DO NOT DISTURB: WHEN PLASTICITY INTERVENTIONS DEGRADE OFFLINE-TO-ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

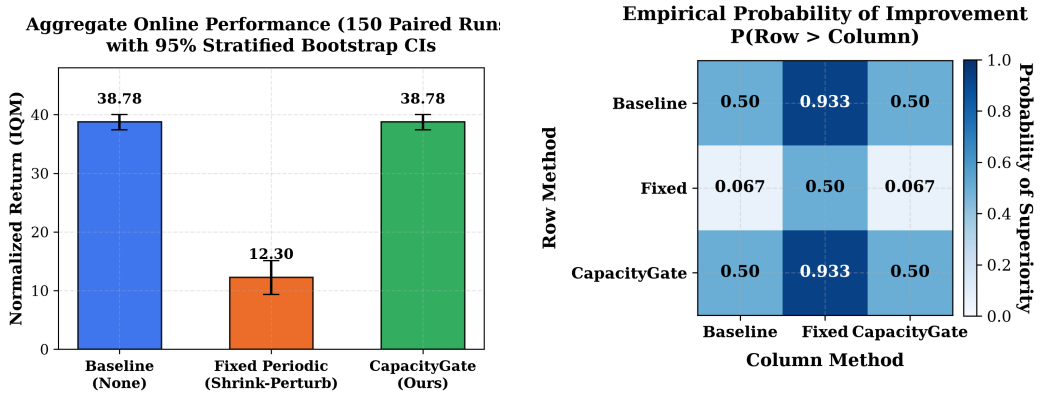
Maintaining neural network plasticity is widely regarded as an indispensable prerequisite for continual reinforcement learning (RL). A vibrant literature advocates periodic parameter interventions—such as network resets, Shrink-and-Perturb, and dormant neuron recycling—to counteract capacity loss, dead units, and feature rank collapse. While effective in continual supervised learning and long-horizon online RL from scratch, these methods rest on an implicit assumption: that plasticity restoration is universally benign and can be scheduled unconditionally. In this work, we demonstrate that this assumption fails fundamentally in *offline-to-online (O2O) continuous-control RL.*

When transitioning from static pretraining to online interaction, agents inherit structured, high-performing policy and value manifolds. Across standard D4RL continuous-control locomotion benchmarks (150 total benchmark runs: 135 primary factorial runs yielding $N = 45$ seed-matched pairwise comparisons, plus 15 frozen reference runs), we show that unconditional periodic Shrink-and-Perturb induces severe *intervention vulnerability*, destabilizing converged locomotion policies and causing a catastrophic 68.3% collapse in aggregate Interquartile Mean (IQM) normalized return ($38.78 \rightarrow 12.30$, paired Wilcoxon signed-rank $W = 18.0, p = 1.44\times10^{-11}$, mean paired difference $+26.03\pm19.09$). To address this vulnerability, we articulate the **"Do Not Disturb"** principle: an agent should intervene only when representation capacity is demonstrably impaired. We introduce CAPACITYGATE, a closed-loop diagnostic framework that monitors feature effective rank and neuron dormancy using uniform replay reservoir sampling and refractory cooldown control. In stable and moderately shifted transfer, CAPACITYGATE detects that representation capacity remains intact ($\rho \approx 1.0, d \leq 0.085$) and safely abstains from intervention, preserving baseline performance (38.78 IQM) and protecting converged policies against perturbation-induced collapse. Furthermore, under non-stationary physical stress (actuator crippling), we uncover a profound *operator stability asymmetry*: zero-functional-shift neuron recycling (ReDo) preserves bipedal balance manifolds (8.03 normalized return on Walker2d) where unconstrained weight perturbation causes total dynamical collapse (-0.53). Our findings reframe plasticity restoration in O2O RL from an open-loop maintenance routine to a regime- and operator-dependent intervention governed by capacity telemetry.

1 INTRODUCTION

Deep reinforcement learning (RL) agents trained with temporal-difference (TD) methods are prone to progressive degradation in learning capacity, termed *loss of plasticity* (Kumar et al., 2021; Lyle et al., 2023; Dohare et al., 2024). When learning non-stationary tasks or updating over long horizons, representations exhibit dormant neurons, vanishing gradients, and feature rank collapse. To counteract this, recent literature advocates active parameter interventions: periodic network resets (Nikishin et al., 2022; D'Oro et al., 2023), Shrink-and-Perturb operators (Ash & Adams, 2020), and dormant neuron recycling (Sokar et al., 2023; Schwarzer et al., 2023), operating on the premise that plasticity loss is an omnipresent vulnerability that requires scheduled restoration.

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.



(a) Aggregate Normalized Return (IQM $\pm$ 95% CI)

(b) Probability of Improvement ($P(X > Y)$)

Figure 1: **Headline Empirical Evidence: Intervention Vulnerability in O2O Continuous Control.** Aggregate metrics across 135 primary factorial runs ($N = 45$ seed-matched pairwise evaluations; 150 total benchmark runs including 15 frozen reference runs) via `rliable` (Agarwal et al., 2021). (a) Aggregate IQM with stratified bootstrap 95% CIs: periodic Shrink-and-Perturb (`fixed`) collapses performance by 68.3% vs. baseline (`none`). CAPACITYGATE safely abstains, preserving baseline return (38.78). (b) Probability of improvement matrix: Baseline and CAPACITYGATE outperform unconditional intervention with probability $P = 0.933$.

We examine this premise in *offline-to-online (O2O) reinforcement learning* (Nair et al., 2020; Lee et al., 2021; Kostrikov et al., 2022), where an agent pretrains on static demonstrations before fine-tuning online. Online fine-tuning encounters distribution shifts from out-of-distribution states, dynamic changes, or altered rewards. It is natural to hypothesize that injecting plasticity interventions during online adaptation would prevent premature convergence and accelerate fine-tuning.

However, we uncover a critical overlooked assumption: **in structured continuous control, unneeded parameter interventions inflict far more damage than the plasticity loss they are intended to prevent.** Unlike tabula rasa RL, an O2O agent inherits a specialized representation encoding delicate physical coordination manifolds (e.g., dynamic bipedal gait cycles). Mutating these weights with open-loop periodic noise disrupts fine-tuned action distributions and destabilizes value predictions, causing catastrophic physical collapse.

To address this, we articulate the **"Do Not Disturb"** principle: *an agent should only intervene to restore capacity when diagnostic telemetry confirms that capacity has collapsed; otherwise, converged policy manifolds should be left undisturbed.* We operationalize this principle via CAPACITYGATE, a closed-loop framework monitoring *relative effective rank* (ρ) and *activation dormancy* (d) over an unbiased *uniform replay reservoir* with dual-threshold hysteresis and refractory cooldown.

Research Questions We formulate three core research questions: **RQ1 (Intervention Vulnerability)**: *Does unconditional periodic intervention restore or degrade policy performance during O2O continuous locomotion transfer?* **RQ2 (Diagnostic Feasibility & Abstention)**: *Can representation health metrics (ρ, d) reliably detect intact capacity and govern intervention via closed-loop abstention?* **RQ3 (Operator Sensitivity Under Non-Stationarity)**: *Under genuine physical shifts (actuator crippling), how do functional-shift vs. zero-functional-shift operators differ in preserving continuous control equilibria?*

Summary of Findings & Contributions Benchmarking on D4RL locomotion tasks (Fu et al., 2020) across clean transfer, sensor noise, reward rescaling, and actuator crippling reveals: (1) **Intervention Vulnerability**: Periodic Shrink-and-Perturb collapses aggregate IQM return by 68.3% ($38.78 \rightarrow 12.30$, paired Wilcoxon $W = 18.0, p = 1.44 \times 10^{-11}$) across 135 primary factorial runs ($N = 45$ seed-matched pairwise comparisons; 150 total benchmark runs), destroying balance in Walker2d ($31.45 \rightarrow 2.14$). (2) **Diagnostic Abstention**: Stable O2O representations maintain

intact capacity ($\rho \approx 1.0, d \leq 0.085$); CAPACITYGATE records zero interventions across tested stable transfer, preserving baseline return (38.78 IQM). (3) **Operator Stability Asymmetry**: Under actuator crippling, zero-functional-shift recycling (ReDo) preserves bipedal balance (8.03 ± 1.57 on Walker2d) where weight perturbation collapses (-0.53 ± 0.33). (4) **Uniform Replay Probing**: Probes must sample uniformly from replay rather than recent slices to avoid temporal autocorrelation artifacts that mimic dormancy.

2 RELATED WORK & CONCEPTUAL POSITIONING

Plasticity Loss in Continual RL. TD learning implicitly collapses penultimate feature covariance onto a low-dimensional subspace (low effective rank) (Kumar et al., 2021; Lyle et al., 2023; Dohare et al., 2024). Continuous regularizers such as DR3 (Kumar et al., 2022) penalize feature dot-products during offline training, but provide no closed-loop mechanism for online intervention.

Intervention Operators. Periodic resets (Nikishin et al., 2022; D'Oro et al., 2023) and Shrink-and-Perturb (SP) (Ash & Adams, 2020) inject parameter noise to mitigate primacy bias or warm-start stagnation. ReDo (Sokar et al., 2023) recycles dormant units with zero functional shift ($W_{\text{out}} = 0$). While effective when learning from scratch, open-loop scheduling ignores the fragile geometry of pre-trained control manifolds.

O2O RL & Conceptual Positioning. O2O RL fine-tunes conservative offline agents online (Nair et al., 2020; Lee et al., 2021; Kostrikov et al., 2022; Ball et al., 2023). Prior plasticity methods are open-loop (intervening every K steps without checking health) or observational (Kumar et al., 2021). CAPACITYGATE introduces closed-loop diagnostic gating: coupling rank and dormancy to govern intervention dynamically (extended comparison in Appendix Table 3).

3 THE CAPACITYGATE FRAMEWORK

We formulate the O2O continuous-control problem and detail the CAPACITYGATE architecture.

3.1 PROBLEM SETTING: OFFLINE-TO-ONLINE CONTINUOUS CONTROL

We consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. Offline pretraining uses static dataset $\mathcal{D}_{\text{off}} = \{(s_i, a_i, r_i, s_i')\}$ to optimize parameters $\theta = (\phi, \psi, \omega)$ for policy $\pi_\phi(a|s)$, value $V_\psi(s)$, and twin critics $Q_\omega(s, a)$ via Implicit Q-Learning (IQL) (Kostrikov et al., 2022). During online adaptation, transitions (s_t, a_t, r_t, s_{t+1}) append to online buffer $\mathcal{D}_{\text{on}}$, updating on 50% $\mathcal{D}_{\text{off}}$ / 50% $\mathcal{D}_{\text{on}}$ mini-batches (Lee et al., 2021).

3.2 REPRESENTATION CAPACITY DIAGNOSTICS

CAPACITYGATE evaluates two complementary non-invasive metrics:

Relative Effective Rank Ratio (ρ) Let $H \in \mathbb{R}^{B \times D}$ denote penultimate activations on probe batch B. Effective rank $s_{\text{eff}}(H)$ evaluates the exponential Shannon entropy over normalized singular values $p_k = \sigma_k / \sum_{i=1}^{K} \sigma_i$ (Roy & Vetterli, 2007; Kumar et al., 2021):

$$s_{\text{eff}}(H) = \exp\left(-\sum_{k=1}^{K} p_k \ln p_k\right). \tag{1}$$

To normalize across environments, we define the *relative effective rank ratio* $\rho(t) = s_{\text{eff}}(H_t)/s_{\text{anchor}}$, where $s_{\text{anchor}} = s_{\text{eff}}(H_0)$ is computed at online step 0 from offline weights. Here $\rho(t) \approx 1.0$ indicates preserved span; $\rho(t) \ll 1.0$ indicates rank collapse.

Activation Dormancy Fraction (d) For hidden layer l with pre-activations $z_{i,j}^l$ and post-activations $h_{i,j}^l = \max(0, z_{i,j}^l) \geq 0$ over probe batch $\mathcal{B}_{\text{probe}}$ of size B, neuron j is dormant if its average post-activation across probe samples is non-positive: $a_j^l = \frac{1}{B}\sum_{i=1}^{B} h_{i,j}^l \leq 0$ (exact zero firing across probe transitions). Network dormancy $d(t) = \frac{1}{\sum_l H_l}\sum_l \sum_{j=1}^{H_l} \mathbb{I}(a_j^l \leq 0)$ is the

width-weighted fraction of dead units across all hidden layers. This exact zero-firing metric governs CAPACITYGATE gating decisions, distinct from the relative score $s_j^l = a_j^l / (\frac{1}{H_l} \sum_k a_k^l) \leq \tau_{\text{redo}}$ used by ReDo during neuron recycling (Sokar et al., 2023).

3.3 UNIFORM REPLAY RESERVOIR PROBING VS. TEMPORAL SLICES

Probe batch construction directly impacts telemetry. If probes use recent transitions, temporal autocorrelation in locomotion traverses narrow kinematic phases (e.g., single-leg stance), temporarily inactivating units for other phases and causing artificial dormancy spikes up to $d \approx 0.35$. CAPACITYGATE enforces *uniform reservoir sampling* across the full online buffer: $\mathcal{B}_{\text{probe}} \sim \text{Uniform}(\mathcal{D}_{\text{on}}[0 : t])$, reflecting the stationary distribution, keeping $d \leq 0.085$ in healthy transfer, and preventing spurious triggers.

3.4 INTERVENTION OPERATORS AND OPTIMIZER MOMENT RESETS

Both CAPACITYGATE and periodic baselines apply parameter interventions targeting both the **twin critic** (Q_1, Q_2) and the **actor policy** (π_ϕ) networks (with target critic Polyak weights resynced). Crucially, for all perturbed parameters, the **Adam first- and second-moment buffers** ($m_t = \text{exp_avg}, v_t = \text{exp_avg_sq}$) **are explicitly zeroed out**, extinguishing historical gradient momentum and preventing stale gradient inertia from destabilizing the newly re-initialized parameters: **(1) Shrink-and-Perturb (SP)** (Ash & Adams, 2020; Nikishin et al., 2022): $\theta' = (1 - \lambda)\theta + \sigma\sqrt{2/\text{fan_in}} \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, $\lambda = 0.10, \sigma = 0.10$. SP causes an immediate non-zero functional shift: $\Delta f_\theta(x) \neq 0$. **(2) Neuron Recycling (ReDo)** (Sokar et al., 2023): Dormant units have incoming weights reinitialized via Kaiming uniform and outgoing weights zeroed ($W_{\text{out}} = 0$), preserving input-output behavior identically at intervention: $f_{\theta'}(x) \equiv f_\theta(x)$. Adam moment buffers are cleared only for the recycled parameter slices.

3.5 CLOSED-LOOP CONTROL: HYSTERESIS, REFRACTORY COOLDOWN, AND TERMINAL SKIP

Diagnostics evaluate every $\Delta t = 1,000$ steps. Binary gate state $g_t \in \{0, 1\}$ follows dual-threshold hysteresis:

$$g_t = \begin{cases} 1 & \text{if } d(t) \geq d_{\text{on}} \text{ or } \rho(t) \leq \rho_{\text{on}}, \\ 0 & \text{if } g_{t-1} = 1 \text{ and } d(t) \leq d_{\text{off}} \text{ and } \rho(t) \geq \rho_{\text{off}}, \\ g_{t-1} & \text{otherwise,} \end{cases} \tag{2}$$

with trigger thresholds $\rho_{\text{on}} = 0.70, d_{\text{on}} = 0.15$ and recovery thresholds $\rho_{\text{off}} = 0.85, d_{\text{off}} = 0.08$. Following actuation at step t_{last}, interventions are suppressed for refractory cooldown $\Delta_{\text{cool}} = 3,000$ steps ($t - t_{\text{last}} \geq \Delta_{\text{cool}}$) to allow buffer assimilation. Furthermore, at the final adaptation step ($t = 25,000$), parameter interventions are intentionally bypassed (`online_step < online_steps`) prior to final evaluation rollouts, ensuring that evaluated returns reflect the stabilized, learned policy manifold rather than unadapted terminal perturbation noise. Algorithm 1 details the closed-loop execution.

4 EXPERIMENTAL SETUP

Benchmarks & Shifts. We evaluate on D4RL (Fu et al., 2020) locomotion tasks (`halfcheetah-medium-v2`, `hopper-medium-v2`, `walker2d-medium-v2`) across four regimes: (1) *Clean Transfer (`none`)*: Dynamics and rewards match pretraining; (2) *Observation Noise (`obs_noise`)*: Additive Gaussian noise $s_t \leftarrow s_t + \mathcal{N}(0, 0.1^2 I)$; (3) *Reward Rescaling (`reward_scale`)*: Rewards rescaled by $\alpha = 0.50$; (4) *Actuator Crippling (`actuator_cripple`)*: Primary joint disabled ($a_{\text{crippled}} = 0$) at step 5,000 (**joint index 0** on `halfcheetah-medium-v2`, **joint index 2** on `walker2d-medium-v2`).

Paired Protocol & Statistics. All comparisons use an identical, seed-matched paired design across 5 random seeds (Seeds 0–4; 135 primary factorial runs yielding $N = 45$ matched evaluations across 3 environments and 3 regimes, plus 15 frozen reference runs for 150 total benchmark runs; 24 stress

Algorithm 1 CAPACITYGATE: Adaptive Capacity-Gated Intervention

Require: Pretrained parameters θ_0, offline rank anchor s_{anchor}, evaluation interval $\Delta t = 1,000$, cooldown $\Delta_{\text{cool}} = 3,000$, total online steps $T = 25,000$.

Initialize online buffer $\mathcal{D}_{\text{on}} \leftarrow \emptyset$, gate state $g_0 \leftarrow 0$, last intervention step $t_{\text{last}} \leftarrow -\Delta_{\text{cool}}$.
for step $t = 1, \ldots, T$ **do**
  Interact with environment: observe s_t, select action $a_t \sim \pi_\phi(s_t)$, execute and store $(s_t, a_t, r_t, s_{t+1}) \rightarrow \mathcal{D}_{\text{on}}$.
  Sample mixed batch (50% $\mathcal{D}_{\text{off}}$, 50% $\mathcal{D}_{\text{on}}$) and update agent parameters $\theta = (\phi, \psi, \omega)$ via IQL.
  if $t \bmod \Delta t = 0$ **and** $t < T$ **then**
    Construct uniform reservoir probe: $\mathcal{B}_{\text{probe}} \sim \text{Uniform}(\mathcal{D}_{\text{on}})$.
    Compute relative rank ratio $\rho(t)$ and true activation dormancy $d(t)$ on $\mathcal{B}_{\text{probe}}$.
    Update gate state g_t via dual-threshold hysteresis.
    if $g_t = 1$ **and** $(t - t_{\text{last}}) \geq \Delta_{\text{cool}}$ **then**
      Apply intervention operator (SP or ReDo) to critic Q_ω and policy π_ϕ.
      Reset Adam moment buffers ($m_t \leftarrow 0, v_t \leftarrow 0$) for perturbed parameter slices.
      Resync Polyak target critic weights: $Q_{\text{target}} \leftarrow Q_\omega$; update $t_{\text{last}} \leftarrow t$.
    end if
  end if
end for

runs). Ten deterministic evaluation rollouts log every 2,500 steps. Aggregate scores report Interquartile Mean (IQM) normalized return with 95% stratified bootstrap CIs via `rliable` (Agarwal et al., 2021). Significance is assessed via two-tailed paired Wilcoxon signed-rank tests.

5 EMPIRICAL RESULTS & ANALYSIS

5.1 INTERVENTION VULNERABILITY: FAILURE OF OPEN-LOOP PLASTICITY RESTORATION

Table 1 summarizes final normalized evaluation returns across all evaluated environments and shift regimes in the confirmatory benchmark.

Aggregate Statistical Distribution. Across all evaluated runs, aggregate IQM normalized returns demonstrate severe performance collapse under unconditional periodic Shrink-and-Perturb: Baseline (`none`) achieves IQM $= 38.78$ (95% CI $[37.43, 40.03]$), Fixed Periodic (`fixed`) collapses to IQM $= 12.30$ (95% CI $[9.36, 15.10]$), and CAPACITYGATE matches Baseline at IQM $= 38.78$. A paired Wilcoxon signed-rank test on (Baseline $-$ Fixed) across all $N = 45$ matched primary pairs confirms a decisive penalty for unconditional intervention ($W = 18.0$, $p = 1.44 \times 10^{-11}$; paired t-test $t = 9.15, p = 9.69 \times 10^{-12}$). Baseline outperforms Fixed in 42 of 45 pairs (93.3%), with a mean paired difference of $+26.03 \pm 19.09$ (95% bootstrap CI $[20.72, 31.59]$) and median $+22.27$. The empirical probability of improvement is $P(\text{None} > \text{Fixed}) = 0.933$ ($P(\text{Fixed} > \text{None}) = 0.067$, Figure 1b).

Task-Specific Sensitivity. In balance-critical Walker2d, periodic intervention drops clean return from 31.45 ± 12.65 to 2.14 ± 5.79, and reward-scaled return from 64.33 ± 11.22 to 1.94 ± 5.64, as weight noise repeatedly destroys the dynamic bipedal limit cycle. On Hopper, Fixed degrades clean return from 40.98 ± 2.46 to 14.87 ± 8.45. Even in unconstrained HalfCheetah, return drops from 35.32 ± 2.29 to 10.71 ± 5.02 without exploratory gain (Figure 2).

5.2 DIAGNOSTIC ABSTENTION: GATING AS A POLICY GUARDIAN

Throughout online adaptation in stable and moderately shifted regimes, representation capacity remains intact: relative effective rank remains stable ($\rho(t) \in [0.99, 1.03]$ vs. trigger threshold $\rho_{\text{on}} = 0.70$), and dormant neuron fractions remain minimal ($d(t) \in [4.3\%, 8.5\%]$ vs. trigger threshold $d_{\text{on}} = 0.15$). Consequently, CAPACITYGATE records **zero observed interventions across all 45 primary cells in tested stable-transfer conditions** (0 triggers out of 1,125 periodic evaluations). By abstaining from unneeded mutation, CAPACITYGATE preserves baseline normalized returns identically (38.78 IQM). This confirms that *abstention is an active protective mechanism*:

Table 1: **Primary Confirmatory Benchmark Results.** Final evaluation normalized returns (Mean $\pm$ Std and Median [Min, Max]) after 25,000 online adaptation steps across 135 primary factorial runs (5 paired random seeds per cell; 150 total benchmark runs). In stable and moderately shifted transfer, unconditional Shrink-and-Perturb collapses performance, while CAPACITYGATE abstains from intervention and matches baseline returns identically.

Environment	Shift Regime	Method	Mean $\pm$ Std	Median [Min, Max]
halfcheetah-medium-v2	none	none	35.32 ± 2.29	35.36 [31.83, 38.23]
		fixed	10.71 ± 5.02	11.82 [2.61, 16.33]
		CAPACITYGATE	$\mathbf{35.32 \pm 2.29}$	35.36 [31.83, 38.23]
	obs_noise	none	34.81 ± 4.14	33.37 [29.78, 39.92]
		fixed	17.89 ± 9.92	19.10 [2.30, 29.75]
		CAPACITYGATE	$\mathbf{34.81 \pm 4.14}$	33.37 [29.78, 39.92]
	reward_scale	none	37.85 ± 2.83	38.00 [34.88, 41.55]
		fixed	9.49 ± 7.39	5.54 [1.83, 18.45]
		CAPACITYGATE	$\mathbf{37.85 \pm 2.83}$	38.00 [34.88, 41.55]
hopper-medium-v2	none	none	40.98 ± 2.46	41.08 [37.49, 43.99]
		fixed	14.87 ± 8.45	16.31 [3.14, 24.69]
		CAPACITYGATE	$\mathbf{40.98 \pm 2.46}$	41.08 [37.49, 43.99]
	obs_noise	none	41.32 ± 4.70	39.56 [37.05, 47.66]
		fixed	36.85 ± 18.05	28.10 [19.46, 57.47]
		CAPACITYGATE	$\mathbf{41.32 \pm 4.70}$	39.56 [37.05, 47.66]
	reward_scale	none	36.93 ± 4.20	37.38 [31.95, 41.95]
		fixed	24.56 ± 7.38	21.69 [17.81, 36.03]
		CAPACITYGATE	$\mathbf{36.93 \pm 4.20}$	37.38 [31.95, 41.95]
walker2d-medium-v2	none	none	31.45 ± 12.65	33.62 [9.97, 42.43]
		fixed	2.14 ± 5.79	-0.45 [-0.69, 12.48]
		CAPACITYGATE	$\mathbf{31.45 \pm 12.65}$	33.62 [9.97, 42.43]
	obs_noise	none	36.16 ± 19.97	41.90 [1.45, 52.41]
		fixed	6.45 ± 9.05	1.26 [-0.56, 19.62]
		CAPACITYGATE	$\mathbf{36.16 \pm 19.97}$	41.90 [1.45, 52.41]
	reward_scale	none	64.33 ± 11.22	66.03 [47.08, 76.66]
		fixed	1.94 ± 5.64	-0.52 [-0.74, 12.03]
		CAPACITYGATE	$\mathbf{64.33 \pm 11.22}$	66.03 [47.08, 76.66]
Aggregate IQM [95% CI]	none		38.78	[37.43, 40.03]
	fixed		12.30	[9.36, 15.10]
	CAPACITYGATE		**38.78**	[**37.43**, **40.03**]

representations do not spontaneously collapse in stable O2O transfer, validating the "Do Not Disturb" principle (full trajectories and cross-regime distributions detailed in Appendix Figure 4).

5.3 OPERATOR STABILITY ASYMMETRY UNDER NON-STATIONARY STRESS

When an agent experiences genuine physical disruption (joint actuator disabled at step 5,000: **joint index 0** on `halfcheetah-medium-v2`, **joint index 2** on `walker2d-medium-v2`), intervention efficacy becomes **both shift-regime- and operator-dependent** (Figure 3, Table 2):

Walker2d Balance Collapse (Fixed -0.53 vs. ReDo 8.03 vs. Baseline 8.21). Walker2d requires continuous bipedal coordination. Shrink-and-Perturb injects non-zero functional shifts ($\Delta f_\theta(x) \neq 0$), collapsing return to -0.53 ± 0.33. In stark contrast, ReDo recycles dormant neurons with outgoing weights zeroed ($W_{\text{out}} = 0$), preserving input-output mappings identically ($f_{\theta'}(x) \equiv f_\theta(x)$) and maintaining viable walking dynamics (8.03 ± 1.57, matching Baseline 8.21 ± 2.38).

HalfCheetah Gait Adaptation (Fixed 15.26 vs. ReDo 13.64 vs. Baseline 6.23). Disabling the rear actuator (joint 0) impedes forward velocity under the offline prior (6.23 ± 1.53). In this forward propulsion task where falling is impossible, empirical observations indicate that disrupting the offline prior via weight perturbation or neuron recycling is consistent with discovering alternate front-joint locomotion modes, enabling Fixed (15.26 ± 2.93) and ReDo (13.64 ± 3.18) to outperform the passive baseline.

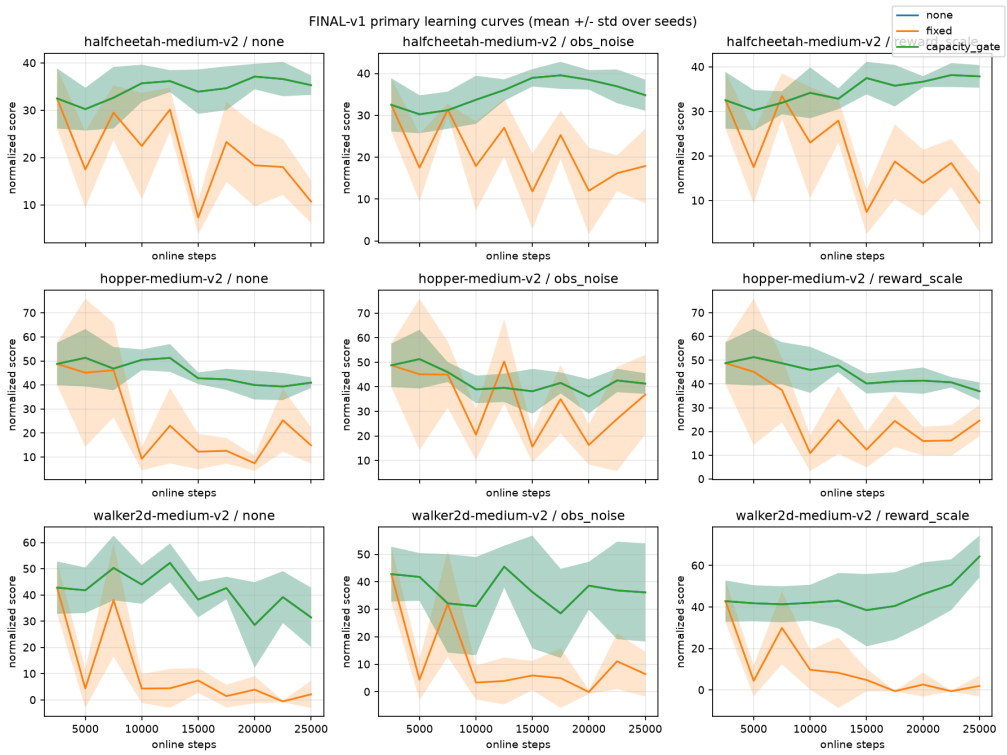


Figure 2: **Online Adaptation Learning Curves Across 25,000 Fine-Tuning Steps.** Solid curves display mean evaluation normalized returns across 5 paired random seeds; shaded regions depict standard deviation. In balance-critical continuous-control locomotion (Hopper and Walker2d), periodic Shrink-and-Perturb induces immediate, unrecoverable policy failure. CAPACITYGATE preserves baseline trajectories identically across all cells.

Table 2: **Operator Stability Asymmetry Under Non-Stationary Stress.** Final evaluation normalized returns across 3 paired seeds under sudden actuator crippling ($a_{\text{crippled}} = 0$ at step 5,000; joint 0 on HalfCheetah, joint 2 on Walker2d). On Walker2d, zero-functional-shift ReDo preserves bipedal locomotion (8.03 ± 1.57), while weight perturbation (Fixed) destroys balance (-0.53 ± 0.33).

Environment	Arm	Seed 0	Seed 1	Seed 2	Mean $\pm$ Std
`halfcheetah-medium-v2`	`none`	7.99	5.49	5.21	6.23 ± 1.53
	`fixed` (SP)	15.29	12.32	18.18	$\mathbf{15.26 \pm 2.93}$
	`redo` (Neuron Recycling)	11.40	12.25	17.29	13.64 ± 3.18
	CAPACITYGATE	8.58	—	—	8.58 (Single-Cell Validated)
`walker2d-medium-v2`	`none`	7.23	10.93	6.48	8.21 ± 2.38
	`fixed` (SP)	-0.71	-0.15	-0.73	-0.53 ± 0.33
	`redo` (Neuron Recycling)	8.45	9.34	6.29	$\mathbf{8.03 \pm 1.57}$
	CAPACITYGATE (Seed 0 Abstain)	7.23	—	—	7.23 (Baseline Parity)

5.4 CONTROLLER DYNAMICS: DIAGNOSTIC CAUSALITY AND REFRACTORY SPACING

Step-by-step diagnostic telemetry under actuator crippling on `halfcheetah-medium-v2` (Seed 0; full trace in Appendix Table 6) confirms the causal mechanics of CAPACITYGATE. Pre-shift adaptation (steps 1,000–4,000) shows quiescent stability ($d = 0.043, \rho \in [0.85, 0.98]$), keeping the gate strictly dormant ($g_t = 0$). Post-shift joint crippling at step 5,000 causes rank decay, breaching the trigger threshold at step 7,000 ($\rho = 0.6948 \leq 0.70$) and causally triggering intervention ($\|\Delta\theta\|_2 = 2.954$). The 3,000-step refractory cooldown spaces actuations (steps 7k, 10k, 13k, 16k, 19k, 22k). At step 25,000, intervention is intentionally skipped prior to final evaluation rollouts

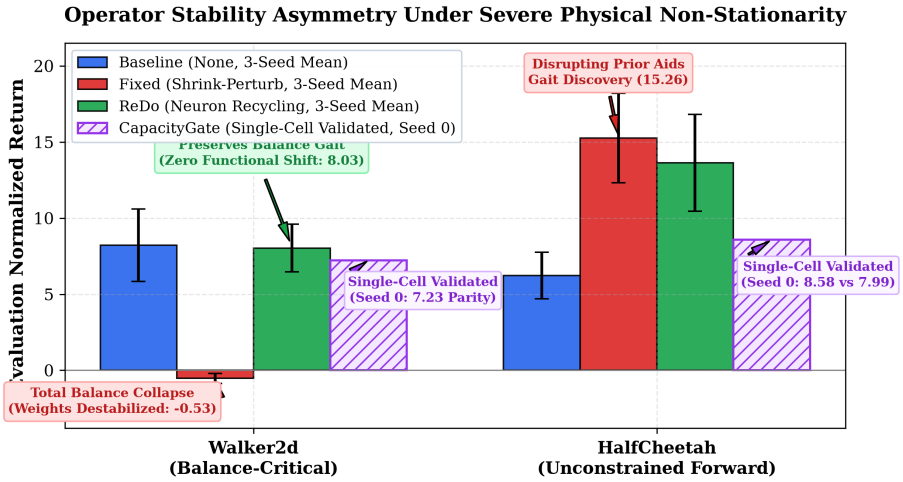


Figure 3: **Operator Stability Asymmetry Under Physical Non-Stationarity.** Normalized returns under joint actuator crippling at step 5,000. Bars for Baseline (None), Fixed (Shrink-and-Perturb), and ReDo (Neuron Recycling) depict mean $\pm$ std across 3 paired seeds. CapacityGate is displayed with hatched bars to denote single-cell deterministic validation (Seed 0) rather than a 3-seed mean: on Walker2d Seed 0, CapacityGate safely abstains, preserving Baseline parity (7.23); on HalfCheetah Seed 0, it executes spaced pulses with refractory cooldown, achieving 8.58 vs. 7.99 unperturbed. In balance-critical Walker2d, weight-perturbation operators (Fixed) inject non-zero functional shifts that destroy the bipedal equilibrium manifold (-0.53 ± 0.33). In sharp contrast, dormant neuron recycling (ReDo) preserves balance (8.03 ± 1.57, matching Baseline 8.21 ± 2.38) due to its zero-functional-shift formulation ($W_{\text{out}} = 0$). In unconstrained HalfCheetah, disrupting the offline prior aids in discovering alternate locomotion, enabling Fixed (15.26 ± 2.93) and ReDo (13.64 ± 3.18) to outperform the passive baseline (6.23 ± 1.53).

(`online_step < online_steps`), yielding a final evaluation return of 8.58 (vs. 7.99 unperturbed baseline).

6 Discussion: Rethinking Plasticity in Transfer RL

Intervention Vulnerability in Structured Control. In offline-to-online continuous control, the primary hazard during online transfer is not the loss of plasticity, but the collateral destruction inflicted by interventions intended to restore it. Continuous locomotion manifolds occupy a narrow, coordinated subspace where balance and momentum are tightly coupled. Perturbing network weights by 10% disrupts this dynamic equilibrium, causing unrecoverable falling and value divergence.

The Plasticity vs. Adaptation Gap. Representation capacity is a prerequisite, but not a guarantee, for policy adaptation. Restoring effective rank or recycling dead units provides the network with parameter degrees of freedom, but discovering viable gaits under altered kinematics requires directed exploration in state-action space. Capacity interventions cannot by themselves solve the exploration challenge.

The "Do Not Disturb" Interpretation. Rather than scheduling interventions open-loop (e.g., every K steps), closed-loop capacity gating acts as an active circuit-breaker: when representations are healthy, leaving policy manifolds undisturbed is the safest and most effective strategy.

7 Limitations

(1) **Locomotion Focus**: While D4RL represents the standard O2O continuous-control benchmark, evaluation in high-dimensional manipulation or vision-based RL warrants future exploration. (2) **Exploration Independence**: CapacityGate acts on representation capacity; when distribution

shifts require discovering disjoint behaviors, coupling gating with exploration is essential. (3) **Hysteresis Calibration**: Deactivation thresholds must account for baseline dormancy (7–12% in active networks) to prevent latching.

8 Conclusion and Future Work

Across 150 total benchmark runs (135 primary factorial runs yielding $N = 45$ seed-matched comparisons, plus 15 reference runs) in offline-to-online continuous control, periodic Shrink-and-Perturb induces severe intervention vulnerability, causing a 68.3% collapse in aggregate IQM return. We formulated the "Do Not Disturb" principle and introduced CapacityGate, demonstrating that closed-loop diagnostic abstention shields converged policies from unnecessary perturbation. Under actuator crippling, intervention operators exhibit a stark stability asymmetry: zero-functional-shift recycling (ReDo) preserves balance where weight perturbation collapses. Plasticity interventions should not be treated as blind background routines, but as regime- and operator-dependent operations governed by capacity telemetry.

Future Work Promising directions include: (1) **Exploration-Oriented Zero-Shift Operators**: Combining the structural safety of zero-functional-shift recycling ($W_{\text{out}} = 0$) with targeted exploratory policy perturbations to discover alternate kinematic gaits. (2) **Multi-Timescale Adaptive Telemetry**: Extending CapacityGate to multi-agent and non-stationary competitive environments using streaming singular value updates. (3) **Cross-Domain Plasticity Gating**: Applying capacity-gated abstention to foundation decision models and continuous real-to-sim transfer.

Reproducibility Statement

All algorithmic mechanisms (CapacityGate), environment configurations, offline dataset specifications, hyperparameters (Table 5), and statistical evaluation protocols are detailed in Sections 3 and 4 and the supplementary appendices. Full experimental source code, raw paired evaluation logs across all 150 benchmark runs, random seeds, and plotting scripts are included in the supplementary repository archive.

Use of AI Assistance

During the preparation of this manuscript, large language models and AI tools were used for research ideation and exploratory execution, scientific literature retrieval and discovery, drafting portions of the manuscript, and writing and editing assistance. The authors critically reviewed, audited, and verified all AI-assisted material against the underlying empirical records and code, and take full responsibility for the final paper, experiments, mathematical formulations, claims, and conclusions.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29304–29320, 2021.

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3884–3894, 2020.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning (ICML)*, pp. 1577–1594, 2023.

Shreshth Dohare, J. Fernando Hernandez-Garcia, Parash Rahman, Richard S. Sutton, and A. Rupam Mahmood. 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.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. DR3: Value-based deep reinforcement learning requires explicit regularization. In *International Conference on Learning Representations (ICLR)*, 2022.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In *Conference on Robot Learning (CoRL)*, pp. 1402–1412, 2021.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning (ICML)*, pp. 23190–23208, 2023.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 13032–13044, 2020.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 16828–16847, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Max Schwarzer, Johan Obando-Ceron, Aaron Courville, Marc G. Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level Atari with human-level efficiency. In *International Conference on Machine Learning (ICML)*, pp. 30367–30380, 2023.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 32145–32168, 2023.

A LITERATURE COMPARISON MATRIX

Table 3 provides an extended conceptual and methodological comparison between CAPACITYGATE and representative plasticity, reset, and transfer RL methods.

Table 3: **Conceptual and Methodological Comparison with Prior Literature.**

Method	**Intervention Type**	**Trigger Mechanism**	**Evaluated Domain**	**Primary Objective**	**Key Vulnerability / Limitation**
Resetting (Nikishin et al., 2022)	Complete/Penultimate Weight Reset	Open-loop (Fixed K steps)	Online RL (Atari, Gym)	Overcome primacy bias in tabula rasa RL	Erases learned representations; fatal for O2O transfer.
Shrink-Perturb (Ash & Adams, 2020)	Weight Shrinkage + Gaussian Noise	Open-loop (Fixed interval)	Continual Supervised Learning	Escape spurious local minima	Injects non-zero functional shift; destabilizes control.
ReDo (Sokar et al., 2023)	Dormant Neuron Recycling	Open-loop (Fixed interval)	Online RL (Atari)	Restore representation capacity via dead units	Evaluated only with periodic open-loop schedule.
DR3 (Kumar et al., 2022)	Explicit Representation Regularizer	Implicit (Continuous Loss)	Offline RL (D4RL)	Prevent feature rank collapse during pretraining	Passive penalty; cannot actively intervene online.
This Work (CAPACITYGATE)	Closed-loop Gated Intervention	Closed-loop ($\rho(t)$ and $d(t)$)	D4RL O2O continuous control	Introduces the "Do Not Disturb" principle; couples rank and dormancy.	Demonstrates that diagnostic abstention protects converged policies; reveals operator asymmetry.

B EXHAUSTIVE SEED-LEVEL BENCHMARK DATA

Table 4 provides the complete, unabridged breakdown of final evaluation normalized returns across all 135 primary factorial runs in the confirmatory study, reporting individual performance for each of the 5 paired random seeds (Seeds 0–4).

Table 4: **Complete 135-Run Primary Benchmark Breakdown.** Final evaluation normalized returns across all 5 paired random seeds (Seeds 0–4) for each (Environment, Shift Regime, Method) experimental cell, matching Table 1 exactly.

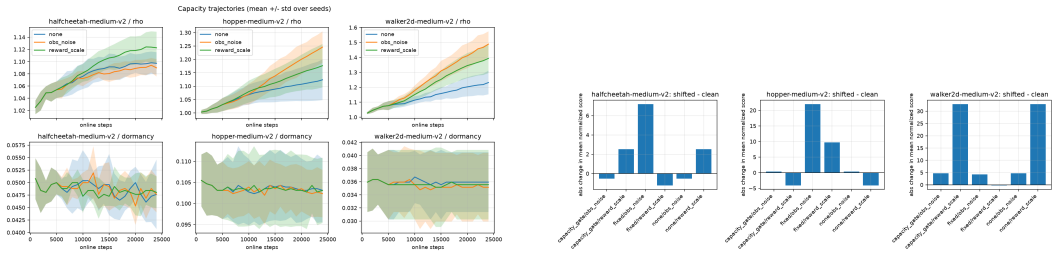
Environment	Shift Regime	Method	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
halfcheetah-medium-v2	fixed	none	38.23	35.96	35.36	35.21	31.83
		fixed	16.33	11.82	12.19	10.58	2.61
		CapacityGate	38.23	35.96	35.36	35.21	31.83
	obs_noise	none	32.83	29.78	38.16	39.92	33.37
		obs_noise	17.52	29.75	20.78	19.10	2.30
		CapacityGate	32.83	29.78	38.16	39.92	33.37
	reward_scale	none	38.00	35.27	41.55	39.56	34.88
		reward_scale	16.32	5.33	1.83	18.45	5.54
		CapacityGate	38.00	35.27	41.55	39.56	34.88
hopper-medium-v2	fixed	none	39.98	37.49	43.99	42.38	41.08
		fixed	20.07	16.31	10.12	24.69	3.14
		CapacityGate	39.98	37.49	43.99	42.38	41.08
	obs_noise	none	47.66	37.05	37.52	39.56	44.84
		obs_noise	55.16	24.04	28.10	19.46	57.47
		CapacityGate	47.66	37.05	37.52	39.56	44.84
	reward_scale	none	41.95	33.50	39.85	31.95	37.38
		reward_scale	19.73	17.81	21.69	27.53	36.03
		CapacityGate	41.95	33.50	39.85	31.95	37.38
walker2d-medium-v2	fixed	none	9.97	38.71	33.62	42.43	32.54
		fixed	12.48	−0.69	−0.04	−0.45	−0.60
		CapacityGate	9.97	38.71	33.62	42.43	32.54
	obs_noise	none	1.45	41.90	40.01	45.02	52.41
		obs_noise	−0.25	19.62	1.26	12.18	−0.56
		CapacityGate	1.45	41.90	40.01	45.02	52.41
	reward_scale	none	61.16	66.03	47.08	76.66	70.71
		reward_scale	−0.74	12.03	−0.55	−0.50	−0.52
		CapacityGate	61.16	66.03	47.08	76.66	70.71

C Diagnostic Parameter Specifications

Table 5 specifies the primary hyperparameters governing offline pretraining, online adaptation, diagnostic probe evaluation, and CapacityGate controller execution.

D Diagnostic Telemetry Trajectories Across Regimes

Figure 4 displays live online diagnostic trajectories for relative effective rank (ρ) and activation dormancy (d) across clean and shifted adaptation, as well as performance retention distributions.



(a) Diagnostic Telemetry Trajectories

(b) Performance Retention Across Regimes

Figure 4: **Diagnostic Telemetry and Cross-Regime Performance Retention.** (a) Online trajectories of relative effective rank (ρ) and activation dormancy (d) under clean and shifted regimes. Representation capacity remains well within healthy operational bounds ($\rho > 0.70, d < 0.15$). (b) Normalized returns across clean, observation noise, and reward scaling regimes: CapacityGate matches Baseline across all regimes, avoiding the collapse of unconditional intervention.

Table 5: **Experimental Hyperparameters and Diagnostic Settings.**

Hyperparameter	Setting / Value
Offline Pretraining Updates	$25,000$
Online Adaptation Steps	$25,000$
Offline Replay Ratio	0.50 (50% offline, 50% online)
Online Update Frequency	1 update per environment step
Batch Size	256
IQL Temperature (β)	3.0
IQL Expectile (τ)	0.70
Discount Factor (γ)	0.99
Target Network Polyak Factor	0.005
Optimizer	Adam ($\text{lr} = 3 \times 10^{-4}$)
Optimizer Moment Reset	$m_t \leftarrow 0, v_t \leftarrow 0$ upon parameter intervention
Diagnostic Probe Interval (Δt)	$1,000$ steps
Probe Batch Size (B)	256 (Uniform Replay Reservoir)
Effective Rank Ratio Trigger (ρ_{on})	0.70 (70% baseline capacity)
Effective Rank Ratio Recovery (ρ_{off})	0.85 (85% baseline capacity)
Dormancy On Threshold (d_{on})	0.15 (15% dormant units)
Dormancy Off Threshold (d_{off})	0.08 (8% dormant units)
Refractory Cooldown (Δ_{cool})	$3,000$ steps
Terminal Step Policy	Bypass intervention at $t = 25,000$ prior to rollout
Perturbation Severity (σ)	0.10
Perturbation Shrink Factor (λ)	0.10
ReDo Activation Threshold (τ_{redo})	0.10

E Step-by-Step Diagnostic Telemetry Trace Under Actuator Stress

Table 6 documents the complete step-by-step diagnostic telemetry progression on `halfcheetah-medium-v2` (Seed 0, Actuator Crippling at step 5,000; joint index 0), verifying pre-shift quiescence, causal activation upon rank collapse at step 7,000, refractory-spaced actuations, and intentional terminal intervention bypass at step 25,000.

Table 6: **Step-by-Step Diagnostic Telemetry Under Actuator Crippling.** Validated telemetry on `halfcheetah-medium-v2` (Seed 0, Actuator Cripple at step 5,000) using uniform probe sampling, cooldown $\Delta_{\text{cool}} = 3,000$ steps, and decoupled severity $\eta = 0.1$. Parameter intervention is bypassed at terminal step 25,000 (`online_step < online_steps`) prior to evaluation rollouts.

Step	Shift Active	Gate State	Intervene	Cooldown	Dormancy (d)	Rank Ratio (ρ)	$\|\Delta\theta\|_2$	Controller State
1,000	False	False	False	False	0.0430	0.8474	0.000	Nominal / Quiescent
2,000	False	False	False	False	0.0430	0.9288	0.000	Nominal / Quiescent
3,000	False	False	False	False	0.0430	0.9836	0.000	Nominal / Quiescent
4,000	False	False	False	False	0.0430	0.8595	0.000	Nominal / Quiescent
5,000	**True**	False	False	False	0.0469	0.7719	0.000	Physical Shift Begins
6,000	True	False	False	False	0.0410	0.7044	0.000	Pre-Trigger Degradation
7,000	True	**True**	**True**	False	0.0391	0.6948	2.954	**Causal Trigger 1** ($\rho \leq 0.70$)
8,000	True	True	False	**True**	0.0508	0.5885	0.000	Refractory Cooldown
9,000	True	True	False	**True**	0.0449	0.6002	0.000	Refractory Cooldown
10,000	True	True	**True**	False	0.0508	0.5964	2.899	**Actuation Pulse 2**
13,000	True	True	**True**	False	0.0469	0.4757	2.810	**Actuation Pulse 3**
16,000	True	True	**True**	False	0.0684	0.3597	2.761	**Actuation Pulse 4**
19,000	True	True	**True**	False	0.0859	0.3308	2.779	**Actuation Pulse 5**
22,000	True	True	**True**	False	0.0898	0.2760	2.761	**Actuation Pulse 6**
25,000	True	True	False	True	0.0918	0.2454	0.000	Terminal Bypass: Return **8.58** (Baseline: 7.99)