Hazard Compression: Catastrophic Forgetting in Diffusion-Based Generative Replay under Distribution Shift

Anonymous Authors[1]

Abstract

Diffusion models trained as generative replay buffers in reinforcement learning are vulnerable to a memorization failure we term *hazard compression*: as a Lagrangian safety penalty suppresses constraint-violating behavior, hazardous transitions vanish from the replay buffer, and the periodically retrained diffusion model catastrophically forgets the constrained region of state-action space. We demonstrate this failure in Prioritized Generative Replay (PGR) (Wang et al., 2025), where the diffusion model's hazard fidelity (measured by our diagnostic probe, DiffHz) collapses from 13.3% to 0.6% under Lagrangian optimization. A rare-event memory buffer that preserves hazardous transitions during diffusion retraining resolves this feedback loop, restoring DiffHz to 8.6% and reducing constraint violations by 99.8% on a velocity-constrained locomotion task. On a second task where the Lagrangian multiplier diverges due to integral windup—a mechanistically distinct failure confirmed by DiffHz remaining high—combined λ-warmup and gradient clipping fully recovers 99.6% of unconstrained reward while reducing cost by 76%. Together, DiffHz and the multiplier trajectory provide a lightweight diagnostic toolkit: low DiffHz signals generative forgetting; diverging λ signals control failure.

1. Introduction

Diffusion models are increasingly integrated into reinforcement learning (RL) pipelines—as generative models of temporal and visual data (Ho et al., 2022), as generative models finetuned via policy gradient on downstream rewards (Black et al., 2024), and as generative replay buffers that densify an agent's experience (Wang et al., 2025). A fundamental question for the foundations of deep generative models is: *under what conditions do these models catastrophically forget minority-class data when retrained on non-stationary distributions?*

We investigate this question in the context of Prioritized Generative Replay (PGR) (Wang et al., 2025), which periodically retrains a conditional diffusion model on the agent's accumulated replay buffer and mixes synthetic transitions with real experience, significantly improving sample efficiency over Soft Actor-Critic (SAC) (Haarnoja et al., 2018) on the DeepMind Control Suite (Tassa et al., 2018). When paired with Lagrangian constrained optimization (Tessler et al., 2019)—the standard approach for safe RL (Achiam et al., 2017; García & Fernández, 2015)—this retraining loop creates a pathological feedback cycle.

We identify *hazard compression*, a memorization failure wherein the diffusion model undergoes catastrophic forgetting of hazardous (constraint-violating) transitions. The mechanism is a feedback loop: (i) the Lagrangian penalty suppresses violations, (ii) hazardous transitions become rare in the replay buffer, (iii) the diffusion model is retrained and loses coverage of the hazardous region, (iv) Q-networks lose calibration at the constraint boundary, and (v) the Lagrangian multiplier λ surges to compensate for this miscalibration, closing the loop.

Hazard compression demonstrates a complementary phenomenon to recently studied diffusion model memorization and generalization dynamics (Buchanan et al., 2025; Pham et al., 2025; Zhang et al., 2025): where prior work has characterized phase transitions under static training sets, we show that *policy-induced distribution shift* in a continual retraining loop causes selective forgetting of a data subpopulation—even when that subpopulation is critical for downstream performance. This places hazard compression in the broader family of failure modes affecting generative models retrained on their own outputs or on non-stationary data streams (Shi et al., 2025; Otani, 2025).

We make four contributions: (1) We identify a new form of catastrophic forgetting in diffusion models—*hazard compression*—driven by policy-induced distribution shift in the training data, and characterize its feedback mechanism (Sec-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

tion 3.4). (2) We introduce DiffHz, a diagnostic probe that measures a diffusion model's retained fidelity to hazardous transitions at the *weight level*, preempting the counterargument that hazards are simply not being requested (Section 3.3). (3) We propose a rare-event memory buffer that resolves hazard compression, reducing violations by 99.8% (Section 3.2). (4) We show hazard compression is mechanistically distinct from integral windup, and that combined anti-windup mechanisms fully recover unconstrained reward while maintaining safety (Section 5.2).

2. Background

Diffusion model memorization. Recent work has established that diffusion models undergo phase transitions between memorization and generalization regimes depending on dataset size, model capacity, and training dynamics (Buchanan et al., 2025; Pham et al., 2025; Zhang et al., 2025). Hazard compression represents a new axis of this phenomenon: distribution shift *within* the training data—caused by policy improvement—selectively erases minority-class modes, analogous to catastrophic forgetting in continual learning (Otani, 2025) but driven by data distribution shift rather than task change.

Prioritized Generative Replay. PGR (Wang et al., 2025) builds upon REDQ-SAC (Chen et al., 2021)—an algorithm utilizing an ensemble of 10 Q-networks with an update-to-data (UTD) ratio of 20—by integrating a conditional diffusion model trained on replay buffer transitions. This diffusion model is periodically retrained and used to generate synthetic transitions, mixed with real experience at a 50% ratio. A curiosity-based conditioning signal prioritizes novel transitions during generation.

Constrained MDPs. We formulate safety as a Constrained MDP (Altman, 1999), where the agent maximizes expected return subject to a cost constraint

$$\mathbb{E}[\textstyle\sum_t c_t] \leq d, \tag{1}$$

for a cost signal c_t and limit d. The Lagrangian relaxation replaces the constrained objective with a penalized reward

$$r_{\text{eff}} = r_t - \lambda c_t, \tag{2}$$

where the multiplier $\lambda \geq 0$ is updated via dual gradient ascent after each episode (Tessler et al., 2019):

$$\lambda \leftarrow \max(0,\, \lambda + \eta(\bar{c} - d))\,. \tag{3}$$

Here η is the dual step size and $\bar{c}$ is the episode-average cost. This update rule is a pure integral controller: λ accumulates the error $(\bar{c} - d)$ with no decay, reset, or clipping.

3. Method

3.1. Constrained Environments

Cheetah-run (upper-bound constraint). We augment DMC Cheetah-Run (Tassa et al., 2018) with a binary cost signal: $c_t = \mathbb{1}[|v_t| > 7.0]$, where v_t is the root forward velocity. A random policy moves too slowly to violate this; violations arise only once the agent becomes competent. The introduction of synthetic cost signals atop reward-maximizing tasks is a standard evaluation protocol in safe RL (Achiam et al., 2017; Ray et al., 2019).

Walker-walk (minimum-viable-behavior constraint). $c_t = \mathbb{1}[|v_t| > 3.0 \vee h_t < 1.0]$, where h_t is the torso height and v_t is the forward velocity. The thresholds sit just outside the nominal operating range of the DMC Walker-walk task (default stand height 1.2 m; default walk speed 1 m/s): a competent walking policy satisfies both bounds, but a random policy does not. Under a random policy, 99% of timesteps violate the height constraint. Compliance requires first learning to balance, making the constraint infeasible during exploration.

3.2. Safety Extensions to PGR

Our full architecture combines two components, presented modularly for strict ablation.

Component 1: Lagrangian Penalty. The agent is trained on the penalized reward of Eq. 2. The multiplier λ is updated after each episode via Eq. 3 with $\eta = 0.01$ and cost limit $d = 2.0$; these correspond to a conservative dual step size and a tight safety target of $\leq 0.2\%$ violation rate per 1,000-step episode. While this theoretically enforces cost-awareness, it inadvertently triggers hazard compression when coupled with a diffusion model (Section 3.4).

Component 2: Rare-Event Memory Buffer. To prevent the generative model from forgetting the constraint boundary, we introduce a 500-slot FIFO buffer that archives all transitions where $c > 0$. During diffusion model retraining, 20% of the training batch is drawn from this buffer with maximum curiosity conditioning. The buffer also anchors 20% of each SAC training batch.

3.3. DiffHz: A Diagnostic Probe for Generative Hazard Fidelity

To quantify the diffusion model's retention of hazardous knowledge, we introduce the Diffusion Hazard Rate (DiffHz). After each diffusion retraining phase, we synthesize 2,000 transitions at high conditioning levels (top quartile of the curiosity conditioning signal, i.e., percentiles P75–P100) and calculate the percentage with reconstructed cost > 0.5.

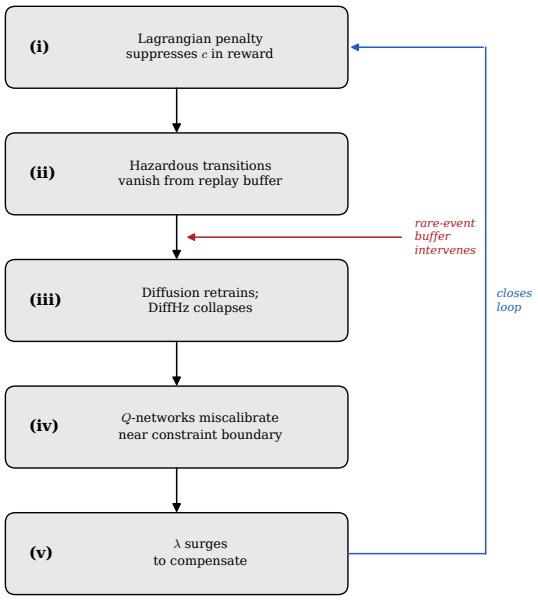

Figure 1. The hazard compression feedback loop. Numbered steps (i)–(v) show how Lagrangian penalization of cost shifts the replay-buffer distribution away from the constraint boundary, which the periodically retrained diffusion model then forgets—miscalibrating Q-networks and driving λ upward, which further suppresses boundary exploration. Our rare-event memory buffer intervenes between (ii) and (iii), preserving hazardous transitions across retraining cycles.

A DiffHz approaching 0% indicates hazard compression: the generative model has catastrophically forgotten the constrained regions. Critically, by probing at high conditioning levels, DiffHz tests whether the model *can* generate hazardous transitions when explicitly requested. A low DiffHz therefore demonstrates forgetting at the *model weight level*—preempting the counterargument that hazards are simply not being sampled because the curiosity signal no longer prioritizes them.

3.4. Hazard Compression: The Feedback Loop

Hazard compression is a distributional shift failure uniquely specific to generative replay. In standard off-policy RL, experience replay acts as an expanding reservoir: historical hazardous transitions remain available regardless of the current policy. Because PGR periodically retrains its diffusion model on the current buffer contents, the generative model's output distribution tightly tracks recent policy behavior.

The loop (Figure 1) operates as follows: the Lagrangian penalty suppresses violations, starving the diffusion model of the transitions it needs to represent the constraint boundary; the retrained model loses hazard coverage (DiffHz collapses), Q-networks miscalibrate near the boundary, and the multiplier λ surges to compensate for that miscalibration rather than for the original violation rate. Elevated λ further suppresses boundary exploration, closing the loop. Our rare-event memory buffer intervenes between steps (ii) and (iii), preserving hazardous transitions across retraining cycles.

We situate this mechanism within the broader literature on diffusion model memorization and self-consuming generative retraining in Section 6.

3.5. Anti-Windup Mechanisms

To address integral windup on Walker-walk, we introduce two modifications to the λ update rule: (1) **λ-warmup**: λ is held at zero for the first $W{=}20$ episodes, preventing accumulation during the random-exploration phase when cost is near-maximal. (2) **Gradient clipping**: $|\Delta\lambda|$ is capped at 0.1 per episode, bounding the rate of integral accumulation regardless of error magnitude.

4. Experimental Setup

All experiments use the PGR codebase with REDQ-SAC (UTD=20, batch size 256, 1M replay buffer). Each run trains for 100,000 environment steps ($\sim$100 episodes of 1,000 steps). Code is available at `https://anonymous.4open.science/r/submission_1`. We evaluate across 3 random seeds (42, 123, 456) on two environments:

- **SAC:** Standard REDQ-SAC without diffusion.
- **PGR:** Unconstrained PGR with diffusion replay.
- **PGR+L:** PGR with Lagrangian penalty only.
- **PGR+L+Buffer (Ours):** Full method. On Walker-walk, we additionally test three anti-windup variants: +Warmup (λ frozen for first 20 episodes), +Clip ($|\Delta\lambda| \leq 0.1$), and +WarmClip (both).

The comparison between PGR+L and PGR+L+Buffer is a controlled ablation: the only difference is the rare-event buffer.

Statistical methods. With $n = 3$ seeds, non-parametric rank tests lack power (minimum $p = 0.10$). We use Welch's t-test and supplement with one-sided exact permutation tests under directional hypotheses, which provide valid finite-sample inference even at small n: with 6 total samples across two conditions, the exhaustive permutation distribution has 20 possible label assignments, yielding minimum one-sided $p = 0.05$. Bootstrap 95% confidence intervals (100K resamples) provide distributional evidence.

Table 1. Cheetah-run results (mean ± std, last 10 episodes, 3 seeds). Significance markers report tests against our full method (PGR+L+Buffer) on safety-relevant cells (cost and DiffHz): $^{\dagger}p = 0.05$ (one-sided permutation); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (Welch's t).

Method	Reward	Ep. Cost	DiffHz	λ
SAC	291±20	0.0±0.0	N/A	–
PGR	679±30	546±70**†	13.3%	–
PGR+L	573±6	4.1±2.1†	0.6%***†	3.08
PGR+L+Buffer	**561**±14	**1.1**±0.4	**8.6%**	**0.80**

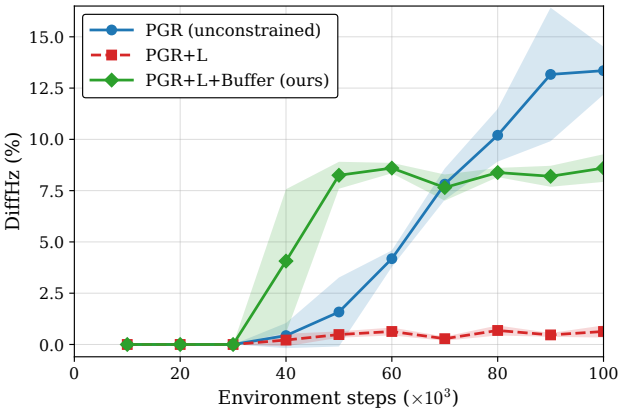

Figure 2. DiffHz over training on Cheetah-run (mean ± 1 std, 3 seeds; ten probe points, one per diffusion retrain). All methods stay near zero for the first ∼30K steps because the policy is not yet skilled enough to breach 7 m/s. After that, the trajectories diverge sharply. Unconstrained PGR climbs toward the environment's true hazard rate (∼13%). PGR+L flatlines below 1%: hazardous transitions vanish from the retraining set and the diffusion model forgets the boundary (*hazard compression*). PGR+L+Buffer sustains DiffHz at ∼8% throughout training, as the rare-event buffer anchors hazardous transitions across retraining cycles.

5. Results

5.1. Cheetah-run: Hazard Compression and Its Resolution

PGR amplifies unsafe behavior. Unconstrained PGR achieves 2.3× the reward of SAC (679 vs. 291) but incurs an episode cost of 546. SAC's safety is incidental: it learns too slowly to breach 7.0 m/s.

Lagrangian alone causes hazard compression. PGR+L reduces episode cost by 99.2% (546 → 4.1). However, DiffHz collapses from 13.3% to 0.6% ($p < 0.01$, one-sided permutation $p = 0.05$); Figure 2 shows the full collapse trajectory. As the agent avoids the boundary, the replay buffer starves the diffusion model of hazardous transitions. The multiplier surges to $\lambda = 3.08$ to maintain safety via blunt penalization.

The rare-event buffer breaks the compression loop. The buffer preserves DiffHz at 8.6% vs. 0.6% ($p < 0.001$, one-sided permutation $p = 0.05$; Figure 2). Q-network calibration improves, allowing λ to drop by ∼4× (3.08 → 0.80). This enables an additional 73% reduction in violations (4.1 → 1.1, one-sided permutation $p = 0.05$). Task reward remains statistically indistinguishable between the two safe variants ($p = 0.36$).

Table 2. Walker-walk results (mean ± std, last 10 episodes, 3 seeds). +WarmClip recovers 99.6% of unconstrained PGR reward.

Method	Reward	Ep. Cost	DiffHz	λ
SAC	855±126	178±184	N/A	–
PGR	933±37	109±102	88.0%	–
PGR+L	179±26	27±3	73.7%	246.8
PGR+L+Buffer	194±7	30±9	74.2%	224.4
+Warmup	219±17	25±8	82.1%	79.5
+Clip	808±118	23±5	85.0%	8.7
+WarmClip	**929**±20	**26**±6	**91.2%**	**6.2**

5.2. Walker-walk: Integral Windup and Its Resolution

Walker-walk provides a critical control condition for the DiffHz probe: DiffHz remains at 74–91% across all Walker variants, confirming that the diffusion model does *not* forget hazardous transitions. The failure lies in the Lagrangian optimizer, not the generative model.

Integral windup. The Lagrangian update is a pure integral controller. Under a random policy, 99% of timesteps violate the height constraint (episode cost ≈ 990), incrementing λ by ∼9.9 per episode. By episode 20, $\lambda \approx 200$. Even once the agent learns to balance (cost ≈ 26), unwinding from $\lambda \approx 224$ requires the multiplier to decrease by $\eta(d - \bar{c}) = 0.02$ per episode at best (when $\bar{c} = 0$), so ∼11,000 episodes—over 100× the training budget. Both PGR+L and PGR+L+Buffer collapse task reward by ∼80%.

DiffHz confirms mechanistic separation. DiffHz remaining at 74–91% across Walker variants while Cheetah's collapses to 0.6% validates the probe as a differential diagnostic: it detects memorization failure specifically, and correctly indicates its absence when the pathology lies elsewhere.

Anti-windup ablation. The ablation reveals that both mechanisms are independently necessary for full recovery. Warmup alone reduces λ from 224 to 80 but only recovers 13% of reward (194 → 219), as post-warmup accumulation remains unbounded. Clipping alone bounds λ at 8.7 and recovers substantial reward (808), but early-episode damage persists. Combined, +WarmClip achieves 929 ± 20 reward—statistically indistinguishable from unconstrained PGR (933 ± 37, $p = 0.91$)—while reducing cost by 76% (109 → 26) with $\lambda = 6.2$.

DiffHz reveals a secondary effect. DiffHz increases monotonically with anti-windup strength (74% → 82% →

85% $\rightarrow$ 91%). Lower λ permits freer exploration, generating diverse transitions including hazardous ones. Anti-windup does not just improve reward—it improves the diffusion model's state-space coverage.

6. Discussion

Connection to memorization and model collapse. Hazard compression is a policy-mediated form of self-consuming generative retraining (Shi et al., 2025): the diffusion model's outputs shape Q-networks, which shape the policy, which determines which transitions enter the next retraining set. Where prior memorization work has characterized phase transitions under static training sets (Buchanan et al., 2025; Pham et al., 2025; Zhang et al., 2025), we show a complementary failure mode driven by distribution shift *within* training. The Walker-walk result adds nuance: DiffHz rising from 74% to 91% under anti-windup indicates that the Lagrangian multiplier itself shapes generative coverage—overly aggressive penalization suppresses policy diversity and indirectly degrades diffusion fidelity even without hazard compression per se.

Beyond diffusion? The mechanism is not diffusion-specific. Any generative model periodically refit to a policy-dependent buffer should face the same pressure, though dynamics may differ across architectures. Empirical confirmation is important future work.

DiffHz as a template. DiffHz instantiates a general class of *conditional fidelity probes*: test whether a DGM retains a mode at the weight level by sampling at extreme values of its conditioning signal. Such probes could be deployed wherever a DGM is suspected of selectively forgetting minority modes.

Limitations. Our evaluation spans two environments with 3 seeds, limiting statistical power. Extending to additional environments and constraint types is important future work. More principled anti-windup approaches such as PID Lagrangian controllers (Stooke et al., 2020) could provide theoretically grounded alternatives to our heuristic warmup and clipping.

7. Conclusion

We identify hazard compression—a catastrophic forgetting failure mode in diffusion-based generative replay under Lagrangian safety constraints—and introduce DiffHz, a diagnostic probe for detecting it. A rare-event memory buffer resolves hazard compression on Cheetah-run, achieving 99.8% cost reduction while preserving 83% of task reward. On Walker-walk, DiffHz correctly diagnoses a mechanistically distinct failure (integral windup), and combined λ-warmup with gradient clipping fully recovers 99.6% of unconstrained reward while reducing cost by 76%. DiffHz and the λ trajectory together provide a lightweight diagnostic toolkit: low DiffHz signals generative forgetting; diverging λ signals control failure. As diffusion models become prevalent in RL pipelines, understanding their memorization failures under non-stationary training distributions is critical for safe deployment.

Impact Statement

This paper presents work whose goal is to advance the understanding of failure modes in diffusion-based generative models applied to reinforcement learning. We identify a safety-relevant failure (hazard compression) and provide diagnostic tools and mitigations. The broader impact is positive: our work contributes to safer deployment of generative models in control systems.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 22–31, 2017.

Altman, E. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Buchanan, S., Pai, D., Ma, Y., and De Bortoli, V. On the edge of memorization in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chen, X., Wang, C., Zhou, Z., and Ross, K. W. Randomized ensembled double Q-learning: Learning fast without a model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

García, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Otani, A. Mitigating catastrophic forgetting and mode collapse in text-to-image diffusion via latent replay. *arXiv preprint arXiv:2509.10529*, 2025.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Ray, A., Achiam, J., and Amodei, D. Benchmarking safe exploration in deep reinforcement learning. Technical report, OpenAI, 2019. `https://cdn.openai.com/safexp-short.pdf`.

Shi, L., Wu, M., Zhang, H., Zhang, Z., Tao, M., and Qu, Q. A closer look at model collapse: From a generalization-to-memorization perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight.

Stooke, A., Achiam, J., and Abbeel, P. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 9133–9143, 2020.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Wang, R., Frans, K., Abbeel, P., Levine, S., and Efros, A. A. Prioritized generative replay. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Zhang, Z., Li, X., Li, X., Tao, M., and Qu, Q. Generalization of diffusion models arises from a regularized representation space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.