SYNTHETIC EXPERIENCE REPLAY

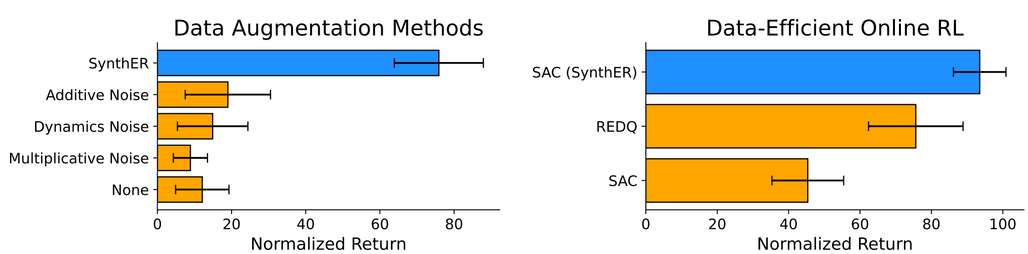
Cong Lu* University of Oxford
Philip J. Ball* University of Oxford
Jack Parker-Holder University of Oxford

ABSTRACT

A key theme in the past decade has been that when large neural networks and large datasets combine they can produce remarkable results. In deep reinforcement learning (RL), this paradigm is commonly made possible through *experience replay*, whereby a dataset of past experiences is used to train a policy or value function. However, unlike in supervised or self-supervised learning, an RL agent has to collect its own data, which is often limited. Thus, it is challenging to reap the benefits of deep learning, and even small neural networks can overfit at the start of training. In this work, we leverage the tremendous recent progress in generative modeling and propose Synthetic Experience Replay (SYNTHER), a diffusion-based approach to arbitrarily upsample an agent's collected experience. We show that SYNTHER is an effective method for training RL agents across offline and online settings. In offline settings, we observe drastic improvements both when upsampling small offline datasets and when training larger networks with additional synthetic data. Furthermore, SYNTHER enables online agents to train with a much higher update-to-data ratio than before, leading to a large increase in sample efficiency, *without any algorithmic changes*. We believe that synthetic training data could open the door to realizing the full potential of deep learning for replay-based RL algorithms from limited data.

1 INTRODUCTION

In the past decade, the combination of large datasets (Deng et al., 2009; Schuhmann et al., 2022) and ever deeper neural networks (Krizhevsky et al., 2012; He et al., 2015; Vaswani et al., 2017; Devlin et al., 2018) has led to a series of more generally capable models (Radford et al., 2019; Brown et al., 2020; Ramesh et al., 2022). In reinforcement learning (RL, Sutton & Barto (2018)), agents typically learn online from their own experience. Thus, to leverage sufficiently rich datasets, RL agents typically make use of *experience replay* (Mnih et al., 2015; Fedus et al., 2020), where training takes place on a dataset of recent experiences. However, this experience is typically limited, unless an agent is distributed over many workers which requires both expensive computational cost and sufficiently fast simulation (Espeholt et al., 2018; Kapturowski et al., 2019).



(a) IQL (Kostrikov et al., 2022) on a reduced 15% subset of walker2d medium-replay (Fu et al., 2020).

(b) SAC (Haarnoja et al., 2018) on 6 DeepMind Control Suite and OpenAI Gym environments.

Figure 1: Upsampling data using SYNTHER greatly outperforms explicit data augmentation schemes for small offline datasets and data-efficient algorithms in online RL *without any algorithmic changes*. Moreover, synthetic data from SYNTHER may readily be added to *any* algorithm utilizing experience replay. Full results are presented in Section 4.

*Equal Contribution. Correspondence to `cong.lu@stats.ox.ac.uk`.

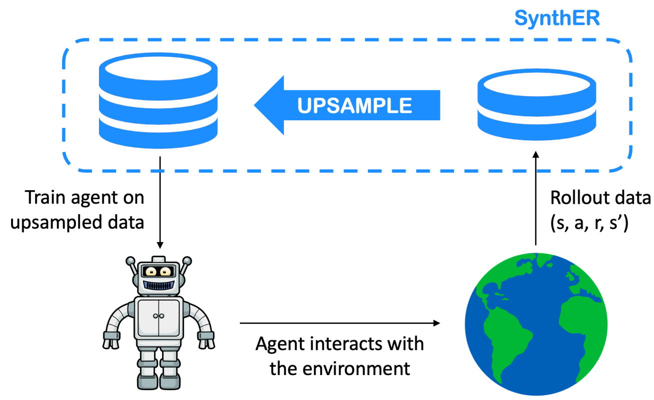


Figure 2: SYNTHER allows any RL agent using experience replay to arbitrarily upsample their experiences and train with synthetic data. By leveraging this increased data, agents can learn more effectively from smaller datasets and can achieve higher sample efficiency.

Another approach for leveraging broad datasets for training RL policies is *offline* RL (Agarwal et al., 2020; Levine et al., 2020), whereby behaviors may be distilled from previously collected data either via behavior cloning (Schaal, 1996), off-policy learning (Kumar et al., 2020; Fujimoto & Gu, 2021) or model-based methods (Yu et al., 2020; Kidambi et al., 2020; Lu et al., 2022a). Offline data can also significantly bootstrap online learning (Hester et al., 2017; Wagenmaker & Pacchiano, 2022; Ball et al., 2023), however, it is a significant challenge to apply these methods when there is a mismatch between offline data and online environment. Thus, many of the successes rely on toy domains with transfer from specific behaviors in a simple proprioceptive environment.

Whilst strong results have been observed in re-using prior data in RL, appropriate data for particular behaviors may simply not exist and thus this approach falls short in generality. We consider an alternative approach—rather than passively reusing data, we leverage tremendous progress in generative modeling to generate a large quantity of new, synthetic data. While prior work has considered upsampling online RL data with VAEs or GANs (Huang et al., 2017; Imre, 2021; Ludjen, 2021), we propose making use of *diffusion* generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Karras et al., 2022), which unlocks significant new capabilities.

Our approach, which we call *Synthetic Experience Replay*, or SYNTHER, is conceptually simple, whereby given a limited initial dataset, we can arbitrarily upsample the data for an agent to use as if it was real experience. Therefore, in this paper, we seek to answer a simple question: *Can the latest generative models replace or augment traditional datasets in reinforcement learning?* To answer this, we consider the following settings: offline RL where we entirely replace the original data with data produced by a generative model, and online RL where we upsample experiences to broaden the training data available to the agent. In both cases, SYNTHER leads to a drastic improvement by utilizing synthetic data, obtaining performance comparable to that of agents trained with substantially more real data. Furthermore, in certain offline settings, the additional data enables effective training of larger policy and value networks, resulting in higher performance by alleviating the representational bottleneck. We thus believe this paper presents sufficient evidence that given additional scale, our approach could enable entirely new efficient training strategies for RL agents.

To summarize, the contributions of this paper are:

- We propose SYNTHER in Section 3, a diffusion-based approach that allows one to generate synthetic experiences and thus arbitrarily upsample data for any reinforcement learning algorithm utilizing experience replay.
- We validate the synthetic data generated by SYNTHER in offline settings in Section 4.1, presenting the first generative approach to show parity with real data on the standard D4RL offline datasets with a wide variety of algorithms. Furthermore, we observe considerable improvements from upsampling for small offline datasets in Section 4.1.1 and scaling up network sizes in Section 4.1.2.
- We show how SYNTHER can arbitrarily upsample an online agent's training data in Section 4.2 by continually training the diffusion model. This allows us to significantly increase an agent's update-to-data (UTD) ratio matching the efficiency of specially designed data-efficient algorithms *without any algorithmic changes*.

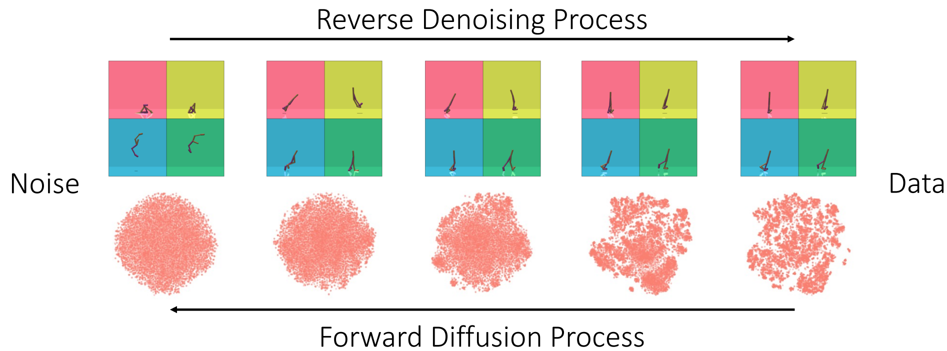


Figure 3: SynthER generates synthetic samples using a diffusion model which we visualize on the walker2d environment. On the **top row**, we render the state component of the transition tuple on a subset of samples; and on the **bottom row**, we visualize a t-SNE (Van der Maaten & Hinton, 2008) projection of 100,000 samples. The denoising process creates cohesive and plausible transitions whilst also remaining diverse, as seen by the multiple clusters that form at the end of the process in the bottom row.

2 Background

2.1 Reinforcement Learning

We model the environment as a Markov Decision Process (MDP, Sutton & Barto (2018)), defined as a tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \rho_0, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces respectively, $P(s'|s, a)$ the transition dynamics, $R(s, a)$ the reward function, ρ_0 the initial state distribution, and $\gamma \in (0, 1)$ the discount factor. The goal in reinforcement learning is to optimize a policy $\pi(a|s)$ that maximizes the expected discounted return $\mathbb{E}_{\pi,P,\rho_0}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$.

2.2 Offline Reinforcement Learning

In *offline RL* (Levine et al., 2020), the policy is not deployed in the environment until test time. Instead, the algorithm only has access to a static dataset $\mathcal{D}_{\text{env}} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{T}$, collected by one or more behavioral policies π_b. We refer to the distribution from which $\mathcal{D}_{\text{env}}$ was sampled as the *behavioral distribution* (Yu et al., 2020). In some of the environments we consider, the environment may be finite horizon or have early termination. In that case, the transition tuple also contains a terminal flag d_t where $d_t = 1$ indicates the episode ended early at timestep t and $d_t = 0$ otherwise.

2.3 Diffusion Models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are a class of generative models inspired by non-equilibrium thermodynamics that learn to iteratively reverse a forward noising process and generate samples from noise. Given a data distribution $p(\mathbf{x})$ with standard deviation σ_{data}, we consider noised distributions $p(\mathbf{x}; \sigma)$ obtained by adding i.i.d. Gaussian noise of standard deviation σ to the base distribution. The forward noising process is defined by a sequence of noised distributions following a fixed noise schedule $\sigma_0 = \sigma_{\max} > \sigma_1 > \cdots > \sigma_N = 0$. When $\sigma_{\max} \gg \sigma_{\text{data}}$, the final noised distribution $p(\mathbf{x}; \sigma_{\max})$ is essentially indistinguishable from random noise.

Karras et al. (2022) consider a probability-flow ODE with the corresponding continuous noise schedule $\sigma(t)$ that maintains the desired distribution as $\mathbf{x}$ evolves through time given by Equation (1).

$$\mathrm{d}\mathbf{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))\mathrm{d}t \tag{1}$$

where the dot indicates a time derivative and $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))$ is the score function (Hyvärinen, 2005), which points towards the data at a given noise level. Infinitesimal forward or backward steps of this ODE either nudge a sample away or towards the data. Karras et al. (2022) consider training a denoiser $D_\theta(\mathbf{x}; \sigma)$ on an L2 denoising objective:

$$\min_\theta \mathbb{E}_{\mathbf{x}\sim p, \sigma, \epsilon\sim\mathcal{N}(0,\sigma^2 I)} \left\| D_\theta(\mathbf{x} + \epsilon; \sigma) - \mathbf{x} \right\|_2^2 \tag{2}$$

and then use the connection between score-matching and denoising (Vincent, 2011) to obtain $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) = (D_\theta(\mathbf{x}; \sigma) - \mathbf{x})/\sigma^2$. We may then apply an ODE (or SDE as a generalization of Equation (1)) solver to reverse the forward process. In this paper, we train our diffusion models to approximate the online or offline behavioral distribution.

Algorithm 1 SYNTHER for online replay-based algorithms. Our additions are highlighted in **blue**.

Input: real data ratio $r \in [0, 1]$
Initialize: $\mathcal{D}_{\text{real}} = \emptyset$ real replay buffer, π agent, $\mathcal{D}_{\text{synthetic}} = \emptyset$ synthetic replay buffer, M diffusion model
for $t = 1, \dots, T$ **do**
    Collect data with π in the environment
    Update diffusion model M with samples from $\mathcal{D}_{\text{real}}$
    Generate samples from M and add them to $\mathcal{D}_{\text{synthetic}}$
    Train π on samples from $\mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{synthetic}}$ mixed with ratio r
end for

3 SYNTHETIC EXPERIENCE REPLAY

In this section, we introduce Synthetic Experience Replay (SYNTHER), our approach to upsampling an agent's collected experience using diffusion. We begin by describing the simpler training process used for offline reinforcement learning and then how that may be adapted to the online setting by continually training the diffusion model.

3.1 OFFLINE SYNTHER

For offline reinforcement learning, we take the data distribution of the diffusion model $p(\mathbf{x})$ to simply be the offline behavioral distribution. For the proprioceptive environments we consider, the full transition is low-dimensional compared with typical pixel-based applications of diffusion. Therefore, the network architecture is an important design choice; we elect to use a residual MLP denoising (Tolstikhin et al., 2021) network, with detailed hyperparameters given in Appendix B. We visualize the denoising process when trained on a representative D4RL (Fu et al., 2020) offline dataset, walker2d medium-replay, in Figure 3. We further validate our diffusion model on the D4RL datasets in Figure 7 in Appendix A by showing that the synthetic data closely matches the original data when comparing the marginal distribution over each dimension.

Next, we conduct a quantitative analysis and show that **the quality of the samples from the diffusion model is significantly better** than with prior generative models such as VAEs (Kingma & Welling, 2014) and GANs (Goodfellow et al., 2014). We consider the state-of-the-art Tabular VAE (TVAE) and Conditional Tabular GAN (CTGAN) models proposed by Xu et al. (2019) for low-dimensional data with the default hyperparameters, and evaluate on the D4RL halfcheetah medium-replay dataset. As proposed in Patki et al. (2016), we compare the following two high-level statistics on the synthetic data:

- **Marginal:** (Number between 0-1) Aggregate Kolmogorov-Smirnov (Massey Jr, 1951) statistic, measuring the maximum distance between cumulative distribution functions, for each individual dimension of the data.
- **Correlation:** (Number between 0-1) Aggregate pairwise Pearson rank correlation (Fieller et al., 1957) between the dimensions of the data.

We also assess down-stream offline RL performance using the synthetic data with two state-of-the-art offline RL algorithms, TD3+BC (Fujimoto & Gu, 2021) and IQL (Kostrikov et al., 2022), in Table 1. The full evaluation protocol is described in Section 4.1.

Table 1: SYNTHER is better at capturing both the high-level statistics of the dataset (halfcheetah medium-replay) than prior generative models and also leads to far higher downstream performance. Metrics (left) computed from 100K samples from each model, offline RL performance (right) computed using 5M samples from each model. We show the mean and standard deviation of the final performance averaged over 4 seeds.

Model	**Metrics**		**Eval. Return**	
	Marginal	**Correlation**	**TD3+BC**	**IQL**
Diffusion (Ours)	**0.989**	**0.998**	**45.4±0.4**	**46.7±0.1**
VAE	0.941	0.976	12.3±5.6	6.7±1.0
GAN	0.947	0.979	11.0±3.3	4.6±2.4

Table 2: A comprehensive evaluation of SYNTHER on a wide variety of D4RL (Fu et al., 2020) datasets and selection of state-of-the-art offline RL algorithms. We show that synthetic data from SYNTHER faithfully reproduces the original performance, which allows us to completely eschew the original training data. We show the mean and standard deviation of the final performance averaged over 4 seeds. **Highlighted** aggregated results show at least parity over each group of results.

Environment	Behavioral Policy	TD3+BC		IQL		EDAC	
		Original	SYNTHER	Original	SYNTHER	Original	SYNTHER
halfcheetah-v2	random	11.3±0.8	10.9±0.4	15.2±1.2	19.4±0.3	-	
	mixed	44.8±0.7	45.4±0.4	43.5±0.4	46.7±0.1	62.1±1.3	62.9±1.8
	medium	48.1±0.2	48.8±0.3	48.3±0.1	49.9±0.2	67.7±1.2	64.2±1.2
	medexp	90.8±7.0	85.9±8.2	94.6±0.2	93.6±1.7	104.8±0.7	94.0±8.3
walker2d-v2	random	0.6±0.3	3.4±1.8	4.1±0.8	4.8±0.7	-	
	mixed	85.6±4.6	91.9±1.4	82.6±8.0	90.2±4.8	87.1±3.2	84.9±1.6
	medium	82.7±5.5	85.2±1.1	84.0±5.4	83.2±5.6	93.4±1.6	88.2±1.6
	medexp	110.0±0.4	110.1±0.3	111.7±0.6	111.8±0.7	114.8±0.9	113.6±0.5
hopper-v2	random	8.6±0.3	17.8±11.2	7.2±0.2	7.5±0.5	-	
	mixed	64.4±24.8	54.0±10.8	84.6±13.5	102.8±0.3	99.7±0.9	97.6±1.6
	medium	60.4±4.0	63.0±4.3	62.8±6.0	71.8±4.3	101.7±0.3	101.0±0.7
	medexp	101.1±10.5	102.5±10.9	106.2±6.1	97.5±8.8	105.2±11.6	109.7±0.2
locomotion average		**59.0±4.9**	**59.9±4.3**	**62.1±3.5**	**64.9±2.3**	**92.9±2.4**	**90.7±1.9**
maze2d-v1	umaze	29.4±14.2	40.5±9.8	37.7±2.0	40.5±1.0	95.3±7.4	101.5±20.6
	medium	59.5±41.9	67.1±36.6	35.5±1.0	34.1±0.2	57.0±4.0	69.4±8.0
	large	97.1±29.3	128.0±41.3	49.6±22.0	48.7±4.6	95.6±26.5	161.6±9.7
maze average		**62.0±28.2**	**78.5±29.2**	**40.9±8.3**	**41.1±1.9**	**82.6±12.6**	**110.8±12.8**

We see that the diffusion model is more faithful to the original data than prior generative models and thus the synthetic data leads to substantially higher returns on both the TD3+BC and IQL algorithms. Thus, we hypothesize a large part of the failure of prior methods (Imre, 2021; Ludjen, 2021) is due to the use of a weaker generative model.

3.2 ONLINE SYNTHER

SYNTHER may be used to upsample an online agent's experiences by continually training the diffusion model on new experiences. We provide pseudocode for how to incorporate SYNTHER into any online replay-based RL agent in Algorithm 1 and visualize this in Figure 2. Concretely, a diffusion model is periodically updated on the real transitions and then used to populate a second synthetic buffer. The agent may then be trained on a mixture of real and synthetic data sampled with ratio r. For the results in Section 4.2, we simply set $r = 0.5$ following Ball et al. (2023). The synthetic replay buffer may also be configured with a finite capacity to prevent overly stale data.

4 EMPIRICAL EVALUATION

In our empirical evaluation, we evaluate SYNTHER across a wide variety of offline and online settings. We first validate our approach on offline RL, where we entirely replace the original data, and further show large benefits from upsampling small offline datasets. Next, we show that SYNTHER leads to large improvements in sample efficiency in online RL where we upsample recent experiences to broaden the training data available to the agent.

4.1 OFFLINE D4RL EVALUATION

We first verify that synthetic samples from SYNTHER faithfully model the underlying distribution from the original D4RL (Fu et al., 2020) datasets. To do this, we evaluate SYNTHER in combination with 3 distinct SOTA offline RL algorithms: TD3+BC (Fujimoto & Gu (2021), explicit policy regularization), IQL (Kostrikov et al. (2022), expectile regression), and EDAC (An et al. (2021), uncertainty-based regularization) on an extensive selection of D4RL datasets. We consider the MuJoCo (Todorov et al., 2012) locomotion (halfcheetah, walker2d, and hopper) and maze2d environments. For these experiments, we use a residual MLP denoising network with full hyperparameters given in Appendix B. All datasets share the same training hyperparameters, with some

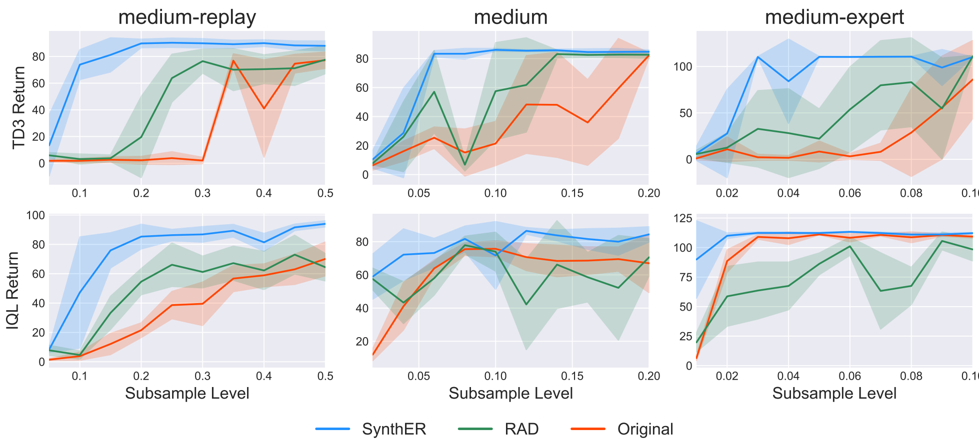


Figure 4: SYNTHER is a powerful method for upsampling reduced variants of the walker2d datasets and vastly improves on competitive explicit data augmentation approaches for both the TD3+BC (top) and IQL (bottom) algorithms. The subsampling levels are scaled proportionally to the original size of each dataset. We show the mean and standard deviation of the final performance averaged over 4 seeds.

larger datasets using a wider network. For every dataset, we upsample the original dataset to **5M samples**; we justify this choice in Appendix C.1. We show the final performance in Table 2.

Our results show that we achieve at least parity for all groups of environments and algorithms as highlighted in the table, *regardless of the precise details of each algorithm*. We note significant improvements to maze2d environments, which are close to the 'best' performance as reported in CORL (Tarasov et al., 2022) (i.e., the best iteration during offline training) rather than the final performance. We hypothesize this improvement is largely due to increased data from SYNTHER, which leads to less overfitting and increased stability. For the locomotion datasets, we largely reproduce the original results, which we attribute to the fact that most D4RL datasets are at least 1M in size and are already sufficiently large. However, as detailed in Table 4 in Appendix A.1, SYNTHER allows the effective size of the dataset to be compressed significantly, up to $12.9\times$ on some datasets.

4.1.1 UPSAMPLING FOR SMALL DATASETS

Next, we investigate the benefit of SYNTHER for small offline datasets and compare it to canonical 'explicit' data augmentation approaches (Laskin et al., 2020; Ball et al., 2021). Concretely, we wish to understand whether SYNTHER generalizes and generates synthetic samples that improve policy learning compared with *explicitly* augmenting the data with hand-designed inductive biases. We focus on the walker2d (medium, medium-replay/mixed, medium-expert) datasets in D4RL and uniformly subsample each at the transition level. We subsample each dataset proportional to the original dataset size so that the subsampled datasets approximately range from 20K to 200K samples. As in Section 4.1, we then use SYNTHER to *upsample* each dataset to 5M transitions. Our denoising network uses the same hyperparameters as for the original evaluation in Section 4.1.

In Figure 4, we can see that for all datasets, SYNTHER leads to a significant gain in performance and vastly improves on explicit data augmentation approaches. For explicit data augmentation, we select the overall most effective augmentation scheme from Laskin et al. (2020) (adding Gaussian noise of the form $\epsilon \sim \mathcal{N}(0, 0.1)$). Notably, with SYNTHER we can achieve close to the original levels of performance on the walker2d-medium-expert datasets starting from **only 3% of the original data**. In Figure 1a, we methodically compare across both additive and multiplicative versions of RAD, as well as dynamics augmentation (Ball et al., 2021) on the 15% reduced walker medium-replay dataset.

Why is SYNTHER better than explicit augmentation? To provide intuition into the efficacy of SYNTHER over canonical explicit augmentation approaches, we compare the data generated by SYNTHER to that generated by the best-performing data augmentation approach in Figure 1a, namely additive noise. We wish to evaluate two properties: 1) How diverse is the data? 2) How accurate is the data for the purposes of learning policies? To measure diversity, we measure the *minimum* L2 distance of each datapoint from the dataset, which allows us to see how far the upsampled data is from the original data. To measure the validity of the data, we follow Lu et al. (2022a) and measure the MSE between the reward and next state proposed by SYNTHER with the true next state and reward defined by the simulator. We plot both these values in a joint scatter plot to compare how they vary with respect to each other. For this, we compare specifically on the reduced 15% subset of walker2d medium-replay as in Figure 1a. As we see in Figure 5, SYNTHER generates a significantly wider marginal distribution over the distance from the dataset, and generally produces samples that are further away from the dataset than explicit augmentations. Remarkably, however, we see that these samples are far more consistent with the true environment dynamics. Thus, SYNTHER generates samples that have significantly lower dynamics MSE than explicit augmentations, even for datapoints that are far away from the training data. This implies that a high level of generalization has been achieved by the SYNTHER model, resulting in the ability to generate **novel, diverse, yet dynamically accurate data** that can be used by policies to improve performance.

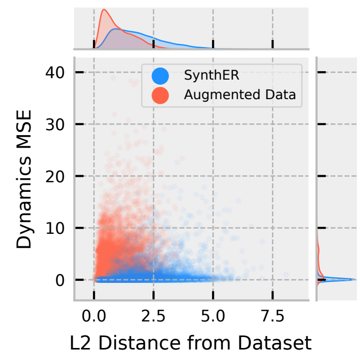


Figure 5: Comparing L2 distance from training data and dynamics accuracy under SYNTHER and augmentations.

4.1.2 SCALING NETWORK SIZE

A further benefit we observe from SYNTHER on the TD3+BC algorithm is that upsampled data can enable scaling of the policy and value networks leading to improved performance. As is typical for RL algorithms, TD3+BC uses a small value and policy network with two hidden layers, and width of 256, and a batch size of 256. We consider increasing the size of both networks to be three hidden layers and width 512 (approximately 6$\times$ more parameters), and the batch size to 1024 to better make use of the upsampled data in Table 3.

We observe a large overall improvement of **12.2%** for the locomotion datasets when using a larger network with synthetic data (Larger Network + SYNTHER). Notably, when using the original data (Larger Network + Original Data), the larger network performs the same as the baseline. This suggests that the bottleneck in the algorithm lies in the representation capability of the neural network and *synthetic samples from* SYNTHER *enables effective training of the larger network*. This could alleviate the data requirements for scaling laws in reinforcement learning (Adaptive Agent Team et al., 2023; Hilton et al., 2023). However, for the IQL and EDAC algorithms, we did not observe an improvement by increasing the network size which suggests that the bottleneck there lies in the data or algorithm rather than the architecture.

Table 3: SYNTHER enables effective training of larger policy and value networks for TD3+BC leading to a **12.2%** gain on the offline MuJoCo locomotion datasets. In comparison, simply increasing the network size with the original data does not improve performance. We show the mean and standard deviation of the final performance averaged over 4 seeds.

Environment	Behavioral Policy	Baseline	Larger Network	
			Original Data	SYNTHER
halfcheetah-v2	random	11.3±0.8	11.0±0.6	**12.8±0.7**
	mixed	44.8±0.7	44.9±0.5	**48.1±0.4**
	medium	48.1±0.2	48.5±0.2	**53.4±0.1**
	medexp	90.8±7.0	91.0±3.8	**101.4±1.1**
walker2d-v2	random	0.6±0.3	1.8±1.8	**3.5±2.0**
	mixed	85.6±4.6	82.5±7.3	**93.6±2.3**
	medium	82.7±5.5	84.5±1.0	**88.0±0.4**
	medexp	110.0±0.4	110.2±0.4	110.3±0.2
hopper-v2	random	8.6±0.3	8.2±0.8	17.0±11.3
	mixed	64.4±24.8	66.2±17.5	**88.8±10.8**
	medium	60.4±4.0	58.7±5.7	65.6±4.1
	medexp	101.1±10.5	97.8±7.9	**111.5±0.5**
locomotion average		59.0±4.9	58.8±4.0	**66.2±2.8**

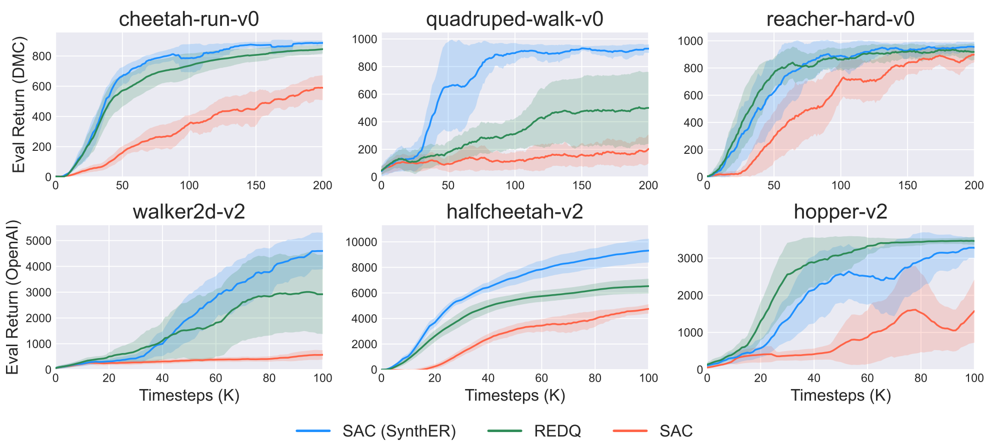


Figure 6: SYNTHER greatly improves the sample efficiency of online RL algorithms by enabling an agent to train on upsampled data. This allows an agent to use an increased update-to-data ratio (UTD=20 compared to 1 for regular SAC) *without any algorithmic changes*. We show the mean and standard deviation of the online return over 4 seeds. DeepMind Control Suite environments are shown in the top row, and OpenAI Gym environments are shown in the bottom.

4.2 Online Evaluation

Finally, we show that SYNTHER can effectively upsample an online agent's continually collected experiences on 3 environments from the DeepMind Control Suite (DMC, Tunyasuvunakool et al. (2020)) (cheetah-run, quadruped-walk, and reacher-hard) and 3 environments the OpenAI Gym Suite (Brockman et al., 2016) (walker2d, halfcheetah, and hopper). We choose the base algorithm to be SAC (Haarnoja et al., 2018), a popular off-policy entropy-regularized algorithm, and benchmark against a SOTA data-efficient variant of itself, REDQ (Chen et al., 2021). REDQ uses an ensemble of 10 Q-functions and computes target values across a randomized subset of them during training. By default, SAC uses an update-to-data ratio of 1 (1 update for each transition collected); the modifications to SAC in REDQ enable this to be raised to 20. Our method, 'SAC (SYNTHER)', augments the training data by generating 1M new samples for every 10K real samples collected and samples them with a ratio $r = 0.5$. We then match REDQ and train with a UTD ratio of 20. We evaluate our algorithms over 200K online steps for the DMC environments and 100K for OpenAI Gym.

In Figure 6, we see that SAC (SYNTHER) matches or outperforms REDQ on the majority of the environments with particularly strong results on the quadruped-walk and halfcheetah-v2 environments. This is particularly notable as D'Oro et al. (2023) found that UTD=20 on average *decreased performance* for SAC compared with the default value of 1, attributable to issues with overestimation and overfitting (Chen et al., 2021; Li et al., 2023). We aggregate the final performance on the environments in Figure 1b, normalizing the DMC returns following Lu et al. (2022b) and OpenAI returns as in D4RL. Moreover, due to the fast speed of training our diffusion models and fewer Q-networks, our approach is in fact faster than REDQ based on wall-clock time, whilst also requiring fewer algorithmic design choices, such as large ensembles and random subsetting. Full details on run-time are given in Appendix D.2.

5 Related Work

Whilst generative training data has been explored in reinforcement learning; in general, synthetic data has not previously performed as well as real data on standard RL benchmarks.

Generative Training Data. Imre (2021); Ludjen (2021) had both considered using VAEs and GANs to generate synthetic data for online reinforcement learning. However, we note that both works failed to match the original performance on simple environments such as CartPole—this is likely due to the use of a weaker class of generative models which we explored in Section 3.1. Huang et al. (2017) considered using GAN samples to *pre-train* an RL policy, observing a modest improvement in sample efficiency for CartPole. Yu et al. (2023); Chen et al. (2023) consider augmenting the image observations of robotic control data using a text-guided diffusion model whilst maintaining the same action. This differs from our approach which models the entire transition and *can synthesize novel action and reward labels*.

Outside of reinforcement learning, He et al. (2022) consider generative training data for image classification and pre-training. They also find that synthetic data is useful for data-scarce settings which are especially prevalent in reinforcement learning. Sehwag et al. (2022) consider generative training data to improve adversarial robustness in image classification. In continual learning, "generative replay" (Shin et al., 2017) has been considered to compress examples from past tasks to prevent forgetting.

Generative Modeling in RL. Prior work in diffusion modeling for offline RL has largely sought to supplant traditional reinforcement learning with "upside-down RL" (Schmidhuber, 2019). Diffuser (Janner et al., 2022) models long sequences of transitions or full episodes and can bias the whole trajectory with guidance towards high reward or a particular goal. It then takes the first action and re-plans by receding horizon control. Decision Diffuser (Ajay et al., 2022) similarly operates at the sequence level but instead uses conditional guidance on rewards and goals. (Du et al., 2023) present a similar trajectory-based algorithm for visual data. In contrast, SYNTHER operates at the transition level and seeks to be readily compatible with existing reinforcement learning algorithms.

Pearce et al. (2023) consider a diffusion-based approach to behavioral cloning, whereby a state-conditional diffusion model may be used to sample actions that imitate prior data. Azar et al. (2012); Li et al. (2020) provide theoretical sample complexity bounds for model-based reinforcement learning given access to a generative model.

Model-Based Reinforcement Learning. We note the parallels between the synthetic data we generate and model-based reinforcement learning (Janner et al., 2019; Yu et al., 2020; Lu et al., 2022a); model-based methods generate synthetic samples by rolling out from previously seen states. Two key differences to our method are: SYNTHER synthesizes new experiences without the need to start from a real state and the generated experiences are distributed exactly according to the data, rather than subject to compounding model errors due to inaccuracies and policy exploitation. Furthermore, SYNTHER is an orthogonal approach and could in fact be *combined with* model-based RL by sampling initial states to rollout from, which could lead to increased diversity.

6 CONCLUSION

In this paper, we proposed SYNTHER, a powerful and general method for upsampling agent experiences in any reinforcement learning algorithm using experience replay. We integrated SYNTHER with ease on four distinct algorithms, each fine-tuned for its own use case, **with no algorithmic modification**. Our results show the potential of synthetic training data when combined with modern diffusion models. In offline reinforcement learning, SYNTHER allows training from extremely small datasets, scaling up policy and value networks, and high levels of data compression. In online reinforcement learning, the additional data allows online agents to use higher update-to-data ratios leading to increased sample efficiency.

We believe that scaling SYNTHER to more settings would unlock exciting new capabilities for RL agents. SYNTHER could readily be extended to n-step formulations of experience replay by simply expanding the input space of the diffusion model. Furthermore, as diffusion models have to date been largely applied in the visual domain, we expect the findings of this paper to transfer to an even greater extent in the visual setting. One could leverage the generalization capability of image-based diffusion models to synthesize novel views and configurations of a pixel-based environment. Finally, by leveraging guidance for diffusion models (Ho & Salimans, 2021), the generated synthetic data could be biased towards certain modes, resulting in new, transferable, and composable sampling strategies for RL algorithms.

ACKNOWLEDGMENTS

Cong Lu is funded by the Engineering and Physical Sciences Research Council (EPSRC). Philip Ball is funded through the Willowgrove Studentship. The authors would like to thank the anonymous Reincarnating Reinforcement Learning Workshop at ICLR 2023 reviewers for positive and constructive feedback which helped to improve the paper.

REFERENCES

Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei Zhang. Human-timescale adaptation in an open-ended task space, 2023. URL `https://arxiv.org/abs/2301.07608`.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/3d3d286a8d153a4a58156d0e02d8570c-Paper.pdf`.

Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 619–629. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/ball21a.html`.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data, 2023. URL `https://arxiv.org/abs/2302.02948`.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=AY8zfZm0tDd`.

Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation, 2023. URL `https://arxiv.org/abs/2302.06671`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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*, 2023. URL `https://openreview.net/forum?id=OpC-9aBBVJe`.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023. URL https://arxiv.org/abs/2302.00111.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470–481, 1957. ISSN 00063444. URL http://www.jstor.org/stable/2332878.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2022. URL https://arxiv.org/abs/2210.07574.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John Agapiou, Joel Z. Leibo, and Audrunas Gruslys. Deep q-learning from demonstrations, 2017. URL https://arxiv.org/abs/1704.03732.

Jacob Hilton, Jie Tang, and John Schulman. Scaling laws for single-agent reinforcement learning, 2023. URL https://arxiv.org/abs/2301.13442.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6nbpPqUCIi7.

Vincent Huang, Tobias Ley, Martha Vlachou-Konchylaki, and Wenfeng Hu. Enhanced experience replay generation for efficient reinforcement learning, 2017. URL https://arxiv.org/abs/1705.08245.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Baris Imre. An investigation of generative replay in deep reinforcement learning, January 2021. URL http://essay.utwente.nl/85772/.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.

Steven Kapturowski, Georg Ostrovski, Will Dabney, John Quan, and Remi Munos. Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, 2012.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19884–19895. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e615c82aba461681ade82da2da38004a-Paper.pdf.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

Qiyang Li, Aviral Kumar, Ilya Kostrikov, and Sergey Levine. Efficient deep reinforcement learning requires regulating statistical overfitting. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=14-kr46GvP-`.

Cong Lu, Philip Ball, Jack Parker-Holder, Michael Osborne, and Stephen J. Roberts. Revisiting design choices in offline model based reinforcement learning. In *International Conference on Learning Representations*, 2022a. URL `https://openreview.net/forum?id=zz9hXVhf40`.

Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A. Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations, 2022b. URL `https://arxiv.org/abs/2206.04779`.

A.C.P.P. Ludjen. Generative replay in deep reinforcement learning, June 2021. URL `http://essay.utwente.nl/87315/`.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas 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*, 518:529–33, 02 2015. doi: 10.1038/nature14236.

N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, Oct 2016. doi: 10.1109/DSAA.2016.49.

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Pv1GPQzRrC8`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL `https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf`.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Stefan Schaal. Learning from demonstration. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL `https://proceedings.neurips.cc/paper/1996/file/68d13cf26c4b4f4f932e3eff990093ba-Paper.pdf`.

Juergen Schmidhuber. Reinforcement learning upside down: Don't predict rewards–just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=M3Y74vmsMcY`.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WVX0NNVBBkV.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/0efbe98067c6c73dba1250d2beaa81f9-Paper.pdf.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. CORL: Research-oriented deep offline reinforcement learning library. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022. URL https://openreview.net/forum?id=SyAS49bBcv.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=EI2KOXKdnP.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning, 2022. URL https://arxiv.org/abs/2211.04974.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Dee M, Jodilyn Peralta, Brian Ichter, Karol Hausman, and Fei Xia. Scaling robot learning with semantically imagined experience, 2023. URL https://arxiv.org/abs/2302.11550.

SUPPLEMENTARY MATERIAL

A DATA MODELING

In this section, we provide further details for our data modeling. Our diffusion model generates full environment transitions i.e., a concatenation of states, actions, rewards, next states, and terminals where they are present. For the purposes of modeling, we normalize each continuous dimension (non-terminal) to have 0 mean and 1 std. We visualize the marginal distributions over the state, action, and reward dimensions on the standard halfcheetah medium-replay dataset in Figure 7 and observe that the synthetic samples accurately match the high-level statistics of the original dataset.

We note the difficulties of appropriately modeling the terminal variable which is a binary variable compared to the rest of the dimensions which are continuous for the environments we investigate. This is particularly challenging for "expert" datasets where early termination is rare. For example, walker2d-expert only has $\approx 0.0001\%$ terminals. In practice, we find it sufficient to leave the terminals un-normalized and round them to 0 or 1 by thresholding the continuous diffusion samples in the middle at 0.5. A cleaner treatment of this variable could be achieved by leveraging work on diffusion with categorical variables (Hoogeboom et al., 2021).

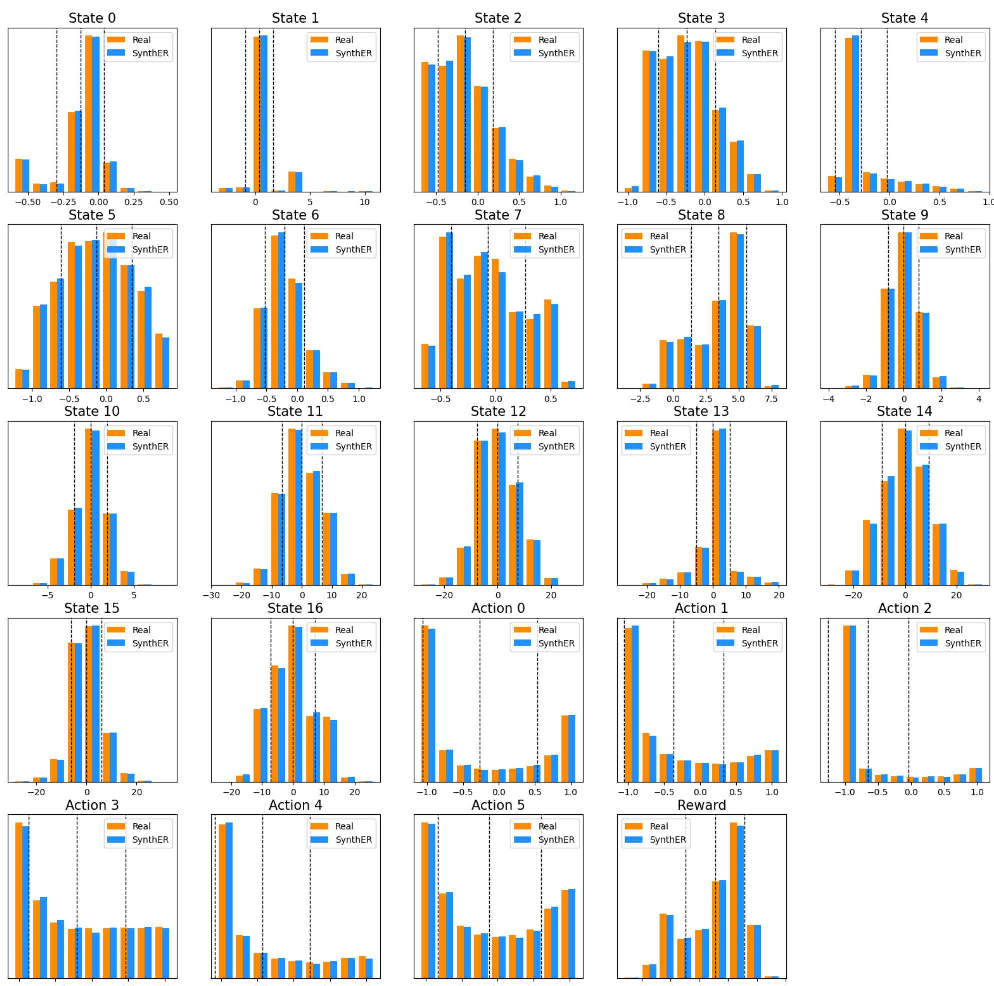


Figure 7: Histograms of the empirical marginal distribution of samples from SYNTHER in **blue** on the halfcheetah medium-replay dataset against the original data in **orange**. Dashed lines indicate the mean $\pm$ one standard deviation in the original dataset. SYNTHER faithfully reproduces the high-level statistics of the dataset.

A.1 DATA COMPRESSION

An immediate advantage of sampling data from a generative model is compression. In Table 4, we compare the memory requirements of SYNTHER and the original data by the number of 32-bit floating point numbers used by each for some sample D4RL (Fu et al., 2020) datasets. For the original data, this simply scales linearly with the size of the dataset. On other hand, SYNTHER amortizes this in the number of parameters in the denoising network, resulting in a high level of dataset compression, at the cost of sampling speed. This property was also noted in the continual learning literature with generative models summarizing previous tasks (Shin et al., 2017). As we discuss in Appendix B.2, sampling is fast with 100K transitions taking around 90 seconds to generate.

Table 4: SYNTHER provides high levels of dataset compression *without sacrificing downstream performance* in offline reinforcement learning. Statistics shown are for the standard D4RL MuJoCo walker2d datasets which has a transition dimension of 42, and the residual denoiser used for evaluation on these environments in Section 4.1. Figures are given to 1 decimal place.

Dataset	**# FP32s in Original Dataset**	**# Diffusion Parameters**	**Compression**
mixed	12.6M		1.9×
medium	42M	6.5M	6.5×
medium-expert	84M		12.9×

B HYPERPARAMETERS

The formulation of diffusion we use in our paper is the Elucidated Diffusion Model (EDM, Karras et al. (2022)). We parametrize the denoising network D_θ as a simple residual MLP.

B.1 DENOISING NETWORK

Our denoising network D_θ is an MLP with skip connections from the previous layer as in Tolstikhin et al. (2021). Thus each layer has the form given in Equation (3).

$$x_{L+1} = \text{linear}(\text{activation}(x_L)) + x_L \tag{3}$$

The hyperparameters are listed in Table 5. The noise level of the diffusion process is encoded by a Random Fourier Feature (Rahimi & Recht, 2007) embedding. The base size of the network uses a width of 1024 and depth of 6 and thus has $\approx$ 6M parameters. We adjust the batch size for training based on dataset size. For online training and offline datasets with fewer than 1 million samples (medium-replay datasets) we use a batch size of 256, and 1024 otherwise.

For the following offline datasets, we observe more performant samples by increasing the width up to 2048: halfcheetah medium-expert, hopper medium, and hopper medium-expert. This raises the network parameters to ≈ 25M, which remains fewer parameters than the original data as in Table 4. We provide ablations on the depth and type of network used in Table 7.

Table 5: Default Residual MLP Denoiser Hyperparameters.

Parameter	**Value(s)**
no. layers	6
width	1024
batch size	{ 256 for online and medium-replay, 1024 otherwise }
RFF dimension	16
activation	relu
optimizer	Adam
learning rate	3×10^{-4}
learning rate schedule	cosine annealing
model training steps	100K

B.2 Elucidated Diffusion Model

For the diffusion sampling process, we use the stochastic SDE sampler of Karras et al. (2022) with the default hyperparameters used for the ImageNet, given in Table 6. We use a higher number of diffusion timesteps at 128 for improved sample fidelity. We use the implementation at `https://github.com/lucidrains/denoising-diffusion-pytorch` which is released under an Apache license.

Table 6: Default ImageNet-64 EDM Hyperparameters.

Parameter	**Value**
no. diffusion steps	128
σ_{min}	0.002
σ_{max}	80
S_{churn}	80
S_{tmin}	0.05
S_{tmax}	50
S_{noise}	1.003

The diffusion model is fast to train, taking approximately 17 minutes for 100K training steps on a standard V100 GPU. It takes approximately 90 seconds to generate 100K samples with 128 diffusion timesteps.

C SynthER Ablations

We consider ablations on the number of generated samples and type of denoiser used for our offline evaluation in Section 4.1.

C.1 Size of Upsampled Dataset

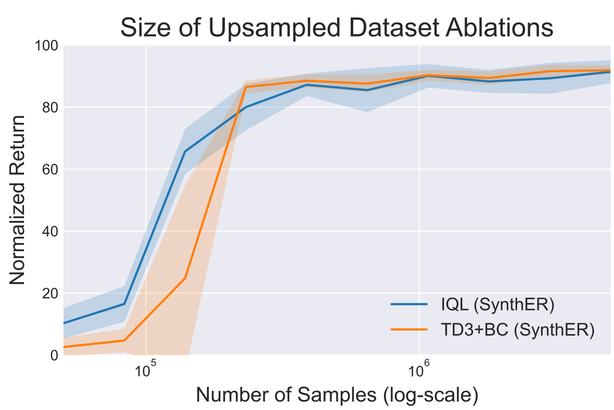


Figure 8: Ablations on the number of samples generated by SynthER for the offline walker medium-replay dataset. We choose 10 levels log-uniformly from the range [50K, 5M]. We find that performance eventually saturates at around 5M samples.

In our main offline evaluation in Section 4.1, we upsample each dataset to 5M. We investigate this choice for the walker medium-replay dataset in Figure 8 and choose 10 levels log-uniformly from the range [50K, 5M]. Similarly to He et al. (2022), we find that performance gains with synthetic data eventually saturate and that 5M is a reasonable heuristic for all our offline datasets.

C.2 Network Ablations

We ablate the hyperparameters of the denoising network, comparing 3 settings of depth from $\{2, 4, 6\}$ and analyze the importance of skip connections. The remaining hyperparameters follow Appendix B.1. We choose the hopper medium-expert dataset as it is a large dataset of 2M. As we can see in Table 7, we see a positive benefit from the increased depth and skip connections which leads to our final choice in Table 5.

Table 7: Ablations on the denoiser network used for SYNTHER on the hopper medium-expert dataset. We observe that greater depth and residual connections are beneficial for downstream offline RL performance. We show the mean and standard deviation of the final performance averaged over 4 seeds.

Network	Depth	Final Performance
MLP	2	86.8±18.7
	4	89.9±17.9
	6	**100.4± 6.9**
Residual MLP	2	78.5±11.3
	4	**99.3±14.7**
	6	**101.1±10.5**

D RL Implementation

For the algorithms in the offline RL evaluation in Section 4.1, we use the 'Clean Offline Reinforcement Learning' (CORL, Tarasov et al. (2022)) codebase. We take the final performance they report for the baseline offline evaluation. Their code can be found at `https://github.com/tinkoff-ai/CORL` and is released under an Apache license.

For the online evaluation, we consider Soft Actor-Critic (SAC, Haarnoja et al. (2018)) and use the implementation from the REDQ (Chen et al., 2021) codebase. This may be found at `https://github.com/watchernyu/REDQ` and is released under an MIT license. We use the 'dmcgym' wrapper for the DeepMind Control Suite (Tunyasuvunakool et al., 2020). This may be found at `https://github.com/ikostrikov/dmcgym` and is released under an MIT license.

D.1 Data Augmentation Hyperparameters

For the data augmentation schemes we visualize in Figure 1a, we define:

1. Additive Noise (Laskin et al., 2020): adding $\epsilon \sim \mathcal{N}(0, 0.1)$ to s_t and s_{t+1}.
2. Multiplicative Noise (Laskin et al., 2020): multiplying s_t and s_{t+1} by single number $\epsilon \sim$ Unif$([0.8, 1.2])$.
3. Dynamics Noise (Ball et al., 2021): multiplying the next state delta $s_{t+1} - s_t$ by $\epsilon \sim$ Unif$([0.5, 1.5])$ so that $s_{t+1} = s_t + \epsilon \cdot (s_{t+1} - s_t)$.

D.2 Online Running Times

Our online implementation in Section 4.2 uses the default training hyperparameters in Appendix B.1 to train the diffusion model every 10K online steps, and generates 1M transitions each time. On the 200K DMC experiments, 'SAC (SynthER)' takes $\approx$ 21.1 hours compared to $\approx$ 22.7 hours with REDQ on a V100 GPU. Both running times are dominated by the use of an update-to-data ratio (UTD) of 20. We expect that this may be heavily sped-up with early stopping on the diffusion training, and leave this to future work. The default SAC algorithm with UTD=1 takes ≈ 2 hours.