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

Wei Gao Ellie Du* Akeel Majeed*

University College London

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: collapsing 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 that an agent can consume (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, substantially 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 multiplier surges to compensate, closing the loop. We measure steps (i)–(iii) and (v) directly; the intervening Q-calibration step (iv) is inferred rather than directly measured (see Limitations).

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., 2026): 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 characterize *hazard compression*—a policy-induced, selective forgetting of

*Equal contribution

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

a safety-critical minority subpopulation in a continually retrained diffusion model—and its feedback mechanism (Section 3.4), situating it alongside known memorization and continual-forgetting phenomena rather than as an entirely new effect. (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).[1]

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., 2026). 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 $\mathbb{E}[\sum_t c_t] \leq d$ for a cost signal c_t and limit d. The Lagrangian relaxation replaces the constrained objective with $r_{\text{eff}} = r_t - \lambda c_t$, where $\lambda \geq 0$ is updated via dual gradient ascent: $\lambda \leftarrow \max(0,\ \lambda + \eta(\bar{c} - d))$ after each episode (Tessler et al., 2019). This update rule is a pure integral controller: λ accumulates the error $(\bar{c} - d)$ with no decay, reset, or clipping.

[1]Code, configs, and all result logs: `https://github.com/Uninterpretable-Evolving-Blackbox/hazard-compression`

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 $r_{\text{eff}} = r - \lambda c$. The multiplier λ is updated after each episode via dual gradient ascent ($\eta = 0.01$, cost limit $d = 2.0$); these correspond to a conservative dual step size and a tight safety target of $\approx 0.2\%$ 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 (the top quartile of the curiosity conditioning signal) and calculate the percentage with reconstructed cost > 0.5.

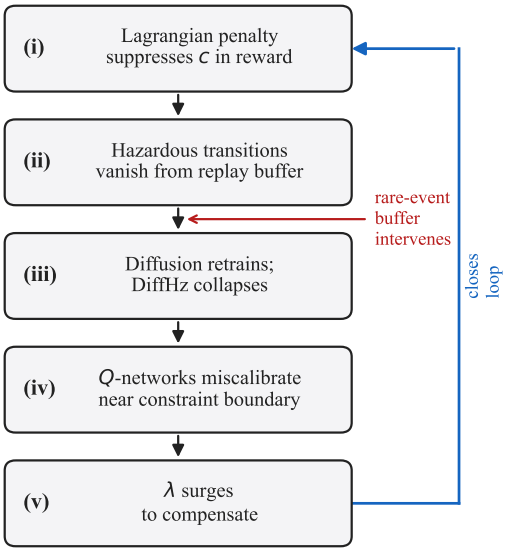

Figure 1. The hazard compression feedback loop. Numbered steps (i)–(v) show how Lagrangian penalization of cost shifts the replay-buffer distribution away from the constraint boundary, which the periodically retrained diffusion model then forgets—miscalibrating Q-networks and driving λ upward, which 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. Probing at high conditioning reflects PGR's curiosity signal, which up-weights *novel* transitions: on an upper-bound constraint (Cheetah) hazards are rare and novel, concentrating at high conditioning, so a low DiffHz sensitively indicates forgetting; under a minimum-viable-behavior constraint (Walker), hazards are common under early policies, so sustained high DiffHz reflects abundant rather than novel coverage. DiffHz is thus interpreted together with the task's conditioning–hazard correspondence.

3.4. Hazard Compression: The Feedback Loop

Hazard compression is a distributional-shift failure characteristic of 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 feedback 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); we infer that Q-networks then 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 diffusion retraining cycles.

This mechanism is a concrete instance of how diffusion models behave under non-stationary training distributions, connecting to work on model collapse in self-consuming generative loops (Shi et al., 2025) and catastrophic forgetting in continually retrained generative models (Otani, 2025).

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). 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

Table 1. Cheetah-run results (mean ± std, last 10 episodes, 3 seeds). $^{\dagger}p = 0.05$ (permutation); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (Welch's t) vs. our method.

Method	Reward	Ep. Cost	DiffHz	λ
SAC	$291_{\pm 20}$	$0.0_{\pm 0.0}$	N/A	–
PGR	$679_{\pm 30}$	$546_{\pm 70}{}^{**\dagger}$	13.3%	–
PGR+L	$573_{\pm 6}$	$4.1_{\pm 2.1}{}^{*\dagger}$	$0.6\%^{***\dagger}$	3.08
PGR+L+Buffer	**561**$_{\pm 14}$	**1.1**$_{\pm 0.4}$	**8.6%**	**0.80**

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 exact permutation tests, 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 $p = 0.05$. Bootstrap 95% confidence intervals (100K resamples) provide distributional evidence. Given $n = 3$, we lead with effect sizes and bootstrap CIs and report p-values only as directional support.

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 ∼22× (13.3% → 0.6%; bootstrap 95% CIs $[12.0, 14.8]$ vs. $[0.4, 1.1]$, non-overlapping; 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% $[8.1, 9.6]$ vs. 0.6% $[0.4, 1.1]$ (non-overlapping; 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). Task reward remains statistically indistinguishable between the two safe variants (573 → 561, $p = 0.36$): the buffer's safety gain comes at negligible task-reward cost.

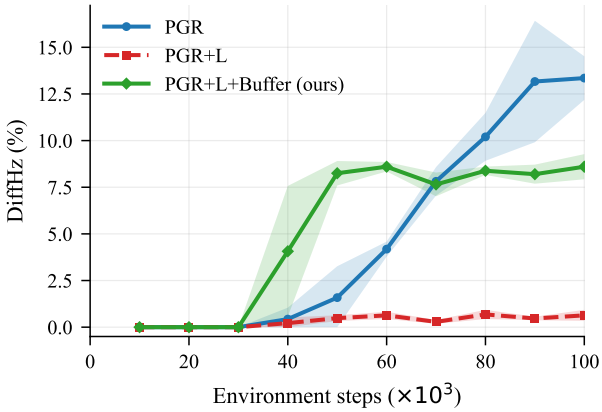

Figure 2. DiffHz over training on Cheetah-run (mean ± 1 std, 3 seeds; ten probe points, one per diffusion retrain). All methods stay near zero for the first ∼30K steps because the policy is not yet skilled enough to breach 7 m/s. After that, the trajectories diverge sharply. Unconstrained PGR climbs 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.

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	**929**$_{\pm 20}$	**26**$_{\pm 6}$	**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% throughout training, 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 λ to decrease by at most $\eta(d - \bar{c}) = 0.02$ per episode (when $\bar{c} = 0$), i.e. ∼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% 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 ($194 \rightarrow 219$), 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 reveals a secondary effect. DiffHz increases monotonically with anti-windup strength (74% $\rightarrow$ 82% $\rightarrow$ 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., 2026), 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? We conjecture 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. We do not verify this across architectures; empirical confirmation (e.g., a VAE or GMM replay buffer, or a controlled refit on a shrinking-support distribution) 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; we accordingly emphasize effect sizes and bootstrap CIs over p-values. Extending to additional environments, constraint types, native safety benchmarks (e.g., Safety-Gym), and other generative architectures is important future work. We attempted to measure the boundary Q-miscalibration (step iv) directly, partitioning the Monte-Carlo Q-bias by distance to the constraint boundary, but the signal was inconclusive at our evaluation scale: near-boundary states are rare, leaving that bin sparse and noisy, and we observed no boundary-specific elevation that distinguished the DiffHz-collapsed from the DiffHz-preserved condition. We therefore retain (iv) as an inferred step. A controlled comparison against a non-generative hazard archive, to isolate the role of the generative refit, is likewise natural 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 mode in diffusion-based generative replay under Lagrangian safety constraints, and introduce DiffHz to distinguish it from integral windup. Together, collapsing DiffHz and diverging λ form a lightweight diagnostic toolkit for safety-constrained diffusion-based RL.

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 learn-

ing. 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. *arXiv preprint arXiv:2505.21777*, 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.

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., Shi, L., Wu, M., Tao, M., and Qu, Q. Generalization of diffusion models arises with a balanced representation space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.