Hazard Compression: Catastrophic Forgetting in Diffusion-Based Generative Replay for Safe Reinforcement Learning

Anonymous Authors[1]

Abstract

Generative replay accelerates online RL by training a generative model on accumulated experience and densifying that experience with synthetic transitions—an emerging paradigm for sample-efficient online learning. We show that this paradigm has a critical safety failure mode under constrained optimization: when a Lagrangian 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—a phenomenon we term *hazard compression*. We demonstrate this failure in Prioritized Generative Replay (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 for diffusion-based generative replay: low DiffHz signals generative forgetting; diverging λ signals control failure.

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

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

1. Introduction

Generative replay is an emerging paradigm in online reinforcement learning: a generative model is periodically retrained on accumulated experience, and synthetic transitions from this model are mixed with real experience during policy updates (Wang et al., 2025). The diffusion model effectively serves as a learned, continuously updated synthetic experience buffer that densifies the agent's experience and improves sample efficiency over standard online methods (Haarnoja et al., 2018). As such methods move toward deployment in safety-critical domains, a key question arises: *do generative replay buffers preserve the data needed for safe behavior under constrained optimization, or does the act of training the agent to be safe corrupt the buffer's coverage of unsafe states?*

We show the answer is the latter, in a way that creates a critical safety failure. We investigate Prioritized Generative Replay (PGR) (Wang et al., 2025)—which uses a conditional diffusion model as its generative buffer and substantially improves sample efficiency on the DeepMind Control Suite (Tassa et al., 2018)—paired with Lagrangian constrained optimization (Tessler et al., 2019; Achiam et al., 2017; García & Fernández, 2015), the standard approach for safe RL. This combination creates a pathological feedback cycle.

We identify *hazard compression*: the Lagrangian penalty suppresses violations, hazardous transitions become rare in the replay buffer, the diffusion model retrains and loses coverage of the constrained region, Q-networks miscalibrate at the boundary, and the Lagrangian multiplier λ surges to compensate—further suppressing exploration of the hazardous region. The result is an agent that appears safe by reward-cost trade-off but whose replay buffer can no longer represent the safety boundary it is constrained against.

This failure mode is specific to generative replay (standard experience replay retains historical transitions regardless of current policy), and has implications for the safe use of diffusion-based generative replay in safety-constrained settings. We make four contributions: (1) We identify hazard compression and characterize its feedback mechanism (Section 3.4). (2) We introduce DiffHz, a lightweight diagnostic probe that measures a diffusion model's retained

fidelity to hazardous transitions at the *weight level* (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—a separate failure of the Lagrangian optimizer—validated by DiffHz, and that combined anti-windup mechanisms fully recover unconstrained reward while maintaining safety (Section 5.2).

2. Background

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: $c_t = \mathbf{1}[|v_t| > 7.0]$, 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). $c_t = \mathbf{1}[|v_t| > 3.0 \vee h_t < 1.0]$, 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.

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 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 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.

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/new_submit-97C4`. 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 $\pm$ 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$\pm$20	0.0$\pm$0.0	N/A	–
PGR	679$\pm$30	546$\pm$70**†	13.3%*†	–
PGR+L	573$\pm$6	4.1$\pm$2.1†	0.6%***†	3.08
PGR+L+Buffer	**561**$\pm$14	**1.1**$\pm$0.4	**8.6%**	**0.80**

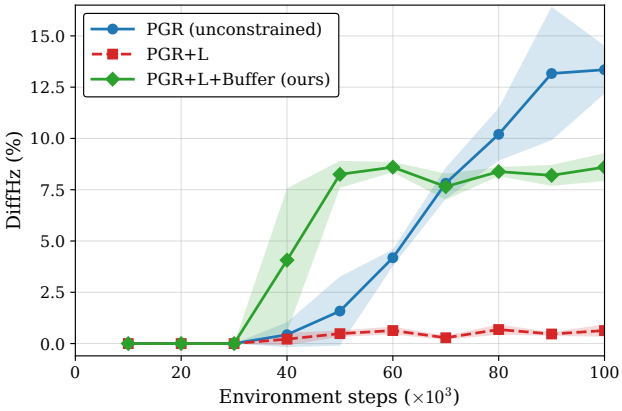

Figure 1. DiffHz over training on Cheetah-run (mean $\pm$ 1 std, 3 seeds; ten probe points, one per diffusion retrain). All methods stay near zero for the first $\sim$30K steps because the policy is not yet skilled enough to breach 7 m/s. After that, the trajectories diverge sharply. Unconstrained PGR climbs toward the environment's true hazard rate ($\sim$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 $\sim$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$\times$ 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 $\rightarrow$ 4.1). However, DiffHz collapses from 13.3% to 0.6% ($p < 0.01$, one-sided permutation $p = 0.05$); Figure 1 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

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	$855_{\pm 126}$	$178_{\pm 184}$	N/A	–
PGR	$933_{\pm 37}$	$109_{\pm 102}$	88.0%	–
PGR+L	$179_{\pm 26}$	$27_{\pm 3}$	73.7%	246.8
PGR+L+Buffer	$194_{\pm 7}$	$30_{\pm 9}$	74.2%	224.4
+Warmup	$219_{\pm 17}$	$25_{\pm 8}$	82.1%	79.5
+Clip	$808_{\pm 118}$	$23_{\pm 5}$	85.0%	8.7
+WarmClip	$\mathbf{929}_{\pm 20}$	$\mathbf{26}_{\pm 6}$	**91.2%**	**6.2**

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 1). Q-network calibration improves, allowing λ to drop by $\sim 4\times$ ($3.08 \rightarrow 0.80$). This enables an additional 73% reduction in violations ($4.1 \rightarrow 1.1$, one-sided permutation $p = 0.05$). Task reward remains statistically indistinguishable between the two safe variants ($p = 0.36$).

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, episode cost ≈ 990, incrementing λ by ~ 9.9 per episode—by episode 20, $\lambda \approx 200$. Once balanced (cost ≈ 26), unwinding requires λ to decrease by at most $\eta(d - \bar{c}) = 0.02$ per episode (when $\bar{c} = 0$): $\sim$11,000 episodes, over $100\times$ the training budget. Both PGR+L and PGR+L+Buffer collapse task reward by $\sim 80\%$.

DiffHz confirms mechanistic separation. The contrast—Walker DiffHz at 74–91% vs. Cheetah collapsing to 0.6%—validates DiffHz as a differential diagnostic: it specifically detects generative forgetting, not optimizer pathologies.

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 ($194 \rightarrow 219$), 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% ($109 \rightarrow 26$) with $\lambda = 6.2$. DiffHz also rises monotonically with anti-windup strength ($74\% \rightarrow 91\%$), suggesting that overly aggressive penalization itself indirectly degrades buffer coverage.

6. Discussion

Implications for generative replay deployment. Generative replay buffers are increasingly used to densify limited interaction data and accelerate online learning. Our results indicate that under safety-constrained optimization, this class of methods has a distinct failure mode that classical experience replay does not: the buffer's coverage is coupled to the policy's behavior, so improving safety can erase the very data needed to evaluate it. Hazard compression is conceptually related to recent work on memorization phase transitions in diffusion models (Buchanan et al., 2025; Pham et al., 2025; Zhang et al., 2025) and self-consuming generative loops (Shi et al., 2025), but the failure here is policy-mediated and arises within a single training run rather than from static training-set properties. Practitioners deploying diffusion-based replay in safety-critical pipelines should monitor weight-level hazard fidelity (e.g., DiffHz or analogues), not only headline cost metrics.

Generality. DiffHz instantiates a general class of *conditional fidelity probes*—testing whether a generative model retains a mode at the weight level by sampling at extreme conditioning values—and the hazard compression mechanism is not diffusion-specific: any generative model periodically refit to a policy-dependent buffer should face the same pressure. Empirical confirmation across architectures is important future work.

Limitations. Our evaluation spans two environments with 3 seeds, limiting statistical power. Extending to additional environments, constraint types, and other generative architectures 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 safety failure in diffusion-based generative replay where Lagrangian-driven safety pressure erases the buffer's coverage of the constraint boundary—and introduce DiffHz, a lightweight diagnostic probe. A rare-event memory buffer resolves the failure on Cheetah-run (99.8% cost reduction, 83% reward retention); 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%.

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.

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.

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.