RECALL: RECOVERY EXPERIENCE COLLECTION FOR ACTIVE LIFELONG LEARNING IN VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models are typically trained offline on large demonstration datasets and updated post-deployment through additional fine-tuning. A more capable system should instead improve continually, actively identifying states in which it should collect new demonstrations and without forgetting previously learned behaviors. We present RECALL, the first empirical study of active, lifelong learning in VLAs, investigating two coupled questions: where to collect new supervision and how to incorporate it into a pretrained policy. We evaluate a VLA's ability to retain prior behaviors while adapting to new data, comparing start-state demonstrations against policy-induced recovery states, uncertainty-guided versus non-uncertainty-based state selection, sparse versus dense recovery collection, and continual adaptation strategies including reduced learning rates, elastic weight consolidation, and replay. Key findings include: recovery-state supervision is often more useful than repeated start-state demonstrations, though this advantage does not require uncertainty-based selection; more recovery data is not always better; targeted recovery data alone can cause severe forgetting; and preserving prior behavior requires maintaining appropriate behavioral coverage. We further demonstrate uncertainty-triggered recovery collection as a pre-execution intervention on a physical robot. Together, these findings establish an empirical foundation for active lifelong learning in VLAs and offer practical guidance for building robot policies that improve continuously from deployment experience.

1 INTRODUCTION

Recent Vision-Language-Action (VLA) models such as RT-1 (Brohan et al., 2023b), RT-2 (Brohan et al., 2023a), Gemini Robotics (Team et al., 2025), π_0-FAST (Pertsch et al., 2025), and $\pi_{0.5}$ (Intelligence et al., 2025) demonstrate that large-scale robot learning can produce broadly capable manipulation policies. Yet, a pretrained policy may encounter new tasks, embodiments, environments, objects, or failure modes that were not sufficiently represented in its original training distribution. The process of updating a VLA for these variations is typically *passive and episodic*; when a policy performs poorly, an expert collects additional demonstrations (often from nominal task starts) to fine-tune the model, and then repeats this data collection and fine-tuning process if performance remains insufficient. To enable long-term deployment of generalist robots, a VLA should instead be both *active* and *lifelong* learners: active in choosing what data to collect, and lifelong in incorporating that experience over time without destroying capabilities it has already learned.

Active data selection and continual adaptation have been studied broadly in machine learning (Hoque et al., 2022; Parisi et al., 2019) and increasingly in large language (Kung et al., 2023; Shi et al., 2025) and multimodal (Gupte et al., 2024; Yu et al., 2024) models, but have only recently begun to be explored in VLAs (Liu et al., 2026; Römer et al., 2026b;a). Furthermore, translating these ideas to VLAs introduces additional challenges. Expert supervision is expensive because it requires physical or simulated robot interaction, and errors can place the system in off-nominal states that differ substantially from the original demonstration distribution. At the same time, fine-tuning on a small, targeted set of new behaviors can alter a large pretrained policy and degrade skills that are absent from the adaptation data. Active lifelong learning for VLAs therefore couples two problems

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.

that are often studied separately: *where* should new supervision be collected, and *how* should that supervision be incorporated without forgetting prior behavior?

In this paper, we present a study of Recovery Experience Collection for Active Lifelong Learning (RECALL): **the first empirical study of active, lifelong learning in VLAs**. We conduct six experiments to evaluate the capabilities and failure modes of a pretrained π_0-FAST policy when combined with a range of strategies for active and continual learning. Our contributions are as follows:

1. **A problem formulation for active lifelong VLA adaptation.** We frame post-deployment VLA learning as two coupled design problems: *where* new supervision should be collected and *how* that experience should be incorporated while retaining prior capabilities.
2. **A controlled study of active recovery-data collection.** We compare standard-start demonstrations, uncertainty-triggered recovery demonstrations, non-uncertainty intermediate-state controls, and different data collection densities, separating the broader value of policy-induced recovery supervision from that of uncertainty-guided state selection.
3. **Practical guidance for continual VLA adaptation.** Through recovery-only fine-tuning, reduced learning rates, EWC, replay with different coverage, and physical-robot deployment, we identify recurring stability–plasticity tradeoffs and derive practical recommendations for designing VLAs that can actively improve from deployment experience over time.

2 Related Works

Active and Interactive Data Collection for Robot Policies. Active learning improves data efficiency by identifying where additional supervision is most useful. In imitation learning, DAgger (Ross & Bagnell, 2010) addresses covariate shift by collecting expert labels on learner-induced states, while HG-DAgger (Kelly et al., 2019), SafeDAgger (Zhang & Cho, 2017), and ThriftyDAgger (Hoque et al., 2022) selectively request intervention or supervision to reduce labeling burden. Related approaches similarly use human corrections or takeovers as informative training data (Spencer et al., 2020; Wang et al., 2026). Uncertainty provides one such selection signal: prior work uses Monte Carlo dropout, ensembles, conformal methods, and related uncertainty estimates to query experts selectively (Cui et al., 2019; Menda et al., 2019; Zhao et al., 2025; Wang & Wang, 2024; Lee et al., 2025). These methods have primarily been studied with smaller task-specific policies in autonomous-driving or classical imitation-learning settings (Kelly et al., 2019; Wang & Wang, 2024; Ross & Bagnell, 2010). We instead study a pretrained VLA and separate the value of policy-induced recovery states from any advantage specific to uncertainty-based selection.

Uncertainty Quantification and Help Prediction in VLAs. Uncertainty estimation has been studied through ensembles (Lakshminarayanan et al., 2017), Bayesian approximations (Gal & Ghahramani, 2016), conformal prediction (Angelopoulos & Bates, 2022), and learned failure or intervention predictors. KnowNo (Ren et al., 2023), for example, uses conformal prediction to determine when a vision-language planner should ask for help, but operates over high-level plan selection rather than low-level VLA actions. INSIGHT (Karli et al., 2025) instead performs inference-time introspection for autoregressive VLAs using statistics derived from action-token decoding, including entropy, log-probability, and aleatoric and epistemic uncertainty.

Continual Adaptation and Catastrophic Forgetting. Continual learning studies how models acquire new knowledge without *catastrophic forgetting* (McCloskey & Cohen, 1989), a failure also observed during continual fine-tuning of large language models (Luo et al., 2025). Common approaches include replay, which retains prior examples during optimization (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017), and parameter regularization such as EWC (Kirkpatrick et al., 2017); related alternatives include distillation-based replay such as DER (Buzzega et al., 2020) and parameter-efficient adaptation such as LoRA (Hu et al., 2021). Our study focuses on the stability–plasticity problem created by highly targeted recovery data, studying recovery-only fine-tuning, reduced learning rates, EWC, and replay with different coverage.

3 Study Overview and Research Questions

We study active lifelong learning in VLAs through one controlled collection-and-adaptation cycle, shown in Figure 1: a deployed policy encounters states during execution, selected states are used to

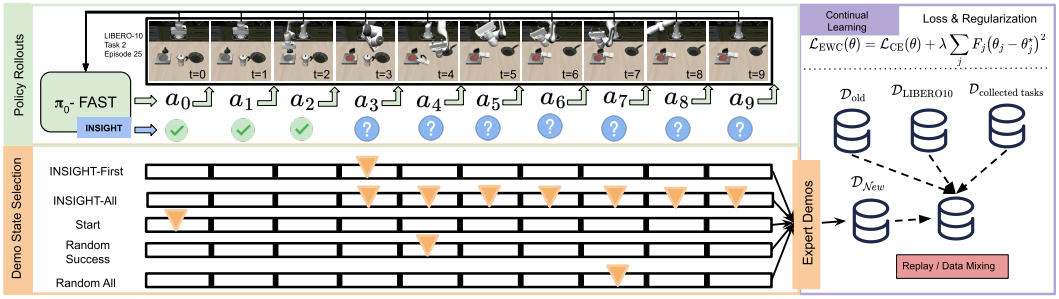


Figure 1: **Overview of the RECALL study.** The green region shows policy rollouts and uncertainty monitoring by INSIGHT. The orange region shows how demonstrations can be prompted at different timesteps (explored in **RQ1**) and densities (explored in **RQ2**) within the set of states visited by the rollout, forming the dataset $\mathcal{D}_{\text{new}}$. Corresponding to the purple region, **RQ3** assesses fine-tuning over $\mathcal{D}_{\text{new}}$ with or without a replay buffer, **RQ4** assesses update constraints, and **RQ5** varies prior-data replay coverage.

obtain additional expert supervision, and the resulting experience is added to a data buffer for future training. After repeating this process for several rollouts, the policy is again fine-tuned over the new data batch. Rather than proposing a single state-selection rule or continual-learning algorithm, we conduct a sequence of experiments to characterize the design choices and research questions that arise across this pipeline:

RQ1: Which demonstration starting states are useful? Do recovery demonstrations from policy-induced intermediate states provide more useful supervision than matched demonstrations from standard task starts, and does uncertainty-guided state selection result in better recovery demonstrations?

RQ2: How densely should recovery demonstrations be collected? Is one complete recovery from the first uncertain state sufficient, or does collecting recoveries from every uncertain state provide additional benefit?

RQ3: Are recovery demonstrations sufficient as a standalone fine-tuning dataset? Can a VLA adapt using only newly collected recovery data without replaying prior experience?

RQ4: Can conservative updates prevent forgetting while preserving adaptation? We test whether reducing update magnitude or constraining parameter changes can preserve prior behavior without suppressing useful learning.

RQ5: How much prior experience must replay cover? We examine whether retaining progressively narrower subsets of prior experience is sufficient to preserve previously learned behaviors.

RQ6: Can recovery-state collection be deployed on a physical robot? We test whether pre-execution help prediction can trigger human takeover and produce recovery data that improves a subsequently deployed physical policy.

RQ1–RQ2 study the *collection* side of active lifelong learning: where supervision should begin and how much should be collected. RQ3–RQ5 study the *adaptation* side: whether targeted data can be used alone and how stability and plasticity depend on the update rule and prior-data coverage. RQ6 evaluates whether the recovery-collection mechanism can be executed in a physical deployment.

4 Shared Experimental Setup and Evaluation Protocol

Unless stated otherwise within an individual research question, the simulation experiments share the following model, training, and evaluation protocol. Experiment-specific collection rules, datasets, regularizers, and replay configurations are introduced in the corresponding RQ sections.

Base VLA and Adaptation. We use the π_0-FAST policy (Pertsch et al., 2025), trained on LIBERO (Liu et al., 2023) for 30k steps using the authors' training recipe. Its final checkpoint serves as the common initial policy π_{θ_0} for all simulation experiments. Unless stated otherwise, all conditions are initialized from π_{θ_0} and fine-tuned with the standard π_0-FAST autoregressive action-token

objective and training schedule. Replay mixtures are sampled without explicit source reweighting or enforced old/new batch balancing. Full optimization details are provided in Appendix B.

Simulation Tasks and Evaluation Split. Recovery data are collected for LIBERO-10 tasks numbered $\mathcal{T}_{\text{collected}} = \{2, 3, 5, 8, 9\}$, while tasks numbered $\mathcal{T}_{\text{retained}} = \{0, 1, 4, 6, 7\}$ receive no new recovery demonstrations. Each checkpoint is evaluated with 50 autonomous rollouts per task (500 total). We report overall success, collected-task success, and retained-task success to distinguish adaptation from preservation of prior behavior. The initial policy achieves 59.8% overall success, with 32.0% on collected tasks and 87.6% on retained tasks.

Checkpoint and Statistical Protocol. To avoid favoring conditions through best-checkpoint selection, our primary simulation summary is the mean over the common checkpoint grid $\mathcal{G} = \{1000, 2000, \ldots, 9000, 9999\}$. Direct grid-mean comparisons include only conditions evaluated at every checkpoint in $\mathcal{G}$; missing evaluations are not imputed, and individual maxima are reported only descriptively. Because checkpoints from one fine-tuning run are correlated and do not represent independent training seeds, RQ1 and RQ2 inference operates on task-level grid means using exact two-sided sign-flip tests, with multiplicity correction defined within each RQ. Full testing details, assumptions, Holm families, and checkpoint-sensitivity analyses are provided in Appendix D.

Normalization and Experimental Scope. Primary replay experiments recompute normalization statistics on each training mixture, whereas several new-only, low-learning-rate, and EWC configurations reuse the original statistics. We therefore compare these as complete adaptation configurations rather than attributing differences solely to the nominal training method; normalization ablations are reported in Appendix G. Each simulation condition is one fine-tuning run of one autoregressive VLA over a single collection-and-adaptation cycle, so checkpoint variation does not measure training-seed variability. RQ6 uses a separate physical-robot protocol (see Section 10).

5 RQ1: Which Demonstration Starting States Are Useful?

We address this RQ via two sub-questions. **RQ1a)** Are demonstrations beginning from policy-induced intermediate states more useful than demonstrations from task initial states? **RQ1b)** Does uncertainty identify more useful intermediate states than non-uncertainty-based sampling?

Method. To construct policy-induced states, the common base policy π_{θ_0} generates 50 autonomous rollouts for each collected LIBERO-10 task, with simulator states saved approximately every five environment steps. We then compare four collection strategies. **Start** collects complete demonstrations from the nominal task-start distribution. **INSIGHT** uses the strongly-supervised variation of INSIGHT (Karli et al., 2025), a previously trained pre-execution help predictor, and begins a recovery at the first state that it flags as *uncertain* in a rollout. The predictor and its decision threshold are fixed before the study collection. **Random-All** samples intermediate states using task-specific Gaussian timing priors centered at the mean rollout duration for each task, while **Random-Success** uses a global timing prior fitted to successful-rollout durations; episodes themselves are sampled from all eligible rollouts. Full sampling, replacement, and acquisition details are provided in Appendix C. Each selected simulator state is restored and the same operator provides a complete SpaceMouse recovery through task completion; unsuccessful or invalid collections are replaced as detailed in Appendix C.

We use each collection strategy to obtain 149 demonstrations with identical task quotas $(36, 32, 13, 31, 37)$ across tasks $(2, 3, 5, 8, 9)$. These are the five tasks on which the initial policy performs worst, and the per-task quotas match the distribution of INSIGHT recovery demonstrations available from those tasks. Each dataset is combined with the same full prior replay dataset and trained using the shared CE configuration from Section 4.

Results. Across the common checkpoint grid, INSIGHT reaches 65.98% overall success, Random-All 67.64%, and Random-Success 67.70%, compared with 57.06% for Start. The corresponding collected-task means are 43.48%, 47.48%, 47.16%, and 27.28%, while retained-task means remain similar, ranging from 86.84% to 88.48%. All three intermediate-state-versus-Start overall contrasts have raw $p < 0.05$, but none remains significant after Holm correction across the retrospectively defined RQ1 family. Balanced-grid descriptive peaks are 72.4% for INSIGHT, 74.0% for Random-All, 73.8% for Random-Success, and 60.2% for Start; these maxima are descriptive because checkpoint selection uses evaluation outcomes.

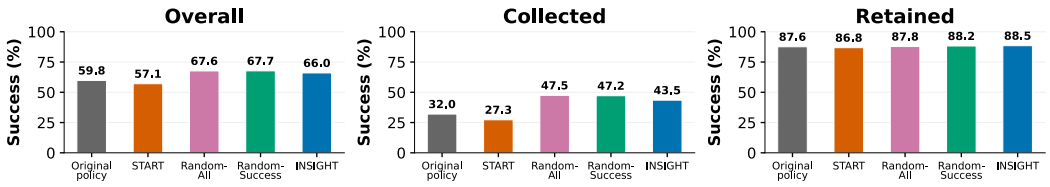


Figure 2: **Recovery-state collection achieves higher observed performance than standard-start recollection.** Bars show performance averaged over the common checkpoint grid for various data collection strategies. Random intermediate-state controls perform similarly to INSIGHT, so the simulation benefit cannot be attributed uniquely to uncertainty-based state selection.

Table 1: **Task-wise acquisition yield:** the % of proposed states that enable new demonstrations.

	T2	T3	T5	T8	T9	Overall
Random-All	87.8	74.4	76.5	46.3	97.4	72.3
Random-Success	100.0	84.2	65.0	96.9	97.4	90.9

The samplers also differ in acquisition burden (Table 1): Random-All requires 206 proposed states to obtain 149 retained demonstrations, compared with 164 for Random-Success, with the largest yield difference on task 8. INSIGHT retains 149 demonstrations from 179 candidates (83.2%), although this fraction is not directly comparable because excluded INSIGHT candidates were not audited using the same invalid-state categories.

Discussion. The observed pattern suggests that the usefulness of the data is not unique to uncertainty-based state selection. All three policy-induced intermediate-state strategies have higher observed means than Start, while Random-All and Random-Success perform similarly to INSIGHT. This experiment therefore does not establish a unique learning advantage for uncertainty-based state selection. The acquisition audit also shows that downstream policy performance alone does not characterize a collection strategy. Random-Success obtains the same retained quota with 20.4% fewer candidate-state proposals than Random-All. Timing analysis in Appendix C shows that the samplers rarely select the same exact state, although they can expose the expert to overlapping remaining portions of a task. This may help explain why distinct starting-state distributions can nevertheless produce similar downstream performance.

Takeaway: *Policy-induced intermediate-state collection yields higher post-tuning performance than standard-start recollection in our experiments, but uncertainty should not be credited for that difference without appropriate intermediate-state controls; acquisition burden should also be measured separately from final dataset size.*

6 RQ2: Is More Recovery Data Always Better?

Method. We now evaluate whether there is a benefit to *densely* collecting additional recovery demonstrations within the same set of task configurations. We compare two versions of INSIGHT: **INSIGHT-First**, which records one complete recovery beginning at the first state flagged as *uncertain* in each rollout, and **INSIGHT-All**, which restores and collects a recovery from *every* uncertain state throughout each rollout. From RQ2 onward, we fix INSIGHT as the recovery-state selector so that subsequent experiments isolate collection density and adaptation strategy rather than state selection. We use INSIGHT because it is the operational mechanism of interest: unlike the random controls in RQ1, it provides a deployable pre-execution signal intended to request help before a likely failure occurs. RQ1 therefore treats the random samplers as controls for the value of intermediate-state supervision, while RQ2–RQ5 study how INSIGHT-triggered recovery data should be collected and incorporated. For both INSIGHT-First and INSIGHT-All, each selected state is restored and the same operator provides a complete recovery demonstration through task completion. During fine-tuning, the resulting recovery demonstrations are mixed with the complete original training dataset used to obtain the base policy (*full prior replay*), trained with the standard CE procedure, and use normalization statistics recomputed on the resulting training mixture.

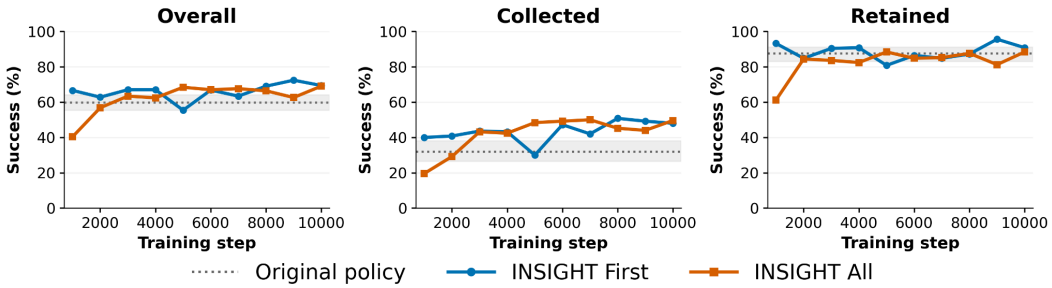


Figure 3: **Denser recovery collection does not improve the observed training-trajectory mean.** INSIGHT-First uses one complete recovery per uncertainty-flagged rollout, while INSIGHT-All collects from every uncertain state. Both use the same replay and training configuration.

Results. INSIGHT-First invokes 149 recovery demonstrations and ∼115k logged collection timesteps. INSIGHT-All invokes 576 demonstrations and ∼889k logged timesteps: $3.87\times$ as many demonstrations and $7.72\times$ the logged timesteps.

Across the common checkpoint grid, INSIGHT-First reaches 65.98% overall mean success compared with 62.40% for INSIGHT-All. Collected-task means are 43.48% and 42.08%, respectively, while retained-task means are 88.48% and 82.72%. The overall difference of $+3.58$ percentage points has raw $p = 0.3535$ and Holm-adjusted $p = 0.7070$. None of the overall, collected, or retained First-versus-All comparisons is statistically significant after correction. The result therefore does not establish superiority of INSIGHT-First, but the substantially larger INSIGHT-All dataset does not produce a higher observed grid mean.

Discussion. A complete recovery beginning at the first intervention point already supervises the remainder of the task. Later uncertain states from the same rollout can therefore invoke demonstrations containing many of the same actions, objects, and remaining subskills, making demonstration count a poor proxy for supervision diversity. This comparison is not a pure data-volume ablation: INSIGHT-All also changes the starting-state distribution, demonstration lengths, and relative exposure to new versus replayed samples. We therefore interpret the result as local diminishing returns under this collection protocol, not as a universal upper bound on useful recovery data.

Takeaway: *More recovery demonstrations are not automatically more useful; one complete recovery per intervention episode provides performance comparable with denser data collection.*

7 RQ3: Can Recovery Data Alone Sustain Adaptation?

Recovery collection deliberately concentrates supervision around difficult policy-induced states. We therefore ask whether the resulting recovery demonstrations can serve as the entire fine-tuning dataset, or whether prior training data must also remain represented during adaptation.

Method. The **new-only** configurations fine-tune only on the recovery demonstrations (prompted via INSIGHT-First or -All) and reuse the original normalization statistics. The **replay-based** configurations instead mix those recovery demonstrations with the complete original training dataset used to train the base policy (*full prior replay*) and recompute normalization statistics on this combined fine-tuning dataset. This comparison therefore tests complete adaptation configurations rather than isolating replay as a single causal factor; Appendix G examines normalization sensitivity separately.

Results. New-only fine-tuning can degrade performance substantially. For example, an INSIGHT-First new-only checkpoint using the original normalization statistics reaches only 28.4% overall success, compared with 59.8% for the original policy. Its collected-task success is 0.4%, while retained-task success falls from 87.6% to 56.4%. Performance deteriorates further with continued new-only training, whereas the corresponding INSIGHT-First configuration with full prior replay maintains substantially higher collected- and retained-task performance.

Discussion. Recovery demonstrations can be informative precisely because they concentrate on states where the original policy is weak, but this also makes their distribution narrow. They omit

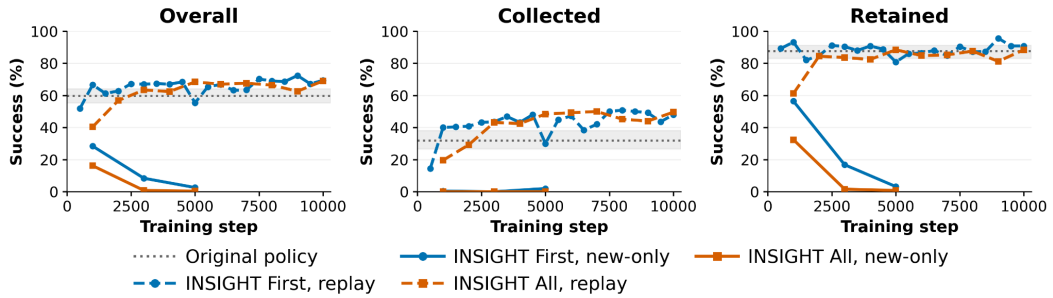


Figure 4: **Recovery-only fine-tuning loses previously learned competence and fails to sustain adaptation.** Solid curves use only newly collected recovery demonstrations; replay configurations retain prior data during fine-tuning. Normalization differs between these configurations as described in Section 4.

many nominal transitions, task prefixes, and behaviors needed for complete task execution. Training exclusively on this distribution can therefore damage competence even on the tasks from which the recoveries were collected. The failure is consequently stronger than the usual picture of "learning the new task while forgetting old ones": the model can lose broad competence before successfully incorporating the targeted recovery behavior. Active state selection and continual adaptation cannot therefore be treated as independent problems.

Takeaway: *Targeted recovery demonstrations cover only a limited subset of difficult policy-induced states, making them insufficient as the sole fine-tuning dataset for preserving broader prior behavior.*

8 RQ4: Can Conservative Updates Balance Adaptation and Retention?

Method. We ask whether catastrophic degradation can be mitigated without broad replay by constraining the update itself. Using the INSIGHT-First recovery dataset, we compare standard new-only CE training with a constant low learning rate $\alpha = 2.5 \times 10^{-8}$, two to three orders of magnitude below the standard fine-tuning schedule, and diagonal elastic weight consolidation (EWC) (Kirkpatrick et al., 2017). With diagonal Fisher information F computed at the original parameters θ_0, EWC optimizes: $\mathcal{L}_{\text{EWC}}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda \sum_j F_j(\theta_j - \theta_{0,j})^2$.

We evaluate multiple values of λ (which controls the strength of the constraint on parameter updates) and Fisher reference datasets (which indicate the prior behaviors used to identify parameters that should be protected). We therefore compare Fisher estimates computed from broader prior data with estimates that exclude the collected tasks. These configurations reuse the original normalization statistics, while replay recomputes them; comparisons therefore reflect complete adaptation configurations, not regularization alone.

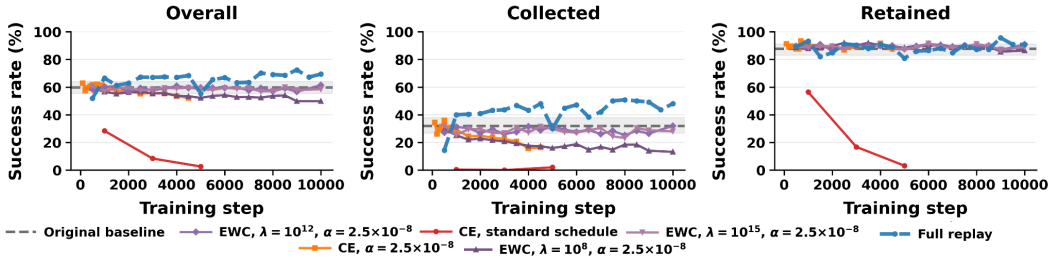


Figure 5: **Conservative updates improve stability but can suppress useful adaptation.** Low learning rates and EWC better preserve retained-task behavior than standard recovery-only fine-tuning, but collected-task improvement remains limited. Gray bands denote the original policy's Wilson interval, not variability across training runs.

Results. Both low-rate training and EWC preserve retained-task performance substantially better than standard new-only fine-tuning, but collected-task adaptation remains limited. For example, EWC with $\lambda = 10^{12}$ and the low learning rate achieves grid means of 58.70% overall, 28.44% collected, and 88.96% retained success, leaving collected-task performance below the original policy's 32.0%. The low-learning-rate CE run reaches a descriptive maximum of 62.8% at step 100 before deteriorating; because it does not cover the complete common grid, this peak is not directly comparable to the primary grid means. Additional EWC and filtered-Fisher sweeps are in Appendix H.

Discussion. These experiments expose the stability–plasticity tradeoff: strong constraints can preserve prior behavior while suppressing useful adaptation, whereas weaker constraints reintroduce forgetting. Retention alone is therefore not evidence of successful continual adaptation. We do not interpret EWC as universally ineffective, nor the replay comparison as an algorithmic ranking, particularly because the configurations also differ in normalization.

Takeaway: *Successful continual adaptation requires both stability and plasticity; parameter constraints that preserve prior behavior can still fail if they prevent useful learning.*

9 RQ5: What Must the Replay Buffer Cover?

Method. RQ4 shows that restricting parameter movement does not by itself guarantee useful adaptation. We therefore return to replay and ask which prior behaviors must remain represented during fine-tuning. Using the same INSIGHT-First recovery dataset, we compare three replay scopes: **Full replay** mixes the recovery demonstrations with the complete prior training dataset; **LIBERO-10 replay** retains only prior demonstrations from the benchmark suite; and **Collected-task replay** retains prior demonstrations only from the five tasks receiving new recovery data. We additionally evaluate EWC and filtered-EWC variants on top of collected-task replay to test whether parameter regularization can compensate for reduced behavioral coverage.

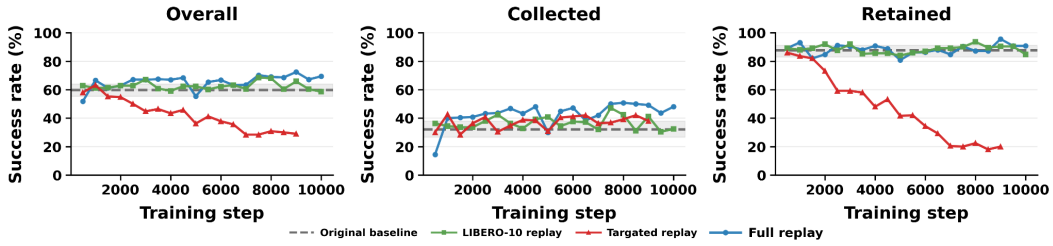


Figure 6: **Replay coverage controls which behaviors remain stable during adaptation.** Full and LIBERO-10 replay retain broad in-suite performance, whereas restricting replay to collected tasks allows performance on unrepresented tasks to deteriorate.

Results. Full replay achieves grid means of 65.98% overall, 43.48% collected, and 88.48% retained success. LIBERO-10 replay reaches 62.80%, 36.96%, and 88.64%, respectively. Thus, replay restricted to the benchmark suite preserves similar mean performance on the five retained LIBERO-10 tasks at this operating point. LIBERO-10 replay reaches a descriptive peak of 68.6% but declines to 58.6% at the final checkpoint, again illustrating why a selected maximum can obscure the training trajectory. Collected-task replay initially improves targeted tasks but subsequently loses performance on non-collected tasks. EWC and filtered-EWC variants do not consistently prevent this deterioration (Appendix I, Fig. 13).

Discussion. The comparison between full and LIBERO-10 replay shows that reproducing the complete original training distribution is not necessary to preserve the particular in-suite behaviors evaluated here. However, it does not establish retention of broader capabilities that are not measured. Collected-task replay reveals the complementary failure mode: once the five non-collected tasks disappear from the adaptation distribution, those behaviors become vulnerable despite strong prior performance. In these experiments, parameter penalties do not reliably substitute for replay.

Takeaway: *Replay should represent the behaviors a policy is expected to preserve; a smaller buffer can be sufficient only relative to the capabilities actually covered and evaluated.*

10 RQ6: Can Recovery Collection Work on a Physical Robot?

Method. We finally test whether INSIGHT-triggered recovery collection can operate in a physical deployment using a UFactory xArm-7. Before an action chunk is executed, INSIGHT evaluates the proposed chunk. If help is requested, the chunk is withheld, the robot holds its current pose, and a human operator uses a GELLO leader arm (Wu et al., 2024) to demonstrate the remainder of the task. The resulting trajectory is recorded as one recovery demonstration; policy parameters are updated only afterward through batch fine-tuning. The physical collection interface is shown in Appendix Fig. 14, with additional evaluation details in Appendix J.

We collect 30 recoveries, ten each for *lift the corn*, *put the corn in the pot*, and *put the pot in the sink*. The resulting fine-tuned **INSIGHT-30** policy is compared with **START-30**, which is fine-tuned on 30 task- and environment-matched demonstrations from standard task starts. Both policies originate from the same base checkpoint and use the same replay data, batch size, number of updates, and normalization statistics. The final evaluation uses fixed policy checkpoints and 60 matched scene pairs, balanced with 20 pairs per task. Because the policies are paired by scene, aggregate success is compared using a two-sided exact McNemar test; task-specific rates are descriptive.

Results. Across the 60 scene pairs, START-30 succeeds in 32/60 trials (53.3%) and INSIGHT-30 in 43/60 (71.7%), an absolute difference of 18.3 percentage points. The paired outcomes contain 26 scenes where both policies succeed, 11 where both fail, 17 where only INSIGHT-30 succeeds, and 6 where only START-30 succeeds. The two-sided exact McNemar test gives $p = 0.0347$.

The aggregate difference varies by task. INSIGHT-30 improves *lift-corn* success from 65% to 90% and *corn-in-pot* success from 35% to 75%, while START-30 performs better on *pot-in-sink* (60% vs. 50%). These task-level comparisons are descriptive because each contains only 20 matched pairs.

During recovery collection, mean robot-side π_0-FAST request latency is 1.126 s, while the local INSIGHT decision adds 0.039 s, approximately 3.5% additional inference overhead.

Discussion. The hardware study demonstrates that recovery-state collection can be implemented as a practical pre-execution intervention loop and that the resulting recovery-trained policy can improve over an equal-count standard-start control under a fresh matched evaluation. Uncertainty also has an operational role on hardware that is distinct from the unresolved selector-specific learning question in simulation. Simulator states can be restored retrospectively, whereas arbitrary past physical states cannot be reconstructed after a rollout. INSIGHT instead makes the intervention decision before the proposed action is executed, providing a deployable mechanism for real-time takeover even though the simulation study does not establish that INSIGHT selects uniquely better training states.

Takeaway: *Uncertainty-guided data collection is practical and effective for physical robots, but target-task adaptation, retention, safety, and operator burden remain distinct evaluation objectives.*

11 Practical Lessons and Conclusion

We study active lifelong VLA learning as a coupled collection and adaptation problem. Across our experiments, policy-induced intermediate-state demonstrations achieve higher observed performance than matched standard-start recollection, although random intermediate-state controls perform similarly to INSIGHT. Collection strategy also affects acquisition burden, and collecting more recoveries from related states does not necessarily improve downstream performance.

Targeted recovery data alone can cause severe forgetting, while conservative updates can preserve prior behavior at the cost of useful adaptation. Replay provides the strongest observed operating point among the complete configurations tested here, but its effectiveness depends on maintaining coverage of the behaviors that should remain stable. On hardware, pre-execution uncertainty provides a practical mechanism for requesting recovery supervision, with INSIGHT-30 improving over an equal-count START-30 control in our matched evaluation, without establishing that uncertainty is the uniquely best state-selection rule.

Together, these findings provide an empirical first step toward VLAs that actively identify useful experience and incorporate it over time. Active lifelong learning cannot be treated as a state-selection problem alone: state selection, acquisition burden, supervision diversity, plasticity, and retention must be addressed jointly.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation and execution, including refining experimental designs, discussing analysis choices, supporting implementation and data-processing code, and helping interpret experimental results. We also used generative AI tools to draft and revise portions of the manuscript, improve grammar and readability, assist with scientific figure and image creation, and suggest edits to software code.

We did not use generative AI to generate synthetic experimental datasets or to formulate or prove mathematical claims. Other required-disclosure categories, including translation and qualitative or thematic data analysis, were not applicable to this work.

All AI-assisted research decisions, code, analyses, figures, and manuscript text were reviewed by the authors. AI-generated or AI-modified code was tested on the relevant experimental pipelines, numerical results were checked against the underlying experiment logs and data, and scientific claims were verified by the authors before inclusion in the manuscript. Generative AI tools did not independently determine experimental conclusions or reported results. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023b. URL https://arxiv.org/abs/2212.06817.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Yuchen Cui, David Isele, Scott Niekum, and Kikuo Fujimura. Uncertainty-aware data aggregation for deep imitation learning, 2019. URL https://arxiv.org/abs/1905.02780.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Sanket Rajan Gupte, Josiah Aklilu, Jeffrey J. Nirschl, and Serena Yeung-Levy. Revisiting active learning in the era of vision foundation models, 2024. URL https://arxiv.org/abs/2401.14555.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 598–608. PMLR, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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, 2025. URL https://arxiv.org/abs/2504.16054.

Ulas Berk Karli, Ziyao Shangguan, and Tesca FItzgerald. Insight: Inference-time sequence introspection for generating help triggers in vision-language-action models, 2025. URL https://arxiv.org/abs/2510.01389.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8077–8083. IEEE Press, 2019. doi: 10.1109/ICRA.2019.8793698. URL https://doi.org/10.1109/ICRA.2019.8793698.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1813–1829, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.112. URL https://aclanthology.org/2023.emnlp-main.112/.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Sung-Wook Lee, Xuhui Kang, and Yen-Ling Kuo. Diff-dagger: Uncertainty estimation with diffusion policy for robotic manipulation, 2025. URL https://arxiv.org/abs/2410.14868.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

Huihan Liu, Changyeon Kim, Bo Liu, Minghuan Liu, and Yuke Zhu. Pretrained vision-language-action models are surprisingly resistant to forgetting in continual learning. *arXiv preprint*, 2026.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,

and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL https://arxiv.org/abs/2308.08747.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning, 2019. URL https://arxiv.org/abs/1807.08364.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.09747.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017. URL https://arxiv.org/abs/1611.07725.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

Ralf Römer, Maximilian Seeliger, Saida Liu, Ben Sturgis, Marco Bagatella, Daniel Marta, Andreas Krause, and Angela P. Schoellig. Uncertainty quantification for flow-based vision-language-action models. *arXiv preprint arXiv:2606.18043*, 2026a.

Ralf Römer, Yi Zhang, Yuming Li, and Angela P. Schoellig. Clare: Continual learning for vision-language-action models via autonomous adapter routing and expansion. *arXiv preprint arXiv:2601.09512*, 2026b.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 661–668, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/ross10a.html.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5), November 2025. ISSN 0360-0300. doi: 10.1145/3735633. URL https://doi.org/10.1145/3735633.

David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Is your imitation learning policy better than mine? policy comparison with near-optimal stopping. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.077.

Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. 07 2020. doi: 10.15607/RSS.2020.XVI.055.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

Anjiabei Wang, Shuangge Wang, and Tesca Fitzgerald. Enhancing goal inference via correction timing. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 2744–2753, 2026.

Changquan Wang and Yun Wang. Uncertainty-driven data aggregation for imitation learning in autonomous vehicles. *Information*, 15(6), 2024. ISSN 2078-2489. doi: 10.3390/info15060336. URL https://www.mdpi.com/2078-2489/15/6/336.

Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2024. URL https://arxiv.org/abs/2309.13037.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230, June 2024.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 2891–2897. AAAI Press, 2017.

Michelle Zhao, Reid Simmons, Henny Admoni, Aaditya Ramdas, and Andrea Bajcsy. Conformalized interactive imitation learning: Handling expert shift and intermittent feedback, 2025. URL https://arxiv.org/abs/2410.08852.

APPENDIX

A LIMITATIONS

Our study has several limitations. First, each simulation condition uses a single fine-tuning run. Multiple checkpoints characterize one optimization trajectory but do not measure variability across training seeds. Moreover, the common-grid summaries, task-level sign-flip tests, and RQ1/RQ2 multiplicity families were defined retrospectively. We therefore treat the statistical comparisons as exploratory evidence rather than definitive estimates of general performance differences.

Second, the state-selection experiments do not isolate every explanation for the value of recovery supervision. Random-All and Random-Success use timing priors rather than exact semantic-progress matching, and the acquisition audit describes proposed states rather than the final retained distributions after replacement. Random-All additionally required a post-hoc replacement fallback on task 8. Similar downstream performance across INSIGHT and random controls therefore does not imply equivalent demonstrations or that uncertainty provides no useful signal. Stronger progress-, failure-, and representation-matched controls remain important future work.

Third, several adaptation comparisons differ in more than the nominal continual-learning method. Primary replay conditions recompute normalization statistics on their training mixtures, whereas several new-only, low-learning-rate, and EWC conditions reuse the original statistics. These experiments therefore compare complete adaptation configurations rather than cleanly isolating replay, regularization, learning rate, or normalization, and our tested methods are not an exhaustive benchmark of continual-learning approaches. More broadly, we study one autoregressive VLA, one benchmark, five collected and five retained tasks, and one collection-and-adaptation cycle. Retention is therefore established only for the evaluated LIBERO-10 tasks, not for the model's broader capabilities, and repeated lifelong adaptation remains unexplored.

Finally, the physical study is limited to one xArm-7, one operator, one training seed, and three target tasks. It evaluates target-task adaptation but not retention of unrelated physical skills, transfer across robots or operators, safety, failure severity, or operator workload, and does not compare random intermediate-state collection on hardware. The results therefore establish the feasibility of first-trigger recovery collection, not that INSIGHT is the uniquely best intervention rule. This study also depends on the quality and operating point of the help predictor; we keep its threshold fixed and do not systematically study calibration or the false-positive/false-negative tradeoff.

Overall, RECALL should be interpreted as an empirical study of design choices for one cycle of active lifelong VLA adaptation rather than a complete solution to lifelong robot learning. Broader conclusions require additional training seeds, architectures, benchmarks, embodiments, stronger state-selection controls, and repeated real-world adaptation.

B TRAINING DETAILS

All simulation policies are initialized from the final 30k-step π_0-FAST checkpoint and optimized with the standard autoregressive action-token prediction objective. Unless otherwise stated, fine-tuning uses the original π_0-FAST learning-rate schedule, with linear warmup to $\alpha = 2.5 \times 10^{-5}$ followed by cosine decay to $\alpha = 2.5 \times 10^{-6}$. When prior and newly collected datasets are mixed, samples are drawn without explicit source reweighting or enforced old/new batch balancing.

C RANDOM INTERMEDIATE-STATE CONTROLS AND ACQUISITION AUDIT

The random intermediate-state controls in RQ1 are designed to separate the general value of collecting demonstrations from policy-induced intermediate states from the narrower question of whether uncertainty provides a uniquely useful state-selection signal. This appendix describes their sampling distributions, human recovery procedure, acquisition yield, and relationship to INSIGHT first-trigger states.

C.1 Policy-Induced State Bank and Initial Sampling

Random-All and Random-Success use the same policy-induced simulator state bank used for INSIGHT collection: 50 autonomous rollouts for each collection task, with simulator states saved approximately every five environment steps. The random selectors operate on this fixed state bank and do not rerun the policy.

For the initial draw, rollout episodes are sampled uniformly without replacement within each task using seed 42, and exactly one state is selected from each sampled episode. Sampling episodes before timesteps avoids implicitly weighting longer rollouts more heavily because they contain more saved states.

The final target quotas are

$$(36, 32, 13, 31, 37) \tag{1}$$

for LIBERO-10 tasks $(2, 3, 5, 8, 9)$, respectively, yielding 149 retained demonstrations per RQ1 condition.

Eligible states begin at environment step 10 and exclude the final five environment steps of each rollout. After selecting an episode, a target intervention time is sampled from the condition-specific timing prior, truncated to that episode's eligible interval, and snapped to the nearest stored simulator state. Neither Random-All nor Random-Success uses INSIGHT predictions during state selection. Random-Success also does not restrict episode selection itself to successful rollouts.

C.2 Timing Priors

Random-All. Random-All fits a Gaussian independently for each task using the endpoint lengths of all autonomous rollouts, including both successful and failed episodes. Failed episodes reach the 525-step horizon, whereas successful episodes can terminate earlier.

Table 2: **Autonomous rollout endpoint statistics used by Random-All.**

Task	Rollouts	Endpoint mean	Endpoint SD	Range
2	50	436.6	129.7	185–525
3	50	462.8	118.3	205–525
5	50	373.5	171.5	140–525
8	50	525.0	0.0	525–525
9	50	485.9	90.9	240–525

Task 8 is degenerate because none of its autonomous rollouts terminates before the horizon. Its zero-variance endpoint fit, together with the selector's minimum variance, consequently places most Random-All proposals very late in the episode.

After repeated invalid task-8 proposals, we introduced a *replacement-only* fallback that fits a timing prior to the pooled raw endpoint lengths from tasks 2, 3, 5, and 9:

$$\mu = 439.7, \qquad \sigma = 137.43. \tag{2}$$

This fallback is post hoc and is used only when replacing invalid task-8 proposals. The original Random-All selections are not modified retrospectively. The fallback proposes ten additional states and yields three retained recoveries.

Random-Success. Random-Success uses a global Gaussian timing prior fitted to the endpoint lengths of autonomous episodes inferred to be successful because they terminate before the failure horizon. Pooling these durations across the collection tasks gives

$$\mu = 225.70, \qquad \sigma = 47.52. \tag{3}$$

Task 8 contains no successful autonomous episode, so a task-specific successful-duration prior cannot be estimated. We therefore use the same global prior for every collection task, truncated to the eligible interval of the selected episode. Episodes are still sampled from all eligible rollouts, not only successful ones.

Random-Success should therefore be interpreted as a coarse *successful-duration timing control*, not as exact task-progress or semantic-state matching.

Table 3: **Mean timestep of original state proposals.** Values precede random-state replacements and INSIGHT retention filtering and therefore describe candidate distributions rather than final training-set start-state distributions.

Task	INSIGHT First candidates	Random-All	Random-Success
2	215.5	313.6	218.5
3	197.6	363.0	222.2
5	186.5	216.9	192.7
8	221.9	512.3	229.5
9	160.1	395.0	223.8

C.3 Human Recovery Collection and Replacement Sampling

For each selected intermediate state, the simulator is restored exactly and the first author uses a SpaceMouse to attempt a complete expert recovery through task completion. The same operator and teleoperation interface are used for all simulation demonstration conditions. There is no artificial teleoperation timeout.

A successful trajectory is saved immediately. States already satisfying the task or requiring no meaningful expert action are labeled `invalid_no_action_needed`, while states judged structurally unrecoverable are labeled `invalid_unreachable`. Operator mistakes can be marked `retryable`, allowing another attempt from the same state. Only successful recoveries enter the final training dataset.

Each invalid proposal creates one replacement requirement. Replacement sampling follows the original condition's timing rule except for the documented Random-All task-8 fallback. Sampling first prefers previously unused episodes. If those are exhausted, an episode that produced only invalid proposals may be revisited at a different timestep, preferably separated from earlier proposals by at least 25 environment steps. Episodes already contributing a successful demonstration are not reused.

Thus, the final RQ1 datasets have matched task quotas and retained demonstration counts, while the number of candidate states required to obtain those datasets can differ.

Table 4: **Acquisition audit for the random intermediate-state controls.** Counts refer to proposed simulator states, not elapsed human teleoperation time or the number of retry attempts.

Measure	Random-All	Random-Success
Original proposed states	149	149
Retained from original draw	120	136
Original usable yield (%)	80.5	91.3
Replacement proposed states	57	15
Retained from replacements	29	13
Replacement yield (%)	50.9	86.7
Replacement rounds	12	2
Total proposed states	206	164
Final retained demonstrations	149	149
Overall usable yield (%)	72.3	90.9
No-action invalid states	3	6
Unreachable invalid states	54	9

Random-Success requires 42 fewer proposed states than Random-All to obtain the same 149-demonstration training set, corresponding to 20.4% fewer total proposals. Considering only proposals beyond the original target of 149, Random-Success requires 73.7% fewer additional proposals.

Most of this difference arises from task 8, where Random-All requires 67 proposed states to obtain 31 retained recoveries, while Random-Success requires 32. Earlier timing is not uniformly easier, however: on task 5, Random-Success requires 20 proposals for 13 recoveries compared with 17 under Random-All.

Table 5: **Acquisition yield by task.**

Task	Demos	RA proposals	RA yield (%)	RS proposals	RS yield (%)
2	36	41	87.8	36	100.0
3	32	43	74.4	38	84.2
5	13	17	76.5	20	65.0
8	31	67	46.3	32	96.9
9	37	38	97.4	38	97.4

Strong INSIGHT First retains 149 recoveries from 179 candidate first-trigger states, corresponding to a candidate-to-retained fraction of 83.2%. However, the 30 excluded INSIGHT candidates were not annotated using the same invalid-state categories as the random controls. We therefore do not interpret this fraction as a directly comparable acquisition-yield estimate.

C.4 Relationship to INSIGHT First-Trigger States

We additionally examine how the original random proposals relate temporally to INSIGHT first-trigger candidates. This analysis compares the 149 original proposals from each random sampler with the 179 INSIGHT first-trigger candidates. It excludes random replacement states and does not identify which 149 INSIGHT candidates ultimately form the retained training set.

Exact overlap denotes the same rollout episode and stored simulator timestep. It does not measure visual, semantic, or behavioral similarity between nearby states.

Table 6: **Original random-state proposals relative to INSIGHT first-trigger candidates.** The first three rows use all 149 proposals per random method. Remaining timing statistics are calculated only for proposals from episodes containing an INSIGHT flag ($n = 115$ for Random-All and $n = 111$ for Random-Success).

Measure	Random-All	Random-Success
Exact first-trigger match (%)	0.0	0.0
Exact match to any INSIGHT flag (%)	2.0	2.0
Episode contains an INSIGHT flag (%)	77.2	74.5
Before / equal / after first trigger (%)	9.6 / 0.0 / 90.4	36.9 / 0.0 / 63.1
Mean absolute distance (steps)	217.5	100.1
Median absolute distance (steps)	215	80
Mean signed random − INSIGHT	+204.2	+19.2
Median signed displacement	+215	+35
Within 5 / 10 / 25 / 50 steps (%)	0.9 / 1.7 / 7.0 / 11.3	2.7 / 8.1 / 17.1 / 36.9

Random-Success is temporally closer to INSIGHT First than Random-All, but neither sampler reproduces the same first-trigger states. The relatively small signed displacement for Random-Success also partly reflects cancellation between samples occurring before and after the corresponding first trigger.

Normalized rollout phase in this audit is defined as the sampled timestep divided by the last stored timestep of the rollout and should therefore be interpreted as relative elapsed rollout time rather than semantic task progress. On task 8, for example, the mean original-draw phases are 97.6% for Random-All, 43.7% for Random-Success, and 42.3% for INSIGHT candidates.

These observations suggest that different state-selection rules can begin from distinct states while still expose the expert to overlapping remaining subskills or trajectory segments. Successful-demonstration filtering and replacement sampling may further reduce differences among the final

datasets. We treat this as a hypothesis rather than evidence that the retained recovery trajectories are equivalent. Direct comparison of the final trajectory contents remains future work.

RQ2 provides complementary evidence that additional related recovery demonstrations can exhibit diminishing returns at the evaluated operating point, but it does not establish that the different RQ1 datasets contain identical supervision, that a performance ceiling has been reached, or that demonstration count alone causes the observed behavior.

D Statistical Analysis and Checkpoint Sensitivity

The original analysis emphasized the best observed checkpoint for each training condition. Because selecting a maximum from a training trajectory can produce optimistic estimates, the revised primary simulation analysis instead summarizes performance over the common checkpoint grid

$$\mathcal{G} = \{1000, 2000, \ldots, 9000, 9999\}, \tag{4}$$

where checkpoint 9999 denotes the final approximately 10k-step evaluation.

For methods with complete coverage, the primary summary is the equally weighted success mean across the checkpoints in $\mathcal{G}$. We additionally use the late window

$$\mathcal{G}_{\text{late}} = \{8000, 9000, 9999\} \tag{5}$$

as a descriptive sensitivity summary. Fixed-horizon performance at checkpoint 9999 and balanced-grid maxima are also reported where useful. Selected maxima are descriptive because checkpoint selection uses evaluation outcomes. Missing evaluations are never imputed.

Although every checkpoint is evaluated with 50 rollouts per LIBERO-10 task, successive checkpoints from one training trajectory are not independent model replicates. Checkpoint repetition therefore characterizes variation along one optimization trajectory rather than variation across independent fine-tuning seeds.

For configured grid-mean comparisons, we use exact two-sided sign-flip tests over task-level checkpoint means. If d_j denotes the grid-mean difference between two methods on task j, the observed statistic is

$$\left| \frac{1}{m} \sum_{j=1}^{m} d_j \right|, \tag{6}$$

and the reference distribution enumerates all 2^m possible sign assignments. Tasks therefore serve as inferential blocks rather than checkpoints or rollout episodes. The test assumes that task-level differences are independent blocks and sign-exchangeable under the null. With five tasks in either the collected or retained subset, the smallest attainable two-sided exact p-value is 0.0625.

We define two exploratory multiplicity families corresponding to the research questions. The **RQ1 family** contains the six pairwise comparisons among START, INSIGHT First, Random-All, and Random-Success, crossed with overall, collected-task, and retained-task outcomes, for 18 hypotheses. The **RQ2 family** contains INSIGHT First versus INSIGHT All across the same three outcomes, for three hypotheses. Holm correction is applied separately within each family.

These families were defined retrospectively to align the statistical analysis with the paper's research questions after earlier experimental results were available. The corrections therefore control family-wise error within each declared exploratory family, not jointly across every statistical statement in the paper. Dependence among overall, collected, and retained outcomes does not prevent use of the Holm procedure.

A nonsignificant result is interpreted only as a failure to detect a difference under this analysis; it does not establish equivalence or noninferiority. No significance tests are applied to selected best-checkpoint values. The reported task-block tests also do not quantify variability across independent training seeds.

The automatically generated tables that follow report the complete grid-mean and late-window summaries, the RQ1 and RQ2 exploratory comparisons with their family-specific Holm adjustments, and the raw per-checkpoint task results used by the analysis.

E Additional Experimental Results

The following tables are generated directly from the final experiment registry and analysis outputs. The grid mean is the primary simulation summary, while the late-window mean and selected maxima are descriptive sensitivity summaries. Missing checkpoint windows are not imputed.

E.1 Checkpoint-Summary Sensitivity

Table 7 compares the primary common-grid mean with a late-window mean, the final checkpoint, and the maximum observed overall success on the balanced grid. Peak values are selected using evaluation outcomes and are therefore descriptive only.

Table 7: Overall success across alternative checkpoint summaries. Values are percentages; peaks are descriptive.

Method	Grid mean	Late mean	9999	Peak / step
INSIGHT First	65.98	70.27	69.40	72.40 / 9000
START	57.06	57.07	57.40	60.20 / 3000
INSIGHT All	62.40	66.00	69.00	69.00 / 9999
Random-All	67.64	68.27	74.00	74.00 / 9999
Random-Success	67.70	70.40	68.80	73.80 / 9000

E.2 Grid Means and Coverage Across Configurations

O, C, and R denote overall, collected-task, and retained-task success, respectively. A dash indicates that the specified checkpoint window is incomplete; missing evaluations are not treated as zero and are not imputed.

Table 8: Grid-mean and late-window success across experimental configurations. Values are percentages.

Condition	Grid O	Grid C	Grid R	Late O	Late C	Late R
Strong First All-Mix New Norm	65.98	43.48	88.48	70.27	49.33	91.20
Strong First All-Mix Old Norm	–	–	–	–	–	–
Strong First Baseline All-Mix New Norm	57.06	27.28	86.84	57.07	26.00	88.13
Strong First Baseline All-Mix Old Norm	–	–	–	–	–	–
Strong First Only New Norm	–	–	–	–	–	–
Strong First Only Old Norm	–	–	–	–	–	–
Strong All All-Mix New Norm	62.40	42.08	82.72	66.00	46.27	85.73
Strong All All-Mix Old Norm	–	–	–	–	–	–
Strong All Only New Norm	–	–	–	–	–	–
Strong All Only Old Norm	–	–	–	–	–	–
Weak First All-Mix New Norm	66.90	48.52	85.28	67.67	48.93	86.40
Weak First All-Mix Old Norm	60.48	33.20	87.76	62.47	36.93	88.00
Weak First Baseline All-Mix Old Norm	56.60	23.88	89.32	56.87	22.13	91.60
Weak First Baseline All-Mix New Norm	–	–	–	58.00	25.20	90.80
Weak First Only Old Norm	–	–	–	–	–	–
Fisher 1e6 Low LR	–	–	–	–	–	–
Fisher 1e8 Low LR	53.44	18.36	88.52	51.07	15.20	86.93
Fisher 1e9 Low LR	–	–	–	54.27	19.07	89.47
Fisher 1e10 Low LR	58.02	26.60	89.44	58.20	26.40	90.00
Fisher 1e12 Low LR	58.70	28.44	88.96	58.33	27.87	88.80
Fisher 1e15 Low LR	58.46	27.60	89.32	57.87	26.67	89.07
Filtered Fisher 1e8 Low LR	53.74	18.28	89.20	51.20	14.00	88.40
Filtered Fisher 1e13 Low LR	60.00	30.40	89.60	59.87	29.07	90.67
Filtered Fisher 1e14 Low LR	–	–	–	–	–	–
Random-All	67.64	47.48	87.80	68.27	49.87	86.67

Continued on next page

Table 8: Grid-mean and late-window success across experimental configurations. Values are percentages.

Condition	Grid O	Grid C	Grid R	Late O	Late C	Late R
Random-Success	67.70	47.16	88.24	70.40	52.53	88.27
Filtered Fisher 1e15 Low LR	58.42	27.68	89.16	59.73	29.73	89.73
Strong First Only Low LR	–	–	–	–	–	–
Fisher 1e10	–	–	–	–	–	–
Fisher 1e12	59.32	29.04	89.60	60.40	30.67	90.13
Fisher 1e15	59.68	28.96	90.40	60.20	30.40	90.00
Fisher 1e10 Strong First Libero 10 Mix	–	–	–	–	–	–
Strong First Libero 10 Mix	62.80	36.96	88.64	64.13	38.67	89.60
Strong First Libero10 Collected Mix	–	–	–	–	–	–
Strong First Libero10 Collected Mix Fisher 1e8	–	–	–	–	–	–
Strong First Libero10 Collected Mix Fisher 1e10	–	–	–	–	–	–
Strong First Libero10 Collected Mix Filtered Fisher 1e9	–	–	–	–	–	–
Strong First Libero10 Collected Mix Filtered Fisher 1e10	–	–	–	–	–	–

E.3 RQ1: State-Selection Comparisons

Table 9 reports exact two-sided task-block sign-flip tests on the common-grid means. The RQ1 family contains six state-selection comparisons crossed with overall, collected, and retained outcomes, for 18 exploratory tests. Holm correction is applied within this family.

Table 9: Exact task-block sign-flip comparisons for RQ1 grid means. Δ is Method A minus Method B in percentage points.

Comparison	Subset	m	Δ pp	Raw p	Holm p
INSIGHT First − START	overall	10	8.92000	0.00977	0.17578
INSIGHT First − START	collected	5	16.20000	0.06250	0.93750
INSIGHT First − START	retained	5	1.64000	0.31250	1.00000
INSIGHT First − Random-All	overall	10	-1.66000	0.58984	1.00000
INSIGHT First − Random-All	collected	5	-4.00000	0.50000	1.00000
INSIGHT First − Random-All	retained	5	0.68000	0.68750	1.00000
INSIGHT First − Random-Success	overall	10	-1.72000	0.38281	1.00000
INSIGHT First − Random-Success	collected	5	-3.68000	0.25000	1.00000
INSIGHT First − Random-Success	retained	5	0.24000	0.93750	1.00000
Random-All − START	overall	10	10.58000	0.02148	0.34375
Random-All − START	collected	5	20.20000	0.06250	0.93750
Random-All − START	retained	5	0.96000	0.68750	1.00000
Random-Success − START	overall	10	10.64000	0.01367	0.23242
Random-Success − START	collected	5	19.88000	0.06250	0.93750
Random-Success − START	retained	5	1.40000	0.43750	1.00000
Random-Success − Random-All	overall	10	0.06000	0.99219	1.00000
Random-Success − Random-All	collected	5	-0.32000	0.93750	1.00000
Random-Success − Random-All	retained	5	0.44000	0.87500	1.00000

E.4 RQ2: First-Trigger versus All-Trigger Collection

Table 10 reports the corresponding First-versus-All comparison. The RQ2 family contains overall, collected, and retained outcomes, for three exploratory tests. Holm correction is applied within this family.

Table 10: Exact task-block sign-flip comparisons for RQ2 grid means. Δ is INSIGHT First minus INSIGHT All in percentage points.

Comparison	Subset	m	Δ pp	Raw p	Holm p
INSIGHT First − INSIGHT All	overall	10	3.58000	0.35352	0.70703
INSIGHT First − INSIGHT All	collected	5	1.40000	0.93750	0.93750
INSIGHT First − INSIGHT All	retained	5	5.76000	0.12500	0.37500

E.5 Raw Per-Checkpoint Results

The following tables report the unmodified per-task success counts used by the analysis. T0–T9 denote the ten LIBERO-10 tasks. Each task is evaluated with 50 autonomous rollouts per checkpoint. Overall, Collected, and Retained are success percentages computed from these counts. Missing checkpoints were not evaluated and are not imputed.

Table 11: Raw per-checkpoint success counts and percentages: Strong First All-Mix New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	45	48	0	15	43	19	41	46	0	2	51.8	14.4	89.2
1000	47	50	4	42	44	28	45	47	4	22	66.6	40.0	93.2
1500	45	49	15	40	32	25	36	43	6	15	61.2	40.4	82.0
2000	40	44	14	31	43	36	39	46	8	13	62.8	40.8	84.8
2500	46	47	20	30	45	35	43	47	12	11	67.2	43.2	91.2
3000	50	47	28	38	45	28	40	44	5	10	67.0	43.6	90.4
3500	42	48	19	37	39	29	45	46	9	23	67.4	46.8	88.0
4000	46	46	21	34	43	36	44	48	2	15	67.0	43.2	90.8
4500	41	49	33	45	41	22	45	46	8	12	68.4	48.0	88.8
5000	40	49	6	26	37	25	33	43	7	11	55.4	30.0	80.8
5500	42	46	28	31	43	27	38	46	5	21	65.4	44.8	86.0
6000	43	48	28	32	43	31	36	46	4	23	66.8	47.2	86.4
6500	46	43	15	31	41	18	42	48	10	22	63.2	38.4	88.0
7000	43	48	16	29	39	32	34	48	0	28	63.4	42.0	84.8
7500	48	48	18	39	46	30	35	49	13	25	70.2	50.0	90.4
8000	44	47	26	34	42	35	37	48	10	22	69.0	50.8	87.2
8500	45	47	26	34	36	35	41	49	4	26	68.6	50.0	87.2
9000	48	48	28	36	48	29	46	49	7	23	72.4	49.2	95.6
9500	43	50	19	42	45	28	43	46	6	14	67.2	43.6	90.8
9999	45	48	23	36	47	31	39	48	7	23	69.4	48.0	90.8

Table 12: Raw per-checkpoint success counts and percentages: Strong First All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	44	49	11	15	47	28	39	49	0	10	58.4	25.6	91.2
3000	48	49	15	29	39	23	38	48	0	19	61.6	34.4	88.8
5000	46	0	24	22	43	15	36	47	0	15	49.6	30.4	68.8

Table 13: Raw per-checkpoint success counts and percentages: Strong First Baseline All-Mix New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	45	50	3	22	42	26	29	47	2	10	55.2	25.2	85.2
2000	44	50	0	25	41	33	39	44	0	19	59.0	30.8	87.2
3000	37	48	4	33	45	24	43	45	4	18	60.2	33.2	87.2
4000	40	49	3	17	42	27	41	49	0	8	55.2	22.0	88.4
5000	41	49	3	23	34	31	37	49	0	10	55.4	26.8	84.0
5500	41	49	3	28	41	24	37	48	0	18	57.8	29.2	86.4
6000	42	49	15	14	37	30	42	45	0	14	57.6	29.2	86.0
6500	38	48	7	17	44	29	35	46	1	14	55.8	27.2	84.4
7000	42	47	7	18	40	25	41	45	0	19	56.8	27.6	86.0
8000	36	49	9	15	47	27	37	48	1	13	56.4	26.0	86.8
9000	47	49	7	27	42	24	31	50	0	10	57.4	27.2	87.6
9999	47	47	3	21	47	28	35	49	2	8	57.4	24.8	90.0

Table 14: Raw per-checkpoint success counts and percentages: Strong First Baseline All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	45	50	10	24	46	29	39	49	0	11	60.6	29.6	91.6

Continued on next page

Table 14: Raw per-checkpoint success counts and percentages: Strong First Baseline All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
3000	0	48	11	16	45	15	39	48	1	20	48.6	25.2	72.0
5000	37	48	2	24	42	17	45	48	1	8	54.4	20.8	88.0

Table 15: Raw per-checkpoint success counts and percentages: Strong First Only New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
3000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
5000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0

Table 16: Raw per-checkpoint success counts and percentages: Strong First Only Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	34	32	0	1	16	0	25	34	0	0	28.4	0.4	56.4
3000	25	13	0	0	0	0	0	4	0	0	8.4	0.0	16.8
5000	2	6	0	5	0	0	0	0	0	0	2.6	2.0	3.2

Table 17: Raw per-checkpoint success counts and percentages: Strong All All-Mix New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	0	43	2	20	34	24	30	46	0	3	40.4	19.6	61.2
2000	41	48	0	29	45	39	28	49	2	3	56.8	29.2	84.4
3000	42	47	6	34	45	46	34	41	1	21	63.4	43.2	83.6
4000	39	50	9	41	47	46	23	47	6	4	62.4	42.4	82.4
5000	41	45	17	38	42	45	46	47	10	11	68.4	48.4	88.4
6000	43	47	7	44	38	49	40	44	3	20	67.0	49.2	84.8
7000	45	47	22	41	41	45	30	50	3	14	67.6	50.0	85.2
8000	44	47	14	30	43	38	36	49	5	26	66.4	45.2	87.6
9000	44	44	10	37	26	47	42	47	6	10	62.6	44.0	81.2
9999	45	48	13	40	34	45	45	49	6	20	69.0	49.6	88.4

Table 18: Raw per-checkpoint success counts and percentages: Strong All All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	42	46	11	25	46	20	36	48	0	10	56.8	26.4	87.2
3000	47	48	3	21	48	37	39	45	0	6	58.8	26.8	90.8
5000	36	50	7	18	39	19	43	47	0	1	52.0	18.0	86.0

Table 19: Raw per-checkpoint success counts and percentages: Strong All Only New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
3000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
5000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0

Table 20: Raw per-checkpoint success counts and percentages: Strong All Only Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	15	7	0	0	11	0	18	30	0	0	16.2	0.0	32.4
3000	1	0	0	0	0	0	0	3	0	0	0.8	0.0	1.6
5000	0	2	0	0	0	0	0	0	0	0	0.4	0.0	0.8

Table 21: Raw per-checkpoint success counts and percentages: Weak First All-Mix New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	45	50	0	3	39	32	40	43	1	0	50.6	14.4	86.8
1000	44	45	25	30	44	41	46	44	5	5	65.8	42.4	89.2
1500	38	47	25	38	43	41	33	42	2	23	66.4	51.6	81.2
2000	44	49	17	36	42	36	33	41	5	23	65.2	46.8	83.6
2500	42	47	18	44	45	40	31	49	9	20	69.0	52.4	85.6
3000	38	44	23	31	46	29	39	46	4	19	63.8	42.4	85.2
3500	38	46	25	37	36	24	35	45	3	11	60.0	40.0	80.0
4000	23	47	28	39	35	33	37	43	6	12	60.6	47.2	74.0
4500	44	46	21	34	36	39	36	38	3	23	64.0	48.0	80.0
5000	39	45	38	41	44	31	38	46	5	20	69.4	54.0	84.8
5500	43	45	26	40	34	33	38	43	6	18	65.2	49.2	81.2
6000	46	47	26	29	42	35	38	47	11	33	70.8	53.6	88.0
6500	43	47	32	32	36	40	34	46	3	31	68.8	55.2	82.4
7000	49	45	36	36	41	32	43	44	4	22	70.4	52.0	88.8
7500	41	46	21	38	44	42	42	47	2	28	70.2	52.4	88.0
8000	40	46	32	41	45	21	37	47	1	24	66.8	47.6	86.0
8500	31	46	29	42	49	35	39	46	7	24	69.6	54.8	84.4
9000	40	48	29	38	41	31	34	47	8	28	68.8	53.6	84.0
9500	39	47	29	36	43	39	40	48	5	28	70.8	54.8	86.8
9999	43	49	28	38	47	28	38	46	4	16	67.4	45.6	89.2

Table 22: Raw per-checkpoint success counts and percentages: Weak First All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	43	50	25	22	48	28	41	49	3	20	65.8	39.2	92.4
1000	45	46	2	19	43	39	47	48	3	21	62.6	33.6	91.6
1500	46	49	20	36	42	29	39	43	0	11	63.0	38.4	87.6
2000	41	46	11	23	45	29	39	49	0	12	59.0	30.0	88.0
2500	45	47	11	23	40	24	41	45	0	17	58.6	30.0	87.2
3000	43	48	22	36	42	24	35	48	0	7	61.0	35.6	86.4
3500	46	49	24	13	37	27	34	48	2	15	59.0	32.4	85.6
4000	41	50	18	16	43	29	37	49	0	11	58.8	29.6	88.0
4500	43	50	17	25	46	38	38	46	0	8	62.2	35.2	89.2
5000	44	48	16	16	41	24	40	47	0	11	57.4	26.8	88.0
5500	44	48	10	33	43	28	45	46	1	16	62.8	35.2	90.4
6000	37	46	13	15	38	23	38	49	0	15	54.8	26.4	83.2
6500	44	48	12	31	44	18	37	46	2	15	59.4	31.2	87.6
7000	44	49	22	21	45	28	35	48	2	25	63.8	39.2	88.4
7500	48	47	24	20	43	32	34	43	2	18	62.2	38.4	86.0
8000	42	49	16	23	44	35	42	45	1	22	63.8	38.8	88.8
8500	45	47	19	22	44	29	39	43	1	9	59.6	32.0	87.2
9000	47	49	17	25	40	20	41	48	1	19	61.4	32.8	90.0
9500	43	48	20	27	39	22	38	47	0	23	61.4	36.8	86.0
9999	36	47	26	34	47	23	37	46	1	14	62.2	39.2	85.2

Table 23: Raw per-checkpoint success counts and percentages: Weak First Baseline All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	46	49	8	24	45	26	41	48	0	14	60.2	28.8	91.6
1000	45	48	17	25	46	27	41	45	1	16	62.2	34.4	90.0
1500	45	49	5	26	46	17	39	46	1	12	57.2	24.4	90.0
2000	39	49	7	10	43	22	33	44	0	16	52.6	22.0	83.2

Continued on next page

Table 23: Raw per-checkpoint success counts and percentages: Weak First Baseline All-Mix Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
2500	42	48	9	9	47	28	41	48	1	20	58.6	26.8	90.4
3000	35	49	2	18	47	24	40	47	0	14	55.2	23.2	87.2
3500	45	50	11	18	41	20	39	46	0	14	56.8	25.2	88.4
4000	45	48	14	15	45	15	44	45	0	13	56.8	22.8	90.8
4500	42	49	5	13	44	22	42	46	0	18	56.2	23.2	89.2
5000	45	49	9	17	46	23	38	44	0	22	58.6	28.4	88.8
5500	40	49	4	15	45	24	40	43	1	9	54.0	21.2	86.8
6000	43	49	3	16	39	24	45	44	0	7	54.0	20.0	88.0
6500	43	50	2	10	44	15	39	46	0	5	50.8	12.8	88.8
7000	42	50	4	17	46	25	41	47	0	8	56.0	21.6	90.4
7500	41	48	8	15	40	15	36	49	0	8	52.0	18.4	85.6
8000	43	50	5	10	48	18	43	48	0	9	54.8	16.8	92.8
8500	42	47	2	17	42	23	37	46	0	9	53.0	20.4	85.6
9000	47	48	2	23	43	17	37	46	1	23	57.4	26.4	88.4
9500	45	47	2	22	36	20	35	46	1	10	52.8	22.0	83.6
9999	44	48	6	12	46	23	48	48	0	17	58.4	23.2	93.6

Table 24: Raw per-checkpoint success counts and percentages: Weak First Baseline All-Mix New Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	46	49	0	15	34	19	36	41	0	5	49.0	15.6	82.4
1000	45	50	0	20	47	21	41	43	2	10	55.8	21.2	90.4
1500	42	50	1	13	40	13	28	47	1	9	48.8	14.8	82.8
2000	46	48	1	21	48	27	37	41	0	13	56.4	24.8	88.0
2500	46	50	4	12	47	25	32	45	2	6	53.8	19.6	88.0
3000	45	47	6	30	39	26	39	41	0	13	57.2	30.0	84.4
3500	42	47	1	29	47	17	36	46	1	8	54.8	22.4	87.2
4000	46	47	3	14	46	13	43	44	0	8	52.8	15.2	90.4
4500	47	49	15	16	42	17	38	47	1	14	57.2	25.2	89.2
5000	41	48	2	22	43	32	41	46	0	1	55.2	22.8	87.6
5500	41	48	5	15	44	18	36	43	1	10	52.2	19.6	84.8
6000	45	46	12	28	41	28	43	44	2	16	61.0	34.4	87.6
7500	47	49	8	19	46	21	33	47	3	15	57.6	26.4	88.8
8000	42	50	4	20	47	24	42	47	0	14	58.0	24.8	91.2
8500	47	48	5	33	47	28	40	47	3	10	61.6	31.6	91.6
9000	45	48	10	15	45	23	43	47	0	18	58.8	26.4	91.2
9500	42	49	9	10	44	30	41	49	0	6	56.0	22.0	90.0
9999	48	49	9	15	42	31	39	47	1	5	57.2	24.4	90.0

Table 25: Raw per-checkpoint success counts and percentages: Weak First Only Old Norm

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	46	45	5	5	30	0	26	38	0	0	39.0	4.0	74.0
1000	47	40	0	4	18	0	28	42	0	0	35.8	1.6	70.0
1500	44	32	0	16	11	6	20	33	0	0	32.4	8.8	56.0
2000	42	19	0	5	7	2	5	31	0	0	22.2	2.8	41.6
2500	19	9	0	3	1	0	1	13	0	0	9.2	1.2	17.2
3000	20	21	0	11	2	0	2	11	0	0	13.4	4.4	22.4
3500	8	15	0	8	0	0	0	10	0	0	8.2	3.2	13.2
4000	4	16	0	12	0	2	0	3	0	0	7.4	5.6	9.2
4500	4	10	0	12	0	0	0	1	0	0	5.4	4.8	6.0
5000	0	6	0	9	0	1	0	2	0	0	3.6	4.0	3.2
5500	0	4	0	5	0	0	0	1	0	0	2.0	2.0	2.0

Table 26: Raw per-checkpoint success counts and percentages: Fisher 1e6 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	43	48	9	22	44	27	41	46	0	11	58.2	27.6	88.8
1500	44	50	8	19	45	29	34	49	1	8	57.4	26.0	88.8
2000	45	49	1	20	46	21	41	48	1	4	55.2	18.8	91.6
2500	39	50	1	18	46	21	37	48	5	3	53.6	19.2	88.0
3000	44	50	3	19	45	27	36	47	2	0	54.6	20.4	88.8
3500	42	48	4	16	45	25	37	48	2	1	53.6	19.2	88.0
4000	45	50	0	18	48	22	36	49	0	3	54.2	17.2	91.2
4500	47	48	2	11	45	25	42	45	1	0	53.2	15.6	90.8
4999	41	48	3	7	45	25	41	45	3	1	51.8	15.6	88.0

Table 27: Raw per-checkpoint success counts and percentages: Fisher 1e8 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	43	49	14	22	44	26	41	45	1	11	59.2	29.6	88.8
1000	37	49	7	20	46	27	41	47	0	9	56.6	25.2	88.0
1500	45	50	4	18	43	25	37	46	1	7	55.2	22.0	88.4
2000	46	49	2	21	46	29	35	48	0	5	56.2	22.8	89.6
2500	44	49	0	21	47	29	41	49	2	2	56.8	21.6	92.0
3000	44	48	4	20	47	25	37	49	3	0	55.4	20.8	90.0
3500	42	50	2	13	46	28	44	48	4	1	55.6	19.2	92.0
4000	42	49	4	16	47	23	40	47	1	0	53.8	17.6	90.0
4500	40	50	3	13	45	26	40	49	0	1	53.4	17.2	89.6
5000	44	50	5	12	45	23	35	47	0	0	52.2	16.0	88.4
5500	43	49	3	10	47	27	39	47	3	0	53.6	17.2	90.0
6000	41	50	6	12	46	27	38	49	2	0	54.2	18.8	89.6
6500	44	49	1	8	45	27	43	46	1	0	52.8	14.8	90.8
7000	42	50	6	8	46	24	36	48	4	0	52.8	16.8	88.8
7500	47	50	4	5	43	24	40	45	4	0	52.4	14.8	90.0
8000	44	49	3	10	39	31	41	49	2	0	53.6	18.4	88.8
8500	47	49	11	9	41	24	38	49	1	1	54.0	18.4	89.6
9000	45	48	7	10	37	17	37	47	1	0	49.8	14.0	85.6
9999	40	48	3	8	45	21	38	45	1	0	49.8	13.2	86.4

Table 28: Raw per-checkpoint success counts and percentages: Fisher 1e9 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	41	49	20	16	46	19	39	45	1	14	58.0	28.0	88.0
1000	43	49	13	19	48	25	34	48	1	7	57.4	26.0	88.8
1500	46	49	6	21	47	30	39	48	1	5	58.4	25.2	91.6
2000	43	49	6	19	43	25	42	47	0	5	55.8	22.0	89.6
2500	41	49	7	20	49	25	39	48	1	1	56.0	21.6	90.4
3000	45	49	4	20	45	25	38	48	1	3	55.6	21.2	90.0
3500	45	50	5	17	44	23	39	48	3	4	55.6	20.8	90.4
4500	41	50	3	12	47	24	36	48	3	2	53.2	17.6	88.8
4999	44	48	1	27	47	17	38	47	2	5	55.2	20.8	89.6
5500	43	49	2	15	44	28	40	48	1	3	54.6	19.6	89.6
6000	45	50	4	16	42	27	37	47	2	7	55.4	22.4	88.4
6500	46	50	6	17	47	22	39	48	3	1	55.8	19.6	92.0
7000	46	49	5	19	45	22	36	49	2	4	55.4	20.8	90.0
7500	42	49	10	17	43	29	41	48	1	5	57.0	24.8	89.2
8000	42	49	6	15	46	25	39	47	2	3	54.8	20.4	89.2
8500	44	49	3	13	45	27	42	49	3	2	55.4	19.2	91.6
9000	44	48	6	13	45	23	43	49	1	1	54.6	17.6	91.6
9999	42	50	5	14	43	25	37	47	2	2	53.4	19.2	87.6

Table 29: Raw per-checkpoint success counts and percentages: Fisher 1e10 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	43	48	12	17	46	23	42	46	1	17	59.0	28.0	90.0
1000	40	49	11	24	45	24	39	46	0	11	57.8	28.0	87.6

Continued on next page

Table 29: Raw per-checkpoint success counts and percentages: Fisher 1e10 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1500	46	49	10	16	45	29	42	46	3	14	60.0	28.8	91.2
2000	37	49	12	18	48	29	42	47	1	9	58.4	27.6	89.2
2500	46	49	5	17	47	26	42	48	1	9	58.0	23.2	92.8
3000	45	48	5	18	45	32	38	46	0	14	58.2	27.6	88.8
3500	42	48	9	23	46	27	43	47	1	17	60.6	30.8	90.4
4000	44	49	13	16	45	32	41	48	2	11	60.2	29.6	90.8
4500	48	49	6	23	46	23	41	47	1	11	59.0	25.6	92.4
5000	43	49	4	18	44	22	37	47	0	10	54.8	21.6	88.0
5500	45	50	6	19	45	27	41	46	0	9	57.6	24.4	90.8
6000	44	49	4	29	45	28	41	46	0	6	58.4	26.8	90.0
6500	45	49	1	14	47	28	39	46	1	14	56.8	23.2	90.4
7000	45	49	6	18	44	26	41	46	1	13	57.8	25.6	90.0
7500	42	49	8	18	43	24	40	47	2	13	57.2	26.0	88.4
8000	38	49	3	21	45	31	41	48	0	18	58.8	29.2	88.4
8500	42	48	6	16	47	28	39	47	3	8	56.8	24.4	89.2
9000	44	49	5	15	46	29	44	46	0	17	59.0	26.4	91.6
9999	44	48	4	19	47	26	39	47	1	9	56.8	23.6	90.0

Table 30: Raw per-checkpoint success counts and percentages: Fisher 1e12 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	48	48	13	15	45	26	40	43	0	14	58.4	27.2	89.6
1000	43	49	17	21	47	25	39	45	0	15	60.2	31.2	89.2
1500	45	49	15	18	45	27	42	46	0	16	60.6	30.4	90.8
2000	42	48	16	17	45	23	35	46	1	10	56.6	26.8	86.4
2500	46	48	21	18	45	25	39	44	0	10	59.2	29.6	88.8
3000	42	48	13	19	47	24	42	45	0	9	57.8	26.0	89.6
3500	43	49	11	16	44	26	42	47	0	16	58.8	27.6	90.0
4000	43	49	15	19	46	32	41	46	0	13	60.8	31.6	90.0
4500	43	49	19	16	46	26	41	46	0	11	59.4	28.8	90.0
5000	43	49	13	17	43	30	38	46	1	18	59.6	31.6	87.6
5500	44	48	14	20	44	27	43	45	0	12	59.4	29.2	89.6
6000	44	50	12	16	45	27	43	46	0	14	59.4	27.6	91.2
6500	43	49	16	15	45	29	44	45	2	11	59.8	29.2	90.4
7000	43	49	13	16	45	27	41	45	0	9	57.6	26.0	89.2
7500	44	49	17	16	44	26	41	47	0	12	59.2	28.4	90.0
8000	42	49	15	17	43	20	39	49	0	11	57.0	25.2	88.8
8500	40	49	16	19	46	24	44	46	0	12	59.2	28.4	90.0
9000	43	48	8	22	46	21	36	44	1	14	56.6	26.4	86.8
9999	43	48	15	22	49	27	43	44	0	16	61.4	32.0	90.8

Table 31: Raw per-checkpoint success counts and percentages: Fisher 1e15 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	42	48	16	26	43	27	45	45	0	11	60.6	32.0	89.2
1000	43	49	9	15	42	25	41	46	0	14	56.8	25.2	88.4
1500	42	49	20	16	46	28	42	44	0	10	59.4	29.6	89.2
2000	45	49	23	17	46	26	39	45	0	8	59.6	29.6	89.6
2500	45	48	12	22	47	25	37	43	0	8	57.4	26.8	88.0
3000	43	49	18	16	45	19	43	45	1	13	58.4	26.8	90.0
3500	44	47	11	22	44	31	37	48	0	13	59.4	30.8	88.0
4000	45	50	14	15	45	27	40	47	0	12	59.0	27.2	90.8
4500	44	49	21	18	46	27	35	45	0	13	59.6	31.6	87.6
5000	41	49	16	22	44	26	40	44	0	10	58.4	29.6	87.2
5500	42	48	17	14	45	29	39	47	0	9	58.0	27.6	88.4
6000	46	49	13	17	46	25	42	46	0	13	59.4	27.2	91.6
6500	43	49	18	19	46	25	43	41	1	9	58.8	28.8	88.8
7000	45	48	18	17	45	28	39	44	1	12	59.4	30.4	88.4
7500	44	49	12	20	45	19	39	46	0	10	56.8	24.4	89.2
8000	45	49	9	15	46	25	41	47	0	9	57.2	23.2	91.2
8500	44	49	15	20	45	28	40	46	0	13	60.0	30.4	89.6
9000	43	49	15	11	44	29	35	47	0	17	58.0	28.8	87.2
9999	46	48	13	15	43	30	40	45	0	12	58.4	28.0	88.8

Table 32: Raw per-checkpoint success counts and percentages: Filtered Fisher 1e8 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	39	49	13	18	48	32	37	48	1	15	60.0	31.6	88.4
1000	45	48	9	26	42	30	37	46	0	11	58.8	30.4	87.2
1500	41	49	5	22	44	25	43	46	1	6	56.4	23.6	89.2
2000	42	49	2	18	46	28	38	48	0	4	55.0	20.8	89.2
2500	40	49	2	18	46	23	37	48	1	3	53.4	18.8	88.0
3000	42	50	2	14	45	29	40	48	1	2	54.6	19.2	90.0
3500	40	49	4	14	47	24	40	50	3	0	54.2	18.0	90.4
4000	41	50	2	13	47	28	37	46	2	2	53.6	18.8	88.4
4500	43	49	2	16	47	25	41	47	3	0	54.6	18.4	90.8
5000	41	49	8	10	46	27	38	50	2	1	54.4	19.2	89.6
5500	44	48	4	12	48	23	43	48	2	0	54.4	16.4	92.4
6000	41	49	3	9	46	28	41	48	0	0	53.0	16.0	90.0
6500	42	50	8	8	42	23	39	47	3	0	52.4	16.8	88.0
7000	46	50	1	10	42	26	43	50	3	1	54.4	16.4	92.4
7500	43	49	4	7	45	25	38	46	1	0	51.6	14.8	88.4
8000	46	50	5	8	39	20	45	42	0	0	51.0	13.2	88.8
8500	43	47	1	6	42	19	42	47	0	0	49.4	10.4	88.4
9000	44	50	6	8	43	21	40	46	1	0	51.8	14.4	89.2
9999	44	49	6	8	44	21	37	44	1	0	50.8	14.4	87.2

Table 33: Raw per-checkpoint success counts and percentages: Filtered Fisher 1e13 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	44	49	19	21	41	25	38	46	0	14	59.4	31.6	87.2
1000	44	48	14	21	47	29	43	47	0	13	61.2	30.8	91.6
1500	42	49	14	16	46	24	37	45	0	9	56.4	25.2	87.6
2000	43	49	17	18	44	28	40	45	0	13	59.4	30.4	88.4
2500	41	48	13	17	45	22	41	44	0	17	57.6	27.6	87.6
3000	40	48	18	18	46	26	36	47	2	21	60.4	34.0	86.8
3500	42	49	19	14	43	22	41	46	0	12	57.6	26.8	88.4
4000	45	49	16	16	46	29	37	44	0	18	60.0	31.6	88.4
4500	45	48	11	23	46	30	43	46	1	8	60.2	29.2	91.2
5000	43	49	11	21	44	23	39	48	0	17	59.0	28.8	89.2
5500	41	49	15	13	46	27	39	48	0	16	58.8	28.4	89.2
6000	45	50	17	16	42	20	44	47	1	23	61.0	30.8	91.2
6500	45	48	14	16	48	26	39	47	1	16	60.0	29.2	90.8
7000	43	49	16	17	45	28	41	43	0	15	59.4	30.4	88.4
7500	44	49	16	16	44	30	39	46	0	15	59.8	30.8	88.8
8000	46	49	19	18	47	26	43	48	0	12	61.6	30.0	93.2
8500	46	49	21	17	46	25	41	48	0	13	61.2	30.4	92.0
9000	42	49	14	19	47	21	38	49	0	15	58.8	27.6	90.0
9999	41	49	16	19	46	24	39	47	0	15	59.2	29.6	88.8

Table 34: Raw per-checkpoint success counts and percentages: Filtered Fisher 1e14 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	43	49	14	21	46	26	40	47	1	15	60.4	30.8	90.0
1000	41	48	16	22	44	28	42	46	0	13	60.0	31.6	88.4
1500	42	49	8	17	44	27	37	44	0	7	55.0	23.6	86.4
2000	40	48	17	23	44	27	36	48	0	11	58.8	31.2	86.4
2500	44	49	16	20	42	24	38	47	0	15	59.0	30.0	88.0
3000	44	48	18	18	45	24	41	45	0	12	59.0	28.8	89.2
3500	44	49	11	16	46	26	40	46	1	11	58.0	26.0	90.0
4000	44	48	13	19	48	22	42	48	0	9	58.6	25.2	92.0
4500	46	49	14	18	45	30	40	45	0	12	59.8	29.6	90.0
5000	42	48	15	13	45	29	38	45	1	18	58.8	30.4	87.2
5500	43	48	16	15	45	22	37	43	0	10	55.8	25.2	86.4
6000	43	49	6	15	46	26	43	46	0	17	58.2	25.6	90.8
6500	42	49	19	25	46	24	45	46	0	10	61.2	31.2	91.2
7000	45	49	15	20	47	26	38	45	2	14	60.2	30.8	89.6
7500	46	49	13	18	43	30	38	44	0	11	58.4	28.8	88.0

Table 35: Raw per-checkpoint success counts and percentages: Random-All

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	46	49	9	19	44	31	42	45	5	21	62.2	34.0	90.4
2000	43	49	29	37	35	28	43	47	20	15	69.2	51.6	86.8
3000	45	48	26	36	42	35	42	50	12	24	72.0	53.2	90.8
4000	41	47	34	36	43	36	46	50	10	11	70.8	50.8	90.8
5000	45	48	35	19	44	33	39	47	4	19	66.6	44.0	89.2
6000	39	48	30	28	37	34	37	49	8	19	65.8	47.6	84.0
7000	42	48	32	30	38	22	40	47	13	13	65.0	44.0	86.0
8000	33	48	21	20	42	19	38	47	6	31	61.0	38.8	83.2
9000	40	44	34	35	48	33	42	48	6	19	69.8	50.8	88.8
9999	44	45	41	38	46	31	39	46	10	30	74.0	60.0	88.0

Table 36: Raw per-checkpoint success counts and percentages: Random-Success

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	40	47	17	36	49	41	42	48	16	14	70.0	49.6	90.4
2000	44	46	8	37	40	44	44	48	11	16	67.6	46.4	88.8
3000	38	49	24	34	39	37	43	48	2	14	65.6	44.4	86.8
4000	45	47	15	29	44	36	38	46	2	12	62.8	37.6	88.0
5000	44	47	17	27	46	40	44	48	9	17	67.8	44.0	91.6
6000	40	45	21	31	45	29	41	48	8	29	67.4	47.2	87.6
7000	42	46	16	37	43	38	35	45	7	14	64.6	44.8	84.4
8000	45	46	23	35	45	42	37	47	5	18	68.6	49.2	88.0
9000	41	47	35	39	43	39	40	50	9	26	73.8	59.2	88.4
9999	41	47	31	27	44	37	41	48	4	24	68.8	49.2	88.4

Table 37: Raw per-checkpoint success counts and percentages: Filtered Fisher 1e15 Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	40	48	19	22	43	22	39	45	0	10	57.6	29.2	86.0
1000	41	50	10	15	44	28	44	47	0	12	58.2	26.0	90.4
1500	45	49	12	18	45	25	44	43	1	15	59.4	28.4	90.4
2000	44	49	17	19	44	25	39	46	0	11	58.8	28.8	88.8
2500	43	49	15	16	45	22	41	48	0	10	57.8	25.2	90.4
3000	43	49	15	15	46	27	36	42	0	11	56.8	27.2	86.4
3500	46	50	18	16	45	22	41	45	0	11	58.8	26.8	90.8
4000	41	49	12	17	40	28	38	46	0	11	56.4	27.2	85.6
4500	42	48	10	17	46	28	38	48	0	13	58.0	27.2	88.8
5000	42	48	15	20	45	23	44	47	0	15	59.8	29.2	90.4
5500	46	49	17	21	45	28	40	49	0	14	61.8	32.0	91.6
6000	48	49	14	13	49	22	36	47	0	11	57.8	24.0	91.6
6500	45	49	19	23	47	24	39	46	0	14	61.2	32.0	90.4
7000	44	49	8	20	44	25	41	45	0	10	57.2	25.2	89.2
7500	45	48	12	18	46	28	43	48	1	10	59.8	27.6	92.0
8000	42	49	21	21	46	20	43	46	0	13	60.2	30.0	90.4
8500	40	49	11	16	45	27	39	45	0	15	57.4	27.6	87.2
9000	43	49	10	20	47	26	44	45	2	16	60.4	29.6	91.2
9999	42	48	13	19	45	27	38	46	0	15	58.6	29.6	87.6

Table 38: Raw per-checkpoint success counts and percentages: Strong First Only Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
100	44	50	20	27	47	27	41	46	1	11	62.8	34.4	91.2
200	42	50	16	16	46	21	38	47	0	12	57.6	26.0	89.2
300	43	49	17	16	40	23	43	49	0	17	59.4	29.2	89.6
400	42	49	17	27	46	28	44	43	0	7	60.6	31.6	89.6
500	42	48	16	31	44	28	43	42	0	15	61.8	36.0	87.6
600	45	50	12	21	43	25	37	44	0	13	58.0	28.4	87.6
700	45	50	16	20	47	31	41	50	2	8	62.0	30.8	93.2
800	40	49	11	20	45	29	41	48	0	14	59.4	29.6	89.2
900	44	49	10	19	47	33	44	46	2	13	61.4	30.8	92.0

Continued on next page

Table 38: Raw per-checkpoint success counts and percentages: Strong First Only Low LR

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	43	50	14	19	46	30	41	47	3	7	60.0	29.2	90.8
1500	44	49	5	23	45	25	41	46	0	8	57.2	24.4	90.0
2000	43	49	4	25	45	26	37	49	2	5	57.0	24.8	89.2
2500	43	49	1	25	46	29	32	47	1	3	55.2	23.6	86.8
3000	43	49	0	24	46	28	38	47	2	2	55.8	22.4	89.2
3500	42	50	2	20	46	25	40	47	3	2	55.4	20.8	90.0
4000	46	48	3	11	46	24	39	50	0	1	53.6	15.6	91.6
4500	43	48	2	12	43	24	38	47	3	0	52.0	16.4	87.6

Table 39: Raw per-checkpoint success counts and percentages: Fisher 1e10

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	42	49	8	19	46	17	36	47	1	14	55.8	23.6	88.0
1000	44	50	7	18	45	24	39	42	2	16	57.4	26.8	88.0
1500	44	50	8	14	47	29	40	46	0	10	57.6	24.4	90.8
2000	41	49	6	17	43	33	46	45	0	3	56.6	23.6	89.6
2500	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
3000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
3500	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
4000	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0
4500	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0

Table 40: Raw per-checkpoint success counts and percentages: Fisher 1e12

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	42	49	19	17	45	28	35	43	0	12	58.0	30.4	85.6
1000	41	48	11	20	46	26	43	47	0	10	58.4	26.8	90.0
1500	42	49	19	18	46	21	38	47	0	19	59.8	30.8	88.8
2000	47	49	15	17	47	20	39	47	0	14	59.0	26.4	91.6
2500	42	49	15	20	45	22	39	44	1	13	58.0	28.4	87.6
3000	45	50	19	12	47	25	42	45	0	12	59.4	27.2	91.6
3500	43	49	17	17	45	22	36	47	0	12	57.6	27.2	88.0
4000	43	47	17	21	45	24	36	45	0	16	58.8	31.2	86.4
4500	44	48	17	17	47	31	42	45	0	14	61.0	31.6	90.4
5000	46	49	13	14	46	23	40	47	0	17	59.0	26.8	91.2
5500	44	49	14	23	46	21	42	50	0	9	59.6	26.8	92.4
6000	43	49	17	19	43	26	38	45	0	12	58.4	29.6	87.2
6500	43	49	16	20	46	26	40	46	0	17	60.6	31.6	89.6
7000	41	49	19	17	43	28	41	45	0	12	59.0	30.4	87.6
7500	39	50	15	24	44	23	39	48	0	13	59.0	30.0	88.0
8000	45	48	17	23	45	29	38	48	0	13	61.2	32.8	89.6
8500	41	50	13	22	46	28	38	47	0	15	60.0	31.2	88.8
9000	45	49	18	17	47	27	39	45	0	9	59.2	28.4	90.0
9500	42	49	16	19	43	25	37	45	0	9	57.0	27.6	86.4
9999	45	48	18	19	46	27	41	47	0	13	60.8	30.8	90.8

Table 41: Raw per-checkpoint success counts and percentages: Fisher 1e15

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	43	49	13	22	47	31	38	47	1	9	60.0	30.4	89.6
1000	44	49	18	17	45	28	46	43	0	11	60.2	29.6	90.8
1500	45	48	21	25	44	30	41	46	0	8	61.6	33.6	89.6
2000	44	49	16	19	46	21	37	46	1	13	58.4	28.0	88.8
2500	38	49	15	20	43	28	43	45	0	15	59.2	31.2	87.2
3000	43	49	18	21	48	24	46	46	0	13	61.6	30.4	92.8
3500	43	49	14	21	45	24	44	44	0	16	60.0	30.0	90.0
4000	45	49	17	18	47	25	38	46	0	12	59.4	28.8	90.0
4500	42	49	13	14	45	24	41	46	0	12	57.2	25.2	89.2
5000	47	49	13	15	48	25	44	46	0	10	59.4	25.2	93.6
5500	42	49	14	18	45	31	40	47	0	12	59.6	30.0	89.2

Continued on next page

Table 41: Raw per-checkpoint success counts and percentages: Fisher 1e15

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
6000	45	49	16	13	46	22	36	43	1	13	56.8	26.0	87.6
6500	43	49	12	21	44	26	38	47	0	10	58.0	27.6	88.4
7000	42	49	16	19	47	29	42	46	0	12	60.4	30.4	90.4
7500	45	47	12	22	46	24	42	47	0	17	60.4	30.0	90.8
8000	45	49	16	20	47	26	39	46	0	12	60.0	29.6	90.4
8500	45	48	17	20	42	25	38	46	0	12	58.6	29.6	87.6
9000	42	49	15	22	44	26	39	46	0	14	59.4	30.8	88.0
9999	42	49	16	22	46	25	43	49	0	14	61.2	30.8	91.6

Table 42: Raw per-checkpoint success counts and percentages: Fisher 1e10 Strong First Libero 10 Mix

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	45	48	11	17	46	23	40	46	1	13	58.0	26.0	90.0
1000	43	49	10	14	45	27	42	43	0	14	57.4	26.0	88.8
1500	44	49	12	16	48	26	42	46	0	15	59.6	27.6	91.6
2000	43	48	9	20	43	25	41	47	0	16	58.4	28.0	88.8
2500	42	49	8	21	43	30	45	44	1	7	58.0	26.8	89.2
3000	44	50	8	16	44	31	42	49	0	8	58.4	25.2	91.6
3500	45	49	7	18	45	29	45	45	0	6	57.8	24.0	91.6
4000	41	50	7	12	44	23	46	45	1	3	54.4	18.4	90.4
4500	42	50	11	10	43	27	36	45	0	0	52.8	19.2	86.4
5000	45	48	4	1	43	14	43	42	0	0	48.0	7.6	88.4
5500	46	48	2	3	42	13	43	46	0	0	48.6	7.2	90.0
6000	43	49	5	0	46	13	40	49	0	0	49.0	7.2	90.8
6500	44	49	6	5	45	14	44	48	0	0	51.0	10.0	92.0
7000	44	48	13	5	43	14	41	47	0	0	51.0	12.8	89.2
7500	46	48	9	4	43	18	40	46	0	1	51.0	12.8	89.2
8000	44	46	8	5	44	17	44	46	0	0	50.8	12.0	89.6
8500	46	49	10	2	44	25	40	46	0	1	52.6	15.2	90.0
9000	46	47	5	2	45	20	38	45	0	2	50.0	11.6	88.4

Table 43: Raw per-checkpoint success counts and percentages: Strong First Libero 10 Mix

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	45	49	21	14	46	31	35	48	1	24	62.8	36.4	89.2
1000	43	47	22	12	42	29	40	48	0	23	61.2	34.4	88.0
1500	41	48	20	21	44	28	44	46	0	15	61.4	33.6	89.2
2000	45	49	16	20	44	29	43	49	0	19	62.8	33.6	92.0
2500	44	47	16	27	44	31	41	43	2	19	62.8	38.0	87.6
3000	45	49	25	29	46	34	42	48	1	17	67.2	42.4	92.0
3500	45	46	21	19	43	25	34	45	0	26	60.8	36.4	85.2
4000	42	47	24	21	45	24	34	46	0	13	59.2	32.8	85.6
4500	47	47	30	22	42	22	32	46	1	23	62.4	39.2	85.6
5000	42	46	22	28	39	33	36	47	0	19	62.4	40.8	84.0
5500	42	46	26	17	40	22	39	48	0	21	60.2	34.4	86.0
6000	43	49	18	27	44	31	38	43	0	18	62.2	37.6	86.8
6500	46	48	22	25	44	27	38	47	1	18	63.2	37.2	89.2
7000	44	50	19	21	40	21	40	49	2	17	60.6	32.0	89.2
7500	47	47	29	29	41	34	43	47	0	26	68.6	47.2	90.0
8000	48	50	20	31	45	27	43	48	3	25	68.0	42.4	93.6
8500	46	46	17	14	44	24	42	46	2	21	60.4	31.2	89.6
9000	44	48	22	25	50	34	38	46	1	21	65.8	41.2	90.4
9500	46	48	18	20	46	27	41	45	1	10	60.4	30.4	90.4
9999	45	47	17	19	45	23	30	45	0	22	58.6	32.4	84.8

Table 44: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	44	46	18	18	41	23	37	47	0	16	58.0	30.0	86.0

Continued on next page

Table 44: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
1000	44	39	24	31	43	33	35	48	0	19	63.2	42.8	83.6
1500	47	32	15	18	41	24	42	43	1	13	55.2	28.4	82.0
2000	47	24	19	22	31	32	38	43	1	17	54.8	36.4	73.2
2500	46	20	20	30	10	34	35	37	2	16	50.0	40.8	59.2
3000	42	29	17	26	10	18	28	39	0	15	44.8	30.4	59.2
3500	48	28	14	28	8	27	30	31	2	16	46.4	34.8	58.0
4000	35	26	16	24	3	36	19	37	3	18	43.4	38.8	48.0
4500	42	22	20	30	3	22	27	39	0	24	45.8	38.4	53.2
5000	40	21	14	18	1	27	19	23	4	14	36.2	30.8	41.6
5500	37	18	34	23	2	26	15	33	1	17	41.2	40.4	42.0
6000	24	22	31	24	0	24	18	22	0	24	37.8	41.2	34.4
6500	22	18	32	23	0	28	9	24	1	21	35.6	42.0	29.2
7000	16	15	24	21	0	23	6	14	1	22	28.4	36.4	20.4
7500	16	13	22	22	0	29	7	14	1	18	28.4	36.8	20.0
8000	14	20	32	19	0	27	4	18	1	19	30.8	39.2	22.4
8500	16	13	29	28	0	28	3	13	0	20	30.0	42.0	18.0
9000	14	16	24	25	0	26	6	14	0	20	29.0	38.0	20.0

Table 45: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix Fisher 1e8

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	48	49	18	18	47	27	42	46	0	17	62.4	32.0	92.8
1000	47	49	1	12	46	26	38	43	0	2	52.8	16.4	89.2
1500	17	15	0	0	6	1	1	3	0	0	8.6	0.4	16.8
2000	9	5	0	0	4	0	0	0	0	0	3.6	0.0	7.2
2500	8	2	0	0	3	0	1	0	0	0	2.8	0.0	5.6
3000	5	1	0	0	3	0	1	0	0	0	2.0	0.0	4.0
3500	12	2	0	0	1	0	0	0	0	0	3.0	0.0	6.0
4000	20	5	0	0	1	5	3	0	0	0	6.8	2.0	11.6
4500	23	9	0	0	6	11	1	0	0	0	10.0	4.4	15.6
5000	8	1	0	0	1	1	0	0	0	0	2.2	0.4	4.0
5500	14	0	0	1	0	2	0	0	0	0	3.4	1.2	5.6
6000	17	3	0	0	1	5	0	0	0	0	5.2	2.0	8.4
6500	5	1	0	1	0	2	0	0	0	0	1.8	1.2	2.4
7000	11	0	0	0	0	5	0	0	0	0	3.2	2.0	4.4
7500	30	14	0	1	0	9	0	0	0	0	10.8	4.0	17.6
8000	15	3	0	1	0	12	0	0	0	0	6.2	5.2	7.2

Table 46: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix Fisher 1e10

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	46	49	9	19	44	26	40	45	0	12	58.0	26.4	89.6
1000	40	49	15	15	46	24	39	41	1	12	56.4	26.8	86.0
1500	45	49	12	16	47	22	38	37	0	14	56.0	25.6	86.4
2000	47	49	4	16	48	27	44	42	1	9	57.4	22.8	92.0
2500	46	49	5	22	46	33	48	49	0	6	60.8	26.4	95.2
3000	42	48	5	21	44	30	43	45	1	2	56.2	23.6	88.8
3500	39	37	2	5	27	1	38	36	0	0	37.0	3.2	70.8
4000	19	13	0	0	7	2	5	3	0	0	9.8	0.8	18.8
4500	40	24	0	1	28	2	20	12	0	0	25.4	1.2	49.6
5000	42	35	4	5	35	2	22	22	0	0	33.4	4.4	62.4
5500	40	33	2	6	37	8	25	27	0	0	35.6	6.4	64.8
6000	41	31	5	8	32	1	20	34	0	0	34.4	5.6	63.2
6500	40	22	4	5	32	14	23	27	0	0	33.4	9.2	57.6
7000	42	20	4	7	32	9	31	28	0	0	34.6	8.0	61.2
7500	42	25	5	12	30	8	24	28	0	1	35.0	10.4	59.6
8000	40	28	8	14	28	11	23	19	0	1	34.4	13.6	55.2
8500	41	23	14	12	28	13	23	21	0	0	35.0	15.6	54.4
9000	41	15	8	11	32	12	23	10	0	1	30.6	12.8	48.4

Table 47: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix Filtered Fisher 1e9

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	45	47	11	21	43	27	40	48	2	10	58.8	28.4	89.2
1000	42	49	7	22	43	23	47	48	0	8	57.8	24.0	91.6
1500	42	49	14	23	43	25	43	46	1	6	58.4	27.6	89.2
2000	44	46	13	21	42	25	43	44	6	5	57.8	28.0	87.6
2500	44	42	9	27	38	30	37	48	0	5	56.0	28.4	83.6
3000	45	40	4	17	43	25	36	45	3	6	52.8	22.0	83.6
3500	47	39	11	12	42	20	28	42	3	11	51.0	22.8	79.2
4000	40	33	10	8	37	22	17	43	2	9	44.2	20.4	68.0
4500	42	32	10	14	37	20	16	43	3	11	45.6	23.2	68.0

Table 48: Raw per-checkpoint success counts and percentages: Strong First Libero10 Collected Mix Filtered Fisher 1e10

Step	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall	Collected	Retained
500	44	49	4	17	47	30	36	48	0	10	57.0	24.4	89.6
1000	44	48	6	22	45	25	42	45	2	13	58.4	27.2	89.6
1500	45	49	4	17	46	26	40	45	0	10	56.4	22.8	90.0
2000	42	50	5	19	44	26	41	47	2	18	58.8	28.0	89.6
2500	41	50	17	20	46	30	34	46	0	7	58.2	29.6	86.8
3000	40	49	13	10	42	25	40	49	1	1	54.0	20.0	88.0
3500	15	2	0	0	14	0	18	12	0	0	12.2	0.0	24.4
4000	14	20	0	0	13	2	8	7	0	0	12.8	0.8	24.8
4500	31	34	0	1	26	1	14	17	0	0	24.8	0.8	48.8
5000	34	32	0	5	28	2	15	18	0	0	26.8	2.8	50.8
5500	37	26	0	8	35	5	18	22	0	0	30.2	5.2	55.2
6000	42	22	3	7	31	5	16	16	1	0	28.6	6.4	50.8
6500	36	23	4	4	32	11	21	15	0	0	29.2	7.6	50.8
7000	43	30	3	13	31	12	22	18	0	2	34.8	12.0	57.6
7500	39	29	5	4	28	15	20	15	0	1	31.2	10.0	52.4
8000	33	28	3	6	28	8	20	9	1	0	27.2	7.2	47.2
8500	40	32	5	9	27	13	18	17	0	3	32.8	12.0	53.6
9000	40	28	1	6	30	13	15	16	0	1	30.0	8.4	51.6

F Compute Infrastructure

Models were trained on an institutional high-performance computing cluster using NVIDIA H200 GPUs. Most evaluations were also run on the cluster, with additional local testing on an NVIDIA RTX 6000 Ada workstation.

The previously reported estimate of approximately 2,154 H200 GPU-hours predates the Random-All and Random-Success training runs and the additional balanced-checkpoint evaluations. We therefore omit a revised aggregate compute total until the final accounting includes these experiments.

G Normalization-Statistics Ablations and Secondary INSIGHT Comparisons

The primary replay-based RQ1 conditions recompute normalization statistics on the exact replay-plus-new-data mixture used for each training condition. Several new-only, low-learning-rate, and EWC configurations instead reuse the normalization statistics of the original dataset. Consequently, comparisons across these families change both the adaptation procedure and the numerical state/action transform.

The experiments in this section characterize sensitivity to that choice. They do not establish that retaining the original normalization statistics acts as an implicit continual-learning regularizer.

G.1 Strong INSIGHT with Original Normalization

The Strong INSIGHT First and matched START conditions using original normalization statistics are evaluated only at checkpoints 1000, 3000, and 5000. Because they do not cover the complete

common grid, a grid-mean comparison is unavailable. Their historically selected overall maxima are 61.6% and 60.6%, respectively, and are reported only descriptively.

These sparse old-normalization maxima should not be compared directly with the complete new-normalization RQ1 grid means as a controlled estimate of the effect of normalization.

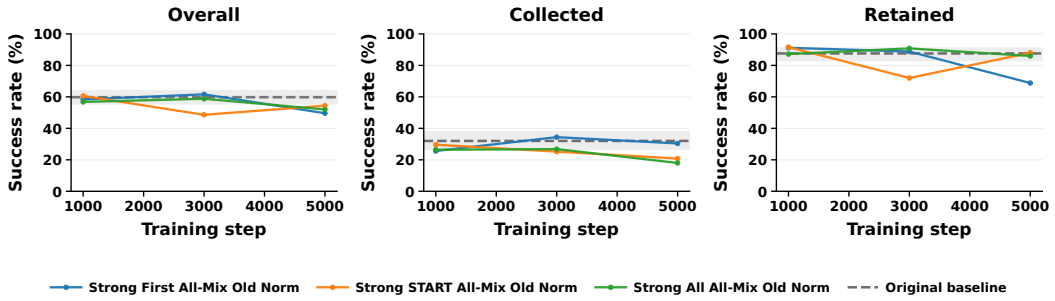


Figure 7: **Strong INSIGHT First and START with original normalization statistics.** Only available checkpoints are shown; missing evaluations are not imputed and no complete-grid summary is reported.

G.2 WEAK INSIGHT

Weak INSIGHT provides a complementary LIBERO-aligned supervision setting. Using original normalization statistics, Weak First has grid means of 60.48% overall, 33.20% collected-task, and 87.76% retained-task success, compared with 56.60%, 23.88%, and 89.32% for its matched START condition.

With recomputed normalization, Weak First has a 66.90% overall grid mean, but the corresponding START trajectory lacks checkpoint 7000, so a complete-grid comparison is withheld. Their complete late-window means are 67.67% and 58.00%, respectively, and are reported descriptively.

Strong and Weak INSIGHT differ in the supervision used to train the help predictor. They should not be interpreted as a controlled comparison of uncertainty thresholds or calibration operating points.

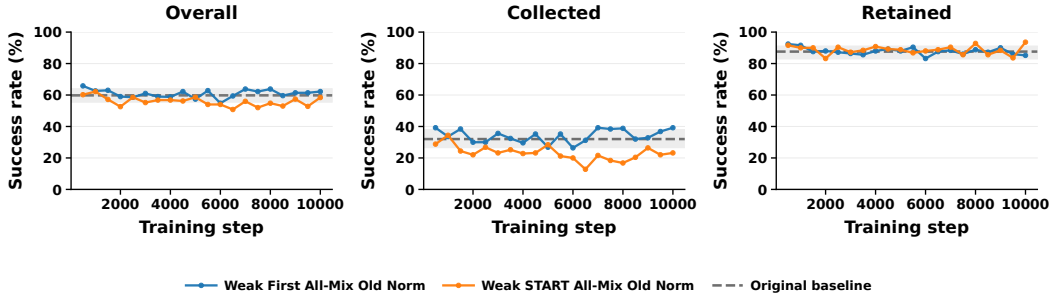


Figure 8: **Weak INSIGHT First and START with original normalization statistics.** Curves show single-run checkpoint evaluations; complete summaries are reported in Appendix E.

H ADDITIONAL EWC ANALYSES

The low-learning-rate EWC sweep uses $\alpha = 2.5 \times 10^{-8}$ across multiple regularization coefficients. We additionally evaluate Fisher estimates computed from filtered reference data.

Across the tested configurations, stronger parameter constraints generally preserve retained-task performance while limiting adaptation on the collected tasks. The detailed trajectories also show sensitivity to coefficient, Fisher reference data, and training horizon. These experiments characterize

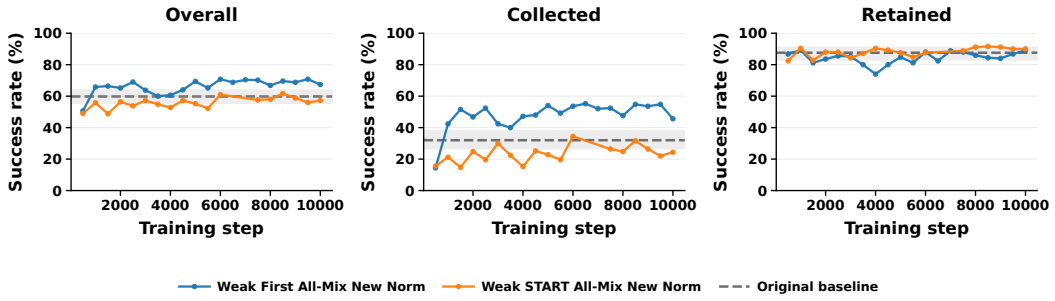


Figure 9: **Weak INSIGHT First and START with recomputed normalization statistics.** The START trajectory lacks checkpoint 7000; the missing evaluation is not imputed.

the stability–plasticity tradeoff observed for the tested configurations; they do not establish that EWC or parameter regularization is ineffective in general.

Comparisons against the primary replay configurations additionally differ in normalization statistics. They should therefore be interpreted as comparisons between complete adaptation configurations rather than clean algorithm-only ablations.

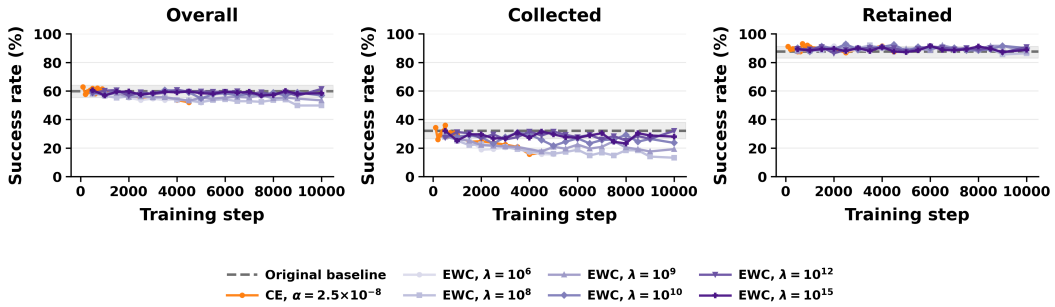


Figure 10: **EWC coefficient sweep with the low learning rate.** Curves show individual training trajectories. Gray bands denote the original policy's Wilson confidence interval rather than uncertainty across independent training seeds.

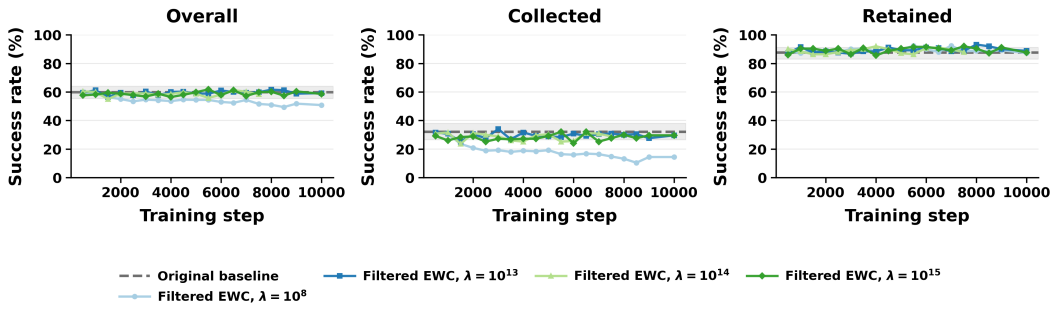


Figure 11: **Filtered-Fisher EWC sweep.** Changing the Fisher reference data alters the observed stability–plasticity trajectory but does not remove the tradeoff in the tested configurations.

I Supplementary Learning Curves

Figure 12 shows the common-grid trajectories underlying RQ1. Figure 13 reports the collected-task replay EWC variants discussed in RQ5. Checkpoints are repeated observations from one training trajectory rather than independent training runs.



Figure 12: **RQ1 state-selection conditions over the common checkpoint grid.** Differences among grid-mean, late-window, final-checkpoint, and peak rankings reflect variation along the observed optimization trajectories.

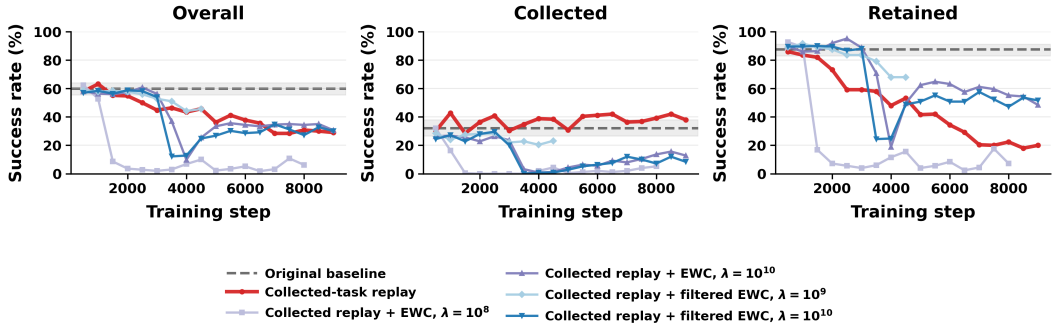


Figure 13: **Collected-task replay with EWC variants.** The trajectories illustrate sensitivity to replay coverage, EWC coefficient, Fisher reference data, and training horizon.

J Physical-Robot Evaluation Details and Prior Sequential Study

The primary physical-robot result in Section 10 comes from the fresh fixed-budget matched evaluation. This experiment was conducted separately from the earlier sequential policy comparison.



Figure 14: **Pre-execution recovery collection.** INSIGHT evaluates the proposed VLA action chunk before execution. When help is requested, the chunk is withheld, the robot holds its current pose, and the human supervisor takes over using the GELLO interface to complete the recovery demonstration.

J.1 Fixed-Budget Matched Evaluation

START-30 and INSIGHT-30 are evaluated on 60 matched scene pairs, balanced across the three physical manipulation tasks with 20 pairs per task. The evaluated checkpoints and sample size were

fixed before collecting these outcomes. No early stopping is used, and policy order is randomized within each pair. The archived evaluation schedule uses random seed 20260923.

Across the 60 matched pairs, both policies succeed in 26 pairs and both fail in 11. INSIGHT-30 succeeds while START-30 fails in 17 pairs, while START-30 succeeds and INSIGHT-30 fails in 6.

Table 49: **Paired outcomes in the fixed-budget physical evaluation.**

Paired outcome	Count
Both succeed	26
Both fail	11
INSIGHT-30 succeeds, START-30 fails	17
START-30 succeeds, INSIGHT-30 fails	6

There are therefore 23 discordant pairs. Under the null hypothesis of equal discordant probabilities,

$$B \sim \mathrm{Binomial}(23, 1/2). \tag{7}$$

The two-sided exact McNemar test gives

$$p = 2\ \Pr(B \leq 6) = 0.0346897. \tag{8}$$

The corresponding one-sided directional value is 0.0173448, but we use the two-sided result as the primary statistic.

Overall, START-30 succeeds on 32/60 trials (53.3%), while INSIGHT-30 succeeds on 43/60 trials (71.7%).

J.2 Earlier Sequential Evaluation

Before the fixed-budget experiment, we conducted a separate sequential policy comparison using STEP (Snyder et al., 2025), a near-optimal stopping procedure for comparing imitation-learning policies.

That study stopped with BASE at 7/19 successes versus INSIGHT-30 at 16/19, and START-30 at 21/39 successes versus INSIGHT-30 at 32/39.

Because the stopping time depends on observed outcomes, these stopped proportions are retained only as historical descriptive measurements. They are not pooled with the fresh 60-pair experiment and are not reused as fixed-sample effect-size estimates. The revised paper's primary physical-robot inference is based on the fixed-budget matched evaluation above.