.19, while the floor changes from 2.00 to 1.93. Full-plan averaging credits correction outside the deployed prefix.

Gripper-only executed-prefix closure is 97.2%, compared with 13.8% for arm joints. Pooling all seven raw coordinates gives 24.9% executed-prefix and 75.1% full-plan closure. Thus aggregation changes the reported magnitude substantially. The horizon effect persists for arm joints in all three lineages (Appendix D); no policy or deployment changes between readings.

For Can, a retrospective bank pools 45,492 observations from six separate retain-only rollout generators. Replacing demonstration observations with this bank, while holding episode evidence fixed, yields fewer null rejections and fewer moderate-dose detections (Table 2). Rank still detects all three candidates at 25% and 50% exposure; the changed action-reference scores reduce combined sensitivity. The 20 null and 12 dose candidates are distinct from the generators. Audits of generators on their own pooled bank are descriptive and excluded here.

A larger bank can reduce both null rejection and exposure detection. Retained-policy rollouts need not cover edited-policy states; report the observation source with every decision.

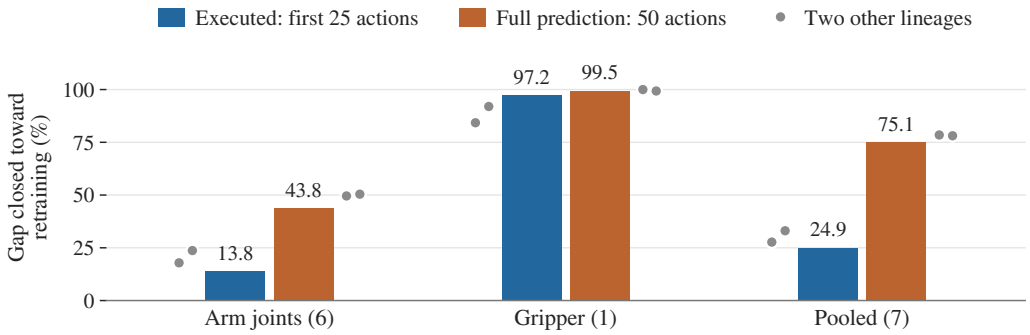


Figure 4: Executed horizon and action coordinates change R200's apparent progress. Columns show primary-lineage gap closure for the 25 executed actions and the full 50-action prediction at fixed observations; gray markers show the two other independently trained original-policy lineages. All use the same three retrains; coordinates are averaged in raw units. This retrospective decomposition changes neither checkpoints nor deployment. Appendix D gives raw distances and further sensitivities.

Table 2: Can joint decisions under different observation banks, with candidate best-success checkpoints and episode evidence fixed ($K = 19$, $\alpha = 0.05$). Rollouts come from six other retain-only models. This retrospective comparison uses one generator pool and one nested exposure draw.

Action bank	Nulls	5% dose	25% dose	50% dose	100% dose
Demonstrations	2/20	0/3	2/3	3/3	3/3
Independent-generator rollouts	0/20	0/3	0/3	1/3	3/3

For the original pooled diagnostic, expanding ACT's references from three to 19 changes R200 closure from 24.9% to 22.6%. Three-reference subsets span a central 95% range of 9.3–36.7%; four BRANCH windows yield 4.4–54.5%, with unchanged joint rejection. These are sensitivities within one fleet and bank, not independent-experiment intervals. Stable rejection need not imply stable effect size (Appendix D).

5 Discussion

The findings change what a successful edit establishes: task recovery demonstrates useful repair, while removal claims require comparison with the specified retained-data counterfactual. Benign withdrawal makes this distinction especially clear because utility need not improve after deletion. Selection and observation coverage also shape the discrepancies an audit can detect.

The empirical scope remains limited. Hardware covers one task family, sequential collection, one deployed lineage per policy class, and ten non-members. Can uses a small null cohort and one nested withdrawal draw. Stronger calibration evidence requires larger fresh null cohorts and independent reference panels; closed-loop coverage requires states visited by edited policies. The coordinate, selection, and coverage analyses remain retrospective.

6 Conclusion

Successful robot repair is insufficient evidence of demonstration unlearning. Physical task recovery, favorable membership rank, and local action correction can each coexist with discrepancies from retain-only retraining. Our audit makes those discrepancies measurable through complementary executed-action and episode-evidence comparisons, grounded in reference variation. Independent-candidate tests show both useful detection and the limits imposed by selection and coverage. Demonstration-unlearning claims should therefore be judged against the specified retained-data counterfactual in both actions and evidence, with task utility reported as a separate outcome.

AI USE STATEMENT

In this work, we used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and interpret results. We have not used generative AI tools to assist with translation or to support qualitative and thematic data analysis, and formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to source and search for information. We have reviewed all AI-assisted work. The physical robot trials were conducted by the authors, and headline results were recomputed from saved model outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The physical experiments use teleoperated demonstrations and robot trials; simulation uses public benchmarks and constructed withdrawal scenarios. The audit is an empirical comparison with a specified retraining procedure and does not establish legal compliance or a privacy guarantee.

REPRODUCIBILITY STATEMENT

The supplementary material provides episode identities, reference and candidate roles, measurement conventions, model configurations, and code for the audit, the figures, and the main analyses. The appendix separates original prospective analyses from retrospective comparisons. Historical runs with a request mismatch are excluded from the corrected baseline evidence.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 141–159, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 1897–1914, 2022.

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

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *NeurIPS*, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, 2020.

Chen Gong, Kecen Li, Jin Yao, and Tianhao Wang. TrajDeleter: Enabling trajectory forgetting in offline reinforcement learning agents. In *Network and Distributed System Security Symposium (NDSS)*, 2025. URL https://arxiv.org/abs/2404.12530.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning (ICML)*, 2020.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519, 2025. URL https://arxiv.org/abs/2403.01218.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *NeurIPS*, 2023.

Jiazhuo Li, Yu Zhang, Yiming Fei, Kangkang Dong, Xiaojun Zhu, Houde Liu, and Jinze Tao. Rethinking demonstration unlearning in imitation learning for robotics. *arXiv preprint arXiv:2608.20784*, 2026. doi: 10.48550/arXiv.2608.20784. URL https://arxiv.org/abs/2608.20784v1.

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)*, pp. 1678–1690, 2021.

Chaofan Pan, Lingfei Ren, Xiangyu Jiang, Yanhua Li, Xuemei Cao, Xiangkun Wang, Hao Yu, Wei Wei, and Xin Yang. TOUR: A trajectory-level unlearning benchmark for offline reinforcement learning. *arXiv preprint arXiv:2607.21111*, 2026. URL https://arxiv.org/abs/2607.21111.

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. In *Conference on Robot Learning (CoRL)*, volume 305 of *PMLR*, pp. 17–40, 2025. URL https://arxiv.org/abs/2504.16054.

Ravi Ranjan and Agoritsa Polyzou. VLA-Forget: Vision-language-action unlearning for embodied foundation models. In *Proceedings of the 4th Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pp. 60–77. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.knowfm-1.5/. arXiv:2604.03956v2.

Mónica Ribero, Antonin Schrab, and Arthur Gretton. Regularized f-divergence kernel tests. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4249–4257. PMLR, 2026. URL https://proceedings.mlr.press/v300/ribero26a.html.

Sahasrajit Sarmasarkar, Anastasia Koloskova, and Sanmi Koyejo. Auditing of unlearning algorithms. *arXiv preprint arXiv:2607.05898*, 2026. URL https://arxiv.org/abs/2607.05898.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *NeurIPS*, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.06460.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Sun Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? Findings from the first NeurIPS unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024. URL https://arxiv.org/abs/2406.09073.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Hao Wang, Jiuzhou Lei, Dayou Li, Bangya Liu, Minghui Zheng, Manling Li, Ruohan Zhang, and Zhiwen Fan. Behavior uncloning: Distilling mode redirection into policy weights without inference-time steering. *arXiv preprint arXiv:2606.29201*, 2026. URL https://arxiv.org/abs/2606.29201.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.

Zhengyang Yan, Junhao Li, Fangqi Zhu, Zijun Wang, Quanxin Shou, Yikun Miao, Xiaoyi Pang, Zicong Hong, and Song Guo. RedFlow: Redirect failure into action-level corrections for flow-matching VLA policy. *arXiv preprint arXiv:2607.27782*, 2026. URL https://arxiv.org/abs/2607.27782.

Jiahao Zhang, Yilong Wang, and Suhang Wang. Attack by unlearning: Unlearning-induced adversarial attacks on graph neural networks, 2026a. URL https://arxiv.org/abs/2603.18570. arXiv:2603.18570.

Jiahao Zhang, Yilong Wang, Zhiwei Zhang, Xiaorui Liu, and Suhang Wang. Unlearning inversion attacks for graph neural networks. In *ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 934–945, 2026b. doi: 10.1145/3773966.3777929. arXiv:2506.00808.

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.

A Experimental Protocol and Interventions

Physical task and data. A PiPER follower arm transports a yellow cup into a red cup. Recorded observations contain front and wrist RGB images (640×480) and seven state coordinates; the action vector contains six joint commands and a gripper command. The dataset is recorded at 30 frames/s. These recording settings do not establish the wall-clock rate of deployed inference. The original training set contains 100 retained demonstrations and 30 withdrawn demonstrations exhibiting a displaced-release mode. The latter are episodes 0–29 of the 50-episode collection; episodes 40–49 are the ten audit non-members. ACT and DP retrain from scratch; the $\pi_{0.5}$ reference repeats adapter fitting from the same fixed pretrained base, so its removal target concerns adapter-training data.

Success requires the yellow cup to remain seated inside the red cup through the end of the evaluated clip. Resting on the rim, being knocked out after placement, and remaining in the gripper at timeout are failures. ACT and $\pi_{0.5}$ each have five deployed conditions with 20 trials per condition. Initial cup positions are matched; models were collected sequentially, so position matching does not remove session drift. The 200 anonymized clips were shuffled and scored before opening the condition mapping. Blind labels are authoritative and agree with live labels on 199/200 clips. DP contributes offline measurements only. After a mid-collection rig fault and a 29-minute gap, the protocol selected the original ACT policy's four complete consecutive five-trial blocks before scoring; no clips within these blocks are dropped.

Matched observations and temporal regions. The ACT action bank contains 24 uniformly spaced valid chunk-start observations from each of 30 withdrawn and 20 retained episodes (1200 observations). All policies receive the same recorded observations; ACT uses deterministic zero-latent decoding and its saved postprocessor to recover raw action units. Its 50 predicted actions are stored, and the first 25 executed actions define the primary comparison. The bank contains 71 withdrawn BRANCH observations. These are demonstration states, not observations collected under edited-policy rollouts.

The branch detector compares recorded 50-action chunks with clean chunks in a $\pm$30-frame grasp-relative phase band. It uses a sustained high-threshold excursion (2.0 for three frames) and a backward low-threshold walk (0.6, with a five-frame gap limit) to locate onset. BRANCH is the fixed interval from five frames before onset through 20 frames after it, and is the same window used to train redirect. It is therefore an *in-sample* regional diagnostic. PRE precedes this window; POST runs from its end through detected release; TAIL contains the remainder. Release is clamped to lie after the window. The 30 episodes contain 9028 frames: 4203 PRE, 780 BRANCH, 702 POST, and 3343 TAIL. All-frame regional loss measurements and the subsampled action bank have different sample counts.

Reference roles and evidence access. ACT's descriptive action, descriptive evidence, and joint references contain respectively three, five, and 19 retrains; the smaller sets are nested in the joint panel. Five evidence seeds each have three frame draws. DP has four action references and three evidence references; its primary action horizon is eight steps. No matched-state pairwise action floor is available for $\pi_{0.5}$; three adapter fits calibrate its evidence. Its 96 recorded loss samples per split are not an action-bank count. The exact DP bank count was not recovered in the summary artifacts used for the original manuscript.

ACT evidence uses masked mean chunk-L1 between deterministic zero-latent predictions and recorded actions in the policy's normalized coordinates; the posterior latent and KL term are excluded. DP and $\pi_{0.5}$ use recorded denoising and flow-matching losses, respectively. Can uses action-residual losses from queried outputs and recorded demonstration targets. The Can implementation does not establish equivalence to diffusion- or flow-loss attacks. Frame scores are averaged within episodes; AUC compares withdrawn episodes with non-members, and retrain models score the same cohorts. Calibrated member loss divides by the pooled retrain level. For $\pi_{0.5}$, first divide member loss by non-member loss within each model, then compare with the pooled retrain ratio, because model normalization statistics differ.

Interventions. Retain fine-tuning, localized redirect, and calibrated loss ascent use the recorded budgets and retained-data stream. They are interventions for evaluating measurements, not a claimed new removal algorithm. ACT and PushT use AdamW. ACT uses learning rate 3×10^{-5}; the stable

PushT and benign-request setting uses 2.5×10^{-6} and ratio 0.20. Edit monitors compare retained loss with 1.25 times its original level and calibration loss with 1.5 times its original level; run configurations determine monitoring frequency, patience, and whether a trip aborts the run. For example, ACT R200 records 50-update monitoring, patience two, and guard-abort override enabled, with no recorded trip and termination at its 200-update budget. The $\pi_{0.5}$ implementation follows its separate recorded adapter-training configuration.

For ascent, a forget example with loss ℓ contributes $\mathrm{softplus}(\tau - \ell)\mathbf{1}[\ell < \tau]$. The cap τ is the evolving policy's 0.65-quantile calibration loss, recomputed every 50 updates in the reported setting. ACT weights the forget term by retained-neighbor action discrepancy, adds a retained-neighbor training-loss anchor, and penalizes the MSE between current and frozen-original deterministic action chunks on retained states (anchor and proximity weights 3). The proximity penalty is in function space. A global stop occurs if the forget/calibration loss gap falls to 10% of its initial value or crosses zero. This is not a cap fixed to the retain-only retrain level.

Redirect minimizes deployment-path L1 distance to retrieved clean action targets within BRANCH. One clean episode supplies each entire window, selected by phase-compatible state and lag-five velocity matching, with a cap of two withdrawn episodes per clean source. The target window plus its 50-step action chunk must fit without padding. The forget gradient is scaled to $\rho\|g_r\|$ (ACT $\rho = 0.32$), then its conflicting component is removed:

$$g_f \leftarrow g_f - \frac{\langle g_f, g_r \rangle}{\|g_r\|^2} g_r \quad \text{if } \langle g_f, g_r \rangle < 0.$$

The PushT norm-PCGrad ascent variant uses the same projection; ACT's capped-ascent implementation optimizes its summed loss directly. The supplementary material includes the episode, branch, and region manifests.

B Audit Construction and Statistical Interpretation

Symmetric candidate comparison. Equation 3 defines the coordinates for the augmented candidate/reference set. Each reference's complement includes the candidate; all coordinate vectors are computed before standardization. For each coordinate, use complement mean $\mu_{-i,k}$ and sample SD $\sigma_{-i,k}$ to form

$$s_i = \max_k \frac{|z_{ik} - \mu_{-i,k}|}{\widetilde{\sigma}_{-i,k}}, \quad \widetilde{\sigma}_{-i,k} = \max\left\{\sigma_{-i,k}, 10^{-12}, 10^{-3}\,\mathrm{mean}_{j \neq i}\,|z_{jk}|\right\}.$$

The conservative rank in Eq. 4 is finite-sample valid under exchangeable augmented inputs (Vovk et al., 2005; Shafer & Vovk, 2008). The guarantee is marginal, not conditional on every frozen panel. A different editing procedure is admissible if null exchangeability holds. Checkpoint, region, and sensitivity sweeps have no familywise guarantee. With $K = 19$, the minimum p-value is 0.05; three-marginal Bonferroni has minimum adjusted p-value 0.15.

ACT's reference inclusion rule admits every retain-only replica of the same recipe with both bank and score files, fixed before scoring the candidates. The 19-reference panel includes the 14 replicas added after coordinate design under this rule. The code checks duplicate model identities, candidate/reference overlap, split/region alignment, and finite scores; episode and frame order across banks were checked separately. Table 3 reports all nine candidates. Retrospectively restricting saved actions to six arm joints and recomputing every candidate/reference score preserves every pooled-coordinate joint and marginal p-value. Episode/frame identities, evidence, horizon, and all 19 references are held fixed. This is a metric-sensitivity check, not independent calibration evidence.

Rank, level, and uncertainty. A strictly increasing score transformation preserves all member/non-member pairwise orderings, hence AUC, while changing absolute score levels. Consequently AUC cannot recover absolute proximity to the retrain loss level. This observation motivates the loss coordinate; the empirical B200/B300 comparison has different, rather than mathematically identical, rankings and is not a literal transformation example.

Episode bootstrap intervals (2000 resamples) describe the sampled cohorts. Retrain seeds and frame draws are different sources of variation: ACT's descriptive AUC has frame-draw SD 0.039, seed SD

Table 3: ACT marginal and joint p-values, identical for the recomputed arm-only and original pooled metrics. Each candidate uses 19 retrains. Values are unadjusted across checkpoints; 0.05 is the attainable minimum.

checkpoint	action	loss level	rank AUC	joint
θ_0	0.050	0.050	0.050	0.050
redirect R200	0.050	0.050	0.050	0.050
redirect R400	0.050	0.050	0.050	0.050
ascent B50	0.050	0.050	0.050	0.050
ascent B100	0.050	0.050	0.050	0.050
ascent B200	0.050	0.100	0.500	0.050
ascent B300	0.050	0.050	0.050	0.050
ascent B400	0.050	0.050	0.050	0.050
FT200	0.050	0.050	0.050	0.050

0.021, and total SD 0.052. Fifteen seed/draw readings are not 15 independent retrains. Ten hardware non-members imply empirical FPR increments of ten percentage points; precise low-FPR evaluation is unsupported. Reference pairs share models and are not independent training replications. A pairwise mean action floor is descriptive, not an equivalence threshold.

Reassigning each of 19 retrains in turn as a candidate against the other 18 only reranks one fixed augmented collection. It is an implementation check, not an independent false-rejection study; its minimum p-value $1/19$ also exceeds 0.05. Independent null candidates are evaluated in Appendix E.

C Complete Hardware and ACT Results

Table 4 contains all deployed conditions and blinded outcomes. Paired contrasts use matched initial positions; their interpretation retains the sequential-session limitation. Physical ACT trials concern one original lineage, whereas offline ACT results include three. The two policy classes share one physical task family.

Table 4: All 200 blind-scored physical trials. Wilson intervals summarize each 20-trial condition. McNemar contrasts are unadjusted, paired by initial position; "ceiling" denotes deployed retain-only retraining. The dominant failure-mode count is not the total failure count.

arm	condition	success	Wilson 95% CI	McNemar vs. ceiling	dominant failure mode
ACT	ceiling	20/20	[0.84, 1.00]	—	—
ACT	redirect	18/20	[0.70, 0.97]	b=2, c=0, p=0.500	dropped_short (1)
ACT	FT	9/20	[0.26, 0.66]	b=11, c=0, p<0.001	sway_and_knock (8)
ACT	ascent	7/20	[0.18, 0.57]	b=13, c=0, p<0.001	sway_and_knock (10)
ACT	θ_0	5/20	[0.11, 0.47]	b=15, c=0, p<0.001	sway_and_knock (14)
$\pi_{0.5}$	ceiling	19/20	[0.76, 0.99]	—	dropped_short (1)
$\pi_{0.5}$	redirect	17/20	[0.64, 0.95]	b=3, c=1, p=0.625	dropped_short (3)
$\pi_{0.5}$	FT	18/20	[0.70, 0.97]	b=2, c=1, p=1.000	dropped_short (1)
$\pi_{0.5}$	ascent	8/20	[0.22, 0.61]	b=12, c=1, p=0.003	sway_and_knock (9)
$\pi_{0.5}$	θ_0	2/20	[0.03, 0.30]	b=17, c=0, p<0.001	sway_and_knock (17)

Table 5: $\pi_{0.5}$ execution summaries. Angles are mean±SD. Release statistics use 20/19/17/18/18 detected events in row order; grasp uses 20 each. Closed intervals are median recorded steps divided by 30, with undetected release ending at the last recorded step; these are nominal seconds, not measured elapsed time.

Condition	Success	Release (°)	Grasp (°)	Closed interval
Retain-only retrain	19/20	15.1 ± 1.6	-20.6 ± 0.8	2.95
Redirect	17/20	14.3 ± 1.5	-16.6 ± 5.8	2.43
Retain FT	18/20	13.8 ± 2.3	-15.0 ± 8.8	2.87
Ascent	8/20	13.3 ± 1.8	-15.5 ± 7.0	2.33
Original	2/20	11.3 ± 1.5	-17.2 ± 5.9	1.92

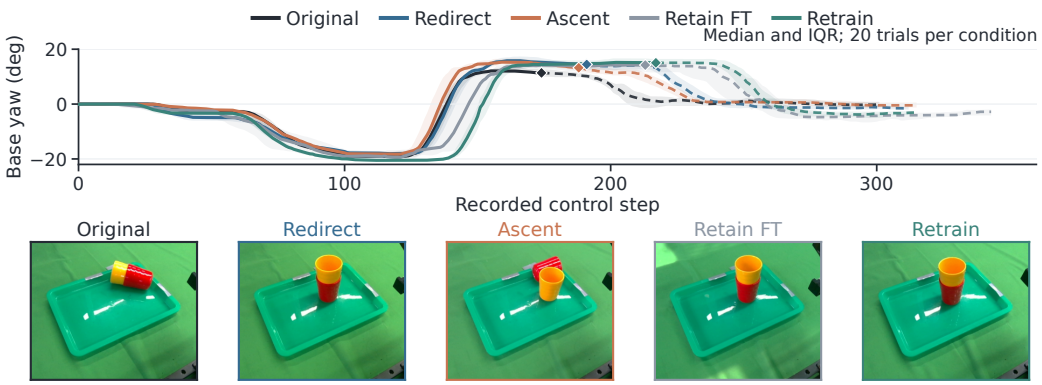


Figure 5: Measured $\pi_{0.5}$ execution and representative terminal states. Top: median base yaw and interquartile range over 20 trials, aligned by recorded control step. Diamonds mark median estimated release; dashed segments follow it. Bottom: actual terminal frames nearest each condition's median endpoint centroid, selected from all 20 trials with a fixed crop. These illustrations do not determine the aggregate success counts in Table 4.

The primary ACT membership results appear in Table 6. All three original ACT lineages have AUC 1.000 and calibrated member loss 0.19–0.22. Their final ascent loss ratios are 1.83, 0.91, and 1.19, with AUC 0.533, 0.913, and 0.863. The strong overshoot is therefore a primary-lineage counterexample, not a uniform ascent endpoint. The evolving cap saturates at 0.547/0.362/0.413, against retrain member loss 0.461; all three runs reach full budget without guard trips. These are associations with the stopping rule, not evidence of a causal mechanism.

Table 6: ACT episode evidence, primary lineage. AUC intervals are episode-bootstrap 95% intervals. The descriptive reference AUC is 0.639, from five retrain seeds and three frame draws. The loss ratio uses the pooled level of five references.

Candidate	AUC [95% CI]	AUC minus reference	Calibrated loss
Original	1.000 [1.00,1.00]	+0.36	0.22
FT200	1.000 [1.00,1.00]	+0.36	0.19
R200	1.000 [1.00,1.00]	+0.36	0.40
R400	0.990 [0.96,1.00]	+0.35	0.44
B50	0.997 [0.98,1.00]	+0.36	0.38
B100	0.940 [0.86,1.00]	+0.30	0.57
B200	0.650 [0.44,0.84]	+0.01	0.84
B300	0.557 [0.34,0.76]	−0.08	1.59
B400	0.533 [0.30,0.75]	−0.11	1.83

Table 7: Pooled seven-coordinate ACT closure across three lineages (%). Each lineage uses its own original policy and the same three-reference floor. The executed horizon is primary; full-plan values are diagnostic. These descriptive point estimates are not confidence intervals.

	primary lineage		lineage b		lineage c	
	exec. 25	full 50	exec. 25	full 50	exec. 25	full 50
redirect R200	+24.9	+75.1	+33.1	+78.5	+27.7	+78.1
redirect R400	+23.4	+75.0	+9.7	+71.3	−1.1	+67.9
ascent B200	−57.8	−52.4	+48.4	+62.3	+8.9	+42.2
ascent B300	−184.1	−128.5	+18.5	+61.4	−96.0	+20.1

Regional evidence in Table 8 describes the localization of changes in measured loss. It does not identify where information is stored. The DP ascent example has AUC 0.560 versus retrain mean 0.534 and calibrated loss 1.52. Its original-to-floor action gap is not resolved relative to reference

Table 8: All-frame regional member loss relative to retraining. The denominator is the pooled per-region mean of three retrains. BRANCH is redirect's training window; closeness to one is not a formal equivalence decision.

region (frame share)	θ_0	FT200	redirect R200	ascent B300
PRE (46.6%)	0.33	0.32	0.52	2.31
BRANCH (8.6%)	0.26	0.21	1.07	2.19
POST (7.8%)	0.20	0.17	0.68	1.27
TAIL (37.0%)	0.17	0.13	0.20	1.30

spread, so no gap-closure fraction is interpreted. For $\pi_{0.5}$, the original and all audited edits retain AUC 1.000 against reference 0.558 ± 0.012, with calibrated losses 0.10–0.44.

D Action and Auditor Sensitivity

Table 9: Primary-lineage ACT divergences pooling seven raw postprocessed coordinates on withdrawn regions. The floor is mean±SD over three reference pairs; candidate entries are mean divergence to three retrains. Both horizons use the same aligned bank.

	executed slice (25 steps, primary)				full plan (50 steps, diagnostic)			
	BRANCH	POST	PRE	TAIL	BRANCH	POST	PRE	TAIL
floor (3 pairs)	1.81 ± 0.07	2.01 ± 0.08	2.83 ± 0.14	7.27 ± 0.97	1.83 ± 0.06	3.27 ± 0.03	3.44 ± 0.11	8.93 ± 0.64
θ_0	3.99	13.25	4.48	16.51	9.70	16.86	6.03	22.73
R200	3.45	7.21	4.10	15.90	3.79	9.02	5.09	22.12
R400	3.48	6.27	3.84	16.37	3.80	8.06	4.72	22.76
B50	3.96	14.09	4.54	20.58	10.49	20.14	6.02	25.82
B100	4.30	14.28	4.80	21.72	11.96	20.46	6.45	26.93
B200	5.25	15.84	5.70	23.44	13.83	22.21	7.60	28.60
B300	8.01	19.26	9.90	29.12	19.82	26.84	13.73	33.49
B400	7.97	18.68	11.56	30.15	19.98	26.28	16.18	34.43
FT200	3.92	12.23	4.31	16.12	9.58	16.04	5.72	22.24

The pooled 25-step original and R200 BRANCH divergences are 3.99 and 3.45, against floor 1.81; the corresponding 50-step values are 9.70, 3.79, and 1.83. The full-plan closure is larger because the original discrepancy grows mainly outside the executed prefix. Recomputing from the saved banks reproduces the full-plan BRANCH values within 5×10^{-3}.

Table 10: Coordinate-group sensitivity of R200 gap closure (%). Saved-bank reanalysis with the same 71 BRANCH observations. Each group recomputes its own original divergence and reference floor. These retrospective descriptive comparisons do not supply equivalence tests.

Lineage	J	Horizon	All seven	Six arm joints	Gripper
primary	3	25	24.91	13.81	97.23
primary	3	50	75.08	43.75	99.45
b	3	25	33.08	23.69	91.97
b	3	50	78.47	49.57	99.99
c	3	25	27.70	17.88	84.26
c	3	50	78.11	50.42	99.32
primary	19	25	22.64	11.77	101.59
primary	19	50	74.16	42.09	99.33

The raw seven-coordinate mean combines arm joints with the gripper without a common physical-unit conversion. The dataset metadata does not specify the gripper's physical unit. Removing it reduces the primary R200 executed-prefix closure to 13.8%; the gripper coordinate alone reaches 97.2%. The arm-only full-plan/prefix contrast remains 43.8/13.8%, 49.6/23.7%, and 50.4/17.9% across lineages. Thus the horizon effect persists when the gripper is excluded, while its magnitude depends on the coordinate aggregation.

Table 11: Pooled-coordinate ACT reference-size sensitivity. Nested panels are drawn from the same 19 available references on 71 BRANCH observations and 25 executed actions. These are descriptive finite-panel comparisons.

J	Floor	Original	R200	Gap closed (%)
3	1.810	3.992	3.449	24.9
5	1.715	3.990	3.489	22.0
10	1.620	4.030	3.526	20.9
19	1.668	4.003	3.474	22.6

Across all 969 three-reference subsets, the central 95% range of R200 closure is 9.3–36.7%. Sampling 5000 distinct subsets each at five and ten references gives 13.0–32.6% and 16.9–28.1%. These are not confidence intervals from independent fleets. The 19-model panel contains 171 overlapping pairs, not 171 independent retraining replications. The full-fleet R200 closure is 22.6%, versus 24.9% with the original three-reference convention.

Table 12: Pooled-coordinate ACT observation-bank and horizon sensitivity. Entries are gap closure (%); subsample entries show mean±SD over 40 draws. Every draw rejects all nine candidates with both action and joint tests. Window changes use split-aware within-episode ordering; the empty window is not estimable.

perturbation	rejected	R200	R400	B200	B300
episode subsample (same draw for every model)					
25% (8 of 30), 40 draws	9/9	$+25.4 \pm 20.5$	$+25.9 \pm 20.2$	-58.3 ± 12.7	-187.7 ± 32.2
50% (15 of 30), 40 draws	9/9	$+23.7 \pm 15.5$	$+24.0 \pm 15.9$	-58.9 ± 7.5	-187.4 ± 21.2
100% (as reported)	9/9	$+24.9$	$+23.4$	-57.8	-184.1
BRANCH window, banked observations per side					
-2 per side (0 obs.)	–	–	–	–	–
-1 per side (12 obs.)	9/9	$+4.4$	$+17.1$	-28.6	-168.1
0, as reported (71 obs.)	9/9	$+24.9$	$+23.4$	-57.8	-184.1
$+1$ per side (131 obs.)	9/9	$+54.5$	$+57.4$	-39.0	-120.8
$+2$ per side (191 obs.)	9/9	$+48.6$	$+55.1$	-32.0	-100.0
horizon					
1 step	9/9	-82.6	-78.3	-116.9	-311.8
10 steps	9/9	-31.8	-34.6	-57.1	-196.0
25 steps (primary)	9/9	$+24.9$	$+23.4$	-57.8	-184.1
50 steps (full plan)	9/9	$+75.1$	$+75.0$	-52.4	-128.5

Decision stability does not imply stable effect magnitude: R200 closure spans 4.4–54.5% across nonempty window changes. These perturbations reuse demonstration observations; they add no new states reached by an edited policy.

Alternative loss summaries. Standardizing each episode's loss against five retrains preserves the main ACT distinction: redirect remains below the reference (mean $z = -10.8/-10.1$ for R200/R400), whereas B300/B400 exceed it ($+10.3/+14.7$). The cutoff $|z| > 1.96$ is only descriptive with five estimated references, not a calibrated 5% episode test or an IN/OUT LiRA likelihood ratio.

E Independent Can Validation

The prospective Can pilot uses 180 retained episodes, 60 revoked worse-operator episodes, and 25 non-members. A manifest fixed before new candidate training specifies all roles. The reference panel consists of 13 existing and six new same-recipe retrains. Twenty fresh retain-only models are null candidates; alternatives reintroduce 3, 15, 30, or 60 withdrawn episodes with three training seeds per dose and one nested subset draw. The bank contains 18,291 revoked frames and 7526 non-member frames (25,817 total). Deterministic BC outputs are queried once per frame. Every

alternative is audited on all 60 withdrawn episodes; the dose is training exposure, not a proportion of memorization.

The primary recipe trains for 2000 epochs and selects the most recent best-success checkpoint from 30-rollout evaluations every 100 epochs. The 19 selected references span epochs 100–1500 (only one exceeds 700), while null candidates span 100–1600. Six mild recipe variants alter total epochs or selection frequency; their selected epochs remain within the realized reference range.

Table 13: Prospective Can candidates: marginal and joint rejections. All tests use the frozen 19-reference panel and $\alpha = 0.05$; marginals are unadjusted. Historical edit comparisons used a different request and are excluded (Appendix F).

Candidate population	n	Joint	Action	Loss	Rank
Retain-only, reference recipe	20	2	2	0	0
Retain-only, 1000 epochs	3	0	0	0	0
Retain-only, selection every 50 epochs	3	0	0	0	0
5% reintroduced	3	0	0	0	0
25% reintroduced	3	2	0	0	3
50% reintroduced	3	3	0	3	3
100% reintroduced	3	3	1	3	3

The two null rejections (seeds 102 and 106) have joint nonconformity only 1.012 and 1.061 times the most extreme reference in their respective augmented comparisons. The observed 2/20 rate has a one-sided 95% binomial upper bound of 28.3%, conditional on this panel. The pilot therefore provides imprecise calibration evidence. At 50% reintroduction, evidence detects 3/3 while action detects 0/3. At 25%, rank detects 3/3 but joint detects 2/3; the combined score need not dominate each marginal.

Retrospective checkpoint-selection comparison. All 19 reference, 20 null, and 12 alternative training runs reach epoch 2000. Re-evaluating both reference and candidate populations at that epoch changes selection together while preserving training subsets, bank, episode identities, and score definitions. Configurations match except for identifiers, seeds, and prescribed alternative data. The selected-checkpoint results are reproduced first; Table 14 gives all paired outcomes.

Table 14: Paired selection-rule sensitivity on the same runs. Counts are rejections at $\alpha = 0.05$ with 19 references. Three seeds and one nested draw per alternative dose; marginals are unadjusted. This is retrospective sensitivity, not independent validation.

		Original selected checkpoint				Fixed epoch 2000			
Cohort	n	Joint	Action	Loss	Rank	Joint	Action	Loss	Rank
Retain-only null	20	2	2	0	0	0	1	0	0
5% reintroduced	3	0	0	0	0	2	1	1	0
25% reintroduced	3	2	0	0	3	3	3	3	3
50% reintroduced	3	3	0	3	3	3	3	3	3
100% reintroduced	3	3	1	3	3	3	3	3	3

Reference pairwise action divergence changes from 0.07865 to 0.13196, and mean null-candidate divergence from 0.08609 to 0.13284. Fixed-epoch joint rejection is 0/20 (one-sided 95% upper bound 13.9%); action alone rejects seed 109. All original joint detections persist and low-dose detection increases on these reused alternatives. Because the choice followed inspection of the pilot, it does not establish improved calibration or a superior checkpoint rule.

The 44 new pilot training runs took approximately 18 workstation hours with three concurrent jobs. The audit pass took under a minute using 21 cached and 44 newly evaluated banks, not a cold 65-model evaluation. Reference generation dominates cost. References can be reused for edits of the same retained set and recipe; a new removal request generally requires new references.

Rollout-observation coverage. A later sensitivity study replaces the action bank with 45,492 observations pooled from six separately trained retain-only recipe variants, each collecting 30 paired-

start rollouts. The evidence coordinates remain those from the demonstration bank. The reference action floor changes from 0.07865 to 0.09639. Table 2 restricts interpretation to fresh null and reintroduction candidates, which are distinct from the six generators. The generators' own audited rows use self-generated observations and are descriptive only; they are not included here.

At 25% and 50% reintroduction the rank marginal still rejects 3/3, and at 50% loss also rejects 3/3. Changing the action bank changes joint standardization and reference nonconformity, so joint detection can decrease despite unchanged evidence. The lower null rejection count therefore does not establish a uniformly better audit. This pooled bank does not comprehensively cover each edited candidate's own induced-state distribution.

F Additional Requests and Baseline Comparisons

Harmful and benign PushT requests. The harmful setting adds 80 mirrored episodes to 206 clean episodes. Reflection applies $x \mapsto 512 - x$ to agent, actions, and all T keypoints, with keypoint permutation $[1, 0, 3, 2, 5, 4, 7, 6]$. The 80 audit non-members use source episodes disjoint from withdrawn sources and calibration sources. Three original policies and five clean retrains share normalization statistics; each checkpoint has 250 evaluation rollouts. Mean original and retrain success are $57.9 \pm 5.40\%$ and $72.6 \pm 5.25\%$ across training seeds.

At the frozen stable operating point (learning rate 2.5×10^{-6}, ratio 0.20, 200 updates), ascent increases forget loss by 40–78% but changes AUC by at most 0.003. This refutes the frozen prediction of erased membership evidence. The higher ratio 0.50 is exploratory and guard-stopped at 150–200 updates; its task success improves on all three original seeds, but AUC remains 0.964/0.981/0.985 and retain loss worsens. Frozen and exploratory conditions are not pooled.

Table 15: Harmful PushT replication at the stable learning rate. Success changes use each seed's own original checkpoint and 250 rollouts. The exploratory high dose is guard-limited and not budget matched.

seed	operator	Δsuccess (pts)	retain	forget	steps
s1000	ascent @0.20 (frozen)	-0.4	+6.1%	+43.1%	200/200
s1000	ascent @0.50	+8.0	+42.7%	+180.9%	150/200
s1000	FT200	+1.2	+2.2%	+2.8%	200/200
s2000	ascent @0.20 (frozen)	+3.6	+11.3%	+40.4%	200/200
s2000	ascent @0.50	+11.6	+33.8%	+141.0%	150/200
s2000	FT200	-1.2	+3.6%	+3.7%	200/200
s3000	ascent @0.20 (frozen)	+1.6	+19.1%	+77.5%	200/200
s3000	ascent @0.50	+2.4	+68.7%	+249.6%	150/200
s3000	FT200	-6.4	+5.0%	+1.5%	200/200

The benign request trains the original on 166 clean episodes, withdraws 40, and retains 126; 40 additional clean episodes are excluded from all training and serve as non-members. Original success is 80.8%, versus retrain successes 65.2/57.6/64.8% (mean 62.5%). The original's AUC is 0.966 versus reference mean 0.447. At the same frozen operator setting, ascent has success 80.8% and AUC 0.961; FT has success 79.6% and AUC 0.968. Thus withdrawing useful data changes the direction of the task performance comparison without making task success a removal test. The inherited calibration guard scores mirrored clean states; this vestigial choice did not trip in the benign run.

Localized redirect feasibility and invalidated tuning. The frozen feasibility diagnostic requires action discrepancy and clean-state support. Mirrored PushT editable intersections are 3.65/0.85/0.19% across three detector calibrations, below the prespecified 5% threshold. Localized redirect was therefore not run; no all-frame operator substitutes for this prediction. A prior 10^{-5} learning-rate sweep was invalidated by matched FT instability. That grid is excluded from ascent evidence. The preregistration record preserves the failed predictions and dated applicability amendment.

Corrected Can method comparison. The later clean-request campaign contains 86 edits: 14 method families on four original seeds (56), seven tuned configurations on seeds 2–3 (14), and two methods on eight expansion seeds (16). Each has a 500-rollout evaluation. Every recorded edit uses the same 60-episode `dift_forget`, 180-episode `okay_better_train`, and 15-episode `dift_calib` masks, disjoint from the 25 audit non-members. The audit uses the 19-reference demonstration bank (action floor 0.07865, mean AUC 0.41565), not the earlier 13-reference descriptive convention. All methods use local deterministic-BC adaptations and the recorded budgets.

Table 16: Complete corrected Can comparison. Success is mean±sample SD across training seeds, with 500 rollouts per checkpoint. J/A/L/R count joint, action, loss-level, and rank rejections at unadjusted $\alpha = 0.05$. Cohorts are kept separate; 86 edits are not 86 independent tasks or removal requests.

Method	n	Success (%)	J	A	L	R
original comparison, seeds 1–4						
DIFT-pF (calibrated ascent)	4	35.1±6.3	4	4	4	4
retain FT	4	22.6±4.2	4	0	4	4
gradient ascent	4	0.0±0.0	4	4	4	2
EU-k	4	0.0±0.0	4	4	0	0
Blindspot	4	1.3±0.8	3	0	4	2
RMU	4	0.0±0.0	4	4	4	3
NegGrad+	4	0.0±0.0	4	4	4	4
Fisher	4	0.0±0.0	4	4	4	4
SimNPO	4	1.1±0.6	4	4	0	0
NPO	4	10.2±5.7	4	4	4	4
SSD	4	9.8±12.4	4	4	3	4
task vector	4	6.3±6.9	4	3	4	4
SCRUB	4	1.2±1.2	4	4	4	0
relabel	4	18.2±9.7	2	3	0	2
tuned finalists, seeds 2–3						
NPO + tether 10	2	16.0±9.9	2	2	0	0
NPO + tether 3	2	10.0±10.2	2	2	0	0
SimNPO (tuned)	2	9.5±0.1	2	2	2	2
NegGrad+ + tether 10	2	4.4±5.9	2	2	2	2
SSD (tuned)	2	22.5±15.4	2	1	2	2
GA + tether 10	2	2.1±2.1	2	2	2	1
Blindspot (tuned)	2	4.1±1.6	2	0	2	1
expansion, seeds 6–13						
DIFT-pF (expansion)	8	39.4±9.3	8	7	7	7
retain FT (expansion)	8	25.4±13.3	7	5	8	8

The four joint non-rejections ($p = 0.10$) have rejecting marginals: Blindspot seed 2 (loss), relabel seeds 2 and 3 (rank and action, respectively), and expansion FT seed 9 (loss and rank). These unadjusted marginals do not define an acceptance rule. Expansion DIFT-pF seed 9 achieves 58.4% success with AUC 0.4673; action and rank have $p = 0.10$ and 0.20, while loss and joint have $p = 0.05$.

Historical `worse_train` contains 90 episodes: 57/60 audited withdrawals and 22/25 non-members. Original training excludes non-members; forget-dependent editing exposes them. Retain-only updates need not use the loaded examples. Historical tables and four DIFT-pF pilots are excluded; fresh-null/dose cohorts are unaffected. The rerun cannot isolate request correction because retain-only FT also changed.