# CAPABILITY LOSS IN EMBODIED RECURSIVE SELF-IMPROVEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recursive self-improvement (RSI) is emerging as a promising method to improve embodied agents with their own experience. However, we argue that, under a fixed task distribution, RSI can erode tail and error-recovery capabilities while nominal success keeps rising, a failure we call *silent erosion*. We study this failure with a dynamical model over behavior modes and identify two causes of this capability loss: *selection*, which down-weights the modes a policy performs poorly on, and *visitation drift*, which removes the post-failure states in which recovery is exercised. We turn this analysis into RESCUE (REtaining Skills by Covering Underrepresented Experience), which spends a fixed external budget on mastered modes with shrinking data mass, retains them across generations, and forecasts degradation from their coverage. Our experiments separate data recursion from parameter drift, intervene on each mechanism, and replicate across policy architectures, benchmarks, and real-robot deployment. Our code will be publicly available.

## 1 INTRODUCTION

Recursive self-improvement (RSI) is emerging as one promising route to more capable agents. RSI allows an agent to use its current capabilities to improve future versions of itself. Recent progress in LLM-based agents suggests that increasingly capable systems can generate better solutions, feedback, and training experience (Zhang et al., 2026a; Lee et al., 2026). Extending this paradigm to embodied agents is particularly appealing, because embodied agents continually interact with their environments and may turn accumulated experience into progressively stronger control capabilities.

Several embodied systems already attempt to turn their own experience into better policies. An early example is RoboCat, which trained on experience generated by earlier versions of the agent (Bousmalis et al., 2024). More recently, vision–language–action (VLA) models have been improved with their own interaction data through advantage-conditioned training (Physical Intelligence et al., 2025a), reinforcement learning with task-success rewards (Tan et al., 2025), and distillation of data collected by reinforcement-learning specialists (Xiao et al., 2025), and fleet-scale systems now retrain policies on data collected during deployment (Pan et al., 2026; Wang et al., 2026). Although these approaches differ in their learning algorithms, most follow a common loop: the current policy collects experience, a rule based on task outcomes determines how much each trajectory contributes to training, and the next policy is trained on the resulting data. The simplest rule keeps only successful rollouts, i.e., weights each rollout by a binary success indicator; more general rules replace this binary weight with a graded signal derived from task outcomes.

However, iterating this loop does not necessarily lead to sustained improvement. Because the current policy both generates and selects its own training data, behaviors it performs rarely or poorly contribute less data in each generation and gradually weaken, while nominal task success keeps rising and hides the loss. We observe this pattern when self-training the $\pi_{0.5}$ VLA (Physical Intelligence et al., 2025b) on LIBERO (Liu et al., 2023) and evaluating recovery on LIBERO-Recover (Liu et al., 2026b), as shown in Figure 1. The policy becomes more reliable from ordinary initial states but loses its ability to recover from errors. We refer to this phenomenon as *silent erosion*.

To formalize this failure, we model the self-training loop as a dynamical system over behavior modes, in which a mode loses training data in two ways. *Selection* favors modes on which the policy already succeeds, with a strength set by the selection rule, and *visitation drift* reduces how often a mode is reached at all, because an improving policy fails less often. Together they form a feedback



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

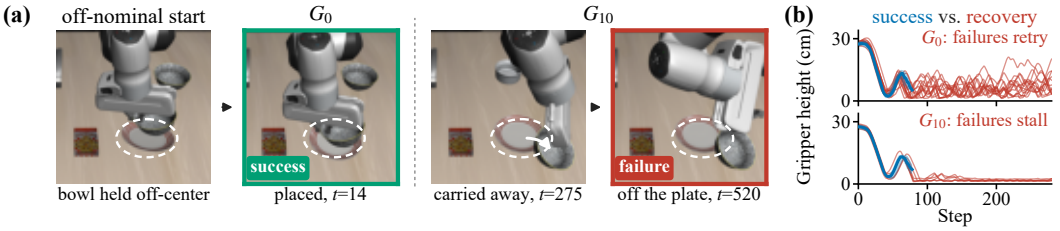



Figure 1: Rollouts of the $\pi_{0.5}$ VLA self-trained on LIBERO-Spatial tasks. (a) From the same off-nominal LIBERO-Recover state (bowl held off-center above the plate), $G_0$ places the bowl within 14 steps, while $G_{10}$ carries it away and exceeds the step limit. (b) Gripper height in 50 episodes per generation: successful episodes nearly coincide, but after a failure $G_0$ retries and $G_{10}$ stalls.

loop: a mode that receives less data becomes harder to perform, and so contributes even less data. Controlled experiments confirm that the erosion is carried by the data rather than the parameters and that restoring the missing modes reverses it. The model further implies that external data help essentially only through the behavior modes they cover, not through their mixing proportion. Building on this analysis, we propose RESCUE (REtaining Skills by Covering Underrepresented Experience), which adds external demonstrations for mastered modes whose coverage is shrinking and retains them across generations. RESCUE preserves recovery capability while keeping the nominal gains. Our main contributions are as follows:

- We identify *silent erosion* in embodied recursive self-improvement: under a fixed task distribution, nominal success keeps rising while tail and error-recovery capabilities decline.
- We model the loop as a dynamical system over behavior modes that decomposes the loss of training data into *selection* and *visitation drift*, and we show how, under shared parameters, less data turns into less competence.
- Building on this analysis, we propose RESCUE, which monitors per-mode coverage, spends a fixed external budget on mastered modes whose data mass is shrinking, retains them across generations, and forecasts degradation from the same coverage trace.
- Experiments show that the erosion is driven by recursively selected data rather than parameter drift, and that RESCUE preserves recovery with a fraction of the external data uniform replay needs while retaining the nominal gain.

## 2 THEORETICAL ANALYSIS

### 2.1 PRELIMINARIES

**The Self-Improvement Loop.** We consider a policy $\pi_g$ that acts on a fixed distribution of tasks. At generation $g$, requests $x \sim P_{\text{req}}$ specify a task instruction and an initial condition; $P_{\text{req}}$ never changes, so any change in the training data originates from the policy itself. A teacher $T_g$ executes a batch of requests, and a selection rule assigns each rollout $\tau$ a retention weight $\omega(\tau) \in [0, 1]$ computed from its outcome. The success filter, $\omega(\tau) = \mathbf{1}[\tau \text{ succeeds}]$, is the most selective rule; advantage-conditioned training (Physical Intelligence et al., 2025a) and on-policy reinforcement learning (Tan et al., 2025) are softer rules that favor the same rollouts. The retained rollouts form $\mathcal{D}^{\text{self}}_{g+1}$, which is mixed with external demonstrations $\mathcal{D}^{\text{ext}}$ at ratio $\alpha$ under a fixed episode budget.

**Definition 1** (Embodied recursive self-improvement)**.** *Given a seed $\pi_0$, generation $g$ collects rollouts with $T_g$ under $P_{\text{req}}$, retains them with weights $\omega$, forms $\mathcal{D}_{g+1} = (1-\alpha)\,\mathcal{D}^{\text{self}}_{g+1} + \alpha\,\mathcal{D}^{\text{ext}}$, and trains $\pi_{g+1}$ on $\mathcal{D}_{g+1}$, initialized from $\pi_g$ (continuing) or from $\pi_0$ (reset). The recursive setting sets $T_g = \pi_g$; the fixed-teacher control sets $T_g = \pi_0$.*

**Capability Measurement.** To test whether a capability survives the loop, we measure it directly on three evaluation suites. The nominal suite $\mathcal{E}_{\text{nom}}$ samples ordinary initial states from $P_{\text{req}}$. The tail suite $\mathcal{E}_{\text{tail}}$ samples conditions from the low-density region of $P_{\text{req}}$, restricted to those on which the seed succeeds. The recovery suite $\mathcal{E}_{\text{rec}}$ starts episodes from post-failure states of LIBERO-Recover, whose levels $L_1$–$L_4$ range from retrying a failed action to restoring an environmental state that blocks the task (Liu et al., 2026b). We denote the success rates on the three suites by $\text{SR}_{\text{nom}}$, $\text{SR}_{\text{tail}}$,

and $\mathrm{SR}_{\mathrm{rec}}$. Two frozen banks of recovery states that the seed solves separate losing an acquired capability from never acquiring it: a bank state *flips* when a later generation fails it; methods may read only the calibration bank $\mathcal{H}_{\mathrm{cal}}$, and the sealed bank $\mathcal{H}_{\mathrm{eval}}$ serves for reporting.

### 2.2 Data Mass over Behavior Modes

Proofs, auxiliary results, and remarks for this section are in Appendix C, and Table 17 there collects the notation. We use $\Phi$ to represent the mapping from each trajectory $\tau$ to a behavior mode $z = \Phi(\tau) \in \mathcal{Z}$. For self-improving generation $g$, under teacher $T_g$ and requests $P_{\mathrm{req}}$, the *visitation* of a mode $z$ is $v_g(z) = P(\Phi(\tau) = z)$, its *selection weight* is $s_g(z) = \mathbb{E}[\omega(\tau) \mid \Phi(\tau) = z]$, and its *retained mass* is

$$m_g(z) = \frac{v_g(z)\, s_g(z)}{\bar{s}_g}, \qquad \bar{s}_g = \textstyle\sum_{z'} v_g(z')\, s_g(z'), \tag{1}$$

where $\bar{s}_g$ is the average retention weight. For a set of modes $\mathcal{S} \subseteq \mathcal{Z}$, the coverage of $\mathcal{S}$ in the next training set is $\mathrm{cov}_{g+1}(\mathcal{S}) = (1-\alpha)\, m_g(\mathcal{S}) + \alpha\, P_{\mathcal{D}^{\mathrm{ext}}}(\mathcal{S})$, and the mode-level coverage $\mathrm{cov}_g^{\mathrm{mode}}(\mathcal{S})$ is the fraction of modes in $\mathcal{S}$ with nonzero mass.

**Proposition 1** (Selection reweighting). *Under the success filter, $s_g(z) = \mathrm{SR}_g(z)$, the per-mode success rate, and $\bar{s}_g = \overline{\mathrm{SR}}_g$; hence $m_g(\mathcal{S})/v_g(\mathcal{S}) = \mathrm{SR}_g(\mathcal{S})/\overline{\mathrm{SR}}_g$ for every set of modes with $v_g(\mathcal{S}) > 0$. If every rollout is retained with weight $\omega \in [\lambda, 1]$, $\lambda > 0$, then $s_g(z)/\bar{s}_g \in [\lambda, 1/\lambda]$; with no selection, $m_g = v_g$.*

Equation 1 separates the two mechanisms. *Selection* is the factor $s_g(z)/\bar{s}_g$, through which modes the policy performs worse than average lose mass. *Visitation drift* is the movement of $v_g$ itself, which does not depend on the rule. To understand the dynamics of such self-improving systems, we study a simplified environment in which the recovery region consists of a single mode.

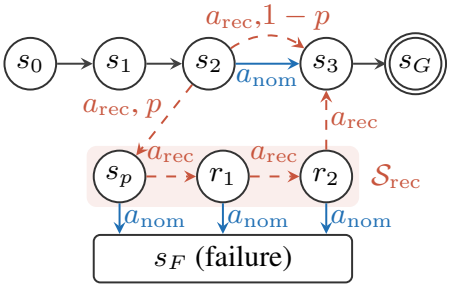



Figure 2: The nominal–recovery MDP.

**Definition 2** (Nominal–recovery MDP). *The MDP in Figure 2 has two actions, $a_{\mathrm{nom}}$ and $a_{\mathrm{rec}}$, and a goal state $s_G$. Episodes follow the nominal path $s_0 \to s_1 \to s_2 \to s_3 \to s_G$. At $s_2$ a perturbation occurs with probability $p$; if it does, $a_{\mathrm{nom}}$ stabilizes it and continues to $s_3$, whereas $a_{\mathrm{rec}}$ is displaced to the perturbed state $s_p$; otherwise both actions continue to $s_3$. In the recovery region $\mathcal{S}_{\mathrm{rec}} = \{s_p, r_1, r_2\}$, $a_{\mathrm{rec}}$ advances along the recovery path $s_p \to r_1 \to r_2 \to s_3$, whereas $a_{\mathrm{nom}}$ falls into the failure sink $s_F$. The shared label is only a device for the minimal conflict: the parameter direction that improves the nominal decision opposes the one that supports recovery. We define the nominal error $\epsilon_n = \pi(a_{\mathrm{rec}} \mid s_2)$, the nominal competence $u = 1 - \epsilon_n$, and the per-decision recovery competence $q = \pi(a_{\mathrm{rec}} \mid s)$ for every $s \in \mathcal{S}_{\mathrm{rec}}$. The success probabilities are $J_{\mathrm{nom}}(\pi) = \Pr_\pi(\text{reach } s_G \mid s_0)$ and $J_{\mathrm{rec}}(\pi) = \Pr_\pi(\text{reach } s_G \mid s_p)$.*

**Proposition 2** (Occupancy identity). *$J_{\mathrm{rec}}(\pi) = q^3$, $J_{\mathrm{nom}}(\pi) = 1 - p\,\epsilon_n(1 - q^3)$ (which equals $\overline{\mathrm{SR}}$ under the success filter), and the per-episode occupancies (visitation probabilities) $d_\pi$ are*

$$d_\pi(s_p) = p\,\epsilon_n, \qquad d_\pi(\mathcal{S}_{\mathrm{rec}}) = p\,\epsilon_n\,(1 + q + q^2). \tag{2}$$

The probability of entering the recovery region, $d_\pi(s_p)$, is thus proportional to the nominal error, so nominal improvement reduces recovery visitation under any selection rule. Visitation alone produces no forgetting, since a tabular learner with no data on $s_p$ leaves $q$ unchanged. Under the success filter the two mechanisms compound. For each collected rollout, unfiltered data contain $p\epsilon_n(1 + q + q^2)$ recovery decisions, labeled $a_{\mathrm{rec}}$ at rate $q$, whereas filtered data keep only the $3p\epsilon_n q^3$ decisions of successful recovery episodes, all labeled $a_{\mathrm{rec}}$. The filter cleans the label but keeps only the fraction $h(q) = \frac{3q^3}{1+q+q^2}$ of recovery decisions, which increases from $h(0) = 0$ to $h(1) = 1$. Whether fewer but cleaner labels become lower competence depends on the learner.

### 2.3 From Less Data to Less Competence

Forgetting requires that nominal and recovery decisions share parameters. We present the mechanism with the simplest such learner, a logistic policy $\pi_\theta(a_{\mathrm{rec}} \mid s) = \sigma(\theta^\top \phi(s))$ with features

$\phi_n = \phi(s_2)$ and $\phi_r = \phi(s_p) = \phi(r_i)$, overlap $c = \langle \phi_n, \phi_r \rangle / (\|\phi_n\| \|\phi_r\|)$, and norm ratio $\rho = \|\phi_n\| / \|\phi_r\|$, so that $\epsilon_n = \sigma(\ell_n)$ and $q = \sigma(\ell_r)$. A generation takes one expected gradient step of size $\eta$ (further steps act, to first order in $\eta$, as a larger $\eta$) on $\mathcal{L}(\theta) = w_n \, \mathrm{CE}(\epsilon_n, y_n) + w_r \, \mathrm{CE}(q, y_r)$, where $(w_n, w_r)$ are the data weights on the two decisions and $(y_n, y_r)$ the rates at which the data label $a_{\mathrm{rec}}$. For the filtered and external sources below, the two loss gradients, $(\epsilon_n - y_n)\phi_n$ with $\epsilon_n \geq y_n$ and $(q - y_r)\phi_r$ with $q \leq y_r$, have cosine $-c$, so positive $c$ is gradient conflict, and the recovery logit $\ell_r = \theta^\top \phi_r$ moves by (Proposition 3)

$$\Delta \ell_r = -\eta \big[ w_n (\epsilon_n - y_n) \, c \, \|\phi_n\| \|\phi_r\| + w_r (q - y_r) \, \|\phi_r\|^2 \big], \tag{3}$$

an interference term from nominal data plus a supervision term from recovery data. The data source fixes $(w_n, w_r, y_n, y_r)$ (Lemma 1); with $w_n = 1$ per retained self-generated episode: (a) unfiltered self-data give $y_n = \epsilon_n$ and $y_r = q$, so self-cloning is a fixed point in expectation; (b) success-filtered self-data give $y_r = 1$ and $w_r = \kappa \, m(\mathcal{S}_{\mathrm{rec}})$, with $\kappa = 3$ recovery decisions per retained recovery episode; (c) filtered data from a fixed teacher $\pi_0$ have constant weights and labels; and (d) an external fraction $\alpha$ with recovery share $\gamma$ adds recovery weight $\beta := \alpha\gamma\kappa$ $(y_r = 1)$ and nominal weight $\alpha(1 - \gamma)$ $(y_n = 0)$. Substituting each source into equation 3 gives the behavior of the loop.

**Theorem 1** (Three regimes of the loop)**.** *Let $0 < c < 1$ and index generations by $g$.*

*(i)* Success-filtered recursion *($\alpha = 0$, source (b)). If $h(q_g) < u_g \, c\rho$, then $u_{g+1} > u_g$, $q_{g+1} < q_g$, and the condition holds at $g + 1$. Hence nominal competence rises and recovery competence falls at every subsequent generation, with per-generation drop*

$$\ell_{r,g} - \ell_{r,g+1} = \eta \, \|\phi_r\|^2 \, \frac{p \, \epsilon_{n,g}}{\overline{\mathrm{SR}}_g} \, (1 - q_g)(1 + q_g + q_g^2) \, \big[ u_g \, c\rho - h(q_g) \big]. \tag{4}$$

*If instead $h(q_g) > u_g \, c\rho$, then $q_{g+1} > q_g$; if moreover $h(q_g) > u_g \, \rho / c$, recovery sharpens at every subsequent generation.*

*(ii)* Fixed-teacher data *(source (c)). The loss is a fixed convex function of $\theta$; for $\eta \leq 4/(\|\phi_n\|^2 + w_r^0 \|\phi_r\|^2)$, $\epsilon_{n,g} \to y_n^0$ and $q_g \to 1$, so any recovery loss is transient.*

*(iii)* Mixed data *(source (b) at weight $1 - \alpha$ plus source (d)). With the* effective recovery mass *$w_g^{\mathrm{eff}} := (1 - \alpha) \, \kappa \, m_g(\mathcal{S}_{\mathrm{rec}}) + \beta$, the total recovery weight in the training set,*

$$\Delta \ell_r \geq 0 \iff w_g^{\mathrm{eff}} \geq (1 - \alpha) \, \frac{p \, \epsilon_{n,g}}{\overline{\mathrm{SR}}_g} \, (1 + q_g + q_g^2) \, u_g \, c\rho \; + \; \alpha(1 - \gamma) \, \frac{\epsilon_{n,g} \, c\rho}{1 - q_g}. \tag{5}$$

The loop thus needs an improvement signal (unfiltered self-cloning is a fixed point), shared parameters with positive overlap $c$, and data weights endogenous to the policy. Erosion is then monotone below the threshold $q^\star = h^{-1}(uc\rho)$ (e.g., $q^\star \approx 0.7$ per decision, or $(q^\star)^3 \approx 0.35$ in episode success, for $u = 0.8$ and $c\rho = 0.6$), and its rate carries the factor $p\epsilon_n / \overline{\mathrm{SR}}$, so at fixed $q$ it scales to first order with the nominal error. Under recursion, equation 2 ties recovery supervision to the nominal error, so the drift never changes sign as the policy improves; fixed-teacher data break this coupling (ii), and external recovery data add supervision that does not shrink with $\epsilon_n$ (iii). The first term on the right of equation 5 is of the order of the recovery visitation $d_\pi(\mathcal{S}_{\mathrm{rec}})$, so little external recovery mass is needed; the second shows that nominal external data add interference, most of all for a highly competent skill. In the erosion regime, therefore, no $\alpha$ protects at $\gamma = 0$, and mixtures with equal $\alpha$ but different $\gamma$ can differ by the full capability gap. In regime (i), $u_g$ rises while $q_g$ falls, so only a tier that starts inside $\mathcal{S}_{\mathrm{rec}}$ sees the loss; this is why the erosion is silent (Remark 3).

Other rules close the loop as well. By a known equivalence (Peters & Schaal, 2007; Zelikman et al., 2022; Singh et al., 2024), regime (i) also describes on-policy reinforcement learning with a success reward (Proposition 4), and advantage-conditioned cloning (Physical Intelligence et al., 2025a) should erode between the unfiltered and filtered cases (Remark 1). Applied to VLA policies, the model depends on two proxies, $1 - \mathrm{SR}_{\mathrm{nom}}$ for $\epsilon_n$ and the measured gradient conflict for $c$ (Remark 2), so the experiments test signs, orderings, and scaling, not the location of the threshold.

## 3 Capability-Preserving Self-Improvement

Equation 5 is also a design rule. Since its threshold depends on unmeasured constants, we propose RESCUE (REtaining Skills by Covering Underrepresented Experience), which raises the effective

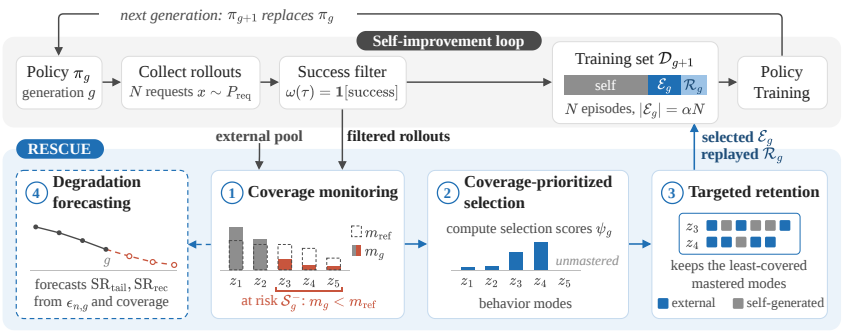



Figure 3: Overview of RESCUE. Top: the self-improvement loop. Bottom: (1) coverage monitoring compares each mode's retained mass $m_g$ with its seed reference $m_{\text{ref}}$ and flags the shrinking set $\mathcal{S}_g^-$; (2) selection scores modes by $\psi_g$ (equation 6) and fills the external budget, skipping modes the seed never mastered; (3) retention keeps the least-covered mastered modes in a buffer and replays them; (4) forecasting predicts $\text{SR}_{\text{tail}}$ and $\text{SR}_{\text{rec}}$ from the nominal error and coverage. Fresh selections $\mathcal{E}_g$ and external-origin replays $\mathcal{R}_g$ together make up $\alpha N$ episodes.

recovery mass under a fixed external budget $\alpha$ by spending the budget where the seed's data mass most exceeds the current one among behaviors the seed had mastered, and by keeping that data from being displaced later. Training remains plain behavior cloning, so any effect is attributable to the data rather than the optimizer. Figure 3 shows the pipeline, and Algorithm 1 in Appendix D gives one recursion chain, including the budget accounting.

**Coverage monitoring.** Each generation, $\Phi$ annotates the rollout pool with modes, plus a recovery tag that records whether a trajectory passes through a failure state and at which LIBERO-Recover level. From the post-selection pool we estimate the retained mass $m_g(z)$ of every mode, $\text{cov}_g^{\text{mode}}$, the recovery visitation $\nu_g$ (share of rollouts entering $\mathcal{S}_{\text{rec}}$), and $w_g^{\text{eff}}$. Two references are fixed at $G_0$: the seed's retained mass $m_{\text{ref}}(z)$ and its success $q_0(z)$ on each mode, the latter measured on $\mathcal{H}_{\text{cal}}$ for recovery modes. The shrinking set $\mathcal{S}_g^-$ collects the modes whose mass has fallen below $m_{\text{ref}}$.

**Coverage-prioritized selection.** External candidate episodes are scored by the mode they realize,

$$\psi_g(z) = \frac{m_{\text{ref}}(z)}{m_g(z) + \delta}\, q_0(z), \tag{6}$$

with a small $\delta > 0$. The density ratio orders modes by how far their mass has fallen below the seed's, and the mastery factor $q_0(z)$ keeps the budget off modes that were never learned. The external budget, net of external-origin replays, is allocated across modes in proportion to $\psi_g$ and filled with distinct trajectories first.

**Targeted retention.** A stratified buffer of fixed capacity keeps episodes of the lowest-coverage mastered modes from the retained self-data and the selected external data. Each replayed episode is debited from the quota of its origin, so every arm trains on exactly $N$ episodes with external-origin fraction $\alpha$; retention changes which external episodes are trained on, not how many. Uniform accumulation of past self-data is no substitute (Appendix D).

**Coverage-based degradation forecasting.** Theorem 1(iii) writes the change of the recovery logit as interference that scales with the nominal error minus supervision that scales with $1 - q$ times the effective recovery mass. The coverage trace therefore also yields a forecast of $\text{SR}_{\text{tail}}$ and $\text{SR}_{\text{rec}}$: a logit hazard with two fitted coefficients per tier, driven by the nominal error $\epsilon_{n,g}$ and by $w_g^{\text{eff}}$ for recovery or the coverage of the shrinking set for the tail (equation 9 in Appendix D). It is a working approximation, judged by whether it forecasts held-out tasks and unseen generations (Section 4.5).

# 4 EXPERIMENTS

We test the analysis of Section 2 through five research questions. **RQ1**: does the loop lower tail and recovery success while nominal success rises, and do distributional metrics detect the loss? **RQ2**: is the cause recursively selected data or parameter drift? **RQ3**: do visitation drift and selection both remove recovery data, and do the rate, sign, and rule dependence of erosion follow Theorem 1(i)? **RQ4**: do external data protect through their recovery mass rather than their proportion (Theorem 1(iii)), and does RESCUE spend a fixed budget better than the alternatives and forecast degradation? **RQ5**: do the findings carry over to other architectures, benchmarks, and real robot?

## 4.1 SETUP

**Experiment setup.** We train $\pi_{0.5}$ (Physical Intelligence et al., 2025b) on LIBERO (Liu et al., 2023), which supplies the nominal and tail tiers, and evaluate recovery on LIBERO-Recover (Liu et al., 2026b). Ten tasks are trained jointly: family A (six tasks from LIBERO-Spatial, -Object, and -Long) supplies all headline, mechanism, and fitting results, and family B (four tasks from LIBERO-Goal and -Long) is never used for fitting or tuning. The seed is fine-tuned from the public checkpoint on ten of the fifty demonstrations per task, so that the nominal tier has headroom, plus thirty teleoperated LIBERO-Recover demonstrations per task. Generations run to $G = 10$ ($G = 20$ for extrapolation); each collects 150 rollouts per task under $P_{\text{req}}$, which places 18% of its mass on at-risk conditions, and every arm trains on exactly $N = 1000$ episodes. The fixed at-risk mode set $\mathcal{A}$ contains the modes the seed realizes with below-average success. External pools are $\mathcal{D}^{\text{ext}}_{\text{plain}}$ (unused LIBERO demonstrations, no recovery content), $\mathcal{D}^{\text{ext}}_{\text{rec}}$ (1000 recovery demonstrations, disjoint from seed data, tiers, and banks), and their union $\mathcal{D}^{\text{ext}}_{\text{mix}}$, whose uniform draw covers $\mathcal{A}$ with probability $\gamma \approx 0.17$.

**Evaluation and statistics.** Each tier runs 50 episodes per task per generation, with $\mathcal{E}_{\text{rec}}$ stratified over levels $L_1$–$L_4$. The banks hold $|\mathcal{H}_{\text{cal}}| = 200$ and $|\mathcal{H}_{\text{eval}}| = 150$ states with re-estimated seed success at least 0.70; a flip is a fall to at most 0.30. The statistical unit is the recursion chain ($n = 5$, seeds paired across arms), and we write an estimate with its 95% interval as $x_{\pm h}$, where $h$ is the interval half-width; Appendix E specifies intervals, equivalence margins, and $c$.

**Arms** (Table 18 in Appendix E). All arms share the seed, the rollouts, $N$, and the update budget, re-initialize optimizer state every generation, and differ only in $T_g$, the initialization, $\omega$, and the composition of $\mathcal{D}_{g+1}$. *Attribution arms* use the success filter at $\alpha = 0$: A (fixed teacher $\pi_0$, reset), B (fixed, continuing), C (recursive teacher $\pi_g$, reset), and D (recursive, continuing; the primary setting), so $D - B$ and $C - A$ measure data recursion and $B - A$ and $D - C$ parameter drift; E (recursive, reset) trains on the union of all past retained self-data. *Mechanism arms* train on arm D's rollout pool and change one factor: no filter (4); patch-back of $0.1N$ successful $G_0$ rollouts of random, lost recovery, or nominal-matched modes (5a–5c); and recovery visitation pinned at its $G_0$ level by rollouts from the seed's $G_0$ failure states, with and without the filter (6, 7). *Loop variants* replace the filter with no selection (U), top-90% truncation (U+T), policy gradient with a success reward (PG), or advantage-conditioned cloning (R). *Protection arms* add external data at fraction $\alpha$ with random, iso-$\beta$ (fixed $\alpha\gamma$), or nominal-only ($\gamma = 0$) composition; uniform replay and RESCUE draw from the same pool $\mathcal{D}^{\text{ext}}_{\text{rec}}$ and differ only in which episodes they select and retain.

Table 1: Attribution arms over generations ($\pi_{0.5}$, family A, mean over 5 chains; s.e.m. $\leq 2.2$; arms defined in Section 4.1). Shading: change from the seed ($G_0$, in headers), red above, blue below. All generations and levels in Table 3.

| | $\text{SR}_{\text{nom}}$ (81.4) | | | $\text{SR}_{\text{tail}}$ (62.0) | | | $\text{SR}_{\text{rec}}$ (44.6) | | |
|---|---|---|---|---|---|---|---|---|---|
| | $G_1$ | $G_5$ | $G_{10}$ | $G_1$ | $G_5$ | $G_{10}$ | $G_1$ | $G_5$ | $G_{10}$ |
| A | 82.3 | 82.5 | 81.5 | 60.8 | 61.2 | 62.2 | 42.9 | 43.5 | 44.9 |
| B | 83.5 | 85.2 | 85.3 | 62.4 | 60.1 | 59.2 | 44.1 | 42.0 | 40.8 |
| C | 82.9 | 86.9 | 88.4 | 58.1 | 48.2 | 43.2 | 38.2 | 24.1 | 16.3 |
| **D** | 83.6 | 88.4 | 90.2 | 59.0 | 46.7 | 37.1 | 40.3 | 22.5 | 11.2 |
| E | 83.8 | 86.8 | 87.6 | 60.2 | 57.0 | 56.4 | 43.4 | 38.9 | 35.7 |

## 4.2 THE SILENT EROSION (RQ1)

Under arm D, nominal success rises while tail and recovery success fall from the first generation on and never recover (Table 1). A fixed teacher removes almost all of this loss: arm A stays within noise of the seed, and arm B loses a little recovery and then plateaus. Theorem 1(ii) allows only a transient loss under a fixed teacher, so we attribute this small sustained remainder to re-optimization

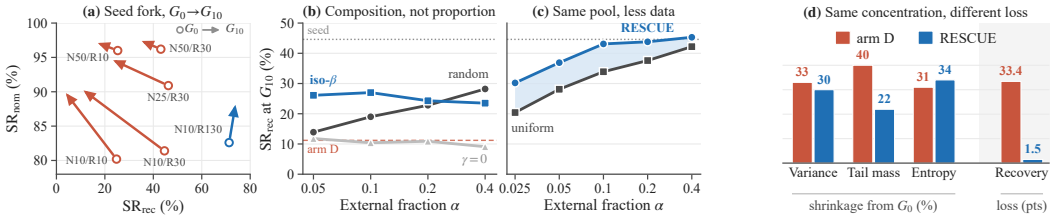



Figure 4: Tests of the model's predictions (family A; mean over 5 chains). (a) Seed fork under arm-D recursion, from $G_0$ to $G_{10}$; N$k$/R$k$ are nominal/recovery demonstrations per task (N10/R30 is the default). Vermilion seeds lose recovery and the blue seed does not (rate and threshold, Theorem 1(i)). (b) $\mathrm{SR}_{\mathrm{rec}}$ at $G_{10}$ against $\alpha$ for random composition, an iso-$\beta$ family ($\alpha\gamma = 0.04$), and nominal-only data ($\gamma = 0$); dashed: arm D, dotted: seed (Theorem 1(iii)). (c) Uniform replay and RESCUE on the same pool $\mathcal{D}_{\mathrm{rec}}^{\mathrm{ext}}$; the band is RESCUE's advantage. (d) Arm D and RESCUE shrink alike on variance and entropy but differ by 31.9 points in $\mathrm{SR}_{\mathrm{rec}}$. Complete results are in Tables 9, 13, 4.

drift. The lost states are ones the seed could solve: arm D flips a majority of the sealed bank $\mathcal{H}_{\mathrm{eval}}$, far above the false-flip floor of arm A (Table 4), which rules out both a capability the seed never had and regression to the mean; the frozen seed and training on external demonstrations alone (Table 11) rule out pipeline drift and repeated fine-tuning. Every recovery level erodes (Table 3d in Appendix A.1). A nominal-only benchmark would therefore certify as improved a loop that has lost three quarters of its recovery success.

Distributional metrics register that the recursive arms concentrate, but not which behaviors they lose (Table 4; Figure 4d). Variance and entropy order the attribution arms roughly as recovery loss does, yet RESCUE concentrates about as much as arm D while keeping recovery near the seed. Across loop conditions, recovery loss tracks the loss of mode-level coverage $\mathrm{cov}^{\mathrm{mode}}(\mathcal{A})$ (Spearman 0.82 [0.63, 0.92]) more closely than variance or entropy shrinkage (0.61 and 0.47; Appendix A.1).

### 4.3 Source of Silent Erosion (RQ2)

The teacher, not the initialization, carries the erosion. In the $2 \times 2$ design of arms A–D, the teacher main effect on $\mathrm{SR}_{\mathrm{rec}}$ at $G_{10}$ (recursive minus fixed) is $-29.1_{\pm 3.9}$ points, against $-4.6_{\pm 2.8}$ for the initialization (continuing minus reset), and their interaction is negligible ($-1.0_{\pm 4.6}$). The teacher contrast is almost as large with reset as with continuing initialization ($C - A$ against $D - B$; Table 1), so the erosion does not require inheriting parameters. Parameter drift is real but secondary: scaling the updates per generation from half to twice the default nearly quadruples arm B's recovery loss yet changes $D - B$ by less than its interval width (Table 5). Dormant-neuron ratio and feature rank stay flat, against a plasticity account (Dohare et al., 2024). Uniform accumulation (arm E) retains much but not all of the capability, as expected from Theorem 1(iii), since the recovery mass of accumulated self-data itself decays (Proposition 2).

### 4.4 Visitation Drift and Selection Mechanism (RQ3)

**Visitation and selection.** Along arm D the request share of at-risk conditions is fixed by design, yet their coverage shrinks in the raw rollouts and shrinks faster after the filter (Table 7). The first effect is visitation drift: as Proposition 2 predicts, the entry probability $\nu_g$ into $\mathcal{S}_{\mathrm{rec}}$ stays proportional to the nominal error, with no trend in $\nu_g/\epsilon_n$ across generations (slope $-0.002_{\pm 0.006}$). The second is selection: retained recovery mass falls faster than visitation, and $m_g(\mathcal{S}_{\mathrm{rec}})/\nu_g$ drops by more than half. The measured overlap $c$ stays positive at every generation, as Theorem 1(i) requires for erosion.

**Rate and threshold.** Theorem 1(i) predicts that the per-generation loss of the recovery logit scales with the nominal error at matched recovery competence, and that its sign is set by recovery competence relative to $h^{-1}(uc\rho)$ at matched nominal competence. A constant proportional hazard also slows within one chain, so we fork the seed by its numbers of nominal and recovery demonstrations and run arm-D recursion from each (Figure 4a; Table 9). At matched recovery competence, the logit loss scales with the nominal error (slope $1.12_{\pm 0.31}$, intercept $-0.018_{\pm 0.045}$), so the relative loss falls sharply with nominal competence instead of staying equal as a constant hazard requires. At matched nominal competence, erosion is ordered by recovery competence, and the recovery-strong

seed N10/R130 does not erode ($+0.024_{\pm 0.034}$ logits per generation); we claim only that it lies at or above the threshold, not that sustained sharpening occurs. The nominally strong, recovery-weak seed N50/R10 erodes slowly: sign follows recovery competence, speed follows nominal error.

**Interventions** (Table 10 in Appendix A.4). Either channel alone restores part of the loss: pinning visitation under the filter (arm 6) raises $\mathrm{SR}_{\mathrm{rec}}$ over arm D by $13.7_{\pm 4.6}$ points, removing the filter (arm 4) restores a comparable amount, and removing it at pinned visitation (arm 7) adds a further $7.8_{\pm 7.2}$. At equal budget from the same source, re-injecting the lost recovery modes (arm 5b) restores most of the loss, whereas random or nominal-matched modes (5a, 5c) restore almost none, so which behaviors are retained matters, not data quality; closed-loop reruns preserve this ordering.

**Selection rule.** The loop variants follow the ordering the model predicts (Table 10, right). Unfiltered self-cloning (U), a fixed point in expectation (Lemma 1), barely drifts, whereas quality truncation (U+T) supplies an improvement signal and already trades recovery for nominal success. Policy gradient with a success reward (PG) erodes like arm D, as Proposition 4 predicts ($\mathrm{PG}-D = 1.8_{\pm 3.2}$; the upper end reaches the 5-point margin, so equivalence is not formally established), and advantage-conditioned cloning (R) lies between PG and U (Remark 1).

### 4.5 Effect of External Data (RQ4)

**Composition, not proportion.** By Theorem 1(iii), protection should barely depend on $\alpha$ at fixed recovery weight $\beta = \alpha\gamma\kappa$, and data without recovery content ($\gamma = 0$) should not protect at any $\alpha$. With random composition, protection rises monotonically with $\alpha$ (Figure 4b), which alone would suggest that proportion matters. Along an iso-$\beta$ family that varies $\alpha$ eightfold at fixed $\alpha\gamma$, however, $\mathrm{SR}_{\mathrm{rec}}$ is nearly flat, and nominal-only external data leave it within noise of arm D at every $\alpha$ tested. Across the recursive arms (Tables 11, 13), $\mathrm{SR}_{\mathrm{rec}}$ tracks the time-averaged $w^{\mathrm{eff}}$ (Spearman 0.84 [0.68, 0.93]) far more closely than $\alpha$ (0.31 [−0.05, 0.60]).

**Data efficiency at a matched pool.** On the same pool $\mathcal{D}_{\mathrm{rec}}^{\mathrm{ext}}$ (Figure 4c; Table 13d–e), RESCUE is ahead of uniform replay at every budget: with half the budget it already exceeds uniform replay at $\alpha = 0.10$, uniform replay needs four times the budget to come within a point of it, and it matches uniform replay with a quarter of the unique recovery demonstrations.

**Baselines** (Table 2). The ranking follows which term of equation 3 a remedy acts on. Parameter-space penalties toward $\pi_0$ (Kirkpatrick et al., 2017; Shenfeld et al., 2026) and weight merging (Yadav et al., 2026) act only on the interference term and slow erosion without stopping it (Table 12). Remedies that add recovery supervision do better even without external data: distillation toward $\pi_0$ on $\mathcal{H}_{\mathrm{cal}}$ states (Li & Hoiem, 2018) retains more, and RESCUE-0 (coverage-balanced filter plus retention, $\alpha = 0$) is on par with uniform replay of external recovery data. Expert corrections on 30% of failures, as in intervention-collecting recipes (Physical Intelligence et al., 2025a), inherit visitation drift: their share falls from 4.7% of episodes at $G_1$ to 3.1% at $G_{10}$. At $\alpha = 0.10$, RESCUE exceeds every baseline with Holm-corrected intervals that exclude zero, keeps $91_{\pm 12}$% of arm D's nominal gain, and is equivalent to arm D in $\mathrm{SR}_{\mathrm{nom}}$ within the 2-point margin.

Table 2: Remedies added to arm D, at $G_{10}$ (family A; 5 chains). Nom./Tail/Rec.: $\mathrm{SR}_{\mathrm{nom}}/\mathrm{SR}_{\mathrm{tail}}/\mathrm{SR}_{\mathrm{rec}}$; $\Delta$: paired $\mathrm{SR}_{\mathrm{rec}}$ difference to RESCUE with Holm-corrected 95% half-width. Complete results in Table 11.

| Condition | Nom. | Tail | Rec. | $\Delta$ |
|---|---|---|---|---|
| Arm D (recursive, primary) | 90.2 | 37.1 | 11.2 | $-31.9_{\pm 3.1}$ |
| *No external data* | | | | |
| Weight merging with $\pi_0$ (RETAIN) | 88.7 | 44.9 | 21.4 | $-21.7_{\pm 2.9}$ |
| Distillation on $\mathcal{H}_{\mathrm{cal}}$ states (LwF) | 88.6 | 50.2 | 29.6 | $-13.5_{\pm 2.9}$ |
| RESCUE-0 (filter + retention) | 89.2 | 55.6 | 33.4 | $-9.7_{\pm 2.6}$ |
| *Expert or external recovery data, $\alpha = 0.10$* | | | | |
| On-failure corrections (RECAP-style) | 89.8 | 50.3 | 29.5 | $-13.6_{\pm 2.9}$ |
| Gated querying (Thrifty-style) | 89.6 | 52.7 | 35.8 | $-7.3_{\pm 2.3}$ |
| Uniform replay of $\mathcal{D}_{\mathrm{rec}}^{\mathrm{ext}}$ | 89.5 | 55.1 | 33.9 | $-9.2_{\pm 2.3}$ |
| KL penalty + uniform replay | 88.3 | 56.4 | 36.1 | $-7.0_{\pm 2.4}$ |
| RESCUE (full) | 89.4 | **59.8** | **43.1** | – |

**Ablations** (Table 11). Averaged over the $2 \times 2$ cells of selection and retention, $\psi$-selection adds 5.2 points of $\mathrm{SR}_{\mathrm{rec}}$ and the buffer, which adds no external data, 4.0. Within $\psi$, the density ratio matters more than the mastery factor $q_0$, and an unsupervised mode dictionary costs a few points.

**Forecasting** (Table 14; Appendix A.6). We forecast one step ahead on held-out family-B tasks, which share chains with family A but are never used in fitting (protocol i), and extrapolate from $G_{10}$ to $G_{11}$–$G_{20}$ (protocol ii). The coverage hazard of equation 9 has the lowest error in three of the four settings, whereas hazards built on $\alpha$ or action entropy do not improve on a generation-only fit.

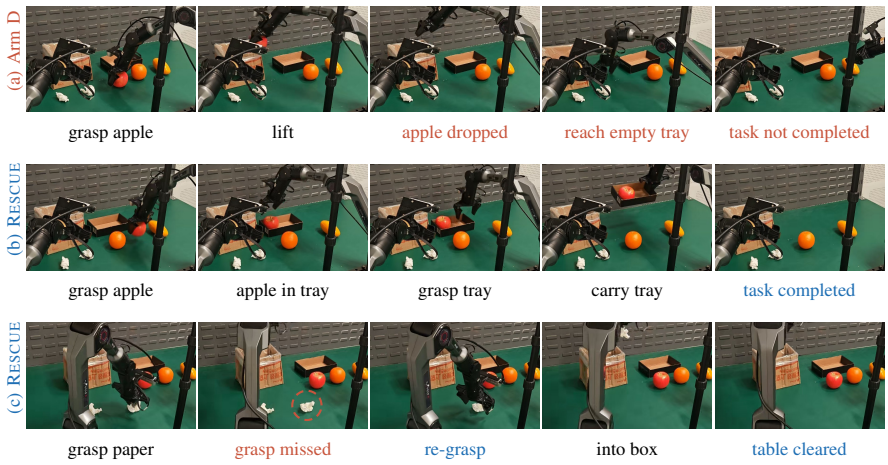



Figure 5: Physical-robot rollouts with $\pi_{0.5}$. (a) After recursive self-improvement (arm D), the policy drops the apple, keeps reaching toward the empty tray without re-grasping, and fails the task. (b) RESCUE places the apple in the tray and carries the tray away. (c) On the paper-clearing task, RESCUE misses its first grasp, later returns to the paper, re-grasps it, and clears the table.

### 4.6 GENERALITY (RQ5)

We evaluate the core arms on the held-out family B, which shares chains with family A, and rerun them with OpenVLA-OFT (Kim et al., 2025), with Diffusion Policy (Chi et al., 2023) trained from scratch, and on four RoboCasa tasks (Nasiriany et al., 2024); settings and full results are in Appendix A.7 (Table 15). In every replication, arm D gains nominal success and loses much recovery success, arm B loses little, and $D - B$ keeps its sign. RESCUE stays within a few points of the seed in four of the five simulated settings, where uniform replay loses clearly more; the advantage of RESCUE excludes zero everywhere except Diffusion Policy (from scratch; largest overlap $c$).

**Real robot.** We also run arm D and RESCUE with $\pi_{0.5}$ on two physical tabletop tasks and report the study qualitatively (Figure 5; Appendix A.7; videos in the supplementary material). After recursive self-improvement, the policy recovers visibly less reliably once a grasp or transport goes wrong, whereas with RESCUE it re-attempts and completes the task.

## 5 RELATED WORK

Recursive self-improvement is DAgger (Ross et al., 2011; Hoque et al., 2022) without the expert, so states visited less often as the policy improves receive fewer labels, and what is sharpened (Huang et al., 2025) is also what is visited. Unlike model collapse (Shumailov et al., 2024; Dohmatob et al., 2025; Zhang, 2026; Gerstgrasser et al., 2024; Schaeffer et al., 2025) and performative prediction (Perdomo et al., 2020; Taori & Hashimoto, 2023), our object is task capability on fixed behavior modes, and the induced distribution is a state occupancy. Forgetting is studied under exogenous task sequences (Wang et al., 2024; Liu et al., 2026a; Zhu et al., 2026) and addressed with penalties (Kirkpatrick et al., 2017; Shenfeld et al., 2026) or merging (Yadav et al., 2026), which we include as baselines; our interference model follows linear analyses of forgetting through feature overlap (Bennani & Sugiyama, 2020; Doan et al., 2021; Evron et al., 2022), with task weights set by the policy's own success. Appendix B gives a full discussion.

## 6 CONCLUSION

Embodied recursive self-improvement can erode rarely exercised capabilities while nominal success rises. We characterized this silent erosion, attributed it to recursively selected data, explained it through selection and visitation drift under shared parameters, and turned the resulting protection condition into RESCUE; any behavior whose supporting states an improving policy learns to avoid is exposed to the same loop.

## AI USE STATEMENT

We used large language models only for language refinement (e.g., grammar, style, and phrasing) and to identify potentially relevant literature. All suggested edits and related papers were independently reviewed and verified by the authors. LLMs were not used for technical ideation, modeling, result selection, or drawing conclusions, and did not affect the substantive content of the work. We take responsibility for the final content of this work.

## REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

Mehdi Abbana Bennani and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020. URL https://openreview.net/forum?id=mzgcEGD-yk4.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JORAfH2xFd.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauzá Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Fernandes Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott E. Reed, Sergio Gómez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Thomas Rothörl, José Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin A. Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. Robocat: A self-improving generalist agent for robotic manipulation. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=vsCpILiWHu.

Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Xiang Li, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, Tiejun Huang, Yu Wang, and Chao Yu. $\pi_{\text{RL}}$: Online rl fine-tuning for flow-based vision-language-action models, 2026. URL https://arxiv.org/abs/2510.25889.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=O4cHTxW9BS.

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

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/doan21a.html.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 08 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07711-7. URL https://doi.org/10.1038/s41586-024-07711-7.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=et5l9qPUhm.

Maximilian Du, Zhanyi Sun, Chen Xu, Paarth Shah, Masha Itkina, and Shuran Song. Memory anchors for continual robot learning, 2026. URL https://arxiv.org/abs/2608.26545.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/evron22a.html.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong HSU, Calvin Luo, Yonglong Tian, and Chen Sun. Self-correcting self-consuming loops for generative model training. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=i0nVanexij.

Prannay Hebbar, Yogendra Manawat, Samuel Verboomen, Alesia Ivanova, Selvam Palanimalai, Kunal Bhatia, and Vignesh Baskaran. Sia: Self improving ai with harness & weight updates, 2026. URL https://arxiv.org/abs/2605.27276.

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 Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 598–608. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/hoque22a.html.

Jiaheng Hu, Jay Shim, Chen Tang, Yoonchang Sung, Bo Liu, Peter Stone, and Roberto Martin-Martin. Simple recipe works: Vision-language-action models are natural continual learners with reinforcement learning, 2026. URL https://arxiv.org/abs/2603.11653.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WJaUkwci9o.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems (RSS)*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 143–156. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/laskey17a.html.

Hyunin Lee, Jinglue Xu, Jeffrey Seely, Donghyun Lee, Matei Zaharia, and Yujin Tang. Recursive harness self-improvement, 2026. URL https://arxiv.org/abs/2607.15524.

Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Inf. Fusion*, 58(C):52–68, June 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.004. URL https://doi.org/10.1016/j.inffus.2019.12.004.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, December 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Huihan Liu, Changyeon Kim, Bo Liu, Minghuan Liu, and Yuke Zhu. Pretrained vision-language-action models are surprisingly resistant to forgetting in continual learning. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=VzdSHEab4G.

Lin Liu, Zhicheng Bao, Lu Zhang, Ziying Song, Wu Yang, Yuzheng Zhuang, Shuai Tao, Wulong Liu, Caiyan Jia, and Huchuan Lu. Libero-recover: Beyond task success towards failure recovery in robotic manipulation models, 2026b. URL https://arxiv.org/abs/2609.05178.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

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

Mingjie Pan, Siyuan Feng, Qinglin Zhang, Xinchen Li, Jianheng Song, Chendi Qu, Yi Wang, Chuankang Li, Ziyu Xiong, Zhi Chen, Yi Liu, and Jianlan Luo. Sop: A scalable online post-training system for vision-language-action models, 2026. URL https://arxiv.org/abs/2601.03044.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 745–750, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith,

Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025a. URL https://arxiv.org/abs/2511.14759.

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

Ralf Römer, Yi Zhang, Yuming Li, and Angela P. Schoellig. CLARE: Continual learning for vision-language-action models via autonomous adapter routing and expansion. In *ICRA 2026 Workshop: From Data to Decisions: VLA Pipelines for Real Robots*, 2026. URL https://openreview.net/forum?id=4qsn6fQ536.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think, 2025. URL https://arxiv.org/abs/2503.03150.

Shuai Shao, Qihan Ren, Dongrui Liu, Chen Qian, Boyi Wei, Dadi Guo, Yang JingYi, Xinhao Song, Linfeng Zhang, Weinan Zhang, and Jing Shao. Your agent may misevolve: Emergent risks in self-evolving LLM agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fd1jgQQW28.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7HNRYT4V44.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759, 2024. doi: 10.1038/S41586-024-07566-Y. URL https://doi.org/10.1038/s41586-024-07566-y.

David Silver and Richard S. Sutton. Welcome to the era of experience. In G. Konidaris (ed.), *Designing an Intelligence*. MIT Press, 2025.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models, 2025. URL https://arxiv.org/abs/2505.17016.

Rohan Taori and Tatsunori B. Hashimoto. Data feedback loops: model-driven amplification of dataset biases. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329. URL https://doi.org/10.1109/TPAMI.2024.3367329.

Yi Wang, Xinchen Li, Pengwei Xie, Pu Yang, Buqing Nie, Yunuo Cai, Qinglin Zhang, Chendi Qu, Jeffrey Wu, Jianheng Song, Xinlin Ren, Jingshun Huang, Mingjie Pan, Siyuan Feng, Zhi Chen, and Jianlan Luo. Learning while deploying: Fleet-scale reinforcement learning for generalist robot policies, 2026. URL https://arxiv.org/abs/2605.00416.

Maciej Wolczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=5qsptDcsdEj.

Yuxuan Wu, Guangming Wang, Zhiheng Yang, Tianchen Deng, Maoqing Yao, Brian Sheil, and Hesheng Wang. Continually evolving skill knowledge in vision language action model, 2026. URL https://arxiv.org/abs/2511.18085.

Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Yuqi Xie, Fengyuan Hu, Jimmy Wu, Zhengyi Luo, Linxi "Jim" Fan, Guanya Shi, and Yuke Zhu. Self-improving vision-language-action models with data generation via residual rl, 2025. URL https://arxiv.org/abs/2511.00091.

Yajat Yadav, Zhiyuan Zhou, Andrew Wagenmaker, Karl Pertsch, and Sergey Levine. Robust fine-tuning of vision-language-action robot policies via parameter merging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uWJwQ5SZoM.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0NphYCmgua.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.

Qixin Zeng, Shuo Zhang, Hongyin Zhang, Renjie Wang, Han Zhao, Libang Zhao, Runze Li, Donglin Wang, and Chao Huang. Crl-vla: Continual vision-language-action learning, 2026. URL https://arxiv.org/abs/2602.03445.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Tjarko Lange, and Jeff Clune. Darwin gödel machine: Open-ended evolution of self-improving agents. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=pUpzQZTvGY.

Qi Zhang, Shaopeng Zhai, Shengzhe Zhang, Litao Liu, Tianyi Zhang, Fuxian Huang, and Ming Zhou. A generalist pair-wise progress critic model for vision-language-action robots. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=i7mfaYYLDf.

Zhipeng Zhang. Silent collapse in recursive learning systems, 2026. URL https://arxiv.org/abs/2605.14588.

Jiarun Zhu, Yijun Hong, Xiaoquan Sun, Zetian Xu, Qijun He, Haijier Chen, Zhiyong Wang, Mingqi Yuan, Wenjun Zeng, and Jiayu Chen. Can vision-language-action models learn from real-world data continually without forgetting?, 2026. URL https://arxiv.org/abs/2605.26820.

# Capability Loss in Embodied Recursive Self-Improvement

## Appendix

## Appendix Contents

**A Additional Experiment Results** **16**
A.1 Recovery Levels, Concentration, Flips, and Banks . . . . . 16
A.2 Full Coverage Series . . . . . 17
A.3 The Seed Fork . . . . . 18
A.4 Full Mechanism and Baseline Tables . . . . . 18
A.5 Parameter-Space Sweeps and Mixing . . . . . 20
A.6 Forecast Validation . . . . . 20
A.7 Replications . . . . . 20
A.8 Per-Chain Results and Plasticity . . . . . 21

**B Extended Related Work** **22**

**C Theory: Auxiliary Results and Proofs** **24**
C.1 Auxiliary Results and Remarks . . . . . 24
C.2 Proof of Proposition 1 . . . . . 26
C.3 Proof of Proposition 2 . . . . . 26
C.4 Proof of Proposition 3 . . . . . 27
C.5 Proof of Lemma 1 . . . . . 27
C.6 Proof of Theorem 1 . . . . . 27
C.7 Proof of Proposition 4 . . . . . 28

**D RESCUE: Algorithm and Forecasting Details** **28**

**E Protocol Details** **29**

## A Additional Experiment Results

### A.1 Recovery Levels, Concentration, Flips, and Banks

By level (Table 3d), the relative loss grows from 63% at $L_1$ to 89% at $L_4$, but in logit units it is nearly uniform (1.9 to 2.3), as the factor $(1 - q^3)[uc\rho - h(q)]$ of equation 4 implies, so we do not claim that harder levels erode faster.

Across the five attribution arms of Table 4, action-chunk variance orders the arms as recovery loss does and nominal-mode entropy does so up to a swap of C and D, so these metrics register that the recursive arms concentrate. They cannot tell harmful concentration from benign sharpening: RESCUE sharpens about as much as arm D while keeping its recovery, and the coverage-balanced chain concentrates as much as arm C on both metrics while losing half as much recovery. Across

Table 3: Generational trajectories of the attribution arms at all reported generations, of which Table 1 shows a subset ($\pi_{0.5}$, family A, mean over 5 chains; s.e.m. $\leq 2.2$). A: fixed teacher, reset; B: fixed, continuing; C: recursive, reset; **D**: recursive, continuing (primary); E: union of past self-data, reset. All arms start from the seed ($G_0$ in panel headers). In (a)–(c), cells are shaded by the change from the seed: red above, blue below, deeper for larger gaps (square-root scale, saturating at $\pm 30$ points). (d) $\mathrm{SR}_{\mathrm{rec}}$ at $G_{10}$ by LIBERO-Recover level for the seed, B, D, and RESCUE ($\alpha$=0.10); the last row is the level mean ($= \mathrm{SR}_{\mathrm{rec}}$) and the last column arm D's loss in logits.

| | (a) $\mathrm{SR}_{\mathrm{nom}}$, $G_0$=81.4 | | | | (b) $\mathrm{SR}_{\mathrm{tail}}$, $G_0$=62.0 | | | | (c) $\mathrm{SR}_{\mathrm{rec}}$, $G_0$=44.6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $G_1$ | $G_3$ | $G_5$ | $G_{10}$ | $G_1$ | $G_3$ | $G_5$ | $G_{10}$ | $G_1$ | $G_3$ | $G_5$ | $G_{10}$ |
| A | 82.3 | 80.6 | 82.5 | 81.5 | 60.8 | 63.1 | 61.2 | 62.2 | 42.9 | 45.8 | 43.5 | 44.9 |
| B | 83.5 | 83.1 | 85.2 | 85.3 | 62.4 | 59.3 | 60.1 | 59.2 | 44.1 | 41.2 | 42.0 | 40.8 |
| C | 82.9 | 86.6 | 86.9 | 88.4 | 58.1 | 54.6 | 48.2 | 43.2 | 38.2 | 32.9 | 24.1 | 16.3 |
| **D** | 83.6 | 87.9 | 88.4 | 90.2 | 59.0 | 50.8 | 46.7 | 37.1 | 40.3 | 27.4 | 22.5 | 11.2 |
| E | 83.8 | 84.9 | 86.8 | 87.6 | 60.2 | 59.4 | 57.0 | 56.4 | 43.4 | 39.8 | 38.9 | 35.7 |

| | (d) $\mathrm{SR}_{\mathrm{rec}}$ by LIBERO-Recover level | | | | |
|---|---|---|---|---|---|
| | Seed | B | D | RESCUE | D, $\Delta$logit |
| $L_1$ | 68.0 | 63.0 | 25.0 | 66.0 | 1.85 |
| $L_2$ | 56.5 | 52.5 | 12.0 | 55.5 | 2.25 |
| $L_3$ | 38.0 | 33.5 | 6.0 | 36.5 | 2.26 |
| $L_4$ | 16.0 | 14.2 | 1.8 | 14.4 | 2.34 |
| Mean | 44.6 | 40.8 | **11.2** | **43.1** | – |

Table 4: Concentration versus capability at $G_{10}$ (family A; mean over 5 chains), plotted in Figure 4d: action-chunk variance, behavior mass on the bottom 5% of request density, and nominal-mode action entropy as percent change from $G_0$; the change in $\mathrm{SR}_{\mathrm{rec}}$ in points; and the flip rate on the sealed bank $\mathcal{H}_{\mathrm{eval}}$ in percent. Lower block: unfiltered self-cloning (U), the closed-loop coverage-balanced filter (2c, Table 10), and RESCUE at $\alpha = 0.10$.

| Arm | Variance | Tail mass | Entropy | $\Delta\mathrm{SR}_{\mathrm{rec}}$ | Flips |
|---|---|---|---|---|---|
| A | −2 | −3 | −1 | +0.3 | 3 |
| B | −8 | −6 | −6 | −3.8 | 7 |
| C | −27 | −36 | −34 | −28.3 | 52 |
| D | −33 | −40 | −31 | −33.4 | 61 |
| E | −24 | −39 | −19 | −8.9 | 19 |
| U | −6 | −9 | −4 | −2.6 | 6 |
| 2c | −29 | −14 | −33 | −13.7 | 24 |
| RESCUE | −30 | −22 | −34 | **−1.5** | **5** |

Table 5: Sensitivity of the attribution to the optimization budget per generation (family A; mean over 5 chains). The budget scales the number of gradient updates relative to the default of 5,000 (Table 20); data and all other settings are unchanged. $\Delta\mathrm{SR}_{\mathrm{rec}}$ is the change from $G_0$ to $G_{10}$ in points; the last column is the paired contrast with its 95% $t$-interval half-width.

| Updates | $\mathrm{SR}_{\mathrm{nom}}$ | | $\Delta\mathrm{SR}_{\mathrm{rec}}$ | | $D - B$ |
|---|---|---|---|---|---|
| | B | D | B | D | |
| ×0.5 (2,500) | 84.1 | 89.1 | −2.1 | −32.4 | $-30.3_{\pm 6.5}$ |
| ×1 (5,000) | 85.3 | 90.2 | −3.8 | −33.4 | $-29.6_{\pm 5.8}$ |
| ×2 (10,000) | 85.9 | 90.6 | −7.9 | −35.0 | $-27.1_{\pm 6.2}$ |

all loop conditions of Tables 1, 11, and 13 at $G_{10}$, the Spearman correlation of recovery loss is 0.61 [0.31, 0.80] with variance shrinkage, 0.47 [0.12, 0.72] with entropy shrinkage, and 0.66 [0.38, 0.83] with tail-mass loss, against 0.82 [0.63, 0.92] with the mode-level coverage $\mathrm{cov}^{\mathrm{mode}}$ of the at-risk set. The intervals of tail mass and $\mathrm{cov}^{\mathrm{mode}}$ overlap, so the ranking of these two is not settled; the separation that matters is that a protected and an unprotected arm show the same concentration and different recovery. In Table 5, moving from half to twice the default update budget nearly quadruples arm B's recovery loss (from 2.1 to 7.9 points), consistent with re-optimization drift that grows with the amount of optimization per generation, while $D - B$ changes by less than its interval width.

Panel (b) illustrates why the held-out re-estimate matters: selecting bank states on a noisy $G_0$ estimate produces a steep gradient of flips across success bands by regression to the mean, and the re-estimate flattens it. Only the flattened gradient may be read as evidence that erosion proceeds from the boundary of the seed's competence inward.

### A.2 FULL COVERAGE SERIES

The request share is fixed by design; raw coverage sits below it from $G_1$ because failed and detour-avoiding episodes leave the at-risk mode during execution (visitation drift); the widening gap between raw and filtered coverage is the selection channel of Proposition 1; and mode-level coverage cannot drop below 100% until a mode is entirely absent from a filtered pool, which first happens at $G_2$.

Table 6: Recovery banks (mean over 5 chains with the s.e.m. as subscript; 20 rollouts per state). A flip is a re-estimated $G_0$ success of at least 0.70 falling to at most 0.30. Panel (a): cumulative flip rate on the calibration bank $\mathcal{H}_{\text{cal}}$ (200 states). Panel (b): arm-D flips at $G_{10}$ stratified by the seed's success band, using the naive selection estimate and the independent re-estimate; only the re-estimate is reported in the text. Panel (c): sealed evaluation bank $\mathcal{H}_{\text{eval}}$ (150 states), whose $G_{10}$ column is the flip column of Table 4. The frozen seed measures evaluation noise alone; arm A adds retraining noise and is the false-flip floor used in the text.

(a) Cumulative flip rate on $\mathcal{H}_{\text{cal}}$ (%)

| Arm | $G_1$ | $G_3$ | $G_5$ | $G_{10}$ |
|---|---|---|---|---|
| D (recursive, main) | $6.5_{\pm 1.4}$ | $24.0_{\pm 2.6}$ | $37.5_{\pm 3.1}$ | $58.5_{\pm 3.3}$ |
| B (fixed teacher, continuing) | $1.5_{\pm 0.8}$ | $3.5_{\pm 1.2}$ | $5.5_{\pm 1.5}$ | $7.5_{\pm 1.8}$ |
| A (fixed teacher, reset) | $1.0_{\pm 0.6}$ | $2.0_{\pm 0.9}$ | $2.5_{\pm 1.0}$ | $3.0_{\pm 1.1}$ |
| Frozen seed | $0.5_{\pm 0.4}$ | $1.0_{\pm 0.6}$ | $1.5_{\pm 0.7}$ | $1.5_{\pm 0.7}$ |
| RESCUE ($\alpha = 0.10$) | $1.0_{\pm 0.6}$ | $2.5_{\pm 1.0}$ | $4.0_{\pm 1.3}$ | $6.0_{\pm 1.6}$ |

(b) Arm-D $G_{10}$ flip rate by seed success band (%)

| Seed band | $[0.70, 0.80)$ | $[0.80, 0.90)$ | $[0.90, 1.00]$ |
|---|---|---|---|
| Naive (selection estimate) | 79.3 | 55.8 | 33.6 |
| Independent re-estimate (reported) | 63.9 | 57.2 | 50.1 |

(c) Cumulative flip rate on $\mathcal{H}_{\text{eval}}$ (%)

| Arm | $G_5$ | $G_{10}$ |
|---|---|---|
| D | $38.0_{\pm 3.4}$ | $60.9_{\pm 3.1}$ |
| B | $4.9_{\pm 1.5}$ | $6.8_{\pm 1.6}$ |
| A | $2.1_{\pm 1.0}$ | $2.7_{\pm 1.1}$ |
| Frozen seed | $1.3_{\pm 0.7}$ | $2.0_{\pm 0.9}$ |
| RESCUE | $3.6_{\pm 1.3}$ | $5.2_{\pm 1.5}$ |

Table 7: Coverage trace along arm D (family A; mean over 5 chains): nominal error $\epsilon_n$, recovery visitation $\nu_g$ (share of rollouts entering $\mathcal{S}_{\text{rec}}$), retained recovery mass $m_g(\mathcal{S}_{\text{rec}})$ (share of training episodes through $\mathcal{S}_{\text{rec}}$), raw and post-filter coverage of the at-risk set, mode-level coverage of the fixed at-risk set $\mathcal{A}$ (36 modes, fixed from the seed's per-mode success; Appendix E), and the measured overlap $c$, the negative cosine between nominal and recovery loss gradients (s.e. $\leq 0.04$). All columns except $\nu_g/\epsilon_n$ and $c$ are in percent. The full $G_1$–$G_{10}$ series is in Table 8.

| | $\epsilon_n$ | $\nu_g$ | $\nu_g/\epsilon_n$ | $m_g(\mathcal{S}_{\text{rec}})$ | At-risk, raw | At-risk, filtered | $\text{cov}_g^{\text{mode}}(\mathcal{A})$ | Overlap $c$ |
|---|---|---|---|---|---|---|---|---|
| $G_0$ | 18.6 | 12.1 | 0.65 | 7.4 | 18 | 14 | 100 | 0.31 |
| $G_1$ | 16.4 | 10.1 | 0.62 | 5.6 | 17 | 13 | 100 | 0.27 |
| $G_3$ | 12.1 | 8.3 | 0.69 | 3.9 | 13 | 9 | 83 | 0.35 |
| $G_5$ | 11.6 | 6.9 | 0.59 | 2.4 | 11 | 6 | 64 | 0.29 |
| $G_{10}$ | 9.8 | 6.4 | 0.65 | 1.5 | 6 | 2 | 31 | 0.33 |

### A.3 THE SEED FORK

At matched recovery competence (upper block), relative recovery loss falls with seed nominal competence, as the rate prediction of Theorem 1(i) requires. At matched nominal competence (middle block, with the default seed), it is ordered by seed recovery competence, and the recovery-strong seed does not lose recovery over ten generations, as the threshold prediction requires of a seed at or above the threshold. The two blocks are consistent only because sign and speed have different determinants: N50/R10 has the recovery competence of N10/R10 and therefore also erodes, but at a fifth of the nominal error it loses 32% rather than 84% over ten generations. The tail tier of the recovery-strong seed still declines (63.4 to 55.9); the model makes no prediction for the tail tier, whose at-risk modes are not recovery modes, and we do not read this decline as evidence for or against it. Whether measured seeds behave this way is the sharpest test of Theorem 1: if nominally strong seeds erode fastest regardless of their recovery competence, or if the recovery-strong seed erodes as fast as the default seed, the model is missing something.

In the upper block the per-generation logit loss divided by $\epsilon_n$ is 1.10, 1.33, and 1.03 for N10, N25, and N50. The three ratios are not ordered, which the model does not require at this resolution, since the factor $u\,c\rho$ of equation 4 differs by under 20% across the seeds. Mapping the per-decision threshold $h^{-1}(uc\rho)$ onto episode success requires the number of recovery decisions per episode, which LIBERO-Recover episodes do not fix, so the fork tests the ordering of seeds around the threshold rather than its location, and the sign change of the drift lies between R30 and R130.

### A.4 FULL MECHANISM AND BASELINE TABLES

Table 10 lists every mechanism arm, including the closed-loop reruns, and Table 11 every control, baseline, and ablation; Table 2 in the main text is a subset of the latter. A short-path-biased filter (arm

Table 8: $G_1$–$G_{10}$ coverage of the fixed at-risk set $\mathcal{A}$ under arm D (mean over 5 chains). $\mathrm{cov}^{\mathrm{raw}}$ is the share of raw rollouts realizing an at-risk mode, $\mathrm{cov}^{\mathrm{filt}}$ the share after the success filter, $\mathrm{cov}^{\mathrm{mode}}$ the fraction of at-risk modes with nonzero mass in the training data, and $N_g^{\mathrm{rec}}$ the number of retained trajectories through $\mathcal{S}_{\mathrm{rec}}$ among the 600 family-A training episodes. Panel (b) gives log-linear fits on the pooled chain-level series ($5 \times 10$ points), with the half-width of the 95% interval of the retention factor as subscript; the fitted curves are reported separately from the data. Table 7 reports the companion quantities $\epsilon_n$, $\nu_g$, $m_g(\mathcal{S}_{\mathrm{rec}})$, and $c$ at selected generations. Proposition 1 describes one generation's reweighting and does not imply a constant per-generation retention factor, so the series need not be exactly geometric.

(a) Coverage series

| Gen. | Request (%) | $\mathrm{cov}^{\mathrm{raw}}$ (%) | $\mathrm{cov}^{\mathrm{filt}}$ (%) | $\mathrm{cov}^{\mathrm{mode}}$ (%) | $N_g^{\mathrm{rec}}$ |
|---|---|---|---|---|---|
| $G_1$ | 18.0 | 17.2 | 12.6 | 100.0 | 34 |
| $G_2$ | 18.0 | 15.9 | 10.2 | 91.7 | 28 |
| $G_3$ | 18.0 | 13.4 | 8.7 | 83.3 | 23 |
| $G_4$ | 18.0 | 12.1 | 7.1 | 72.2 | 19 |
| $G_5$ | 18.0 | 10.6 | 5.8 | 63.9 | 14 |
| $G_6$ | 18.0 | 10.1 | 4.6 | 55.6 | 13 |
| $G_7$ | 18.0 | 8.4 | 4.2 | 47.2 | 12 |
| $G_8$ | 18.0 | 8.2 | 3.3 | 44.4 | 11 |
| $G_9$ | 18.0 | 6.9 | 3.0 | 36.1 | 8 |
| $G_{10}$ | 18.0 | 6.3 | 2.4 | 30.6 | 9 |

(b) Geometric fits ($G_1$–$G_{10}$)

| Series | Retention | $R^2$ |
|---|---|---|
| $\mathrm{cov}^{\mathrm{raw}}$ | $0.894_{\pm 0.015}$ | 0.91 |
| $\mathrm{cov}^{\mathrm{filt}}$ | $0.833_{\pm 0.021}$ | 0.93 |
| $\mathrm{cov}^{\mathrm{mode}}$ | $0.877_{\pm 0.011}$ | 0.96 |

Table 9: All three tiers for the seed fork of Figure 4a under identical arm-D recursion (family A; mean over 5 chains). N$k$ is the number of nominal demonstrations per task and R$k$ the number of recovery demonstrations per task (Appendix E). The upper block (rate) matches seed recovery competence and varies nominal error; the middle block (threshold), with the default seed, matches nominal competence and varies recovery competence; N50/R10 is nominally strong and recovery-weak. $\Delta\mathrm{SR}_{\mathrm{rec}}$ is the change in $\mathrm{SR}_{\mathrm{rec}}$ from $G_0$ to $G_{10}$ in points, with the relative loss in parentheses, and $\Delta\ell_{\mathrm{rec}}$ is the mean per-generation change of the $\mathrm{SR}_{\mathrm{rec}}$ logit over $G_0$–$G_5$. Highlighted: the recovery-strong seed, which does not erode.

| | $G_0$ | | | $G_{10}$ | | | | |
|---|---|---|---|---|---|---|---|---|
| Seed | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{tail}}$ | $\mathrm{SR}_{\mathrm{rec}}$ | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{tail}}$ | $\mathrm{SR}_{\mathrm{rec}}$ | $\Delta\mathrm{SR}_{\mathrm{rec}}$ | $\Delta\ell_{\mathrm{rec}}$ |
| N10/R30 (default) | 81.4 | 62.0 | 44.6 | 90.2 | 37.1 | 11.2 | −33.4 (75%) | −0.204 |
| N25/R30 | 90.9 | 71.8 | 46.2 | 94.6 | 55.3 | 23.4 | −22.8 (49%) | −0.121 |
| N50/R30 | 96.2 | 79.5 | 43.1 | 97.4 | 72.8 | 35.2 | −7.9 (18%) | −0.039 |
| N10/R10 | 80.2 | 60.7 | 24.8 | 89.8 | 33.0 | 3.9 | −20.9 (84%) | −0.251 |
| N10/R130 | 82.6 | 63.4 | 71.3 | 88.1 | 55.9 | 73.9 | **+2.6** (gain) | **+0.024** |
| N50/R10 | 96.0 | 78.9 | 25.3 | 97.1 | 70.4 | 17.1 | −8.2 (32%) | −0.052 |

3) is the worst configuration, and the reverse intervention, deleting recovery-bearing segments from arm 2 and substituting equal nominal data (2r), removes more than half of arm 2's recovery success. Open-loop arms are one-step counterfactuals on a shared pool and cannot show the feedback the model predicts, in which retained competence produces more successful recovery rollouts and hence more recovery data. Rerun as independent chains, the patch-back and forced-visitation interventions retain slightly more than their open-loop counterparts (41.2 against 40.5, and 27.8 against 24.9), as this feedback predicts, so the ordering of the interventions is preserved. The coverage-balanced filter retains less in closed loop (30.9 against 33.5), because its own improving policy still reduces the number of recovery rollouts it can rebalance.

Continued training on plain demonstrations alone erodes recovery moderately, since it covers nothing in $\mathcal{S}_{\mathrm{rec}}$ and adds nominal interference, while adding the recovery pool retains it; repeated fine-tuning per se is not the cause. Replaying the seed's own recovery demonstrations at $\alpha = 0.10$ reaches 32.6, below uniform replay of the larger pool. Without external data, RESCUE-0 also exceeds noise injection during collection (Laskey et al., 2017) (23.8) and the coverage-balanced filter alone (30.9). Distinct-first filling contributes little (1.5 points, with an interval that touches zero). Varying the buffer capacity between half and twice its default, and $\delta$ in equation 6 between 0.01 and 0.2, keeps $\mathrm{SR}_{\mathrm{rec}}$ between 41.9 and 43.4.

Table 10: Mechanism arms at $G_{10}$ (family A; mean over 5 chains; $\mathrm{SR}_{\mathrm{rec}}$ s.e.m. $\leq 2.3$). Nom., Tail, and Rec. are $\mathrm{SR}_{\mathrm{nom}}$, $\mathrm{SR}_{\mathrm{tail}}$, and $\mathrm{SR}_{\mathrm{rec}}$; Cov. is $\mathrm{cov}_{10}^{\mathrm{mode}}$, the mode-level coverage of $\mathcal{A}$ in the $G_{10}$ training data (%). Left: open-loop arms, which keep their own student but train every generation on data drawn from arm D's rollout pool, so requests and visitation are shared. Right: closed-loop reruns of the decisive interventions and the loop variants of Table 18, all independent chains that collect their own rollouts. Arm designs are specified in Appendix E. Highlighted: the decisive patch-back intervention; bold: best open-loop values.

| # | Open loop (shared arm-D pool) | Nom. | Tail | Rec. | Cov. |
|---|---|---|---|---|---|
| 1 | Global success filter (= arm D) | 90.2 | 37.1 | 11.2 | 31 |
| 2 | Coverage-balanced filter | 89.3 | 56.2 | 33.5 | 73 |
| 3 | Short-path-biased filter | 90.8 | 32.4 | 8.9 | 23 |
| 4 | Same pool, no filter | 86.5 | 51.0 | 26.8 | 57 |
| 5a | Patch-back, random modes | 89.8 | 38.9 | 14.6 | 35 |
| 5b | Patch-back, lost recovery modes | 89.6 | **57.3** | **40.5** | **86** |
| 5c | Patch-back, nominal-matched | 90.1 | 38.0 | 12.3 | 32 |
| 6 | Forced visitation + filter | 89.7 | 46.5 | 24.9 | 54 |
| 7 | Forced visitation, no filter | 86.8 | 54.6 | 32.7 | 61 |
| 2r | Arm 2, recovery segments removed | 89.5 | 40.2 | 15.4 | 34 |

| # | Independent chains | Nom. | Tail | Rec. | Cov. |
|---|---|---|---|---|---|
| *Closed-loop interventions* | | | | | |
| 2c | Coverage-balanced filter | 89.0 | 54.1 | 30.9 | 66 |
| 5b-c | Patch-back, lost recovery modes | 89.4 | 58.0 | 41.2 | 87 |
| 6c | Forced visitation + filter | 89.5 | 49.8 | 27.8 | 58 |
| *Loop variants (arm D's selection rule replaced)* | | | | | |
| U | Unfiltered self-cloning | 82.3 | 60.1 | 42.0 | 80 |
| U+T | Unfiltered, top-90% truncation | 84.9 | 55.7 | 34.2 | 62 |
| PG | Policy gradient, success reward | 89.5 | 38.6 | 13.0 | 33 |
| R | Advantage-conditioned cloning | 89.7 | 42.8 | 18.9 | 41 |

### A.5 Parameter-Space Sweeps and Mixing

The strongest settings recover between 11.5 and 21.8 points of $\mathrm{SR}_{\mathrm{rec}}$ relative to arm D, but only by giving up between 3.9 and 5.3 points of the nominal gain, which places them outside the 75% nominal-gain criterion; none reaches RESCUE's 43.1 at any strength.

Within the iso-$\beta$ family, $\mathrm{SR}_{\mathrm{rec}}$ falls with the nominal share $\alpha(1-\gamma)$ with the sign equation 8 predicts, but the interval includes zero ($-0.9_{\pm 1.0}$ points per 0.1). The data-efficiency ratio of RESCUE against random mixing from $\mathcal{D}_{\mathrm{mix}}^{\mathrm{ext}}$ is larger than against uniform replay, but that comparison also credits RESCUE with knowing which demonstrations carry recovery content, and we do not use it as evidence for the method. In panel (e), capability tracks $n_{\mathrm{uniq}}$ for RESCUE and uniform replay, while random mixing gains little between 250 and 500 because its draw contains few recovery episodes at either size.

### A.6 Forecast Validation

Table 14 lists the forecast errors discussed in Section 4.5. On the one-step $\mathrm{SR}_{\mathrm{tail}}$ forecast the $\beta$-only hazard, which uses the external recovery fraction $\alpha\gamma$ fixed at mixing time, is slightly better (2.3 against 2.4). In protocol (ii) the gain comes mostly from the partially protected arms, whose losses level off once the nominal error falls far enough for their fixed external recovery mass to satisfy equation 8; a constant hazard extrapolates their early decline and overshoots. On arm D, where no such levelling occurs, the two models differ by less than one point. The entropy hazard, which has the form of equation 9 with nominal-mode entropy in place of $w^{\mathrm{eff}}$, does not improve on the generation-only fit, the forecasting counterpart of Figure 4d.

### A.7 Replications

OpenVLA-OFT is fine-tuned with LoRA as in its public recipe, on the same reduced demonstration budget. Diffusion Policy is trained from scratch with task-identity conditioning, which removes the pretrained representation. On RoboCasa, the recovery tier is built from failure snapshots of seed rollouts with the four-level LIBERO-Recover protocol, and the recovery pool comes from a privileged-state scripted controller. The physical study runs arm D and RESCUE with $\pi_{0.5}$ on two tabletop tasks, placing an apple in a tray and carrying the tray away, and clearing crumpled paper into a box; it is reported qualitatively through Figure 5 and the videos in the supplementary material. Table 15 lists the results of every simulated replication. The measured overlap is positive in every setting, as Theorem 1 requires for erosion; the seeds differ in both nominal error and recovery competence, so the replications test the direction of the effects and not their magnitude or an ordering of erosion by $c$.

Table 11: Controls, baselines, and ablations at $G_{10}$ (family A; mean over 5 chains; s.e.m. in the range 0.7–2.3). Nom., Tail, and Rec. are $\mathrm{SR}_{\mathrm{nom}}$, $\mathrm{SR}_{\mathrm{tail}}$, and $\mathrm{SR}_{\mathrm{rec}}$; $\Delta$ is the paired difference in $\mathrm{SR}_{\mathrm{rec}}$ against full RESCUE, with the half-width of its Holm-corrected 95% interval as subscript. Left: controls (plain: $\mathcal{D}^{\mathrm{ext}}_{\mathrm{plain}}$; rec.: $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$) and remedies that add no external data; RESCUE-0 is the coverage-balanced filter plus retention at $\alpha = 0$. Right: remedies that add expert or external recovery data at $\alpha = 0.10$ (on-failure corrections: scripted expert on 30% of failures; gated querying: novelty-triggered draws from $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$), and RESCUE ablations on the same pool. Parameter-space methods report the strongest of four settings that keeps at least 75% of arm D's nominal gain ($\mathrm{SR}_{\mathrm{nom}} \geq 88.0$); full sweeps are in Table 12.

| Condition | Nom. | Tail | Rec. | $\Delta$ |
|---|---|---|---|---|
| *Controls* | | | | |
| Frozen seed $\pi_0$ | 81.4 | 62.0 | 44.6 | – |
| Static training on $G_0$ pool | 82.6 | 58.9 | 41.0 | – |
| External only, plain | 83.1 | 57.4 | 36.9 | – |
| External only, plain + rec. | 85.0 | 59.0 | 43.8 | – |
| Arm B (fixed teacher, continuing) | 85.3 | 59.2 | 40.8 | – |
| Arm D (recursive, primary) | 90.2 | 37.1 | 11.2 | $-31.9_{\pm 3.1}$ |
| *Arm D plus a remedy, no external data* | | | | |
| Uniform replay of past self-data | 89.9 | 37.8 | 13.1 | $-30.0_{\pm 3.2}$ |
| Quadratic penalty to $\pi_0$ (EWC) | 88.4 | 40.6 | 15.9 | $-27.2_{\pm 2.9}$ |
| KL penalty to $\pi_0$ | 88.1 | 42.3 | 18.7 | $-24.4_{\pm 3.0}$ |
| Weight merging with $\pi_0$ (RETAIN) | 88.7 | 44.9 | 21.4 | $-21.7_{\pm 2.9}$ |
| Noise injection (DART) | 88.9 | 47.9 | 23.8 | $-19.3_{\pm 3.0}$ |
| Distillation on $\mathcal{H}_{\mathrm{cal}}$ states (LwF) | 88.6 | 50.2 | 29.6 | $-13.5_{\pm 2.9}$ |
| Coverage-balanced filter (2c) | 89.0 | 54.1 | 30.9 | $-12.2_{\pm 2.8}$ |
| RESCUE-0 | 89.2 | 55.6 | 33.4 | $-9.7_{\pm 2.6}$ |

| Condition | Nom. | Tail | Rec. | $\Delta$ |
|---|---|---|---|---|
| *Arm D plus expert or external recovery data* | | | | |
| On-failure corrections (RECAP-style) | 89.8 | 50.3 | 29.5 | $-13.6_{\pm 2.9}$ |
| Gated querying (Thrifty-style) | 89.6 | 52.7 | 35.8 | $-7.3_{\pm 2.3}$ |
| Uniform replay of $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$ | 89.5 | 55.1 | 33.9 | $-9.2_{\pm 2.3}$ |
| $G_0$ recovery-demo replay | 89.6 | 53.8 | 32.6 | $-10.5_{\pm 2.4}$ |
| KL penalty + uniform replay | 88.3 | 56.4 | 36.1 | $-7.0_{\pm 2.4}$ |
| RESCUE (full) | 89.4 | **59.8** | **43.1** | – |
| RESCUE *ablations* | | | | |
| Uniform selection + buffer | 89.6 | 57.1 | 38.4 | $-4.7_{\pm 2.0}$ |
| $\psi$-selection, no buffer | 89.6 | 57.0 | 39.6 | $-3.5_{\pm 1.6}$ |
| $\psi$ without mastery factor $q_0$ | 89.5 | 57.8 | 40.2 | $-2.9_{\pm 1.7}$ |
| $\psi$ without density ratio | 89.5 | 56.3 | 38.1 | $-5.0_{\pm 2.0}$ |
| No distinct-first filling | 89.4 | 58.9 | 41.6 | $-1.5_{\pm 1.6}$ |
| Unsupervised mode dictionary | 89.5 | 57.4 | 39.4 | $-3.7_{\pm 1.9}$ |

Table 12: Full strength sweeps of the parameter-space baselines on arm D at $G_{10}$ (family A; mean over 5 chains). Strengths increase from left to right (Table 19); each cell gives $\mathrm{SR}_{\mathrm{nom}}$ / $\mathrm{SR}_{\mathrm{rec}}$. The setting reported in Table 11 is the strongest one with $\mathrm{SR}_{\mathrm{nom}} \geq 88.0$ and is marked with $^*$.

| Method | Strength 1 | Strength 2 | Strength 3 | Strength 4 |
|---|---|---|---|---|
| Quadratic penalty to $\pi_0$ (EWC-style) | 89.6 / 12.4 | 88.9 / 14.1 | 88.4 / 15.9$^*$ | 86.3 / 22.7 |
| KL penalty to $\pi_0$ | 89.7 / 13.0 | 88.9 / 16.2 | 88.1 / 18.7$^*$ | 85.4 / 27.9 |
| Weight merging with $\pi_0$ (RETAIN-style) | 89.9 / 13.8 | 89.2 / 17.6 | 88.7 / 21.4$^*$ | 84.9 / 33.0 |

### A.8 PER-CHAIN RESULTS AND PLASTICITY

Every chain shows a recovery drop of at least 24 points under recursion relative to the fixed-teacher arm, and the paired interval excludes zero by a wide margin. The chains that lose most under recursion (2 and 4) are also those in which arm B retains most, so the paired differences vary more than either arm alone. The forced-visitation contrast, arm 6 minus arm D, is $+13.7$, with 95% interval half-width 2.1 unadjusted and 4.6 Holm-adjusted (Table 10), and the filter-removal contrast at pinned visitation, arm 7 minus arm 6, is $+7.8$ with a Holm-adjusted interval that barely excludes zero.

The plasticity check tests the optimization-side alternative (Dohare et al., 2024): if erosion were loss of plasticity, the dormant-neuron ratio and the effective feature rank would move substantially

Table 13: Mixing boundary on arm D at $G_{10}$ (family A; mean over 5 chains; $\mathrm{SR}_{\mathrm{rec}}$ s.e.m. $\leq 2.2$); $\gamma$ is the fraction of external episodes covering the at-risk set. (a) Proportion sweep with random composition from $\mathcal{D}^{\mathrm{ext}}_{\mathrm{mix}}$ ($\gamma \approx 0.17$). (b) Iso-$\beta$ family at a fixed external recovery fraction $\alpha\gamma = 0.04$, composed from $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$ and $\mathcal{D}^{\mathrm{ext}}_{\mathrm{plain}}$. (c) Nominal-only external data ($\gamma = 0$; $\mathcal{D}^{\mathrm{ext}}_{\mathrm{plain}}$ with at-risk content removed). (d) $\mathrm{SR}_{\mathrm{rec}}$ against $\alpha$ for uniform replay and RESCUE drawing from the same pool $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$. (e) $\mathrm{SR}_{\mathrm{rec}}$ at $\alpha = 0.10$ when the recovery part of the pool is restricted to $n_{\mathrm{uniq}}$ unique demonstrations; random mixing draws from the correspondingly restricted $\mathcal{D}^{\mathrm{ext}}_{\mathrm{mix}}$. Figure 4b–c plots panels (a)–(d).

(a) $\alpha$ sweep, random composition

| $\alpha$ | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{tail}}$ | $\mathrm{SR}_{\mathrm{rec}}$ | $\mathrm{cov}^{\mathrm{mode}}_{10}$ |
|---|---|---|---|---|
| 0.00 | 90.2 | 37.1 | 11.2 | 31 |
| 0.05 | 90.3 | 40.6 | 13.9 | 37 |
| 0.10 | 89.7 | 43.6 | 19.0 | 44 |
| 0.20 | 89.4 | 46.1 | 22.8 | 53 |
| 0.40 | 87.6 | 51.8 | 28.2 | 64 |

(b) Iso-$\beta$, $\alpha\gamma = 0.04$

| $\alpha$ | $\gamma$ | $\alpha(1-\gamma)$ | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{rec}}$ |
|---|---|---|---|---|
| 0.05 | 0.80 | 0.01 | 90.3 | 26.1 |
| 0.10 | 0.40 | 0.06 | 89.6 | 27.0 |
| 0.20 | 0.20 | 0.16 | 89.7 | 24.3 |
| 0.40 | 0.10 | 0.36 | 88.2 | 23.5 |

(c) $\gamma = 0$

| $\alpha$ | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{rec}}$ |
|---|---|---|
| 0.05 | 90.4 | 11.8 |
| 0.10 | 89.6 | 10.4 |
| 0.20 | 89.8 | 10.9 |
| 0.40 | 88.3 | 9.1 |

(d) Matched pool $\mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$, $\mathrm{SR}_{\mathrm{rec}}$

| $\alpha$ | 0.025 | 0.05 | 0.10 | 0.20 | 0.40 |
|---|---|---|---|---|---|
| Uniform replay | 20.4 | 28.1 | 33.9 | 37.6 | 42.2 |
| RESCUE | 30.2 | 36.9 | 43.1 | 43.8 | 45.3 |

(e) Unique recovery demonstrations, $\mathrm{SR}_{\mathrm{rec}}$

| $n_{\mathrm{uniq}}$ | 250 | 500 | 1000 |
|---|---|---|---|
| RESCUE | 34.5 | 38.1 | 43.1 |
| Uniform replay | 25.2 | 31.3 | 33.9 |
| Random mixing | 14.1 | 14.8 | 19.0 |

Table 14: Forecast validation. Entries are mean absolute errors in success-rate points against chain means; for $\mathrm{SR}_{\mathrm{rec}}$ the subscript is the half-width of the 95% bootstrap interval over chains and tasks (the larger side). The evaluation-noise floor is about 1.5 points. All predictors are fitted on family A over $G_1$–$G_{10}$. Protocol (i): one-step-ahead forecasts of generation $g+1$ from quantities measured at $g \leq 9$, on the four held-out family-B tasks, for all arms of Tables 13 and 11. Protocol (ii): multi-step forecasts of $G_{11}$–$G_{20}$ from origin $G_{10}$ on family A for five arms (D, random mixing, uniform recovery replay, coverage-balanced chain, and RESCUE at $\alpha = 0.10$), using no quantity measured after $G_{10}$. For arm D the realized $(\mathrm{SR}_{\mathrm{nom}}, \mathrm{SR}_{\mathrm{tail}}, \mathrm{SR}_{\mathrm{rec}})$ are $(91.4, 30.2, 6.3)$ at $G_{15}$ and $(90.8, 26.0, 4.1)$ at $G_{20}$; the generation-only fit predicts $\mathrm{SR}_{\mathrm{rec}}$ of 5.6 and 2.8, and the coverage hazard 5.9 and 3.2.

| | (i) Held-out family B, one step | | (ii) $G_{11}$–$G_{20}$ from $G_{10}$ | |
|---|---|---|---|---|
| Predictor | $\mathrm{SR}_{\mathrm{tail}}$ | $\mathrm{SR}_{\mathrm{rec}}$ | $\mathrm{SR}_{\mathrm{tail}}$ | $\mathrm{SR}_{\mathrm{rec}}$ |
| Last-value continuation | 2.9 | $3.6_{\pm 0.7}$ | 5.9 | $7.4_{\pm 1.5}$ |
| Generation-only geometric fit (per arm) | 3.2 | $3.4_{\pm 0.7}$ | 4.8 | $5.8_{\pm 1.3}$ |
| $\alpha$-only hazard | 4.6 | $6.9_{\pm 1.2}$ | 7.7 | $10.2_{\pm 1.9}$ |
| $\beta$-only hazard ($\alpha\gamma$ at mixing time) | **2.3** | $3.0_{\pm 0.7}$ | 4.3 | $5.1_{\pm 1.2}$ |
| Entropy hazard (entropy in place of $w^{\mathrm{eff}}$) | 3.4 | $4.1_{\pm 0.8}$ | 5.1 | $6.2_{\pm 1.3}$ |
| Coverage hazard, equation 9 ($\epsilon_n$, $w^{\mathrm{eff}}$) | 2.4 | $\mathbf{2.2}_{\pm 0.6}$ | **3.1** | $\mathbf{3.4}_{\pm 1.0}$ |

across generations. On arm D from $G_0$ to $G_{10}$, the dormant-neuron ratio moves from $5.8\%$ to $7.1\%$ and the effective feature rank from 806 to 791, both within drift.

## B EXTENDED RELATED WORK

**Self-improvement and its theory.** Success-filtered bootstrapping in language models originates with STaR, which retains generated rationales that lead to correct answers and fine-tunes on them (Zelikman et al., 2022), and extends through self-rewarding training (Yuan et al., 2024), self-play fine-tuning (Chen et al., 2024), and iterative critique and refinement (Madaan et al., 2023; Shinn et al., 2023); system-level recursion that rewrites agent code has also been demonstrated (Zhang et al., 2026a; Lee et al., 2026; Hebbar et al., 2026). Theory characterizes the gains of self-training as sharpening what the model already does rather than creating new capability (Huang et al., 2025), and taxonomies of self-evolving agents document goal drift and behavioral degradation (Shao et al., 2026). That success-filtered cloning is a policy-gradient step with a binary reward has been noted since reward-weighted regression (Peters & Schaal, 2007), in STaR (Zelikman et al., 2022), and in the expectation-maximization view of self-training (Singh et al., 2024); Proposition 4 records this known equivalence. The dynamics of retraining on a model-induced distribution are studied as per-

Table 15: Replications of the core arms. (a) Changes from the seed at the last generation $G$, in points; "Unif." is uniform recovery replay at $\alpha = 0.10$; $c$ is the gradient overlap at $G_0$. (b) Paired contrasts at $G$, in points, with the half-width of the 95% $t$-interval over paired chains as subscript (5 chains on LIBERO with $\pi_{0.5}$, 3 for the other settings). Interval widths follow the chain count. The physical-robot study is qualitative and not listed (Figure 5).

(a) Changes from the seed

| | | | Seed | | | $\Delta\mathrm{SR}_{\mathrm{rec}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model / platform | Tasks | Chains, $G$ | $\mathrm{SR}_{\mathrm{nom}}$ | $\mathrm{SR}_{\mathrm{rec}}$ | $\Delta\mathrm{SR}_{\mathrm{nom}}$ (D) | D | B | Unif. | RESCUE | $c$ |
| $\pi_{0.5}$ / LIBERO family A | 6 | 5, 10 | 81.4 | 44.6 | $+8.8$ | $-33.4$ | $-3.8$ | $-10.7$ | $\mathbf{-1.5}$ | 0.31 |
| $\pi_{0.5}$ / LIBERO family B (held out) | 4 | 5, 10 | 78.2 | 41.3 | $+9.6$ | $-31.0$ | $-4.4$ | $-11.9$ | $\mathbf{-2.8}$ | 0.33 |
| OpenVLA-OFT / LIBERO family A | 6 | 3, 10 | 83.5 | 40.8 | $+7.9$ | $-32.6$ | $-5.1$ | $-13.2$ | $\mathbf{-3.9}$ | 0.38 |
| Diffusion Policy / LIBERO family A | 6 | 3, 10 | 58.7 | 27.4 | $+11.3$ | $-21.9$ | $-4.0$ | $-9.8$ | $\mathbf{-6.9}$ | 0.52 |
| $\pi_{0.5}$ / RoboCasa | 4 | 3, 10 | 58.9 | 36.2 | $+10.4$ | $-24.7$ | $-3.2$ | $-9.6$ | $\mathbf{-2.9}$ | 0.29 |

(b) Paired contrasts

| Setting | $D - B$: $\mathrm{SR}_{\mathrm{rec}}$ | $D - B$: $\mathrm{SR}_{\mathrm{tail}}$ | RESCUE $-$ Unif.: $\mathrm{SR}_{\mathrm{rec}}$ |
|---|---|---|---|
| $\pi_{0.5}$ / LIBERO family A | $-29.6_{\pm 5.8}$ | $-22.1_{\pm 4.9}$ | $+9.2_{\pm 2.3}$ |
| $\pi_{0.5}$ / LIBERO family B (held out) | $-26.6_{\pm 6.3}$ | $-19.8_{\pm 5.8}$ | $+9.1_{\pm 2.9}$ |
| OpenVLA-OFT / LIBERO family A | $-27.5_{\pm 9.4}$ | $-20.4_{\pm 9.4}$ | $+9.3_{\pm 5.5}$ |
| Diffusion Policy / LIBERO family A | $-17.9_{\pm 9.7}$ | $-13.1_{\pm 9.8}$ | $+2.9_{\pm 5.5}$ |
| $\pi_{0.5}$ / RoboCasa | $-21.5_{\pm 8.7}$ | $-15.2_{\pm 9.2}$ | $+6.7_{\pm 5.6}$ |

Table 16: Per-chain $\mathrm{SR}_{\mathrm{rec}}$ at $G_{10}$ for the key arms on family A (300 episodes per chain per arm; successes out of 300 in parentheses) and the paired contrast $D - B$. The last row gives the mean with the s.e.m. as subscript; for $D - B$ the subscript is instead the half-width of the 95% $t$-interval over the five paired chain differences, unadjusted for multiplicity, which includes evaluation noise.

| Chain | Arm D | Arm B | RESCUE ($\alpha = 0.10$) | Arm 6 | $D - B$ |
|---|---|---|---|---|---|
| 1 | 12.7 (38) | 36.7 (110) | 41.3 (124) | 28.0 (84) | $-24.0$ |
| 2 | 9.7 (29) | 42.7 (128) | 45.3 (136) | 20.7 (62) | $-33.0$ |
| 3 | 12.0 (36) | 40.3 (121) | 42.3 (127) | 26.7 (80) | $-28.3$ |
| 4 | 9.0 (27) | 44.7 (134) | 44.3 (133) | 22.3 (67) | $-35.7$ |
| 5 | 12.7 (38) | 39.7 (119) | 42.3 (127) | 27.0 (81) | $-27.0$ |
| Mean | $11.2_{\pm 0.8}$ | $40.8_{\pm 1.4}$ | $43.1_{\pm 0.7}$ | $24.9_{\pm 1.4}$ | $-29.6_{\pm 5.8}$ |

formative prediction (Perdomo et al., 2020) and data feedback loops (Taori & Hashimoto, 2023), which analyze stability and bias amplification when the data depend on the deployed model; the embodied loop makes the induced distribution a state occupancy. The broader agenda of learning from streams of experience supplies the programmatic framing (Silver & Sutton, 2025). Our setting differs from all of these in that the data distribution is generated by acting in an environment, so that what is sharpened is also what is visited.

**Embodied self-improvement systems.** In robotics the loop is already practice. RoboCat folds successful rollouts of earlier versions into the training set and retrains (Bousmalis et al., 2024); $\pi_{0.6}$ learns from deployment experience through advantage-conditioned training with per-cycle checkpoint rollback (Physical Intelligence et al., 2025a); reinforcement learning with task-success rewards (Tan et al., 2025) and distillation of data collected by reinforcement-learning specialists (Xiao et al., 2025) improve VLA models with their own interaction data; fleet-scale post-training systems stream deployment data back into shared policies and list continual acquisition without forgetting as an open question (Pan et al., 2026; Wang et al., 2026); and a VLA critic provides a self-assessment precedent (Zhang et al., 2026b). Their safeguards, rollback and ad hoc mixing ratios, are heuristic because no measurement tells the loop what it is losing. We supply the measurement and a mechanism-derived rule for what to mix.

**Covariate shift, interactive imitation, and correction.** DAgger showed that a learner's own state distribution must be labeled to prevent compounding errors (Ross et al., 2011), and budget-aware variants such as ThriftyDAgger query an expert only on states flagged as novel or risky (Hoque et al., 2022). Recursive self-improvement is the DAgger loop with the external corrector removed:

the policy labels its own visited states with its own actions. Self-consuming generative loops for human motion have been rescued by a physics-based corrector that projects generated motion onto feasible motion (Gillman et al., 2024); because a feasibility projector adds no recovery supervision, our analysis predicts that it does not transfer to task-conditioned recovery, whereas a task-semantic corrector that re-exposes failure and recovery states does.

**Model collapse.** Recursive training on generated data erodes distributional tails (Shumailov et al., 2024; Dohmatob et al., 2025; Zhang, 2026). Self-consuming loops trade fidelity against diversity depending on how samples are filtered (Alemohammad et al., 2024), and the stability of iterative retraining depends on the proportion of fresh data (Bertrand et al., 2024). Accumulating rather than replacing data mitigates collapse (Gerstgrasser et al., 2024), but collapse verdicts depend on the chosen definition (Schaeffer et al., 2025), and even a small fraction of synthetic data can stall generalization (Dohmatob et al., 2025). We fix the object of loss at the capability level, on fixed sets of behavior modes, and dual-report distributional metrics; in our account, accumulation is an instance of external coverage that protects only insofar as the accumulated data covers the at-risk modes, and the small-fraction result corresponds to the corner of the protection condition in which external data covers no recovery mode at all.

**Continual learning and forgetting in VLA models.** Catastrophic forgetting is classically studied under an exogenous sequence of tasks (Wang et al., 2024; Liu et al., 2026a; Zhu et al., 2026), with regularization (Kirkpatrick et al., 2017), replay, and parameter isolation as the main families, and a robotics lineage from lifelong robot learning to continual-learning benchmarks for manipulation (Lesort et al., 2020; Wolczyk et al., 2021). A recent wave of methods addresses forgetting in VLA models under the same task-sequence assumption, through adapter routing and expansion (Römer et al., 2026), skill-conditioned experts (Wu et al., 2026), parameter merging (Yadav et al., 2026), and regularized reinforcement learning (Zeng et al., 2026). The empirical evidence is split: pretrained VLA models appear resistant to forgetting under small replay budgets (Liu et al., 2026a) and on-policy reinforcement learning is reported to forget little (Hu et al., 2026), whereas naive real-robot fine-tuning collapses (Zhu et al., 2026); replay-buffer composition is reported to matter strongly (Du et al., 2026). RL's Razor attributes forgetting to the divergence from the base policy (Shenfeld et al., 2026), and loss of plasticity is the optimization-side alternative (Dohare et al., 2024). Theoretical accounts of forgetting in linear and kernel regimes express it through the overlap of task features or of the neural tangent kernel across tasks (Bennani & Sugiyama, 2020; Doan et al., 2021; Evron et al., 2022), and gradient conflict between tasks is the operational measure that multi-task optimizers act on (Yu et al., 2020); our overlap $c$ is this measure, and our contribution is to make the task weights endogenous to the loop. Our attribution design separates the divergence induced by optimization on a fixed data source (arm B) from the divergence induced by recursively selected data (arm D against arm B), and our results reconcile the split evidence: the benign regime is a fixed data source with continuing optimization, and collapse appears once the data themselves are recursively selected. Parameter-space remedies act on the interference term of Proposition 3 and add no recovery supervision.

**Recovery evaluation.** LIBERO-Recover constructs failure scenarios from failures actually produced by embodied policies and organizes them into four recovery levels, and it reports that leading policies with near-saturated LIBERO success recover poorly and that adding recovery demonstrations improves recovery without improving nominal success (Liu et al., 2026b). It is a static cross-section of recovery capability; we add the generational axis and the frozen banks that separate loss of an acquired capability from failure to acquire it.

# C Theory: Auxiliary Results and Proofs

## C.1 Auxiliary Results and Remarks

Propositions 1 and 2 and Theorem 1 are stated in Section 2. This subsection states the auxiliary results used there and the remarks it refers to; the remaining subsections give the proofs.

**Proposition 3** (One-step logit dynamics)**.** *One expected gradient step of size $\eta$ on $\mathcal{L}$ changes the recovery logit by equation 3 and the nominal logit $\ell_n = \theta^\top \phi_n$ by*

$$\Delta \ell_n = -\eta \big[ w_n(\epsilon_n - y_n) \, \|\phi_n\|^2 + w_r(q - y_r) \, c \, \|\phi_n\| \|\phi_r\| \big]. \tag{7}$$

Table 17: Notation of Section 2. Generations are indexed by $g$; $h(\cdot)$ is the selection factor of equation 2.2.

| | |
|---|---|
| $v_g(z),\ s_g(z),\ \bar{s}_g,\ m_g(z)$ | visitation, selection weight, its mean, retained mass of mode $z$ equation 1 |
| $\epsilon_n = \sigma(\ell_n),\ u = 1 - \epsilon_n$ | nominal error and nominal competence at $s_2$ (per decision) |
| $q = \sigma(\ell_r)$ | recovery competence per decision; episode recovery success is $J_{\mathrm{rec}} = q^3$ |
| $\overline{\mathrm{SR}} = 1 - p\epsilon_n(1 - q^3)$ | nominal episode success $J_{\mathrm{nom}}$ |
| $p$ | perturbation probability at $s_2$ |
| $c,\ \rho$ | feature overlap (gradient conflict, Section 2.3) and norm ratio $\|\phi_n\|/\|\phi_r\|$ |
| $w_n, w_r,\ y_n, y_r$ | data weights and $a_{\mathrm{rec}}$-label rates at the nominal and recovery decisions |
| $\alpha,\ \gamma,\ \kappa,\ \beta = \alpha\gamma\kappa$ | external fraction, its recovery share, recovery decisions per demo, external recovery weight |
| $h(q) = 3q^3/(1 + q + q^2)$ | ratio of filtered to unfiltered recovery decisions per collected rollout |
| $w_g^{\mathrm{eff}} = (1 - \alpha)\kappa\, m_g(\mathcal{S}_{\mathrm{rec}}) + \beta$ | effective recovery mass equation 5 |

In equation 3, the first term is interference from nominal data and the second is supervision from recovery data. Which term dominates is set by the data source through $(w_n, w_r, y_n, y_r)$, which the next lemma computes for the four sources used in this paper.

**Lemma 1** (Content of the data sources)**.** *Normalize $w_n = 1$ per retained self-generated episode and write $\overline{\mathrm{SR}} = 1 - p\epsilon_n(1 - q^3)$, the nominal episode success $J_{\mathrm{nom}}$, which under the success filter coincides with the normalizer $\bar{s}_g$ of Proposition 1.*

*(a)* Unfiltered self-data*: $y_n = \epsilon_n$ and $y_r = q$, so both gradients vanish and self-cloning is a fixed point in expectation.*

*(b)* Success-filtered self-data*: $y_r = 1$, $\epsilon_n - y_n = p\,\epsilon_n(1 - \epsilon_n)(1 - q^3)/\overline{\mathrm{SR}}$, and $w_r = 3p\,\epsilon_n q^3/\overline{\mathrm{SR}} = \kappa\, m(\mathcal{S}_{\mathrm{rec}})$, with $\kappa = 3$ recovery decisions per retained recovery episode and $m(\mathcal{S}_{\mathrm{rec}}) = p\epsilon_n q^3/\overline{\mathrm{SR}}$ the retained mass of equation 1.*

*(c)* Filtered data from a fixed teacher $\pi_0 = (\epsilon_n^0, q^0)$*: $y_n = y_n^0$ and $w_r = w_r^0$ are constants and $y_r = 1$.*

*(d)* External data *at fraction $\alpha$, of which a fraction $\gamma$ are recovery demonstrations that start from post-failure states and contain $\kappa$ recovery decisions each ($\kappa = 3$ here) and the rest are nominal demonstrations: recovery weight $\beta := \alpha\gamma\kappa$ with $y_r = 1$, and nominal weight $\alpha(1 - \gamma)$ with $y_n = 0$.*

Source (d) assumes that demonstrations take the correct nominal action; a demonstration error rate $\epsilon_d$ would replace $y_n = 0$ by $y_n = \epsilon_d$ and shrink the interference term accordingly. Substituting each source of Lemma 1 into Proposition 3 gives Theorem 1.

**Proposition 4** (Filtered cloning is a policy-gradient step)**.** *For rollouts drawn from $\pi_\theta$ itself, the expected success-filtered cloning step at $\theta$ equals a policy-gradient step on $J_{\mathrm{nom}}$ with reward $\mathbf{1}[\mathrm{success}]$ and step size $\eta/\overline{\mathrm{SR}}$, because $\mathbb{E}\big[\mathbf{1}[\mathrm{success}] \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\big] = \nabla_\theta J_{\mathrm{nom}}(\theta)$, and subtracting a state-dependent baseline leaves the expectation unchanged. Under the logistic parameterization, regime (i) of Theorem 1 therefore also describes on-policy reinforcement learning with a task-success reward, with the same threshold and the same rate up to the step size.*

*Remark* 1 (Softer rules). Advantage-conditioned cloning keeps every rollout and routes the improvement signal through a conditioning variable (Physical Intelligence et al., 2025a). It supplies the improvement signal that makes the loop move, but it also keeps cloning failed recovery attempts under their own outcome label, so by Proposition 1 the recovery mass it retains lies between the unfiltered and filtered values. The model predicts erosion between the unfiltered fixed point and regime (i), a prediction we test rather than derive.

*Remark* 2 (What the model omits). The model has two representative states, one expected gradient step per generation, and a logistic policy; it omits perception, multi-step credit assignment, finite-sample noise, and optimizer state. In particular, regime (ii) predicts that a fixed teacher eventually raises recovery competence, so any sustained loss in the fixed-teacher arm measures what the model omits, chiefly re-optimization drift. The overlap $c$ is a property of the representation and of the two data sources that must be measured, not assumed. We estimate it as the *negative* of the mean cosine between the loss gradient on a success-filtered nominal mini-batch and the loss gradient on a recovery mini-batch, so that positive $c$ is gradient conflict (Yu et al., 2020); the model predicts erosion only for $c > 0$, and if the conflict were absent or reversed the same loop would leave recovery unchanged or sharpen it. The equivalence of Proposition 4 holds for one step at the current parameters, so its extension to reinforcement learning shares the logistic parameterization and the one-step approximation. In general environments the occupancy identity becomes the empirical relation that recovery visitation scales with the nominal error, the per-decision error $\epsilon_n$ is proxied by the episode-level error $1 - \mathrm{SR}_{\mathrm{nom}}$, and the per-decision competence $q$ corresponds to episode recovery success $q^3$. The general dictionary version replaces $c$ by pairwise overlaps between modes, which we do not pursue.

*Remark* 3 (Why the erosion is silent). In regime (i), $u_g$ increases at every generation while $q_g$ decreases, so the nominal tier certifies improvement throughout, and only a tier that starts episodes inside $\mathcal{S}_{\mathrm{rec}}$ sees the loss. The model says nothing about entropy or variance; our interpretation, tested in Section 4.2, is that such monitors register the same process as sharpening, because the surviving nominal modes concentrate while rare modes vanish.

*Remark* 4 (The fixed-teacher limit). Theorem 1(ii) is a statement about $(\epsilon_n, q)$: when $q_g \to 1$ the recovery logit $\ell_r$ diverges, as it does for any logistic fit to a deterministic label. The proof also shows that $\Delta\ell_r \geq 0$ at every generation in which $(\epsilon_{n,g} - y_n^0)\, c\rho \leq w_r^0(1 - q_g)$.

*Remark* 5 (The band between the two thresholds). In the band $u_g\, c\rho < h(q_g) \leq u_g\, \rho/c$ of Theorem 1(i), both logits move in the direction that raises competence at the current generation, but the threshold $u_g c\rho$ rises with $u_g$, so the theorem guarantees a self-sustaining sharpening regime only above $h^{-1}(u\rho/c)$. The seed fork (Section 4.4) therefore tests whether a recovery-strong seed avoids erosion, not whether it sharpens indefinitely.

**Threshold form of Theorem 1(iii).** Moving the self-generated recovery term $(1-\alpha)\kappa\, m_g(\mathcal{S}_{\mathrm{rec}})$ of equation 5 to the right-hand side gives an equivalent condition on the external recovery weight alone: $\Delta\ell_r \geq 0$ if and only if

$$\beta \;\geq\; \beta_g^\star := (1-\alpha)\,\frac{p\,\epsilon_{n,g}}{\overline{\mathrm{SR}}_g}\,(1 + q_g + q_g^2)\,\big[u_g\, c\rho - h(q_g)\big] \;+\; \alpha(1-\gamma)\,\frac{\epsilon_{n,g}\, c\rho}{1-q_g}. \tag{8}$$

The first term of $\beta_g^\star$ is of the order of the recovery visitation $d_\pi(\mathcal{S}_{\mathrm{rec}})$ and vanishes at the threshold $h(q_g) = u_g c\rho$; the second is the interference added by nominal external data.

### C.2 Proof of Proposition 1

A rollout that realizes mode $z$ is retained with expected weight $s_g(z) = \mathbb{E}[\omega(\tau) \mid \Phi(\tau) = z]$, so the retained distribution over modes is proportional to $v_g(z)\, s_g(z)$, which is equation 1. Under the success filter $\omega(\tau) = \mathbf{1}[\tau \text{ succeeds}]$, hence $s_g(z) = P(\text{success} \mid z) = \mathrm{SR}_g(z)$ and the normalizer is $\overline{\mathrm{SR}}_g = \sum_z v_g(z)\mathrm{SR}_g(z)$. Summing over $z \in \mathcal{S}$, $m_g(\mathcal{S}) = \sum_{z\in\mathcal{S}} v_g(z)\mathrm{SR}_g(z)/\overline{\mathrm{SR}}_g = v_g(\mathcal{S})\,\mathrm{SR}_g(\mathcal{S})/\overline{\mathrm{SR}}_g$, where $\mathrm{SR}_g(\mathcal{S}) = \sum_{z\in\mathcal{S}} v_g(z)\mathrm{SR}_g(z)/v_g(\mathcal{S})$ is by the law of total probability the success rate conditional on $\Phi(\tau) \in \mathcal{S}$. If $\omega \in [\lambda, 1]$ then $s_g(z) \in [\lambda, 1]$ and $\bar{s}_g = \sum_z v_g(z) s_g(z) \in [\lambda, 1]$, so $s_g(z)/\bar{s}_g \in [\lambda, 1/\lambda]$. Without selection $\omega \equiv 1$, so $s_g \equiv 1$ and $m_g = v_g$. □

### C.3 Proof of Proposition 2

From $s_p$ the episode reaches $s_3$, and hence $s_G$, if and only if $a_{\mathrm{rec}}$ is taken at $s_p$, $r_1$, and $r_2$, which has probability $q^3$; the alternative action reaches $s_F$, so $J_{\mathrm{rec}}(\pi) = q^3$. From $s_0$ the episode reaches $s_2$ with probability one. With probability $p$ a perturbation is injected; the episode is then displaced to $s_p$ if and only if $a_{\mathrm{rec}}$ is taken, which has probability $\epsilon_n$, and succeeds from $s_p$ with probability $q^3$; otherwise it continues to $s_3$ and succeeds. Hence $J_{\mathrm{nom}}(\pi) = 1 - p\,\epsilon_n(1-q^3)$. The expected number of visits to $s_p$ per episode is $p\,\epsilon_n$, to $r_1$ is $p\,\epsilon_n q$, and to $r_2$ is $p\,\epsilon_n q^2$, which gives equation 2. □

### C.4 Proof of Proposition 3

For $\mathrm{CE}(\sigma(\ell), y) = -y \log \sigma(\ell) - (1-y)\log(1-\sigma(\ell))$ one has $\partial\,\mathrm{CE}/\partial\ell = \sigma(\ell) - y$. Therefore $\nabla_\theta \mathcal{L} = w_n(\epsilon_n - y_n)\,\phi_n + w_r(q - y_r)\,\phi_r$ and $\Delta\theta = -\eta\nabla_\theta\mathcal{L}$. Projecting onto $\phi_r$ and $\phi_n$ and writing $\langle \phi_n, \phi_r\rangle = c\,\|\phi_n\|\|\phi_r\|$ gives the two displayed expressions. □

### C.5 Proof of Lemma 1

(a) Without selection the label at a state is the policy's own action distribution, so $y_n = \epsilon_n$ and $y_r = q$, and both factors $(\epsilon_n - y_n)$ and $(q - y_r)$ vanish.

(b) Condition on success. The probability of taking $a_{\mathrm{rec}}$ at $s_2$ and succeeding is $\epsilon_n\big[(1-p) + p\,q^3\big] = \epsilon_n\big[1 - p(1-q^3)\big]$, so

$$y_n = P(a_{\mathrm{rec}} \mid s_2, \text{success}) = \frac{\epsilon_n\,[1 - p(1-q^3)]}{\overline{\mathrm{SR}}},$$
$$\epsilon_n - y_n = \frac{\epsilon_n\,[\overline{\mathrm{SR}} - 1 + p(1-q^3)]}{\overline{\mathrm{SR}}} = \frac{p\,\epsilon_n(1-\epsilon_n)(1-q^3)}{\overline{\mathrm{SR}}},$$

using $\overline{\mathrm{SR}} = 1 - p\epsilon_n(1-q^3)$. A successful episode passes through $\mathcal{S}_{\mathrm{rec}}$ with probability $p\,\epsilon_n q^3/\overline{\mathrm{SR}}$ and then contains three recovery decisions, each labeled $a_{\mathrm{rec}}$ (an $a_{\mathrm{nom}}$ at a recovery state leads to $s_F$ and so is absent from successful episodes); with one nominal decision per episode normalized to $w_n = 1$, this gives $w_r = 3p\,\epsilon_n q^3/\overline{\mathrm{SR}}$ and $y_r = 1$. The share of retained episodes that pass through $\mathcal{S}_{\mathrm{rec}}$ is the retained mass $m(\mathcal{S}_{\mathrm{rec}}) = p\,\epsilon_n q^3/\overline{\mathrm{SR}}$ of equation 1 with $\mathcal{S}_{\mathrm{rec}}$ as a single mode, so $w_r = \kappa\, m(\mathcal{S}_{\mathrm{rec}})$ with $\kappa = 3$.

(c) The teacher $\pi_0$ is fixed, so the statistics in (b) evaluated at $(\epsilon_n^0, q^0)$ are constants.

(d) Recovery demonstrations start from post-failure states and consist of $\kappa$ recovery decisions labeled $a_{\mathrm{rec}}$; at fraction $\alpha\gamma$ of the training set they contribute recovery weight $\beta = \alpha\gamma\kappa$ with $y_r = 1$. Nominal demonstrations pass $s_2$ once with the expert action $a_{\mathrm{nom}}$, so at fraction $\alpha(1-\gamma)$ they contribute nominal weight $\alpha(1-\gamma)$ with $y_n = 0$. □

### C.6 Proof of Theorem 1

**(i)** Substituting Lemma 1(b) into Proposition 3,

$$\Delta\ell_r = -\eta\|\phi_r\|^2\,\frac{p\,\epsilon_n}{\overline{\mathrm{SR}}}\Big[(1-\epsilon_n)(1-q^3)\,c\rho - 3q^3(1-q)\Big]$$
$$= -\eta\|\phi_r\|^2\,\frac{p\,\epsilon_n}{\overline{\mathrm{SR}}}\,(1-q)(1+q+q^2)\Big[u\,c\rho - h(q)\Big],$$

where we used $1 - q^3 = (1-q)(1+q+q^2)$ and $h(q) = 3q^3/(1+q+q^2)$. This is negative if and only if $h(q) < u\,c\rho$, and it is equation 4. Similarly,

$$\Delta\ell_n = -\eta\|\phi_n\|^2\,\frac{p\,\epsilon_n}{\overline{\mathrm{SR}}}\Big[u(1-q^3) - 3q^3(1-q)\,\frac{c}{\rho}\Big]$$
$$= -\eta\|\phi_n\|^2\,\frac{p\,\epsilon_n}{\overline{\mathrm{SR}}}\,(1-q)(1+q+q^2)\Big[u - h(q)\,\frac{c}{\rho}\Big],$$

which is negative if and only if $h(q)\,c < u\rho$; this is implied by $h(q) < u\,c\rho$ because $c \le 1$ gives $h(q)\,c < u\,c^2\rho \le u\rho$. Hence $u_{g+1} > u_g$ and $q_{g+1} < q_g$. Since $h$ is increasing, $h(q_{g+1}) < h(q_g) < u_g\,c\rho < u_{g+1}\,c\rho$, so the condition holds at $g+1$ and the monotone behavior propagates by induction. That $h$ increases on $[0,1]$ follows from $\frac{d}{dq}\big[3q^3/(1+q+q^2)\big] = (9q^2+6q^3+3q^4)/(1+q+q^2)^2 > 0$ for $q > 0$.

For the sharpening statement, if $h(q_g) > u_g\,c\rho$ the bracket of $\Delta\ell_r$ is negative, so $\Delta\ell_r > 0$ and $q_{g+1} > q_g$. If moreover $h(q_g) > u_g\,\rho/c$, then $h(q_g)\,c > u_g\rho$ and the bracket of $\Delta\ell_n$ is negative, so $\Delta\ell_n > 0$ and $u_{g+1} < u_g$. Then $h(q_{g+1}) > h(q_g) > u_g\,\rho/c > u_{g+1}\,\rho/c$, so the stronger condition holds at $g+1$, and since $u\rho/c \ge u\,c\rho$ for $c \le 1$ the weaker one does as well; recovery therefore sharpens at every subsequent generation by induction. In the band $u_g\,c\rho < h(q_g) \le u_g\,\rho/c$ both

logits move in the direction that raises competence and the threshold $u_g c\rho$ rises with $u_g$; we make no claim about persistence there.

**(ii)** Substituting Lemma 1(c), $\Delta\ell_r = -\eta\|\phi_r\|^2\big[(\epsilon_n - y_n^0)\,c\rho - w_r^0(1-q)\big]$, which is non-negative if and only if $(\epsilon_n - y_n^0)\,c\rho \le w_r^0(1-q)$. For the asymptotic statement, note that with a fixed teacher $w_n = 1$, $w_r = w_r^0$, $y_n = y_n^0$, and $y_r = 1$ do not change across generations, so every generation takes a gradient step on the same function $\mathcal{L}(\theta) = \mathrm{CE}(\sigma(\theta^\top\phi_n), y_n^0) + w_r^0\,\mathrm{CE}(\sigma(\theta^\top\phi_r), 1)$. Its Hessian is $\epsilon_n(1-\epsilon_n)\phi_n\phi_n^\top + w_r^0\,q(1-q)\phi_r\phi_r^\top \preceq \frac{1}{4}\big(\|\phi_n\|^2 + w_r^0\|\phi_r\|^2\big)I =: L\,I$, so $\mathcal{L}$ is convex and $L$-smooth. For $\eta \le 1/L$ the descent lemma gives $\mathcal{L}(\theta_{g+1}) \le \mathcal{L}(\theta_g) - \frac{\eta}{2}\|\nabla\mathcal{L}(\theta_g)\|^2$, and since $\mathcal{L} \ge 0$ the gradients are square-summable, so $\nabla\mathcal{L}(\theta_g) = (\epsilon_{n,g} - y_n^0)\phi_n + w_r^0(q_g - 1)\phi_r \to 0$. Because $|c| < 1$, $\phi_n$ and $\phi_r$ are linearly independent, so both coefficients vanish in the limit: $\epsilon_{n,g} \to y_n^0$ and $q_g \to 1$. In particular the condition $(\epsilon_{n,g} - y_n^0)\,c\rho \le w_r^0(1-q_g)$ fails at most while $\epsilon_{n,g}$ is still far from $y_n^0$, and any decrease of $\ell_r$ is transient.

**(iii)** Scale the self-generated terms of (i) by $1-\alpha$ and add the external terms of Lemma 1(d). The external recovery term contributes $-\eta\,\beta\,(q-1)\|\phi_r\|^2 = +\eta\,\beta(1-q)\|\phi_r\|^2$ to $\Delta\ell_r$; the external nominal term contributes $-\eta\,\alpha(1-\gamma)(\epsilon_n - 0)\,c\,\|\phi_n\|\|\phi_r\| = -\eta\|\phi_r\|^2\,\alpha(1-\gamma)\,\epsilon_n\,c\rho$. Hence

$$\Delta\ell_r = -\eta\|\phi_r\|^2\Big\{(1-\alpha)\,\frac{p\,\epsilon_n}{\overline{\mathrm{SR}}}\,(1-q)(1+q+q^2)\big[u\,c\rho - h(q)\big] + \alpha(1-\gamma)\,\epsilon_n\,c\rho - \beta(1-q)\Big\},$$

and $\Delta\ell_r \ge 0$ if and only if the brace is non-positive. Dividing by $1-q > 0$ gives equation 8. For equation 5, expand the bracket: $(1-\alpha)\frac{p\epsilon_n}{\overline{\mathrm{SR}}}(1+q+q^2)\,h(q) = (1-\alpha)\,\frac{3p\epsilon_n q^3}{\overline{\mathrm{SR}}} = (1-\alpha)\,w_r = (1-\alpha)\,\kappa\,m(\mathcal{S}_{\mathrm{rec}})$ by Lemma 1(b). Moving this term to the left of equation 8 and adding $\beta$ gives $w^{\mathrm{eff}} = (1-\alpha)\kappa\,m(\mathcal{S}_{\mathrm{rec}}) + \beta$ on the left and the two interference terms on the right. □

### C.7 Proof of Proposition 4

Let $R(\tau) = \mathbf{1}[\tau \text{ succeeds}]$ and let rollouts be drawn from $\pi_\theta$. Success-filtered cloning minimizes the average negative log-likelihood over retained episodes,

$$\mathcal{L}_{\mathrm{filt}}(\theta') = \frac{\mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\sum_t -\log\pi_{\theta'}(a_t \mid s_t)\big]}{P_\theta(\text{success})},$$

whose gradient at $\theta' = \theta$ is $-\mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\sum_t \nabla_\theta \log\pi_\theta(a_t \mid s_t)\big]/\overline{\mathrm{SR}}$. By the score-function identity, $\nabla_\theta\,\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = \mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\sum_t \nabla_\theta\log\pi_\theta(a_t \mid s_t)\big]$, and $\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = J_{\mathrm{nom}}(\theta)$. Hence the cloning step $-\eta\nabla\mathcal{L}_{\mathrm{filt}}$ equals $(\eta/\overline{\mathrm{SR}})\,\nabla_\theta J_{\mathrm{nom}}(\theta)$. For any state-dependent baseline $b$, $\mathbb{E}\big[b(s_t)\nabla_\theta\log\pi_\theta(a_t \mid s_t)\big] = \sum_s d_\theta(s)\,b(s)\sum_a \nabla_\theta\pi_\theta(a \mid s) = 0$, so subtracting a baseline does not change the expectation. In the nominal–recovery MDP the identity can be checked directly: $\partial J_{\mathrm{nom}}/\partial\ell_r = 3p\,\epsilon_n q^3(1-q)$ equals $\overline{\mathrm{SR}}$ times the recovery term $w_r(1-q)$ of Lemma 1(b), and $\partial J_{\mathrm{nom}}/\partial\ell_n = -p\,u\,\epsilon_n(1-q^3)$ equals $\overline{\mathrm{SR}}$ times $-(\epsilon_n - y_n)$. □

## D Rescue: Algorithm and Forecasting Details

Algorithm 1 gives one recursion chain of Rescue, including the budget accounting of Section 3; in the monitor, $\kappa$ is set to the mean number of recovery decisions per recovery trajectory. Uniform accumulation of past self-data is no substitute: it inherits the selection bias of earlier generations (Proposition 1), and its recovery content decays with the nominal error (Proposition 2).

**Coverage hazard.** Theorem 1(iii) writes the change of the recovery logit as interference that scales with the nominal error minus supervision that scales with $1-q$ times the effective recovery mass. For each tier $m \in \{\mathrm{tail}, \mathrm{rec}\}$, with $\hat{\ell}_g^{(m)}$ the predicted logit of its success rate and $\hat{S}_g^{(m)} = \sigma(\hat{\ell}_g^{(m)})$, we therefore forecast from quantities recorded at generation $g$:

$$\hat{\ell}_{g+1}^{(m)} = \hat{\ell}_g^{(m)} - a^{(m)}\,\epsilon_{n,g} + b^{(m)}\,\big(1 - \hat{S}_g^{(m)}\big)\,v_g^{(m)}, \tag{9}$$

with fitted $a^{(m)}, b^{(m)} > 0$, $v_g^{(\mathrm{rec})} = w_g^{\mathrm{eff}}$, and $v_g^{(\mathrm{tail})} = \mathrm{cov}_g(\mathcal{S}_g^-)$. The form absorbs or drops several factors of equation 4, and its tail version is a heuristic analogue, so equation 9 is a working approximation, judged by whether it forecasts held-out tasks and unseen generations (Section 4.5).

**Algorithm 1** RESCUE: one recursion chain

**Require:** seed $\pi_0$; external pools $\mathcal{D}^{\text{ext}}$ annotated with modes; dictionary $\Phi$; reference mass $m_{\text{ref}}$ and historical competence $q_0$ from $\pi_0$ on $\mathcal{H}_{\text{cal}}$; budget $\alpha$; episodes per generation $N$; buffer capacity $B$

$\mathcal{B}_0 \leftarrow \emptyset$
**for** $g = 0$ **to** $G - 1$ **do**
  Collect rollouts with $\pi_g$; retain successes $\tilde{\mathcal{D}}^{\text{self}}_{g+1}$; estimate $m_g, \text{cov}^{\text{mode}}_g, \nu_g, N^{\text{rec}}_g$ ▷ monitor
  Score modes by $\psi_g$ in equation 6; update $\mathcal{S}^-_g$
  Draw replay $\mathcal{R}_g \subseteq \mathcal{B}_g$, $|\mathcal{R}_g| = b_g = b^{\text{self}}_g + b^{\text{ext}}_g$ by origin, $b^{\text{ext}}_g \leq \alpha N$ ▷ replay
  $\mathcal{E}_g \leftarrow$ top-$\psi_g$ external episodes not in $\mathcal{R}_g$, $|\mathcal{E}_g| = \alpha N - b^{\text{ext}}_g$, allocated across modes in proportion to $\psi_g$, distinct trajectories first ▷ selection
  $\mathcal{D}^{\text{self}}_{g+1} \leftarrow (1-\alpha)N - b^{\text{self}}_g$ episodes of $\tilde{\mathcal{D}}^{\text{self}}_{g+1}$ ▷ replay debited from its origin's quota
  Train $\pi_{g+1}$ on $\mathcal{D}^{\text{self}}_{g+1} \cup \mathcal{E}_g \cup \mathcal{R}_g$ by behavior cloning ($N$ episodes; external-origin fraction exactly $\alpha$)
  $\mathcal{B}_{g+1} \leftarrow$ stratified retention of the lowest-coverage mastered modes from $\mathcal{B}_g \cup \tilde{\mathcal{D}}^{\text{self}}_{g+1} \cup \mathcal{E}_g$ at capacity $B$ ▷ retention
  Evaluate $\pi_{g+1}$ on the tiers and banks; log the coverage trace
**end for**

**Fitting the coverage hazard.** The two coefficients per tier of equation 9 are fitted by least squares on the family-A tasks over $G_1$–$G_{10}$. Multi-step forecasts roll the self-generated part of $v_g$ forward with the occupancy identity $\nu_g \propto \epsilon_{n,g}$ of Proposition 2, and $\epsilon_{n,g}$ with a per-arm geometric fit to nominal success, while the external and buffer parts are fixed by the recipe. Besides absorbing $u_g c\rho$ and $h(q_g)$ into the coefficients, the form drops the factor $(1-q_g)(1+q_g+q_g^2)$ from the interference term of equation 4; the model makes no prediction for the tail tier, whose at-risk modes are not recovery modes.

## E PROTOCOL DETAILS

**Tasks and splits.** Family A (fitting, mechanism studies, all shared-pool arms) consists of two LIBERO-Spatial, two LIBERO-Object, and two LIBERO-Long tasks; family B consists of two LIBERO-Goal and two further LIBERO-Long tasks. All ten tasks are trained jointly in every arm, so the two families share recursion chains; family B is never used for fitting the forecasting model, for tuning any method component, or for choosing hyperparameters. Table 21 lists the ten tasks by suite, task index, and instruction.

**Seed model.** The seed $\pi_0$ is $\pi_{0.5}$ fine-tuned from the public pretrained checkpoint on ten of the fifty LIBERO demonstrations per task and on thirty recovery demonstrations per task drawn from the LIBERO-Recover teleoperated trajectories on the ten tasks. The reduced nominal budget leaves the nominal tier room to improve; the full-budget public recipe saturates the standard LIBERO suites. The seed fork of Figure 4a and Table 9 varies the nominal budget (10, 25, or 50 demonstrations per task; N10, N25, N50) and the recovery budget (10, 30, or 130 recovery demonstrations per task; R10, R30, R130), where R130 adds 100 demonstrations per task from the privileged-state scripted recovery controller, drawn from failure scenarios disjoint from $\mathcal{D}^{\text{ext}}_{\text{rec}}$, the evaluation tiers, and both banks. The fork runs arm-D recursion at $\alpha = 0$ only, so the N50 seed, which uses all fifty demonstrations, needs no $\mathcal{D}^{\text{ext}}_{\text{plain}}$. No seed data is re-injected after $G_0$ except in the arms that state it ($G_0$-demonstration replay and patch-back).

**Data per generation and budget accounting.** Each generation collects 150 rollouts per task (1500 in total) under $P_{\text{req}}$, which places 18% of its mass on at-risk conditions by design. The primary setting applies only the success filter; the top-90% quality truncation is applied only in the U+T variant. In the illustration the $G_0$ pass rate is about 78%, so roughly 1170 retained episodes are available at $G_0$ and more at later generations. Every arm trains on exactly $N = 1000$ episodes per generation (100 per task): $(1-\alpha)N - b^{\text{self}}_g$ fresh self-generated episodes, $b^{\text{self}}_g$ self-origin buffer replays, $\alpha N - b^{\text{ext}}_g$ fresh external episodes, and $b^{\text{ext}}_g$ external-origin buffer replays. Each replay is

debited from the quota of its origin, so the external fraction counted over fresh and replayed episodes is exactly $\alpha$: at $\alpha = 0.10$ and $b_g = 90$, if for instance $b_g^{\text{ext}} = 40$, the split is 850 / 50 / 60 / 40, and the training set holds 100 external-origin episodes, as in every other arm at $\alpha = 0.10$. The origin split $b_g^{\text{self}}$ / $b_g^{\text{ext}}$ is logged per generation. A task whose retained episodes fall short of its quota is filled by resampling its own retained episodes, and the resampled count is logged. Surplus self-generated episodes are discarded and never banked outside the buffer.

**External pools.** $\mathcal{D}_{\text{plain}}^{\text{ext}}$ consists of the LIBERO human demonstrations on the ten tasks that were not used to fine-tune the seed and contains no recovery content. $\mathcal{D}_{\text{rec}}^{\text{ext}}$ contains $n_0 = 1000$ unique recovery demonstrations on the ten tasks, combining the LIBERO-Recover teleoperated trajectories available for these tasks with demonstrations produced by a privileged-state scripted recovery controller on additional failure scenarios constructed with the LIBERO-Recover pipeline; the unique-trajectory axis of Table 13e uses stratified subsets of 250 and 500 unique demonstrations. Both pools are disjoint from the seed's injected demonstrations and from both banks by exact state matching and near-duplicate filtering at the scenario-construction tolerance. Nominal-matched, random, and recovery-enriched compositions draw from these pools in stated proportions, and every control names its pool.

**Mode dictionary.** Modes are approach direction (4) $\times$ grasp style (3) $\times$ detour pattern (3), 36 per task, assigned by geometric rules over privileged simulator state and frozen at $G_0$; each trajectory additionally carries a recovery tag in $\{\text{none}, L_1, L_2, L_3, L_4\}$ that records whether it passes through a failure state and the LIBERO-Recover level of that failure. The at-risk set $\mathcal{A}$ used in the analysis is fixed once from the seed's per-mode success: among the modes the seed realizes in at least 1% of its $G_0$ rollouts on a task, those with below-average seed success on that task, which gives 36 modes across family A. RESCUE's monitor instead uses the time-varying set $\mathcal{S}_g^-$ of modes whose mass has fallen below its reference.

**Tiers and banks.** Each tier runs 50 episodes per task per generation. $\mathcal{E}_{\text{nom}}$ samples independent initial states from $P_{\text{req}}$; $\mathcal{E}_{\text{tail}}$ samples from the at-risk portion of $P_{\text{req}}$, restricted to conditions on which the seed succeeds in at least 5 of 10 screening rollouts that are not reused for evaluation; $\mathcal{E}_{\text{rec}}$ starts from LIBERO-Recover failure states on the ten tasks, stratified equally over levels $L_1$–$L_4$. The banks are drawn from the recovery-state pool with identical retention rules (seed success at least 0.70 over 20 rollouts; stratified by task and level), split once by stratified random draw at $G_0$ into $\mathcal{H}_{\text{cal}}$ (200 states) and $\mathcal{H}_{\text{eval}}$ (150 states), and frozen. Seed success on every bank state is then re-estimated on an independent 20 rollouts, and the re-estimate is used for all flip statistics. $\mathcal{H}_{\text{cal}}$ is the only bank read by any method component ($q_0$ and $\psi_g$); $\mathcal{H}_{\text{eval}}$ is never read at training, calibration, model-selection, or early-stopping time and is evaluated once per generation per arm for reporting.

**Training.** Each generation trains for a fixed update budget, identical across arms, from the arm-specified initialization: the seed $\pi_0$ for reset arms and the previous generation's parameters for continuing arms. Optimizer moments and the learning-rate warmup are re-initialized at every generation in every arm, so continuing arms inherit parameters but no optimizer state. Table 19 lists the protocol parameters and Table 20 the optimization and inference settings. For $\pi_{0.5}$ these follow the public `pi05_libero` configuration of openpi with two changes: the budget is 5,000 updates per generation instead of 30,000, and the 10,000-step warmup, which would exceed that budget, is shortened to 500 steps. OpenVLA-OFT and Diffusion Policy keep the optimizer, batch size, and architecture of their public LIBERO and image-based recipes (Kim et al., 2025; Chi et al., 2023). Their update counts are set so that every policy draws the same number of training frames per generation ($5{,}000 \times 256 \approx 1.28$M), which gives 20,000 updates at batch size 64. Seeds are trained with the same settings on twice the per-generation frame budget (10,000 updates for $\pi_{0.5}$, 40,000 for the other two policies); the Diffusion Policy seed starts from random initialization. The RoboCasa and physical studies use the $\pi_{0.5}$ settings unchanged. The sweep in Table 5 scales only the number of updates. All values in Table 20 are taken from the public recipes or derived from them by the rules above, and none is tuned on either task family.

**Statistics.** The statistical unit is the recursion chain ($n = 5$ for $\pi_{0.5}$ on LIBERO, seeds paired across arms; three chains for OpenVLA-OFT, Diffusion Policy, and RoboCasa). All intervals are

computed from raw per-chain success counts and therefore include evaluation noise. Tier-level intervals are hierarchical bootstrap intervals over chains and episodes; paired contrasts use $t$-intervals over the paired chain differences, Holm-corrected within a table; practical equivalence uses two one-sided tests with $\delta = 2$ points on $\mathrm{SR}_{\mathrm{nom}}$ and $\delta = 5$ points on $\mathrm{SR}_{\mathrm{rec}}$, with an insufficient-evidence zone; mode-disappearance events are reported as one-sided 95% upper bounds. Analyses of covariance with the teacher's three-tier success as covariate are reported as sensitivity analyses only. In the text, an estimate with its 95% interval is written as $x_{\pm h}$, where $h$ is the interval half-width; for the few bootstrap intervals that are not symmetric, $h$ is the larger of the two sides. Rank correlations keep bracketed intervals, since they are bounded by one. The overlap $c$ is estimated each generation as the negative mean cosine between the loss gradients of success-filtered nominal and recovery mini-batches.

**Arms and controls.** The design is read through paired contrasts. $D - B$ and $C - A$ change only the teacher and therefore measure data recursion, with and without inherited parameters; $B - A$ and $D - C$ change only the initialization and therefore measure parameter drift; and arm E asks how much uniform accumulation of past self-data retains. Arm D is the success-filtered core of deployment recipes; recipes that also collect expert corrections (Physical Intelligence et al., 2025a) add recovery supervision and are evaluated as a separate arm (Table 11). The loop variant R trains its value function on the arm's own rollouts, and PG uses a learned value baseline. Controls include the frozen seed re-evaluated each generation, static training on the frozen $G_0$ pool, and continued training on $\mathcal{D}^{\mathrm{ext}}_{\mathrm{plain}}$ alone or on $\mathcal{D}^{\mathrm{ext}}_{\mathrm{plain}} \cup \mathcal{D}^{\mathrm{ext}}_{\mathrm{rec}}$. The flip rate of arm A, whose data source never changes, measures the false-flip floor of the bank procedure including retraining noise, and the frozen seed re-evaluated each generation measures the floor due to evaluation noise alone. The overlap $c$ is estimated every generation over 64 batch pairs, and $\rho$ is the ratio of the two gradient norms.

**Mechanism arms.** The open-loop arms of Table 10 each keep their own student but train every generation on data drawn from arm D's rollout pool, so requests and visitation are shared. Arms 6–7 differ only in replacing part of that pool with rollouts of the same policy started from post-failure states sampled from the seed's own $G_0$ failures (disjoint from both banks), which pins the share of rollouts entering $\mathcal{S}_{\mathrm{rec}}$ at $\nu_0$. Patch-back arms 5a–5c spend an equal budget (10% of $N$) on the seed's own successful $G_0$ rollouts and differ only in which modes they select. Arm 2r deletes the recovery-bearing segments from arm 2's training data and substitutes equal nominal data. Closed-loop arms rerun the decisive interventions as independent chains that collect their own rollouts, and loop variants are independent chains with continuing initialization.

Table 18: Experimental arms. *Teacher*: the policy $T_g$ that collects the rollouts of generation $g$; *Init.*: the parameters from which $\pi_{g+1}$ is trained (reset to $\pi_0$, or continued from $\pi_g$). All arms start from the same seed $\pi_0$, collect 150 rollouts per task per generation under the fixed request distribution $P_{\text{req}}$, and train on exactly $N = 1000$ episodes with the same update budget per generation; optimizer moments and warmup are re-initialized at every generation, so continuing arms inherit parameters but no optimizer state. The external fraction is $\alpha = 0$ except in the last block. Arms A–D form a $2 \times 2$ design over teacher and initialization.

| Arm | Teacher | Init. | Selection rule | Training set $\mathcal{D}_{g+1}$ | What the arm isolates |
|---|---|---|---|---|---|
| *Attribution arms (Sections 4.2–4.3)* | | | | | |
| A | $\pi_0$ | $\pi_0$ | success filter | filtered $\pi_0$ rollouts | nothing changes: retraining and false-flip floor |
| B | $\pi_0$ | $\pi_g$ | success filter | filtered $\pi_0$ rollouts | parameter drift on a fixed data source |
| C | $\pi_g$ | $\pi_0$ | success filter | filtered $\pi_g$ rollouts | data recursion, no inherited parameters |
| D | $\pi_g$ | $\pi_g$ | success filter | filtered $\pi_g$ rollouts | both; primary setting |
| E | $\pi_g$ | $\pi_0$ | success filter | $N$ episodes from the union of all past retained self-data | uniform accumulation of self-data |
| *Loop variants: arm D with another selection rule (Section 4.4)* | | | | | |
| U | $\pi_g$ | $\pi_g$ | none ($\omega \equiv 1$) | unfiltered $\pi_g$ rollouts | no improvement signal (Lemma 1a) |
| U+T | $\pi_g$ | $\pi_g$ | top-90% quality truncation | truncated $\pi_g$ rollouts | a weak improvement signal |
| PG | $\pi_g$ | $\pi_g$ | success reward, learned value baseline | on-policy $\pi_g$ rollouts, policy-gradient update with the Flow-Noise likelihood (Chen et al., 2026) | Proposition 4 |
| R | $\pi_g$ | $\pi_g$ | advantage conditioning, value fit on own rollouts | all $\pi_g$ rollouts with advantage labels (Physical Intelligence et al., 2025a) | a softer rule (Remark 1) |
| *Protection: arm D plus external recovery data, $\alpha = 0.10$ (Section 4.5)* | | | | | |
| Unif. replay | $\pi_g$ | $\pi_g$ | success filter | $(1-\alpha)N$ filtered $\pi_g$ rollouts + $\alpha N$ uniform from $\mathcal{D}_{\text{rec}}^{\text{ext}}$ | proportion, no targeting |
| RESCUE | $\pi_g$ | $\pi_g$ | success filter | $(1-\alpha)N$ filtered $\pi_g$ rollouts + $\alpha N$ top-$\psi$ from $\mathcal{D}_{\text{rec}}^{\text{ext}}$ and buffer (Alg. 1) | coverage-targeted composition |

Table 19: Protocol parameters shared by all arms. Optimization and inference settings for each policy are in Table 20, and the tasks are in Table 21.

| Parameter | Value |
|---|---|
| Generations | $G = 10$ (core), $G = 20$ (extrapolation arms) |
| Recursion chains | 5, seeds paired across arms |
| Rollouts per task per generation | 150 |
| Tasks | 10 (family A: 6; family B: 4), trained jointly |
| Seed demonstrations per task | 10 nominal, 30 recovery (fork: 10/25/50 nominal, 10/30/130 recovery) |
| Selection rule, primary setting | success filter only |
| Quality truncation | U+T variant only, top 90% |
| Training episodes per generation | $N = 1000$ (100 per task, all arms) |
| Mixing ratios | $\alpha \in \{0, 0.05, 0.10, 0.20, 0.40\}$ (plus 0.025 in Table 13d); RESCUE default $\alpha = 0.10$ |
| Unique recovery demonstrations | $n_{\text{uniq}} \in \{250, 500, 1000\}$, $n_0 = 1000$ |
| Buffer capacity | $B = 2000$ episodes; replay $b_g = 90$ at $\alpha = 0.10$, external-origin replays charged to $\alpha N$ |
| Evaluation episodes | 50 per task per tier per generation (300 per tier on family A) |
| Bank sizes | $\lvert\mathcal{H}_{\text{cal}}\rvert = 200$, $\lvert\mathcal{H}_{\text{eval}}\rvert = 150$; 20 rollouts per state |
| At-risk request share | 18% |
| Equivalence margins (TOST) | $\delta = 2$ points on $\text{SR}_{\text{nom}}$, $\delta = 5$ points on $\text{SR}_{\text{rec}}$ |
| Parameter-space baselines | EWC $\lambda \in \{1, 10, 100, 1000\}$; KL coefficient $\in \{0.01, 0.03, 0.1, 0.3\}$; merge weight on $\pi_0 \in \{0.1, 0.2, 0.3, 0.5\}$ |
| Episode step limit (nominal, tail) | 220 / 280 / 300 / 520 for Spatial / Object / Goal / Long, after 10 settling steps, as in the public LIBERO evaluation scripts |

Table 20: Optimization and inference settings per policy. Frames per generation are matched across policies; updates per generation are the $\times 1$ row of Table 5. The $\pi_{0.5}$ column also applies to the RoboCasa and physical studies.

| Setting | $\pi_{0.5}$ | OpenVLA-OFT | Diffusion Policy |
|---|---|---|---|
| Initialization of the seed | public $\pi_{0.5}$ base checkpoint | OpenVLA-7B | random |
| Fine-tuning mode | full | LoRA, rank 32, dropout 0 | full (trained from scratch) |
| Batch size | 256 | 64 (8 per GPU $\times$ 8 GPUs) | 64 |
| Updates per generation | 5,000 | 20,000 | 20,000 |
| Seed updates | 10,000 | 40,000 | 40,000 |
| Optimizer | AdamW, $\beta = (0.9, 0.95)$, weight decay $10^{-10}$, gradient clipping at norm 1.0 | AdamW as in the public fine-tuning script | AdamW, $\beta = (0.95, 0.999)$, weight decay $10^{-6}$ |
| Learning rate and schedule | peak $5 \times 10^{-5}$; 500-step linear warmup, then constant | constant $5 \times 10^{-4}$, no warmup (the public $10\times$ decay at 100k steps is never reached) | peak $10^{-4}$; 500-step warmup, cosine decay over the generation |
| Warmup and optimizer state | re-initialized every generation | re-initialized every generation | re-initialized every generation |
| Weight averaging | EMA 0.999; EMA weights are evaluated and inherited | none; LoRA weights merged for evaluation | EMA (power 0.75, max. 0.9999) |
| Inputs | third-person and wrist images at $224 \times 224$, proprioceptive state, instruction | third-person and wrist images, proprioception, instruction; random-crop augmentation | third-person and wrist images, proprioception, task-identity embedding |
| Predicted chunk | 10 actions | 8 actions | 16 actions, 2 observation steps |
| Executed before replanning | 5 actions | 8 actions | 8 actions |
| Action decoding | 10 flow-matching integration steps | parallel decoding, L1 head | 100 DDPM denoising steps |

Table 21: Task identifiers. Index is the 0-based task index within the suite in the LIBERO benchmark task map; LIBERO-Long is the `libero_10` suite.

| Family | Suite | Index | Instruction |
|---|---|---|---|
| A | LIBERO-Spatial | 0 | pick up the black bowl between the plate and the ramekin and place it on the plate |
| A | LIBERO-Spatial | 3 | pick up the black bowl on the cookie box and place it on the plate |
| A | LIBERO-Object | 4 | pick up the ketchup and place it in the basket |
| A | LIBERO-Object | 6 | pick up the butter and place it in the basket |
| A | LIBERO-Long | 0 | put both the alphabet soup and the tomato sauce in the basket |
| A | LIBERO-Long | 2 | turn on the stove and put the moka pot on it |
| B | LIBERO-Goal | 0 | open the middle drawer of the cabinet |
| B | LIBERO-Goal | 3 | open the top drawer and put the bowl inside |
| B | LIBERO-Long | 3 | put the black bowl in the bottom drawer of the cabinet and close it |
| B | LIBERO-Long | 9 | put the yellow and white mug in the microwave and close it |