RESET COST DOES NOT DETERMINE RESET VALUE: THE ROLE OF POST-RESET UPDATE BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuron resets restore plasticity in deep reinforcement learning, and recent work frames the choice of which units to reset as a cost–benefit problem with an estimable immediate cost. We find that the immediate disruption is measurable but that how harmful it turns out to be depends on the update budget after the reset. From common SAC checkpoints on six DeepMind Control tasks we apply matched reset interventions with an exact no-op control and measure the immediate change in the critic's TD loss and the return over a fixed 150k-step continuation. The immediate disruption is accurately ranked and spans three orders of magnitude across selection rules. Yet the costliest reset is far less harmful under high update pressure than under low pressure on every task ($6/6$), and a random reset of intermediate cost is moderated in the same direction. A history $\times$ recovery-budget factorial on two tasks attributes most of this effect to the post-reset update budget rather than to the training history the reset meets, in a way consistent with update supply. Immediate cost alone therefore does not rank value across post-reset budgets, and no consistent cost–value ordering holds across the tested regimes. Weakening the reset removes most low-budget harm even for high-cost selections. Reset evaluations should report the post-reset update budget and horizon under which a method is judged. Results are for critic-only resets of state-based SAC, with one TD3 replication.

1 INTRODUCTION

Neural networks trained by reinforcement learning lose plasticity as training proceeds: units become dormant, effective rank falls, and the network fits new targets poorly (Kumar et al., 2021; Lyle et al., 2022; 2023; Sokar et al., 2023; Abbas et al., 2023; Lyle et al., 2024; Klein et al., 2024). A widely used remedy is to periodically reinitialize units that appear unused (Sokar et al., 2023); it is double-edged, producing large gains or collapse depending on which units are chosen. The natural response is a better selection criterion (Liu et al., 2025); recent work formalizes selection as a cost–benefit trade-off with an estimable cost and a delayed benefit (Elisii et al., 2026). Update pressure—gradient updates per environment step—is meanwhile the main lever of sample-efficient deep RL (Chen et al., 2021; Hiraoka et al., 2022; D'Oro et al., 2023; Nauman et al., 2024b) and the regime in which resets are most often deployed (Nikishin et al., 2022; Schwarzer et al., 2023).

We find that the immediate disruption of a reset is measurable but that its downstream consequence depends on the recovery budget. A reset damages what the network currently computes and changes what it learns next; it has a *cost*—the immediate loss of what the reset units contributed to the current objective—and a *value*—the return that follows it; the *benefit* of a reset is its value relative to a no-op control. We ask whether the estimable term suffices. The immediate cost can be ranked well (Section 4), and at the extremes of the cost range it orders long-run value under low pressure. But the magnitude, and on some tasks the sign, of a reset's value is set by the update pressure after it (Section 5); a history $\times$ recovery factorial attributes most of the pressure effect to the update budget after the reset rather than to the learner it meets (Section 6); and no consistent cost–value ordering holds across the tested regimes (Section 7). For practice this means that weakening the reset removes most low-budget harm even for high-cost selections, and that reset evaluations should report the post-reset update budget and horizon.

Our design branches every arm from a common checkpoint of a SAC agent on DeepMind Control, with the number of units reset per layer, the reset strength, the operator, the fresh initialization and

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.

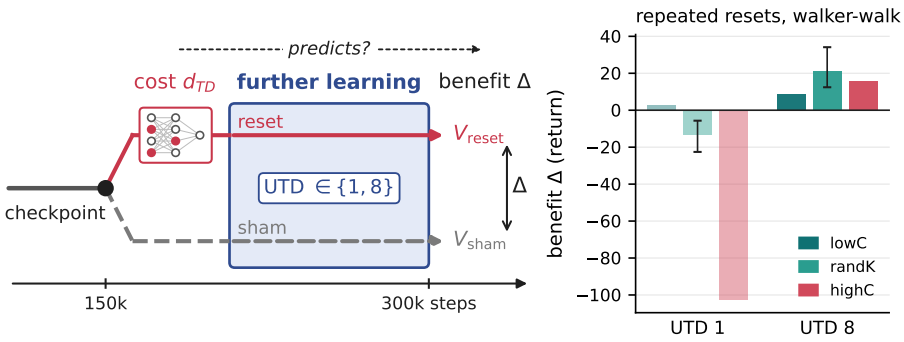


Figure 1: **Does the immediate cost of a reset predict its value after further learning?** *Left:* every arm branches from a common checkpoint. The cost d_{TD} is the change in critic TD loss across the reset (same audit batch, fixed target); the intervention is critic-only, so the return is untouched at that instant. Both arms then learn for the same budget at the update pressure under study, and the benefit Δ is the difference in value from `sham`, which computes every score and resets nothing. *Right:* the benefit of three repeated resets on `walker-walk` at UTD 1 (light) and UTD 8 (dark), AUC in raw return (Section 5; `randK` with seed-blocked CI): at UTD 1 the order follows immediate cost, at UTD 8 it does not.

the random-number streams of the subsequent interaction and evaluation paired across arms, so that the selection rule is the only manipulated variable. A `sham` arm computes every score and resets nothing; it is an exact no-op and the correctness check of the design (Section 3).

Contributions.

- **The post-reset update budget strongly moderates the value of a fixed reset.** On six tasks with fresh seeds, repeatedly resetting the highest-scoring units—the costliest reset, used as a stress test—is far less harmful under high update pressure than under low pressure ($6/6$ tasks; task-level contrast $+32\%$ of baseline return) despite six to ten times the cumulative cost; a random-K reset of intermediate cost is moderated in the same direction ($16/16$ seeds), the effect is graded over UTD 1–8 on one task, and the reversal replicates with TD3 (Section 5).
- **The budget acts after the reset.** A history $\times$ recovery UTD factorial on two tasks attributes most of the `highC` pressure contrast to the post-reset UTD ($16/16$ seeds) and a smaller, task-dependent share to the pre-reset history; a single-reset factorial reproduces the budget channel and does not detect it at matched gradient updates, consistent with update supply (Section 6).
- **No consistent cost–value ordering holds across the tested regimes.** Within a regime, the cost–value relation of a single reset changes sign with pressure on `walker-walk` and is weak on `cheetah-run` (Section 7). As a supporting result, the sensitivity score used as the cost term ranks the held-out loss change of random masks (Spearman 0.83; signed first-order term 0.10) and separates the cheapest from the costliest units by about $3{,}100\times$ (Section 4).

2 Related Work

Plasticity loss and its remedies. Loss of plasticity has been characterized through rank collapse (Kumar et al., 2021; Lyle et al., 2022), feature and gradient statistics (Lyle et al., 2023; 2024; Berariu et al., 2021; Igl et al., 2021), in on-policy (Juliani & Ash, 2024) and visual (Ma et al., 2024; Lee et al., 2023) settings (survey: Klein et al., 2024). Remedies include resets (Nikishin et al., 2022; Schwarzer et al., 2023), regularizers and optimizer interventions (Lyle et al., 2023; Nauman et al., 2024a; Elsayed & Mahmood, 2024), activation-based recycling of dormant units (ReDo; Sokar et al., 2023), and dormant-ratio-guided perturbation and exploration (DrM; Xu et al., 2024). We quantify the cost of selection and show that activation and contribution diverge on it.

Question	Section	Tasks	Seeds per cell	Reset	Pressure
Does the score rank immediate cost?	§4	walker; both	8	single	UTD 1, 8
Does pressure moderate reset value?	§5	six DMC tasks	8, 8, 6, 6, 6, 6	repeated	UTD 1, 8
Does cost order value within a regime?	§7	walker, cheetah	8	single (+ repeated)	UTD 1, 8
History or recovery budget?	§6	walker, cheetah	8	repeated; single	$h, r \in \{1, 8\}$

Table 1: The experiments of the main text; every arm branches from a common checkpoint per seed. Seed lineage and pre-registration records are in Appendix U.

`sham`	scores computed, nothing reset (exact no-op)	d_{TD}	*cost*: change in critic TD loss at the reset
`lowC` / `highC`	bottom-/top-K units by contribution	V	*value*: post-branch return summary (AUC)
`lowA`	bottom-K units by activation (ReDo-like)	$\Delta^a_{u,s}$	*benefit*: $V_{a,u,s} - V_{\texttt{sham},u,s}$
K	10% of units per critic hidden layer	$\tilde{\Delta}, \tilde{I}$	the same, relative to the cell's mean `sham` value
u; h, r	UTD; history and recovery UTD	I^a_s	pressure contrast $\Delta^a_{8,s} - \Delta^a_{1,s}$
V_{reg}	pre-specified value (mean of the 18 evaluations)	I_H, I_R	history and recovery marginal contrasts

Table 2: Notation; "harmful" and "beneficial" refer to the sign of Δ.

Gradient-based diagnostics and utility-based replacement. Gradient magnitude has been argued to reflect learning capacity better than activation (Liu et al., 2025), and gradient pathologies have been studied at the network (Creus Castanyer et al., 2025) and data-distribution (Wu et al., 2026) level. Continual Backprop replaces low-utility units (Dohare et al., 2024) and ReGraMa resets units by gradient magnitude (Liu et al., 2025). Closest to our estimator, Elisii et al. (2026) formalize reset selection as a cost–benefit trade-off, $B_t(S) - \lambda C_t(S)$, treat the benefit as a delayed counterfactual and the cost as a directly estimable term, and validate an intervention-aligned cost estimator in supervised continual learning. We use $|\nabla_{\theta_i} L \cdot d_i|$ (d_i the displacement the operator will apply) as the cost term of that trade-off and test how far a good cost estimator helps reset selection in deep RL.

Separating immediate effect from later benefit. Plasticity Injection (Nikishin et al., 2023) reads plasticity loss off learning after a prediction-preserving intervention; Sustainable Backup Propagation (Zhou et al., 2025) adds a recovery mechanism to cyclic resets. We hold the intervention fixed and measure cost and later value separately.

Update pressure. High update-to-data ratios buy sample efficiency (Chen et al., 2021; Hiraoka et al., 2022; D'Oro et al., 2023; Nauman et al., 2024b) at the price of value divergence and overfitting (Hussing et al., 2024; Nauman et al., 2024a), and periodic resets are the standard countermeasure in that regime (Nikishin et al., 2022; D'Oro et al., 2023). Those works established one direction: resetting makes high replay ratios usable. We study the converse, holding the operator fixed and asking how pressure changes what a reset is worth, and separate the update budget after the reset from the history before it (churn across updates is one route by which history could matter: Tang & Berseth, 2024; Tang et al., 2025). **Soft resets.** Shrink-and-perturb (Ash & Adams, 2020) underlies BBF's periodic soft resets (Schwarzer et al., 2023); recency-weighted replay (Wu et al., 2026) is a sampling correction, whereas we hold the sampler fixed.

3 Experimental setup

Agent and tasks. We use SAC (Haarnoja et al., 2018) with twin critics on six DeepMind Control tasks (Tassa et al., 2018) from state observations with action repeat 2: `walker-walk` and `cheetah-run` carry every experiment (the single-event cost experiment of Section 4 uses `walker-walk` only), and `cartpole-swingup`, `finger-spin`, `ball_in_cup-catch` and `reacher-hard` join for the repeated-schedule test of Section 5. Seeds per experiment are listed in Table 1 and Appendix U; Table 2 collects the notation. Update pressure is the update-to-data ratio UTD $\in \{1, 8\}$, and nothing else differs between pressure levels. A trunk is trained to a branch point of 150k environment steps at both pressures (matched data budget, different update budget), and its complete state, replay buffer and random-number streams are saved.

Intervention. Resets are applied to the critic hidden layers, where the TD loss defines contribution. For unit i we compute $c_i = \mathrm{mean}_m \, |\nabla_{\theta_i} L_{TD}(m) \cdot d_i|$ over microbatches m, with $d_{\mathrm{in}} = \theta^0 - W_{\mathrm{in}}$ and $d_{\mathrm{out}} = -W_{\mathrm{out}}$: the block scored is the unit's incoming weights and outgoing column (bias term omitted; immaterial, Appendix V). Activation is the ReDo dormancy statistic. The arms are

UTD	sham	lowC	lowA	highC	randK†
1	0 (exact)	5.6	12.1	18,150	1,133
8	0 (exact)	12.9	28.1	39,540	1,507

Table 3: Median immediate critic disruption d_{TD} over 8 seeds. lowC and highC differ by $\sim$3,100$\times$ in $8/8$ seeds at both pressures. †Random K mask, per-event median on the repeated schedule of Section 5 (walker-walk seeds 40–47; Appendix J).

sham, lowC (bottom-K by contribution), highC (top-K by contribution) and lowA (bottom-K by activation), with K equal to 10% of units per layer; all arms share the checkpoint, the environment stream, K, the operator and the fresh draw θ^0, and both scores are computed in every arm.

Measurement. Immediate cost d_{TD} is the change in TD loss on a fixed audit batch of $4{,}096$ replay transitions drawn once per event, scored against a bootstrap target computed once and held fixed; it is also the scoring batch, so the cost is in-sample with respect to selection (a disjoint batch orders masks the same way, Appendix A). Value is the time-normalized area under the evaluation-return curve over the fixed 150k-step post-branch budget (AUC: the trapezoidal integral over environment steps divided by the budget), a fixed-horizon quantity. V_{reg}, the plain mean of the 18 evaluations, was the pre-specified metric; the body reports the AUC because it is time-normalized (the irregular evaluation times give the registered mean a $4/18$ weight on the first 10k post-branch steps, Appendix V); the two agree on every headline result and disagree on one secondary comparison (Appendix H; both metrics for every headline result in Table 17). Throughout, the *benefit* of an arm is the paired difference $\Delta^a_{u,s} = V_{a,u,s} - V_{\mathtt{sham},u,s}$ against the sham arm of the same seed s and pressure u; it is a signed quantity in raw return units and is negative when the reset is harmful. All statistics use the AUC unless V_{reg} is named; central tendencies are means with 95% paired bootstrap intervals (Agarwal et al., 2021) unless a median is stated. The design, estimand and decision rule of every confirmatory experiment were fixed in advance, and the adjudication scripts were written before any run (Appendix U).

4 Contribution predicts the immediate cost of a reset

Can the immediate cost of a reset be ranked before the reset is made? Table 3 and Figure 18 report the immediate cost of a single reset. Selecting the lowest- rather than the highest-contribution units changes the cost of an otherwise identical intervention by more than three orders of magnitude, in every seed at both pressures (paired bootstrap CI excluding zero). The sham arm returns exactly zero, the correctness condition of the design.

Activation is not a substitute for contribution. The two rules agree on about half of the selected set (Jaccard 0.53), and the disagreement is costly: lowA disturbs the objective 2.2$\times$ more than lowC ($8/8$ seeds, CI excluding zero).

Scope. d_{TD} measures critic TD disruption, not behavioural damage: the intervention touches only the critic, so the immediate change in return is exactly zero in every arm. Pressure raises the cost by less than the raw scale suggests: normalized by the pre-reset TD loss, which is itself 1.36$\times$ larger at UTD 8, the amplification is 1.5–1.9$\times$ (Figure 18; per-arm intervals in Appendix V). The cost term is therefore measurable and well ranked; the rest of the paper asks whether it suffices.

What the score measures. c_i is the absolute value of a per-unit first-order term, not the signed first-order change $\nabla_\theta L^\top d$. On 32 checkpoints with 100 random 10% masks each, scored on one batch and evaluated on a disjoint held-out batch, the score ranks the held-out loss change at Spearman $\rho = 0.83$ (median and mean; CI of the mean $[0.80, 0.86]$), while the signed prediction does not ($\rho = 0.10$; worse on $32/32$ checkpoints) and predicts a loss decrease for a median 28% of masks although none decreases it; gradient magnitude ranks nearly as well (0.75) and activation less well (median 0.56, mean 0.55). We therefore call c_i an absolute first-order sensitivity whose ordering, not level, is calibrated (Appendix A; alternative scores, draw dependence and set-cost additivity in Appendix B).

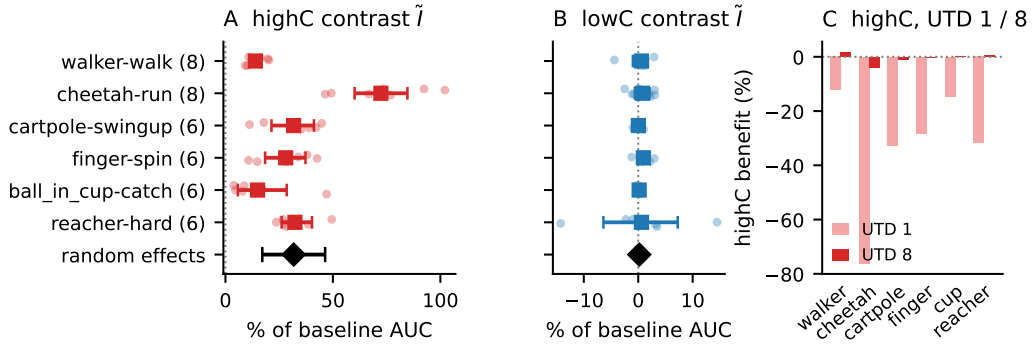


Figure 2: Pressure moderation of reset value on six DeepMind Control tasks (AUC; seeds per task in parentheses; V_{reg} values in Table 17). **A, B** task-level pressure contrast $\tilde{I} = \tilde{\Delta}_8 - \tilde{\Delta}_1$ for the repeated `highC` and `lowC` schedules, with seed-blocked bootstrap CIs, per-seed points and the random-effects estimate (diamond; `highC` $+31.8\%$ $[+17.1, +46.4]$, `lowC` $+0.2\%$ $[-0.0, +0.4]$; both $6/6$ tasks, Holm 0.031). **C** `highC` benefit at UTD 1 and UTD 8 per task.

	UTD 1				UTD 8			
task (seeds)	sham V	$\Delta_{\texttt{lowC}}$	$\Delta_{\texttt{randK}}$	$\Delta_{\texttt{highC}}$	sham V	$\Delta_{\texttt{lowC}}$	$\Delta_{\texttt{randK}}$	$\Delta_{\texttt{highC}}$
`walker-walk` (8)	840	+3	−13	−103	931	+9	+21	+16
`cheetah-run` (8)	590	+8	−33	−449	784	+17	+16	−30
`cartpole-swingup` (6)	871	−2	–	−285	870	−2	–	−9
`finger-spin` (6)	660	−2	–	−186	762	+5	–	−2
`ball_in_cup-catch` (6)	974	+0	–	−143	970	+2	–	+3
`reacher-hard` (6)	847	−9	–	−269	915	−4	–	+5

Table 4: Six tasks, repeated schedule: mean `sham` value (AUC, raw return) and the benefit of the low-cost, random-K (two tasks) and high-cost rules at each pressure; at UTD 8 the costliest reset is at most slightly positive and still harmful on two.

5 Does immediate cost order long-run value?

Does the immediate cost of a reset order its long-run value once update pressure changes? We repeat the matched intervention every 10k environment steps and measure its value on the six tasks, each with its own fresh seeds (8, 8, 6, 6, 6, 6), at UTD 1 and UTD 8, with arms `sham`, `lowC` and `highC` (and `lowA` on the first two tasks). `highC` serves as a stress test: it resets the units of maximal immediate cost and so exposes the pressure effect most clearly (an intermediate-cost arm follows below). The benefit $\Delta^a_{u,s}$ of an arm is placed on a relative scale by dividing by the mean `sham` value of the same task and pressure ($\tilde{\Delta}$, a constant within a task–pressure cell), and the seed-level pressure contrast is $\tilde{I}^a_s = \tilde{\Delta}^a_{8,s} - \tilde{\Delta}^a_{1,s}$. On the first two tasks the `highC` pressure contrast is positive in $16/16$ seeds (exact $p = 2^{-16}$; pooled magnitude in Appendix F). Across all six tasks, with the task as the unit of analysis, the across-task mean of the task-level contrasts is tested by an exact task-level sign-flip test (2^6 assignments, one-sided, Holm-corrected over the two arms), with a DerSimonian–Laird random-effects estimate reported.

Result (Figure 2). The most expensive reset is far less harmful under high pressure on every task: the `highC` pressure contrast is positive on $6/6$ tasks (task-level $p = 0.016$, Holm 0.031), across-task mean $+32.2\%$ of baseline return, random-effects estimate $+31.8\%$ (CI $[+17.1\%, +46.4\%]$, $I^2 = 0.95$; Hartung–Knapp $[+9.8\%, +53.8\%]$, Appendix F). The normalized cumulative critic disruption of the `highC` schedule is 6–10$\times$ larger at UTD 8 on the two core tasks (cumulative; single-event amplification 1.5–1.9$\times$, Section 4), and UTD 8 is nevertheless the far better outcome in every task: cumulative local cost did not order these outcomes; pressure did. Whether the schedule becomes *beneficial* at high pressure is task-dependent: at UTD 8 the `highC` benefit is slightly positive on `walker-walk`, `ball_in_cup-catch` and `reacher-hard` and negative on the other three (Figure 2C), so "far less harmful" is the result that generalizes and "beneficial" is not. Table 4 gives

the underlying values: at UTD 8 the costliest reset is near neutral on four tasks and remains harmful on two, so the pressure contrast is a reduction of harm. A TD3 replication on `walker-walk` gives a `highC` pressure contrast of $+11.8\%$ of baseline return (CI $[+7.8\%, +16.7\%]$, $6/6$ seeds, exact $p = 1/64$; Appendix G).

The low-cost schedule. `lowC` is pressure-moderated in the same direction on every task but by little ($6/6$ tasks positive, $p = 0.016$, Holm 0.031; random-effects $+0.2\%$, CI $[-0.0\%, +0.4\%]$): consistent in direction but small. The activation and contribution arms do not differ detectably, and the repeated low-cost arm re-selects almost the same units at every event (93% overlap; Appendix R; origin of the hypothesis in Appendix E).

An intermediate-cost reset. A random K mask redrawn at every event (`randK`) costs $1{,}133$ and $1{,}507$ per event on `walker-walk` at UTD 1 and 8, three orders of magnitude above `lowC` and 6–24$\times$ below `highC` in the same runs (Appendix J). Its pressure contrast is detected in the same direction (pooled $+5.8\%$ of baseline return, CI $[+4.2\%, +7.6\%]$, $16/16$ seeds, exact $p = 2 \times 10^{-5}$): harmful at UTD 1 (-13 and -33) and positive at UTD 8 on both tasks ($+21$ and $+16$), better than `lowC` on `walker-walk` ($+12$, CI $[+4, +25]$) and not detectably different on `cheetah-run` (-0.4, $[-6, +5]$). The selection $\times$ regime interaction (`randK`$-$`lowC` at UTD 8 minus at UTD 1) is $+29$ $[+12, +50]$ and $+41$ $[+28, +55]$ ($8/8$ seeds each; descriptive): between these two regimes the preferable reset changes (Figure 1; curves in Figure 9). The change is not reproduced at a fixed training history: on the `walker-walk` checkpoints of the low-cost factorial, no difference between `randK` and `lowC` is detected at either recovery UTD (`randK`$-$`lowC` $+0.2$ $[-1.8, +2.3]$ at recovery UTD 1, -3.6 $[-12.6, +4.6]$ at UTD 8; interaction -3.8 $[-12.8, +3.5]$; Appendix P), and `randK` is nearly harmless there even at recovery UTD 1. Which selection is preferable therefore depends on the checkpoint cohort as well as on the regime.

Dose–response, and where the harm goes. On `walker-walk` the pressure effect is graded: with fresh seeds at UTD 2 and 4 (combined with the UTD 1 and 8 cells above), the `highC` benefit rises monotonically from -12.2% of baseline at UTD 1 through -4.5% and $+0.7\%$ to $+1.7\%$ at UTD 8 (Spearman $\rho = 1$ over four levels, exact permutation $p = 1/24$), crossing zero between UTD 2 and 4 (Appendix K). Splitting the paired return difference into loss and gain areas (Appendix S), at UTD 8 the loss area of `highC` and `randK` on `walker-walk` is under one return unit and their gain area ($+16$, $+21$) exceeds `lowC`'s ($+9$): high pressure does not merely shorten the harm. Immediate cost thus orders value at UTD 1 (`lowC` $>$ `randK` $>$ `highC`) but not at UTD 8; immediate cost alone did not consistently order value across regimes.

6 Where the pressure effect comes from: training history versus recovery budget

Does pressure act through the learner the reset meets or through the updates that follow it? Section 5 varied UTD before and after the intervention together, so its contrast cannot separate them. We cross the two: the history UTD h trains the trunk and the recovery UTD r is used after the branch, with the repeated `highC` schedule against `sham` in every cell, on the checkpoints of the earlier low-cost factorial (`walker-walk` and `cheetah-run`, eight seeds each; Appendix M). With $\Delta_{h,r,s} = V_{\texttt{highC},h,r,s} - V_{\texttt{sham},h,r,s}$, $I_{H,s} = \frac{1}{2}[(\Delta_{8,1,s} - \Delta_{1,1,s}) + (\Delta_{8,8,s} - \Delta_{1,8,s})]$, $I_{R,s} = \frac{1}{2}[(\Delta_{1,8,s} - \Delta_{1,1,s}) + (\Delta_{8,8,s} - \Delta_{8,1,s})]$, and $\Delta_{8,8,s} - \Delta_{1,1,s} = I_{H,s} + I_{R,s}$: a contrast of the same form as the pressure contrast of Section 5, on these seeds, splits algebraically into a history and a recovery marginal contrast. Each cell benefit $\Delta_{h,r,s}$ is first placed on a relative scale by dividing by the mean `sham` return of its own task and cell (a constant within a task–cell), and the marginal contrasts $\tilde{I}_{H,s}, \tilde{I}_{R,s}$ are then formed from these $\tilde{\Delta}_{h,r,s}$. We test the pooled pair $\{\bar{\tilde{I}}_H, \bar{\tilde{I}}_R\}$ with an exhaustive paired sign-flip test with within-task flips (2^{16} assignments, Holm-corrected over the pair); per-task tests are secondary.

Result (Figure 3). Pooled over sixteen seeds the diagonal contrast is $+44.8\%$ of baseline return (AUC; $16/16$ seeds) and both marginal contrasts are detected, but unequally: $\bar{I}_R = +38.1\%$ (CI $[+24.2\%, +52.4\%]$, $16/16$, exact $p = 2 \times 10^{-5}$, Holm 3×10^{-5}) against $\bar{I}_H = +6.6\%$ (CI $[+2.0\%, +10.8\%]$, $15/16$, $p = 0.005$, Holm 0.005). The recovery-budget contrast is dominant and consistent across tasks (`walker-walk` $+97$, `cheetah-run` $+441$ return, $8/8$ each); the history contrast is smaller and task-dependent, detected on `walker-walk` ($+49$, $8/8$) and not on

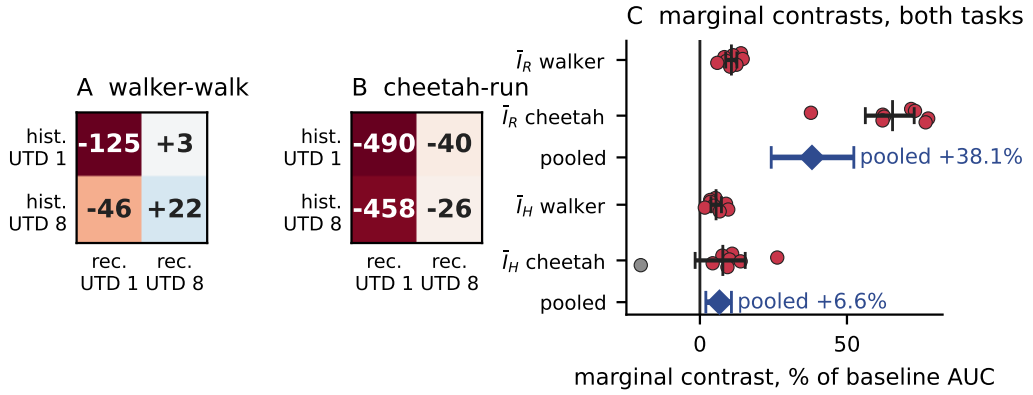


Figure 3: History × recovery factorial for the repeated `highC` schedule, on the time-normalized AUC. **A, B** cell means of the `highC` benefit $\Delta_{h,r}$ (raw return) on `walker-walk` and `cheetah-run`, eight seeds each. **C** per-seed marginal contrasts on the relative scale, with per-task seed-blocked bootstrap intervals and the pooled estimates as diamonds ($\bar{I}_R = +38.1\%$ $[+24.2, +52.4]$, $\bar{I}_H = +6.6\%$ $[+2.0, +10.8]$; sixteen-seed bootstrap; V_{reg} values in Table 17).

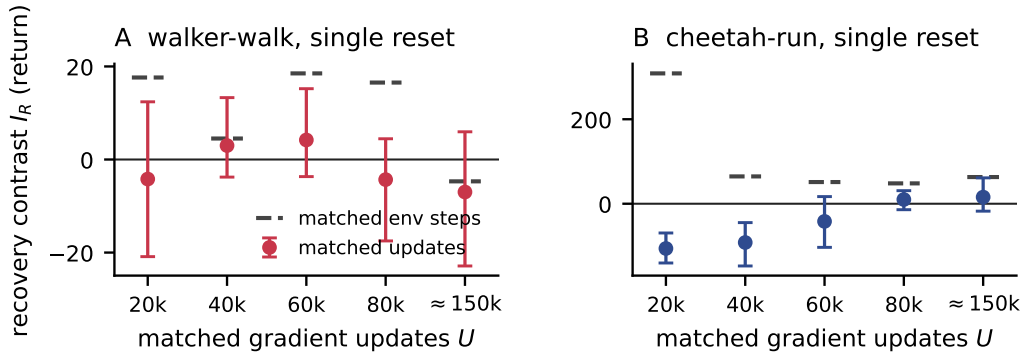


Figure 4: A single `highC` reset with no later intervention (Appendix O): the recovery contrast I_R re-formed at matched gradient updates (recovery UTD 8 after t steps against UTD 1 after $8t$; dots, seed-blocked bootstrap CIs) against the same evaluations compared at matched environment steps (dashes). The last pair (+20k against +150k) is approximately matched.

`cheetah-run` ($+23$, $6/8$, $p = 0.25$). In the pooled split the post-reset UTD marginal accounts for 85% of the diagonal contrast (seed-blocked CI $[78\%, 95\%]$; 88% on raw returns, 87% under V_{reg}; Appendix M). These shares are marginal contrasts of a manipulated post-reset UTD, which also changes actor and target update frequency and the data collected; they are not a mediation share. Within each history the recovery effect is $+128$ and $+67$ on `walker-walk` (history UTD 1, 8) and $+450$ and $+432$ on `cheetah-run` (Appendix S). With a recovery UTD of 8 the schedule is close to neutral on `walker-walk` ($+3$ and $+22$) and far less harmful on `cheetah-run` (-40 and -26, against -490 and -458 at recovery UTD 1), whatever the history. At the endpoint the recovery contrast remains detected while the history contrast does not (Table 17).

Update count, or something more? A recovery UTD of 8 also supplies eight times as many gradient updates per environment step. At matched gradient updates (UTD 8 after t steps against UTD 1 after $8t$; Appendix N) the repeated-schedule recovery contrast vanishes for `lowC` but survives for `highC` (surviving fraction 0.87–1.06 and 0.90–1.05 from 40k and 60k matched updates on the two tasks). At $+8t$ steps, however, the UTD-1 arm has undergone two to fourteen further resets, so the repeated schedule confounds update count with reset count. A single-reset factorial removes this confound (one `highC` reset at the branch, nothing after, new `sham` arms, 128 runs; Appendix O). It reproduces the recovery channel at matched steps ($\bar{I}_R = +10.3\%$ of baseline return, CI $[+7.8\%, +12.9\%]$,

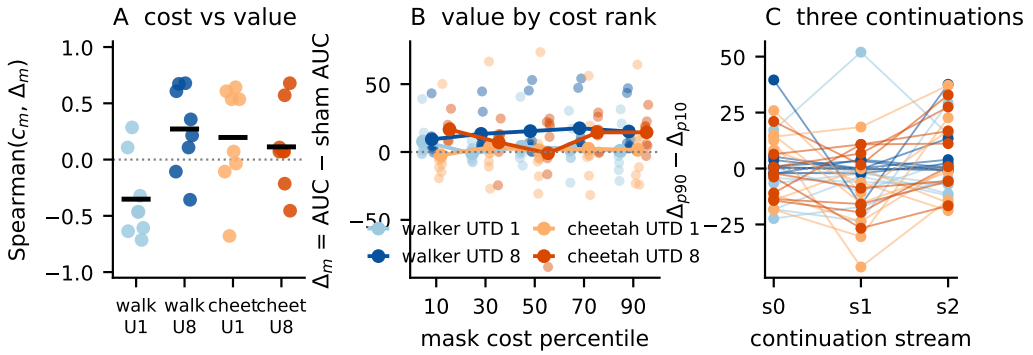


Figure 5: A single reset from a common checkpoint, no later intervention; value on the time-normalized AUC. **A** per-checkpoint Spearman correlation between the immediate cost and the value of seven masks, 32 checkpoints (two tasks, two pressures, eight fresh seeds each); bars are cell means (-0.35, +0.27, +0.20, +0.11). **B** value by cost percentile, cell means with per-checkpoint points. **C** the costliest-minus-cheapest value difference of each checkpoint under three continuation streams (same checkpoint, same masks, different post-reset randomness; streams s0–s2): its sign agrees across all three in $15/32$ checkpoints (chance under no effect: $8/32$).

$16/16$ seeds, Holm 0.00003; history -0.4%, $10/16$, not detected) and does not detect it at matched updates: from 60k updates every CI includes zero on both tasks, point estimates -42 to $+16$ return against $+53$ and $+102$ at matched steps, and earlier the UTD-8 arm is behind on `cheetah-run` (Figure 4; Appendix O). The repeated-schedule survival is consistent with the reset-count confound rather than with update density, and for a single reset the recovery budget is consistent with update supply (single-evaluation contrasts with wide intervals; a fixed replay stream would separate update count from replay age, Appendix N).

What the finding implies. If the budget decides whether a costly reset is repaid, scaling the reset by the expected budget should avoid the low-budget harm. We test one untuned rule, $\alpha = \min(1, \mathrm{UTD}/8)$ (soft-reset strength), on the repeated `highC` schedule and seeds of Section 5 (Table 7, Appendix H). At UTD 1, scaling down reduces the distance to `sham` by 97% (`walker-walk`) and 97% (`cheetah-run`): pooled $\Delta_{0.125} - \Delta_1 = +43.1\%$ of baseline return (CI $[+27.3\%, +59.4\%]$, $16/16$ seeds, exact $p = 2 \times 10^{-5}$). The scaled `highC` at UTD 1 ($+3$ on `walker-walk`, -15 on `cheetah-run`) lands in the range of `lowC` at UTD 1 on the same checkpoints ($+3$, $+8$; AUC, with V_{reg} in Table 7), a comparison for which no equivalence margin was defined. At UTD 8 the full reset is not detectably better than the scaled one (pooled $\Delta_1 - \Delta_{0.125} = -1.0\%$, CI $[-2.5\%, +0.4\%]$, two-sided $p = 0.22$; under V_{reg} the scaled reset is slightly better, -2.8%, $p = 0.040$), so a constant small strength accounts for the effect as well as budget-aware scaling does; the test shows that a small strength removes most of the low-pressure harm.

The low-cost schedule. The same factorial, run first for the repeated `lowC` schedule (Appendix L), detects both channels in the pooled family ($\bar{I}_H = +0.99\%$, Holm 0.015; $\bar{I}_R = +1.24\%$, Holm 0.0014) at about one percent of baseline return, with one per-task detection (`cheetah-run` $\bar{I}_R$, Holm 0.016) and none at the endpoint.

7 Within a regime: does the cost of a single reset order its value?

The comparisons so far differ in pressure. Within the *same* learner state and pressure, does the immediate cost of a reset order its value? A useful low-cost selector would make the relation negative. We test this with a single reset on fresh seeds (eight per task, UTD 1 and 8). At each checkpoint we measure the immediate cost of 100 random 10% masks, select the masks at cost ranks 10, 30, 50, 70 and 90 and one pair of equal-cost, nearly disjoint masks, apply each once, and train on for the full budget against `sham` from the same checkpoint (Figure 5); the cheapest and costliest masks and `sham` are also continued under three independent post-reset random streams. The estimand is

the within-checkpoint Spearman correlation between cost and value (AUC), averaged over the four task×pressure cells, under a cell-stratified two-sided sign-flip test.

Result. The cost–value relation depends on the condition. On `walker-walk` it is negative at UTD 1 (-0.35, seed-blocked CI $[-0.57, -0.10]$), the direction a useful selector would produce, and positive at UTD 8 ($+0.27$, $[+0.02, +0.51]$); the cost × UTD interaction is $+0.62$ (CI $[+0.25, +0.97]$). On `cheetah-run` it is weakly positive at both pressures ($+0.20$, $+0.11$). Pooled over the four cells the mean is $+0.06$ (stratified CI $[-0.09, +0.19]$, $p = 0.47$): a null that averages opposite-signed relations rather than showing their absence. The random masks span the interior of the cost range rather than the selectors' extremes, and a single reset's value varies at least as much across continuation streams as across cost ranks, so the test excludes a strong monotonic relation, not a weak one (Appendix Q). A policy-facing damage measure (the change of the actor-loss gradient across the reset) is only weakly related to d_{TD} and orders value no better within a regime (Appendix S).

Repeated schedules. On the repeated schedule of Section 5 the same checkpoints and masks give mean correlation -0.00 (stratified CI $[-0.13, +0.12]$, $p = 1.00$; V_{reg}: -0.09, $[-0.19, +0.02]$), with the same walker-at-UTD-1 cell negative (-0.19) and the others near zero (Appendix Q). The two tests agree: no consistent cost–value ordering holds across the tested regimes; the relation is condition-dependent and, inside the cost range, weak.

8 Limitations

All results are for critic-only resets of state-based SAC (one TD3 replication) at two UTD levels (four on one task); the calibration, the factorials and the single-reset test rest on `walker-walk` and `cheetah-run`, and only the pressure moderation extends to six tasks. The repeated arms are a near-no-op (`lowC`, `lowA`), a random intermediate-cost reset (`randK`) and a destructive extreme (`highC`); threshold-based selectors are untested. Whether the budget decides *which* selection is preferable is open: the ranking of the random and low-cost rules changed between UTD regimes on one cohort but was not reproduced at a fixed history on another (Appendix P). Within these bounds, the sign of the `highC` value at high pressure is task-dependent, the history contrast is detected on one of two tasks, and replay age remains confounded with update count at matched updates (Appendix O).

9 Conclusion

Immediate disruption was measurable, and its downstream consequence depended on the recovery budget. The costliest reset was far less harmful under high pressure on all six tasks and with TD3, a random reset of intermediate cost was moderated in the same direction, and the effect was graded over four pressure levels on one task; a factorial attributed most of the reversal to the post-reset budget, which for a single reset was consistent with update supply; and no consistent cost–value ordering held across the tested regimes. Immediate cost orders the magnitude of outcomes at the extremes; the learning conditions after the reset set that magnitude and, on some tasks, the sign (at high pressure the costliest reset is still harmful on two of six tasks). **For practitioners:** among the resets tested, the post-reset update budget strongly moderated the size of the harm, the ranking of rules differed between UTD regimes, and weakening the reset removed most low-budget harm even for high-cost selections. **For evaluating reset methods:** state the post-reset update budget and horizon, include a no-op arm under each condition, and report outcomes at matched steps and matched updates. A cost estimator settles which reset is cheap, not which one is worth making.

Reproducibility statement

Every confirmatory experiment was pre-registered before it was run: the design, the estimand, the decision rule and the failure framing were fixed in advance, and the adjudication scripts were written before any run. The calibration and the confirmatory replication pre-committed the headline sentence for each possible outcome. Amendments are separate timestamped documents that never modify a frozen file. All figures and all numbers in the text are produced by frozen scripts from the raw run logs and checked against the text by an automated consistency script; none is transcribed by hand. Runs use fixed seeds, and the branching design saves and reloads the complete state—networks, optimizers, replay buffer and every RNG stream—so an arm can be re-derived exactly from its checkpoint. Seed

lineage, and which experiments share checkpoints, are listed in Appendix U. An anonymized artifact with the code, every pre-registration and amendment, the adjudication scripts and their decision outputs is provided as supplementary material; raw logs and checkpoints will be released with the code.

USE OF AI ASSISTANCE

AI assistants were used under the authors' direction to assist with grammatical editing, literature searches, and debugging selected portions of code. All study designs, decision rules, statistical standards, and scientific claims were established, reviewed, and approved by the authors. The authors also reviewed and approved the final manuscript and take full responsibility for its content.

REFERENCES

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2023.

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

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tudor Berariu, Wojciech Czarnecki, Soham De, Jörg Bornschein, Samuel L. Smith, Razvan Pascanu, and Claudia Clopath. A study on the plasticity of neural networks. *arXiv preprint arXiv:2106.00042*, 2021.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double Q-learning: Learning fast without a model. In *International Conference on Learning Representations (ICLR)*, 2021.

Roger Creus Castanyer, Johan Obando-Ceron, Lu Li, Pierre-Luc Bacon, Glen Berseth, Aaron Courville, and Pablo Samuel Castro. Stable gradients for stable learning at scale in deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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, 2024.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *International Conference on Learning Representations (ICLR)*, 2023.

Patrick Elisii, Lucas Beauchemin, and Dawer Jamshed. Attribution-based neuron utility for plasticity restoration in deep networks. *arXiv preprint arXiv:2605.06834*, 2026.

Mohamed Elsayed and A. Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Takuya Hiraoka, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, and Yoshimasa Tsuruoka. Dropout Q-functions for doubly efficient reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Marcel Hussing, Claas Voelcker, Igor Gilitschenski, Amir-massoud Farahmand, and Eric Eaton. Dissecting deep RL with high update ratios: Combatting value divergence. In *Reinforcement Learning Conference (RLC)*, 2024.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Arthur Juliani and Jordan T. Ash. A study of plasticity loss in on-policy deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Timo Klein, Christoph Luther, Manus McAuliffe, Lukas Miklautz, Claudia Plant, and Sebastian Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832v3*, 2024.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Hojoon Lee, Hanseul Cho, Hyunseung Kim, Daehoon Gwak, Joonkee Kim, Jaegul Choo, Se-Young Yun, and Chulhee Yun. PLASTIC: Improving input and label plasticity for sample efficient reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiashun Liu, Zihao Wu, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Ling Pan. Measure gradients, not activations! enhancing neuronal activity in deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2024.

Guozheng Ma, Lu Li, Sen Zhang, Zixuan Liu, Zhen Wang, Yixin Chen, Li Shen, Xueqian Wang, and Dacheng Tao. Revisiting plasticity in visual reinforcement learning: Data, modules and training stages. In *International Conference on Learning Representations (ICLR)*, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Michal Nauman, Michał Bortkiewicz, Piotr Miłoś, Tomasz Trzciński, Mateusz Ostaszewski, and Marek Cygan. Overestimation, overfitting, and plasticity in actor-critic: the bitter lesson of reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Miłoś, and Marek Cygan. Bigger, regularized, optimistic: scaling for compute and sample-efficient continuous control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. Deep reinforcement learning with plasticity injection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Max Schwarzer, Johan Samir Obando Ceron, Aaron Courville, Marc G. Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level Atari with human-level efficiency. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Hongyao Tang and Glen Berseth. Improving deep reinforcement learning by reducing the chain effect of value and policy churn. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*, 2018.

Zihao Wu, Hongyao Tang, Yi Ma, Jiashun Liu, Yan Zheng, and Jianye Hao. The rank and gradient lost in non-stationarity: Sample weight decay for mitigating plasticity loss in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Guowei Xu, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, Shuzhen Li, Yanjie Ze, Hal Daumé III, Furong Huang, and Huazhe Xu. DrM: Mastering visual reinforcement learning through dormant ratio minimization. In *International Conference on Learning Representations (ICLR)*, 2024.

Huaicheng Zhou, Zifeng Zhuang, and Donglin Wang. Stay hungry, keep learning: Sustainable plasticity for deep reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

A PREDICTIVE CALIBRATION OF THE SCORE ON RANDOM MASKS

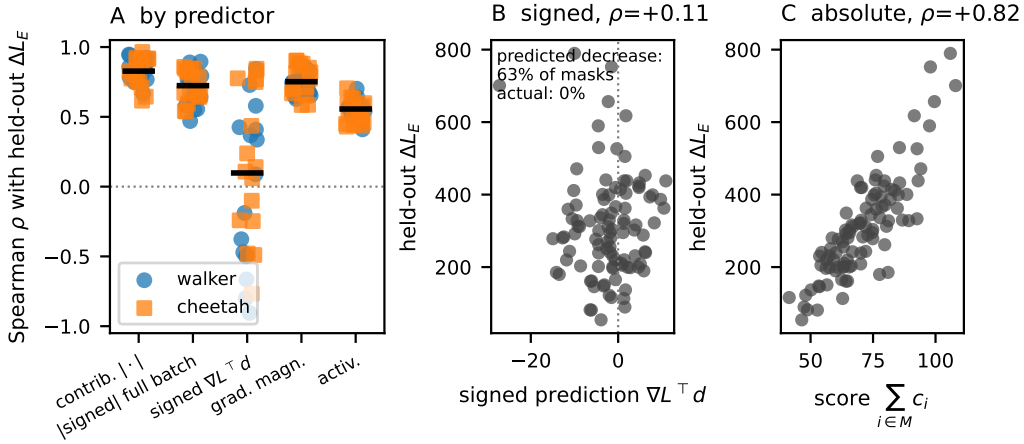


Figure 6: Predictive calibration on 32 checkpoints (two tasks, two pressures, eight seeds each), 100 random 10% masks per checkpoint, predictors computed on a scoring batch and the loss change measured on a disjoint held-out batch against a fixed target. **A** per-checkpoint Spearman correlation with the held-out loss change; bars are medians. **B, C** the checkpoint with the median signed-predictor ρ: the signed first-order prediction is two orders of magnitude smaller than the actual change and predicts a loss *decrease* for 63% of masks although no mask decreases the loss, while the absolute score orders the masks.

The score c_i takes the absolute value of a per-unit first-order term, so it is not the signed first-order change of the loss under the full reset, $\nabla_\theta L^\top d$. To test whether that matters, on 32 checkpoints (both tasks, both pressures) we draw 100 random masks of 10% of units per layer, compute each predictor on the scoring batch, apply the reset, and measure the loss change on a disjoint held-out batch (Figure 6).

The score ranks the held-out loss change at median Spearman $\rho = 0.83$ over the 32 checkpoints (IQR $[0.77, 0.91]$; bootstrap CI of the mean $[0.80, 0.86]$). The signed first-order prediction does not: $\rho = 0.10$ (IQR $[-0.47, 0.47]$, CI $[-0.13, 0.25]$), worse than the absolute score on $32/32$ checkpoints and no better in-sample ($\rho = 0.09$), so the failure is one of linearization, not of batch generalization. It predicts a loss *decrease* for a median 28% of masks although no mask decreases the loss: a hard reset is a large displacement, and the absolute term behaves as a proxy for the magnitude of a unit's influence, not as a first-order cost. Gradient magnitude alone ranks nearly as well (0.75, CI $[0.72, 0.79]$) and activation less well (0.56, CI $[0.52, 0.57]$). We therefore describe c_i as an *absolute first-order sensitivity*, not a first-order cost: its level is not calibrated (OLS slope ≈ 28 on the raw scale), its ordering is.

Appendix B compares c_i with activation, gradient magnitude, pruning saliency and continual-backprop utility on the same protocol, and shows that the selected set depends on the draw θ^0 (Jaccard 0.52 across draws) and that set costs are additive only for bottom-contribution sets (5–13$\times$ the sum for random and top sets, Appendix C).

B Alternative scores, draw dependence and set costs

Alternatives and draw dependence. The same protocol compares c_i with other unit scores (Table 5). Weight$\times$gradient pruning saliency $|\theta_i \cdot g_i|$ ranks nearly as well (mean Spearman 0.77 against 0.83), the continual-backprop utility (Dohare et al., 2024) and gradient magnitude less well (0.71, 0.75), activation worst (0.55). Because c_i is computed with the fresh draw θ^0 the operator will actually apply, the selected bottom-K set depends on that draw: across two independent draws the score's ranking correlates at Spearman 0.88 but the bottom-K sets overlap at only Jaccard 0.52. Replacing θ^0 by the expectation over eight draws that are *not* the one applied gives a draw-independent score that ranks held-out cost at least as well (0.86; better in $26/32$ checkpoints) and selects a set overlapping the draw-specific one at Jaccard 0.61. A practitioner therefore need not see the draw before selecting.

Set costs are not sums of unit costs. The score is a per-unit quantity, but a reset removes a set. On the same protocol we compared the loss change of resetting a set jointly with the sum of the single-unit changes, $\eta = (\Delta_{\text{joint}} - \Delta_{\text{add}})/|\Delta_{\text{joint}}|$. Joint reset costs were approximately additive for bottom-contribution masks (joint/sum ratio 1.05, median $\eta = +0.05$, IQR $[-0.05, +0.27]$) but strongly superadditive for random masks ($5\times$, $\eta = 0.81$) and for the top-contribution mask ($13\times$, $\eta = 0.92$, IQR $[0.91, 0.94]$). We interpret this as mask-level non-additivity, which may arise from network interactions and/or from aligned prediction perturbations amplified by the squared loss, rather than as direct evidence of neuronal compensation; the decomposition in Appendix C attributes most of it to the latter (at the output level the single-unit perturbations combine nearly additively, $\eta_Q \approx 0.1$–0.2), and shows that the signed first-order term is negligible for every kind of set. Whatever its origin, it means that a set's cost cannot be obtained by summing per-unit scores except where nothing important is removed, and that c_i is a *conditional* reset sensitivity—the cost of removing a unit while the rest of the network stays.

score	sees the draw	Spearman with held-out ΔL_E	bottom-K Jaccard vs. c_i	s/ckpt
c_i (registered)	yes	$0.83\ [0.80, 0.86]$	1.00	1.1
$\mathbb{E}_{\theta^0}\, c_i$ (8 draws)	no	$0.86\ [0.83, 0.88]$	0.61	8.3
$\lvert\theta_i \cdot g_i\rvert$ saliency	no	$0.77\ [0.72, 0.81]$	0.35	< 0.1
gradient magnitude	no	$0.75\ [0.72, 0.79]$	0.52	0.6
continual-backprop utility	no	$0.71\ [0.68, 0.74]$	0.49	< 0.1
activation (ReDo)	no	$0.55\ [0.52, 0.57]$	0.44	< 0.1

Table 5: Unit scores compared on the calibration protocol: mean over 32 checkpoints of the per-checkpoint Spearman correlation with the held-out loss change of 100 random masks (seed-level bootstrap CI in brackets), median overlap of the selected bottom-K set with the registered score's, and median scoring time per checkpoint (4 CPU threads).

C Decomposition of the joint reset cost

On the evaluation batch with its fixed target, the joint loss change of a mask S decomposes exactly as $\Delta_{\text{joint}} = F + R_{\text{self}} + R_{\text{cross}}$, with $F = g^\top d_S$ the signed first-order term of the whole reset, $R_{\text{self}} = \sum_{i \in S}(\Delta_i - g_i^\top d_i)$ the single-unit non-linearity and $R_{\text{cross}} = \Delta_{\text{joint}} - \sum_{i \in S} \Delta_i$ the within-mask interaction. At the level of the critic outputs, $\eta_Q = \|\delta Q_S - \sum_{i \in S} \delta Q_i\| / \|\delta Q_S\|$ over the 4096 held-out transitions measures whether the outputs themselves combine non-additively; a small η_Q with a large loss-level η indicates aligned output shifts amplified by the squared loss rather than parameter entanglement. Over the 32 checkpoints (medians, IQR in brackets), the signed first-order term is negligible for every kind of set (F/Δ_{joint}: 0.02 for the bottom-K set, 0.00 for random and top-K sets), which restates the calibration result. For the bottom-K set the change is single-unit non-linearity ($R_{\text{self}}/\Delta_{\text{joint}} = 0.92\ [0.73, 1.05]$) with a small cross term ($0.05\ [-0.05, 0.27]$). For random and top-K sets the cross term dominates ($0.80\ [0.71, 0.85]$ and $0.92\ [0.91, 0.94]$), yet at the output level the single-unit perturbations combine nearly additively ($\eta_Q = 0.09\ [0.08, 0.12]$ for random, $0.23\ [0.17, 0.30]$ for top-K; $0.49\ [0.38, 0.66]$ for bottom-K, whose perturbations are tiny). Most of the loss-level superadditivity of random and top-K sets is therefore aligned output shift amplified by the squared loss rather than parameter entanglement; a smaller, non-zero part of it ($\eta_Q \approx 0.1$–0.2) is genuine non-additivity of the outputs.

D Sources of run-to-run variance

Reset arms are often more dispersed across seeds than the untouched baseline, which suggests that a reset is partly a matter of chance: one reinitializes and hopes the draw is favourable. We tested this in the continuous-control cell by isolating, one at a time, the two sources of randomness that a reset method controls.

The initialization draw. Holding the checkpoint, the seed, the environment stream and the selection rule fixed and varying only the RNG that generates the fresh initialization, eight draws of the same seed move the outcome by CV 0.002–0.022, against an across-seed CV of 0.037–0.115. The variance ratio $\rho = \text{Var}_{\text{draw}}/(\text{Var}_{\text{draw}} + \text{Var}_{\text{seed}})$ is 0.018 at UTD 1 and 0.051 at UTD 8, against a pre-registered threshold of 0.2.

The selection boundary. The contribution score is near-deterministic across independent scoring batches (Spearman 0.98–0.99), yet the selected bottom-K set still has Jaccard distance 0.19–0.25 to it (Jaccard index 0.75–0.81; about 10–14% of the units exchanged at equal K), because a hard threshold on a near-continuous ranking exchanges units at the cutoff. Isolating the scoring batch in the same way gives $\rho = 0.007$ and 0.020: the churn is among units whose contribution is near-tied and which are therefore functionally interchangeable.

Scope. In the tested cell, neither the draw nor the bottom-K boundary is a major source of outcome variance. It does not follow that the variance is irreducible: a fully deterministic rule can still increase dispersion if its effect depends on the pre-reset learner state, and that channel is untested here.

E ORIGIN OF THE PRESSURE HYPOTHESIS, AND THE REPEATED LOW-COST ARM

The prediction tested in Section 5 originated in discrete control. Figure 7 shows a DQN (Mnih et al., 2015) Atari sweep of the same variable at matched seeds, comparing binary top-K recycling with a contribution-scaled soft reset. At UTD 1–2 the paired differences straddle zero ($4/8$ and $2/5$ positive); at UTD 4–8 on Phoenix, and on Asterix at UTD 4, they are positive in 15 of 16 matched pairs (seed-clustered permutation $p = 0.016$; cell-clustered $p = 0.125$, limited by the three independent high-pressure cells). The continuous-control experiments in this paper were designed to ask what this regularity consists of, on a different algorithm family and with the two channels of pressure separated.

On `walker-walk` the pre-registered `lowC` difference-in-differences was $I = +8.2$ under the registered metric (CI $[+0.3, +15.8]$, $7/8$ seeds, one-sided sign test $p = 0.035$; on the AUC $+8.0$, CI $[-0.8, +16.4]$, $6/8$).

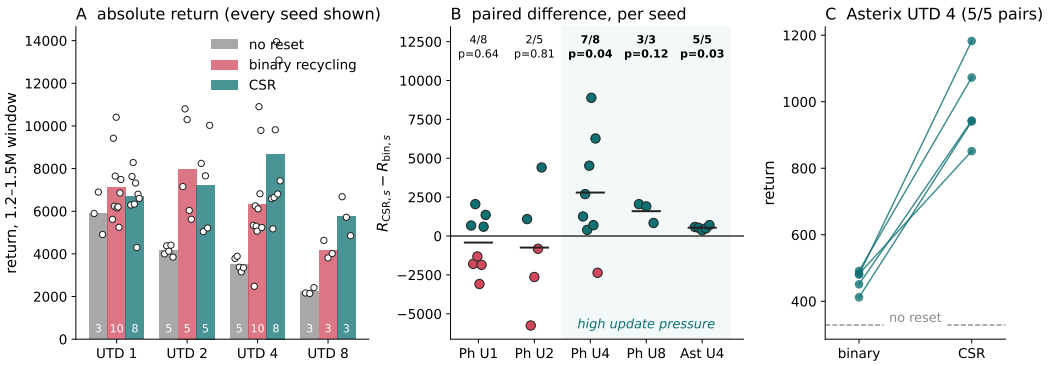


Figure 7: The DQN–Atari sweep that motivated the pressure hypothesis (CSR in the legend is the contribution-selective reset, the `lowC` rule; "binary recycling" is the hard reset). **A** absolute return, all seeds shown. **B** paired differences per cell with within-cell sign tests. **C** Asterix at UTD 4.

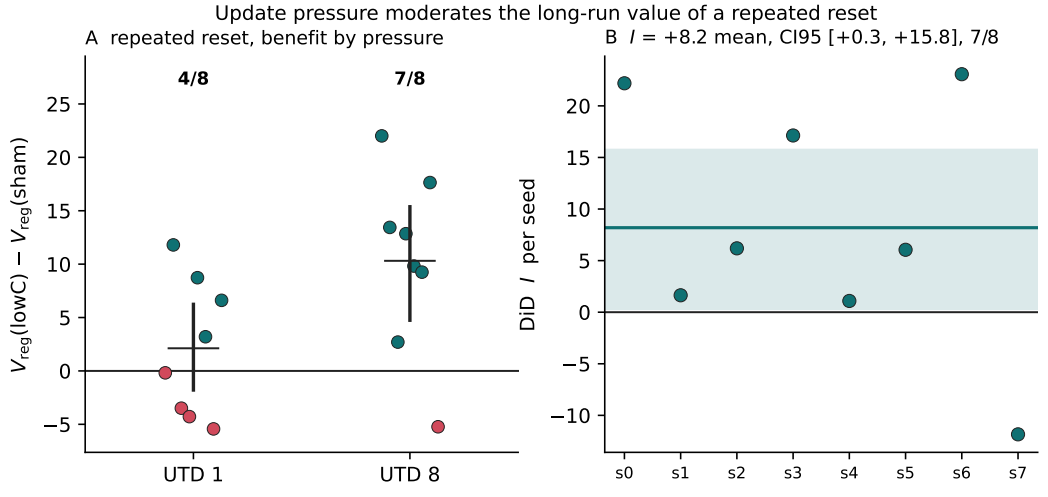


Figure 8: Repeated `lowC` reset every 10k steps. **A** V_{reg} benefit relative to the paired `sham` arm at each pressure, all seeds shown, bars are means with bootstrap CIs. **B** per-seed difference-in-differences.

Figure 8 shows the repeated `lowC` arm of Section 5 on the original seeds.

F CONFIRMATORY REPLICATION OF THE HIGH-CONTRIBUTION REVERSAL ON THE FIRST TWO TASKS

An exploratory observation. An arm added after this outcome was observed, run on fresh seeds (8–15) and reported as non-confirmatory, suggested a sharper statement. Repeated `highC`, the most expensive selection, was harmful at UTD 1 (-175.2, $0/8$ seeds positive) and, on `walker-walk`, beneficial at UTD 8 ($+14.2$, $8/8$), with about three times the cumulative critic disruption at UTD 8. Because that arm was added after inspecting the data, we pre-registered its pressure dependence as a confirmatory hypothesis—derived from the exploratory result, which is stated as such—and tested it prospectively.

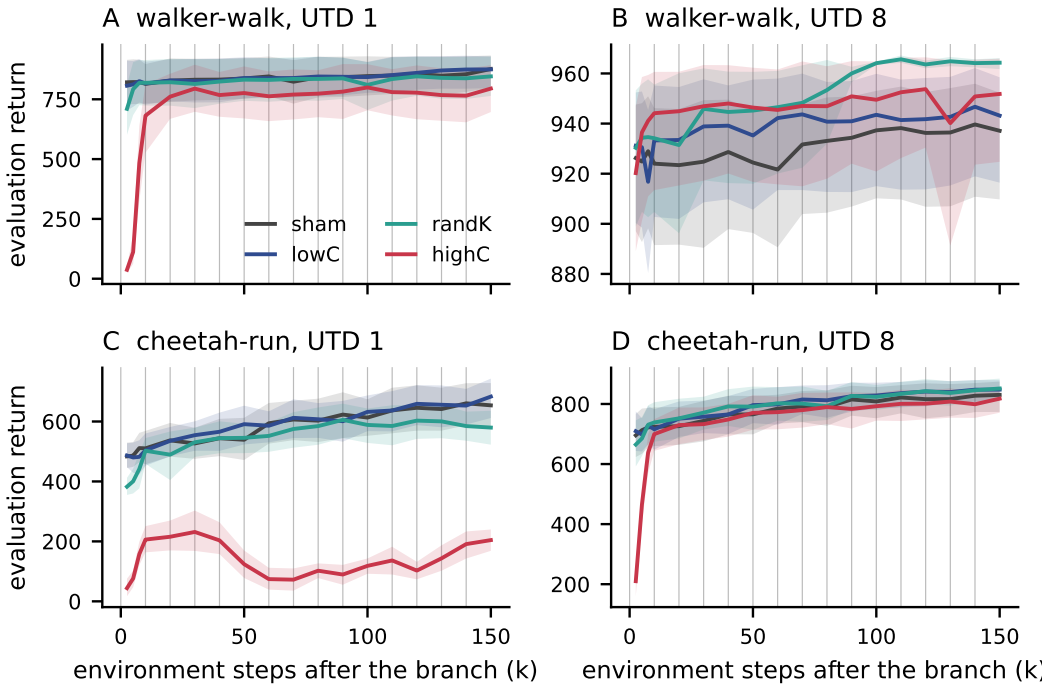


Figure 9: Evaluation return after the branch for the repeated `sham`, `lowC`, `randK` and `highC` schedules (means over eight seeds with seed-blocked 95% bands; vertical lines mark the reset events every 10k steps). At UTD 1 the costly resets collapse the return and recover slowly or not at all within the budget; at UTD 8 the same resets recover within one or two intervals.

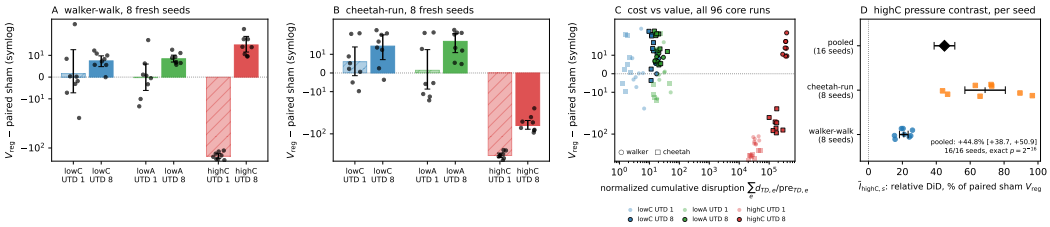


Figure 10: Confirmatory replication on fresh seeds, on the registered V_{reg} (the AUC values of the same contrasts are in the text and Table 17). **A, B** V_{reg} relative to the paired `sham` arm for repeated `lowC`, `lowA` and `highC` at UTD 1 (hatched) and UTD 8, every seed shown, bars are means with seed-blocked bootstrap CIs (symmetric-log axis). **C** normalized cumulative critic disruption $\sum_e d_{TD,e}/\text{pre}_{TD,e}$ against benefit for all 96 core runs: the runs with the largest cumulative cost (`highC` at UTD 8) are not the runs with the worst outcome. **D** the confirmatory quantity: the per-seed relative pressure contrast $\tilde{I}_{\texttt{highC},s} = \tilde{\Delta}^{\texttt{highC}}_{8,s} - \tilde{\Delta}^{\texttt{highC}}_{1,s}$, where $\tilde{\Delta}$ divides the paired benefit by the mean `sham` V_{reg} of the same task and pressure, shown for all eight seeds of each task with the per-task mean and seed-blocked bootstrap CI, and the pooled sixteen-seed estimate (diamond, stratified CI).

Design and confirmatory result. On sixteen new seeds (40–47 on `walker-walk`, 50–57 on `cheetah-run`), disjoint from every seed used elsewhere, we ran `sham`, `lowC`, `highC` and `lowA` at UTD 1 and UTD 8 with the same schedule and the same fresh draw (Figure 10). The registered primary estimand is the seed-level difference-in-differences of `highC` between pressures, $I_{\texttt{highC},s} = \Delta^{\texttt{highC}}_{8,s} - \Delta^{\texttt{highC}}_{1,s}$. Because the tasks differ in return scale, the pooled analysis uses the relative benefit $\tilde{\Delta}^{a}_{u,s} = \Delta^{a}_{u,s} \big/ \frac{1}{8}\sum_{s'} V_{\texttt{sham},u,s'}$ (V the value definition in use, V_{reg} or AUC), whose denominator is the mean `sham` AUC of the same task and pressure over its eight seeds (a constant within a task–pressure cell), and $\tilde{I}_{\texttt{highC},s} = \tilde{\Delta}^{\texttt{highC}}_{8,s} - \tilde{\Delta}^{\texttt{highC}}_{1,s}$, expressed as a percentage of baseline return. **The confirmatory family is a single test**: the mean of the two per-task means of $\tilde{I}_{\texttt{highC},s}$, pooled over sixteen seeds, under an exhaustive one-sided sign-flip test with signs flipped within task (2^{16} assignments). Per-task tests are secondary. The pooled estimate is $\bar{\tilde{I}}_{\texttt{highC}} = +44.8\%$ of baseline return (V_{reg}; seeds 40–47 and 50–57) (stratified bootstrap CI $[+38.7\%, +50.9\%]$), positive in $16/16$ seeds, exact $p = 2^{-16} \approx 1.5 \times 10^{-5}$, the smallest value the test can return (this V_{reg} value coincides numerically with the AUC diagonal contrast of Section 6, a different quantity); on the AUC, $\bar{\tilde{I}}_{\texttt{highC}} = +43.1\%$ of baseline return (stratified CI $[+36.9\%, +49.5\%]$). Here the registered construction resamples the eight seeds within each task and averages the task means, with the `sham` denominator held fixed; re-estimating it inside every resample gives $[+39.7\%, +46.8\%]$. This two-task pooled contrast is subsidiary to the six-task headline of Section 5. In raw V_{reg} units the per-task contrasts are $+176.4$ on `walker-walk` (CI $[+154.8, +197.6]$) and $+383.8$ on `cheetah-run` (CI $[+313.9, +453.9]$), $8/8$ seeds each (secondary exact $p = 2^{-8}$ each). The normalized cumulative critic disruption of the `highC` schedule is $10\times$ (walker) and $6\times$ (cheetah) larger at UTD 8 than at UTD 1, and UTD 8 is in both cases the far better outcome. In this matched comparison, cumulative local cost did not order the outcomes; pressure did.

The registered secondary quantities qualify the sign. On `walker-walk`, `highC` is harmful at UTD 1 (-161.6, CI $[-183.1, -139.8]$, $0/8$) and beneficial at UTD 8 ($+14.8$, CI $[+11.5, +18.7]$, $8/8$; the exploratory cohort gave $+14.2$), where it also beats the low-cost arm (`highC` − `lowC` $= +7.3$, $8/8$). On `cheetah-run` it remains harmful at UTD 8 (-56.5, CI $[-72.8, -40.1]$, $0/8$; -7% of baseline) and is merely an order of magnitude less harmful than at UTD 1 (-440.3, -76% of baseline). What replicates across tasks is therefore the moderation of value by pressure; whether the most expensive reset is net beneficial at high pressure is task-dependent. The low-cost arms are pressure-moderated in the same direction: `lowC` gives $+7.5$ (walker, $7/8$) and $+14.4$ (cheetah, $7/8$) at UTD 8, and `lowA` $+8.4$ and $+16.9$ ($8/8$ each), and no difference between the activation and contribution arms is detected at either pressure (paired CIs spanning zero).

Small-k meta-analysis sensitivity. With six tasks and $I^2 = 0.95$ (AUC) the DerSimonian–Laird interval of the across-task `highC` contrast is optimistic. On the AUC, a REML estimate of the

between-task variance ($\tau^2 = 0.033$) gives a mean of $+31.8\%$ with a Hartung–Knapp t_5 interval of $[+9.8\%, +53.8\%]$; for `lowC` the corresponding interval is $[-0.1\%, +0.5\%]$. Under the registered V_{reg} ($I^2 = 0.93$; REML $\tau^2 = 0.028$ against 0.020 for DerSimonian–Laird) the `highC` mean is $+34.5\%$ with Hartung–Knapp interval $[+15.8\%, +53.2\%]$, and the `lowC` interval is $[-0.1\%, +0.8\%]$. The confirmatory verdict is the exact task-level sign test ($6/6$ tasks, $p = 1/64$ for each arm), which these sensitivity intervals do not touch.

G Replication of the pressure reversal with TD3

Pre-registered before any run. TD3 (Fujimoto et al., 2018) was implemented with the same twin-Q critic module as SAC (two hidden layers of 256 ReLU units, Adam 3×10^{-4}, $\gamma = 0.99$, $\tau = 0.005$), a deterministic tanh actor with Gaussian exploration noise ($\sigma = 0.1$), target-policy smoothing ($\sigma = 0.2$, clip 0.5) and delayed actor and target updates (every second critic update); the SAC code path is byte-identical to the version used for every other experiment. Scores, operator, audit batch, cost and value are unchanged. `walker-walk`, UTD $\in \{1, 8\}$, seeds 120–125 (the seed numbers coincide with SAC trunks of Appendix I; the TD3 runs share no checkpoint, network or trajectory with them and are new trunks), trunk 150k and branch 150k steps, arms `sham`, `lowC`, `highC` repeated every 10k with $K = 10\%$. Primary: $\bar{\tilde{I}}_{\texttt{highC}} > 0$, exact 2^6 sign-flip, one-sided. Table 6 gives the result under both value definitions.

value	arm	$\tilde{\Delta}_1$	$\tilde{\Delta}_8$	$\bar{\tilde{I}}$ [CI], positive seeds	exact p	cum. cost UTD 8 / UTD 1
V_{reg}	`highC`	-16.3%	$+3.3\%$	$+19.6\%$ $[+14.4, +25.8]$, $6/6$	0.0156	$3.86\times$
V_{reg}	`lowC`	$+0.0\%$	$+1.9\%$	$+1.9\%$ $[+1.0, +3.5]$, $6/6$	0.0156	$4.06\times$
AUC	`highC`	-8.4%	$+3.4\%$	$+11.8\%$ $[+7.8, +16.7]$, $6/6$	0.0156	
AUC	`lowC`	$+0.1\%$	$+2.0\%$	$+1.9\%$ $[+0.9, +3.7]$, $6/6$	0.0156	

Table 6: TD3 on `walker-walk`, six seeds per UTD. Benefits and pressure contrasts on the relative scale (% of the `sham` mean of the UTD cell), seed-blocked bootstrap CIs, exact sign-flip p (minimum attainable $1/64$). The immediate cost at event 0 orders as in SAC: median d_{TD} 0 (`sham`), 21.7 (`lowC`), 18170 (`highC`).

H Recovery-budget-aware reset strength

Pre-registered before any run; not a proposed method and not tuned. That a small reset strength is safe is already known from shrink-and-perturb and soft resets (Ash & Adams, 2020; Schwarzer et al., 2023); what is new here is only that the harm removed scales with the post-reset budget, and we do not claim that budget-aware scaling beats a constant small α. The rule $\alpha = \min(1, \text{UTD}/8)$ sets the soft-reset strength ($W_{\text{in}} \leftarrow (1-\alpha)W_{\text{in}} + \alpha\theta^0$, bias, outgoing column and Adam moments $\leftarrow (1-\alpha)\times$) from the recovery UTD. Design: the HC cohort (`walker-walk` seeds 40–47, `cheetah-run` 50–57, UTD 1 and 8), reusing the stored `sham` and `highC` ($\alpha = 1$) cells (re-run check, Appendix U) and adding `highC` with $\alpha = 0.125$ at both pressures (32 runs). Primary: at UTD 1, $\Delta_{0.125} - \Delta_1 > 0$ pooled over 16 seeds on the relative scale, exact 2^{16} sign-flip with within-task flips, one-sided; secondary at UTD 8, two-sided, descriptive. Results (registered V_{reg}; AUC in parentheses): primary $+46.0\%$ (CI $[+32.1\%, +60.6\%]$, $16/16$, $p = 2 \times 10^{-5}$; AUC $+43.1\%$), secondary -2.8% (CI $[-5.3\%, -0.5\%]$, $8/16$, $p = 0.040$; AUC -1.0%, $p = 0.22$). Cumulative critic cost of the $\alpha = 0.125$ schedule is 15–17$\times$ smaller than at $\alpha = 1$ in every cell.

	walker-walk			cheetah-run		
UTD	sham V	highC $\alpha = 1$	highC $\alpha = 0.125$	sham V	highC $\alpha = 1$	highC $\alpha = 0.125$
1	839	-162	$\mathbf{+2}$	579	-440	$\mathbf{-20}$
8	931	$\mathbf{+15}$	$+2$	775	$\mathbf{-57}$	-2

Table 7: Benefit Δ (registered V_{reg}, raw return, eight seeds per cell) of the repeated highC schedule at two reset strengths; bold marks the cell the rule $\alpha = \min(1, \text{UTD}/8)$ selects. Scaling by the budget avoids the low-pressure harm; at high pressure the difference full minus scaled, $\Delta_1 - \Delta_{0.125}$, is -1.0% (AUC, not detected) and -2.8% (V_{reg}, $p = 0.040$), i.e. in favour of the scaled reset, so the full reset is never detectably better.

I Redistributing a fixed update budget after a reset

If the recovery channel is a matter of update supply, a low-pressure learner should benefit from *moving* updates to right after a low-cost reset without adding any. We pre-registered a compute-matched 2×2 on fresh seeds (walker-walk 120–127, cheetah-run 130–137, base UTD 1; shared pre-intervention checkpoints, fresh downstream randomness and outcomes): {sham, lowC} $\times$ {uniform schedule, recovery burst}, where the burst performs 8 updates per step for $L = 1000$ steps after each 10k-step event and the remaining steps of the interval receive proportionally fewer, so every interval and every arm receives exactly 10k updates (verified from the logged update counts). L was chosen on four development seeds by a frozen rule; the three candidate values (250, 500, 1000) were indistinguishable there. The primary estimand, the seed-level interaction $(D - B) - (C - A)$ (reset with burst minus burst alone, net of reset alone minus nothing), pooled over the sixteen seeds under a one-sided within-task sign-flip test, was not detected: $+0.93\%$ of baseline return, CI $[-0.05\%, +1.89\%]$, $11/16$ seeds positive, $p = 0.053$ (walker-walk $+2.8$, $5/8$; cheetah-run $+9.0$, $6/8$). The method arm did not beat no reset ($D - A$: -0.1 on walker-walk, -6.6 on cheetah-run) nor an activation-based baseline on the uniform schedule ($D - \text{lowA}$: -3.4, -9.9), and on cheetah-run the burst schedule by itself was harmful ($B - A = -15.9$). Under the pre-registered promotion rule this is a negative method result: knowing that recovery supply moderates reset value does not, by itself, say how to redistribute a fixed update budget to exploit it.

J A random intermediate-cost repeated reset

Pre-registered before any run. The arm randK resets $K = 10\%$ of units per critic hidden layer drawn uniformly at random at every event of the repeated schedule (own random stream; scores computed as in every arm; operator, strength, fresh draw and schedule unchanged). It runs on the confirmatory cohort of Section 5 (walker-walk seeds 40–47, cheetah-run 50–57, UTD 1 and 8), reusing the stored sham, lowC, lowA and highC cells (re-run check, Appendix U). The registered test is the pooled relative pressure contrast $\tilde{I}_{\text{randK}} = \tilde{\Delta}_{8,s} - \tilde{\Delta}_{1,s}$ over sixteen seeds under a one-sided exact sign-flip test with within-task flips. Pooled: AUC $+5.79\%$ (CI $[+4.15\%, +7.58\%]$, $16/16$, exact $p = 1.5e-05$); V_{reg} $+6.33\%$ ($[+4.51\%, +8.32\%]$, $16/16$). Median per-event d_{TD} of randK is 1,133 / 1,507 (walker-walk, UTD 1 / 8) and 323 / 1,064 (cheetah-run), against highC 7,146 / 35,844 and 2,914 / 17,280 and lowC at most 1 in the same runs (the repeated lowC re-selects units already reset, Appendix R).

	UTD 1: Δ			UTD 8: Δ			
value, task	lowC	randK [CI]	highC	lowC	randK [CI]	highC	$\tilde{I}_{\text{randK}}$ (%) [CI], pos
AUC, walker-walk	+2.8	−13.5 [−22.6, −5.7]	−102.7	+8.6	+20.9 [+12.4, +34.1]	+15.8	+3.8 [+2.6, +5.3], 8/8
AUC, cheetah-run	+8.1	−33.4 [−50.0, −18.4]	−449.2	+16.7	+16.2 [+8.4, +26.1]	−30.0	+7.7 [+5.2, +10.2], 8/8
V_{reg}, walker-walk	+1.6	−17.2 [−27.9, −9.1]	−161.6	+7.5	+19.1 [+11.8, +30.8]	+14.8	+4.1 [+2.8, +5.6], 8/8
V_{reg}, cheetah-run	+6.0	−40.1 [−58.5, −22.7]	−440.3	+14.4	+12.6 [+5.0, +21.8]	−56.5	+8.6 [+5.8, +11.3], 8/8

Table 8: Random-K repeated reset: cell means of the benefit (raw return) beside the stored low- and high-cost arms, and the per-task relative pressure contrast with seed-blocked bootstrap CIs. At UTD 1 the ordering lowC > randK > highC holds on both tasks (randK−lowC −16.3 [−37.0, −2.8] and −41.5 [−56.2, −27.2]); at UTD 8 randK is better than lowC on walker-walk and not detectably different on cheetah-run (+12.2 [+4.4, +25.2]; −0.4 [−5.6, +4.8]), and better than highC on cheetah-run but not detectably on walker-walk (+5.1 [−3.4, +19.5]; +46.2 [+40.5, +52.1]).

K Dose–response of the pressure effect in UTD

Pre-registered before any run. Fresh walker-walk trunks at UTD 2 and 4 (seeds 140–147, eight per level; 150k steps, complete state saved) with the repeated sham, lowC and highC schedule of Section 5 branched from each, combined with the UTD 1 and 8 cells of the confirmatory cohort (seeds 40–47); the four levels therefore come from two seed sets. The registered statistic is the Spearman correlation between $\log_2$ UTD and the level means of the relative highC benefit, tested by exact permutation over the 4! orderings (one-sided; minimum $p = 1/24$).

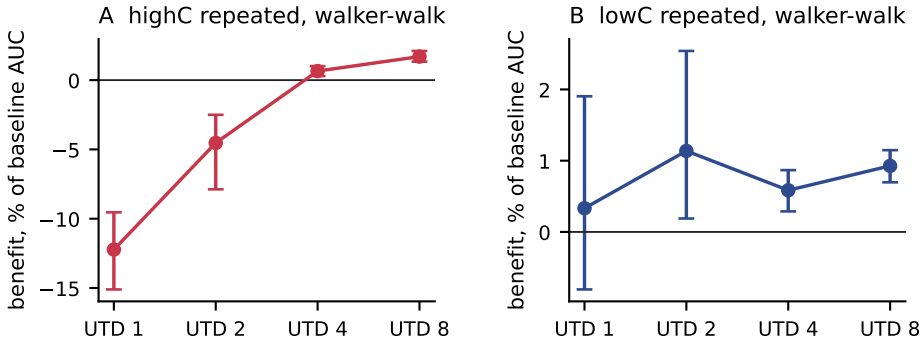


Figure 11: Relative benefit of the repeated highC (**A**) and lowC (**B**) schedules on walker-walk at four update-to-data ratios (AUC; means with seed-blocked bootstrap CIs, eight seeds per level; Spearman ρ with $\log_2$ UTD and exact permutation p in Table 9).

arm	value	UTD 1	UTD 2	UTD 4	UTD 8	ρ	p
highC	AUC	−12.2 [−15.1, −9.5], 0/8	−4.5 [−7.9, −2.5], 0/8	+0.7 [+0.3, +1.0], 7/8	+1.7 [+1.3, +2.1], 8/8	+1.00	0.042
highC	V_{reg}	−19.3 [−21.8, −16.7], 0/8	−8.0 [−10.5, −6.4], 0/8	+0.2 [−0.3, +0.7], 5/8	+1.6 [+1.2, +2.0], 8/8	+1.00	0.042
lowC	AUC	+0.3 [−0.8, +1.9], 3/8	+1.1 [+0.2, +2.5], 6/8	+0.6 [+0.3, +0.9], 7/8	+0.9 [+0.7, +1.1], 8/8	+0.40	0.375
lowC	V_{reg}	+0.2 [−0.9, +1.5], 4/8	+1.0 [+0.1, +2.2], 6/8	+0.6 [+0.2, +0.9], 7/8	+0.8 [+0.5, +1.1], 7/8	+0.40	0.375

Table 9: Relative benefit (% of the level's mean sham value) with seed-blocked CIs and positive-seed counts, the Spearman correlation with $\log_2$ UTD and its exact permutation p-value. The highC benefit is monotone in UTD under both value definitions and crosses zero between UTD 2 and 4; lowC is not monotone and stays within about one percent. Cumulative normalized disruption of the highC schedule rises with UTD (1.6×10^5, 3.7×10^5, 5.9×10^5, 5.6×10^5 at UTD 1, 2, 4, 8).

L THE LOW-COST PRESSURE-SWAP FACTORIAL

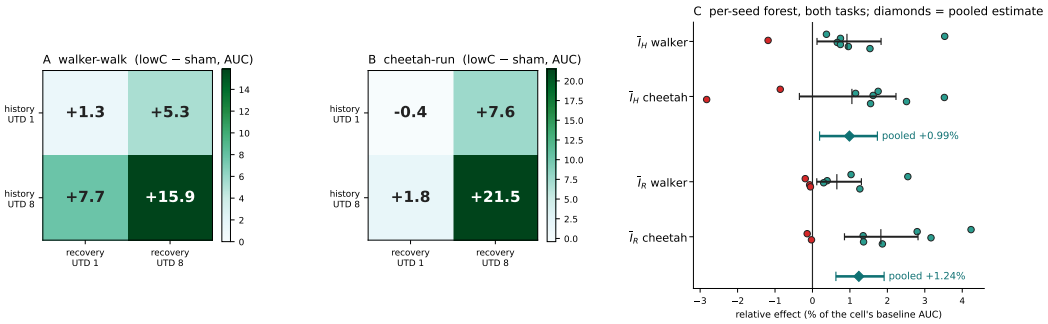


Figure 12: History × recovery factorial for the repeated `lowC` schedule on `walker-walk` and `cheetah-run`, eight fresh seeds each, on the time-normalized AUC. **A,B** absolute cell means of the `lowC` benefit per task. **C** per-seed forest plot on the relative scale with the pooled estimates as diamonds ($\bar{I}_H = +0.99\%$ $[+0.19, +1.73]$, $\bar{I}_R = +1.24\%$ $[+0.63, +1.91]$; the registered V_{reg} statistics are below).

This factorial preceded the `highC` one of Section 6 and used the same design with the repeated `lowC` schedule against `sham` in every cell, on `walker-walk` (seeds 16–23) and `cheetah-run` (seeds 30–37), all fresh at the time (Figure 12). Estimands are formed at seed level and aggregated once. With $\Delta_{h,r,s} = V_{\texttt{lowC},h,r,s} - V_{\texttt{sham},h,r,s}$,

$$I_{H,s} = \tfrac{1}{2}\big[(\Delta_{8,1,s} - \Delta_{1,1,s}) + (\Delta_{8,8,s} - \Delta_{1,8,s})\big], \quad I_{R,s} = \tfrac{1}{2}\big[(\Delta_{1,8,s} - \Delta_{1,1,s}) + (\Delta_{8,8,s} - \Delta_{8,1,s})\big].$$

Because the tasks differ in return scale, each contrast is placed on a relative scale,

$$\tilde{\Delta}_{h,r,s} = \frac{\Delta_{h,r,s}}{\frac{1}{|S_{\mathcal{T}}|}\sum_{s' \in S_{\mathcal{T}}} V_{\texttt{sham},h,r,s'}},$$

whose denominator is the mean `sham` return of that task and cell over its own seeds (a constant within a task–cell, adding no per-seed noise); $\tilde{I}_{H,s}, \tilde{I}_{R,s}$ are formed from $\tilde{\Delta}$ as above. The confirmatory family is the pooled pair $\{\bar{\tilde{I}}_H, \bar{\tilde{I}}_R\}$ under an *exhaustive paired sign-flip test under within-task exchangeability*: all 2^{16} sign assignments are enumerated with the mean of the two per-task means as the statistic, Holm-corrected over the pair. Intervals are stratified paired bootstrap intervals that resample seeds within task; per-task intervals are seed-blocked. The second task and the pooled family were fixed after the first task had been analysed.

Result. Pooled over the sixteen seeds, both average marginal effects on the trajectory-integrated value are detected: $\bar{I}_H = +0.89\%$ of baseline return (stratified CI_{95} $[+0.21\%, +1.53\%]$, $13/16$ seeds, exhaustive $p = 0.012$, Holm-adjusted 0.012) and $\bar{I}_R = +1.13\%$ (CI_{95} $[+0.48\%, +1.76\%]$, $12/16$, $p = 0.0033$, Holm-adjusted 0.0067). Per task: `walker-walk` cells $+0.7 \to +5.2 \to +7.2 \to +14.6$, $\bar{I}_H = +7.9$, CI $[+0.8, +15.2]$, $7/8$ seeds; $\bar{I}_R = +5.9$, CI $[+0.9, +12.0]$, $5/8$; `cheetah-run` $\bar{I}_H = +7.05$ (CI $[-0.84, +13.95]$), $\bar{I}_R = +12.14$ (CI $[+3.85, +19.83]$); per-task Holm-adjusted $p = 0.070$, 0.070, 0.063 and 0.031. On `walker-walk` the registered bootstrap test gave Holm-adjusted $p = 0.032$ each, and an exhaustive sign-flip test one-sided $p = 0.035$ and 0.047, Holm-adjusted ($p = 0.070$ each), which is why the second task and the pooled exact family were added. The interaction is not detected on either task ($I_{HR} = +2.8$, CI $[-11.1, +18.3]$ on `walker-walk`).

M PRESSURE-SWAP FACTORIAL FOR THE HIGH-CONTRIBUTION ARM

Pre-registered before any run, after the low-cost factorial had been analysed. Design, estimands and tests are those of Section 6 (identical to the earlier `lowC` factorial of Appendix L); checkpoints and `sham` controls are those of the low-cost factorial (re-run check, Appendix U). Table 10 gives the cell means and channel estimates under the registered V_{reg} and the trapezoid AUC.

value	task	$\Delta_{1,1}$	$\Delta_{1,8}$	$\Delta_{8,1}$	$\Delta_{8,8}$	$\bar{I}_H$ [CI], pos	$\bar{I}_R$ [CI], pos	I_{HR}
V_{reg}	walker-walk	−193.8	+1.3	−99.1	+20.5	+56.9 [+43.8, +71.0], 8/8	+157.4 [+140.0, +175.6], 8/8	−75.5
V_{reg}	cheetah-run	−481.4	−52.8	−475.8	−53.2	+2.6 [−61.7, +58.3], 6/8	+425.6 [+363.0, +473.9], 8/8	−6.0
AUC	walker-walk	−125.3	+2.6	−45.6	+21.5	+49.3 [+33.7, +65.8], 8/8	+97.5 [+80.0, +114.5], 8/8	−60.7
AUC	cheetah-run	−490.2	−40.4	−457.8	−26.2	+23.3 [−42.3, +78.8], 6/8	+440.7 [+376.6, +489.9], 8/8	−18.2
V_{reg}	pooled (%)					+6.07 [+1.36, +9.90], 15/16, Holm 0.0077	+40.90 [+28.78, +53.37], 16/16, Holm 0.00003	
AUC	pooled (%)					+6.65 [+2.02, +10.76], 15/16, Holm 0.0050	+38.11 [+24.19, +52.37], 16/16, Holm 0.00003	

Table 10: Repeated `highC` schedule, history × recovery UTD. Cell means of the benefit $\Delta_{h,r}$ in raw return; channel estimates with seed-blocked bootstrap CIs and positive-seed counts; pooled estimates on the relative scale with exact stratified sign-flip p-values, Holm-corrected over the pair. The registered one-sided test is for a positive interaction and does not detect one (exact $p = 0.996$ on `walker-walk`, 0.56 on `cheetah-run`); descriptively, I_{HR} is negative on `walker-walk`, i.e. the cell pattern is not additive, the gains of the two factors together being less than the sum of their separate gains.

Interval construction. The registered intervals of this factorial (Table 10, Figure 3 and the text) are built from the sixteen per-seed relative contrasts $\tilde{I}_s$, each divided by the observed mean `sham` value of its own task and cell (held fixed), resampled with replacement as one pool of sixteen; the test resamples signs within task. Because the per-seed contrasts differ in level between the two tasks, pooling widens the interval. As sensitivity analyses, resampling the eight seeds within each task and averaging the two task means gives $[+33.4\%, +41.9\%]$ for $\bar{I}_R$ and $[+2.0\%, +10.6\%]$ for $\bar{I}_H$ (registered: $[+24.2\%, +52.4\%]$ and $[+2.0\%, +10.8\%]$), and additionally re-estimating the `sham` denominator inside every resample gives $[+35.5\%, +40.3\%]$ and $[+3.5\%, +9.7\%]$. The registered construction is the main analysis throughout; none of the readings changes under the alternatives.

N UPDATE SUPPLY OR COMPUTE?

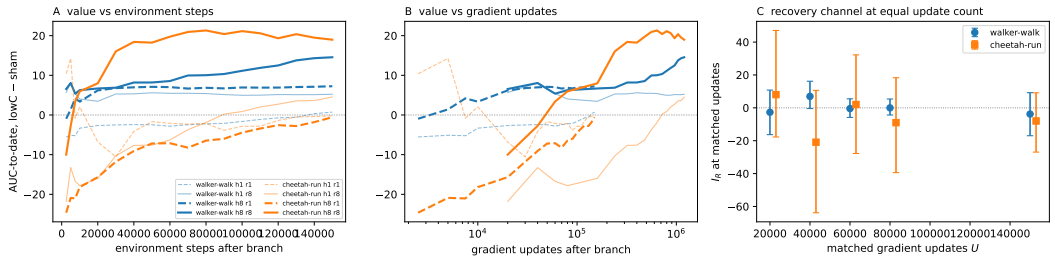


Figure 13: The recovery curves of the `lowC` pressure-swap factorial on two axes. **A** value to date (running mean of the paired `lowC−sham` return) against environment steps after the branch and **B** against gradient updates, for all four history/recovery cells of both tasks. **C** the recovery-channel contrast I_R re-formed at *matched gradient updates* (recovery UTD 8 after t environment steps against recovery UTD 1 after $8t$), mean with seed-blocked bootstrap CI.

A recovery UTD of 8 supplies eight times as many gradient updates per environment step, so the recovery channel could be a statement about update supply or merely about compute. The logged evaluations allow the comparison to be redone at matched gradient updates: recovery UTD 8 evaluated $t \in \{2.5, 5, 7.5, 10, 20\}$k environment steps after the branch against recovery UTD 1 evaluated after $8t$ steps (the last pair uses 150k, the end of the budget). At matched updates the contrast is not detected on either task at any of the five counts (all CIs include zero; point estimates -3.8 to $+7.0$ raw return on `walker-walk` and -20.9 to $+8.0$ on `cheetah-run`, single-evaluation differences, Figure 13C), whereas the same evaluations compared at matched environment steps give positive estimates whose CIs exclude zero at two of the five points on each task. This is consistent with the recovery channel being an effect of update supply—the reset learner needs a certain number of updates to repay the reset, and pressure delivers them sooner—rather than of anything else UTD changes after the branch. It is not established: the matched-update contrasts are single-evaluation differences on eight seeds, their intervals are wide enough to contain the matched-step effect, and at matched updates the UTD-8 learner has also seen eight times fewer new transitions, so replay age is

confounded with update count. The decisive design, a fixed replay stream delivering the same total number of updates at different rates, is left to future work.

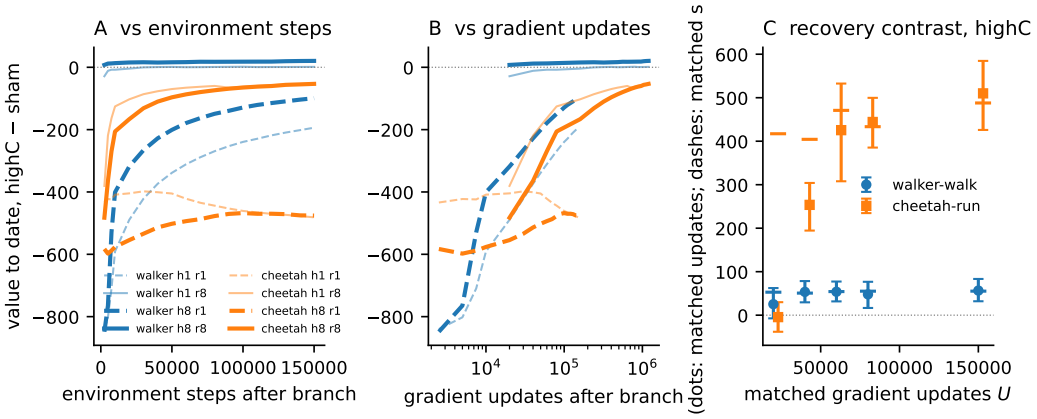


Figure 14: The same analysis for the `highC` factorial of Section 6. **A, B** value to date of the paired `highC−sham` return against environment steps and against gradient updates. **C** the recovery contrast I_R at matched gradient updates (dots, seed-blocked bootstrap CIs) and, for reference, at matched environment steps (dashes).

The `highC` factorial. Re-reading the `highC` logs the same way, after the factorial had been adjudicated, gives a different answer (Figure 14C). At matched gradient updates the recovery contrast is `walker-walk`: $U = 20$k: $+25$ $[-7, +62]$ (5/8) against $+53$ at matched steps; $U = 40$k: $+54$ $[+30, +78]$ (8/8) against $+51$ at matched steps; $U = 60$k: $+54$ $[+32, +77]$ (8/8) against $+54$ at matched steps; $U = 80$k: $+48$ $[+17, +77]$ (7/8) against $+55$ at matched steps; $U = 150$k: $+57$ $[+32, +83]$ (8/8) against $+55$ at matched steps; `cheetah-run`: $U = 20$k: -4 $[-38, +30]$ (4/8) against $+417$ at matched steps; $U = 40$k: $+254$ $[+195, +304]$ (8/8) against $+404$ at matched steps; $U = 60$k: $+425$ $[+308, +533]$ (8/8) against $+471$ at matched steps; $U = 80$k: $+444$ $[+385, +500]$ (8/8) against $+433$ at matched steps; $U = 150$k: $+510$ $[+426, +585]$ (8/8) against $+488$ at matched steps. Except at the earliest match (20k updates), the contrast at matched updates equals the contrast at matched environment steps. This does not identify a mechanism. Because resets recur every 10k environment steps, the two arms of a matched-update pair differ in how many resets they have undergone: at $U = 20$k, 40k, 60k, 80k and 150k updates the UTD-1 arm (evaluated at $+20$k, $+40$k, $+60$k, $+80$k, $+150$k steps) has undergone 2, 4, 6, 8 and 14 further resets, the UTD-8 arm (at $+2.5$k, $+5$k, $+7.5$k, $+10$k, $+20$k) none, none, none, one and one. The surviving contrast is therefore consistent with fewer re-damages under UTD 8 as much as with update density, and the UTD-8 arm has also seen eight times fewer new transitions at each count (replay age). What the repeated schedule does establish is that under a fixed environment-step reset schedule, pressure reduces the harm at equal gradient-update counts. The single-reset factorial of Appendix O is the control that removes the reset-count confound; a fixed-replay-stream design would additionally remove replay age.

O SINGLE-RESET HISTORY × RECOVERY FACTORIAL

Pre-registered before any run. The design is that of the repeated `highC` factorial (Section 6, Appendix M) with one change: a single `highC` reset at the branch ($K = 10\%$) and no later intervention. Because the score computation consumes random numbers at each event, the stored repeated `sham` runs are not byte-identical to a single-event `sham`, so both arms were run afresh: 2 tasks × 4 cells × 8 seeds × 2 arms = 128 runs on the PSWAP trunks (`walker-walk` seeds 16–23, `cheetah-run` 30–37). Estimands and tests are those of Appendix M (AUC primary, V_{reg} reported; relative-scale pooled contrasts with exact 2^{16} within-task sign-flip p-values, Holm over the pair). The matched-update contrast I_R^U compares recovery UTD 8 at $+t$ against recovery UTD 1 at $+8t$ for $t \in \{2.5, 5, 7.5, 10, 20\}$k (single evaluations; the last pair, $+20$k against $+150$k, compares 160k with 150k updates and is only approximately matched), and the surviving fraction is I_R^U divided by the

matched-step contrast formed from the same $+8t$ evaluations. The outcome is consistent with update supply: the recovery contrast is detected at matched steps and not at matched updates.

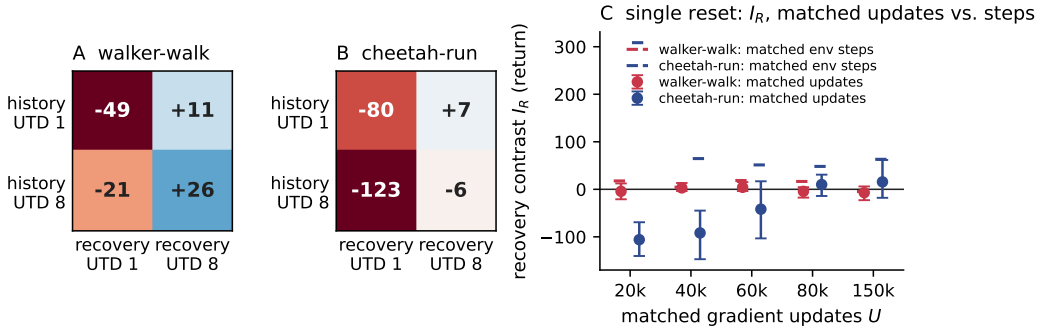


Figure 15: Single `highC` reset, history × recovery UTD. **A, B** cell means of the benefit $\Delta_{h,r}$ (AUC, raw return). **C** the recovery contrast I_R at matched gradient updates (dots, seed-blocked bootstrap CIs) and at matched environment steps (dashes, the same $+8t$ evaluations), both tasks.

value	task	$\Delta_{1,1}$	$\Delta_{1,8}$	$\Delta_{8,1}$	$\Delta_{8,8}$	$\bar{I}_H$ (pos; p)	$\bar{I}_R$ (pos; p)
AUC	walker-walk	−48.7	+11.1	−20.7	+25.7	+21.3 (7/8; 0.035)	+53.1 (8/8; 0.0039)
AUC	cheetah-run	−79.5	+7.4	−122.6	−5.6	−28.1 (3/8; 0.945)	+102.0 (8/8; 0.0039)
V_{reg}	walker-walk	−127.8	+8.3	−77.3	+24.0	+33.1 (7/8; 0.008)	+118.7 (8/8; 0.0039)
V_{reg}	cheetah-run	−124.3	−11.1	−184.3	−34.4	−41.6 (1/8; 0.984)	+131.5 (8/8; 0.0039)
AUC	pooled (%)					−0.39 [−3.06, +2.01] (10/16; Holm 0.607)	+10.31 [+7.77, +12.91] (16/16; Holm 0.00003)
V_{reg}	pooled (%)					−0.31 [−3.61, +2.56] (10/16; Holm 0.571)	+16.37 [+14.15, +18.63] (16/16; Holm 0.00003)

Table 11: Single `highC` reset: cell means of the benefit in raw return, per-task channel estimates with positive-seed counts and exact 2^8 sign-flip p-values, and pooled estimates on the relative scale (per cent of the `sham` cell) with seed-blocked bootstrap CIs and Holm-corrected exact stratified p-values. The recovery channel is detected on both tasks under both value definitions; the history channel is positive on `walker-walk` and negative on `cheetah-run` and is not detected pooled. A single `highC` reset costs nothing in AUC when the recovery UTD is 8 (cells $\Delta_{\cdot,8}$ near zero) and is harmful when it is 1.

task	U	I_R^U [CI]	pos	I_R at matched steps	surviving
walker-walk	20k	-4.2 $[-21.0, +12.4]$	3/8	$+17.6$	-0.24
walker-walk	40k	$+3.0$ $[-3.8, +13.3]$	3/8	$+4.5$	$+0.66$
walker-walk	60k	$+4.2$ $[-3.7, +15.2]$	3/8	$+18.5$	$+0.23$
walker-walk	80k	-4.3 $[-17.6, +4.4]$	5/8	$+16.5$	-0.26
walker-walk	≈ 150k	-7.0 $[-22.7, +6.0]$	6/8	-4.7	$+1.49$
cheetah-run	20k	-105.9 $[-140.2, -69.3]$	0/8	$+308.5$	-0.34
cheetah-run	40k	-91.8 $[-146.8, -44.5]$	0/8	$+64.7$	-1.42
cheetah-run	60k	-41.8 $[-103.9, +17.5]$	3/8	$+51.3$	-0.81
cheetah-run	80k	$+10.0$ $[-14.0, +30.8]$	6/8	$+48.2$	$+0.21$
cheetah-run	≈ 150k	$+15.7$ $[-17.8, +61.7]$	5/8	$+63.2$	$+0.25$

Table 12: Single reset, recovery contrast at matched gradient updates (single evaluations, seed-blocked bootstrap CIs, positive-seed counts) against the contrast at matched environment steps from the same $+8t$ evaluations, and their ratio. On `walker-walk` every CI includes zero; on `cheetah-run` the UTD-8 arm is behind at 20k and 40k updates (it has seen only 2.5k and 5k new transitions) and the contrast is not detected from 60k. The surviving fraction is below 0.5 at 60k and 80k on both tasks and at 150k on `cheetah-run`; the last `walker-walk` ratio has a negative matched-step reference (-4.7, a single evaluation) and is not interpretable. The first row of the reading table (fraction ≥ 0.5 from 40k on both tasks) is excluded and the second applies. The matched-step reference on `walker-walk` is small at these single evaluations, so its ratios are noisy; the CIs of I_R^U carry the result.

P RANDOM-K VERSUS LOWC AT A FIXED TRAINING HISTORY

Pre-registered before any run, as a follow-up to the ranking change observed in Appendix J, where history and recovery UTD change together. Here the history is held fixed: on the `walker-walk` trunks of the low-cost factorial (history UTD 1, seeds 16–23; Appendix L) the repeated `randK` schedule was run at recovery UTD 1 and 8 (16 continuations, 150k steps, events every 10k), against the stored `sham` and `lowC` cells of the same checkpoints (re-run check, Appendix U). The registered estimands are the paired differences $D_r = \mathrm{AUC}_{\texttt{randK}} - \mathrm{AUC}_{\texttt{lowC}}$ at each recovery UTD and the interaction $J = D_8 - D_1$ (seed-blocked bootstrap CIs, exact 2^8 two-sided sign-flip tests). Result: neither difference nor the interaction is detected (the experiment does not establish equivalence), and `randK` is nearly harmless on these checkpoints even at recovery UTD 1 ($+1.5$ against `sham`; `lowC` $+1.3$; at recovery UTD 8 $+1.7$ and $+5.3$), unlike on the confirmatory cohort (Appendix J).

value	D_1 [CI], pos, p	D_8 [CI], pos, p	$J = D_8 - D_1$ [CI], p
AUC	$+0.2$ $[-1.8, +2.3]$, 4/8, $p = 0.87$	-3.6 $[-12.6, +4.6]$, 3/8, $p = 0.50$	-3.8 $[-12.8, +3.5]$, $p = 0.49$
V_{reg}	-3.1 $[-7.0, -0.0]$, 1/8, $p = 0.13$	-3.0 $[-11.2, +5.0]$, 3/8, $p = 0.54$	$+0.1$ $[-9.8, +11.1]$, $p = 0.96$

Table 13: Random-K minus `lowC` at a fixed history (`walker-walk`, history UTD 1, seeds 16–23) at each recovery UTD, and their interaction; raw return, seed-blocked bootstrap CIs, exact two-sided sign-flip p-values.

Q EVENT-0 COST VERSUS REPEATED-SCHEDULE VALUE

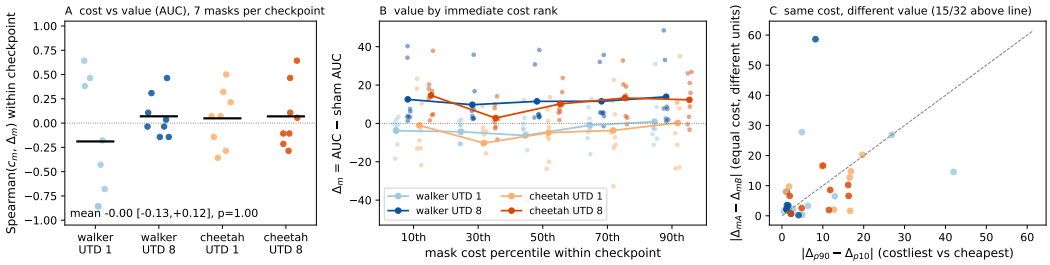


Figure 16: Cost and value of random masks from the same checkpoint on the repeated schedule, on the time-normalized AUC. **A** per-checkpoint Spearman correlation between the event-0 immediate cost and the value of seven rolled-out masks, 32 checkpoints; bars are cell means (mean -0.00, stratified CI $[-0.13, +0.12]$; the registered V_{reg} version, -0.09 $[-0.19, +0.02]$, is the statistic in the text). **B** value by immediate-cost percentile. **C** the value difference between two masks of equal cost and disjoint units against the difference between the costliest and cheapest mask of the same checkpoint.

This pre-registered test preceded the single-reset experiment of Section 7 and used the same checkpoints and masks on the repeated schedule. Figure 17 shows where the 100 random masks of each checkpoint sit in cost relative to the sets a real selector picks: their range lies strictly between the `lowC`/`lowA` sets and the `highC` set on every checkpoint (recomputed in the measurement state of the branch with the bootstrap target re-sampled; median ratios $61\times$ and $8\times$), so both within-regime tests concern the interior of the cost range.

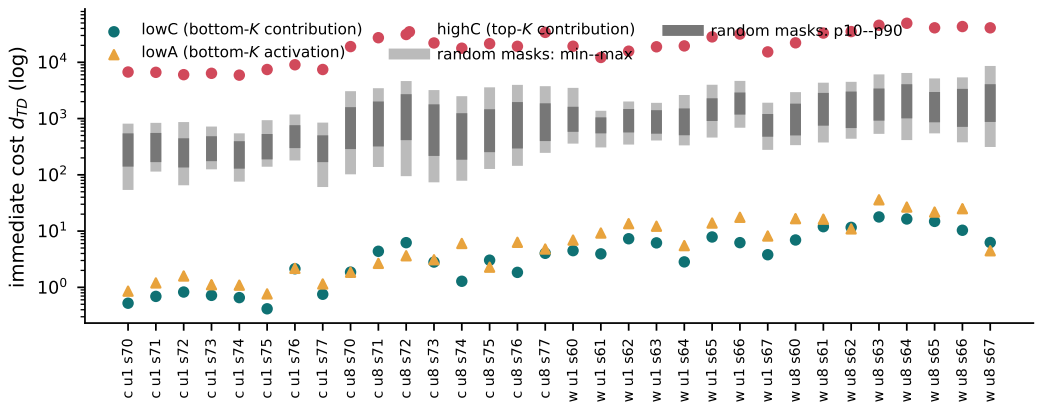


Figure 17: Immediate cost of the 100 random 10% masks (bars: min–max and p10–p90) against the bottom-K contribution (`lowC`), bottom-K activation (`lowA`) and top-K contribution (`highC`) sets at each of the 32 checkpoints of the within-regime experiments (w = walker-walk, c = cheetah-run; u = UTD; s = seed).

One may ask whether, at the *same* learner state and the same pressure, immediate cost predicts long-run value. If a low-cost selector were useful within a regime, the expected relation is negative: cheaper masks should be worth more. We pre-registered a test on fresh seeds (60–67 on `walker-walk`, 70–77 on `cheetah-run`, UTD 1 and 8): at each checkpoint we measured the immediate cost of 100 random 10% masks in the exact measurement state of the branch, selected the masks at cost ranks 10, 30, 50, 70 and 90 and one pair of masks with equal cost (relative difference below 0.1%) and nearly disjoint units (Jaccard ≤ 0.15), and rolled all seven out on the repeated schedule against `sham` from the same checkpoint (Figure 16). The primary estimand is the within-checkpoint Spearman correlation between cost and value, averaged over the four task×pressure cells; the test is a sign-flip test stratified by cell (two-sided, 10^6 Monte-Carlo assignments). No stable, general monotonic relation is detected: mean -0.09 (stratified CI $[-0.19, +0.02]$, $p = 0.21$), with cell means -0.36, $+0.05$, -0.05, $+0.02$ for walker UTD 1 and 8 and cheetah UTD 1 and 8 (on the AUC of Figure 16: mean -0.00, CI $[-0.13, +0.12]$, walker UTD 1 -0.19)—one cell, walker at low pressure, is consistent with the negative relation a useful selector would produce, and the other three are near zero. On the repeated schedule and under the registered metric, the costliest mask beat the cheapest in $13/32$ checkpoints (Clopper–Pearson $[0.24, 0.59]$; $16/32$ for the single reset on the AUC, Section 7), and two masks of equal cost differed in value by at least as much as the costliest and cheapest mask in $13/32$. The upper 95% confidence bound of $+0.02$ is a bound, not a demonstration of independence: the estimate is compatible with a weak negative relation at low pressure and with none at high pressure. Two caveats apply. The cost is that of the first event, whereas the value is that of a schedule that re-scores and re-selects at fourteen later events, so the weak correlation cannot by itself distinguish a weak cost–value relation of the first reset from a first-reset effect overwritten by later, different interventions; the single-reset experiment of Section 7 resolves this, and its continuation replicates measure the ranking noise that this design does not.

Per-cell detail of the single-reset test. The random masks span a narrow band of cost: their cheapest costs about $61\times$ the bottom-K set a real `lowC` selector picks, and the top-K `highC` set costs about $8\times$ their costliest. The costliest-minus-cheapest sign agrees across the three continuation streams in $15/32$ checkpoints (chance under no effect $8/32$; intraclass correlation median 0.42): a reproducible cost-rank effect exists at some checkpoints but is small relative to continuation noise. The costliest mask beat the cheapest in $16/32$ checkpoints, and two masks of equal cost differed in value by at least as much as the costliest and cheapest mask in $12/32$. Seed-blocked CIs of the four cell means of Section 7 are $[-0.57, -0.10]$, $[+0.02, +0.51]$, $[-0.12, +0.48]$, $[-0.12, +0.36]$ (walker UTD 1 and 8, cheetah UTD 1 and 8); under the registered V_{reg} the mean is -0.03, $[-0.15, +0.09]$, $p = 0.69$. The cost × UTD interaction is $+0.62$, CI $[+0.25, +0.97]$ on `walker-walk` and -0.08 (interaction -0.08, CI $[-0.35, +0.20]$) on `cheetah-run`. Averaged over continuation streams the costliest-minus-cheapest difference is $+3.8$, CI $[+0.5, +8.8]$, on walker at UTD 8 and within ± 3 of zero elsewhere.

R REPEATED LOW-COST RESETS RE-SELECT THE SAME UNITS

On the paper's schedule (a reset every 10k environment steps), the bottom-K set of one event overlaps the previous event's set by $R_1 = 0.93$ on average over walker-walk seeds 40–43 at UTD 1 and 8 (0.99 at UTD 1, 0.87 at UTD 8; the expectation for independent selection of 10% is 0.10), and only 9% of all selections over a run are distinct units. The reason is visible in the units' gradients: after a reset, a unit's incoming-weight gradient norm is 2.2–2.9% of the layer median at 100, 1k and 5k critic updates later and 14% just before the next event, because its zeroed outgoing weights leave almost no gradient path into it; it therefore remains the lowest-contribution unit and is selected again. Changing from an environment clock to an update clock (20k critic updates per event) altered the event frequency (8 events at UTD 1, 61 at UTD 8) but did not resolve within-unit re-selection ($R_1 = 0.98$, 8% distinct). This is consistent with the small immediate cost of the later events of the repeated `lowC` arm: the schedule mostly re-initializes units that have not yet re-grown, so little additional functional support is affected. Whether rotating the targets would raise long-run value is a counterfactual this diagnostic does not test.

S DECOMPOSITION OF THE BENEFIT, A POLICY-FACING DAMAGE MEASURE, AND A CRITIC-RECOVERY PILOT

These analyses are descriptive and post hoc.

Per-history recovery effects. The 85% share of Section 6 summarizes two marginal contrasts; the recovery effect $\Delta_{h,8} - \Delta_{h,1}$ is $+127.8$ $[+96.2, +159.6]$ at history UTD 1 and $+67.1$ $[+51.6, +87.1]$ at history UTD 8 on `walker-walk`, and $+449.8$ $[+406.1, +488.4]$ and $+431.6$ $[+339.1, +498.1]$ on `cheetah-run` (8/8 seeds each; AUC, raw return).

Loss and gain areas. With $\Delta R(t) = R_{\text{reset}}(t) - R_{\texttt{sham}}(t)$ interpolated linearly between evaluations, the time-normalized integrals of its negative and positive parts add up to the AUC benefit. Evaluation noise contributes a positive gain area even for a harmless arm, so `lowC` is the reference for that floor.

arm	loss area [CI]	gain area [CI]	net
`walker` lowC UTD 1	-5.3 $[-9.2, -2.6]$	$+8.1$ $[+1.4, +19.6]$	$+2.8$
`walker` randK UTD 1	-16.3 $[-26.9, -8.1]$	$+2.8$ $[+0.6, +5.4]$	-13.5
`walker` highC UTD 1	-103.0 $[-127.0, -80.7]$	$+0.3$ $[+0.0, +0.7]$	-102.7
`walker` lowC UTD 8	-0.6 $[-1.1, -0.2]$	$+9.2$ $[+7.3, +11.0]$	$+8.6$
`walker` randK UTD 8	-0.5 $[-1.4, -0.0]$	$+21.4$ $[+13.0, +34.4]$	$+20.9$
`walker` highC UTD 8	-0.7 $[-1.5, -0.1]$	$+16.5$ $[+13.2, +20.0]$	$+15.8$
`cheetah` lowC UTD 1	-10.1 $[-14.6, -5.8]$	$+18.1$ $[+9.2, +28.0]$	$+8.1$
`cheetah` randK UTD 1	-42.1 $[-55.9, -28.2]$	$+8.6$ $[+3.2, +15.1]$	-33.4
`cheetah` highC UTD 1	-449.2 $[-520.1, -380.2]$	$+0.0$ $[+0.0, +0.0]$	-449.2
`cheetah` lowC UTD 8	-4.1 $[-7.1, -1.4]$	$+20.8$ $[+15.9, +25.5]$	$+16.7$
`cheetah` randK UTD 8	-5.1 $[-7.6, -2.8]$	$+21.3$ $[+12.7, +31.1]$	$+16.2$
`cheetah` highC UTD 8	-36.2 $[-44.4, -28.5]$	$+6.2$ $[+2.3, +11.1]$	-30.0
`walker` single `highC` $h = 1, r = 1$	-56.0 $[-65.2, -49.2]$	$+7.3$ $[+2.0, +16.3]$	-48.7
`walker` single `highC` $h = 1, r = 8$	-3.3 $[-4.8, -1.8]$	$+14.4$ $[+1.8, +38.5]$	$+11.1$
`walker` single `highC` $h = 8, r = 1$	-34.6 $[-37.1, -32.5]$	$+14.0$ $[+9.3, +20.2]$	-20.7
`walker` single `highC` $h = 8, r = 8$	-0.6 $[-1.7, -0.0]$	$+26.3$ $[+11.0, +49.7]$	$+25.7$
`cheetah` single `highC` $h = 1, r = 1$	-86.2 $[-100.2, -70.0]$	$+6.7$ $[+1.9, +12.9]$	-79.5
`cheetah` single `highC` $h = 1, r = 8$	-21.6 $[-27.9, -14.5]$	$+29.0$ $[+8.5, +56.4]$	$+7.4$
`cheetah` single `highC` $h = 8, r = 1$	-125.0 $[-150.2, -94.6]$	$+2.4$ $[+0.6, +4.5]$	-122.6
`cheetah` single `highC` $h = 8, r = 8$	-27.0 $[-36.4, -17.0]$	$+21.4$ $[+15.5, +27.9]$	-5.6

Table 14: Loss and gain areas (raw return, seed-blocked bootstrap CIs, eight seeds each) for the repeated schedules of Section 5 (confirmatory cohort) and for the single `highC` reset of Appendix O (history h, recovery r). At recovery UTD 8 the loss area of the costly resets shrinks by one to two orders of magnitude and the gain area is at or above the `lowC` floor.

Critic recovery on a fixed learning problem (pilot). To ask whether a larger initial disruption must take longer to repair, we froze the actor and the target critic of eight history-UTD-1 checkpoints (`walker-walk` seeds 40–43, `cheetah-run` 50–53), applied one `sham`, `lowC`, `randK` or `highC` reset, and ran 1,000 critic-only updates with an identical minibatch sequence and identical precomputed targets in every arm, measuring the TD loss on a held-out batch (disjoint from the

scoring batch, target fixed once) at 0, 10, 100, 500 and 1,000 updates. Table 15 gives the excess over `sham` at equal update counts. `randK` and `highC` repair more than 99.8% of their excess loss within 1,000 updates (their residual is about 1,000× smaller than the initial excess); `lowC`, whose excess is small to begin with, repairs 78–93%. The ordering of the residual follows the initial cost (`highC` > `randK` > `lowC`) on each of the eight checkpoints individually. On the online schedule 1,000 updates correspond to 1,000 environment steps at UTD 1 and 125 at UTD 8, a correspondence in update count only.

task, arm	0	10	100	500	1,000 updates
`walker` lowC	7.2	4.0	1.4	0.8	0.8
`walker` randK	1,327.1	134.2	7.4	2.2	1.5
`walker` highC	30,386.2	3,309.3	162.8	32.9	17.9
`cheetah` lowC	0.9	0.9	0.4	0.2	0.1
`cheetah` randK	277.3	30.6	2.0	0.7	0.4
`cheetah` highC	6,995.3	507.1	76.1	10.6	6.3

Table 15: Held-out TD loss of the reset arm minus `sham` at equal critic-update counts on a fixed learning problem (mean over four checkpoints per task; actor and target critic frozen).

A policy-facing damage measure. The immediate cost d_{TD} measures critic disruption; the policy is changed only through the actor update. On every single-reset checkpoint and mask we therefore computed the actor-loss gradient on the audit-batch states with fixed reparameterized action noise before and after the critic reset, and used $1 - \cos(g_{\mathrm{pre}}, g_{\mathrm{post}})$ and the relative change of the gradient vector, $\|g_{\mathrm{post}} - g_{\mathrm{pre}}\|/\|g_{\mathrm{pre}}\|$, as alternative costs. Across the 224 masks they are only weakly related to d_{TD} (Spearman 0.11 and 0.12), so the two kinds of damage differ; but neither improves on d_{TD} in ordering the realised benefit within a regime; the per-cell intervals are wide, and the per-cell values, not the split means, are the ones to read.

predictor	walker UTD 1	walker UTD 8	cheetah UTD 1	cheetah UTD 8	dev. mean	val. mean
d_{TD} (recomputed)	-0.38 $[-0.57, -0.15]$	$+0.27$ $[+0.02, +0.51]$	$+0.20$ $[-0.13, +0.48]$	$+0.13$ $[-0.11, +0.37]$	-0.05	$+0.16$
$1 - \cos(g_{\mathrm{pre}}, g_{\mathrm{post}})$	$+0.01$ $[-0.39, +0.36]$	$+0.08$ $[-0.28, +0.43]$	-0.10 $[-0.29, +0.11]$	-0.18 $[-0.52, +0.19]$	$+0.05$	-0.14
$\|g_{\mathrm{post}} - g_{\mathrm{pre}}\|/\|g_{\mathrm{pre}}\|$	-0.22 $[-0.48, +0.03]$	$+0.09$ $[-0.31, +0.49]$	$+0.09$ $[-0.29, +0.47]$	-0.02 $[-0.37, +0.31]$	-0.06	$+0.03$

Table 16: Mean within-checkpoint Spearman correlation between a damage measure and the AUC benefit of the seven masks (stream 0), with seed-blocked CIs over eight checkpoints per cell; the d_{TD} row uses the recomputed cost and reproduces Section 7.

T Robustness to the definition of value

The registered value is the mean of 18 evaluations at irregular times (Appendix V). Table 17 recomputes the two headline results with a time-normalized trapezoidal AUC over environment steps $[150\mathrm{k}, 300\mathrm{k}]$ and with the endpoint (mean of the last three evaluations), from the same logs. The `highC` pressure reversal holds under every definition on both tasks ($8/8$ seeds each): weighting the early evaluations less shrinks the UTD-1 harm on `walker-walk` (the collapse is largest early) but does not change any sign or dominance count. The two `lowC` channels hold under the time-normalized AUC (Holm-adjusted $p = 0.015$ and 0.0014) but are not detected at the endpoint ($p = 0.31$ and 0.31): they are effects on the recovery trajectory, not on where the learner ends up after 150k steps. For `highC` the recovery contrast is detected under every definition and the history contrast under V_{reg} and the AUC but not at the endpoint.

quantity	task	registered V_{reg} (mean of 18)	time-normalized AUC	endpoint (last 3)
highC DiD (raw)	walker-walk	$+176.4$ [$+154.5$, $+197.5$], 8/8	$+118.5$ [$+96.3$, $+142.6$], 8/8	$+94.1$ [$+50.1$, $+148.0$], 8/8
highC DiD (raw)	cheetah-run	$+383.8$ [$+314.4$, $+454.0$], 8/8	$+419.3$ [$+345.6$, $+493.4$], 8/8	$+456.1$ [$+358.3$, $+551.7$], 8/8
highC at UTD 8	walker-walk	$+14.8$	$+15.8$	$+9.9$
highC at UTD 8	cheetah-run	-56.5	-30.0	-16.4
highC $\bar{I}_H$ pooled (% of baseline)	both	$+6.07$, Holm 0.0077	$+6.65$, Holm 0.0050	$+4.77$, Holm 0.074
highC $\bar{I}_R$ pooled (% of baseline)	both	$+40.90$, Holm 0.00003	$+38.11$, Holm 0.00003	$+37.63$, Holm 0.00003
lowC $\bar{I}_H$ pooled (% of baseline)	both	$+0.89$, Holm 0.012	$+0.99$, Holm 0.015	$+0.77$, Holm 0.31
lowC $\bar{I}_R$ pooled (% of baseline)	both	$+1.13$, Holm 0.0067	$+1.24$, Holm 0.0014	$+0.79$, Holm 0.31

Table 17: Headline results under three definitions of value (V_{reg} is the registered metric; "AUC" is the time-normalized trapezoid). Means with seed-blocked bootstrap CIs and positive-seed counts; pooled channel effects with Holm-adjusted exact p-values.

U SEED REGISTRY AND PRE-REGISTRATION LINEAGE

Two kinds of experiment must be distinguished. *Fresh-seed* experiments (the pressure-swap factorial on both tasks, the confirmatory highC replication, the four further tasks, the within-checkpoint cost–value cohort, the recovery-burst test and the exploratory highC cohort) use trunks that appear in no other experiment; the later experiments that reuse those trunks (the highC factorial, the single-reset test, the budget-aware strength test) are listed with the trunks they reuse. The remaining experiments *reuse* checkpoints: the single-event cost experiment, the repeated low-cost arm, the draw and boundary isolation, the calibration and the set-cost diagnostics all start from the seed 0–7 walker-walk trunks (and the seed 30–37 cheetah-run trunks of the pressure swap), so agreement among them is not independent replication. Each confirmatory experiment has a pre-registration document, frozen before its first run, and an adjudication script frozen before its data existed. Amendments are separate documents that never modify a frozen file. Where an experiment reuses stored sham cells, one stored cell per task was re-run with the current code and reproduced its evaluation sequence exactly ("re-run check").

Experiment (section)	Task	Seeds	Registered before run	Runs
Single-event cost (§4)	walker-walk	0–7 (fresh trunks)	design, STOP/GO rules	16 trunks, 64 branches
Repeated `lowC` (§5)	walker-walk	0–7 (reuses cost trunks)	DiD estimand, one-sided test	32 branches
Exploratory `highC` (§5)	walker-walk	8–15	*outcome-triggered; non-confirmatory*	16 trunks, 32 branches
Draw / boundary isolation (App. D)	walker-walk	0–3 (reuses cost trunks)	ρ threshold 0.2	128 branches
Pressure swap (§6)	walker-walk	16–23	estimands, Holm-2	16 trunks, 64 branches
Pressure swap, cheetah (§6)	cheetah-run	30–37	pooled family, relative scale, exact test	16 trunks, 64 branches
Calibration and set-cost diagnostics (§4, App. A)	both	0–7, 30–37 (reuses trunks)	predictors, masks, wording table	32 checkpoints, no rollout
Confirmatory `highC` (§5)	both	40–47, 50–57	single pooled test, wording table	32 trunks, 128 branches
Within-checkpoint cost–value (§7)	both	60–67, 70–77	masks, estimand, reading table	32 trunks, 256 branches
Four further tasks (§5)	cartpole, finger, ball_in_cup, reacher	80–85, 90–95, 110–115, 100–105	task-level family, floor rule	48 trunks, 144 branches
Recovery-burst method test (App. I)	both	120–127, 130–137 (fresh trunks)	compute-matched 2×2, promotion rule	32 trunks, 80 branches
`highC` factorial (§6)	both	16–23, 30–37 (reuses pressure-swap trunks and `sham` cells; re-run check)	estimands, pooled exact test, reading table	64 branches, 2 validity re-runs
Single reset (§7)	both	60–67, 70–77 (reuses within-checkpoint trunks and masks)	AUC primary, stratified test, reading table	448 branches
TD3 replication (App. G)	walker-walk	120–125 (new TD3 trunks; seed numbers shared with the SAC recovery-burst trunks)	one-sided exact test, reading table	12 trunks, 36 branches
Single-reset factorial (§6, App. O)	both	16–23, 30–37 (reuses pressure-swap trunks; new `sham` and `highC` arms)	estimands, pooled exact test, matched-update contrast, reading table	128 branches
Budget-aware strength (App. H)	both	40–47, 50–57 (reuses confirmatory trunks, `sham` and `highC` cells; re-run check)	one-sided exact test, reading table	32 branches, 2 validity re-runs
Random-K repeated reset (§5, App. J)	both	40–47, 50–57 (reuses confirmatory trunks and stored arms; re-run check)	estimand, pooled exact test, reading table	32 branches, 2 validity re-runs
Dose–response UTD 2, 4 (§5, App. K)	walker-walk	140–147 (fresh trunks)	estimand, permutation test, reading table	16 trunks, 48 branches
Random-K at fixed history (§5, App. P)	walker-walk	16–23 (reuses low-cost-factorial trunks, `sham` and `lowC` cells; re-run check)	estimands, exact tests, reading table	16 branches, 1 validity re-run

Table 18: Seed sets and what was fixed before each experiment ran. The one decision-rule change made after outcomes were read (moving two stop criteria from the single-event to the repeated cell) is recorded as a timestamped amendment.

V IMPLEMENTATION DETAILS

Agent. State-based SAC (Haarnoja et al., 2018) with twin critics. Actor: MLP $\mathrm{obs} \to 256 \to 256 \to 2 \cdot \mathrm{act}$ with ReLU, tanh-squashed Gaussian policy, $\log \sigma$ clamped to $[-5, 2]$. Each critic: MLP $(\mathrm{obs} + \mathrm{act}) \to 256 \to 256 \to 1$ with ReLU; the two 256-unit hidden layers of each critic are the resettable layers (4 layers, 1024 units in total; $K = 26$ per layer). Target critics updated by Polyak averaging with $\tau = 0.005$ after every update; $\gamma = 0.99$; entropy coefficient learned with target entropy $-\dim(\mathrm{act})$, initialized at $\alpha = 1$. Three Adam optimizers (actor, critics, entropy coefficient), learning rate 3×10^{-4}, default betas, no weight decay. Batch size 256, one critic and one actor update per gradient step, UTD gradient steps per environment step. Replay buffer holds every transition of the run (capacity $\geq$ total steps; nothing is evicted); sampling is uniform with replacement. The first 5k environment steps use uniform random actions. Environment: DeepMind Control (Tassa et al., 2018) with proprioceptive state observations, action repeat 2, episodes of 1000 environment steps (500 agent steps); no terminal state is ever signalled to the learner (time-limit truncation only). Evaluation: 10 episodes with the deterministic (mean) action on held-out environment seeds $10{,}000$–$10{,}009$, every 10k environment steps and additionally at $+2.5$k, $+5$k and $+7.5$k after the branch. Value, as registered, is the arithmetic mean of these 18 evaluation returns. Because the evaluation times are irregular, this mean weights the first 10k post-branch steps by $4/18$ rather than by $1/15$; it is therefore not a time-normalized area under the curve. All registered results use this definition. Appendix T recomputes the main results with a time-normalized trapezoidal AUC over environment steps and with the endpoint.

Branching. A trunk is trained for 150k environment steps at the history UTD and the complete state is saved: actor, critics, target critics, all three optimizer states, the entropy coefficient, the replay

buffer, and the NumPy and PyTorch RNG states. Every arm restores this state exactly, so arms of one seed see byte-identical environment streams and replay draws until the intervention changes the critic.

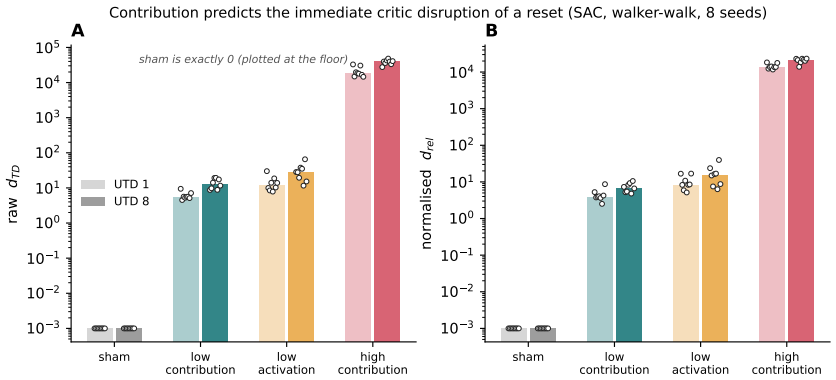


Figure 18: Immediate cost of a single matched reset (8 seeds, log scale, all seeds shown). **A** raw d_{TD}; **B** normalized by the pre-reset TD loss. The ordering `sham` < `lowC` < `lowA` < `highC` holds at both pressures and under normalization.

Scores. On a fixed audit batch of 4096 transitions drawn once from the checkpoint's buffer (generator seeded by the run seed), the contribution score is computed in microbatches of 8: for critic hidden unit i with incoming weights $W_{\text{in},i}$ and outgoing column $W_{\text{out},\cdot i}$, $c_i = \text{mean}_m \, |\langle \nabla_{W_{\text{in},i}} L_m, \; \theta_i^0 - W_{\text{in},i} \rangle - \langle \nabla_{W_{\text{out},\cdot i}} L_m, \; W_{\text{out},\cdot i} \rangle|$, where L_m is the twin-critic TD loss on microbatch m and θ^0 is the fresh draw the operator will use. The activation score is the ReDo statistic: mean $|h_i|$ over the audit batch divided by the layer mean.

Reset operator (all arms). With index set M_ℓ per layer ℓ and strength α ($\alpha = 1$ in every arm except the budget-scaled test of Appendix H; $\alpha = 0$ for `sham`):

```
theta0 ← fresh_draw(seed, event)  # Kaiming-uniform bound of
the layer, deterministic per (seed, event), shared by all
arms
for each resettable layer ℓ with incoming Linear W_in, b and
next Linear W_out:
W_in[M_ℓ, :] ← (1 − α) W_in[M_ℓ, :] + α θ_ℓ^0[M_ℓ, :]
b[M_ℓ] ← (1 − α) b[M_ℓ]
W_out[:, M_ℓ] ← (1 − α) W_out[:, M_ℓ]
Adam first and second moments of the same entries ← (1 − α)×
themselves
```

The mean paired amplification of the normalized cost d_{TD}/pre_{TD} from UTD 1 to UTD 8 is 1.7× for `lowC` (CI $[1.4, 2.1]$), 1.5× for `highC` (CI $[1.3, 1.7]$) and 1.9× for `lowA` (CI $[1.1, 2.9]$). Scores are computed in every arm, `sham` included, before the operator is applied; the immediate cost d_{TD} is the change in the audit-batch TD loss against a bootstrap target computed once before the reset; this is the batch on which the scores were computed. The calibration runs (Appendix A) repeat the measurement on a disjoint batch: over the 32 calibration checkpoints the score-decile masks are ordered identically by the two measurements (Spearman median 1.00, minimum 0.99) and the top-to-bottom decile cost ratio is 7,647 in-sample against 8,309 held out (per-checkpoint ratio median 0.97, range 0.75–1.28). Two details of the operator matter for how the score should be read. First, the score omits the bias term $-\nabla_{b_i} L \cdot b_i$ although the operator zeroes the bias: on the calibration checkpoints the bias term is 7% of the score's magnitude, the unit ranking with and without it correlates at 0.996, the bottom-K sets overlap at Jaccard 0.93, and the held-out ranking of random masks is 0.84 with the term against 0.83 without. Second, layers are processed in order, so when a layer-ℓ unit and a layer-$(\ell{+}1)$ unit are both reset, the entries of W_{out} at their intersection are first zeroed and then overwritten by the fresh row of the layer-$(\ell{+}1)$ unit (1352 of 131072 entries of the two inter-layer matrices at $K = 26$, about 1%). Because that unit's own outgoing column is zeroed by the same reset, the overwritten

entries do not affect the loss at the reset (the immediate cost is identical whether they are kept at zero or re-drawn, ratio 1.0000 on all 32 checkpoints); they do change the initial condition of the reset unit's subsequent learning, an effect on later dynamics that this diagnostic does not measure. The total displacement of a joint reset is therefore not exactly the sum of the per-unit displacements d_i used by the score, which is one more reason to treat c_i as a conditional sensitivity. In the repeated protocol the operator is re-applied every 10k environment steps after the branch (15 events over the 150k budget), re-scoring at each event and drawing a new θ^0 per event (still identical across arms).

DQN–Atari (Appendix E). The discrete-control cells are a separate code base and protocol, summarized here for completeness: the Nature-DQN convolutional network and preprocessing (Mnih et al., 2015), Adam, one game per cell, UTD $\in \{1, 2, 4, 8\}$ as gradient updates per environment step, and periodic recycling of $K = 128$ hidden units per event. Its full configuration is released with the code.