When Does Non-Uniform Replay Matter in Reinforcement Learning?

Michal Korniak[† 1,2] **Mikołaj Czarnecki**[† 2] **Yarden As**[1]
Piotr Miłoś[2] **Pieter Abbeel**[3,4] **Michal Nauman**[2,3]

[1]ETH Zurich [2]University of Warsaw [3]UC Berkeley [4]Amazon FAR

[†]Equal contribution

Abstract

Modern off-policy reinforcement learning algorithms often rely on simple uniform replay sampling and it remains unclear *when and why* non-uniform replay improves over this strong baseline. Across diverse RL settings, we show that the effectiveness of non-uniform replay is governed by three factors: *replay volume*, the number of replayed transitions per environment step; *expected recency*, how recent sampled transitions are; and the *entropy* of the replay sampling distribution. Our main contribution is clarifying when non-uniform replay is beneficial and providing practical guidance for replay design in modern off-policy RL. Namely, we find that non-uniform replay is most beneficial when replay volume is low, and that high-entropy sampling is important even at comparable expected recency. Motivated by these findings, we adopt a simple Truncated Geometric replay that biases sampling toward recent experience while preserving high entropy and incurring negligible computational overhead. Across large-scale parallel simulation, single-task, and multi-task settings, including three modern algorithms evaluated on five RL benchmark suites, this replay sampling strategy improves sample efficiency in low-volume regimes while remaining competitive when replay volume is high.

1 Introduction

Experience replay is one of the central components of off-policy RL [21, 33]. Many replay sampling strategies have been proposed, from uniform replay to distributions biased by error [29] or recency [37]. Yet recent continuous-control algorithms still often use uniform replay by default [4, 22, 16, 24]. Understanding when this choice is sufficient, and when it can be improved, is especially important in regimes such as parallel simulation, multi-task learning, and other settings where data collection outpaces optimization. In these regimes, individual transitions may be replayed only a few times, so the replay sampling distribution can have a larger effect on learning.

A natural way to analyze replay is to count how often each transition is reused over the course of training. This whole-training exposure view underlies prior analyses of recency-biased replay [37, 38]: changing the sampler changes how replay updates are distributed across transitions over their lifetime. However, whole-training exposure conflates two distinct factors. A learner can consume more recent data either by shifting the sampling distribution toward newer transitions, or by increasing the number of replayed transitions per environment step through a larger update-to-data ratio (UTD) or batch size. Moreover, whole-training exposure does not fully explain empirical behavior observed in off-policy RL. In runs where replay capacity is comparable to the training horizon, exposure can

Corresponding author: michael.korniak@gmail.com
Code available at https://github.com/M-Korniak/when-does-replay-matter

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea (ICML 2026). Workshop: RLxF: Reinforcement Learning from World Feedback.

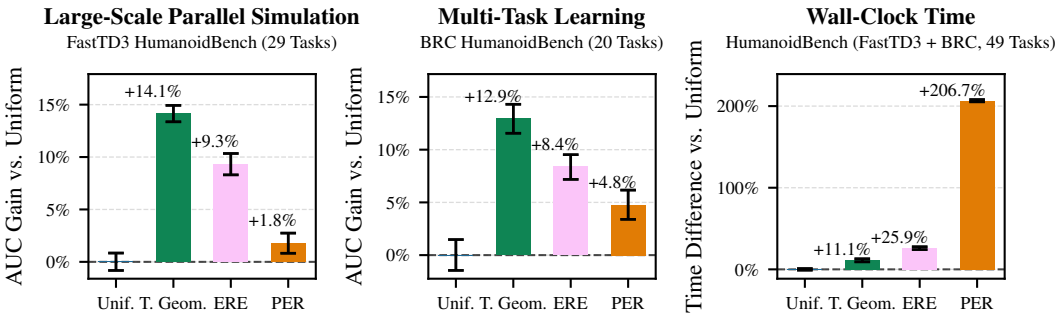

Figure 1: ***Performance and runtime trade-offs on HumanoidBench.*** Relative sample efficiency gains measured as area under (learning) curve (AUC) for single-task (*left*) and multi-task BRC (*middle*), and wall-clock time difference aggregated across both settings (*right*) compared to uniform sampling. Error bars show 95% stratified bootstrap CI. We report results in large-scale parallel FastTD3 and multitask BRC because they feature relatively low experience reuse per transition, where non-uniform replay is expected to have the greatest impact (Subsection 3.1). The recency-biased truncated geometric (described in detail in Section 4) sampling yields consistent gains with minimal overhead, while PER exhibits higher computational load. Experiment details are provided in Section 5. We refer to Section E.1 for discussion on AUC as a suitable metric for *sample efficiency*.

be highly skewed, yet non-uniform replay often provides little benefit [22, 17]. Conversely, as we show in this paper, in high-throughput simulation, exposure can be nearly flat throughout training, yet recency-biased replay can substantially outperform uniform sampling. These observations suggest that whole-training exposure alone does not predict when non-uniform replay improves performance.

In this paper, we expand on this perspective by showing that replay performance also depends on factors that whole-training exposure does not account for. This view separates three factors: **expected recency** captures how recent sampled transitions are on average; **replay volume**, given by UTD times batch size, captures how many replayed transitions are consumed between consecutive environment steps; and **sampling entropy** captures how concentrated the replay sampling distribution is at each update. This decomposition separates changing the sampler from changing the amount of optimization, and distinguishes recency bias from the diversity of replayed samples.

Our empirical study reveals two main findings about replay in RL. First, the benefit of recency bias is strongly mediated by replay volume: when replay volume is small, biasing replay toward recent transitions yields substantial gains [38]. Conversely, when replay volume is large, the same bias offers limited benefit. Second, expected recency alone does not determine performance: smoothly biased, high-entropy samplers outperform strategies that obtain comparable expected recency by restricting the support of the replay sampling distribution. Together, these findings explain why non-uniform replay is especially effective in optimization-limited regimes such as multi-task learning and large-scale parallel simulation, while uniform replay remains competitive when data reuse is already high.

Motivated by these observations, we adopt a simple recency-biased replay sampler based on a truncated geometric distribution. The sampler directly targets the regime identified by our analysis: it increases expected recency smoothly while preserving high sampling entropy, requires no costly priority recomputation, and admits constant-time sampling. Across our experiments, it improves sample efficiency over uniform, priority-biased, and recency-biased replay in low-replay volume regimes (Figure 1) while remaining effectively neutral when replay volume is high. This makes it a practical replacement for uniform replay in off-policy RL, especially when optimization is limited relative to data collection. We list our contributions below:

- We provide a practical guide to replay sampling in modern off-policy RL through a large-scale empirical study across single-task, multi-task, and large-scale parallel training regimes.
- We propose a per-step view of replay that separates expected recency, replay volume, and sampling entropy. This decomposition shows that recency-bias matters most when replay volume is low, while high sampling entropy remains important even at comparable expected recency.
- We show that a simple truncated geometric performs competitively across diverse benchmarks, improving sample efficiency over uniform and priority-based replay in low-replay volume regimes with negligible computational overhead, while remaining neutral in high-replay volume regimes.

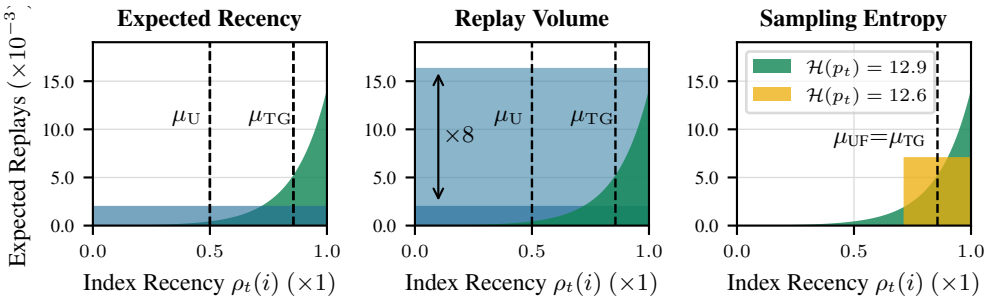

Figure 2: ***Effect of expected recency, replay volume, and sampling entropy.*** (*left*) By biasing samples toward recent data, the Truncated Geometric sampler (Section 4) achieves substantially higher expected recency than uniform replay. (*middle*) Increasing replay volume through UTD or batch size can make uniform replay match or exceed recency-biased sampling in the number of updates applied to recent transitions. (*right*) Replay distributions can have the same expected recency but different sampling entropy. The Truncated Geometric sampler achieves recency through a smooth full-support bias, increasing sampling entropy relative to a uniform buffer with the same expected recency; this is comparable to increasing the uniform sampling buffer size by roughly one third.

2 Background

Off-Policy RL. We consider infinite-horizon MDPs, where an agent interacts with an environment by selecting actions according to a policy and receives scalar rewards [26, 35]. We focus on off-policy deep RL methods, which improve sample efficiency by reusing data collected under previous policies [21, 9, 11]. By decoupling data collection from optimization, off-policy algorithms allow transitions to be leveraged across many gradient updates and have enabled applications including robotic control and real-world online learning during deployment [34, 30, 2]. Replay is critical in deep RL: it enables data reuse and helps to mitigate instability from the "deadly triad" [35, 21].

Replay Sampling Distributions. Many RL methods rely on first-in-first-out (FIFO) uniform replay, which is simple and effective across a wide range of tasks [19, 12, 31]. A substantial body of prior work has explored *non-uniform replay sampling*, in which transitions are sampled with unequal probability [14, 3, 10]. Representative approaches include TD-error-biased Prioritized Experience Replay [PER, 29] or Reliability-Adjusted Prioritized Experience Replay [ReaPER, 25], which bias sampling based on temporal-difference error, and recency-biased Emphasizing Recent Experience [ERE, 37], which biases sampling toward recent transitions. While non-uniform replay has shown improvements in some settings, it often introduces additional complexity, and it remains unclear in which regimes they consistently improve over uniform replay [22].

Recency-Biased Sampling. A key motivation for recency-biased replay is that newer transitions are more related to the current policy and as such may contain more informative signal, particularly when the agent adapts to new regions of the state space [38]. Prior work analyzes this through the lens of whole-training exposure: the expected number of times a transition is sampled while it remains eligible for replay. Crucially, when the buffer size is larger than the total number of steps the agent collects over its lifetime, older transitions are sampled disproportionately more often than newer ones. ERE addresses this, by skewing back the sampling distribution towards recent transitions [37, 38]. While the exposure framing intuitively motivates recency-biased methods, we argue it can obscure factors that matter during optimization: two sampling strategies can share identical exposure yet differ greatly in performance, or exposure can be skewed while the choice of strategy barely matters. In contrast to prior work, we analyze replay per environment step, separating expected recency, replay volume, and sampling entropy.

3 What Matters in Non-Uniform Sampling?

Prior work has shown that increasing replay recency can improve learning in some settings [37, 38]. Our goal is therefore not to re-establish the importance of recency, but to identify the factors that determine *when* recency improves over other replay sampling distributions.

Let p_t denote the replay sampling distribution at environment step t over buffer indices $i = 0, \ldots, N$, where $i = 0$ denotes the oldest sample and N is the total size of the buffer. Let N_t be the number of

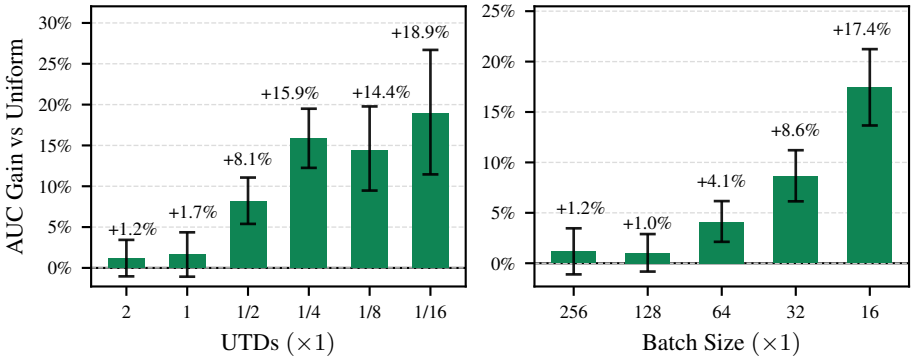

Figure 3: ***Replay volume matters.*** Improvement of recency-biased sampling over uniform replay as replay volume is varied through UTD (*left*) and batch size (*right*) (Section 3.1). In both panels, reducing replay volume increases the advantage of recency-biased replay: when replay volume is high, both methods perform similarly, whereas when replay volume is low, recency-biased sampling yields substantially larger gains. These results show that the benefit of recency bias depends on how much replay is performed per environment step. As such, the sampling distribution is especially important in optimization-limited regimes, such as multi-task learning or large-scale parallel simulation.

transitions collected up to time t. W define the normalized recency of index i as $\rho_t(i) = i/(N_t - 1)$, where larger values correspond to more recent transitions. We define the **expected recency** of the replay sampling distribution p_t as $\mu_t = \mathbb{E}_{i\sim p_t}[\rho_t(i)]$. This quantity captures how recent the sampled transitions are on average. Higher expected recency means that the learner updates more often on the recent experience. Unlike whole-training exposure, which aggregates counts over a transition's lifetime, expected recency describes the "age" composition of the update at a single training step.

To determine the aggressiveness of optimization at a given timestep, we define **replay volume** as $\mathrm{UTD}_t \times B_t$, that is, the number of replayed transitions between two consecutive environment steps. Increasing replay volume increases the amount of optimization performed on the stored data without changing the sampling distribution itself. Thus, replay can change along two separable axes: expected recency determines *where* replay samples come from, while replay volume determines *how many times* a given transition is consumed. Importantly, when replay volume is high, even low-expected recency samplers can still update on many recent transitions; when replay volume is low, the sampling distribution plays a larger role in determining whether recent transitions are replayed at all.

Furthermore, expected recency does not fully characterize the shape of the replay sampling distribution. Two strategies can have the same expected recency while differing substantially in how concentrated their sampling probabilities are. We capture this concentration using the instantaneous **sampling entropy** $\mathcal{H}(p_t)$. As illustrated in the right panel of Figure 2, a sampler can obtain high expected recency by sampling only from a recent window, yielding low entropy, or by smoothly biasing probabilities over the full buffer, yielding higher entropy. These strategies may have similar expected recency, but differ in update diversity and optimization dynamics.

This per-step decomposition guides the rest of our analysis. In Section 3.1, we vary replay volume while keeping the replay sampling distributions fixed, showing that recency bias helps most when replay volume is low and has limited effect when replay volume is high. In Section 3.2, we compare replay strategies that bias sampling toward recent data but differ in sampling entropy, showing that recency alone is not sufficient: high-entropy, full-support sampling is also important. Together, these experiments demonstrate that separating expected recency, replay volume, and sampling entropy, accounts for the observed behavior across different off-policy RL regimes.

3.1 Does Replay Volume Matter?

We first test whether replay volume modulates the effectiveness of recency-biased sampling.

Experiment. To isolate the role of replay volume, we vary UTD and batch size while keeping the replay buffer and replay sampling distributions fixed. Varying UTD changes the number of gradient updates performed per environment step, while varying batch size changes the number of transitions consumed per update. Both therefore change replay volume without changing the expected recency or sampling entropy induced by a fixed sampling strategy. Both UTD and batch size are

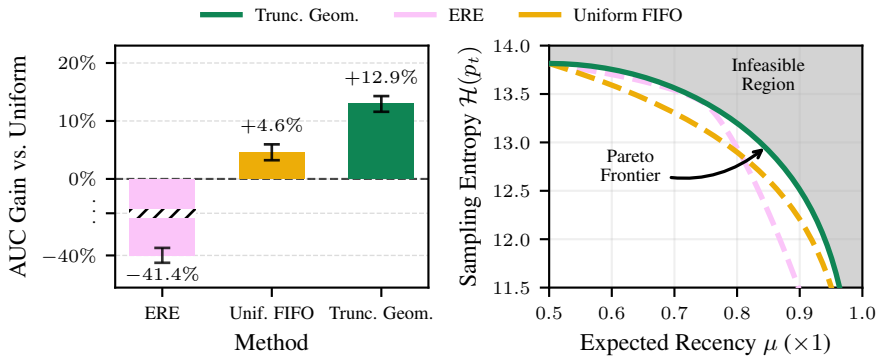

Figure 4: ***Sampling entropy matters.*** (*left*) ERE, Uniform FIFO (300k), and Truncated Geometric are matched to the same expected recency ($\mu \approx 0.85$), yet performance differs substantially with the sampling entropy. At this μ, ERE, has the lowest entropy and falls below the Uniform baseline (see Appendix E.4 for discussion of ERE shortcomings), while Truncated Geometric, which has the highest entropy at this μ, performs best. Interestingly, Uniform FIFO (300k) also outperforms the larger Uniform baseline (1M) despite using a smaller buffer. (*right*) The solid curve shows the maximum-entropy frontier for a 1M buffer: among all discrete replay sampling distributions with fixed expected recency, the Truncated Geometric distribution achieves the highest possible sampling entropy. The shaded region above the curve is infeasible. ERE and Uniform FIFO trace lower-entropy tradeoffs across hyperparameter settings, showing that they obtain recency less efficiently than Truncated Geometric; we discuss entropy interpretation in Appendix E.3.

practically relevant hyperparameters and both reflect compute-data tradeoffs in modern off-policy RL systems [27, 8]. As such, contemporary algorithms operate across a wide range of UTD regimes and batch size settings.* We compare uniform replay to recency-biased Truncated Geometric sampling using SimbaV2 [18] on 13 HumanoidBench tasks [32]. We sweep UTD values $\{2, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ and batch sizes $\{256, 128, 64, 32, 16\}$. As UTD or batch size decreases, replay volume decreases, entering regimes where each transition is replayed relatively few times. We detail the setup in Appendix C.

Results. As shown in Figure 3, uniform and recency-biased replay perform similarly at high UTD values and large batch sizes. In this regime, recency-biased replay has higher expected recency than uniform replay, but replay volume is also large, so even uniform replay updates on many recent transitions. Consequently, increasing expected recency provides limited additional benefit. As UTD or batch size decreases, however, the performance gap widens, with recency-biased replay increasingly outperforming uniform replay. Since the sampling strategies remain fixed across the sweep, their relative expected recency and sampling entropy stay unchanged, while replay volume decreases. The observed shift therefore shows that the benefit of recency bias depends critically on replay volume: when replay volume is limited, the replay sampling distribution becomes more consequential; when replay volume is large, changing expected recency has little effect. These results show that replay volume modulates when non-uniform replay improves over uniform replay.

Takeaway 1 – Replay Volume Matters

The benefit of increasing expected recency depends on replay volume. When replay volume is high, uniform replay already samples many recent transitions, so recency offers limited benefits. When replay volume is low, the sampling distribution becomes more consequential, and recency-biased replay can yield substantial gains. Thus, the replay distribution is especially important in optimization-limited regimes such as multi-task learning or large-scale parallel simulation.

3.2 Does Sampling Entropy Matter?

The preceding subsection shows that replay volume modulates the effectiveness of sampling. We now study the second factor: whether replay strategies with comparable expected recency can nevertheless differ in performance due to the entropy of their replay sampling distributions.

*E.g., BRO [22] uses UTD $= 10$, whereas FastTD3 [31] uses UTD $= \frac{1}{64}$ due to large-scale parallel data collection.

Experiment. To assess whether sampling entropy matters beyond recency, we compare four replay strategies on HumanoidBench using BRC [23]. Three strategies are matched for expected recency ($\mu \approx 0.85$) but differ in entropy: ERE with a 1M buffer, Uniform FIFO with a $\approx$ 300K buffer, and Truncated Geometric with a 1M buffer. Matching μ allows us to probe the effect of entropy while holding average recency fixed. As shown in Figure 4, entropy decreases substantially from Truncated Geometric to Uniform FIFO to ERE, with ERE being the most concentrated. We also include Uniform replay with a 1M buffer as a standard baseline. We detail the setup in Appendix C.

Results. As shown in Figure 4, performance tracks sampling entropy among the μ-matched recency-biased strategies: ERE, which has the lowest entropy, underperforms even the Uniform baseline; Uniform FIFO (300k) performs better; and Truncated Geometric, which has the highest entropy, achieves the best performance overall. Notably, Uniform FIFO (300k) outperforms Uniform (1M) despite using a 3 times smaller buffer, which we attribute to its higher expected recency. These results show that *how* recency is achieved matters. Strategies that obtain recency by sharply restricting sampling support sacrifice entropy of the replay sampling distribution and can underperform even when their average recency is favorable. In contrast, smooth full-support bias, as in Truncated Geometric replay, preserves update diversity while shifting replay toward recent data.

Takeaway 2 – Sampling Entropy Matters

Expected recency is informative but not sufficient to determine performance. Strategies with comparable expected recency can differ substantially depending on whether they preserve high-entropy, full-support sampling. Effective non-uniform replay should therefore bias sampling smoothly while maintaining the diversity in the replay sampling distribution.

4 Efficient Truncated Geometric Sampling Procedure

The analysis in Section 3 suggests that an effective replay sampler should increase expected recency without sacrificing sampling entropy or adding computational overhead. We instantiate this principle with a truncated geometric distribution over replay-buffer indices. This sampler smoothly biases updates toward recent transitions, admits constant-time inverse-CDF sampling, and maximizes entropy among distributions with fixed expected recency. We summarize the procedure in Algorithm 1.

Truncated Geometric Sampler. To smoothly increase expected recency while retaining constant-time sampling, we use a truncated geometric distribution over replay-buffer indices. Let $N_{\max}$ denote the maximum replay-buffer capacity and N_{current} its size at optimization step t. Transitions are indexed from oldest ($i = 0$) to newest ($i = N_{\text{current}} - 1$). We sample an index $i \in \{0, \ldots, N_{\text{current}} - 1\}$ according to

$$p(i) \;\propto\; 2^{\alpha \frac{i}{N_{\max}-1}},$$

where $\alpha > 0$ controls the strength of the recency bias. Larger values of α place more probability mass on recent transitions, while $\alpha \to 0$ recovers uniform replay. This sampler directly addresses the factors identified in Section 3. Firstly, it admits closed-form inverse-CDF sampling and therefore constant-time implementation. Secondly, in contrast to approaches like ERE, it increases expected recency smoothly, without restricting support to a recent window. Finally, among all distributions over buffer indices with fixed expected recency, the truncated geometric distribution maximizes sampling entropy [15, 6]. Thus, for a given expected recency, it preserves the highest update diversity. We visualize the tradeoff between expected recency and sampling entropy in the right panel of Figure 4.

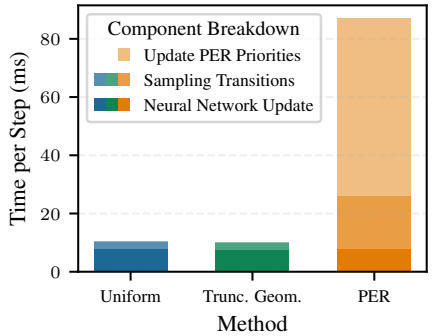

Figure 5: ***Decomposition of total latency.*** We compare the computational overhead of Uniform, Truncated Geometric, and PER. While the network update time is constant across methods, PER introduces significant latency due to priority tree management. In contrast, the truncated geometric sampling maintains a profile nearly identical to uniform due to efficient probability calculation.

Efficient Sampling. Sampling is performed using an efficient inverse transform. A single uniform random variable $U \sim \text{Unif}(0, 1)$ is mapped through the closed-form inverse cumulative distribution

Algorithm 1 Efficient Truncated Geometric Sampling

Require: Replay buffer $\mathcal{D}$, capacity $N_{\max}$, current size $N_{\text{current}} \leq N_{\max}$, recency parameter $\alpha > 0$.

Index transitions in $\mathcal{D}$ as $i = 0, \ldots, N_{\text{current}} - 1$, where $i = 0$ is the oldest and $i = N_{\text{current}} - 1$ is the most recent.
Set $k \leftarrow \alpha/(N_{\max} - 1)$.
Sample $U \sim \text{Unif}(0, 1)$.
Compute normalization constant $Z \leftarrow \frac{2^{kN_{\text{current}}}-1}{2^k-1}$.
Compute $r \leftarrow U \cdot Z$.
Compute index via inverse CDF: $i \leftarrow \lfloor \frac{\log_2(1+r(2^k-1))}{k} - 1 \rfloor$.
Clip i to $[0, N_{\text{current}} - 1]$.
return transition $\mathcal{D}[i]$.

function of the truncated geometric distribution. The sampled index is given by:

$$i = \left\lfloor \frac{\log_2\big(1 + U \cdot Z \cdot (2^k - 1)\big)}{k} \right\rfloor - 1$$

clipped to $[0, N_{\text{current}} - 1]$, with $k = \alpha \,/\, (N_{\max} - 1)$, and $Z = (2^{kN_{\text{current}}} - 1) \,/\, (2^k - 1)$. While non-uniform replay is sometimes assumed to introduce additional wall-clock overhead, the proposed sampling procedure requires a constant time, yielding $\mathcal{O}(1)$ complexity. This contrasts with the cost incurred by strategies whose sampling probabilities depend on updated priorities supported by segment trees [28], which yield $\mathcal{O}(\log N)$ complexity. As shown in Figure 5, the empirical wall-clock performance of the proposed truncated geometric sampling is comparable to uniform replay.

Interpretation of α. The proposed truncated geometric sampling is governed by a single hyperparameter α. As we now show, it provides intuitive control over **expected recency** in sampling. Approximately, for $\Delta \approx (N_{\max} - 1)/\alpha$,

$$\frac{p(i + \Delta)}{p(i)} \approx 2.$$

Thus, moving forward by roughly $(N_{\max} - 1)/\alpha$ positions doubles the sampling probability. Smaller α values approach uniform replay, while larger values increasingly emphasize recent transitions.

5 Experiments

The analysis in Section 3 predicts that recency-biased sampling should be most effective when replay volume is limited, provided the sampler preserves high entropy. We test this prediction across three representative off-policy protocols: large-scale parallel simulation, multi-task learning, and single-task learning. Parallel simulation and multi-task learning are the primary low-replay volume regimes, since data collection is distributed across many environments or tasks.

For each regime, we start from the original algorithm and training setup of the corresponding source paper, modifying only the replay sampler. We compare the original uniform replay baseline against PER [29], ERE [37], and the proposed Truncated Geometric sampler. All base algorithm hyperparameters and training protocols are kept unchanged. PER and ERE hyperparameters are tuned following Appendix C.2, while Truncated Geometric uses a fixed $\alpha = 10$ across all experiments. Unless otherwise specified, all methods use a replay buffer of capacity 1M. We report 95% stratified bootstrap CIs computed with RLiable [1]. Performance is evaluated using normalized episodic returns and success rates where applicable, across multiple seeds. Full experimental details, including hyperparameter settings, detailed figure descriptions, computational resources, environment lists, normalization procedures, and evaluation protocols are provided in Appendices C and D.

Large-Scale Parallel Simulation. We first evaluate replay sampling in a large-scale parallel simulation setting using FastTD3 [31]. We consider 29 high-dimensional continuous-control tasks from HumanoidBench [32] and 6 tasks from Isaac Lab [20]; full task lists are provided in Appendix D.2. In this setting, experience is collected rapidly from many environments in parallel, while relatively few gradient updates are performed per environment step, resulting in limited replay volume. On

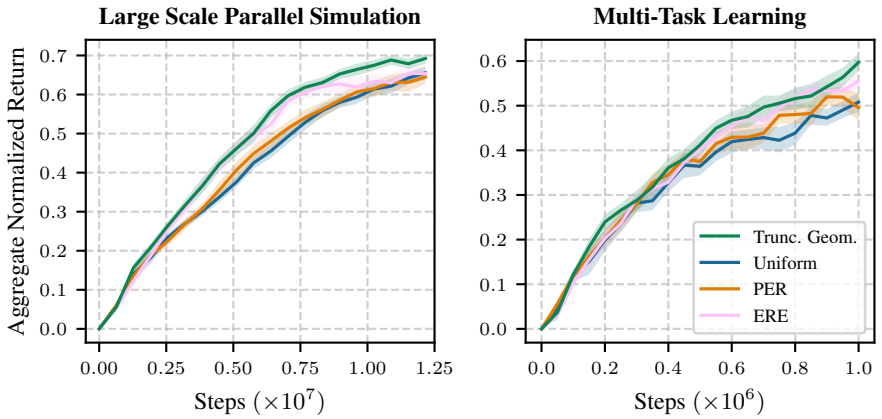

Figure 6: ***High-dimensional humanoid locomotion and manipulation tasks.*** We report aggregate mean return across 29 HumanoidBench tasks trained in a parallel simulation setup with FastTD3 (*left*) and 20 HumanoidBench tasks trained in a multi-task setup with BRC (*right*). Shaded regions show 95% CIs. Truncated Geometric sampling significantly improves over uniform replay on both benchmarks and outperforms PER and ERE, despite FastTD3 and BRC already being strong optimized baselines. These gains come with negligible computational and complexity overhead relative to uniform sampling. We detail the experimental setup in Section 5 and Appendix C.

HumanoidBench, shown in the left panel of Figure 6, Truncated Geometric sampling improves sample efficiency over uniform replay, achieving up to $1.35\times$ faster learning in terms of environment interactions. PER performs comparably to uniform replay, while ERE provides intermediate gains between uniform replay and Truncated Geometric. Because Truncated Geometric sampling incurs small computational overhead relative to uniform, these sample-efficiency gains translate directly into faster wall-clock convergence, reaching comparable performance in approximately 80% of the training time required by standard FastTD3 with uniform replay. We observe similar trends on Isaac Lab, where recency-biased samplers also outperform uniform replay. In this benchmark, Truncated Geometric performs comparably to ERE in aggregate, with ERE benefiting from particularly strong performance on one task. We report the full Isaac Lab results in Appendix F.8.

Multi-Task Learning. We next evaluate replay sampling in the multi-task setting using BRC [23] on the HumanoidBench-Hard tasks listed in Appendix F.2. As in large-scale parallel simulation, multi-task learning naturally limits replay volume, since experience is distributed across many tasks while optimization per task remains limited. As shown in the right panel of Figure 6, Truncated Geometric sampling achieves the best overall performance and improves substantially over uniform replay. ERE is competitive in this setting, though slightly below Truncated Geometric, in contrast to the low-entropy ERE configurations studied in Section 3.2, where aggressive concentration hurts performance. PER provides modest gains over uniform replay but incurs substantial computational overhead. These results further support the analysis in Section 3: when replay volume is limited relative to data collection, increasing expected recency through smooth, high-entropy recency bias improves performance while preserving computational efficiency. We also evaluate BRC with different replay strategies on DMC Dog [36] and Meta-World [39]; full results are provided in Appendix F.

Single-Task Learning. In Section 3.1, we evaluated SimbaV2 on 13 HumanoidBench Nohands tasks across UTD values from 2 to $1/16$ and batch sizes from 256 to 16, showing that Truncated Geometric sampling yields larger gains as replay volume becomes limited. Here, we return to the original SimbaV2 benchmark of 20 tasks under its standard high-UTD, standard-batch-size setting, where replay volume is substantial [18]. This experiment tests whether Truncated Geometric sampling remains safe in the high-replay volume regime, where our analysis predicts limited gains from increasing expected recency. As shown in Figure 14, Truncated Geometric sampling performs on par with uniform replay, yielding a modest 1.2% improvement. These results support the prediction from Section 3: recency-biased replay improves performance when replay volume is limited, while remaining effectively neutral when replay volume is high.

Ablation Studies. We ablate key design choices of Truncated Geometric sampling in Figure 7 using BRC on a variety of multi-task benchmarks. First, we evaluate sensitivity to the recency parameter α described in Section 4. Truncated Geometric sampling improves over uniform replay across a wide range of values, with gains between 7.5% and 12.9%, indicating that the Truncated Geometric sampling is robust to its only hyperparameter. Second, we ask whether the gains from recency-biased

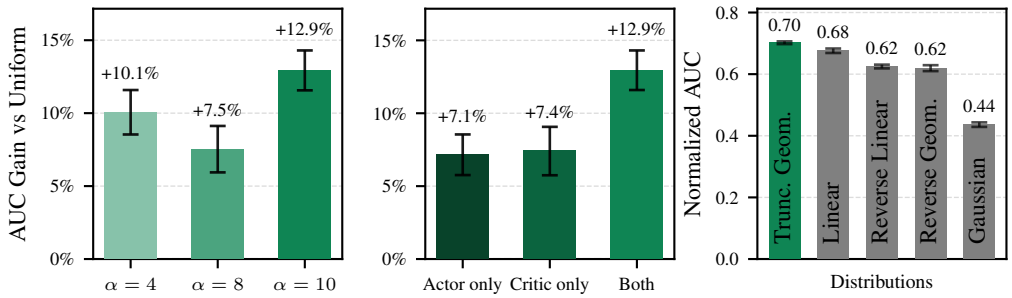

Figure 7: ***Ablations on Truncated Geometric replay.*** (*left*) Sample efficiency gains for Truncated Geometric replay with different recency parameter α, using 20 Humanoidbench tasks. Truncated Geometric consistently improves over uniform replay across all tested values of α, demonstrating robustness to hyperparameter selection. (*middle*) Sample efficiency gains when Truncated Geometric replay is applied separately to actor or critic, using 20 HumanoidBench tasks. These results show that Truncated Geometric sampling benefits actor and critic roughly equally. (*right*) Sample efficiency gains for different sampling distributions investigated on the 4 DMC dog tasks, with Truncated Geometric as the top-performing method.

replay arise primarily through critic learning or through policy updates. This distinction is important because prior work has often found critic learning to be especially sensitive to algorithmic design choices in off-policy RL [7, 22]. We therefore apply Truncated Geometric sampling separately to the actor and critic updates, allowing for updates on independent batches of data. Shapley value analysis attributes 6.6% of the improvement to the critic and 6.3% to the actor, suggesting that both components benefit comparably from recency-biased replay. Finally, we compare against alternative non-uniform sampling distributions. Among the evaluated strategies that we describe in Appendix C.4, smooth recency-biased distributions achieve the strongest performance, supporting the role of high-entropy recency bias. We describe the experimental settings in detail in Appendices C.2.

Summary of the Experimental Results

Recency-biased replay provides the largest gains in low-replay volume regimes, such as large-scale parallel simulation and multi-task learning, while its benefits diminish when replay volume is high, as in high-UTD single-task learning. The results also support the role of sampling entropy: smooth, full-support recency bias outperforms strategies such as ERE that increase expected recency by sharply restricting sampling support. The same analysis explains Truncated Geometric sampling: it improves sample efficiency in low-replay volume regimes, incurs negligible overhead, and preserves performance when replay volume is high. Together, these findings support the analysis in Section 3 and clarify when recency-biased replay improves over uniform replay.

6 Conclusions

We revisited non-uniform replay sampling in off-policy reinforcement learning and showed that its effectiveness cannot be understood from expected recency alone. Instead, replay performance is mediated by two additional factors: replay volume and sampling entropy. Recency-biased replay is most beneficial in low-replay volume regimes, where transitions are replayed infrequently, and when the sampling distribution preserves high entropy rather than sharply restricting support. Practically, these findings suggest that replay design should be treated as a function of the learner–collector ratio: uniform replay is often sufficient when optimization is abundant, while smooth recency-biased replay is a simple and effective default when data collection outpaces learning.

Guided by this analysis, we introduced a simple Truncated Geometric replay sampler that smoothly increases expected recency while maintaining high entropy and constant-time sampling. Across large-scale parallel simulation, multi-task learning, and single-task settings, this approach improves sample efficiency in low-replay volume regimes, incurs negligible computational overhead with $\mathcal{O}(1)$ sampling complexity, and remains neutral when replay volume is high. Our results clarify when and why recency-biased replay improves learning, and provide a practical alternative to uniform and priority-based replay for modern off-policy RL. We discuss limitations of our analysis in Appendix A.

Acknowledgments

We gratefully acknowledge the Polish high-performance computing infrastructure, PLGrid (HPC Center: ACK Cyfronet AGH), for providing computational resources and support under grant no. PLG/2025/018597. Mikołaj Czarnecki was supported by the National Science Centre under Grant No. 2023/50/E/ST6/00665. Yarden As received funding from a grant of the Hasler foundation (grant no. 21039) and the ETH AI Center. Pieter Abbeel holds concurrent appointments as a Professor at UC Berkeley and as an Amazon Scholar. This paper describes work performed at UC Berkeley and is not associated with Amazon.

References

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

[2] Yarden As, Dhruva Tirumala, René Zurbrügg, Chenhao Li, Stelian Coros, Andreas Krause, and Markus Wulfmeier. What matters for simulation to online reinforcement learning on real robots. *arXiv preprint arXiv:2602.20220*, 2026.

[3] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, TB Dhruva, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations*, 2018.

[4] Aditya Bhatt, Daniel Palenicek, Boris Belousov, Max Argus, Artemij Amiranashvili, Thomas Brox, and Jan Peters. CrossQ: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity. In *International conference on learning representations (ICLR)*, 2024.

[5] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.

[6] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[7] 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 *The Eleventh International Conference on Learning Representations*, 2022.

[8] Preston Fu, Oleh Rybkin, Zhiyuan Zhou, Michal Nauman, Pieter Abbeel, Sergey Levine, and Aviral Kumar. Compute-optimal scaling for value-based deep rl, 2025. URL https://arxiv.org/abs/2508.14881.

[9] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 2018.

[10] Scott Fujimoto, Pierluca D'Oro, Amy Zhang, Yuandong Tian, and Michael Rabbat. Towards general-purpose model-free reinforcement learning. *arXiv preprint arXiv:2501.16142*, 2025.

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 2018.

[12] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

[13] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[14] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[15] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[16] Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. *arXiv preprint arXiv:2410.09754*, 2024.

[17] Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyperspherical normalization for scalable deep reinforcement learning. *arXiv preprint arXiv:2502.15280*, 2025.

[18] Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyperspherical normalization for scalable deep reinforcement learning. *arXiv preprint arXiv:2502.15280*, 2025.

[19] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2015.

[20] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

[21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015. URL http://dx.doi.org/10.1038/nature14236.

[22] Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Miłoś, and Marek Cygan. Bigger, regularized, optimistic: scaling for compute and sample efficient continuous control. *Advances in neural information processing systems*, 2024.

[23] Michal Nauman, Marek Cygan, Carmelo Sferrazza, Aviral Kumar, and Pieter Abbeel. Bigger, regularized, categorical: High-capacity value functions are efficient multi-task learners. *arXiv preprint arXiv:2505.23150*, 2025.

[24] Daniel Palenicek, Florian Vogt, Joe Watson, Ingmar Posner, and Jan Peters. Xqc: Well-conditioned optimization accelerates deep reinforcement learning. *arXiv preprint arXiv:2509.25174*, 2025.

[25] Leonard S. Pleiss, Tobias Sutter, and Maximilian Schiffer. Reliability-adjusted prioritized experience replay, 2025. URL https://arxiv.org/abs/2506.18482.

[26] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[27] Oleh Rybkin, Michal Nauman, Preston Fu, Charlie Snell, Pieter Abbeel, Sergey Levine, and Aviral Kumar. Value-based deep rl scales predictably. *arXiv preprint arXiv:2502.04327*, 2025.

[28] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2015.

[29] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2016. URL https://arxiv.org/abs/1511.05952.

[30] Younggyo Seo, Carmelo Sferrazza, Juyue Chen, Guanya Shi, Rocky Duan, and Pieter Abbeel. Learning sim-to-real humanoid locomotion in 15 minutes. *arXiv preprint arXiv:2512.01996*, 2025.

[31] Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. *arXiv preprint arXiv:2505.22642*, 2025.

[32] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.

[33] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[34] Laura Smith, Ilya Kostrikov, and Sergey Levine. A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. *arXiv preprint arXiv:2208.07860*, 2022.

[35] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[36] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[37] Che Wang and Keith Ross. Boosting soft actor-critic: Emphasizing recent experience without forgetting the past, 2019. URL https://arxiv.org/abs/1906.04009.

[38] Che Wang, Yanqiu Wu, Quan Vuong, and Keith Ross. Striving for simplicity and performance in off-policy drl: Output normalization and non-uniform sampling. In *International Conference on Machine Learning*, pp. 10070–10080. PMLR, 2020.

[39] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *CoRR*, abs/1910.10897, 2019. URL http://arxiv.org/abs/1910.10897.

A Limitations

Our experiments focus on continuous-control off-policy RL, where replay buffers are central and low-replay volume regimes arise naturally in large-scale parallel and multi-task training. Extending the proposed per-step replay analysis to discrete-action domains such as Atari, and to value-based agents with different exploration and target-network dynamics, remains an important direction for future work. More broadly, our study is empirical: the decomposition into expected recency, replay volume, and sampling entropy provides an explanatory and practical framework, but not a complete theory of replay. Developing such a theory could clarify how replay sampling interacts with optimization, bootstrapping, distribution shift, and function approximation. Finally, although our ablations show that Truncated Geometric replay is robust to the recency parameter, adaptive schedules for recency bias may further improve performance.

B Truncated Geometric Sampling Procedure Details

In this section, we discuss the implementation details of the proposed Truncated Geometric Sampling Procedure.

B.1 Python Implementation

We provide a concise NumPy [13] implementation of the Truncated Geometric sampling. The implementation uses inverse-CDF sampling, making it $\mathcal{O}(1)$ and thus as computationally efficient as uniform sampling.

```
import numpy as np

def geometric_sample(size, batch_size, alpha=10.0, capacity=1_000_000, offset=0):
    k = alpha / (capacity - 1)
    base = 2.0 ** k
    Z = (base ** size - 1.0) / (base - 1.0)
    r = np.random.uniform(0.0, Z, size=batch_size)
    i = np.log2(1.0 + r * (base - 1.0)) / k - 1.0
    i = np.clip(np.floor(i).astype(np.int64), 0, size - 1)
    if offset != 0:
        i = (i + offset) % capacity
    return i
```

B.2 Parallel Buffer Pseudocode

Algorithm 2 extends the single-task Truncated Geometric sampling procedure to the multi-task setting, where each task maintains an independent replay buffer. At each sampling step, a task is selected uniformly at random, and a transition is sampled from the corresponding buffer using the Truncated Geometric distribution, preserving recency bias across all tasks without introducing additional computational overhead.

C Experimental Details

C.1 Figures

Figure 1. We report the results in ***High Replay Volume, High Expected Recency*** setting on HumanoidBench [32], with FastTD3 [31] evaluated across all 29 Hands tasks and BRC [23] across all 20 available tasks. Both algorithms operate in the low replay volume regime — FastTD3 [31] with an UTD ratio of $1/64$ and BRC [23] with $1/10$ — where recency bias is expected to yield the greatest benefit (Section 3.1). Wall-clock time is measured as the total running time of the algorithm and is aggregated across both settings, covering 49 tasks in total.

Figure 2. We show expected amount of replays we sample from each index in between each environment step. Index recency is normalized index in replay buffer with 0 corresponding to the oldest sample and 1 corresponding to the newest sample in the buffer. By default all figures correspond to expected recency of training with UTD 2 and Batch Size 128. On left and middle plot mean of

Algorithm 2 Efficient Truncated Geometric Sampling Procedure

Require: Per-task replay buffer $\mathcal{D}$, max capacity $N_{\max}$, current size $N_{\text{current}} \leq N_{\max}$, number of tasks M, recency parameter $\alpha > 0$.

Sample task $j \sim \text{Unif}(\{0, \ldots, M-1\})$.
Index transitions in $\mathcal{D}[j]$ as $i = 0, \ldots, N_{\text{current}} - 1$, where $i = 0$ is the oldest and $i = N_{\text{current}} - 1$ is the most recent.
Set $k \leftarrow \alpha/(N_{\max} - 1)$.
Sample $U \sim \text{Unif}(0, 1)$.
Compute normalization constant $Z \leftarrow \frac{2^{k N_{\text{current}}} - 1}{2^k - 1}$.
Compute $r \leftarrow U \cdot Z$.
Compute index via inverse CDF: $i \leftarrow \lfloor \frac{\log_2(1 + r(2^k - 1))}{k} - 1 \rfloor$.
Clip i to $[0, N_{\text{current}} - 1]$.
return transition $\mathcal{D}[j][i]$.

Uniform sampling distribution, μ_{U} has index recency $\rho_t(i) = 0.5$ and mean of Truncated Geometric distribution μ_{TG} has index recency $\rho_t(i) = 0.857$. On middle plot higher Uniform expected recency corresponds to training with UTD 16 and Batch Size 128. On right plot we provide numeric values of Entropy for Truncated Geometric distribution of size $1,000,000$ and Uniform FIFO distribution of size $288,000$. They both have the same mean $\mu_{\text{TG}} = \mu_{\text{UF}} = 0.857$.

Figure 3. We report the results in varying ***Replay Volume*** comparing ***High Expected Recency*** Truncated Geometric and ERE replays and ***Low Expected Recency*** Uniform and PER replays on HumanoidBench Nohands [32] using SimbaV2 [17] with 95% stratified bootstrap CI computed via RLiable [1]. All experiments were run on $1,000,000$ steps with hyperparameters described in Table 1 in column 'SimbaV2'.

Figure 4. Results are reported in ***Low Replay Volume*** comparing ***High Expected Recency*** Truncated Geometric and ERE replays and ***Low Expected Recency*** Uniform and PER replays on HumanoidBench [32] using BRC [23], with 95% stratified bootstrap CI computed via RLiable [1]. The entropy-recency frontier shown in the right panel is computed for a buffer of capacity 1M. Visualizations of the corresponding sampling distributions are provided in Figure 8. We discuss Uniform FIFO as a recency-biased strategy in Appendix E.2.

Figure 5. Measurements are conducted in ***Low Replay Volume*** comparing ***High Expected Recency*** Truncated Geometric replay with **Low** ***Expected Recency*** Uniform and PER replay on HumanoidBench using BRC across 20 tasks. PER maintains a separate segment tree per task, implemented in NumPy [13], to avoid Q-value scale mismatch across tasks. Latency is measured using `time.time` in Python; each operation is timed over 1000 consecutive steps and averaged to obtain a reliable estimate. All measurements are taken approximately at the midpoint of training.

Figure 6. Results on both plots are reported in ***Low Replay Volume*** setting on HumanoidBench [32] comparing ***High Expected Recency*** Truncated Geometric and ERE replays with ***Low Expected Recency*** Uniform and PER replays. Large Scale Parallel Simulation uses FastTD3 [31] algorithm. Multi-Task Learning uses BRC [23]. Error bars are 95% stratified bootstrap CI computed via RLiable [1]. On Large Scale Parallel Simulation we report total number of environment steps on x-axis. We tuned Episode Length and Decay Constant η in ERE buffer to achieve better results.

Figure 7. Results on α and actor/critic ablations are reported in ***Low Replay Volume*** setting on HumanoidBench [32], results on distribution ablations are reported on DMC Dogs [36]. All ablations use BRC [23] algorithm with 95% stratified bootstrap CI computed via RLiable [1]. Details of sampling distributions from right ablation are discussed in Section C.4.

C.2 Hyperparameters

Table 1 reports the full set of hyperparameters used across all experiments. For each algorithm, we distinguish between default hyperparameters (shown in **bold**), which are taken directly from

the original implementations, and tuned hyperparameters (shown in *italic*), which were adjusted to accommodate the different replay sampling strategies.

Hyperparameter	FastTD3	BRC	SimbaV2
Environment and Execution			
Environment suite	HumanoidBench/Isaac Lab	HumanoidBench/DMC	**HumanoidBench Nohands**/DMC/MYOSuite
Parallel environments	**128 (HumanoidBench)/2048 (Isaac Lab)**	**20 (HB), 3 (DMC/Humanoids), 4 (DMC Dogs)**	–
Evaluation environments	**128 (HumanoidBench)/2048 (Isaac Lab)**	–	–
Total environment steps	**100,000**	**1,000,000**	**1,000,000**
Number of seeds	**5**	**3–5**	**5**
TD-n	**1 (HumanoidBench)/1–8 (Isaac Lab)**	**1**	**1**
Optimization			
Update-To-Data ratio	**1/64 (HumanoidBench)/1/128 (Isaac Lab)**	**2**	**2**, 1, 1/2, 1/4, 1/8, 1/16
Actor learning rate	3×10^{-4}	3×10^{-4}	1×10^{-4}
Critic learning rate	3×10^{-4}	3×10^{-4}	1×10^{-4}
Learning rate schedule	**Cosine annealing** ($3 \times 10^{-4} \rightarrow 3 \times 10^{-5}$)	**Constant**	**Linear annealing** ($1 \times 10^{-4} \rightarrow 5 \times 10^{-5}$)
Batch size	**32768**	**1024**	**256**
Discount factor γ	**0.99**	**0.99**	**0.95 (HumanoidBench/DMC), 0.995 (MYOSuite)**
Target update coefficient τ	**0.1**	**0.005**	**0.005**
Policy update frequency	**2**	**2**	**2**
Learning starts	**10 steps**	**5000 steps**	**5000**
Weight decay	**0.1**	–	–
Replay Buffer			
Replay buffer capacity	**50,000 (HumanoidBench) / $5 \times 1024 - 10 \times 1024$ (Isaac Lab)**	100,000, **1,000,000**	**1,000,000**
Sampling strategies	**Uniform** / PER / ERE / TG	**Uniform** / PER / ERE / TG	**Uniform** / PER / ERE / TG
Method-Specific Hyperparameters			
Truncated Geometric α	**10.0**	***10.0***	**10.0**
PER exponent α	**0.6**	*0.4*	**0.6**
PER importance-sampling β	**0.4 → 1.0**	**0.4 → 1.0**	**0.4 → 1.0**
ERE decay constant η	**0.996**	***0.996* (HB, DMC Dogs), *0.994* (DMC Humanoids)**	0.994
ERE episode length	*100*	**1000**	**1000**
ERE mini-batch updates	*200*	**2000**	**2000**
ERE minimum buffer size $c_{\min}$	**10** × batch size	10 × batch size	5000
Architecture			
Actor hidden dimension	**512**	**256**	**128**
Critic hidden dimension	**1024**	**4096**	**512**
Critic blocks	**2**	**2**	**2**
Actor blocks	**1**	**1**	**1**
Distributional atoms	**101**	**101**	**101**
Value support $(v_{\min}, v_{\max})$	**$(-250, 250)$ (HumanoidBench) / $(-50, 50)$, $(-10, 10)$**	$(-10, 10)$	$(-5, 5)$
Task embedding dimension	–	**32**	–
Stability and Systems			
Observation normalization	**Enabled**	**Disabled**	**Disabled**
Reward normalization	**Disabled**	**Enabled**	**Enabled**
Clipped Double Q	**Enabled**	**Disabled**	**Enabled**
AMP	**Enabled (bf16)**	–	–
`torch.compile`	**Enabled** (Excluding ERE)	–	–
Gradient clipping	**Disabled**	**Disabled**	**Disabled**

Table 1: Algorithm-specific training and replay hyperparameters. FastTD3 values are taken directly from the default configuration (`BaseArgs`). BRC column are taken from BRC flag defaults. SimbaV2 is default configuration (`configs`). In Table 1 by **bold** we denote default hyperparameters, by *italic* we denote tuned hyperparameters.

C.3 Sampling Distributions

We visualize example sampling distributions for all used replay strategies in Figure 8. The Truncated Geometric, ERE, and Uniform FIFO distributions are matched to the same expected recency μ and sampling entropy $\mathcal{H}(p_t)$ as used in Section 3.2.

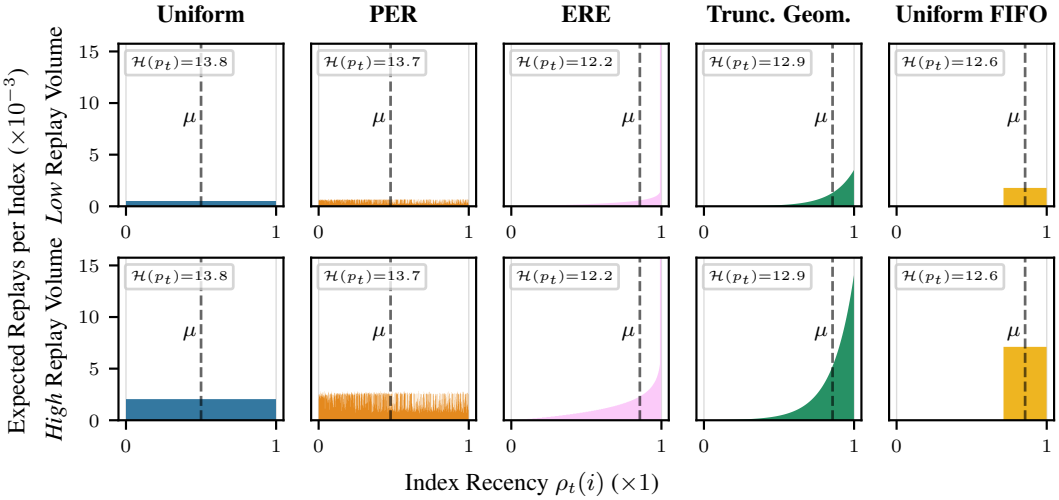

Figure 8: ***Sampling distributions and expected replay counts across methods.*** Each column shows the sampling distribution of a replay strategy visualized as expected number of replays per buffer index, in the *low* (top) and *high* (bottom) replay volume regimes. The dashed vertical line denotes the expected recency μ for each strategy, and $\mathcal{H}(p_t)$ reports the sampling entropy. ERE, Truncated Geometric, and Uniform FIFO are matched to the same expected index μ, yet differ noticeably in their sampling entropy. Truncated Geometric achieves the highest entropy at this μ. We discuss the interpretation of entropy differences in Appendix E.3.

C.4 Details of Distributions From Ablations

In distribution ablation from Figure 7 we analyze how BRC [23] with different distributions from which we sample replay, the distributions are as follow:

- ***High Expected Recency*** Linear.
- ***Low Expected Recency*** Reverse Linear.
- ***Low Expected Recency*** Gaussian.
- ***Low Expected Recency*** Reverse Truncated Geometric.

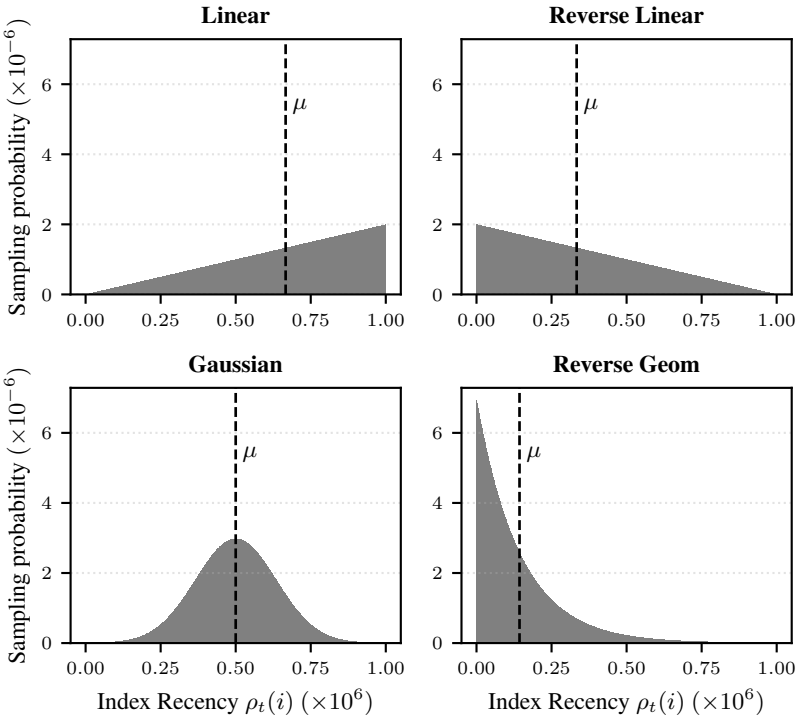

Figure 9: **Sampling distributions from Ablation 7.** All four distributions are normalized so that the highest-probability transition is 2^{10} times more likely to be sampled than the lowest-probability transition.

C.5 Computational Resources

All experiments were conducted on NVIDIA A100 GPUs. Single-task and large-scale parallel runs required approximately 3 hours to complete, while multi-task runs required approximately 40 hours.

C.6 Reproducibility

All environment lists, hyperparameter configurations, training scripts, and evaluation code are provided in the supplementary material and will be released publicly upon acceptance.

D Evaluation Protocols

This section provides full details of the experimental setup used throughout the paper, including replay buffer configurations, algorithm-specific hyperparameters, normalization procedures, and evaluation protocols. All experiments are conducted using the original implementations of the corresponding algorithms, with minimal modifications required to integrate alternative replay sampling strategies.

D.1 Policy Evaluation

All policies are evaluated periodically during training without exploration noise. Each evaluation consists of multiple episodes per task, and results are averaged. All reported curves show the mean performance over random seeds, with 95% confidence intervals computed using nonparametric bootstrapping via RLiable [1]. Unless otherwise specified, all experiments are run with three to five random seeds (see Table 1).

D.2 Benchmarks

In this section we describe benchmarks considered in our studies:

D.2.1 Multi-task benchmarks

All tasks are available in BRC [23] repository `https://github.com/naumix/BiggerRegularizedCategorical`.

- **BRC HumanoidBench Hands** - we use benchmark of all 20 tasks from [32],

  Tasks: *walk, stand, run, stair, crawl, pole, slide, hurdle, maze, sit_simple, sit_hard, balance_simple, balance_hard, reach, spoon, window, insert_small, insert_normal, bookshelf_simple, bookshelf_hard*

- **BRC MetaWorld** - we use benchmark of all 50 tasks from [39].

  Tasks: *assembly-v2-goal-observable, basketball-v2-goal-observable, bin-picking-v2-goal-observable, box-close-v2-goal-observable, button-press-topdown-v2-goal-observable, button-press-topdown-wall-v2-goal-observable, button-press-v2-goal-observable, button-press-wall-v2-goal-observable, coffee-button-v2-goal-observable, coffee-pull-v2-goal-observable, coffee-push-v2-goal-observable, dial-turn-v2-goal-observable, disassemble-v2-goal-observable, door-close-v2-goal-observable, door-lock-v2-goal-observable, door-open-v2-goal-observable, door-unlock-v2-goal-observable, hand-insert-v2-goal-observable, drawer-close-v2-goal-observable, drawer-open-v2-goal-observable, faucet-open-v2-goal-observable, faucet-close-v2-goal-observable, hammer-v2-goal-observable, handle-press-side-v2-goal-observable, handle-press-v2-goal-observable, handle-pull-side-v2-goal-observable, handle-pull-v2-goal-observable, lever-pull-v2-goal-observable, pick-place-wall-v2-goal-observable, pick-out-of-hole-v2-goal-observable, pick-place-v2-goal-observable, plate-slide-v2-goal-observable, plate-slide-side-v2-goal-observable, plate-slide-back-v2-goal-observable, plate-slide-back-side-v2-goal-observable, peg-insert-side-v2-goal-observable, peg-unplug-side-v2-goal-observable, soccer-v2-goal-observable, stick-push-v2-goal-observable, stick-pull-v2-goal-observable, push-v2-goal-observable, push-wall-v2-goal-observable, push-back-v2-goal-observable, reach-v2-goal-observable, reach-wall-v2-goal-observable, shelf-place-v2-goal-observable, sweep-into-v2-goal-observable, sweep-v2-goal-observable, window-open-v2-goal-observable, window-close-v2-goal-observable*

- **BRC DMC Dogs** - we use benchmark of all 4 DMC Dogs tasks [36].

  Tasks: *dog-stand, dog-walk, dog-trot, dog-run*

- **BRC DMC Humanoids** - we use benchmark of all 3 DMC Humanoid tasks [36].

  Tasks: *humanoid-stand, humanoid-walk, humanoid-run*

D.2.2 Large Scale Parallel Learning benchmarks

All tasks are available in FastTD3 [31] repository `https://github.com/younggyoseo/FastTD3`

- **FastTD3 HumanoidBench Hands** - we use benchmark of 29 HumanoidBench Hands tasks from [32]

  Tasks: *balance_ hard, balance_ simple, basketball, bookshelf_ hard, bookshelf_ simple, cabinet, crawl, cube, door, hurdle, insert_ normal, insert_ small, maze,*

package, pole, powerlift, push, reach, room, run, sit_ hard, sit_ simple, slide, spoon, stair, stand, truck, walk, window

- **FastTD3 Isaac Lab** - we use benchmark of 6 Isaac Lab [20] tasks

  Tasks: *Velocity-Flat-G1, Velocity-Flat-H1, Velocity-Rough-G1, Velocity-Rough-H1, Repose-Cube-Allegro-Direct, Repose-Cube-Shadow-Direct*

D.2.3 Single-task benchmarks

All tasks are available for use in SimbaV2 [17] repository https://github.com/DAVIAN-Robotics/SimbaV2

- **SimbaV2 HumanoidBench No Hands** - we use benchmark of 13 HumanoidBench No Hands tasks from [32].

  Tasks: *walk, stand, run, hurdle, crawl, maze, sit_simple, sit_hard, balance_simple, balance_hard, stair, slide, pole*

- **SimbaV2 MyoSuite** - we use benchmark of 10 MyoSuite tasks from [5].

  Tasks: *reach,reach-hard,pose,pose-hard,obj-hold,obj-hold-hard,key-turn,key-turn-hard,pen-twirl,pen-twirl-hard*

- **SimbaV2 DMC** - we use 7 tasks from combined DMC Dogs and DMC Humanoids from [36].

  Tasks: *humanoid-stand, humanoid-walk, humanoid-run, dog-stand, dog-walk, dog-trot, dog-run*

D.3 Normalization and Metrics

Performance is evaluated using standard task-specific metrics, including normalized episodic returns and success rates where applicable. Normalized returns are computed as

$$\text{Score} = \frac{R - R_{\text{random}}}{R_{\text{expert}} - R_{\text{random}}},$$

where R_{random} and R_{expert} are task-specific constants provided by the benchmark.

Table 2 reports normalization constants and success thresholds for all tasks sourced from Sferrazza et al. [32].

Task	Random Score R_{random}	Success Threshold R_{expert}
h1-balance_hard-v0	9.044	800.0
h1-balance_simple-v0	9.391	800.0
h1-crawl-v0	272.658	700.0
h1-hurdle-v0	2.214	700.0
h1-maze-v0	106.441	1200.0
h1-pole-v0	20.090	700.0
h1-reach-v0	260.302	12000.0
h1-run-v0	2.020	700.0
h1-sit_hard-v0	2.448	750.0
h1-sit_simple-v0	9.393	750.0
h1-slide-v0	3.191	700.0
h1-stair-v0	3.112	700.0
h1-stand-v0	10.545	800.0
h1-walk-v0	2.377	700.0
h1hand-balance_hard-v0	10.032	800.0
h1hand-balance_simple-v0	10.170	800.0
h1hand-basketball-v0	8.979	1200.0
h1hand-bookshelf_hard-v0	14.848	2000.0
h1hand-bookshelf_simple-v0	16.777	2000.0
h1hand-cabinet-v0	37.733	2500.0
h1hand-crawl-v0	278.868	800.0
h1hand-cube-v0	4.787	370.0
h1hand-door-v0	2.771	600.0
h1hand-hurdle-v0	2.371	700.0
h1hand-insert_normal-v0	1.673	350.0
h1hand-insert_small-v0	1.653	350.0
h1hand-kitchen-v0	0.000	4.0
h1hand-maze-v0	106.233	1200.0
h1hand-package-v0	-10040.932	1500.0
h1hand-pole-v0	19.721	700.0
h1hand-powerlift-v0	17.638	800.0
h1hand-push-v0	-526.800	700.0
h1hand-reach-v0	-50.024	12000.0
h1hand-room-v0	3.018	400.0
h1hand-run-v0	1.927	700.0
h1hand-sit_hard-v0	2.477	750.0
h1hand-sit_simple-v0	10.768	750.0
h1hand-slide-v0	3.142	700.0
h1hand-spoon-v0	4.661	650.0
h1hand-stair-v0	3.161	700.0
h1hand-stand-v0	11.973	800.0
h1hand-truck-v0	562.419	3000.0
h1hand-walk-v0	2.505	700.0
h1hand-window-v0	2.713	650.0

Table 2: Normalization constants (R_{random}) and success thresholds (R_{expert}) for HumanoidBench tasks.

E Discussion

E.1 Normalized AUC for Sample Efficiency

To evaluate sample efficiency, we report the *Normalized Area Under the Curve* (Normalized AUC). Rather than measuring performance only at the final training step, this metric captures learning progress throughout training and relates to the cumulative regret that the agent incurs during learning.

Formally, let R_t denote the normalized return at evaluation step t. Given equally spaced evaluation intervals $\{t_1, \ldots, t_K\}$, we compute

$$\text{Normalized AUC} = \frac{1}{K} \sum_{k=1}^{K} R_{t_k}.$$

For each run and task, we first compute the temporal mean of normalized returns across evaluation checkpoints. We then average these per-task Normalized AUC values across tasks and finally report the mean across runs. Confidence intervals are estimated using stratified bootstrap resampling over runs, following the RLiable protocol [1].

Normalized AUC provides a more faithful measure of sample efficiency than reporting only the final performance. Final-step metrics can obscure differences in convergence speed and learning stability, whereas Normalized AUC captures both how quickly and how consistently an agent improves during training.

This evaluation protocol follows prior work that emphasizes area-under-curve metrics for sample-efficient reinforcement learning [1].

E.2 Uniform FIFO as a Recency-Biased Method

A Uniform FIFO replay buffer can be interpreted as a recency-biased method. At each step, the oldest transition is discarded, which is equivalent to assigning it a sampling probability of zero. In this sense, FIFO implements recency bias through the discarding of oldest transitions rather than through non-uniform sampling probabilities.

Interestingly, in our preliminary experiments, alternative discarding strategies — such as discarding transitions based on Q-values, estimated advantages, or novelty estimated via random network distillation — did not improve over standard FIFO. This suggests that FIFO may be a surprisingly strong discarding baseline, consistent with the broader literature [37, 38] suggesting that recency — implicitly induced by FIFO through discarding — is an important factor in replay buffer design. We leave a systematic investigation of discarding strategies as an interesting direction for future work.

E.3 Entropy Scale and Buffer Size Equivalence

Since sampling entropy $\mathcal{H}(p_t)$ is a logarithmic quantity, small numerical differences correspond to substantial changes in sampling diversity. For intuition, consider a uniform distribution over a buffer of size N, which has entropy $\ln N$. A change of 1 nat in entropy is therefore equivalent to scaling the effective buffer size by a factor of $e \approx 2.71$. Consequently, differences in entropy of 0.3 nats — which may appear modest — correspond to a factor of $e^{0.3} \approx 1.35$ change in effective buffer size, i.e., a 35% increase or decrease in sampling diversity.

An alternative metric is perplexity, defined as $e^{\mathcal{H}(p_t)}$, which trades the logarithmic scale of entropy for a linear one, potentially making scalar values easier to interpret directly. However, perplexity and entropy are bijectively equivalent, and since entropy is the more standard metric in the literature, we report entropy throughout this work. We encourage readers to keep the logarithmic scale in mind when interpreting entropy differences reported in our figures.

E.4 ERE Buffer Shortcomings

On the left panel of Figure 4, the ERE buffer [37] underperforms severely. We believe this is caused by the low entropy of the distribution from which ERE samples. To make its expected recency match that of the Truncated Geometric with $\alpha = 10$, we tune the Decay Constant η by decreasing it. In the ERE buffer, replays are sampled in episodes k increasing cyclically from 1 to the episode length K; in each episode, we sample uniformly from $c_k = \min(N \cdot \eta^{1000k/K}, c_{\text{min}})$ most recent samples, where N is the length of the buffer and c_{min} is the minimal sampling size, usually set to 5000 for a buffer size of $1{,}000{,}000$. By decreasing the base of the exponent, i.e., the decay constant η, we increase the number of episodes for which we sample from the minimal window $c_k = c_{\text{min}}$; this drastically reduces sampling entropy, as visible for high expected recency μ in the right subplot of Figure 4. For the experiments in Figure 6, we tune the ERE parameters for every algorithm and benchmark

individually to achieve competitive performance comparable to other methods. In comparison, Truncated Geometric replay used a fixed hyperparameter of $\alpha = 10$ across all experiments.

F Learning Curves

We provide learning curves for the experiments presented in Section 5. In Figure 10 we present the performance of all sampling baselines when using FastTD3. In Figures 11 to 13 we illustrate the learning curves for our experiments using BRC on HumanoidBench and DeepMind Control Suite. UTD in these experiments was 1/64 (consistent with experiments from the original paper).

F.1 FastTD3

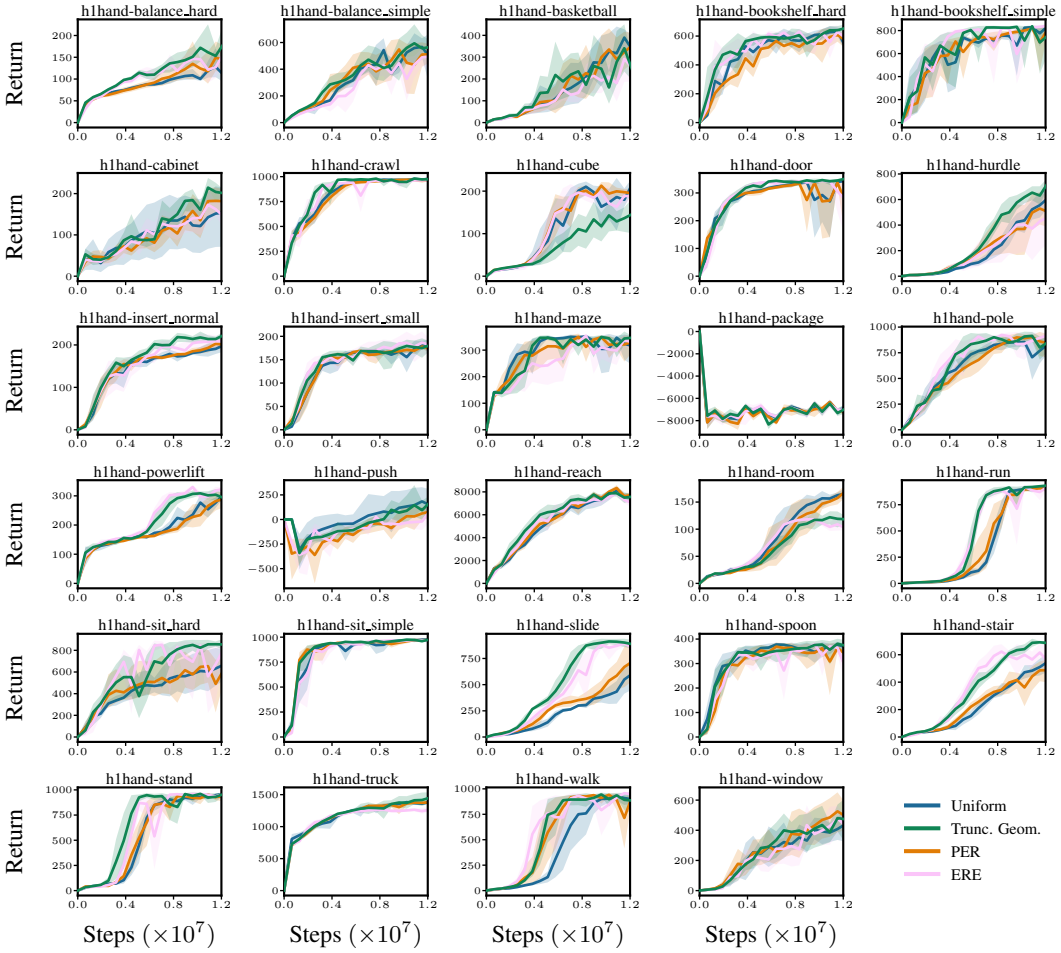

Figure 10: ***(Low Replay Volume) Performance across all 29 tasks.*** We compare FastTD3 [31] with different replay strategies: ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink). Solid lines represent the mean return over 5 seeds, and shaded regions denote the 95% bootstrap confidence intervals computed via `rliable` on HumanoidBench. The dashed grey line in each subplot indicates the task-specific success threshold. All x-axes represent environment steps ($\times 10^5$). On average, Truncated Geometric achieves a 14.1% gain over Uniform in AUC.

F.2 BRC HumanoidBench

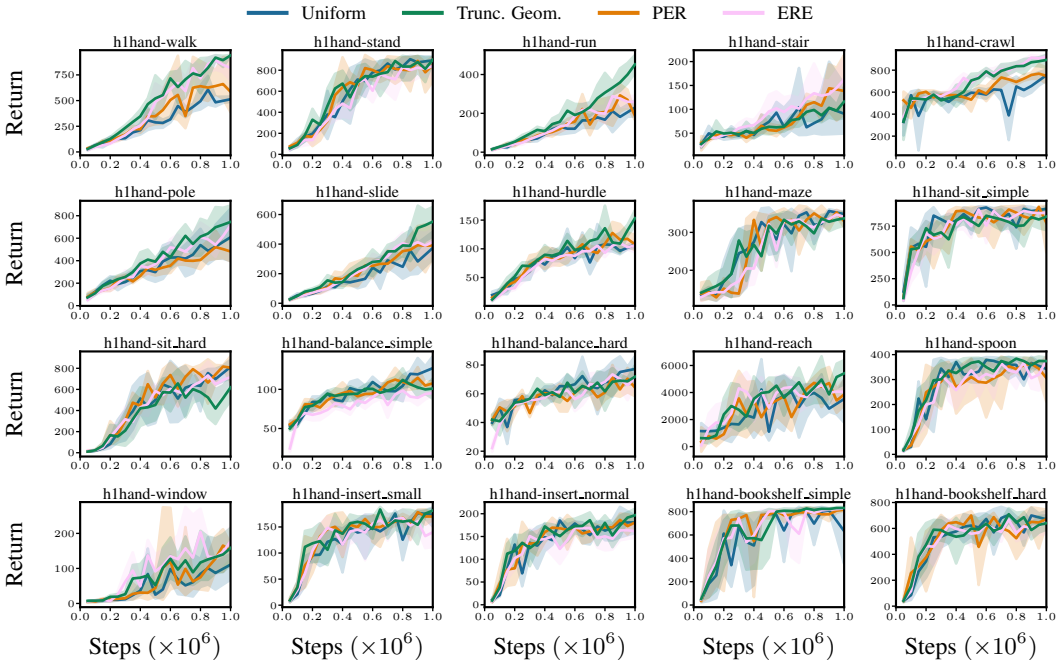

Figure 11: ***(Low Replay Volume)*** *Performance across all 20 BRC tasks.* We compare different replay strategies: ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink). Solid lines represent the mean return over 5 seeds, and shaded regions denote the 95% bootstrap confidence intervals computed via `rliable` on HumanoidBench. The dashed grey line in each subplot indicates the task-specific success threshold. All x-axes represent environment steps ($\times 10^6$). On average, Truncated Geometric achieves a 12.9% gain over Uniform in AUC and approximately 20% improvement in final performance.

F.3 BRC DeepMind Suite Control Humanoids

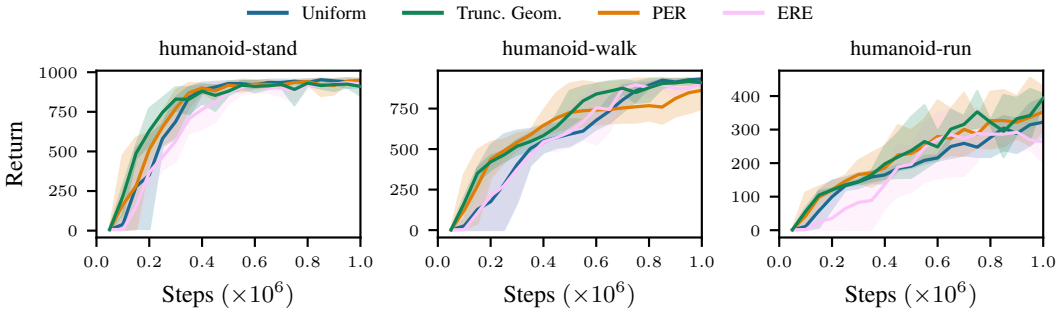

Figure 12: ***(Moderate Replay Volume) Performance on DMC Humanoids tasks. [36]*** We compare different replay strategies: ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink) across three humanoid tasks: `stand`, `walk`, and `run`. Solid lines show the mean return over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. All x-axes represent environment steps ($\times 10^6$), and y-axes show the denormalized task returns. On average, Truncated Geometric achieves a 10% gain over Uniform in AUC.

‘

F.4 BRC DeepMind Suite Control Dogs

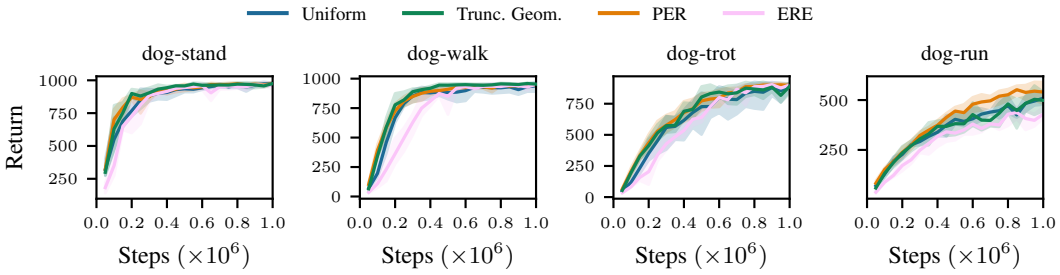

Figure 13: ***(Moderate Replay Volume) Performance on DMC Dog tasks.*** We compare different replay strategies on BRC [23]: ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink) across four dog locomotion tasks: `stand`, `walk`, `trot`, and `run`. Solid lines show the mean return over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. All x-axes represent environment steps ($\times 10^6$), and y-axes show the denormalized task returns. On average, Truncated Geometric achieves a 4.2% gain over Uniform in AUC.

F.5 SimbaV2 regular UTD learning curve

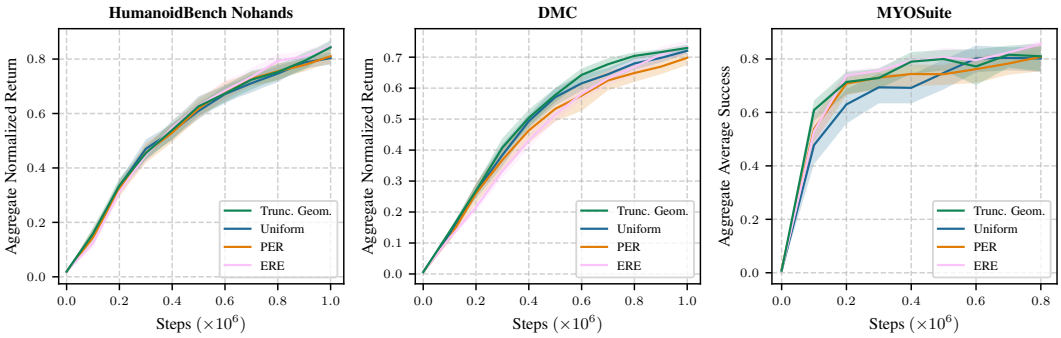

Figure 14: ***(High Replay Volume) Performance on HumanoidBench Nohands tasks.*** We compare different replay strategies on SimbaV2 [17]: ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink). We use UTD=2, standard UTD value taken from original paper on benchmarks HumanoidBench Nohands, DMC. Solid lines show the mean return over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. All x-axes represent environment steps ($\times 10^6$), and y-axes show the denormalized task returns. Truncated Geometric remains effectively neutral when replay volume is high.

F.6 SimbaV2 UTD

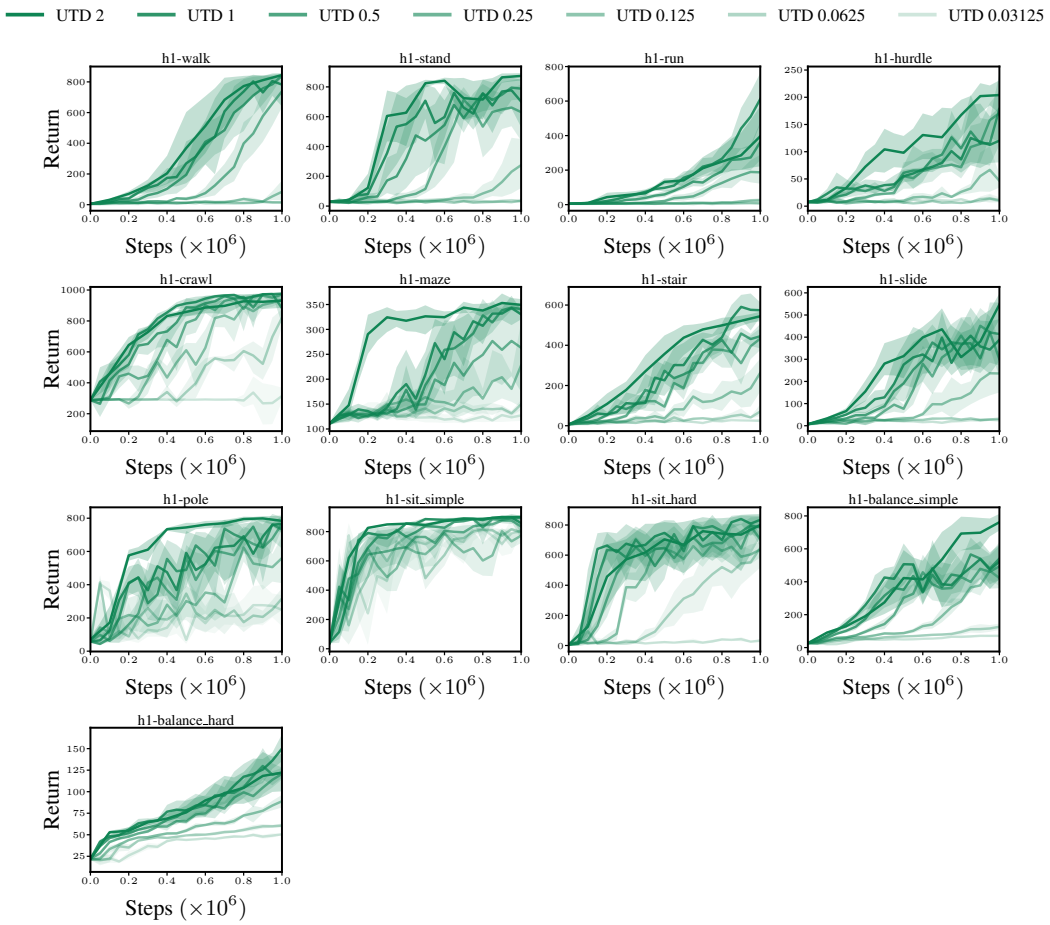

Figure 15: ***(Changing Replay Volume, High Expected Recency) Performance on HumanoidBench Nohands tasks.*** We compare performance of SimbaV2 with Truncated Geometric replay for all UTDs reported in Figure 3. Solid lines show the mean return over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. All x-axes represent environment steps ($\times 10^6$), and y-axes show the denormalized task returns.

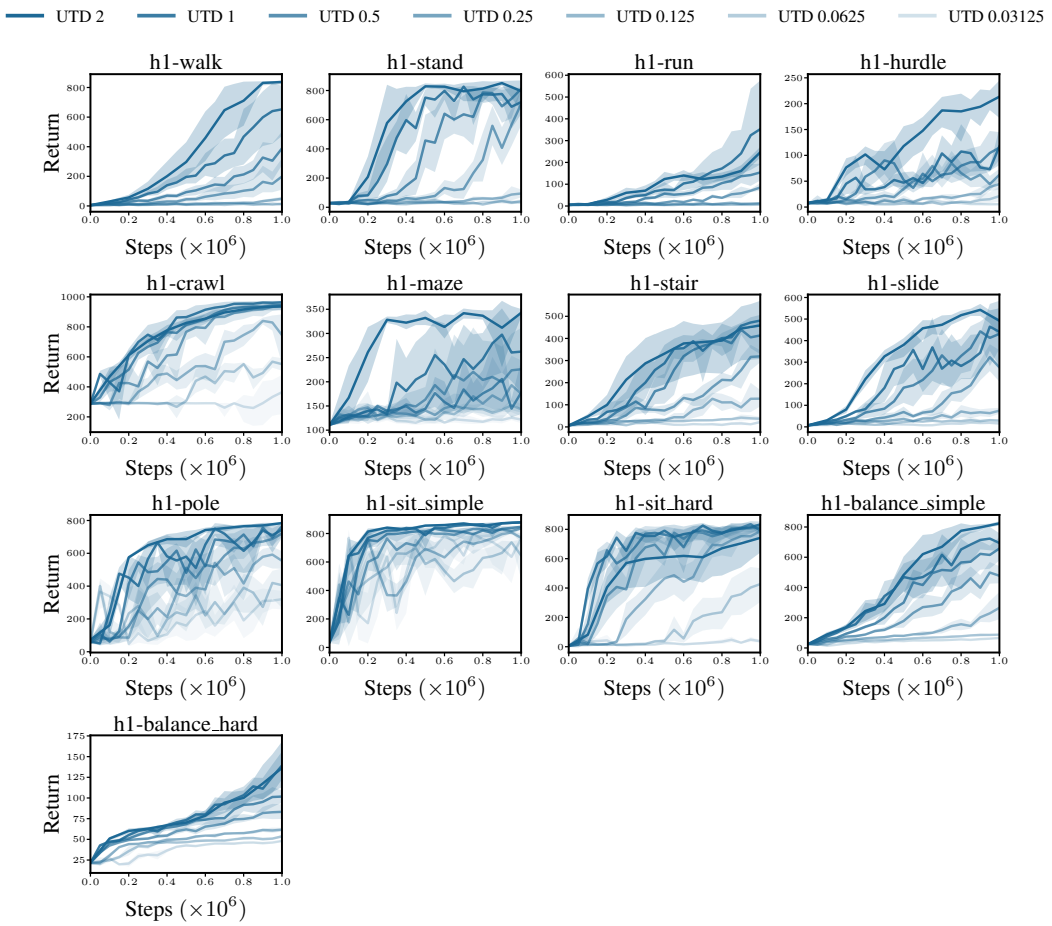

Figure 16: ***(Changing Replay Volume, Low Expected Recency) Performance of SimbaV2 with Uniform replay on HumanoidBench Nohands tasks.*** We report performance for all UTDs from Figure 3. Solid lines show the mean return over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. All x-axes represent environment steps ($\times 10^6$), and y-axes show the denormalized task returns.

F.7 BRC Meta-World

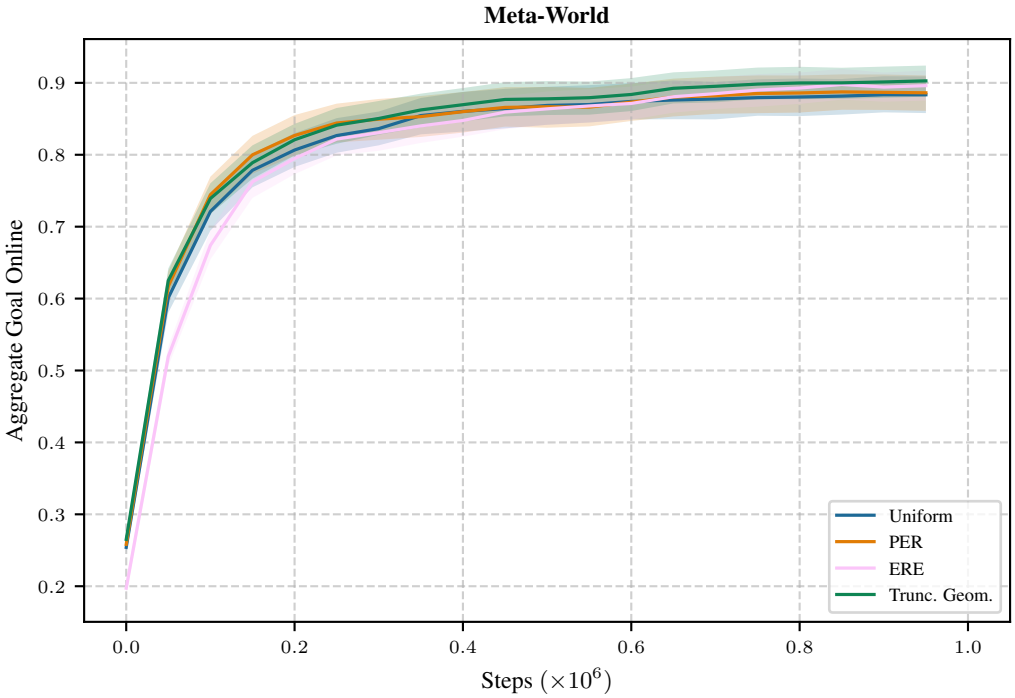

Figure 17: ***(High Replay Volume) Mean performance of BRC [22] on Meta-World*** with ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink) replay strategies. Solid lines show the mean goal online over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`. Truncated Geometric achieves a 2 percentage point improvement over Uniform (90% vs. 88% aggregate success rate), which is noticeable given that Meta-World is largely saturated at this performance level, leaving limited room for improvement.

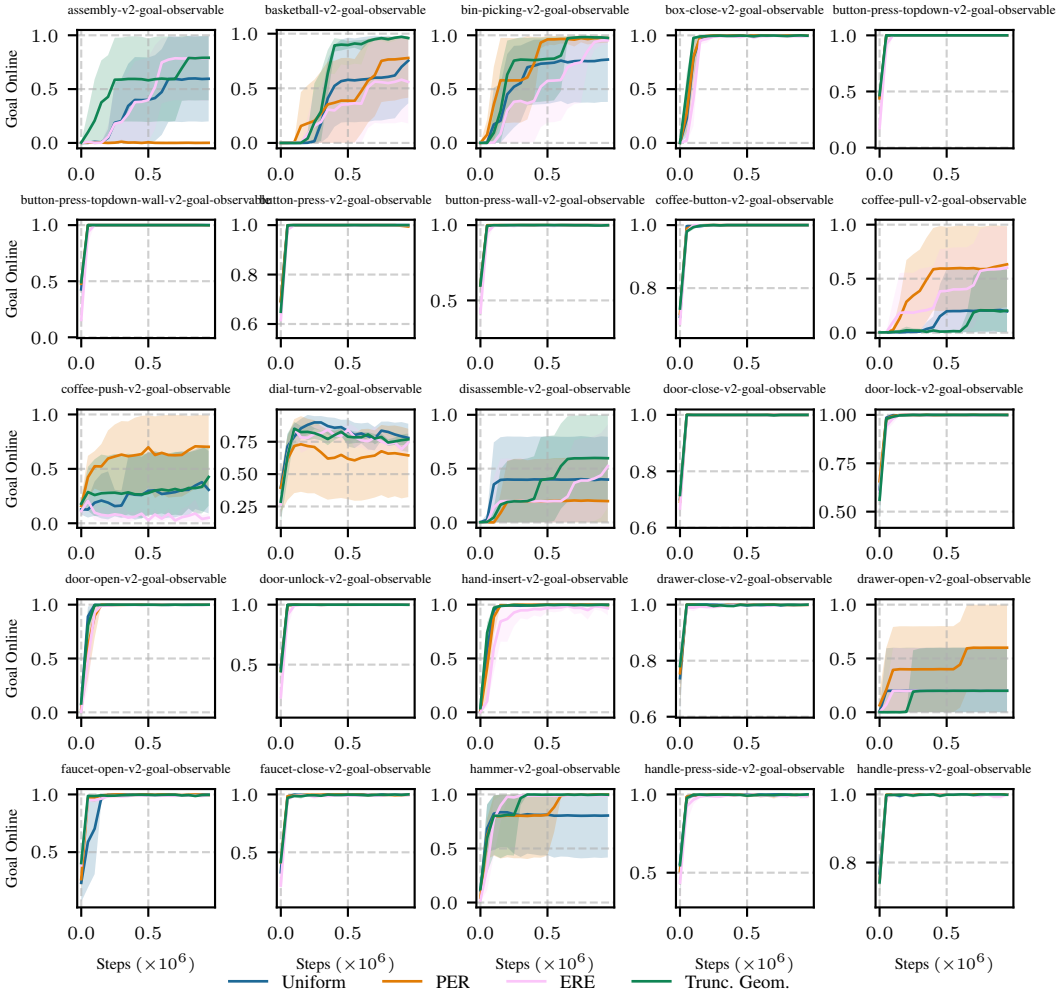

Figure 18: ***(High Replay Volume) Mean performance of BRC [22] with various replay strategies on Meta-World.*** We use ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink). Solid lines show the mean goal online over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`.

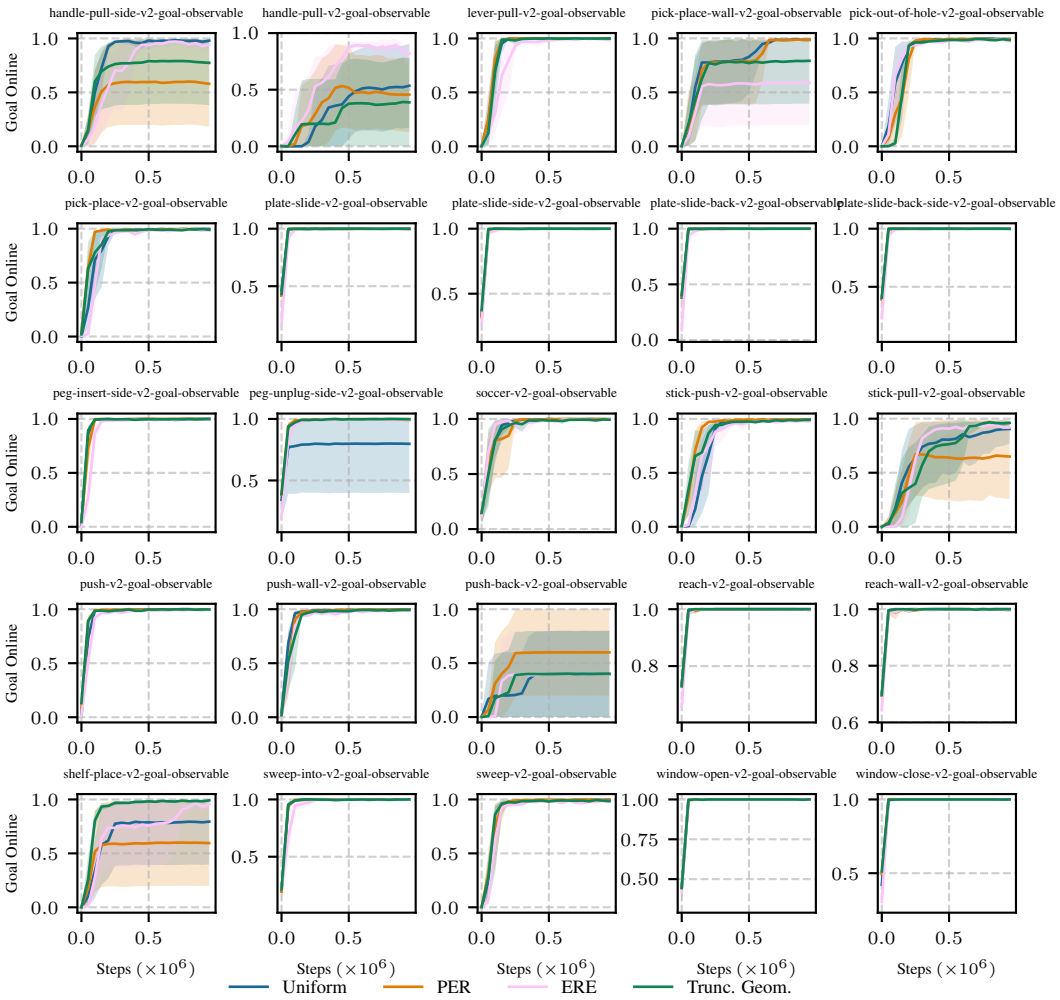

Figure 19: ***(High Replay Volume) Mean performance of BRC [22] with various replay strategies on Meta-World.*** We use ***Low expected recency*** - Uniform (blue), PER (orange) and ***High expected recency*** - Truncated Geometric (green), and ERE (pink). Solid lines show the mean goal online over all available seeds for each method, and shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`.

F.8 FastTD3 Isaac Lab

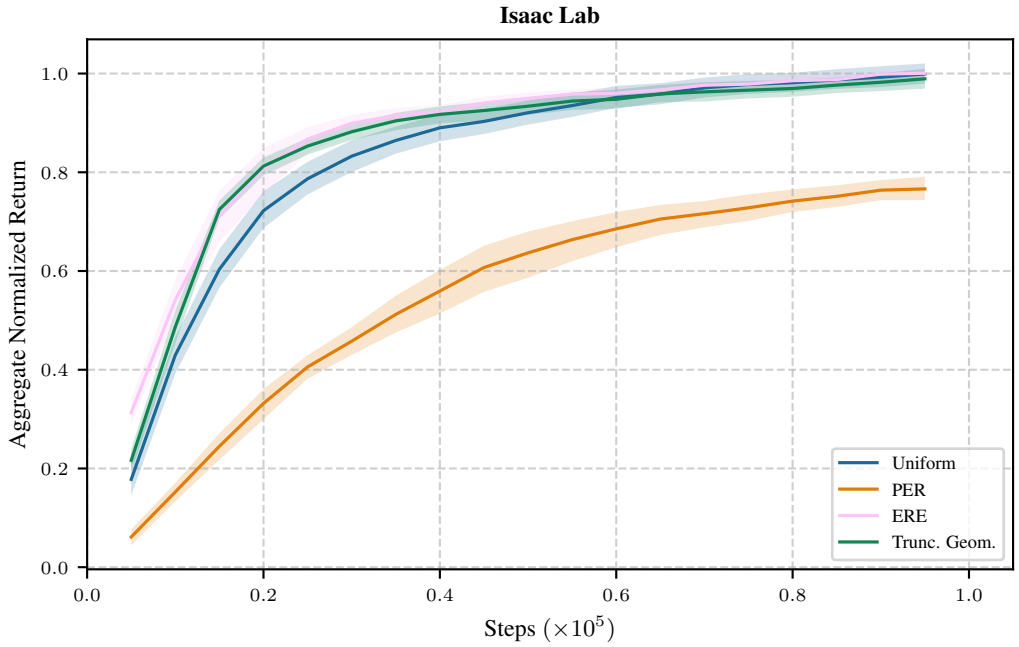

Figure 20: ***(Low Replay Volume, High Expected Recency) Mean performance of FastTD3 [31] with various replay strategies on Isaac Lab.*** In this setting due to low buffer size all replay schemes produce high expected recency replay. Solid lines show mean returns over 5 seeds normalized for each task by dividing returns by mean return from final timestep from uniform sampling strategy. Shaded regions indicate 95% bootstrap confidence intervals computed with `rliable` [1]. Truncated Geometric performs comparably to ERE in aggregate, with ERE benefiting from particularly strong performance on one task.

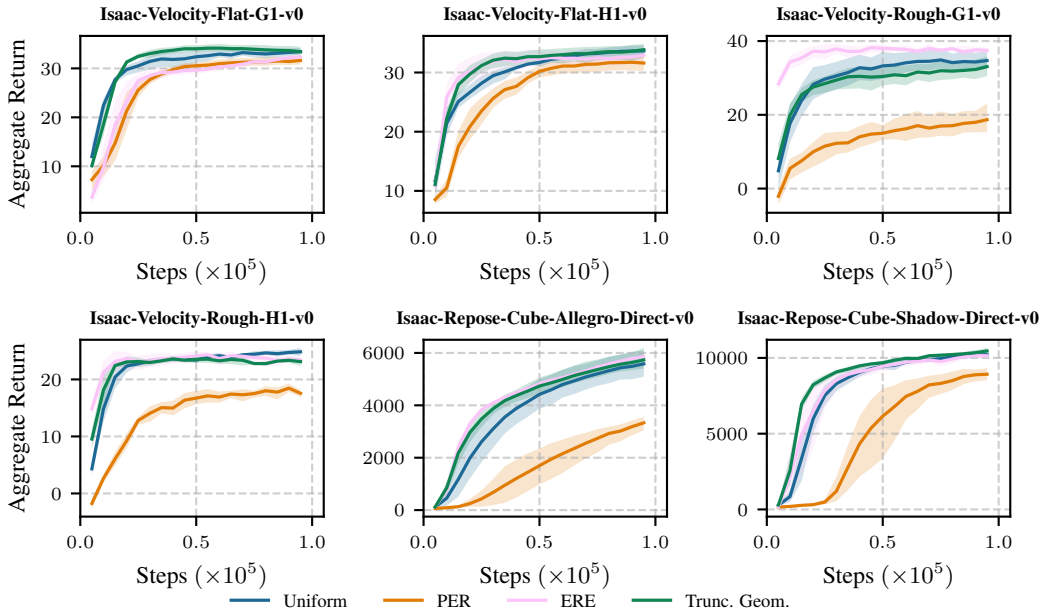

Figure 21: ***(Low Replay Volume, High Expected Recency) Mean performance of FastTD3 on various replay strategy for each Isaac Lab task.*** Solid lines show mean returns over 5 seeds. Shaded regions indicate 95% bootstrap confidence intervals computed with `rliable`.