ENHANCING OFFLINE-TO-ONLINE REINFORCEMENT LEARNING BY ADAPTIVE EXPERIENCE ALIGNED DIFFUSION SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretraining models on diverse prior data and fine-tuning them on domain-specific tasks is an efficient training paradigm to obtain promising performance on scenarios with limited data or interaction. In the context of reinforcement learning (RL), such a paradigm is named offline-to-online (O2O) RL, where the pretrained agent needs to revise and improve the offline pretrained policy based on its own experience in the online environment. Although prior works in the literature have proven the efficiency of fine-tuning the offline-pretrained agent without offline data, they often require additional designs to overcome the unstable online fine-tuning induced by the discrepancy between the offline and online data. Moreover, existing works demonstrate that introducing offline data when training an online agent from scratch is sample-efficient. Therefore, reusing the knowledge from the offline data properly should be favorable to O2O RL. In this paper, we introduce **A**daptive **D**ata **A**ligned **D**iffusion **S**ampling (AD2S), attempting to accelerate the O2O RL fine-tuning from a perspective of data generation. Our method comprises three key components: distance-based experience alignment, curiosity-driven data prioritization, and data regeneration with amplified guidance. AD2S is a plug-in approach and can be combined with existing methods in the offline-to-online RL setting. By implementing AD2S to off-the-shelf methods, Cal-QL, empirical results indicate improvement in commonly studied datasets.

1 INTRODUCTION

Reinforcement learning (RL) has demonstrated exceptional performance across diverse decision-making and reasoning tasks (DeepSeek-AI et al., 2025; Wang et al., 2018; Zhao et al., 2018; Ling et al., 2024; Lai et al., 2025; Peng et al., 2020; Liu et al., 2025a). However, when implementing the RL paradigm in real-world applications, practitioners often confront a critical challenge: the prohibitive costs and safety risks associated with massive online interaction in safety-critical domains such as autonomous driving or healthcare robotics. This fundamental constraint has catalyzed the development of an efficient learning framework where agents are first pre-trained on comprehensive historical datasets and then fine-tuned on targeted environments — an approach now formally named as offline-to-online (O2O) RL (Liu et al., 2024; Zhang et al., 2024; Zhou et al., 2024).

Nevertheless, when deploying the offline pretrained agent to the online environment, two critical challenges emerge, resulting in unstable online Q-learning: (1) Due to the penalization of out-of-distribution (OOD) actions during offline training, the inherent pessimism of offline-pretrained Q networks to OOD actions often leads to overly conservative policy updates; (2) The non-negligible distributional discrepancy between the offline dataset and the online replay buffer induces catastrophic forgetting and suboptimal convergence. Existing methods replay offline data and introduce specific learning paradigms to address the significant distribution gaps between offline datasets and online samples during offline-to-online fine-tuning. For example, they may consider aligning the policy to be consistent with the behavior policies in both offline and online datasets (Nair et al., 2020), leveraging the capacity of the model ensemble to balance the agent performance and training stability (Lee et al., 2021; Zhao et al., 2022), introducing regularization on Q networks (Zhang et al., 2024), constructing a unified learning paradigm for sequential modeling (Zheng et al., 2022), or introducing policy expansion (Zhang et al., 2023; Uchendu et al., 2023).

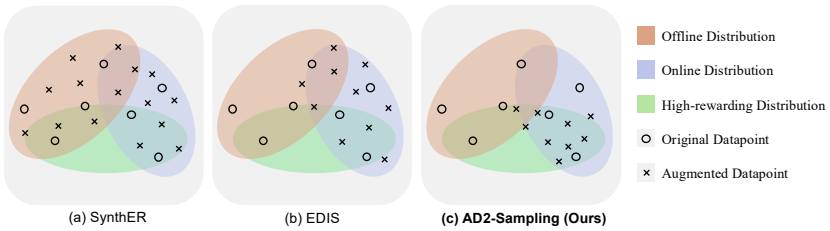

Figure 1: Comparison of previous diffusion-based data generator in O2O RL (Lu et al., 2023c; Liu et al., 2024) and AD2S. AD2S incorporates adaptive data reuse and diffusion-based regeneration, pushing the data towards high-rewarding, under-explored regions for sample-efficient O2O RL.

Recent works have established that agent fine-tuning without offline data replay consistently outperforms the aforementioned fine-tuning methods (Zhou et al., 2024; Liu et al., 2024). However, state-of-the-art solutions typically demand computationally intensive operations, such as high-frequency Q-net updates with model ensembles (Zhou et al., 2024; Zhang et al., 2024) or energy-guided diffusion sampling (Liu et al., 2024). In this paper, we attempt to reuse the key knowledge from offline data and accelerate the online fine-tuning from a data generation perspective. Our key insight is, *Can we adaptively generate synthetic data that is beneficial to O2O RL fine-tuning?*

To verify our insight and accelerate the online fine-tuning phase in O2O RL, we introduce **A**daptive **D**ata **A**ligned **D**iffusion **S**ampling (AD2S or AD^{2}S). Our approach comprises three key components: *distance-based experience alignment*, *curiosity-driven data prioritization*, and *data regeneration with amplified guidance*. Firstly, AD2S aligns offline data that is close to the online experiences, facilitating stable Q-learning through effective dataset reuse. Secondly, AD2S incorporates a curiosity-driven mechanism to assess buffer novelty, adaptively identifying high-novelty transitions (see Figure 1). Thirdly, AD2S utilizes partial noising on the pre-aligned data and conditions the diffusion model to regenerate synthetic data with amplified guidance. These mechanisms enable AD2S to replay the near on-policy, high-novelty experience from the seen data and ensure sufficient online exploration.

Overall, our key contributions are: (1) We introduce AD2S, a simple yet effective framework incorporating a data alignment mechanism and a diffusion model to adaptively generate high-fidelity training data. (2) AD2S replays historical samples based on advantage-weighted relative metrics and regenerates the aligned data towards high-rewarding and under-explored regions for sample-efficient online fine-tuning in O2O RL. (3) Through extensive experiments on popular O2O tasks, empirical results demonstrate that AD2S achieves superior performance compared to previous SOTA methods without any modifications to the backbone algorithms. (4) We assess the synthetic dataset generated by AD2S with data quality metrics, proving its alignment with the objective of O2O RL. These findings validate AD2S as an effective paradigm for accelerating online fine-tuning in O2O RL.

2 Preliminaries

2.1 Reinforcement Learning

Reinforcement Learning (RL) is formulated as a Markov decision process (MDP) described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, consisting of state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, reward function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{R}$, and discount factor $\gamma \in [0, 1)$ (Sutton & Barto, 1998). At each timestep t, the agent selects an action a_t according to the policy π conditioned on the state s_t. Consequently, the agent receives a reward r_t for the action a_t taken in the state s_t, and the environment transforms to the next state $s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t)$. The goal of RL is to learn a policy π^*, which maximizes expected discounted return, $J(\pi) = E_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$. Generally, there are two learning paradigms of RL: online RL, where the agent can learn from interacting with the environment; and offline RL, where the agent can only learn from a fixed dataset $\mathcal{D}^{\text{off}} = \{(s, a, r, s')\}$, which has been collected using an unknown behavior policy π_β.

Offline-to-online (O2O) Reinforcement Learning. O2O RL bridges offline pretraining with online fine-tuning, aiming to leverage historical data to train a near-optimal policy under limited online

interaction. The agent is first pretrained on a fixed dataset $\mathcal{D}^{\text{off}} = \{(s, a, r, s')\}$, then explores in the online environment to recover from suboptimal behaviors and refine its policy.

2.2 Diffusion Models

Score-based diffusion models. Diffusion models (Ho et al., 2020; Karras et al., 2022) are a class of generative models inspired by non-equilibrium thermodynamics. Consider a data distribution $p(\mathbf{x})$ with standard deviation σ_{data}, diffusion models gradually add i.i.d. Gaussian noise of standard deviation σ on the base distribution from time 0 to K and obtain noised distributions $p(\mathbf{x}; \sigma)$. The forward noising process is defined by a sequence of noised distributions following a fixed noise schedule $\sigma_0 = \sigma_{\max} > \sigma_1 > \cdots > \sigma_N = 0$ so that at each noise level, $\mathbf{x}^k \sim p(\mathbf{x}^k; \sigma_k)$. When $\sigma_{\max} \gg \sigma_{\text{data}}$, the final noised distribution $p(\mathbf{x}^K; \sigma_{\max})$ is essentially indistinguishable from random noise. The diffusion model is trained to iteratively denoise samples from a Gaussian distribution and ultimately recover the target distribution, which is formally named the reverse process. Karras et al. (Karras et al., 2022) consider this process as a probability-flow ODE and formulate as below:

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

where $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma k)$ denotes the score function, which points towards the data for a given noise level, and the dot indicates a time derivative. The denoiser $G_\theta(\mathbf{x}_t; \sigma)$ is trained on an L2 denoising minimization objective:

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

and the score can be calculated by $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) = (G_\theta(\mathbf{x}; \sigma) - \mathbf{x})/\sigma^2$. In this paper, we sample data via solving Eq. 1 with the learned denoising network.

Conditional score-based diffusion model. For additional controllability, diffusion models naturally enable conditioning on some signal y (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Classifier-free guidance (CFG) (Ho & Salimans, 2022) is a common post-training technique that further promotes sample fidelity to the condition y in exchange for more complete mode coverage. The guidance distribution $\tilde{p}_\theta$ is interpreted as $\tilde{p}_\theta(\mathbf{x}|y) \propto p_\theta(\mathbf{x}|y) \cdot p_\theta(y|\mathbf{x})^\eta$. Subsequently, considering the equivalence relationship between score matching and the denoising process $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}|y) \propto \epsilon_\theta(\mathbf{x}, y)$ with the implicit classifier $p_\theta(y|\mathbf{x}) \propto p_\theta(\mathbf{x}|y)/p_\theta(\mathbf{x})$, the CFG score $\tilde{\epsilon}_\theta$ can be formed as:

$$\tilde{\epsilon}_\theta(\mathbf{x}^k|y) = (\eta + 1) \cdot \epsilon_\theta(\mathbf{x}^k, y) - \eta \cdot \epsilon_\theta(\mathbf{x}^k, \varnothing), \tag{3}$$

where η is a hyparameter called the *guidance scale*. The training objective of the CFG is to concurrently train the conditional and unconditional score functions as follows, where λ is the dropout rate of condition y:

$$\mathcal{L}(\theta) = \mathbb{E}_{k,\epsilon,\mathbf{x}^0\sim\mathcal{D},y,\lambda\sim\text{Bernoulli}(\lambda)} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}^k, (1-\lambda)y + \lambda\varnothing)\|^2\right]. \tag{4}$$

3 Method

In this section, we introduce Adaptive Data Alignment Diffusion Sampling (AD2S). At its core, AD2S accelerates online fine-tuning in O2O RL through three key mechanisms: (1) distance-based data alignment by reusing near on-policy data from the offline and online samples, (2) curiosity-driven data prioritization from aligned data to enhance online exploration, and (3) amplified condition guided diffusion synthesizer to push the data towards high-rewarding and under-explored regions. We first provide motivation for the AD2S, and concretize how it can be instantiated. Next, we elaborate on the data alignment and generation pipeline. Finally, we present the overall training procedure.

3.1 Motivations

The deployment of the offline pretrained agent in online environments presents two fundamental challenges that hinder effective policy improvement. Firstly, the significant distribution shift between offline datasets and online collected samples induces an unstable Q-learning procedure, leading to catastrophic forgetting of the pretrained Q-function. Moreover, almost all offline pretrained Q-functions are overly pessimistic about OOD actions, as they attempt to penalize these actions during offline training, which creates exploration barriers that prevent effective online fine-tuning.

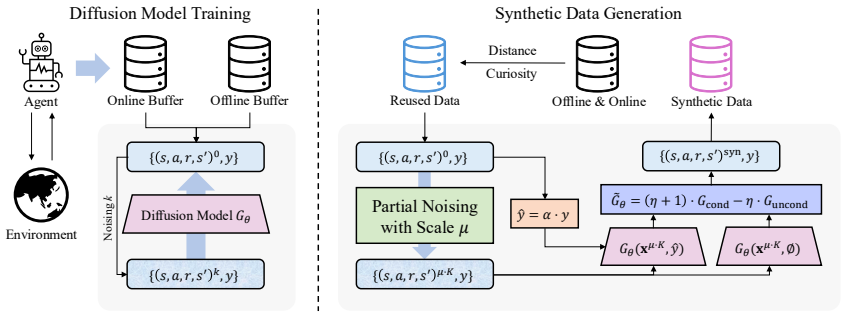

Figure 2: Overall framework of the AD2S. The diffusion model G_θ is trained on seen data $\mathcal{D}^{\text{off}} \cup \mathcal{D}^{\text{on}}$. AD2S conducts key improvements during data generation: (1) adaptively aligning seen data to near on-policy, high-novelty data based on density ratio (DR) alignment and curiosity alignment, (2) partial nosing on aligned data via diffusion forward process, and (3) leverage the diffusion model to regenerate high-rewarding, under-explored synthetic data by amplified condition guidance.

To enable a stable online Q-function fine-tuning, we introduce a distance-based metric (i.e., density ratio) to identify near on-policy samples from the offline data and online experience. The aligned data helps the pretrained Q-function to mitigate the distribution discrepancy and avoid catastrophic forgetting. As for the inherent pessimistic in Q-functions, we introduce a curiosity-driven mechanism to prioritize and select high-novelty data from the distance-based aligned data as the reused data. Reusing the high-novelty data enables the agent to enhance online exploration, thus accelerating the online fine-tuning. Moreover, we introduce a diffusion-based generator to enrich the reused data. Generating the reused data parametrically not only empowers the extrapolation of Q-functions but also interpolates the data distribution to more impoverished, high-rewarding data regions.

Figure 2 gives a brief illustration of AD2S. We first train the conditional diffusion model on offline and online collected samples. The condition y is defined by an advantage-weighed curiosity function. Then, we select the data for generation (named *reused data*) by leveraging a distance and curiosity metric. We use the diffusion model to conditionally regenerate them on amplified condition signals. This enables targeted density of the buffer distribution to high-rewarding, under-explored regions.

3.2 Distance-based Metric for Data Alignment

Adaptively replaying near on-policy experience from offline data can stabilize the Q-learning and significantly enhance sample efficiency in online learning (Liu et al., 2025b; Ball et al., 2023). In O2O RL, we aim to dynamically balance the reuse of offline samples with online experience, therefore mitigating distributional shift while enriching the online replay buffer and preserving policy improvement potential. We use the advantage-weighted priority u to represent the distance between online and offline samples,

$$u = u(s,a,r,s') = w(s,a,r,s') \cdot \exp(\beta \cdot A(s,a)) \tag{5}$$

where $A(s,a)$ is the advantage term, which indicates the potential of the transition for policy improvement, $\beta > 0$ represents a temperature value, and $w(\cdot)$ denotes the density ratio that measures the relative distance of the transition which can be formulated as below,

$$w(s,a,r,s') := d^{\text{on}}(s,a,r,s') / d^{\text{off}}(s,a,r,s') \tag{6}$$

for a given transition (s,a,r,s'), where $d^{\text{on}}(\cdot)$ denotes the transition distribution of online samples in the online buffer $\mathcal{D}^{\text{on}}$ and the $d^{\text{off}}(\cdot)$ represents the offline samples in the offline dataset $\mathcal{D}^{\text{off}}$. This distance metric provides an efficient way to identify near on-policy transitions from online and offline samples. To estimate the proposed density ratio, we approximate $w(\cdot)$ by training a neural network $w_\psi(\cdot)$ and use the variational representation of f-divergences (Nguyen et al., 2007). Consider P and M as probability measures on a measurable space $\mathcal{X}$, with P being absolutely continuous $w.r.t$ M. We define the function $f(y) := y \log \frac{2y}{y+1} + \log \frac{2}{y+1}$. Then we could define the Jensen-Shannon (JS) divergence as $D_{JS}(P\|M) = \int_{\mathcal{X}} f(dP(\mathbf{x})/dM(\mathbf{x}))dM(\mathbf{x})$. Therefore, the density ratio $\frac{dP}{dM}$ could be formed by $w_\psi(\mathbf{x})$ and be estimated by maximizing the lower bound of $D_{JS}(P\|M)$,

$$\mathcal{L}_{\text{DR}}(\psi) = \mathbb{E}_{\mathbf{x}\sim P}\left[f'(w_\psi(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{x}\sim M}\left[f^*(f'(w_\psi(\mathbf{x})))\right], \tag{7}$$

where $w_\psi(\mathbf{x}) \geq 0$ is represented by a neural network, f' is the derivative of f and f^* indicates the convex conjugate of f. We sample from $\mathcal{D}^{\text{on}}$ for $\mathbf{x} \sim P$ and from $\mathcal{D}^{\text{off}}$ for $\mathbf{x} \sim M$.

The traditional way of advantage estimation $A(s,a)$ is to train an advantage function $A(s,a) = Q(s,a) - V(s)$. However, the distribution discrepancy between online and offline samples leads to inaccurate estimation during the early online fine-tuning phase (Zhang et al., 2024; Zhou et al., 2024). To overcome this issue, we introduce the statistical-based relative advantage estimation, which can be formulated as follows,

$$A(s,a) = (r(s,a) - r_{\text{mean}})/r_{\text{std}}, \tag{8}$$

where r_{mean} and r_{std} are calculated from online and offline samples. The relative advantage estimation provides a calibrated and stable advantage estimation, which proves particularly crucial during the early stage of the online fine-tuning phase in O2O RL. In this way, by measuring u proposed in Eq. 5 with parametric density ratio w_ψ, we could construct aligned data $\mathcal{D}^{\text{aligned}}$ for curiosity-driven prioritization and diffusion-based regeneration.

3.3 Curiosity-driven Data Alignment and Diffusion-based Data Generation

Implementing curiosity-driven data generation is an effective way to overcome the inherent pessimism in offline-pretrained Q-functions by enhancing online exploration. In AD2S, we use a forward dynamics model g as the curiosity estimator to construct the reused buffer $\mathcal{D}^{\text{re}}$ from aligned data $\mathcal{D}^{\text{aligned}}$, and a diffusion model to regenerate data towards high-rewarding, under-explored regions. To train the forward dynamics model $g_\phi(s,a)$, we use the data $\mathbf{x}$ sampled from offline and online samples $\mathcal{D}^{\text{off}} \cup \mathcal{D}^{\text{on}}$ and minimize the transition error between the real and predicted next state,

$$e(s,a,s',r) = \|s' - \hat{s}'\|^2 \text{ where } \hat{s}' = g_\phi(s,a). \tag{9}$$

We also integrate the relative advantage metric (Eq. 8) into the error measurement $y(\mathbf{x}) = \exp(\beta \cdot A(s,a)) \cdot e(s,a,r,s')$ to prioritize high potential reward in under-explored regions. We utilize the advantage-weighted metric to perform curiosity estimation for seen samples. Therefore, based on the aforementioned estimations, our framework could adaptively identify near on-policy data with high-novelty and construct the reused data $\mathcal{D}^{\text{re}}$.

For the diffusion model training, we randomly sample data $\mathbf{x}$ from offline and online buffers $\mathcal{D}^{\text{off}} \cup \mathcal{D}^{on}$ and require the diffusion model $G_\theta(\mathbf{x}|y(\mathbf{x}))$ to approximate the conditional distribution $p(\mathrm{x}|y(\mathrm{x}))$, where the condition signal is also defined by the advantage-weighted curiosity: $y(\mathrm{x}) = \exp(\beta \cdot A(s,a)) \cdot e(s,a,r,s')$. Training on $\mathcal{D} = \mathcal{D}^{\text{off}} \cup \mathcal{D}^{on}$ enables G_θ to learn the whole conditioned distribution. Considering the equivalence relationship between score matching and the denoising process described in Section 2.2, The objective for updating parameter θ with the dropout rate λ of condition y is below,

$$\theta^* \leftarrow \arg\min_\theta \mathbb{E}_{k,\epsilon,\mathbf{x}\sim\mathcal{D},\lambda\sim\text{Bernoulli}(\lambda)} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}^k|(1-\lambda)y + \lambda\varnothing)\|^2\right]. \tag{10}$$

To push the aligned data $\mathcal{D}^{\text{aligned}}$ towards high-rewarding, under-explored regions, we add partial noise to these data and use the conditional diffusion model to regenerate them with amplified guidance (Lee et al., 2024; Huang et al., 2024b). Specifically, let $\mathbf{x} = \mathbf{x}^0 \sim \mathcal{D}^{\text{re}}$ denotes the original aligned data, and $k \in [1, K]$ denotes the diffusion timestep. Our method first injects controlled noise into $\mathbf{x}^0$ through a truncated forward process $\mathbf{x}^{\mu \cdot K} \sim \mathcal{N}(\mathbf{x}; \mathbf{x}^0, \sigma(\mu \cdot K)^2 I)$, where the exploration parameter $\mu (0 < \mu = \frac{k}{K} \leq 1)$ governs the noise intensity, trading off between preserving original transition features ($\mu \to 0$) and enabling novel sample generation ($\mu \to 1$). Crucially, we amplify the guidance signal y of the state s during the reverse diffusion process by $\hat{y} = \alpha \cdot y$ where $\alpha > 1$ to enhance the novelty of the regenerated transitions. This denoising process in each step k is formally defined by,

$$\tilde{\epsilon}_\theta(\tilde{\mathbf{x}}^k|y) = (\eta + 1) \cdot \epsilon_\theta(\tilde{\mathbf{x}}^k, \hat{y}) - \eta \cdot \epsilon_\theta(\mathbf{x}^k, \varnothing), \tag{11}$$

where $\tilde{\mathbf{x}}$ is the regenerated samples guided by amplified guidance, η controls the scale of the guidance. This generation mechanism promotes the synthetic data to retain fidelity to task-relevant patterns while extrapolating toward under-explored, high-rewarding regions of the transition space.

3.4 Framework Summary

Finally, we provide a concrete overview of our framework in Algorithm 1. After being pretrained on the offline dataset $\mathcal{D}^{\text{off}}$, the agent interacts with the environment, collecting a stream of real data and

Algorithm 1 Overview of AD2S framework.

Require: synthetic ratio p, density-based alignment ratio p_{DR}, curiosity-driven alignment ratio p_{Curi}, conditional guidance scale η, amplified scale α, offline pretrained agent π
Initialize w_ψ, G_θ, offline buffer D^{off}, online buffer D^{on}, dynamics model g_ϕ
while in *online training phase* **do**
  Collect transitions (s, a, r, s') with π in the environment and add to $\mathcal{D}^{\mathrm{on}}$
  Update w_ψ and g_ϕ using $\mathcal{D}^{\mathrm{on}}$ via Eqs. 7 and 9
  if steps meets G_θ update frequency **then**
    Update G_θ using samples from $\mathcal{D}^{\mathrm{on}} \cup \mathcal{D}^{\mathrm{off}}$ via Eq. 4
    Construct aligned buffer $\mathcal{D}^{\mathrm{re}}$ by calculating u and y using w_ψ, g_ϕ, p_{DR}, and p_{Curi}
    Conditionally generate synthetic data by G_θ with data from $\mathcal{D}^{\mathrm{re}}$ and amplified condition using Eq. 11
    Construct synthetic buffer $\mathcal{D}^{\mathrm{syn}}$ using data generated by G_θ
  Train π on samples from $\mathcal{D}^{\mathrm{on}} \cup \mathcal{D}^{\mathrm{syn}}$ mixed with ratio p
end while

constructing the online replay buffer $\mathcal{D}^{\mathrm{on}}$. We also update the parametric density ratio network w_ψ and forward dynamics model g using samples from $\mathcal{D}^{\mathrm{off}}$ and $\mathcal{D}^{\mathrm{on}}$, via the loss function given by Eqs. 7 and 9. The conditional diffusion model G_θ is trained on the mixed dataset $\mathcal{D}^{\mathrm{on}} \cup \mathcal{D}^{\mathrm{off}}$ using Eq. 10. For the data generation, we first build the $\mathcal{D}^{\mathrm{aligned}}$ by selecting data from offline and online data with the highest metric for $u(s, a, r, s')$ based on ratio p_{DR}, then we construct $\mathcal{D}^{\mathrm{re}}$ by measuring the highest curiosity $y(\mathbf{x})$ on $\mathcal{D}^{\mathrm{aligned}}$ and select data with ratio p_{Curi}. Then we utilize the conditional diffusion model to generate under-explored, high-rewarding synthetic data $\mathcal{D}^{\mathrm{syn}}$ using Eq. 11 with data from $\mathcal{D}^{\mathrm{re}}$ and amplified condition $\hat{y}$. The $\mathcal{D}^{\mathrm{syn}}$ is ultimately used for online fine-tuning.

4 Experiments

In this section, we conduct extensive experiments across commonly studied benchmarks to answer the following questions: (1) How much performance gain does AD2S exhibit across various tasks? (2) What are the underlying mechanisms by which AD2S brings about performance gains? (3) Does AD2S synthesize high-fidelity data?

4.1 Experimental Setup

Datasets and environments. We evaluate the performance of AD2S on three commonly studied benchmarks from the canonical D4RL dataset (Fu et al., 2020), such as MuJoCo Locomotion and Maze2D. These benchmarks help us to validate AD2S under different scenarios. In all tasks, we allow 200K environment interactions for online fine-tuning, which facilitates direct comparison to existing methods (Liu et al., 2024).

Baselines. We compare AD2S with existing augmentation methods based on diffusion models: (1) **SynthER.** Lu et al. (2023c) unconditionally generates synthetic data based on the diffusion model, which can be deployed on both offline and online stages. Here, we directly implement SynthER during the fine-tuning stage for online data generation. (2) **EDIS.** Liu et al. (2024) leverages an energy model to capture the distribution of online data, regarding it as the classifier-guidance of the diffusion model to generate near on-policy data. (3) **PGR.** Wang et al. (2024) considers multiple relevance functions to prioritize the online data, and utilizes the diffusion model to interpolate the replay distribution to more impoverished data regions.

For all D4RL benchmarks, we implement AD2S and baselines on top of base algorithms Cal-QL (Nakamoto et al., 2023), a state-of-the-art O2O method that effectively calibrates over-conservatism of CQL (Kumar et al., 2020). All methods are pretrained on the offline dataset and fine-tuned on the online environment for 0.2M steps. Implementation details are referred to Appendix A.

4.2 Main Results

Our results in Table 1 show that AD2S outperforms existing diffusion-based data synthesizers in various tasks, especially for policies trained on low-quality datasets. Compared to SynthER and PGR, AD2S achieves significant improvement on MuJoCo random datasets, indicating that our proposed

Table 1: Normalized average scores on O2O RL tasks over five random seeds. Here we report the best results for SynthER, PGR and AD2S, and use the results from Liu et al. (2024) for Cal-QL and EDIS. We highlight the best scores in **bold**, and underline the AD2S's score close to the best score.

Dataset	Cal-QL	SynthER	PGR	EDIS	AD2S (Ours)
hopper-random-v2	17.6± 3.1	33.1± 1.7	51.9±37.7	98.1±12.3	**110.7± 4.3**
hopper-medium-replay-v2	102.2± 4.6	108.8± 1.5	102.1± 1.8	**109.9± 0.8**	108.9± 2.2
hopper-medium-v2	97.6± 1.4	106.8± 1.4	108.1± 3.0	105.0± 4.1	**109.0± 3.8**
hopper-medium-expert-v2	107.9± 9.6	111.6± 0.6	111.8± 1.0	109.7± 1.4	**112.1± 1.4**
halfcheetah-random-v2	74.8± 3.2	60.1± 8.0	79.0± 6.8	86.3± 1.8	**90.1± 4.5**
halfcheetah-medium-replay-v2	76.6± 1.2	79.3± 2.0	83.1± 2.1	86.7± 1.4	**92.3± 1.9**
halfcheetah-medium-v2	72.3± 2.1	88.2± 1.8	84.5± 0.9	83.9± 1.0	**90.8± 4.7**
halfcheetah-medium-expert-v2	91.0± 0.6	86.6± 4.2	98.3± 1.3	**98.6± 0.5**	93.9± 1.5
walker2d-random-v2	15.1± 3.5	42.7±29.7	66.0±16.9	61.6±12.6	**99.3±11.6**
walker2d-medium-replay-v2	87.3± 8.5	109.7± 3.7	118.6± 1.1	112.9± 6.4	**121.0± 7.6**
walker2d-medium-v2	84.2± 0.3	108.4± 2.3	114.8± 3.0	103.5± 1.8	**119.4± 1.2**
walker2d-medium-expert-v2	111.1± 0.6	112.6± 1.2	122.2± 8.2	118.5± 4.0	**129.1± 4.6**
locomotion total	937.7	1047.8	1140.6	1174.7	**1276.6**
maze2d-umaze-v1	51.4±17.7	**171.3± 5.2**	157.5± 8.1	162.9± 4.7	169.6± 4.8
maze2d-medium-v1	25.4± 2.2	185.1± 5.5	179.3±15.5	186.4± 5.0	**188.6± 6.5**
maze2d-large-v1	3.9± 7.0	211.1±14.0	163.5±19.8	209.3±30.5	**228.2±15.3**
maze2d total	80.6	567.5	500.3	558.6	**586.4**

Table 2: D4RL normalized scores over five random seeds on 3 tasks, with the highest scores highlighted in **bold**. We conduct experiments on MuJoCo locomotion tasks with low data quality to investigate the effectiveness of different components in AD2S.

Dataset	w/o DA	w/o PN	w/o CG	AD2S (Ours)
hopper-random-v2	20.7± 8.0	27.6± 6.3	105.1±11.9	**110.7± 4.3**
halfcheetah-random-v2	50.0± 4.5	60.3±24.7	86.4± 1.4	**90.1± 4.5**
walker2d-random-v2	20.6± 3.5	68.5±10.6	56.6±31.4	**99.3±11.6**

framework can reduce the impact of low-quality offline datasets on the diffusion-based synthesizer. Meanwhile, our method demonstrates that enriching the high-curiosity data from advantage-weighted near on-policy samples is more sample-efficient than only generating high-curiosity data. Moreover, AD2S obtains further improvement compared to EDIS, indicating the effectiveness of pushing the synthetic data to high-rewarding novel regions. More empirical results are referred to Appendix B.

4.3 Ablation Studies

To verify the effectiveness of each component in AD2S, we conduct ablation studies on walker2d-random-v2, hopper-random-v2, and halfcheetah-random-v2 datasets with the following variants of AD2S: (1) without data alignment (w/o DA), (2) without partial noising (w/o PN) (3) without condition guidance (w/o CG). We report the results in Table 2 and below are key findings: (i) Data alignment can lead the synthesizer to generate near on-policy data, further improving the performance. (ii) When considering amplified condition guidance, partial noising constrains the distance between the synthetic and ground-truth samples, thus acquiring a more sample-efficient online fine-tuning. (iii) Unconditional diffusion sampling encounters performance degradation on low-quality data, demonstrating the effectiveness of our proposed conditional diffusion sampling. More ablation studies are referred to Appendix C.

4.4 Synthetic Data Analysis

To provide an intuition into the efficacy of our proposed method, we follow previous works (Lu et al., 2022; 2023c), using the ground-truth simulator to measure the dynamics distance (i.e., MSE error) between AD2S or EDIS (Liu et al., 2024) with the real next state to verify the transition-level curiosity and validity of the synthetic samples. We also measure the curiosity from the perspective of

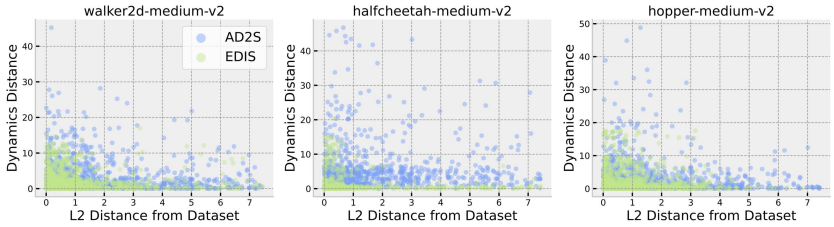

Figure 3: We plot the L2 distance and the dynamic distance under AD2S or EDIS from the online collected data. Compared to EDIS, which is energy model guided diffusion sampling, AD2S can adaptively generate higher novelty data than the existing SOTA method.

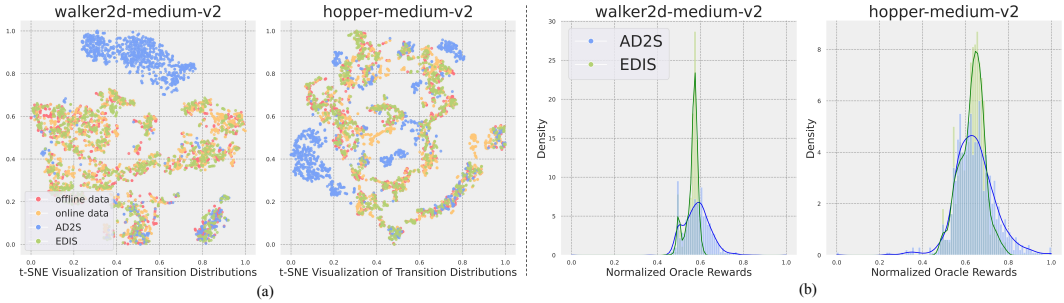

Figure 4: We visualize the transitions of AD2S and EDIS data by t-SNE **(a)**. We also plot the Oracle rewards for them **(b)**. The results demonstrate that AD2S not only generates data with higher curiosity but also pushes the synthetic data towards higher-rewarding regions.

a distance metric by calculating the L2 distance between the synthetic data and the mean of the real datasets. The results are presented in Figure 3. Compared to EDIS, AD2S generates data with a larger L2 distance. Moreover, data generated by AD2S has a larger dynamic distance, which encourages the agent to explore the online environment and refines the knowledge learned from synthetic data (Wang et al., 2024), thus speeding up the fine-tuning process.

To further verify the fidelity of the synthetic samples, we visualize the synthetic data distribution between AD2S and EDIS using t-SNE (van der Maaten & Hinton, 2008) on walker2d-medium and hopper-medium tasks in Figure 4. We also plot the Oracle rewards defined by the simulator. The results demonstrate that AD2S not only generates data with higher curiosity but also pushes the synthetic data towards higher-rewarding regions.

4.5 One-step Advantage Analysis

The traditional advantage $A(s,a) = Q(s,a) - V(s)$ estimates the additional return from taking action a in state s. In contrast, Eq. 8 evaluates transitions relative to the entire buffer, identifying those most critical transitions for training. To validate its efficacy, we approximate the advantage in AD2S using the Q net from the agent itself: $A(s,a) = Q(s,a) - \mathbb{E}[Q(s,\tilde{a})]$, where $\tilde{a}$ denotes sampled random actions. Results on the Walker2d environment (Table 3) demonstrate its effectiveness. We regard that the neural-network-based advantage estimation may introduce instability during early-stage fine-tuning, necessitating additional regularization for Q models, which can lead to training difficulties and increased computational requirements.

5 Related Work

5.1 Offline-to-online RL

Offline-to-online RL methods are developed to bridge the high asymptotic performance in online RL and the low exploration cost in offline RL. The learning process focuses on leveraging the offline

Table 3: D4RL normalized scores over five seeds on the Walker2d task with 4 data qualities. We conduct experiments to validate the efficacy of our short-term advantage proposed in Eq. 8.

Method	random	medium	medium-replay	medium-expert
AD2S	**99.3**±**11.6**	**119.4**± **1.2**	**121.0**± **7.6**	**129.1**± **4.6**
AD2S (variant)	12.6± 5.9	85.1± 1.8	112.5± 7.1	110.0± 1.4

dataset to pre-train an agent to run online RL as sample-efficiently as possible (Lee et al., 2021; Nair et al., 2020; Liu et al., 2025b; Ball et al., 2023; Tarasov et al., 2023a). The commonly studied paradigms utilize offline pretraining followed by a particularly designed fine-tuning phase, such as policy expansion (Zhang et al., 2023; Uchendu et al., 2023), value function calibration (Nakamoto et al., 2023), Q-ensemble techniques (Lee et al., 2021; Wang et al., 2023a), regularization (Zhang et al., 2024), and constraint methods (Nair et al., 2020; Kostrikov et al., 2022; Li et al., 2023). Although retaining offline data during fine-tuning can tackle the over-conservatism of the agent (Fujimoto et al., 2019; Fujimoto & Gu, 2021; Kumar et al., 2020) and prevent catastrophic forgetting, recent works show that fine-tuning the pretrained agent without offline data achieves a better asymptotic performance. Zhou et al. (2024) proposes a simple but effective way to revise the Q function during fine-tuning. Liu et al. (2024) leverages the capacity of the energy model, guiding the diffusion model to generate near on-policy data for sample-efficient fine-tuning. In this paper, we take the advantages of both sides, integrating a weighted density ratio mechanism to select near on-policy data from historical data and leveraging the conditional diffusion model to generate high-fidelity data.

5.2 Diffusion Models as Data Generator in RL

Diffusion models have demonstrated outstanding capabilities in modeling complex distributions (Ho et al., 2020; Saharia et al., 2022; Nichol et al., 2022; Nichol & Dhariwal, 2021; Song et al., 2023). Recent works have employed diffusion models in offline RL for action execution, with extensions to multi-task settings and the alignment of human preferences (Janner et al., 2022; Ajay et al., 2023; Ren et al., 2024; Wang et al., 2023b; Lu et al., 2023a; He et al., 2023b; Jain & Ravanbakhsh, 2024; Mao et al., 2024; He et al., 2023a; Dong et al., 2024). Besides that, another idea is to utilize the capabilities to generate synthetic data in both offline and online RL (Lu et al., 2023c; Lee et al., 2024; Li et al., 2024; Jackson et al., 2024; Liu et al., 2024). GTA (Lee et al., 2024) and TD (Huang et al., 2024a) introduce partial noising on some trajectories, treating the diffusion model as an optimizer to generate high-fidelity trajectories. Moreover, previous works also leverage the capabilities for trajectory stitching (Ghugare et al., 2024), generating the trajectories that do not exist in the dataset (Li et al., 2024; Yang & Wang, 2025; Yuan et al., 2025). Recently, several concurrent works have investigated the potential of learning a world model by diffusion sampling. PolyGRAD (Rigter et al., 2024) and PGD (Jackson et al., 2024) introduce the diffusion model to model the transition function, and embed the policy for classifier-guided trajectory generation. In contrast, DWM (Ding et al., 2024) offers long-horizon predictions in a single forward pass, effectively reducing the compounding error and eliminating the need for recursive queries. In this paper, we follow the generation strategy in (Lu et al., 2023c; Liu et al., 2024; Wang et al., 2024; Lee et al., 2024), focus on adaptively synthesizing data for efficient offline-to-online RL fine-tuning.

6 Conclusion

In this paper, we propose AD2S, a diffusion-based data synthesizer for offline-to-online RL fine-tuning. With the data alignment and amplified guidance, AD2S can reuse high-novelty near on-policy data and enrich the data in high-rewarding regions. As a versatile solution, AD2S seamlessly integrates with prevalent offline-to-online frameworks, with no algorithmic modification. Our extensive experiments on commonly studied benchmarks exhibit considerable performance improvements compared to other diffusion-based data synthesizers. We show that AD2S successfully generates high-quality data from ground-truth datasets, leading to a sample-efficient online fine-tuning.

7 Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation were involved. All datasets used were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

8 Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. The source code has been submitted in the supplementary material to facilitate replication and verification. We leave the detailed description of implementation details in Appendix A.

References

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=sP1fo2K9DFG.

Philip J. Ball, Laura M. Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1577–1594. PMLR, 2023. URL https://proceedings.mlr.press/v202/ball23a.html.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model. *CoRR*, abs/2402.03570, 2024. doi: 10.48550/ARXIV.2402.03570. URL https://doi.org/10.48550/arXiv.2402.03570.

Zibin Dong, Yifu Yuan, Jianye Hao, Fei Ni, Yao Mu, Yan Zheng, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Alignddiff: Aligning diverse human preferences via behavior-customisable

diffusion model. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=bxfKIYfHyx.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020. URL https://arxiv.org/abs/2004.07219.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20132–20145, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a8166da05c5a094f7dc03724b41886e5-Abstract.html.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 2019. URL http://proceedings.mlr.press/v97/fujimoto19a.html.

Raj Ghugare, Matthieu Geist, Glen Berseth, and Benjamin Eysenbach. Closing the gap between TD learning and supervised learning - A generalisation point of view. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=qg5JENs0N4.

Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/ccda3c632cc8590ee60ca5ba226a4c30-Abstract-Conference.html.

Longxiang He, Linrui Zhang, Junbo Tan, and Xueqian Wang. Diffcps: Diffusion model based constrained policy search for offline reinforcement learning. *CoRR*, abs/2310.05333, 2023b. doi: 10.48550/ARXIV.2310.05333. URL https://doi.org/10.48550/arXiv.2310.05333.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Renming Huang, Yunqiang Pei, Guoqing Wang, Yangming Zhang, Yang Yang, Peng Wang, and Hengtao Shen. Diffusion models as optimizers for efficient planning in offline RL. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, volume 15109 of *Lecture Notes in Computer Science*, pp. 1–17. Springer, 2024a. doi: 10.1007/978-3-031-72983-6_1. URL https://doi.org/10.1007/978-3-031-72983-6_1.

Xingshuai Huang, Di Wu, and Benoit Boulet. Goal-conditioned data augmentation for offline reinforcement learning. *CoRR*, abs/2412.20519, 2024b. doi: 10.48550/ARXIV.2412.20519. URL https://doi.org/10.48550/arXiv.2412.20519.

Matthew Thomas Jackson, Michael T. Matthews, Cong Lu, Benjamin Ellis, Shimon Whiteson, and Jakob Nicolaus Foerster. Policy-guided diffusion. *RLJ*, 4:1855–1872, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper233.html.

Vineet Jain and Siamak Ravanbakhsh. Learning to reach goals via diffusion. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3JhmHCVPa8.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. Llmlight: Large language models as traffic signal control agents. In Yizhou Sun, Flavio Chierichetti, Hady W. Lauw, Claudia Perlich, Wee Hyong Tok, and Andrew Tomkins (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, pp. 2335–2346. ACM, 2025. doi: 10.1145/3690624.3709379. URL https://doi.org/10.1145/3690624.3709379.

Jaewoo Lee, Sujin Yun, Taeyoung Yun, and Jinkyoo Park. GTA: generative trajectory augmentation with guidance for offline reinforcement learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/67ea314d1df751bbf99ab664ae3049a5-Abstract-Conference.html.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1702–1712. PMLR, 2021. URL https://proceedings.mlr.press/v164/lee22d.html.

Guanghe Li, Yixiang Shan, Zhengbang Zhu, Ting Long, and Weinan Zhang. Diffstitch: Boosting offline reinforcement learning with diffusion-based trajectory stitching. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=phGHQOKmaU.

Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, and Ya-Qin Zhang. PROTO: iterative policy regularized offline-to-online reinforcement learning. *CoRR*, abs/2305.15669, 2023. doi: 10.48550/ARXIV.2305.15669. URL https://doi.org/10.48550/arXiv.2305.15669.

Haotian Ling, Zhihai Wang, and Jie Wang. Learning to stop cut generation for efficient mixed-integer linear programming. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 20759–20767. AAAI Press, 2024. doi: 10.1609/AAAI.V38I18.30064. URL https://doi.org/10.1609/aaai.v38i18.30064.

Sicong Liu, Yang Shu, Chenjuan Guo, and Bin Yang. Learning generalizable skills from offline multi-task data for multi-agent cooperation. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xu-Hui Liu, Tian-Shuo Liu, Shengyi Jiang, Ruifeng Chen, Zhilong Zhang, Xinwei Chen, and Yang Yu. Energy-guided diffusion sampling for offline-to-online reinforcement learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hunSEjeCPE.

Xuefeng Liu, Hung T. C. Le, Siyu Chen, Rick Stevens, Zhuoran Yang, Matthew R. Walter, and Yuxin Chen. Active advantage-aligned online reinforcement learning with offline data. *CoRR*, abs/2502.07937, 2025b. doi: 10.48550/ARXIV.2502.07937. URL https://doi.org/10.48550/arXiv.2502.07937.

Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22825–22855. PMLR, 2023a. URL https://proceedings.mlr.press/v202/lu23d.html.

Cong Lu, Philip J. Ball, Jack Parker-Holder, Michael A. Osborne, and Stephen J. Roberts. Revisiting design choices in offline model based reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 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. *Trans. Mach. Learn. Res.*, 2023, 2023b. URL https://openreview.net/forum?id=1QqIfGZOWu.

Cong Lu, Philip J. Ball, Yee Whye Teh, and Jack Parker-Holder. Synthetic experience replay. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/911fc798523e7d4c2e9587129fcf88fc-Abstract-Conference.html.

Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. Diffusion-dice: In-sample diffusion guidance for offline reinforcement learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b2fea79b1137d917e8b7cce9434ab5fa-Abstract-Conference.html.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359, 2020. URL https://arxiv.org/abs/2006.09359.

Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline RL pre-training for efficient online fine-tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c44a04289beaf0a7d968a94066a1d696-Abstract-Conference.html.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis (eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1089–1096. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/72da7fd6d1302c0a159f6436d01e9eb0-Abstract.html.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Edward Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. In Marc Toussaint, Antonio Bicchi, and Tucker Hermans (eds.), *Robotics: Science and Systems XVI, Virtual Event / Corvalis, Oregon, USA, July 12-16, 2020*, 2020. doi: 10.15607/RSS.2020.XVI.064. URL https://doi.org/10.15607/RSS.2020.XVI.064.

Allen Z. Ren, Justin Lidard, Lars Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *CoRR*, abs/2409.00588, 2024. doi: 10.48550/ARXIV.2409.00588. URL https://doi.org/10.48550/arXiv.2409.00588.

Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=9CcgO0LhKG.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9(5):1054–1054, 1998. doi: 10.1109/TNN.1998.712192. URL https://doi.org/10.1109/TNN.1998.712192.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/26cce1e512793f2072fd27c391e04652-Abstract-Conference.html.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. CORL: research-oriented deep offline reinforcement learning library. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/62d2cec62b7fd46dd35fa8f2d4aeb52d-Abstract-Datasets_and_Benchmarks.html.

Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, Sergey Levine, and Karol Hausman. Jump-start reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34556–34583. PMLR, 2023. URL https://proceedings.mlr.press/v202/uchendu23a.html.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://api.semanticscholar.org/CorpusID:5855042.

Renhao Wang, Kevin Frans, Pieter Abbeel, Sergey Levine, and Alexei A. Efros. Prioritized generative replay. *CoRR*, abs/2410.18082, 2024. doi: 10.48550/ARXIV.2410.18082. URL https://doi.org/10.48550/arXiv.2410.18082.

Shenzhi Wang, Qisen Yang, Jiawei Gao, Matthieu Gaetan Lin, Hao Chen, Liwei Wu, Ning Jia, Shiji Song, and Gao Huang. Train once, get a family: State-adaptive balances for offline-to-online reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/9318763d049edf9a1f2779b2a59911d3-Abstract-Conference.html.

Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. A reinforcement learning framework for explainable recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pp. 587–596. IEEE Computer Society, 2018.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=AHvFDPi-FA.

Qianlan Yang and Yu-Xiong Wang. RTDiff: Reverse trajectory synthesis via diffusion for offline reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0FK6tzqV76.

Lei Yuan, Yuqi Bian, Lihe Li, Ziqian Zhang, Cong Guan, and Yang Yu. Efficient multi-agent offline coordination via diffusion-based trajectory stitching. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EpnZEzYDUT.

Haichao Zhang, Wei Xu, and Haonan Yu. Policy expansion for bridging offline-to-online reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=-Y34L45JR6z.

Yinmin Zhang, Jie Liu, Chuming Li, Yazhe Niu, Yaodong Yang, Yu Liu, and Wanli Ouyang. A perspective of q-value estimation on offline-to-online reinforcement learning. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 16908–16916. AAAI Press, 2024. doi: 10.1609/AAAI.V38I15.29633. URL https://doi.org/10.1609/aaai.v38i15.29633.

Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 1040–1048. ACM, 2018.

Yi Zhao, Rinu Boney, Alexander Ilin, Juho Kannala, and Joni Pajarinen. Adaptive behavior cloning regularization for stable offline-to-online reinforcement learning. In *30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2022, Bruges, Belgium, October 5-7, 2022*, 2022. doi: 10.14428/ESANN/2022.ES2022-110. URL https://doi.org/10.14428/esann/2022.ES2022-110.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27042–27059. PMLR, 2022. URL https://proceedings.mlr.press/v162/zheng22c.html.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. *CoRR*, abs/2412.07762, 2024. doi: 10.48550/ARXIV.2412.07762. URL https://doi.org/10.48550/arXiv.2412.07762.

A Implementation Details

A.1 Task Description

MuJoCo LocoMotion. MuJoCo locomotion encompasses several standard locomotion tasks commonly utilized in RL research, such as Hopper, Halfcheetah, and Walker2d. In each task, the RL agent controls a robot to achieve forward movement. D4RL (Fu et al., 2020) benchmark provides four qualities of datasets for each task: random-v2, medium-v2, medium-replay-v2, medium-expert-v2.

Maze2D The Maze2D domain is a navigation task requiring a 2D agent to reach a fixed goal location. The tasks are designed to provide a simple test of the ability of offline RL algorithms to stitch together previously collected subtrajectories to find the shortest path to the evaluation goal. The variations of this environment can be initialized with different maze configurations and increasing levels of complexity Three maze layouts are provided: umaze, medium, and large. The task in the environment is for a 2-DoF ball that is force-actuated in the cartesian directions x and y, to reach a target goal in a closed maze.

AntMaze Navigation. Our tests on AntMaze navigation benchmark consist of 4 datasets, namely umaze-diverse-v2, medium-play-v2, medium-diverse-v2, and large-play-v2 from D4RL (Fu et al., 2020). The objective is for an ant to learn how to walk and navigate from the starting point to the destination in a maze environment, with only sparse rewards provided. This task poses a challenge for online RL algorithms to explore high-quality data effectively without access to offline datasets or additional domain knowledge.

Adroit Manipulation. Our empirical evaluation on Adroit manipulation contains 2 domains: pen, door. The RL agent is required to solve dexterous manipulation tasks, including rotating a pen in specific directions, opening a door, and moving a ball, respectively. The offline datasets are clone-v1 datasets in D4RL Fu et al. (2020) benchmark, which only contain a few successful non-Markovian human demonstrations. Therefore, it is pretty difficult for most offline RL approaches to acquire reasonable pre-training performances.

A.2 Implementations and Hyperparameters in AD2S

Our Cal-QL implementation is based on previous work (Liu et al., 2024; Tarasov et al., 2023b), and primarily followed their recommended RL algorithm settings. The code can be found at `https://github.com/tinkoff-ai/CORL` and `https://github.com/liuxhym/EDIS`, which are released under an Apache license. The hyperparameters used in our AD2S's other module are detailed in the Table 4. Our method only introduces two MLP models on top of the diffusion model with corresponding sampling ratio to calculate condition, and we keep the same parameters across different tasks in each domain (e,g, Walker2d, Hopper, Halfcheetah in locomotion domain). On the contrary, EDIS has to train three independent energy models and tune three grad scales for each task. We believe that our method does not require an obvious computational budget and is easy to find the optimal sampling ratio in different tasks.

For all diffusion-based baselines, we use a 6-layer residual MLP as the denoising network. The residual denoising MLP not only provides high-fidelity data generation, but also enables a friendly computational cost in the online fine-tuning stage compared to other popular denoising networks such as U-net (Lee et al., 2024) or transformer (He et al., 2023a). During online fine-tuning, the diffusion synthesizer is retrained on offline and online samples for every $10,000$ environment steps. We also use $5,000$ steps at the start of online fine-tuning to warm up the online replay buffer in AD2S and PGR. For the diffusion sampling process, we follow previous works (Lu et al., 2023c; Wang et al., 2024; Liu et al., 2024), using the stochastic SDE sampler of Karras et al. (Karras et al., 2022) with the same hyperparameter used in EDIS (Liu et al., 2024).

Computation Resources We train AD2S integrated with the base algorithm on an NVIDIA GeForce RTX 3090 GPU and a 32-core CPU.

Table 4: Hyperparameters of AD2S for offline-to-online RL.

Hyperparameter	**Setting**
Network Type (Denoising)	Residual MLP
Denoising Network Depth	6 layers
Denoising Steps	128 steps
Denoising Network Learning Rate	3×10^{-4}
Denoising Network Hidden Dimension	1024 units
Denoising Network Batch Size	256
Denoising Network Activation Function	ReLU
Denoising Network Optimizer	Adam
CFG Scale	2.0
Condition Dropout Rate	0.25
Learning Rate Schedule (Denoising Network)	Cosine Annealing
Training Epochs (Denoising Network)	$50,000$ epochs
Training Interval Environment Step (Denoising Network)	Every $10,000$ steps
Replay Buffer Warm Up Step	$5,000$ steps
Density Ratio Network w_ψ Hidden Dimension	256 units
Density Ratio Network w_ψ Activation Function	ReLU
Dynamics Prediction Network g_ϕ Hidden Dimension	256 units
Dynamics Prediction Network g_ϕ Activation Function	Swish
w_ψ & g Learning Rate	3×10^{-4}
w_ψ & g Optimizer	Adam
Amplified Ratio α	1.2
Partial Noising Scale μ	0.5 in Locomotoin & AntMaze 0.8 in Maze2D
Density-based Prioritized Sampling Ratio p_{DR}	0.5 in Locomotoin & AntMaze 0.8 in Maze2D
Curiosity-based Prioritized Sample Ratio p_{Curi}	0.5 in Locomotoin & AntMaze 0.8 in Maze2D
Advantage Weight for Density Ratio β_{DR}	10
Advantage Weight for Curiosity β_{curi}	10

B Additional Experiments

B.1 Results on Other Environments

To evaluate AD2S in sparse-reward and complex environments, we conduct experiments on the AntMaze navigation and Adroit manipulation benchmarks. Empirical results (Table 5) demonstrate that AD2S consistently outperforms baseline methods in settings with sparse rewards and complex action spaces. However, we observe that all methods exhibit unstable online fine-tuning, which we attribute to the inherent challenges of the AntMaze and Adroit benchmarks.

B.2 Versatility of AD2S

To show the versatility of AD2S, we integrate our method with IQL (Kostrikov et al., 2022) and WSRL (Zhou et al., 2024). Experimental results on the Walker2d environment (Table 6) demonstrate consistent performance improvements. These results validate AD2S's ability to enhance O2O RL across different baseline methods.

Table 5: D4RL normalized scores over five seeds on the antmaze and adroit tasks with the highest scores highlighted in **bold**. We also underline the AD2S's score when it is close to the best score in each task. We conduct experiments to demonstrate the performance of AD2S on sparse rewards and complex environments.

Dataset	Cal-QL	EDIS	AD2S
antmaze-umaze-diverse-v2	93.4± 4.6	95.9± 2.8	**96.8± 3.0**
antmaze-medium-play-v2	86.8± 1.6	93.9± 2.7	**94.4± 5.2**
antmaze-medium-diverse-v2	81.4± 3.9	**89.3± 4.8**	85.0±10.0
antmaze-large-play-v2	42.5± 5.2	66.1± 8.2	**72.5±11.8**
antmaze-large-diverse-v2	42.3± 2.2	57.1± 2.8	**64.0±11.4**
door-clone-v1	-0.3± 0.1	55.8±25.7	**78.0±17.6**
pen-clone-v1	10.7±10.2	81.7±14.9	**93.9± 8.2**

Table 6: D4RL normalized scores over five seeds on the walker2d task with 4 data qualities. We conduct experiments to show the versatility of AD2S by combining it with other backbone algorithms.

Dataset	IQL	IQL + AD2S	WSRL	WSRL + AD2S
walker2d-random-v2	6.5± 0.7	**12.1± 4.1**	65.4±18.2	**82.8±19.8**
walker2d-medium-v2	83.6± 2.0	**98.2± 2.6**	114.3± 4.5	**112.3± 8.7**
walekr2d-medium-replay-v2	83.6± 2.1	**93.6± 4.7**	86.4±12.5	**96.2±10.2**
walker2d-medium-expert-v2	108.9± 2.9	**118.6± 1.3**	118.8± 2.5	**121.5± 1.6**

Table 7: D4RL normalized scores over five seeds on the walker2d task with 4 data qualities. Here, we investigate the sensitivity of the distance alignment ratio p_{DR} in AD2S.

Dataset	$p_{\mathrm{DR}} = 0.1$	$p_{\mathrm{DR}} = 0.3$	$p_{\mathrm{DR}} = 0.5$	$p_{\mathrm{DR}} = 0.7$
walker2d-random-v2	36.8±26.9	91.8± 2.5	**99.3±11.6**	62.4±23.6
walker2d-medium-v2	86.1± 1.9	118.6± 2.8	**119.4± 1.2**	117.5± 1.7
walekr2d-medium-replay-v2	120.3± 3.4	121.3± 3.7	**121.0± 7.6**	119.2± 6.7
walker2d-medium-expert-v2	110.7± 1.4	123.6± 1.7	**129.1± 4.6**	119.8± 3.0

C Additional Ablation Study

C.1 Sensitivity Analysis.

Distance-based alignment ratio. We conduct experiments on the Walker2d task with 4 dataset qualities to perform a sensitivity analysis on p_{DR} in AD2S. We choose p_{DR} from $[0.1, 0.3, 0.5, 0.7]$ and the results are presented in Table 7. As demonstrated in our results, simple grid search on p_{DR} is sufficient for tuning AD2S. The constrained alignment ratio narrows the range of reusable data (e.g., $p_{\mathrm{DR}} = 1$), thereby compromising the diversity of the synthesized distribution.

Curiosity prioritization ratio. We also investigate the choice of p_{Curi} for walker2d task with 4 data qualities in Table 8 and choose 4 levels p_{Curi} from $[0.1, 0.3, 0.5, 0.7]$. The results demonstrate that the proposed AD2S does not require heavy hyperparameter tuning, and performs well reproducibility.

Amplified scale. We provide the sensitivity analysis of α in AD2S on walker2d task with 4 dataset qualities and 5 levels α from $[0.8, 1.0, 1.2, 1.5, 2.0]$. Table 9 reveals that AD2S struggles to synthesize samples whose advantage-weighted data distributions are distant from the ground-truth data, especially on a low-quality dataset. Furthermore, our analysis demonstrates that while moderate conditioning amplification improves performance on medium- and high-quality datasets, a more conservative conditional guidance yields better results for low-quality datasets.

Table 8: D4RL normalized scores over five seeds on the walker2d task with 4 data qualities. Here, we investigate the sensitivity of the curiosity alignment ratio p_{Curi} in AD2S.

Dataset	$p_{\text{Curi}} = 0.1$	$p_{\text{Curi}} = 0.3$	$p_{\text{Curi}} = 0.5$	$p_{\text{Curi}} = 0.7$
walker2d-random-v2	78.9± 7.6	77.4± 1.6	**99.3±11.6**	81.6± 9.8
walker2d-medium-v2	117.7± 4.9	119.4± 3.0	**119.4± 1.2**	117.8± 4.9
walekr2d-medium-replay-v2	119.7± 2.5	119.2± 5.4	**121.0± 7.6**	114.8± 2.5
walker2d-medium-expert-v2	124.2± 1.2	124.7± 4.1	**129.1± 4.6**	123.4± 1.8

Table 9: D4RL normalized scores over five seeds on the walker2d task with 4 data qualities. We conduct experiments to investigate the sensitivity of amplified scale α in AD2S.

Dataset	$\alpha = 1.0$	$\alpha = 1.2$	$\alpha = 1.5$	$\alpha = 2.0$
walker2d-random-v2	89.1±15.6	**99.3±11.6**	90.8± 5.2	74.4± 2.5
walker2d-medium-v2	110.8± 1.2	**119.4± 1.2**	108.5± 5.2	104.9± 3.7
walekr2d-medium-replay-v2	117.5± 5.7	**121.0± 7.6**	119.4± 4.6	117.1± 6.1
walker2d-medium-expert-v2	120.8± 2.5	**129.1± 4.6**	120.5± 3.6	120.6± 4.0

Table 10: D4RL normalized scores over five seeds on the walker2d task with 4 data qualities. We conduct experiments to investigate the sensitivity of the advantage temperature β in AD2S.

Dataset	$\beta = 1$	$\beta = 10$	$\beta = 100$
walker2d-random-v2	85.0±11.7	**99.3±11.6**	16.4± 4.9
walker2d-medium-v2	119.2± 5.8	**119.4± 1.2**	85.1± 1.8
walekr2d-medium-replay-v2	119.7± 8.8	**121.0± 7.6**	72.5±19.8
walker2d-medium-expert-v2	127.9± 1.9	**129.1± 4.6**	113.5± 6.9

Table 11: D4RL normalized scores over five seeds on the walker2d-medium task. We conduct experiments to validate the efficacy of our proposed advantage-weighted alignment.

Dataset	AD2S	DR w/o Adv.	Curiosity w/o Adv.
walker2d-medium-v2	**119.4± 1.2**	103.1± 8.5	106.7± 2.3

Advantage temperature. For the advantage temperature β, we conduct an ablation study across values on the Walker2d task and present the results in Table 10. In AD2S, the role of $A(\cdot, \cdot)$ is to find out transitions in the aligned data that possess high policy improvement potential. Empirical findings indicate $\beta = 10$ delivers optimal performance, and there is little difference between $\beta = 1$ and $\beta = 10$ on *medium*, *medium-expert*, and *medium-replay*, while $\beta = 100$ causes obvious performance degradation. Therefore, our method only needs to ensure that the scaled advantage does not dominate either the density ratio or the curiosity metric. In our experimental settings, we maintain a consistent temperature for all tasks within the same benchmark domain (i.e., the locomotion tasks: Walker2d, Hopper, and HalfCheetah).

C.2 Advantage-weighted Alignment

To illustrate the reason for using the relative advantage metric in both steps, we conduct experiments on the walker2d-medium task with two variants of AD2S: density ratio without advantage (**DR w/o Adv.**) and curiosity without advantage (**Curiosity w/o Adv.**). Results in Table 11 validate the effectiveness of our method. Both the density ratio and the curiosity measurement lack regard for the value of a sample held in the RL environment, and the advantage estimation reflects the potential improvement that the current sample can bring to the policy. Thus, we incorporated the advantage metric in both steps to improve online fine-tuning efficiency.

Table 12: Results on v-d4rl (Lu et al., 2023b) over five seeds on the Walker Walk task with 3 data qualities. We highlight the highest scores in **bold**.

Dataset	DrQ-BC	SynthER	PGR	AD2S
random	459.7±29.0	484.2±52.6	470.8±47.0	**510.1±28.5**
medium	490.6±18.5	494.1±50.9	530.9±23.1	**527.0±38.7**
medium-replay	488.7±32.7	490.2±56.7	503.0±5.3	**508.7±14.1**

C.3 Experiments on Visual Environment

We conduct experiments on pixel-based environments using the v-d4rl benchmark (Lu et al., 2023b), following the data generation paradigm proposed by Lu et al. (2023c). Firstly, we pretrain the DrQ-BC for 1M steps on the offline dataset. Then we freeze the image encoder, generate latent observations using diffusion models, and fine-tune the policy and Q-network in the online environment for 200k steps. We use the same architecture and hyperparameters as used in the D4RL locomotion benchmark, other than changing the partial noising scale μ to 0.1, amplified ratio α to 2.0, and CFG scale to 1.0. Table 12 shows the results on the Walker Walker task with 3 datasets. In future work, we will explore even more complex environments to refine our method and validate its generalization.

C.4 Synthetic Data Analysis

To verify the validity of the synthetic samples generated by AD2S, we use the ground-truth simulator to measure the dynamics distance (i.e., MSE error) between AD2S or other diffusion-based baselines with the real next state. Empirical results are presented in Figures 5. These measurements provide insight into the efficacy of our proposed method.

While AD2S incurs a higher dynamic distance due to its curiosity-driven data prioritization, it synthesizes data further than all baseline methods in the distance metric. This capability helps mitigate the inherent pessimism in offline-pretrained Q-functions and improves online exploration. This demonstrates that AD2S is better at pushing synthetic data towards regions with higher novelty.

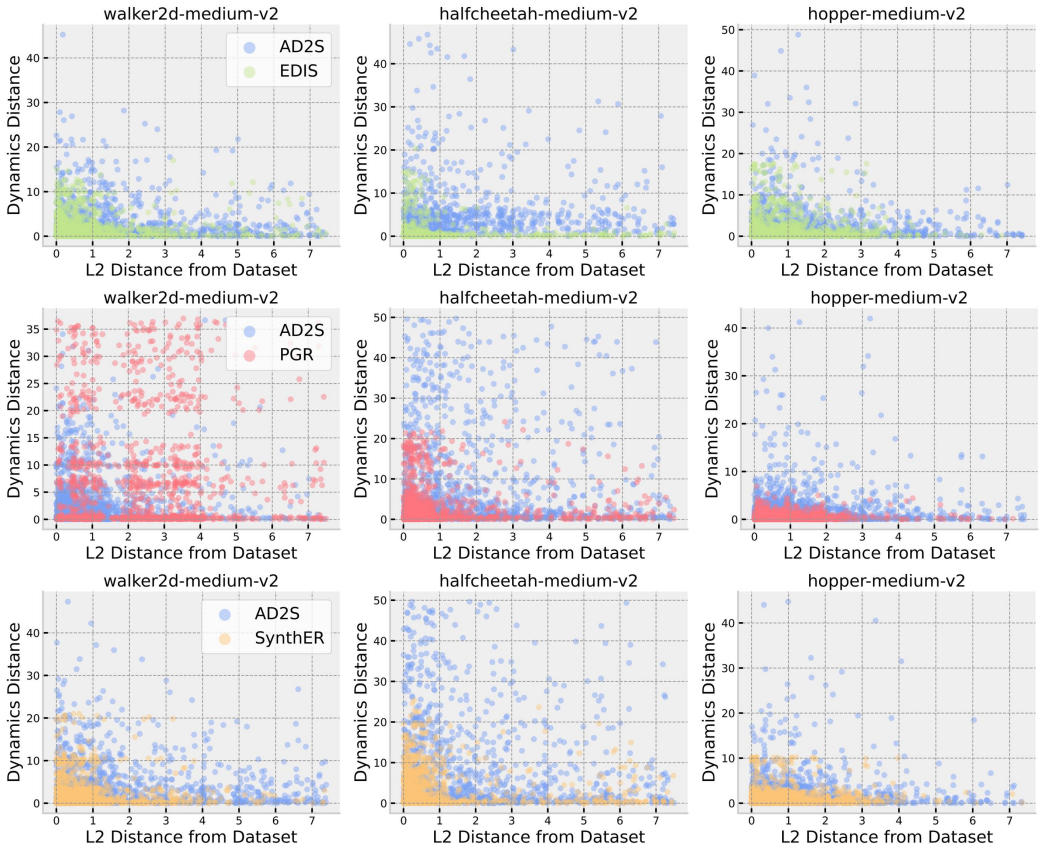

Figure 5: We plot L2 distance from online collected data, and dynamic distance under AD2S or diffusion-based baselines. **Top**: EDIS, **Middle**: PGR, and **Bottom**: SynthER. AD2S can adaptively generate higher novelty data than the existing SOTA method This indicates that our method can adaptively generate higher novelty data than the existing SOTA method.