

Boosting Continuous Control with Consistency Policy

Yuhui Chen

Institute of Automation, Chinese
Academy of Sciences
BEIJING, CHINA
School of Artificial Intelligence,
University of Chinese Academy of
Sciences
BEIJING, CHINA
chenyuhui2022@ia.ac.cn

Haoran Li

Institute of Automation, Chinese
Academy of Sciences
BEIJING, CHINA
School of Artificial Intelligence,
University of Chinese Academy of
Sciences
BEIJING, CHINA
lihaoran2015@ia.ac.cn

Dongbin Zhao

Institute of Automation, Chinese
Academy of Sciences
BEIJING, CHINA
School of Artificial Intelligence,
University of Chinese Academy of
Sciences
BEIJING, CHINA
dongbin.zhao@ia.ac.cn

ABSTRACT

Due to its training stability and strong expression, the diffusion model has attracted considerable attention in offline reinforcement learning. However, several challenges have also come with it: 1) The demand for a large number of diffusion steps makes the diffusion-model-based methods time inefficient and limits their applications in real-time control; 2) How to achieve policy improvement with accurate guidance for diffusion model-based policy is still an open problem. Inspired by the consistency model, we propose a novel time-efficiency method named Consistency Policy with Q-Learning (CPQL), which derives action from noise by a single step. By establishing a mapping from the reverse diffusion trajectories to the desired policy, we simultaneously address the issues of time efficiency and inaccurate guidance when updating diffusion model-based policy with the learned Q-function. We demonstrate that CPQL can achieve policy improvement with accurate guidance for offline reinforcement learning, and can be seamlessly extended for online RL tasks. Experimental results indicate that CPQL achieves new state-of-the-art performance on 11 offline and 21 online tasks, significantly improving inference speed by nearly 45 times compared to Diffusion-QL. Code is available at <https://github.com/cccedric/cpql>.

1 INTRODUCTION

Offline reinforcement learning (RL) offers a promising approach for learning policies from pre-collected datasets to solve sequential decision-making tasks, but it requires conservative behaviors to alleviate value overestimation [11, 29]. The policy representation needs to be powerful enough to cover the diverse behaviors, thereby alleviating the value function overestimation caused by querying out-of-distribution (OOD) actions [14, 30]. Traditional unimodal policy struggles to model multi-modal behaviors in datasets, leading to the failure of policy constraints [7, 37, 49].

The diffusion model [21] showcases remarkable attributes in terms of high training stability and its ability to provide strong distributional representations, achieving impressive outcomes in the generation of high-quality image samples endowed with diverse features. In light of the imperative demand for strong expression ability, the diffusion model is introduced to offline RL for imitating behaviors [8, 31, 48], trajectories modeling [3, 22, 40], and building expected policy [34, 49, 54]. Since the diffusion model can cover diverse behaviors [53] and enhance the KL divergence constraint between the behavior policy and the expected policy [49], these

methods achieve impressive performance by alleviating value overestimation with reducing OOD actions.

However, the adoption of the diffusion model also presents two prominent challenges. Firstly, the training and inference of the diffusion-model-based policy is time inefficient. This is because the inherent sampling process of the diffusion model relies on a Markov chain for a large number of steps (e.g., 1,000 steps) to capture intricate dependencies in the data. This inadequacy translates into protracted training durations and sluggish real-time inference capabilities. The ramifications of this time inefficiency severely hinders practical utility in real-time decision-making domains such as robot control. Secondly, it is hard to improve the parameterized policy by diffusion model accurately under the actor-critic framework. Due to the absence of data from the optimal behavior, updating diffusion model-based policy needs additional guidance to achieve better behavior. Though the Q-value is employed to guide the reverse diffusion process [23, 49], it is theoretically impossible to access the desired policy since the Q-value is inaccurate for intermediate diffusion actions. How to achieve accurate guidance for the expected policy is still an open problem [34].

Instead of a large number of reverse diffusion steps, the consistency model [43] is based on probability flow ordinary differential equation (ODE) and achieves the one-step generation process. Inspired by the consistency model, we propose a novel approach called Consistency Policy with Q-learning (CPQL) that can generate action directly from noise in a single step. Due to this one-step generation, CPQL significantly outperforms previous diffusion-model-based methods in training and inference speed. By establishing a mapping from the reverse diffusion trajectories to the desired policy, CPQL avoids explicit access to inaccurate guidance functions in multi-step diffusion processes, and we theoretically prove that it can achieve policy improvement with accurate guidance for offline RL. Moreover, with the inherent sampling randomness of the stochastic generation process, CPQL can be seamlessly extended for online RL tasks without relying on additional exploration strategies.

In summary, our contribution involves three main aspects. Firstly, we propose a novel time-efficient method named CPQL and improve training and inference speeds by nearly 15x and 45x compared to Diffusion-QL [49], while also improving the performance. Secondly, we conduct a theoretical analysis to demonstrate that the CPQL is capable of achieving policy improvement with accurate guidance and propose an empirical loss to replace consistency loss for stabilizing consistency policy training. Finally, CPQL can seamlessly extend to online RL tasks. As experimented, CPQL achieves

state-of-the-art (SOTA) performance on 11 offline and 21 online tasks.

2 RELATED WORK

Offline RL. The main problem offline RL faces is the overestimation of the value function caused by accessing OOD actions. The previous offline RL methods broadly are categorized as policy constraints [13, 14, 50], value function regularization [28, 36], or in sample learning [26, 35, 51, 52]. Our work belongs to policy constraints. Different from the previous method based on unimodal distribution, we propose a sampling efficient policy based on the consistency model, which is better suited to constrain the policy to meet the multi-modal characteristics of the offline dataset.

Diffusion Models for Imitation Learning. When expert trajectories are accessible, imitation learning is a powerful method for building the expected policy. The diffusion model is employed to tackle diverse behaviors in the expert dataset [9, 31, 37]. When the reward for the typical task is sparse or inaccessible, goal-conditioned imitation learning is an alternative solution, and the diffusion model can be used to build the goal-conditioned policy [41]. Moreover, since the diffusion model has a strong text-to-image ability, it is also used to generate the behavior goal with the language as the input [16, 24].

Diffusion Models for RL. The diffusion model has been widely used in RL, especially offline RL tasks. The diffusion model can be used to model trajectories [3, 22], build the world model [6], augment the dataset [33] and estimate the action distribution conditioned on the state. This distribution can be either a behavioral policy distribution in the dataset [7, 20, 37] or an expected policy distribution [23, 34, 49]. Diffusion-QL [49] and EDP [23] are similar to our work. There are several differences between our work and these methods. First and foremost, they did not clearly point out what distribution the diffusion model fits, while our work starts with modeling the solution of the constrained policy search problem and proposes the consistency policy. Second, these methods are damaged by inaccurate guiding during the reverse diffusion process. Finally, a large number of reverse diffusion steps is necessary for these methods and harm real-time for robot control. QGPO [34] also focuses on solving the problem of diffusion model fitting the optimal solution. Our work is different in that we solve this problem from the perspective of establishing ODE trajectories mapping rather than the perspective of accurate guidance function during the diffusion process. In addition to offline tasks, DIPO [53] is first proposed to use the diffusion model to solve online RL problems. This work proposes the action gradient to update the actions in the replay buffer and uses the diffusion model to fit the updated action distribution. Our approach directly updates the consistency policy using the gradient of the value function.

3 DIFFUSION POLICY FOR OFFLINE RL

3.1 Offline RL

A decision-making problem in RL is usually represented by a Markov Decision Process (MDP), defined as a tuple: $\{S, A, P, r, \gamma\}$. S and A represent the state and action spaces, respectively. $P(s_{t+1}|s_t, a_t)$ represents the transition probability from state s_t to next state s_{t+1} after

taking the action a_t , and $r(s_t, a_t, s_{t+1})$ represents the corresponding reward. γ is the discount factor. A policy $\pi(a_t|s_t)$ describes how an agent interacts with the environment. And we use subscripts $t \in \{1, \dots, T\}$ to denote trajectory timesteps.

Policy Constraint for Offline Learning. Without interactions with the environment, agents under the offline RL scenario learn an expected policy entirely from a previously collected static dataset of transitions denoted as $\mathcal{D}_{offline} = \{(s_t, a_t, r_t, s_{t+1})\}$ by a behavior policy $\mu(a_t|s_t)$. In order to alleviate the value overestimation caused by visiting OOD actions [27], offline RL methods [30] often constrain the learned policy to the behavior policy μ . The goal is to find the desired policy around behavior policy that can achieve the following objective:

$$\begin{aligned} \mathcal{J}(\pi) = & \mathbb{E}_{s_t \sim \mathcal{D}_{offline}} [\mathbb{E}_{a_t \sim \pi(\cdot|s_t)} [Q_v(s_t, a_t)] \\ & - \lambda D_{KL}(\pi(\cdot|s_t) || \mu(\cdot|s_t))] \end{aligned} \quad (1)$$

where $Q_v(s_t, a_t)$ is a learned Q-function of the current policy π , and λ determines the relative importance of the KL divergence against the value function, controlling the balance between exploitation and policy constraints.

Policy Improvement with Constraint Policy Search. In the following sections, we omit subscripts t to simplify the demonstrations if they are unnecessary. With the optimization objective claimed in Eq. (1), we have the closed form of the solution π^* [7, 38, 39]:

$$\pi^*(a|s) \propto \mu(a|s) \exp\left(\frac{1}{\lambda} Q_v(s, a)\right) \quad (2)$$

The aforementioned solution is frequently employed to tackle offline RL problems. These methods typically explicitly estimate the probability distribution of the behavioral policy to compute the analytical form of Eq. (2). However, estimating the probability distribution of the behavioral policy becomes exceedingly challenging when the policy is complex. Therefore, diffusion models are often utilized to model this behavioral policy, which is straightforward due to the availability of a dataset for the behavioral policy. Consequently, the aforementioned diffusion model-based approaches are incapable of estimating the analytical form of Eq. (2), necessitating additional importance sampling for the computation of the policy corresponding to Eq. (2) [7, 17, 20]. Furthermore, in these methods, it is usually possible to only estimate the Q-value of the behavioral policy; thus, such methods are also referred to as one-step bootstrapping. In this paper, we will employ a diffusion model to directly model π^* , thereby avoiding reliance on Q-value and resampling techniques during the inference process, and achieving improved performance by implementing multi-step bootstrapping.

3.2 Diffusion Policy

In this section, we define the expected policy of Eq.(2) as diffusion policy and describe how to use the diffusion model to build it. Before formally introducing this process, we first declare that there are two different types of timesteps in the following sections. Rather than using subscripts t denoting the trajectory timesteps, we use superscripts $k \in [0, K]$ to denote the diffusion timesteps.

The diffusion model starts by diffusing $p_{data}(a)$ with a stochastic differential equation (SDE):

$$da^k = \mu(a^k, k) dk + \sigma(k) dw^k \quad (3)$$

where $\mu(\cdot, \cdot)$ and $\sigma(\cdot)$ are the drift and diffusion coefficients respectively, and $\{w^k\}_{k \in [0, K]}$ denotes the standard Brownian motion. Starting from x^K , the diffusion model aims to recover the original data x^0 by solving a reverse process from K to 0 with the Probability Flow (PF) ODE [44]:

$$da^k = [\mu(a^k, k) - \frac{1}{2}\sigma(k)^2 \nabla \log p^k(a^k)]dk \quad (4)$$

where the only unknown term is the score function $\nabla \log p^k(a^k)$ at each timestep. Thus, the diffusion model trains a neural network parameterized by ϕ to estimate the score function: $s_\phi(a^k, k) \approx \nabla \log p^k(a^k)$. By setting $\mu(a^k, k) = 0$ and $\sigma(k) = \sqrt{2}$, we can obtain an empirical estimate of the PF ODE:

$$\frac{da^k}{dk} = -ks_\phi(a^k, k) \quad (5)$$

When the diffusion model is used to represent the expected policy directly, the challenge lies in estimating the score function because it's impossible to obtain data from the optimal policy straightforwardly, making the estimation intractable. Nevertheless, we can use the diffusion process through reference policy with guidance to estimate the score function [10]. With Eq. (2), we can derive the following score function with guidance for the solution of Eq. (2):

$$\nabla \log \pi^*(a^k|s) = \nabla \log \mu(a^k|s) + \frac{1}{\lambda} \nabla Q_v(s, a^k) \quad (6)$$

Note that we have to calculate the Q-function for each diffusion timestep. However, we only have the learned Q-value at diffusion timestep 0, and it is challenging to accurately estimate the Q-value for each diffusion step. Consequently, achieving the correct policy through accurate diffusion guidance is nontrivial [34].

4 CONSISTENCY POLICY WITH Q-LEARNING

4.1 Consistency Policy

The previous section notes that modeling the solution using the classifier-guided diffusion process is quite challenging due to the inaccuracy of guidance. In this section, we will show how the consistency policy can avoid this problem and achieve policy improvement with accurate guidance. Drawing inspiration from the consistency model, we could directly map the ODE trajectories to the policy in the inverse diffusion process. Following the consistency model [43], we define the consistency policy:

$$\begin{aligned} \pi_\theta(a|s) &\equiv f_\theta(a^k, k|s) \\ &= c_{skip}(k)a^k + c_{out}(k)F_\theta(a^k, k|s) \end{aligned} \quad (7)$$

where $a^k \sim \mathcal{N}(0, kI)$. The $F_\theta(a^k, k|s)$ is a trainable network that takes the state s_t as an condition and outputs an action of the same dimensionality as the input a^k . c_{skip} and c_{out} are differentiable functions such that $c_{skip}(\epsilon) = 1$, and $c_{out}(\epsilon) = 0$ to ensure the consistency policy is differentiable at $k = \epsilon$ if $F_\theta(a^k, k|s)$ is differentiable, which is critical for training process described later. We stop solving the reverse process at $k = \epsilon$, where ϵ is a small positive constant to avoid numerical instability. The sampled action $\hat{a}^\epsilon \sim \pi_\theta(a|s)$ of the consistency policy is used for controlling the agent. The relationship between the forward diffusion process represented by the ODE trajectories and consistency policy is shown in Figure 1.

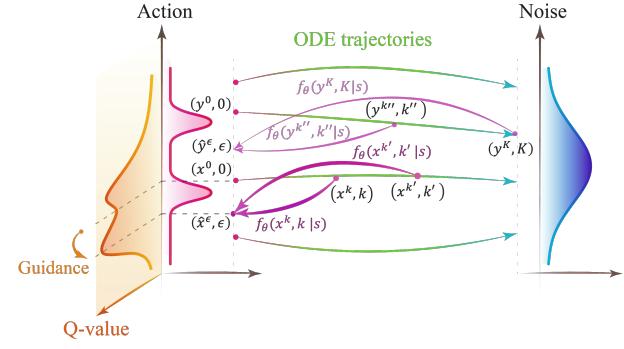


Figure 1: Given an ODE that smoothly converts from actions of the reference policy (e.g., $x^0, y^0 \in A$) to Gaussian noises, the consistency policy f_θ maps any point (e.g., $x^k, x^{k'}, y^{k''}, y^K$) on the PF ODE trajectory to the desired actions (e.g., $\hat{x}^\epsilon, \hat{y}^\epsilon$). Since consistency policy generates the actions from the noise by one step, it reduces an enormous amount of time for policy training and inference.

4.2 Training Loss for Consistency Policy

Training consistency policy with consistency loss from the consistency model is nontrivial. Let us begin by assuming the presence of a pre-trained diffusion model with a score function, denoted as $s_{\phi^*}(a^k, k)$ with parameter ϕ^* , representing the optimal diffusion-model-based policy. Following the consistency model, we can use a consistency policy to distill the inverse diffusion process. However, since we cannot access the data from the optimal policy, the assumptive pre-trained diffusion model is unavailable.

To tackle the above problem, we propose CPQL, which establishes a relationship between the above distillation process and the consistency training for reference policy with the Q-function based on Theorem 1, with full proof in Appendix A.1. In order to determine the solution trajectory of action $\{a_t^k\}_{k \in [\epsilon, K]}$, we discretize the diffusion horizon $[\epsilon, K]$ into M sub-intervals with boundaries $k_1 = \epsilon < k_2 < \dots < k_M = K$. Note that here we use subscripts $m \in \{1, \dots, M\}$ to denote time sub-intervals. To learn the consistency policy, we minimize the objective with stochastic gradient descent on the parameter θ , while updating θ^- with exponential moving average.

THEOREM 1. Let $\Delta k = \max_{m \in [1, M-1]} \{|k_{m+1} - k_m|\}$. We have the assumptions: 1) Distance function d , value function Q and f_{θ^-} are all twice continuously differentiable with bounded second derivatives; 2) There is a pre-trained score function representing the desired policy: $\forall k \in [\epsilon, K] : s_{\phi^*}(a^k, k) = \nabla \log p^k(a^k)$, which cannot be accessed; 3) f_θ satisfies the Lipschitz condition: there exists $L > 0$ such that for all $k \in [\epsilon, K]$ and $x, y \in A$, we have $\|f_\theta(x, k) - f_\theta(y, k)\|_2 \leq L\|x - y\|_2$. The distillation loss for the distillation process of training the consistency policy is defined as:

$$LCD(\theta, \theta^-; \phi^*) = \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(\hat{a}_{\phi^*}^{k_m}, k_m|s))] \quad (8)$$

where $\hat{a}_{\phi^*}^{k_m}$ is calculated with Euler solver and the optimal score function $s_{\phi^*}(a^k, k)$. The training loss for training the consistency policy is

sufficient to replace the distillation loss:

$$L(\theta, \theta^-) = \alpha L_{CT}(\theta, \theta^-) - \beta \mathbb{E}[Q(s, \hat{a}^\epsilon)] + o(\Delta k) \quad (9)$$

where the consistency loss for consistency training is defined as:

$$L_{CT}(\theta, \theta^-) = \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a + k_m z, k_m|s))] \quad (10)$$

and

$$\begin{aligned} \alpha &= \mathbb{E}[1 + Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}] \\ \beta &= \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a + k_{m+1}z, k_{m+1}|s))] \\ &\quad * \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2})] \end{aligned} \quad (11)$$

The expectation is taken with respect to $(s, a) \sim \mathcal{D}_{offline}$, $\hat{a}^\epsilon \sim \pi_\theta$, $m \sim \mathcal{U}[1, M-1]$, $z \sim \mathcal{N}(0, I)$, and d stands for the Euclidean distance $d(x, y) = \|x - y\|_2^2$. Theorem 1 demonstrates that we can improve policy by distilling the assumed optimal score function using behavioral policy data and the Q-function. Eq. (9) provides the loss function for training the consistency policy. $L_{CT}(\theta, \theta^-)$ represents the consistent training loss regarding the behavior policy. Unlike consistency models, this part includes α related to the Q-value as the weight to improve the probability of good actions sampled by the policy. The second term is the Q-value of the estimated action. Its purpose is to guide the policy to generate higher Q-value, leading to better performance. As seen from the above loss function, estimating the Q-value in the intermediate diffusion steps is no longer necessary, which avoids inaccurate guidance.

As experienced, we find that the loss function mentioned above may cause instability during the training process, resulting in the deterioration of policy performance. We speculate that the reason may be that consistency loss needs to sufficiently sample the diffusion time interval to achieve consistency with the original data. However, in practice, the limited sampling steps and the randomness of sampling may not meet this condition. In order to achieve more stable training, we propose a simpler loss function called reconstruction loss:

$$L_{RC}(\theta, \theta^-) = \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), a)] \quad (12)$$

This loss is intuitive. It drives the consistency policy to directly recover the original action, rather than indirectly achieving the recovery by maintaining consistency. We prove that this loss function has the same convergence objective with $L_{CT}(\theta, \theta^-)$ in Appendix A.2. Therefore, we have the final optimization objective:

$$\begin{aligned} L(\theta, \theta^-) &= \alpha L_{RC}(\theta, \theta^-) \\ &\quad - \frac{\eta}{\mathbb{E}_{(s,a) \sim \mathcal{D}_{offline}}[Q(s, a)]} \mathbb{E}_{s \sim \mathcal{D}_{offline}, \hat{a}^\epsilon \sim \pi_\theta}[Q(s, \hat{a}^\epsilon)] \end{aligned} \quad (13)$$

where $\frac{\eta}{\mathbb{E}_{(s,a) \sim \mathcal{D}_{offline}}[Q(s, a)]}$ corresponds to the β in Eq. (11). In practice, we set α as an adjustable parameter ignoring the Q value in α , and η is another adjustable parameter. The (α, η) are set according to the characteristics of different domains.

4.3 Policy Evaluation

For policy evaluation, the objective function uses the KL divergence to constrain the learned policy from accessing the OOD action. The

Algorithm 1 CPQL and CPIQL for Offline RL Tasks

```

Initialize the policy network  $\pi_\theta$ , critic networks  $V_\psi$ ,  $Q_{v_1}$  and  $Q_{v_2}$ , and target networks  $\pi_{\theta^-}$ ,  $Q_{v_1^-}$  and  $Q_{v_2^-}$ 
for each iteration do
    Sample transition batch  $B = \{(s_t, a_t, r_t, s_{t+1})\} \sim \mathcal{D}_{offline}$ 
    # Q-learning
    Update  $Q_{v_i}$  by Eq. (14) (CPQL) or Update  $V_\psi$ ,  $Q_{v_i}$  by Eq. (15) and Eq. (16) (CPIQL)
    # Consistency policy learning
    Sample  $\hat{a}_t^\epsilon \sim \pi_\theta(a_t | s_t)$  by Eq. (7)
    Update  $\pi_\theta$  by minimizing Eq. (13)
    # Update target networks
     $\theta^- = \rho\theta + (1 - \rho)\theta$ 
     $v_i^- = \rho v_i^- + (1 - \rho)v_i$  for  $i \in \{1, 2\}$ 
end for

```

Q-function of CPQL is learned in a conventional way, with the Bellman operator [32] and the double Q-learning trick [47]. We build two Q-networks Q_{v_1} and Q_{v_2} and target networks $Q_{v_1^-}$ and $Q_{v_2^-}$ and minimizing the objective:

$$\begin{aligned} L_Q(v_i) = & \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \mathcal{D}_{offline}, a_{t+1}^\epsilon \sim \pi_\theta} [|r(s_t, a_t) \\ & + \gamma \min_{j=1,2} Q_{v_j^-}(s_{t+1}, a_{t+1}^\epsilon) - Q_{v_i}(s_t, a_t)|^2] \end{aligned} \quad (14)$$

To evaluate the effectiveness of the consistency policy in constraining the sampling of OOD actions that cause overestimation of the Q-value, we further propose Consistency Policy Implicit Q-Learning (CPIQL) as a comparison. Using implicit Q-learning [26] can better restrict accessing the OOD action. Meanwhile, implicit Q-learning avoids sampling the action at the next timestep when calculating the critic loss function, reducing the time needed for critic training. CPIQL builds the value-network V_ψ in addition. Typically, the value function Q and V are given by:

$$L_V(\psi) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}_{offline}} [L_2^\tau(\min_{j=1,2} Q_{v_j^-}(s_t, a_t) - V_\psi(s_t))] \quad (15)$$

$$\begin{aligned} L_Q(v_i) = & \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \mathcal{D}_{offline}} [|r(s, a) + \gamma V_\psi(s_{t+1}) - Q_{v_i}(s_t, a_t)|_2^2], \\ & \text{for } i = 1, 2 \end{aligned} \quad (16)$$

where $L_2^\tau(u) = |\tau - \mathbb{1}(u < 0)|u^2$ is expectile regression function.

In contrast to other RL methods based on the diffusion model, the consistency policy affords the distinct advantage of facilitating one-step sampling. This feature significantly reduces the time costs associated with both the sampling process and the computation of Q-function gradient back-propagation, resulting in better time efficiency. We summarize the complete algorithm procedure of CPQL and CPIQL in Algorithm 1 for offline tasks.

4.4 Extension for Online RL

Previous sessions have demonstrated how CPQL learns the policy of Eq. (2) to solve offline RL problems. In this session, we will demonstrate that the CPQL method can seamlessly extend to online RL problems. We define the objective for online tasks as:

$$\begin{aligned} \mathcal{J}(\pi) = & \mathbb{E}_{s_t \sim \mathcal{D}_r} [\mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q_v(s_t, a_t)] \\ & - \lambda D_{KL}(\pi(\cdot | s_t) || \pi_r(\cdot | s_t))] \end{aligned} \quad (17)$$

Table 1: The performance of CPQL and CPIQL and SOTA baselines on D4RL Locomotion and Adroit tasks. For Diffusion-QL, CPQL, and CPIQL, we here provide the "best" scores, representing the best performance during training. More experimental results, including the final performance and the standard deviation of each task, are provided in Appendix D. The bold values are the highest among each row.

Dataset	TD3+BC	IQL	SfBC	IDQL	Diffusion-QL	EDP	QGPO	Diffuser	DD	CPIQL	CPQL
halfcheetah-medium-v2	48.3	47.4	45.9	49.7	51.1	52.1	54.1	42.8	49.1	55.3	57.9
hopper-medium-v2	59.3	66.3	57.1	63.1	90.5	81.9	98.0	74.3	79.3	101.5	102.1
walker2d-medium-v2	83.7	78.3	77.9	80.2	87.0	86.9	86.0	79.6	82.5	88.4	90.5
halfcheetah-medium-replay-v2	44.6	44.2	37.1	45.1	47.8	49.4	47.6	37.7	39.3	49.8	48.1
hopper-medium-replay-v2	60.9	94.7	86.2	82.4	101.3	101.0	96.9	93.6	100.0	101.7	101.7
walker2d-medium-replay-v2	81.8	73.9	65.1	79.8	95.5	94.9	84.4	70.6	79.0	95.0	94.4
halfcheetah-medium-expert-v2	90.7	86.7	92.6	94.4	96.8	95.5	93.5	88.9	90.6	90.2	98.8
hopper-medium-expert-v2	98.0	91.5	108.6	105.3	111.1	97.4	108.0	103.3	111.8	113.4	114.2
walker2d-medium-expert-v2	110.1	109.6	109.8	111.6	110.1	110.2	110.7	106.9	108.8	112.3	111.5
Average	75.3	77.0	75.6	79.1	87.9	85.5	86.5	77.5	81.8	89.7	91.0
pen-human-v1	0.6	71.5	-	-	72.8	48.2	-	-	-	58.2	89.3
pen-cloned-v1	-2.5	37.3	-	-	57.3	15.9	-	-	-	77.4	83.3
Average	-1.0	54.4	-	-	65.1	32.1	-	-	-	67.8	86.3

where \mathcal{D}_r refers to the replay buffer, and π_r is the policy for collecting the data in \mathcal{D}_r . When π_r is the uniform policy, the above problem is equivalent to maximum entropy RL [18]. It is readily observable that the objective function for online RL tasks bears a striking resemblance to Eq. (1), and Eq. (2) can also serve as a closed-form solution for Eq. (17) with \mathcal{D}_r replacing $\mathcal{D}_{offline}$ and π_r replacing μ . Therefore, CPQL can be seamlessly extended to online tasks. In maximum entropy RL methods, careful tuning of λ is necessary to strike a balance between exploration and exploitation. In CPQL, this situation is alleviated. Firstly, our consistency policy is inherently stochastic, and we find in experiments that this inherent randomness allows for sufficient exploration of the environment without the need for additional exploration strategies. Secondly, the data collected by this policy incorporates this stochasticity and makes the consistency policy maintain this randomness through Eq. (13). Lastly, as the policy asymptotically converges, the proportion of good samples in the data increases, reducing the randomness in the replay buffer, consequently decreasing the policy's randomness and achieving policy convergence. It is worth noting that the aforementioned process does not require manual adjustment of λ ; rather, it is implicit in the CPQL policy iteration. Furthermore, unlike previous maximum entropy reinforcement learning methods, we find that CPQL's performance is not sensitive to λ for different tasks, greatly alleviating the tuning complexity associated with maximum entropy reinforcement learning. We summarize the complete algorithm procedure of CPQL in Algorithm 2 for online tasks, and we also provide an illustrative description of how CPQL works for both offline and online RL tasks in Appendix B.

5 EXPERIMENTS

In this section, we conduct several experiments on the D4RL benchmark [12], dm_control tasks [46], Gym MuJoCo tasks [45] to evaluate the performance and time efficiency of the consistency policy.

Algorithm 2 CPQL for Online RL Tasks

```

Initialize the policy network  $\pi_\theta$ , critic networks  $Q_{v_1}$  and  $Q_{v_2}$ , and target networks  $\pi_{\theta^-}$ ,  $Q_{v_1^-}$  and  $Q_{v_2^-}$ 
Initialize the dataset  $\mathcal{D}_r \leftarrow \emptyset$ 
# Warm up
for  $i \in 0, \dots, W$  do
    Generate  $\hat{a}_t^\epsilon$  by Eq. (7) with  $s_t$ 
    Play  $\hat{a}_t^\epsilon$  and get  $s_{t+1} \sim \mathbb{P}(\cdot | s_t, \hat{a}_t^\epsilon)$ 
     $\mathcal{D}_r \leftarrow \mathcal{D}_r \cup \{s_t, \hat{a}_t^\epsilon, r_t, s_{t+1}\}$ 
end for
for each iteration do
    # Update the dataset
    Sample  $\hat{a}_t^\epsilon \sim \pi_\theta(a_t | s_t)$  by Eq. (7) with  $s_t$ 
    Play  $\hat{a}_t^\epsilon$  and get  $s_{t+1} \sim \mathbb{P}(\cdot | s_t, \hat{a}_t^\epsilon)$ 
     $\mathcal{D}_r \leftarrow \mathcal{D}_r \cup \{s_t, \hat{a}_t^\epsilon, r_t, s_{t+1}\}$ 
    # Policy training
    Sample transition batch  $B = \{(s_t, a_t, r_t, s_{t+1})\} \sim \mathcal{D}_r$ 
    # Q-learning
    Update  $Q_{v_i}$  by Eq. (14)
    # Consistency policy learning
    Update policy by minimizing Eq. (13)
    # Update target networks
     $\theta^- = \rho\theta^- + (1 - \rho)\theta$ 
     $v_i^- = \rho v_i^- + (1 - \rho)v_i$  for  $i \in \{1, 2\}$ 
end for

```

We also provide various ablation studies for a better understanding of how the hyperparameter (α, η) in Eq. (13) and new loss Eq. (12) affect the performance. Throughout this paper, the results are reported by averaging five random seeds. For a detailed look at the experimental setup and corresponding hyperparameters, please refer to Appendix C.

Table 2: The performance of CPQL and SOTA baselines on dm_control tasks under 500K environment steps. The results of MPO, DMPO, D4PG, and DreamerV3 are from the paper of DreamerV3 [19]. The bold values are the highest among each row.

Tasks	TD3	SAC	PPO	MPO	DMPO	D4PG	DreamerV3	CPQL
Acrobot Swingup	46.8	33.2	34.4	80.6	98.5	125.5	154.5	183.1
Cartpole Balance	982.6	961.9	997.6	958.4	998.5	998.8	990.5	999.2
Cartpole Balance Sparse	992.0	993.5	1000.0	998.0	994.0	979.6	996.8	1000.0
Cartpole Swingup	829.2	781.4	760.7	857.7	857.8	874.6	850.0	860.7
Cheetah Run	546.2	530.9	560.4	612.3	581.6	623.5	575.9	727.6
Finger Spin	847.1	825.4	369.9	766.9	744.3	818.4	937.2	965.4
Finger Turn Easy	337.6	371.4	275.2	430.4	593.8	524.5	745.4	874.1
Finger Turn Hard	334.4	344.8	5.06	250.8	384.5	379.2	841.0	864.6
Hopper Hop	40.0	41.7	0.0	37.5	71.5	67.5	111.0	130.1
Hopper Stand	322.7	270.9	2.2	279.3	519.5	755.4	573.2	902.1
Reacher Easy	968.8	973.4	541.6	954.4	965.1	941.5	947.1	981.2
Reacher Hard	965.7	928.2	518.0	914.1	956.8	932.0	936.2	963.6
Walker Run	274.5	445.9	131.7	539.5	462.9	593.1	632.7	683.8
Walker Stand	957.7	973.4	528.4	960.4	971.6	935.2	956.9	983.6
Walker Walk	934.7	922.8	408.9	924.9	933.1	965.1	935.7	952.6
Average	625.3	626.6	408.9	637.7	675.6	700.9	745.6	804.9

Baselines for Offline RL Tasks. For offline scenario, we evaluate on two domains of D4RL benchmark and compare CPQL with current methods that achieve SOTA performance, including Q-learning with policy constraints such as TD3+BC [13], implicit Q-learning such as IQL [26], and diffusion-model-based methods, such as Diffuser [22], DecisionDiffuser (DD) [3], SfBC [7], IDQL [20], Diffusion-QL[49], EDP [23], QGDO [34]. Notably, Diffuser and DD employ the diffusion model to capture trajectory dynamics, while the remaining methods focus on modeling policy distributions.

Baselines for Online RL Tasks. For the online scenario, we evaluate our method on several tasks, such as dm_control tasks and Gym MuJoCo tasks. We compared CPQL with the current methods of achieving SOTA. These methods include off-policy methods such as TD3 [15], SAC [18], MPO [2], D4PG [4], DMPO [1], and on-policy methods such as PPO [42]. In addition, the comparison methods also include model-based methods such as DreamerV3 [19], and diffusion-model-based methods such as DIPO [53].

5.1 Overall Results

In this section, we illustrate the overall results on the offline tasks (D4RL) and the online tasks (dm_control, Gym Mujoco), showing that CPQL achieves competitive performance on both offline and online tasks compared with current SOTA methods. All training curves for CPQL and CPIQL can be found in Appendix E.

5.1.1 Results on D4RL (Offline Tasks). Firstly, we evaluate the performance of the methods on offline D4RL tasks. The experimental results are shown in Table 1. From the experimental results, we can see that the methods based on the diffusion model have significant advantages in performance compared with the unimodal distribution policy method (TD3+BC and IQL). Compared with Diffusion-QL and EDP, QGPO achieves excellent performance by using an accurate guided value function with contrastive loss. Our proposed CPQL avoids inaccurate guidance problems by modeling

ODE trajectory mapping, thus achieving competitive results by approximately 4% improvement (compared with Diffusion-QL). By comparing CPQL and CPIQL, we can find that consistency policy can effectively constrain the sampling of OOD actions without the help of implicit Q-learning. Additionally, CPIQL outperforms IQL by a significant 16.5% on locomotion tasks, indicating that the strong representation ability of consistency policy can effectively achieve policy improvement with accurate diffusion guidance. In conclusion, CPQL achieves better guidance with value function while meeting constraints, thus making their performance outstanding on offline tasks.

5.1.2 Results on dm_control (Online Tasks). Next, we evaluate the methods on 15 dm_control tasks. Part of the training curves and all experimental results are shown in Figure 2 and Table 2, respectively. As shown in Figure 2, CPQL has shown an improvement in both training stability and sample efficiency with better performance when compared to other algorithms under a small number of interactions with the environment (500 K). Also, the experimental results show that CPQL performance outperforms previous SOTA methods, including DreamerV3 on 15 tasks, especially on "Cheetah Run" and "Finger Turn Easy" tasks. On average, there has been an improvement of approximately 8% as compared to DreamerV3. In addition, CPQL directly models the policy and is not plagued by inaccurate diffusion guidance, resulting in better performance on complex tasks, with great representation ability for the policy. It is worth mentioning that DMPO and D4PG have adopted the distributional value function [5], which has proved to achieve better performance than the expected value function. While CPQL only uses the expected value function, it has shown significant performance leadership, which also means the importance of exploration in RL, and the consistency model has great potential in RL.

5.1.3 Results on Gym MuJoCo (Online Tasks). We also evaluate the methods on 6 tasks of Gym MuJoCo. Part of the training curves

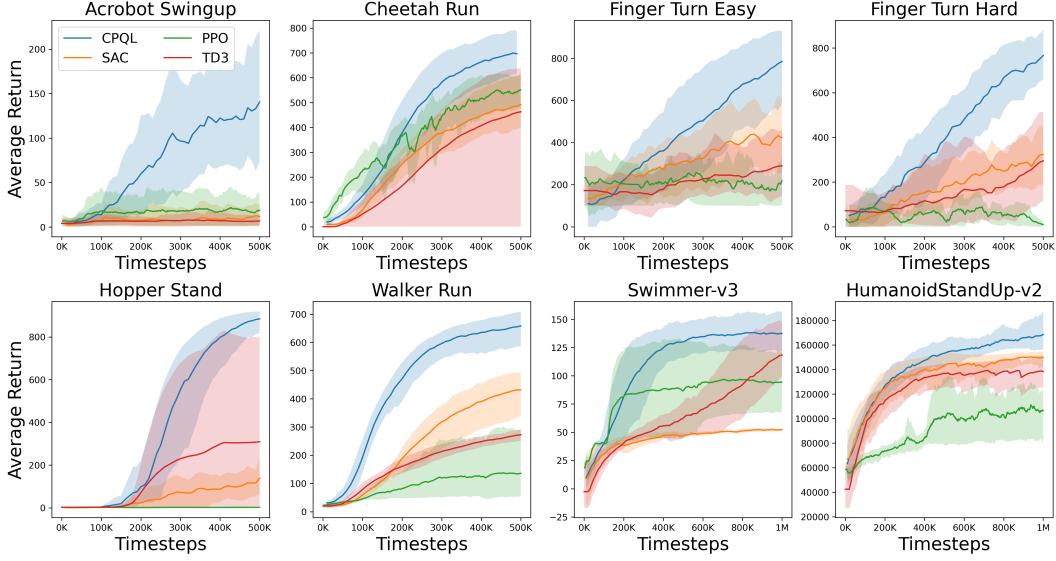


Figure 2: Training curves for 8 online tasks, including 6 tasks from dm_control (top row, and leftmost two of the bottom row) and 2 tasks from Gym Mujoco (rightmost two of the bottom row). CPQL, PPO, SAC and TD3 are compared on each task with 5 random seeds.

Table 3: The performance of CPQL and SOTA baselines on Gym MuJoCo tasks under 1M environment steps. The bold values are the highest among each row.

Tasks	TD3	SAC	PPO	DIPO	CPQL
Swimmer-v3	108.3	51.7	94.6	72.2	137.2
Walker2d-v3	4127.0	4631.9	3751.8	4409.6	5139.9
Ant-v3	5421.9	5665.5	2921.0	5620.2	6209.4
HalfCheetah-v3	10779.9	11287.9	2449.5	10475.2	12195.2
Humanoid-v3	5253.0	4993.3	704.1	4878.5	5394.6
HumanoidStandup-v2	136897.9	150934.8	105654.4	145350.2	174480.6

and all experimental results are shown in Figure 2 and Table 3. Consistent with the conclusion on the dm_control tasks, the methods based on the diffusion model maintain dominance on multiple RL tasks. As we can see from the rightmost two graphs of the bottom row in Figure 2, CPQL has significant performance advantages compared with other algorithms within 1M interactions. It is worth mentioning that DIPO proposed the concept of action gradient and used a diffusion model to fit the replay buffer updated by action gradient. However, on one hand, it will be affected by the coverage of the initial dataset; on the other hand, only fitting the optimal data is likely to cause the diffusion model to lose sampling diversity to reduce its exploration ability. Therefore, CPQL performance is better than DIPO.

5.2 Time Efficiency

We then evaluate the time efficiency of policy training and inference. Following the evaluation criterion in EDP [23], we compare methods including Diffusion-QL, EDP, DIPO, and CPQL by iterations-per-second (IPS) for training speed and steps-per-second (SPS) for inference speed. We choose "walker2d-medium-expert-v2" as the

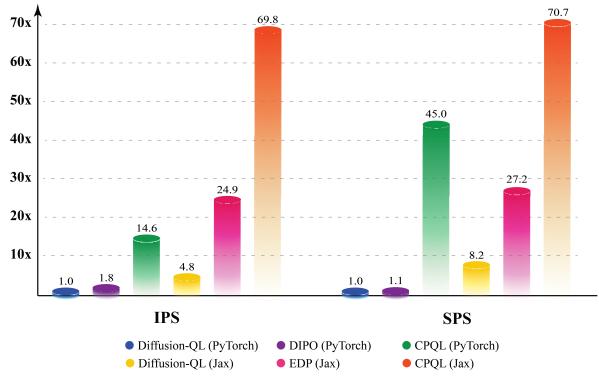
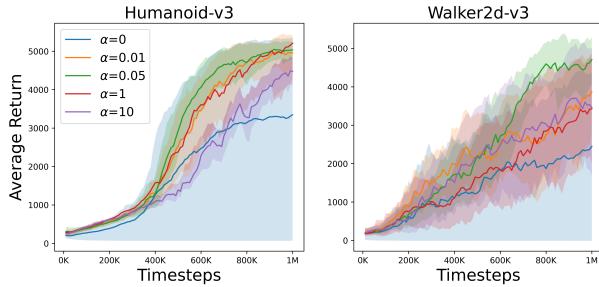


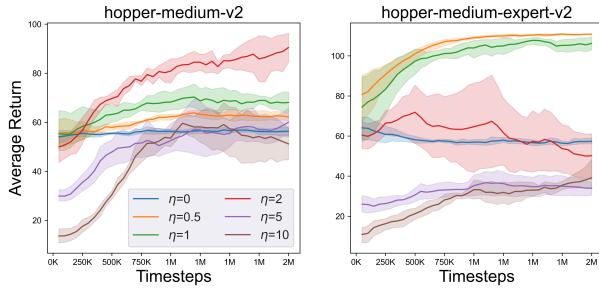
Figure 3: Training and inference speedup on D4RL locomotion tasks. We choose the Diffusion-QL (Pytorch) as the baseline with all original data provided in Appendix D. Implementation for Diffusion-QL(Pytorch), DIPO(Pytorch), EDP(Jax) are from the official repository, with websites listed in Appendix C. And we implement Diffusion-QL(Jax) by ourselves for comparison.

testbed and run each method for 100K iterations of policy updates to calculate the IPS. Then, we sample 100K transitions by interacting with the environment to calculate the SPS.

From the result in Figure 3, CPQL's training and sampling speeds are far faster than other methods based on the diffusion model. Especially compared to Diffusion-QL, CPQL has improved its training speed by nearly 15 times and inference speed by nearly 45 times. Even though EDP also uses one-step diffusion in actor training to approximate the multi-step diffusion process in Diffusion-QL, as well as DPM-solver to speed up sampling, multi-step sampling is



(a) $\eta = 1$ and different α for selected online RL tasks



(b) $\alpha = 1$ and different η for selected offline RL tasks

Figure 4: The effect of the different hyperparameters η and α on the online and offline policy learning.

still necessary during Q-function training and inference sampling. Hence the consistency policy also has a noteworthy improvement in both the training and inference speed compared to the EDP.

5.3 Ablation Study

In this section, we conduct various ablation studies on both online and offline continuous control tasks to determine the factors affecting the performance, including evaluating the impact of different parameters on the policy training and analyzing the stability introduced by reconstruction loss during consistency policy training.

Hyperparameter. η and α respectively control the impact of the Q-function on policy and the KL divergence between learned policy and reference policy. In online tasks, we set $\eta = 1$ and investigate the impact of different α on policy performance. The experimental results are shown in Figure 4 (a). According to the results of two different tasks, an appropriate α is crucial for obtaining better performance. If α is too small or too large, it will cause performance degradation. But the reasons for performance degradation are different. When α is small, the consistency policy is more affected by the gradient of the Q-function, leading to premature loss of sampling diversity and affecting the exploration ability. When α is relatively large, the data distribution greatly affects the consistency policy, which leads to the policy ignoring the guidance of the Q-function.

In offline tasks, we set $\alpha = 1$ and investigate the impact of different η on policy performance. The results are shown in Figure 4 (b). The results of two different tasks indicate that when the η ratio is small, the impact of the Q-function decreases, and the performance of the policy tends to approach that of behavioral

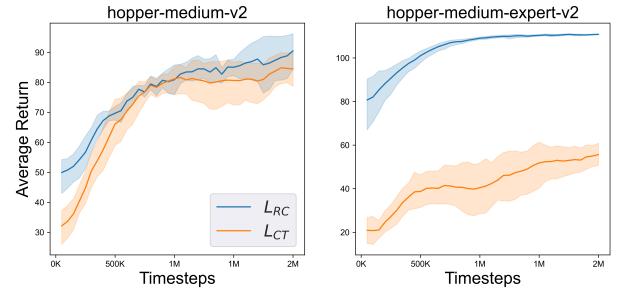


Figure 5: The effect of the different training losses on policy learning.

cloning. When η is relatively large, the constraint effect of the dataset on the policy decreases, leading to more OOD actions in the sampling process, resulting in deviation of Q-value estimation, which also leads to performance degradation.

Training Loss for Consistency Model. We propose replacing consistency training loss L_{CT} with reconstruction loss L_{RC} . Here we study the influence of two different loss functions on policy training as shown in Figure 5. During the policy training, we find that consistency training loss would lead to policy collapse. The results show that reconstruction loss makes the training process more stable and can achieve better results.

6 CONCLUSION

In this work, we propose the time-efficiency consistency policy with Q-learning (CPQL), which constructs the mapping from the PF ODE trajectories to the desired policy and achieves policy improvement with accurate guidance for both offline and online RL tasks. We also introduce an empirical loss for stabilizing consistency policy training. Experimental results show that CPQL achieves about 4% improvement on D4RL locomotion tasks (compared to Diffusion-QL) and 8% improvement on dm_control tasks (compared to DreamerV3). Meanwhile, CPQL significantly improves inference speed by nearly 45 times compared to Diffusion-QL. We show the potential of the consistency model in RL and believe this impact is profound. Its efficient sampling speed has greatly expanded real-time applications such as robot control based on the diffusion model. Of course, there are still many problems, especially for online RL. For instance, how to better control the diversity of diffusion model in the training process and whether it is possible to build a multi-step consistency policy or not to enhance the representation of the policy is worth studying.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants No. 62136008, No. 62103409, the Strategic Priority Research Program of Chinese Academy of Sciences (CAS) under Grant XDA27030400 and in part by the International Partnership Program of the Chinese Academy of Sciences under Grant 104GJHZ2022013GC.

REFERENCES

- [1] Abbas Abdolmaleki, Sandy H. Huang, Leonard Hasen clever, Michael Neunert, H. Francis Song, Martina Zambelli, Murilo F. Martins, Nicolas Heess, Raia Hadsell, and Martin A. Riedmiller. 2020. A distributional view on multi-objective policy optimization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 11–22. <http://proceedings.mlr.press/v119/abdlmaleki20a.html>
- [2] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. 2018. Maximum a Posteriori Policy Optimisation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=S1ANxQW0b>
- [3] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. 2023. 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. <https://openreview.net/pdf?id=sP1f02k9DFG>
- [4] Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva B. Alistair Muldal, Nicolas Heess, and Timothy P. Lillicrap. 2018. Distributed Distributional Deterministic Policy Gradients. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=SyZipzbCb>
- [5] Marc G. Bellemare, Will Dabney, and Rémi Munos. 2017. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*. Doina Precup and Yee Whye Teh (Eds.). PMLR, 449–458. <http://proceedings.mlr.press/v70/bellemare17a.html>
- [6] Johann Brehmer, Joey Bose, Pim de Haan, and Taco Cohen. 2023. EDGI: Equivariant Diffusion for Planning with Embodied Agents. *CoRR* abs/2303.12410 (2023). <https://doi.org/10.48550/arXiv.2303.12410> arXiv:2303.12410
- [7] Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. 2023. Offline Reinforcement Learning via High-Fidelity Generative Behavior Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=42zs3qz2kpy>
- [8] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (Eds.). <https://doi.org/10.15607/RSS.2023.XIX.026>
- [9] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (Eds.). <https://doi.org/10.15607/RSS.2023.XIX.026>
- [10] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.), 8780–8794. <https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html>
- [11] Xing Fang, Qichao Zhang, Yinfeng Gao, and Dongbin Zhao. 2022. Offline Reinforcement Learning for Autonomous Driving with Real World Driving Data. In *25th IEEE International Conference on Intelligent Transportation Systems, ITSC 2022, Macau, China, October 8-12, 2022*. IEEE, 3417–3422. <https://doi.org/10.1109/ITSC55140.2022.922100>
- [12] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *CoRR* abs/2004.07219 (2020). arXiv:2004.07219 <https://arxiv.org/abs/2004.07219>
- [13] Scott Fujimoto and Shixiang Shane Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 20132–20145. <https://proceedings.neurips.cc/paper/2021/hash/a8166da05c5a0947d0c03724b41886e5-Abstract.html>
- [14] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*. Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2052–2062. <http://proceedings.mlr.press/v97/fujimoto19a.html>
- [15] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*. Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1582–1591. <http://proceedings.mlr.press/v80/fujimoto18a.html>
- [16] Jiali Gao, Kaize Hu, Guowei Xu, and Huazhe Xu. 2023. Can Pre-Trained Text-to-Image Models Generate Visual Goals for Reinforcement Learning? *CoRR* abs/2307.07837 (2023). <https://doi.org/10.48550/arXiv.2307.07837> arXiv:2307.07837
- [17] Wonjoon Go and Scott Niekum. 2022. Know Your Boundaries: The Necessity of Explicit Behavioral Cloning in Offline RL. *CoRR* abs/2206.00695 (2022). <https://doi.org/10.48550/arXiv.2206.00695> arXiv:2206.00695
- [18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*. Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1856–1865. <http://proceedings.mlr.press/v80/haarnoja18b.html>
- [19] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. 2023. Mastering Diverse Domains through World Models. *CoRR* abs/2301.04104 (2023). <https://doi.org/10.48550/arXiv.2301.04104> arXiv:2301.04104
- [20] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzień Kuba, and Sergey Levine. 2023. IDQL: Implicit Q-Learning as an Actor-Critic Method with Diffusion Policies. *CoRR* abs/2304.10573 (2023). <https://doi.org/10.48550/arXiv.2304.10573> arXiv:2304.10573
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>
- [22] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 9902–9915. <https://proceedings.mlr.press/v162/janner22a.html>
- [23] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. 2023. Efficient Diffusion Policies for Offline Reinforcement Learning. *CoRR* abs/2305.20081 (2023). <https://doi.org/10.48550/arXiv.2305.20081> arXiv:2305.20081
- [24] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. 2023. DALL-E-Bot: Introducing Web-Scale Diffusion Models to Robotics. *IEEE Robotics Autom. Lett.* 8, 7 (2023), 3956–3963. <https://doi.org/10.1109/LRA.2023.3272516>
- [25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html
- [26] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. 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. <https://openreview.net/forum?id=68n2s9ZJWF8>
- [27] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 11761–11771. <https://proceedings.neurips.cc/paper/2019/hash/c2073ffa77b3537a498057413bb09d3a-Abstract.html>
- [28] Aviral Kumar, Justin Fu, Aviral Kumar, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/0db2061826a5df3221116a5085a6052-Abstract.html>
- [29] Sascha Lange, Thomas Gabel, and Martin A. Riedmiller. 2012. Batch Reinforcement Learning. In *Reinforcement Learning*. Marco A. Wiering and Martijn van Otterlo (Eds.). Adaptation, Learning, and Optimization, Vol. 12. Springer, 45–73. https://doi.org/10.1007/978-3-642-27645-3_2
- [30] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *CoRR* abs/2005.01643 (2020). arXiv:2005.01643 <https://arxiv.org/abs/2005.01643>
- [31] Xiang Li, Varun Belaguli, Jinghuai Shang, and Michael S. Ryoo. 2023. Crossway Diffusion: Improving Diffusion-based Visuomotor Policy via Self-supervised Learning. *CoRR* abs/2307.01849 (2023). <https://doi.org/10.48550/arXiv.2307.01849> arXiv:2307.01849

- [32] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1509.02971>
- [33] Cong Lu, Philip J. Ball, and Jack Parker-Holder. 2023. Synthetic Experience Replay. *CoRR* abs/2303.06614 (2023). <https://doi.org/10.48550/arXiv.2303.06614> arXiv:2303.06614
- [34] Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. 2023. Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning. *CoRR* abs/2304.12824 (2023). <https://doi.org/10.48550/arXiv.2304.12824> arXiv:2304.12824
- [35] Xiaoteng Ma, Yiqin Yang, Hao Hu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, Bin Liang, and Qihai Liu. 2022. Offline Reinforcement Learning with Value-based Episodic Memory. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=RCZqv9NXZ>
- [36] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. 2019. AlgaeDICE: Policy Gradient from Arbitrary Experience. *CoRR* abs/1912.02074 (2019). arXiv:1912.02074 <http://arxiv.org/abs/1912.02074>
- [37] Tim Pearce, Tabish Rashid, Anssi Kanervisto, David Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. 2023. Imitating Human Behaviour with Diffusion Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=Pv1GPQzRrc8>
- [38] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *CoRR* abs/1910.00177 (2019). arXiv:1910.00177 <http://arxiv.org/abs/1910.00177>
- [39] Jan Peters, Katharina Mülling, and Yasemin Altun. 2010. Relative Entropy Policy Search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, Maria Fox and David Poole (Eds.). AAAI Press. <http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1851>
- [40] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. 2023. Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 13756–13766. <https://doi.org/10.1109/CVPR52729.2023.01322>
- [41] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. 2023. Goal-Conditioned Imitation Learning using Score-based Diffusion Policies. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (Eds.). <https://doi.org/10.15607/RSS.2023.XIX.028>
- [42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 <http://arxiv.org/abs/1707.06347>
- [43] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency Models. *CoRR* abs/2303.01469 (2023). <https://doi.org/10.48550/arXiv.2303.01469> arXiv:2303.01469
- [44] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. <https://openreview.net/forum?id=PxTIG12RRHS>
- [45] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*. IEEE, 5026–5033. <https://doi.org/10.1109/IROS.2012.6386109>
- [46] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy P. Lillicrap, Nicolas Heess, and Yuval Tassa. 2020. dm_control: Software and tasks for continuous control. *Softw. Impacts* 6 (2020), 100022. <https://doi.org/10.1016/j.simpa.2020.100022>
- [47] Hado van Hasselt. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta (Eds.). Curran Associates, Inc., 2613–2621. <https://proceedings.neurips.cc/paper/2010/hash/091d584fcfed301b442654dd8c23b3fc9-Abstract.html>
- [48] Hsiang-Chun Wang, Shang-Fu Chen, and Shao-Hua Sun. 2023. Diffusion Model-Augmented Behavioral Cloning. *CoRR* abs/2302.13335 (2023). <https://doi.org/10.48550/arXiv.2302.13335> arXiv:2302.13335
- [49] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. 2023. 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. <https://openreview.net/pdf?id=AHvFDPi-FA>
- [50] Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior Regularized Offline Reinforcement Learning. *CoRR* abs/1911.11361 (2019). arXiv:1911.11361 <http://arxiv.org/abs/1911.11361>
- [51] Chenjun Xiao, Han Wang, Yangchen Pan, Adam White, and Martha White. 2023. The In-Sample Softmax for Offline Reinforcement Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=u-RuvyDqCM>
- [52] Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Wai Kin Victor Chan, and Xianyuhan Zhan. 2023. Offline RL with No OOD Actions: In-Sample Learning via Implicit Value Regularization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=ueYYgo2pSSU>
- [53] Long Yang, Zhixiong Huang, Fenghao Lei, Yucun Zhong, Yiming Yang, Cong Fang, Shiting Wen, Binbin Zhou, and Zhouchen Lin. 2023. Policy Representation via Diffusion Probability Model for Reinforcement Learning. *CoRR* abs/2305.13122 (2023). <https://doi.org/10.48550/arXiv.2305.13122> arXiv:2305.13122
- [54] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. 2023. Guided Conditional Diffusion for Controllable Traffic Simulation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*. IEEE, 3560–3566. <https://doi.org/10.1109/ICRA48891.2023.10161463>

APPENDICES

A MATHEMATICAL PROOFS

Notations. For a decision-making problem in RL, we use s and a to represent the state and action, respectively. $p(s'|s, a)$ represents the transition probability from state s to next state s' after taking the action a . We use subscripts $t \in \{1, \dots, T\}$ to denote trajectory timesteps and we use superscripts $k \in [0, K]$ to denote diffusion timesteps with $m \in \{1, \dots, M\}$ denoting time sub-intervals.

Given a multi-variate function $h(x, y)$, we let $\partial_1(x, y)$ denote the Jacobian of h over x , and analogously $\partial_2(x, y)$ denote the Jacobian of h over y . Unless otherwise stated, a is supposed to be a random variable sampled from the data distribution $p_{data}(a)$, m is sampled uniformly at random from $[1, \dots, M - 1]$.

Preliminaries. We use $f_\theta(a^k, k|s)$ to denote a consistency policy parameterized by θ . The outputs of the consistency policy are consistent for all arbitrary pairs of (a^k, k) on the same PF ODE with an empirical estimate: $\frac{da^k}{dk} = -ks_\phi(a^k, k)$ where the pre-trained score function is parameterized by ϕ . By utilizing a numerical ODE solver such as an Euler solver, we can accurately estimate a^{k_m} from $a^{k_{m+1}}$, which is denoted as:

$$\hat{a}_{\phi^*}^{k_m} = a^{k_{m+1}} + (k_m - k_{m+1})s_{\phi^*}(a^{k_{m+1}}, k_{m+1}) \quad (19)$$

The distillation loss is defined as:

$$L_{CD}(\theta, \theta^-; \phi^*) = \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(\hat{a}_{\phi^*}^{k_m}, k_m|s))] \quad (20)$$

The consistency loss is defined as:

$$L_{CT}(\theta, \theta^-) = \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a + k_m z, k_m|s))] \quad (21)$$

A.1 The Proof of Training Loss for Consistency Policy

THEOREM 2. Let $\Delta k = \max_{m \in [1, M-1]} \{|k_{m+1} - k_m|\}$. We have the assumptions: 1) Distance function d , value function Q and f_{θ^-} are all twice continuously differentiable with bounded second derivatives; 2) There is a pre-trained score model represents the desired policy: $\forall k \in [\epsilon, K] : s_{\phi^*}(a^k, k) = \nabla \log p^k(a^k)$, which cannot be accessed; 3) f_θ satisfies the Lipschitz condition: there exists $L > 0$ such that for all $k \in [\epsilon, K]$, and $x, y \in A$, we have $\|f_\theta(x, k) - f_\theta(y, k)\|_2 \leq L\|x - y\|_2$.

The following training loss for the consistency policy is sufficient to replace the distillation loss:

$$L(\theta, \theta^-) = \alpha L_{CT}(\theta, \theta^-) - \beta \mathbb{E}[Q(s, \hat{a}^\epsilon)] + o(\Delta k) \quad (22)$$

where

$$\begin{aligned} \alpha &= \mathbb{E}\left[1 + Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}\right] \\ \beta &= \mathbb{E}\left[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a + k_{m+1}z, k_{m+1}|s)) * \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}\right] \end{aligned} \quad (23)$$

PROOF. The Lemma 1 in Appendix A.3 of [43] provides a theoretical justification for the distillation loss and the consistency training loss. The expression of the score function can be further simplified to yield:

$$\nabla \log p^k(a^k) = -\mathbb{E}\left[\frac{a^k - a}{k^2} | a^k\right] \quad (24)$$

where $k \in [\epsilon, K]$.

According to Eq. (24) and Eq. (2) claimed in Section 3.1, we can derive the following expression of the score function for the consistency policy:

$$\nabla \log \pi^k(a^k) = \nabla \log \mu_k(a^k) + \nabla \frac{1}{\lambda} Q(s, a^k) = -\mathbb{E}\left[\frac{a^k - a}{k^2} | a^k\right] + \nabla \frac{1}{\lambda} Q(s, a^k) \quad (25)$$

With Taylor expansion, we have:

$$\begin{aligned}
L_{CD}(\theta, \theta_-; \phi^*) &= \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a_{\phi^*}^{k_m}, k_m|s))] \\
&= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}} + (k_{m+1} - k_m)k_{m+1}\nabla \log \pi_{k_{m+1}}(a^{k_{m+1}}, k_m|s)))] \\
&= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s) \\
&\quad + \partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)\nabla \pi_{k_{m+1}}(a^{k_{m+1}}) + \partial_2 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_m - k_{m+1})) + o(|k_{m+1} - k_m|)] \\
&= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))] \\
&\quad + \partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)\nabla \pi_{k_{m+1}}(a^{k_{m+1}}) \\
&\quad + \partial_2 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_m - k_{m+1})] + o(|k_{m+1} - k_m|)] \\
&= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))] \\
&\quad + \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_{m+1} - k_m)k_{m+1}\nabla \pi_{k_{m+1}}(a^{k_{m+1}})]] \\
&\quad + \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_2 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_m - k_{m+1})] + o(|k_{m+1} - k_m|)] \\
&= \mathcal{A} + \mathcal{B}
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
\mathcal{A} &= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))] \\
&\quad + \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_m - k_{m+1})k_{m+1}\mathbb{E}\left[\frac{a^{k_{m+1}} - a}{k^2} | a^{k_{m+1}}\right]]] \\
&\quad + \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_2 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_m - k_{m+1})]] + \mathbb{E}[o(|k_{m+1} - k_m|)]
\end{aligned} \tag{27}$$

and

$$\mathcal{B} = \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_{m+1} - k_m)k_{m+1}\frac{1}{\lambda}\nabla Q(s, a^{k_{m+1}})]] \tag{28}$$

The Theorem 2 in [43] has proofed that:

$$\mathcal{A} = L_{CT}(\theta, \theta^-) + o(\Delta k) \tag{29}$$

We will then focus on the term \mathcal{B} . With Taylor expansion, we have:

$$Q(s, a) = Q(s, a^{k_{m+1}}) + \partial_1 Q(s, a^{k_{m+1}})(a - a^{k_{m+1}}) \tag{30}$$

Thus we have:

$$\partial_1 Q(s, a^{k_{m+1}}) = \frac{Q(s, a) - Q(s, a^{k_{m+1}})}{a - a^{k_{m+1}}} \tag{31}$$

According to Eq. (28) and Eq. (31), we can further derive the following relation:

$$\begin{aligned}
\mathcal{B} &= \frac{1}{\lambda} \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_{m+1} - k_m)k_{m+1}\frac{Q(s, a) - Q(s, a^{k_{m+1}})}{a - a^{k_{m+1}}}]]) \\
&\stackrel{(i)}{=} \frac{1}{\lambda} \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)(k_{m+1} - k_m)\frac{Q(s, a^{k_{m+1}}) - Q(s, a)}{z}]] \\
&= -\frac{1}{\lambda} \mathbb{E}[\partial_2 d(f_\theta(a^{k_{m+1}}, k_{m+1}), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}))[\partial_1 f_{\theta^-}(a^{k_{m+1}}, k_{m+1})(k_m - k_{m+1})z(Q(s, a^{k_{m+1}}) - Q(s, a))\frac{k_{m+1}^2}{(a^{k_{m+1}} - a)^2}]]
\end{aligned} \tag{32}$$

where (i) is due to $a^{k_{m+1}} \sim \mathcal{N}(a, k_{m+1}^2 I)$. Then we use the Taylor expansion in the reverse direction to obtain:

$$\begin{aligned}
\mathcal{B} &= \frac{1}{\lambda} \mathbb{E}[(d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) \\
&\quad - d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}} + (k_m - k_{m+1})z, k_{m+1}|s)))(Q(s, a^{k_{m+1}}) - Q(s, a))\frac{k_{m+1}^2}{(a^{k_{m+1}} - a)^2}] \\
&= \frac{1}{\lambda} \mathbb{E}[(d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) - d(f_\theta(a^{k_{m+1}}, k_{m+1}), f_{\theta^-}(a + k_m z, k_{m+1}|s)))(Q(s, a^{k_{m+1}}) - Q(s, a))\frac{k_{m+1}^2}{(a^{k_{m+1}} - a)^2}] \\
&\stackrel{(ii)}{\approx} \frac{1}{\lambda} \mathbb{E}[(d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) - d(f_\theta(a^{k_{m+1}}, k_{m+1}), f_{\theta^-}(a^{k_m}, k_m|s)))(Q(s, a^{k_{m+1}}) - Q(s, a))\frac{k_{m+1}^2}{(a^{k_{m+1}} - a)^2}]
\end{aligned} \tag{33}$$

where (ii) takes the approximation of $k_m \approx k_{m+1}$, which will result in an error bounded by $L||k_{m+1} - k_m||_2$.

Following [34], we could have:

$$\begin{aligned}
Q(s, a^{k_m}) &= \log \mathbb{E}_{p_{data}(a|a^{k_m})} [e^{\beta Q(s,a)}] \\
&= \log \int \mu(a) e^{Q(s,a)} da \\
&= \log \int \pi(a) da \\
&= 0
\end{aligned} \tag{34}$$

Then we have:

$$\mathcal{B} = -\frac{1}{\lambda} \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) - d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_m}, k_m|s))] Q(s, a) \frac{k_{m+1}^2}{(a^{k_{m+1}} - a)^2}] \tag{35}$$

Thus, we can derive that:

$$\begin{aligned}
\mathcal{A} + \mathcal{B} &= \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_m}, k_m|s))(1 + Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}) \\
&\quad - d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2})] + o(\Delta k) \\
&\stackrel{(iii)}{\approx} \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_m}, k_m|s))(1 + Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}) \\
&\quad - d(f_\theta(a^{k_{m+1}}, k_{m+1}), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) Q(s, \hat{a}^\epsilon) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2})] + o(\Delta k)
\end{aligned} \tag{36}$$

where (iii) takes the approximation of $a \approx \hat{a}^\epsilon$

For the consistency model training, we define the hyperparameter:

$$\alpha = \mathbb{E}[1 + Q(s, a) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}] \tag{37}$$

and

$$\beta = \mathbb{E}[d(f_\theta(a^{k_{m+1}}, k_{m+1}|s), f_{\theta^-}(a^{k_{m+1}}, k_{m+1}|s)) \frac{k_{m+1}^2}{\lambda(a^{k_{m+1}} - a)^2}] \tag{38}$$

Then we have:

$$L_{CD}(\theta, \theta^-; \phi^*) = \mathcal{A} + \mathcal{B} = \alpha L_{CT}(\theta, \theta^-) - \beta \mathbb{E}[Q(s, \hat{a}^\epsilon)] + o(\Delta k) \tag{39}$$

Note that the a for calculating $L_{CT}(\theta, \theta^-)$ is sampled from the dataset, and \hat{a}^ϵ for the Q-function guidance term is generated from the consistency policy. The above proves the formulation of the final optimization objective. \square

A.2 The Proof of Reconstruction Loss

Let $\Delta k = \max_{m \in [1, M-1]} \{|k_{m+1} - k_m|\}$. We have the following assumptions:

- Assume we have a pre-trained score function represents the desired policy: $\forall k \in [\epsilon, K] : s_{\phi^*}(a^k, k) = \nabla \log p^k(a^k)$, which cannot be accessed. And let $f(a^k, k; \phi^*|s)$ be the ground truth conditional consistency policy of this score model.

Then, if $L_{RC}(\theta, \theta^-) = 0$, we have:

$$L_{CT}(\theta, \theta^-) = L_{RC}(\theta, \theta^-)$$

PROOF. From $L_{RC} = 0$, we have:

$$\mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), a)] \equiv 0 \tag{40}$$

where the expectation is taken with respect to $a \sim p_{data}$, $m \sim \mathcal{U}[1, M-1]$, $z \sim \mathcal{N}(0, 1)$.

Considering that $p(a) > 0$ for any $a \in p_{data}$, it entails:

$$d(f_\theta(a + k_m z, k_m|s), a) \equiv 0 \tag{41}$$

for $m \in [1, M-1]$.

With respect to $d(x, y) = 0 \Leftrightarrow x - y = 0$, we have:

$$\begin{aligned}
f_\theta(a + k_{m+1}z, k_{m+1}|s) &\equiv a \\
&\equiv f(a + k_{m+1}z, k_{m+1}; \phi^*|s) \\
&= f(a + k_m z, k_m; \phi^*|s)
\end{aligned} \tag{42}$$

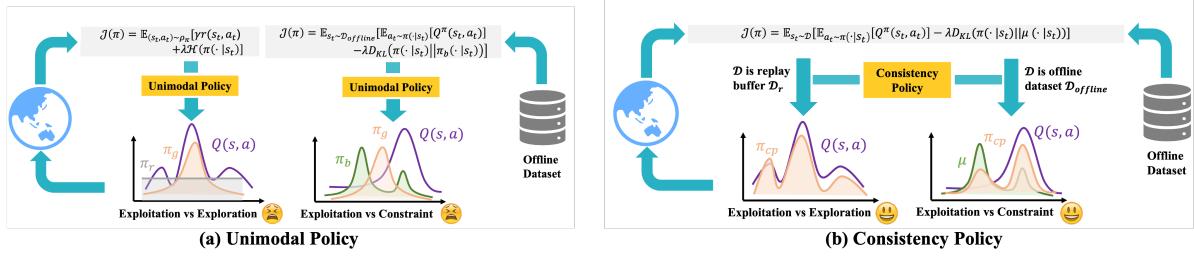


Figure 6: Illustration of online and offline RL with (a) previously used unimodal policy and (b) our proposed consistency policy.

Then we have:

$$\begin{aligned}
 L_{CT}(\theta, \theta^-) &= \mathbb{E}[d(f_\theta(a + k_{m+1}z, k_{m+1}|s), f_{\theta^-}(a + k_m z, k_m|s))] \\
 &= \mathbb{E}[d(f(a + k_m z, k_m; \phi^*|s), f_{\theta^-}(a + k_m z, k_m|s))] \\
 &= \mathbb{E}[d(f_{\theta^-}(a + k_m z, k_m|s), a)] \\
 &= L_{RC}(\theta, \theta^-)
 \end{aligned} \tag{43}$$

□

B CONSISTENCY POLICY FOR OFFLINE RL AND ONLINE RL

We provide an illustrative description of how CPQL works for both offline and online RL tasks, as shown in Figure 6. One one hand, previously used unimodal policy, such as Gaussian policy π_g , has to make the trade-off between better performance and diverse behaviors. The Gaussian policy learns with extra exploration policy π_e for online tasks and struggles with constraints from behavior policy π_b for offline tasks. On the other hand, the consistency policy π_{cp} has a powerful ability to represent complex data distribution. The policy μ refers to the behavior policy under offline scenario while representing the policy π_r for collecting the data in replay buffer \mathcal{D}_r under online scenario. This consistency policy concurrently handles greedy behavior and diverse exploratory behaviors or policy constraints in a seamless manner.

C EXPERIMENT SETUP

In this section, we provide the setup for all experiments tested in the paper and the source code we used for reporting the results of other methods, such as TD3, SAC, PPO, DIPO, and Diffusion-QL.

The consistency policy is based on a multilayer perceptron (MLP) that takes the state as input and provides the corresponding action as output. For the policy networks, we establish the backbone network architecture to be a 3-layer MLP (hidden size 256) with the Mish function serving as the activation function. The value networks also consist of a 3-layer MLP with Mish function, while additionally using LayNorm between each layer.

Table 4: The hyperparameters of all selected tasks for CPQL and CPIQL.

Tasks	α	η	τ	learning rate
All online tasks	0.05	1	-	$3 * 10^{-4}$
halfcheetah-medium-v2	1	2.0	0.7	$3 * 10^{-4}$
hopper-medium-v2	1	2.0	0.6	$3 * 10^{-4}$
walker2d-medium-v2	1	1.0	0.6	$3 * 10^{-4}$
halfcheetah-medium-replay-v2	1	3.0	0.7	$3 * 10^{-4}$
hopper-medium-replay-v2	1	1.0	0.6	$3 * 10^{-4}$
walker2d-medium-replay-v2	1	1.0	0.6	$3 * 10^{-4}$
halfcheetah-medium-expert-v2	1	1.0	0.7	$3 * 10^{-4}$
hopper-medium-expert-v2	1	0.5	0.6	$3 * 10^{-4}$
walker2d-medium-expert-v2	1	1.0	0.6	$3 * 10^{-4}$
pen-human-v1	1	0.15	0.7	$3 * 10^{-5}$
pen-cloned-v1	1	0.1	0.7	$3 * 10^{-5}$

For the consistency policy, we set $k \in [0.002, 80.0]$ and the number of sub-intervals $M = 40$. Following Karras diffusion model [25], the sub-interval boundaries are determined with the formula $k_i = (\epsilon^{\frac{1}{\rho}} + \frac{i-1}{M-1}(T^{\frac{1}{\rho}} - \epsilon^{\frac{1}{\rho}}))^{\rho}$, where $\rho = 7$. And we use the Euler solver as the ODE solver. For the offline scenario, we train for 1000 epochs (2000 for Locomotion tasks), each consisting of 1000 gradient steps with batch size 256 on D4RL tasks. For the online scenario, we train for 1M iterations on Gym MuJoCo tasks and 500k iterations on dm_control tasks. For learning rate, we set $3 * 10^{-5}$ for Adroit tasks of D4RL tasks and $3 * 10^{-5}$ for all other tasks. And we use a fixed learning rate, $3 * 10^{-4}$, for Q-function networks. Considering the α and η , for online tasks such as Gym MuJoCo tasks and dm_control tasks, we set $\alpha = 0.05$ and $\eta = 1$. For offline tasks such as D4RL tasks, we set $\alpha = 1$ and give the optimal η for different tasks. Typically for CPIQL on offline tasks, we provide the hyperparameter τ for the expectile regression function for different tasks.

The above-mentioned hyperparameters we used for different tasks are shown in Table 4. And the Python codes for reporting the results of TD3, SAC, PPO, DIPO, EDP, and Diffusion-QL are the following:

- TD3: <https://github.com/sfujim/TD3>;
- SAC: <https://github.com/toshikwa/soft-actor-critic.pytorch>;
- PPO: <https://github.com/ikostrikov/pytorch-a2c-ppo-acktr-gail>;
- DIPO: <https://github.com/BellmanTimeHut/DIPO>;
- EDP: <https://github.com/sail-sg/edp>;
- Diffusion-QL: <https://github.com/Zhendong-Wang/Diffusion-Policies-for-Offline-RL>.

Table 5: The performance of Diffusion-QL, CPQL and CPIQL on D4RL Locomotion and Adroit tasks. For CPQL and CPIQL, we report the mean and standard deviation of normalized scores across five random seeds. The results of Diffusion-QL(best) are from the paper of Diffusion-QL [49], and the results of Diffusion-QL(final) are from the paper of QGPO [34].

Dataset	Diffusion-QL(best)	Diffusion-QL(final)	CPIQL(best)	CPIQL(final)	CPQL(best)	CPQL(final)
halfcheetah-medium-v2	51.1	50.6	55.3 ± 0.4	54.6 ± 1.0	57.9 ± 0.3	56.9 ± 0.9
hopper-medium-v2	90.5	82.4	101.5 ± 2.1	99.7 ± 2.0	102.1 ± 2.4	99.9 ± 4.5
walker2d-medium-v2	87.0	85.1	88.4 ± 0.5	86.2 ± 0.6	90.5 ± 2.2	82.1 ± 2.4
halfcheetah-medium-replay-v2	47.8	47.5	49.8 ± 1.0	48.0 ± 1.4	48.1 ± 0.7	46.6 ± 0.8
hopper-medium-replay-v2	101.3	100.7	101.7 ± 0.6	100.6 ± 1.5	101.7 ± 0.9	97.7 ± 4.6
walker2d-medium-replay-v2	95.5	94.3	95.0 ± 0.6	91.8 ± 2.8	94.4 ± 1.3	93.6 ± 5.6
halfcheetah-medium-expert-v2	96.8	96.1	90.2 ± 1.3	81.0 ± 1.7	98.8 ± 0.4	97.8 ± 0.5
hopper-medium-expert-v2	111.1	110.7	113.4 ± 0.8	110.6 ± 1.4	114.2 ± 0.4	110.4 ± 3.2
walker2d-medium-expert-v2	110.1	109.7	112.3 ± 0.3	110.9 ± 0.2	111.5 ± 0.1	110.9 ± 0.1
Average	87.9	86.3	89.7	87.0	91.0	88.4
pen-human-v1	72.8	-	58.2 ± 2.7	48.0 ± 8.5	89.3 ± 6.9	56.7 ± 4.9
pen-cloned-v1	57.3	-	77.4 ± 6.9	57.4 ± 5.0	83.3 ± 6.1	65.3 ± 2.5
Average	65.1	-	67.8	52.7	86.3	61.0

Table 6: The original training IPS and sampling SPS of CPQL and other methods on Gym MuJoCo tasks using Nvidia 2080Ti.

-	Diffusion-QL (PyTorch)	DIPO (PyTorch)	CPQL (PyTorch)	Diffusion-QL (Jax)	EDP (Jax)	CPQL (Jax)
IPS	4.7	8.2	68.0	22.3	116.2	325.2
SPS	15.1	17.0	680.2	123.7	411.9	1069.7

D MORE EXPERIMENT RESULTS

In this section, we provide more experiment results for D4RL tasks under the offline scenario, as shown in Table 5. Different from Table 3 in the paper, we here provide 2 columns (best and final) for CPQL and CPIQL, where "best" represents the best performance during training with online evaluation, and "final" refers to the performance at the end of training. In most cases, the final performance can reach the best performance unless the training is unstable, leading to performance collapse in the later training process. In this paper, we choose the "best" scores for method comparison.

Also, we provide the IPS for training and SPS for inference for chosen methods are provided in Table 6.

E TRAINING CURVES FOR OFFLINE AND ONLINE REINFORCEMENT LEARNING

In this section, we provide curves of the training process for each selected task mentioned in this paper. We plot the mean and standard deviation of results across five random seeds as shown in Figure 8, Figure 9 and Figure 7.

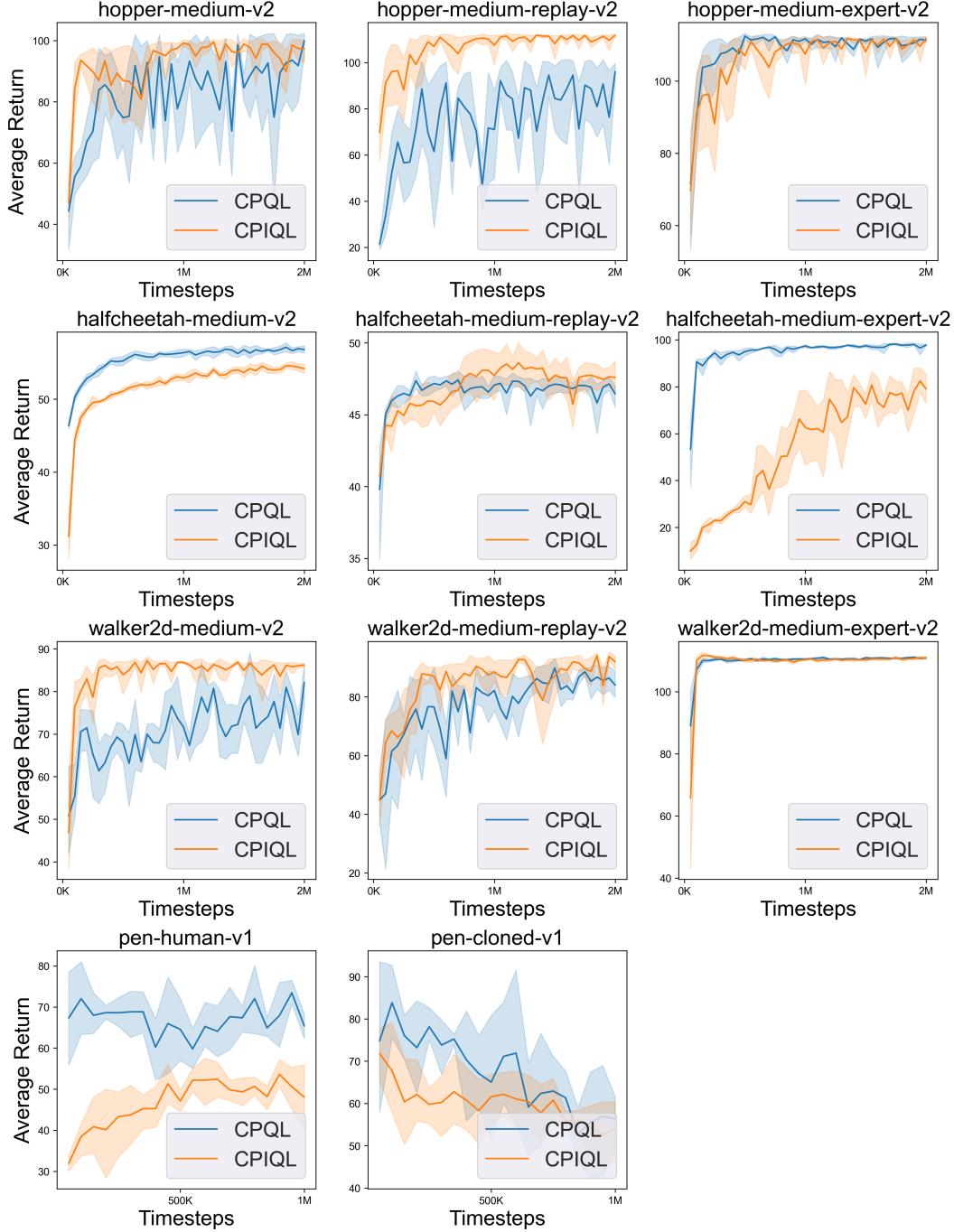


Figure 7: Training curves for D4RL tasks under the offline scenario.

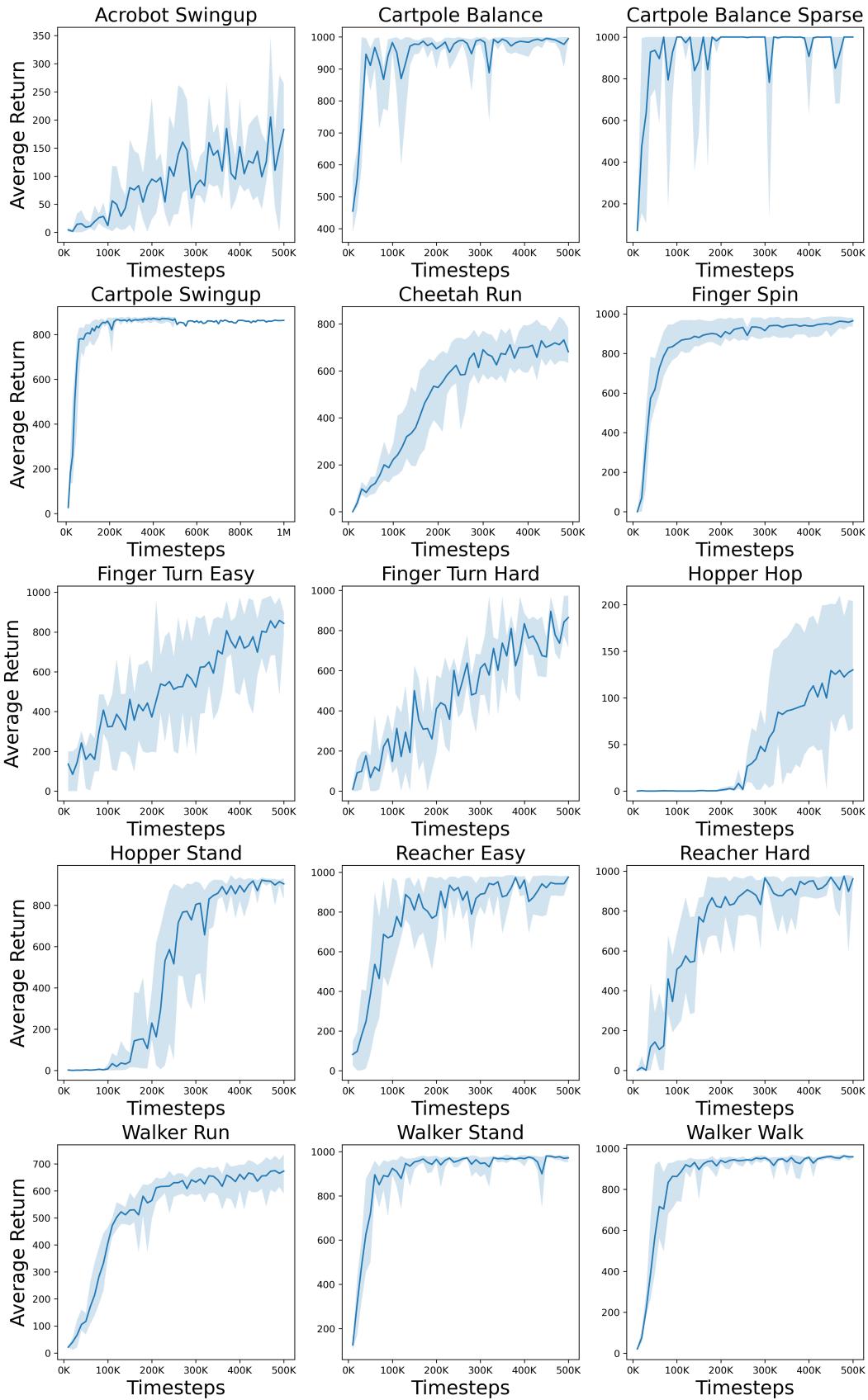


Figure 8: Training curves for dm_control tasks under the online scenario

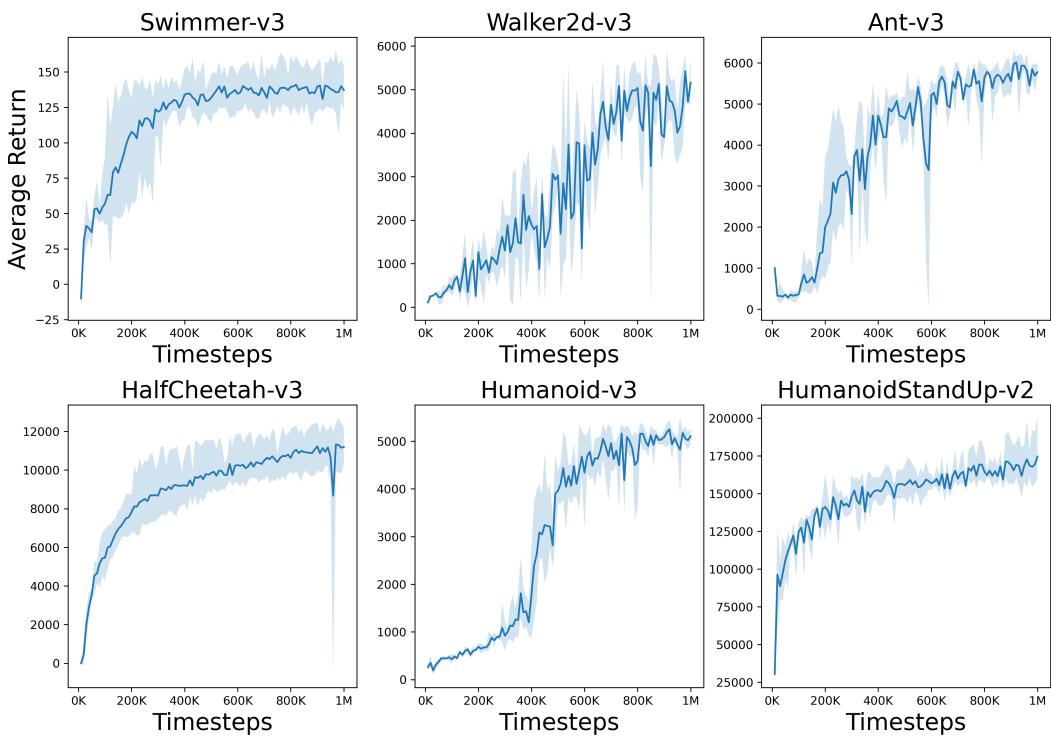


Figure 9: Training curves for Gym MuJoCo tasks under the online scenario.