

DOMAIN: mildly cOnservative Model-bAsed offlINe reinforcement learning

Xiao-Yin Liu, Xiao-Hu Zhou*, *Member, IEEE*, Xiao-Liang Xie, *Member, IEEE*, Shi-Qi Liu, Zhen-Qiu Feng, Hao Li, Mei-Jiang Gui, Tian-Yu Xiang, De-Xing Huang, Zeng-Guang Hou*, *Fellow, IEEE*

Abstract—Model-based reinforcement learning (RL), which learns environment model from offline dataset and generates more out-of-distribution model data, has become an effective approach to the problem of distribution shift in offline RL. Due to the gap between the learned and actual environment, conservatism should be incorporated into the algorithm to balance accurate offline data and imprecise model data. The conservatism of current algorithms mostly relies on model uncertainty estimation. However, uncertainty estimation is unreliable and leads to poor performance in certain scenarios, and the previous methods ignore differences between the model data, which brings great conservatism. Therefore, this paper proposes a mildly cOnservative Model-bAsed offlINe RL algorithm (DOMAIN) without estimating model uncertainty to address the above issues. DOMAIN introduces adaptive sampling distribution of model samples, which can adaptively adjust the model data penalty. In this paper, we theoretically demonstrate that the Q value learned by the DOMAIN outside the region is a lower bound of the true Q value, the DOMAIN is less conservative than previous model-based offline RL algorithms and has the guarantee of security policy improvement. The results of extensive experiments show that DOMAIN outperforms prior RL algorithms on the D4RL dataset benchmark, and achieves better performance than other RL algorithms on tasks that require generalization.

Index Terms—Model-based offline reinforcement learning; Mildly conservative; Adaptive sampling distribution

I. INTRODUCTION

REINFORCEMENT learning (RL) aims to maximize user-defined rewards and acquire optimal behavioral skill. Since the 1970s, RL research has attracted widespread attention, and the advent of deep neural networks has greatly facilitated its development. Today, RL has been applied in various

fields, including robot control [1], [2], autonomous driving [3], recommendation systems [4], and resource scheduling [5].

The paradigm of online RL involves an agent interacting with environment directly through trial-and-error [6]. However, online RL suffers from low sampling efficiency and high costs, making it impractical in certain scenarios [7]. For instance, in situations such as surgery and autonomous driving, real-time interaction between the agent and the environment may pose a threat to human safety. Since the robot may damage its components or surrounding objects, the cost of trial-and-error can be prohibitive in robot control [8]. Offline RL can effectively address these issues by learning policy from a static dataset without direct interaction with the environment.

Offline RL faces a significant challenge due to distribution shift, which is caused by differences between the policies in offline dataset and those learned by the agent. The selection of out-of-distribution (OOD) actions introduces extrapolation errors that may lead to unstable training of the agent [9]. To address the above issue, two types of methods have been developed: policy constraints [10]–[12] and value penalization [13], [14]. Huang *et al.* [40] proposed a mild policy evaluation method by constraining the difference between the Q values of actions supported by target policy and those of actions contained within offline dataset. Yang *et al.* [41] treated different samples with different policy constraint intensities, which effectively improves the agent performance. The goal of the above methods is to minimize the discrepancy between the learned policy and that in the dataset. While they achieve remarkable performance, these methods fail to consider the effects of OOD actions, hampering the exploration and performance improvement. In contrast, Lyu *et al.* [15] proved that the agent performance can be effectively enhanced by exploring OOD region.

The aforementioned offline RL methods are model-free, which constrains exploration in OOD region, hindering performance improvements for agents. In contrast, model-based offline RL enables agents to continually interact with a trained environment model, generating data with broader coverage, thereby enhancing exploration in OOD region [16]. Moreover, agents acquire imaginative abilities through environmental modeling, enabling effective planning and improving control efficacy [17], [18]. Since the uncertainty of environment model, exploration by agents also carries risks. It is critical to strike a balance between return and risk.

Performing conservative policy optimization has become an effective way to the above problem. To incorporate the conservatism into policy, the typical method is estimating model

This work was supported in part by the National Natural Science Foundation of China under Grant 62003343, Grant 62222316, Grant U1913601, Grant 62073325, Grant 61720106012, Grant 62003198, Grant U20A20224, and Grant U1913210; in part by the Beijing Natural Science Foundation under Grant M22008; in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2020140. (*Corresponding authors: Xiao-Hu Zhou and Zeng-Guang Hou).

Xiao-Yin Liu, Xiao-Hu Zhou, Xiao-Liang Xie, Shi-Qi Liu, Zhen-Qiu Feng, Hao Li, Mei-Jiang Gui, Tian-Yu Xiang and De-Xing Huang are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: liuxiaoyin2023@ia.ac.cn, xiaohu.zhou@ia.ac.cn).

Zeng-Guang Hou is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with CASIA-MUST Joint Laboratory of Intelligence Science and Technology, Institute of Systems Engineering, Macau University of Science and Technology, Macao, China. (e-mail: zengguang.hou@ia.ac.cn).

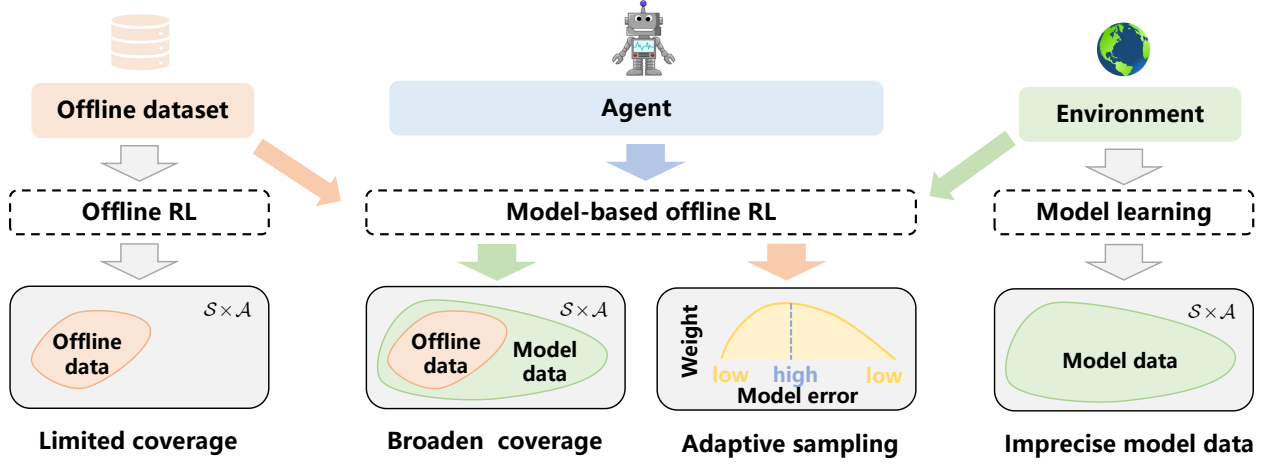


Fig. 1. The framework of mildly conservative model-based offline RL. This method integrates offline RL with model-based RL, aiming to leverage the model to broaden data coverage. The key lies in designing a model data adaptive sampling distribution based on the magnitude of model errors, which enables the adaptive adjustment of model data penalty.

uncertainty [19], [20], [39] or penalizing state-action pairs with high uncertainty [24], [25] during training. Experimental results have shown that model-based offline RL methods outperform most current model-free offline RL methods. However, measuring model uncertainty through neural networks results in prediction errors and low reliability. To address this issue, Yu *et al.* [21] proposed the COMBO algorithm based on CQL [13], which regularizes model data to suppress unreliable data and enhance reliable offline data. Rigter *et al.* [24] and Bhardwaj *et al.* [25] used an adversarial training framework, forcing the agent to act conservatively in OOD region.

However, the above methods ignore differences between the model data and are overly conservative. Therefore, we design an adaptive sampling distribution of model data, and propose a mildly conservative model-based offline RL algorithm (DOMAIN) without estimating model uncertainty (as illustrated in Fig. 1). Here, the mild conservatism is reflected in the following: higher exploration in OOD region and better performance of the agent. This algorithm imposes heavier penalties on model data with high errors while reducing the penalty for accurate model data to alleviate conservatism. Our main contributions are summarized as follows:

- 1) A new model-based offline RL method is proposed that adaptively adjusts the model data penalty, named DOMAIN, where the adaptive sampling distribution of model samples is designed. To our best knowledge, this is the first to incorporate adaptive sampling distribution to conservative value estimation in model-based offline RL.
- 2) Theoretical analyses demonstrate that the value function learned by the DOMAIN in the OOD region is a lower bound of the true value function, the DOMAIN is less conservative than previous offline model-based RL algorithms, and DOMAIN has a safety policy improvement guarantee.
- 3) Extensive experiments indicate that DOMAIN outperforms prior RL algorithms on the D4RL dataset bench-

mark, and achieves the best performance than other RL algorithms on tasks that require generalization.

The framework of this paper is as follows: Section 2 introduces the fundamental theory of RL. Section 3 shows the framework and implementation details of the DOMAIN algorithm. Section 4 gives some relevant theoretical analysis. Section 5 presents the algorithm performance on D4RL and generalization tasks. Finally, Section 6 summarizes the entire work and provides directions for future research.

II. PRELIMINARIES AND RELATED WORKS

A. Reinforcement Learning

The framework of RL is based on the Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, T_{\mathcal{M}}, \rho_0, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $r(s, a)$ is the reward function, $T_{\mathcal{M}}(s' | s, a)$ represents the dynamic system under $\mathcal{M}$, ρ_0 is the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor [8]. The goal of RL is to learn a policy that maximizes the cumulative reward $J(\mathcal{M}, \pi) = \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi}(s,a)} [r(s, a)] / (1 - \gamma)$, where $d_{\mathcal{M}}^{\pi}(s, a) = d_{\mathcal{M}}^{\pi}(s) \pi(a | s)$ is the marginal distribution of states and actions under the learned policy π , and $d_{\mathcal{M}}^{\pi}(s)$ is the discounted marginal state distribution under the learned policy π .

Considering the Bellman operator $\mathcal{T}^{\pi}Q(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim T_{\mathcal{M}}(s' | s, a)} [\max_{a' \in \mathcal{A}} Q^{\pi}(s', a')]$, its sample-based counterpart is $\hat{\mathcal{T}}^{\pi}Q(s, a) := r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s', a')$ [21]. The Actor-Critic algorithm minimizes the Bellman error to learn the Q -function and maximizes the Q -function to optimize the policy π . The formulas are expressed as follows:

$$\begin{aligned} \hat{Q} &\leftarrow \arg \min_Q \mathbb{E}_{s,a,s' \sim \mathcal{D}} \left[\left(Q(s, a) - \hat{\mathcal{T}}^{\pi}Q(s, a) \right)^2 \right] \\ \hat{\pi} &\leftarrow \arg \max_{\pi} \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi} \left[\hat{Q}(s, a) \right] \end{aligned} \quad (1)$$

where $\mathcal{D}$ can be a fixed offline dataset or replay buffer generated by the current policy $\hat{\pi}$ interacting with the environment.

B. Offline RL

The goal of offline RL is to learn a policy $\pi_\theta(\cdot | s)$ from an offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ without interacting with the environment, to maximize the expected discounted reward. However, there exists the distribution shift between the learned policy $\pi_\theta(\cdot | s)$ and the behavioral policy $\pi_b(\cdot | s)$ of the dataset, leading to unstable training of the agent [8]. To alleviate this issue, previous studies [10]–[14] have introduced regularization on the value function of OOD actions to reduce the probability of selecting OOD actions. Here primarily focuses on CQL [13]. The formulas for policy improvement and policy evaluation are as follows:

$$\begin{aligned} \hat{Q} &\leftarrow \arg \min_Q \frac{1}{2} \mathbb{E}_{s,a,s' \sim \mathcal{D}} \left[\left(Q(s,a) - \hat{T}^\pi \hat{Q}(s,a) \right)^2 \right] \\ &\quad + \beta \left\{ \mathbb{E}_{s \sim \mathcal{D}, a \sim \mu(\cdot | s)} [Q(s,a)] - \mathbb{E}_{s,a \sim \mathcal{D}} [Q(s,a)] \right\} \quad (2) \\ \hat{\pi} &\leftarrow \arg \max_{\pi} \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi_k} [\hat{Q}(s,a)] \end{aligned}$$

where β represents the regularization coefficient, $\mu(\cdot | s)$ denotes the distribution from which actions are sampled, which can be a uniform distribution or a learned policy. CQL effectively penalizes the Q -values in the OOD region, providing a conservative estimate of the value function and alleviating the challenges of overestimation and distribution shift.

C. Model-based Offline RL

Model-based offline RL refers to training an environment model $\hat{T}_\phi(s', r | s, a)$ using the offline dataset, where the model is typically parameterized as gaussian distribution $\hat{T}_\phi(s', r | s, a) = \mathcal{N}[\mu_\phi(s, a), \Sigma_\phi(s, a)]$. The loss function for training the environment model employs maximum likelihood estimation as: $\mathcal{L}_M(\phi) = -\mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} [\log \hat{T}_\phi(s', r | s, a)]$. With this learned model, an estimated MDP $\hat{\mathcal{M}}$ can be obtained. Subsequently, the agent interacts with the learned policy $\hat{\pi}$ and the environment model $\hat{T}_\phi$ to generate model data $\mathcal{D}_{\text{model}}$ through H -step rollouts. The policy learning is then performed using the combined dataset $\mathcal{D}_{\text{offline}} \cup \mathcal{D}_{\text{model}}$, where the data is sampled from model data $\mathcal{D}_{\text{model}}$ with the probability f , $f \in [0, 1]$, and from the offline data $\mathcal{D}_{\text{offline}}$ with the probability $1 - f$.

However, offline dataset fails to cover the entire state-action space, resulting in disparities between the learned environment model $\hat{T}_\phi(s', r | s, a)$ and the actual environment model $T(s', r | s, a)$, which leads to prediction errors in the model. To address these challenges, Janner *et al.* [23] employed a model ensemble approach to reduce biases introduced by probabilistic networks. Yu *et al.* [19] proposed the MOPO algorithm, which used model uncertainty as reward penalizes. The MOREL algorithm [20], similar to MOPO, incorporated penalty terms in rewards for OOD region data. Chen *et al.* [39] proposed offline model-based adaptable policy learning method that can adapt its behavior in out-of-support regions instead of learning in in-support regions. Furthermore, Yu *et al.* [21] considered the estimation errors of neural networks for model uncertainty and proposed the COMBO algorithm. This method utilizes value function regularization, aiming to

push down unreliable model data and push up reliable offline data.

III. METHODOLOGY

Incorporating model uncertainty as a penalty in reward can alleviate the impact of imprecise model data. However, accurately estimating model uncertainty remains a challenging task. The COMBO algorithm effectively achieves a balance between return and risk by incorporating regularization based on CQL without estimating model uncertainty. Nonetheless, prevailing methodologies ignore the correlation between the magnitude of model errors and the sampling distribution. Consequently, such approaches tend to adopt overly conservative policies. Therefore, based on COMBO, a mildly conservative algorithm for model-based offline RL (DOMAIN) is proposed in this study.

A. Mildly Conservative Value Estimation

Conservative policy evaluation within the framework of DOMAIN algorithm can be described as follows: Given an offline dataset $\mathcal{D}_{\text{offline}} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$, a learned policy $\hat{\pi}$, and a learned environmental model $\hat{T}_\phi$, the primary goal is to obtain a conservative estimation of $Q(s, a)$. To accomplish this, the policy evaluation step in the DOMAIN employs the CQL framework (see Eq. (2)). This principle can be mathematically expressed as:

$$\begin{aligned} \min_Q \max_p \mathcal{L}_\beta \left(Q, \hat{T}^\pi \hat{Q} \right) + \lambda \left\{ \mathbb{E}_{s,a \sim \rho(s,a)} [Q(s,a)] \right. \\ \left. - \mathbb{E}_{s,a \sim \mathcal{D}_{\text{offline}}} [Q(s,a)] + \mathcal{H}(\rho) \right\} \quad (3) \end{aligned}$$

where λ is a scaling factor used to adjust the weight of the regularization term, $\rho(s, a)$ denotes the sampling distribution of the model data, and $\mathcal{H}(\rho)$ is a constraint term for $\rho(s, a)$. The term $\mathbb{E}_{s,a \sim \rho(s,a)} [Q(s,a)]$ is employed to penalize unreliable model data, while $\mathbb{E}_{s,a \sim \mathcal{D}_{\text{offline}}} [Q(s,a)]$ is used to reward reliable offline data. The term $\mathcal{L}_\beta(Q, \hat{T}^\pi \hat{Q})$ represents the Bellman error. In the Actor-Critic framework, the goal of policy evaluation is to minimize the Bellman error [26]. In the DOMAIN algorithm, the Bellman error consists of two components: the Bellman error from the offline data and the model data. It can be expressed as follows:

$$\begin{aligned} \mathcal{L}_\beta \left(Q, \hat{T}^\pi \hat{Q} \right) \\ = \frac{1}{2} \left\{ (1-f) \mathbb{E}_{s,a,s' \sim \mathcal{D}_{\text{offline}}} \left[\left(Q - \hat{T}^\pi \hat{Q} \right)(s,a) \right]^2 \right. \\ \left. + f \mathbb{E}_{s,a,s' \sim \mathcal{D}_{\text{model}}} \left[\left(Q - \hat{T}^\pi \hat{Q} \right)(s,a) \right]^2 \right\} \quad (4) \end{aligned}$$

where f can be determined by two guidelines. One is the error of model data, and the other is the algorithm itself. If the model data is accurate and the algorithm can effectively penalize model data with large errors, f can be set larger. On the contrary, if the error of model data is large and the algorithm fails to effectively penalize them, f needs to be set smaller to guarantee the stability and security of the agent.

The value function Q and the policy π in Eq. (4) can be approximated using neural networks, denoted as Q_ω and

π_θ , respectively. $\hat{T}^\pi$ represents the actual Bellman operator used in the algorithm, while $\mathcal{T}_{\hat{\mathcal{M}}}^\pi$ and $\mathcal{T}_{\mathcal{M}}^\pi$ represent the operator applied under the empirical MDP and the learned model, respectively. The value function update follows the SAC framework [27], and thus the Bellman operator can be further expressed as:

$$\hat{T}^\pi \hat{Q}(s, a) \approx r + \gamma \left(\min_{i=1,2} Q_{\omega_i}(s', a') - \alpha \log \pi_\theta \right) \quad (5)$$

where α is the entropy regularization coefficient, determining the relative importance of entropy. By employing a mildly conservative policy evaluation, the policy improvement scheme can be obtained:

$$\pi_\theta = \arg \max_{\theta} \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi_\theta} \left[\min_{i=1,2} Q_{\omega_i}(s, a) - \alpha \log \pi_\theta \right] \quad (6)$$

Here, entropy regularization is employed to prevent policy degradation. The two Q value networks are utilized to select the minimum value to address overestimation in agent updates [38]. In Eq. (3), the constraint term $\mathcal{H}(\rho)$ for $\rho(s, a)$ is typically approximated using KL divergence to approach the desired sampling distribution $\omega(s, a)$, i.e., $\mathcal{H}(\rho) = -D_{KL}(\rho \parallel \omega)$. Thus, the following optimization problem can be formulated:

$$\begin{aligned} \max_{\rho} \mathbb{E}_{s, a \sim \rho(s, a)} [Q(s, a)] - D_{KL}(\rho \parallel \omega) \\ \text{s.t. } \int_S \int_{\mathcal{A}} \rho(s, a) ds da = 1, \rho(s, a) \geq 0 \end{aligned} \quad (7)$$

The above optimization problem can be solved using the Lagrange method [13], yielding the following solution:

$$\rho^*(s, a) = \omega(s, a) \exp[Q(s, a)] / Z \quad (8)$$

where Z is the normalization factor. Substituting the solution obtained from solving Eq. (7) back into Eq. (3), the following expression can be obtained:

$$\begin{aligned} \min_Q \mathcal{L}_\beta \left(Q, \hat{T}^\pi \hat{Q} \right) + \lambda \left\{ \log \mathbb{E}_{s, a \sim \omega(s, a)} \exp [Q(s, a)] \right. \\ \left. - \mathbb{E}_{s, a \sim \mathcal{D}} [Q(s, a)] \right\} \end{aligned} \quad (9)$$

By optimizing Eq. (9), a mildly conservative value estimation can be achieved, and if the sampling distribution of model data $\omega(s, a)$ is known, the above expression can be optimized directly. Moreover, $\omega(s, a)$ reflects the degree of penalization for model data. Note that mildly conservative value estimation for model data doesn't mean a more accurate value estimation. Since the errors in model data, the agent may explore dangerous areas. The penalty is often added to the Q value of the inaccurate model data to alleviate this problem, which causes that the estimated value is lower than the true value for inaccurate model data. The following part provides further derivation for model data sampling distribution.

B. Adaptive Value Regularization on Model Data

When performing policy evaluation, we want to assign higher penalty coefficients to larger error model data, thereby yielding an increased value for $\omega(s, a)$. Drawing inspiration from [28], [29], the KL divergence is employed to quantify the extent of model data errors:

$$g(s, a) = \mathbb{E}_{s' \sim T_{\hat{\mathcal{M}}}(s' | s, a)} \left[\frac{T_{\hat{\mathcal{M}}}(s' | s, a)}{T_{\mathcal{M}}(s' | s, a)} \right] \quad (10)$$

where $T_{\hat{\mathcal{M}}}$ represents the state transition model in the learned environment $\hat{\mathcal{M}}$, while $T_{\mathcal{M}}$ corresponds to that in the true environment $\mathcal{M}$. Since sampling distribution for model data needs to satisfy the equation: $\int_S \int_{\mathcal{A}} \omega(s, a) ds da = 1$, the sampling distribution $\omega(s, a)$ based on finite model dataset can be denoted as:

$$\omega(s, a) = \frac{g(s, a)}{\int_S \int_{\mathcal{A}} g(s', a') ds' da'} \quad (11)$$

Hence, to design an accurate sampling distribution, it is crucial to compute the dynamic ratio value $T_{\hat{\mathcal{M}}}/T_{\mathcal{M}}$ accurately. Following [28], the following equation can be derived by applying Bayes theorem:

$$\frac{T_{\hat{\mathcal{M}}}(s' | s, a)}{T_{\mathcal{M}}(s' | s, a)} = \frac{p(\text{model} | s, a, s') p(\text{offline} | s, a)}{p(\text{offline} | s, a, s') p(\text{model} | s, a)} \quad (12)$$

As shown in the above equation, it is easy to find the probabilities of the data pair (s, a, s') and (s, a) being model data and offline data, respectively, are the key factors. To estimate these probabilities, two discriminators are constructed, namely $D_{\Phi_{sas}}(\cdot | s, a, s')$ and $D_{\Phi_{sa}}(\cdot | s, a)$, which approximate the probabilities using parameters Φ_{sas} and Φ_{sa} , respectively. The network loss function employed for training the discriminators is the cross-entropy loss [30]:

$$\begin{aligned} \mathcal{L}(\Phi_{sas}) &= -\mathbb{E}_{\mathcal{D}_{\text{offline}}} [\log D_{\Phi_{sas}}(\text{offline} | s, a, s')] \\ &\quad -\mathbb{E}_{\mathcal{D}_{\text{model}}} [\log D_{\Phi_{sas}}(\text{model} | s, a, s')] \\ \mathcal{L}(\Phi_{sa}) &= -\mathbb{E}_{\mathcal{D}_{\text{offline}}} [\log D_{\Phi_{sa}}(\text{offline} | s, a)] \\ &\quad -\mathbb{E}_{\mathcal{D}_{\text{model}}} [\log D_{\Phi_{sa}}(\text{model} | s, a)] \end{aligned} \quad (13)$$

By training the discriminator networks, the gap between the learned and the actual environment can be estimated, leading to the estimation of the sampling distribution. Based on this, the different sampling probabilities among model data lead to adaptive value regularization for the model data.

C. Algorithm Details of DOMAIN

To incorporate conservatism into algorithm, the typical method is applying reward penalties for OOD data by uncertainty quantification directly [19], [20] or value function regularization indirectly [24], [25]. However, the above methods ignore differences between the model data, which may prevent the agent from choosing a better action since the Q value is over-conservative. DOMAIN imposes heavier penalties on model data with high errors while reducing the penalty for accurate model data to alleviate conservatism. By reducing the penalty on the Q value of model data with small errors, the agent may choose action a with higher reward r through $\arg \max_a Q(s, a)$, thus improving the performance of the learned policy.

Our algorithm DOMAIN, summarized in Algorithm 1, is based on SAC [27] and MOPO [19]. The training approach for the environment model closely resembles that of MOPO, employing a maximum likelihood estimation approach. N dynamic environment models are trained and denoted as $\{\hat{T}_\phi^i = \mathcal{N}[\mu_\phi^i(s, a), \Sigma_\phi^i(s, a)]\}_{i=1}^N$. Subsequently, a subset of M models with superior training accuracy is selected,

Algorithm 1: Mildly Conservative Model-Based Offline Reinforcement Learning

Input: offline dataset $\mathcal{D}_{\text{offline}}$, critic $Q_{\omega_1}, Q_{\omega_2}$, policy π_θ , discriminator $D_{\Phi_{sas}}, D_{\Phi_{sa}}$, environment model $\hat{T}_\phi$, rollout policy $\mu(\cdot | s)$ and rollout length H .

Output: learned policy $\pi_\theta(\cdot | s)$.

- 1: **Initialization:** target network $Q_{\bar{\omega}_1} \leftarrow Q_{\omega_1}$, $Q_{\bar{\omega}_2} \leftarrow Q_{\omega_2}$, model dataset $\mathcal{D}_{\text{model}} = \emptyset$.
 - 2: Use $\mathcal{D}_{\text{offline}}$ to train N environment model $\{\hat{T}_\phi^i = \mathcal{N}[\mu_\phi^i(s, a), \sum_{\phi}^i(s, a)]\}_{i=1}^N$.
 - 3: **for** $t = 1, 2, 3 \dots, T$ **do**
 - 4: Collect model data: rollout $(\mu(\cdot | s), \hat{T}_\phi, H)$, and add rollout data into $\mathcal{D}_{\text{model}}$;
 - 5: Update discriminator $D_{\Phi_{sas}}, D_{\Phi_{sa}}$ according to Eq. (13), and calculate adaptive sampling distribution;
 - 6: Update policy network π_θ according to Eq. (6);
 - 7: Update critic network $Q_{\omega_1}, Q_{\omega_2}$ according to Eq. (9);
 - 8: **if** $t \% \text{update period} = 0$:
 - 9: Soft update periodically:
 - 10: $\bar{\omega}_{1,2} \leftarrow \tau \omega_{1,2} + (1 - \tau) \bar{\omega}_{1,2}$;
 - 11: **end if**
 - 12: **end for**
-

and their predicted values are aggregated to yield the final prediction. Eq.s (5) and (6) adhere to the well-established SAC framework, with the parameter α being automatically adapted throughout the training procedure [27]. The weight λ of the regularization term and the rolling horizon H exert a substantial influence on the agent training.

It employs Eq. (9) to perform mildly conservative policy evaluation, Eq. (6) for policy improvement, and Eq. (13) to train the discriminator. Here, $g(s, a) = \mathbb{E}_{s' \sim \hat{T}_{\hat{\mathcal{M}}}} [T_{\hat{\mathcal{M}}}(s' | s, a) / T_{\mathcal{M}}(s' | s, a)]$ represents the expectation of the dynamic ratio. This expectation is estimated by sampling m states from a Gaussian distribution $\hat{T}_\phi(s', r | s, a) = \mathcal{N}[\mu_\phi(s, a), \sum_\phi(s, a)]$ obtained from the trained model. The computation of sampling distribution $\omega(s, a)$ for model data is based on the finite model dataset $\mathcal{D}_{\text{model}}$, which fails to cover all action space. Therefore, in practical implementation, the estimated sampling distribution $\omega(s, a)$ is approximated by $g(s, a) / \sum_{(s', a')} g(s', a')$.

Note that the above training of algorithm DOMAIN only depends on the offline dataset. If the offline dataset is a little, the agent performance is difficult to improve. The trained environment model is trained from the offline dataset. A small dataset will lead to a large gap between the true and estimated environments, which brings large model data errors. The proposed method will fail to achieve the balance between accurate offline data and inaccuracy model data. Therefore, the offline dataset cannot be too small and needs to cover a certain state-action space.

IV. THEORETICAL ANALYSIS OF DOMAIN

In this section, we analyze the DOMAIN algorithm and demonstrate that the adaptive sampling distribution of model data is equivalent to the adaptive adjustment of penalties

for rewards. Furthermore, we prove that the learned value function in DOMAIN serves as a lower bound for the true value function in OOD region, effectively suppressing the value function in highly uncertain region. In comparison to COMBO, DOMAIN exhibits weaker conservatism. By employing a lower bound optimization strategy, DOMAIN guarantees safety policy improvement.

A. Adaptive Adjustment of Penalties for Rewards

Before the theoretical analysis, we give the following theorem, which proves that the Bellman operator $\hat{T}^\pi$ can converge to a specific Q value.

Theorem 1: Let $\|\cdot\|_\infty$ is the $\mathcal{L}_\infty$ norm and $(\mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}, \|\cdot\|_\infty)$ is complete space. Then, the Bellman operator $\hat{T}^\pi$ is a γ -contraction mapping operator, i.e. $\|\hat{T}^\pi Q_1 - \hat{T}^\pi Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$.

The detailed proof can be found in Appendix A. Theorem 1 shows that the operator $\hat{T}$ is a γ -contraction mapping in the $\mathcal{L}_\infty$ norm. According to the fixed point theorem, any initial Q function can converge to a unique fixed point Q^* by repeatedly applying the Bellman operator $\hat{T}^\pi$. By setting the derivative of the Eq. (9) concerning Q^k to zero during the k -th iteration, the equation can be obtained:

$$\hat{Q}^{k+1}(s, a) = \left(\hat{T}^\pi \hat{Q}^k \right)(s, a) - \lambda \left[\frac{\omega(s, a) - d(s, a)}{d_\beta(s, a)} \right] \quad (14)$$

Here, the distribution for offline data sampling, denoted as $d(s, a)$, is commonly defined as $d(s, a) = d^{\pi_b}(s, a) = d^{\pi_b}(s) \pi_b(a | s)$, where $d^{\pi_b}(s, a)$ represents the marginal distribution of states and actions under the behavioral policy π_b . On the other hand, the data sampling distribution used for computing the Bellman error, denoted as $d_\beta(s, a)$, is defined as $d_\beta(s, a) = (1 - f) d^{\pi_b}(s, a) + f d_{\hat{\mathcal{M}}}^\pi(s, a)$. Here, $d_{\hat{\mathcal{M}}}^\pi(s, a) := d_{\hat{\mathcal{M}}}^\pi(s) \pi(a | s)$ represents the marginal distribution of states and actions under the learning policy π , and $d_{\hat{\mathcal{M}}}^\pi(s)$ represents the discounted marginal state distribution when executing policy π in the learned model $\hat{\mathcal{M}}$.

Let us designate the second term in Eq. (14) as $\eta(s, a)$, which serves as the adjustment of penalties for rewards. When the model data (s, a) exhibits high errors, the value of the designed sampling distribution $\omega(s, a)$ increases. This indicates that the data point (s, a) either resides in the OOD region or a region with a relatively low density of offline dataset, implying that $d(s, a)$ approaches zero. Consequently, $\omega(s, a) > d(s, a)$, resulting in $\eta(s, a) > 0$ that serves as a penalty term. Moreover, the magnitude of penalization in the Q value increases with larger errors in the model data. In contrast, when the model data error is small, suggesting that the data is situated in a region proximal to true environment, it is possible for $\omega(s, a)$ to be smaller than $d(s, a)$, leading to $\eta(s, a) < 0$ which acts as a reward term. This encourages exploration in accurate model data region during training process.

The adaptive adjustment of reward is reflected as follows: During the value iteration process, the adjustment term for reward, $\eta(s, a)$, adapts the rewards based on the magnitude of $\omega(s, a)$ (assuming the offline data set is fixed and thus $d(s, a)$ can be considered fixed as well). This adaptive adjustment

allows for penalizing model data with high errors and reduces the conservatism of the algorithm.

B. DOMAIN Optimizes a Low Bound Value

Considering the influence of sampling errors and model errors, we present the conditions for the existence of a lower bound on the value function in the DOMAIN, demonstrating that DOMAIN does not underestimate the value function for all states. Additionally, we discuss a sufficient condition under which the lower bound of DOMAIN is tighter than that of COMBO, indicating that DOMAIN exhibits weaker conservatism on offline dataset.

Theorem 2: The value function learned using Eq. (9) serves as a lower bound on the true value function, i.e. $V^\pi(s) \geq \hat{V}^\pi(s)$, given that the following conditions are satisfied: $\int_{\mathcal{A}} \omega(s, a) da > \xi(s) \int_{\mathcal{A}} d(s, a) da$ and $\lambda \geq \delta_l$, where the state coefficient $\xi(s)$ and δ_l are defined as:

$$\begin{aligned} \xi(s) &= \frac{(1-f)d^{\pi_b}(s) \max_a [\pi_b(a|s)/\pi(a|s)] + f d_{\mathcal{M}}^\pi(s)}{(1-f)d^{\pi_b}(s) \min_a [\pi_b(a|s)/\pi(a|s)] + f d_{\mathcal{M}}^\pi(s)} \geq 1 \\ \delta_l &= \frac{(1-f) \frac{C_{r,T,\delta} R_{\max}}{(1-\gamma) \min_{s,a} \{\sqrt{|\mathcal{D}|}\}}}{\min_s \left\{ \frac{(\xi(s)-1) \int_{\mathcal{A}} d(s,a) da}{[(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + f d_{\mathcal{M}}^\pi(s)]} \right\}} \\ &\quad + \frac{f[\max_{s,a} \{r_{\mathcal{M}} - r_{\mathcal{M}}\} + \frac{2\gamma R_{\max}}{1-\gamma} \max_{s,a} \{D_{TV}(T_{\mathcal{M}}, T_{\mathcal{M}})\}]}{\min_s \left\{ \frac{(\xi(s)-1) \int_{\mathcal{A}} d(s,a) da}{[(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + f d_{\mathcal{M}}^\pi(s)]} \right\}} \end{aligned} \quad (15)$$

The detailed proof can be found in Appendix B. From the theorem, it can be concluded that for a sufficiently large λ and when the aforementioned conditions are met, $V^\pi(s) \geq \hat{V}^\pi(s)$ holds. When there is a sufficient amount of data, δ_l primarily stems from model errors, $C_{r,T,\delta} R_{\max}/((1-\gamma) \min_{s,a} \{\sqrt{|\mathcal{D}|}\})$ approaching zero. Therefore, when the model error is small and the model data ratio f is appropriately set, a sufficiently small λ can ensure $V^\pi(s) \geq \hat{V}^\pi(s)$.

Furthermore, DOMAIN does not underestimate the value function for all states s . It can be inferred that the value function is underestimated only when the condition in Eq. (15) is satisfied. However, there is a possibility of overestimating the data with high probabilities in the distribution $d(s, a)$ of the offline data set. Since the data points (s, a) with a higher frequency of occurrence in region $\mathcal{D}_{\text{offline}}$ are considered as in-distribution data, the learned environment model $\hat{T}_\phi$ exhibits smaller prediction errors for such data, resulting in smaller values of $\omega(s, a)$. As a result, it becomes challenging to satisfy the condition $\int_{\mathcal{A}} \omega(s, a) da > \xi(s) \int_{\mathcal{A}} d(s, a) da$, leading to overestimation. This aligns with the adaptive adjustment of penalties for reward.

Theorem 3: Let κ_1 and κ_2 denote the average value functions of DOMAIN and COMBO methods respectively. Specifically, $\kappa_1 = \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)}[\hat{Q}_1^\pi]$, $\kappa_2 = \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)}[\hat{Q}_2^\pi]$. It holds that $\kappa_1 > \kappa_2$ if the following condition is satisfied:

$$\mathbb{E}_{s \sim d(s)} \left[d_{\mathcal{M}}^\pi(s) \right] \geq \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)} \left[\omega(s, a) \right] \quad (16)$$

where $\hat{Q}_1^\pi$ and $\hat{Q}_2^\pi$ are the Q -value function of DOMAIN and COMBO in learned policy π respectively. $d(s)$ represents the state distribution of any data set, and $d_{\mathcal{M}}^\pi(s)$ represents the distribution of states under the learned model. The detailed proof of this theorem can be found in Appendix C.

When the data set corresponds to an offline dataset, the data points (s, a) with higher densities in $d(s, a)$ exhibit smaller model prediction errors, implying $w(s, a)$ approaches zero. The generated model data is more likely to be concentrated in-distribution region, indicating $d_{\mathcal{M}}^\pi(s)$ is large. Hence, Eq. (16) is readily satisfied, indicating that the conservatism of DOMAIN is weaker than that of COMBO.

C. DOMAIN Guarantees the Safety of Policy Improvement

The goal of DOMAIN is to learn an optimal policy $\pi^*(a|s)$ in the actual MDP $\mathcal{M}$, maximizing the cumulative return $J(\mathcal{M}, \pi) := \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^\pi(s,a)}[r(s, a)]/1-\gamma$. Building upon previous works [15], [21], [32], [33], we prove that the DOMAIN provides a ζ -safe policy improvement for the behavioral policy $\pi_b(a|s)$.

Theorem 4: Let $\pi^*(a|s)$ be the policy optimized by DOMAIN. Then, $\pi^*(a|s)$ represents a ζ -safe policy improvement over $\pi_b(a|s)$ in the actual MDP $\mathcal{M}$, satisfying $J(\mathcal{M}, \pi^*) - J(\mathcal{M}, \pi_b) \geq \zeta$ with high probability $1 - \delta$. The value of ζ is given by:

$$\begin{aligned} \zeta &= \underbrace{\frac{\lambda}{1-\gamma} [\varpi(\pi^*, f) - \varpi(\pi_b, f)]}_{:=\mu_1} \\ &\quad - \underbrace{2 \frac{1-f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right]}_{:=\mu_2} \\ &\quad - \underbrace{2 \frac{f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} [\varepsilon_r(s, a) + R_{\max} D(s, a)]}_{:=\mu_3} \end{aligned} \quad (17)$$

The proof of this theorem can be found in Appendix D. Here, μ_2 represents the sampling error term, which becomes negligible when the data volume is sufficiently large. μ_3 denotes the model error term, with a larger value indicating a larger discrepancy between the trained environment model $\hat{T}_\phi$ and the true model T . The term $\varpi(\pi, f) = \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^\pi(s,a)}[\eta(s, a)]$ measures the expected rewards and penalties under a specific policy and model data ratio, capturing the difference in the expected penalty between different policies. Therefore, μ_1 represents the discrepancy in the expected penalties under different policies.

Since the environment model is trained using an offline dataset $\mathcal{D}_{\text{offline}}$ generated by the behavioral policy $\pi_b(a|s)$, the learned model approximates the model data distribution $d_{\mathcal{M}}^{\pi_b}$ generated by the interaction between the behavioral policy and the environment, which is close to the distribution d^{π_b} of the offline dataset. Consequently, the generated model data exhibits high precision, resulting in smaller values of $\omega(s, a)$. Considering that the value of $D_{KL}(\pi^*||\pi_b)$ is not expected to be large, the distributions $d_{\mathcal{M}}^{\pi_b}$ and $d_{\mathcal{M}}^\pi$ can be

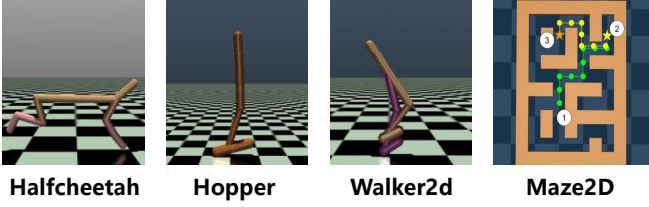


Fig. 2. Description of Gym-Mujoco and Maze2D Tasks. The Gym-Mujoco benchmark includes Halfcheetah, Hopper and Walker2d tasks. The Maze2D benchmark includes three maze layout tasks.

TABLE I
BASE HYPERPARAMETER CONFIGURATION OF DOMAIN.

Parameter		Value
Shared Parameter	Number of hidden units per layer	256/200
	Number of iterations	1M/0.5M
	Batch size	256
	Optimizer	Adam
Environment Model	Model learning rate	1e-4
	Number of hidden layers	4
	Number of model networks N	7
	Number of elites M	5
	Ratio of model date f	0.5
Policy Learning	Learning rate (policy/critic)	1e-4/3e-4
	Number of hidden layers	2
	Discount factor γ	0.99
	Soft update parameter τ	5e-3
Discriminator Training	Target entropy	$-\dim(A)$
	Discriminator learning rate	3e-4
	Number of hidden layers	1
	KL Divergence clipping range	$[10^{-45}, 10]$
	Dynamics ratio clipping range	$[10^{-45}, 1]$
	Output layer	$2 \times \text{Tanh}$

approximated as equal, leading to a higher expected value of $\varpi(\pi^*, f) - \varpi(\pi_b, f) > 0$ with high probability. Thus, by selecting a sufficiently large λ , it can be ensured that $\zeta > 0$. Furthermore, by choosing an appropriate value for f , even a smaller λ can guarantee the safety improvement of the behavioral policy.

V. EXPERIMENTS

This section aims to address the following questions: 1) How does the DOMAIN method compare to state-of-the-art RL algorithms in standard offline RL benchmarks? 2) Can the DOMAIN method handle tasks that require generalization to OOD behaviors?

A. The Implementation Details of Experiments

Experimental Datasets. We employ Gym-Mujoco and Maze2D benchmarks to test the performance of DOMAIN. The Gym-Mujoco tasks, including Halfcheetah-v2, Hopper-v2, and Walker2d-v2, are commonly studied in OpenAI Gym [35], simulated using the MuJoCo physics engine. The Maze2D domain, including Umaze-v1, Medium-v1 and Large-v1 three

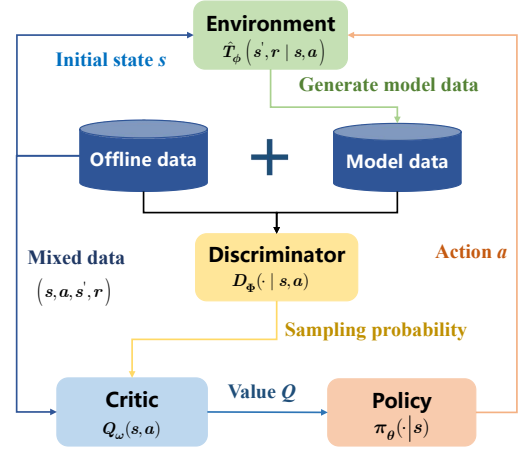


Fig. 3. Network composition of the DOMAIN algorithm. The network is composed of four parts: environment, discriminator, critic, and policy.

maze layouts, is a navigation task requiring the agent to reach a fixed goal location.

Fig. 2 depicts the Halfcheetah, Hopper, Walker2d and Maze2D four tasks. For three Gym-Mujoco tasks, the goal is to control the robot’s various joints to achieve faster and more stable locomotion. By applying RL methods to robot control, the agent observes the state of the MuJoCo environment and controls the robot’s joints based on its learned policy. For Maze2D tasks, the goal is to find the shortest path to reach the fixed goal location. The offline datasets of Gym-Mujoco and Maze2D benchmarks are based on the D4RL dataset [36].

For three Gym-Mujoco tasks, three types of recorded datasets are considered: Medium, Medium-replay, and Medium-expert. The “Medium” dataset is generated by collecting 1 million samples from a policy that undergoes early stopping during online training using SAC [27]. The “Medium-replay” dataset records all samples during the training process, reaching a medium level. The “Medium-expert” dataset is a combination of expert and sub-optimal data [36].

Practical algorithm implementation details. Fig. 3 gives the network composition of the algorithm, including environment, discriminator, critic and policy network. Through the discriminator network, the model data sampling probability is calculated for the critic network update; the critic network guides the policy network update; the actions generated on the strategy network generate more model data through the environment model. Table I provides the hyper-parameters for the network model. The detailed parameter settings for each network update are given below:

1) *Model Training*: The environment model is approximated using a probabilistic network that outputs a Gaussian distribution to obtain the next state and reward. Similar to the approach in [19], we set N as 7 and select the best 5 models. Each model consists of a 4-layer feedforward neural network with 200 hidden units. The model training employs maximum likelihood estimation with a learning rate of 1e-4 and Adam optimizer. The 256 data samples are randomly drawn from the offline dataset $\mathcal{D}_{\text{offline}}$ to calculate the maximum likelihood term.

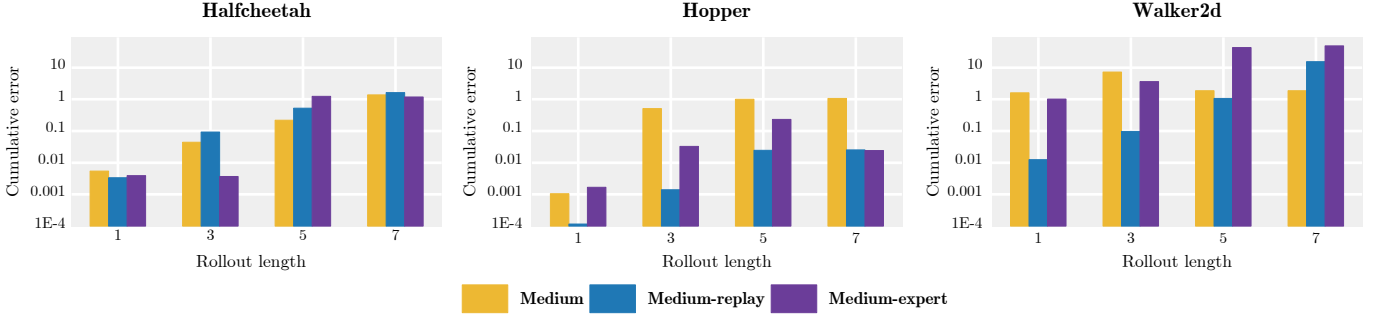


Fig. 4. Quantitative analysis of the cumulative model errors. We compare the trained model with the actual MuJoCo environment, with the latter remaining unknown throughout the policy learning process. The initial state is initialized randomly within the MuJoCo environment, and the rollout policy is based on the final learned policy. To facilitate visualization, the logarithmic transformation is employed on the accumulated errors.

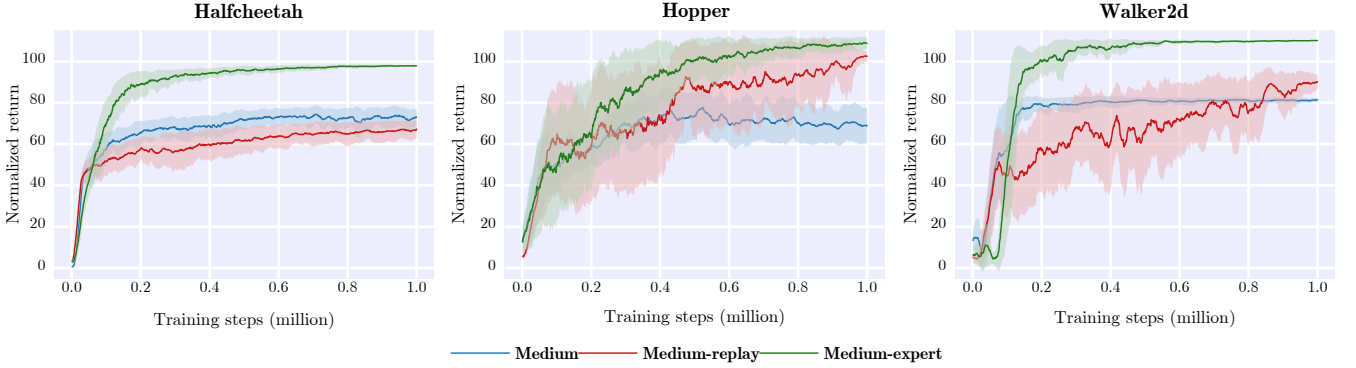


Fig. 5. Corresponding learning curves for Table II. Each figure shows the training curve for a specific task under different datasets, where the shaded area represents the standard variance of the return.

2) *Discriminator Training*: Two discriminator networks, $D_{\Phi_{s,a,s}}(\cdot | s, a, s')$ and $D_{\Phi_{s,a}}(\cdot | s, a)$, are trained to approximate the probabilities $p(\cdot | s, a, s')$, $p(\cdot | s, a)$. The networks use the prediction “ $2 \times \text{Tanh}$ ” activation function to map the output values to the range $[-2, 2]$, and then pass the clipped results through a softmax layer to obtain the final probabilities.

3) *Policy Optimization*: The training step is set as 1 million and half a million for Gym-Mujoco tasks and Maze2D tasks respectively. Policy optimization is based on the SAC framework. The critic network Q_ω and the policy network π_θ adopt a 2-layer feedforward neural network with 256 hidden units. The learning rates for them are set to $3e-4$ and $1e-4$ respectively. The ratio of model data in the dataset is $f = 0.5$. The entropy regularization coefficient α in Eq. (5) is automatically adjusted, with the entropy target set to $-\dim(A)$, and the learning rate for the self-tuning network is set to $1e-4$.

In practical policy optimization, the obtained adaptive sampling distribution is utilized to select data with high uncertainty and penalize their Q -values. Following [13], [21], we sample 10 actions in every high uncertainty state based on the uniform policy $\text{Unif}(a)$ and the learned policy $\pi(a | s)$, and then estimate the log-sum-exp term in (9) using importance sampling.

Parameter Tuning. The quality of model data is influenced by the environment model $\hat{T}_\phi$, the rollout policy $\mu(\cdot | s)$,

and the rollout horizon H . The weight coefficient λ of the regularization term significantly affects the stability of agent training. The rollout policy for Gym-Mujoco and Maze2D benchmarks are the current learned policy and the uniform policy, respectively. Table III and Table VI give the results of Gym-Mujoco tasks and Maze2D tasks for different parameters, respectively. The detailed parameter tuning results in different tasks are as follows.

1) *Rollout horizon H* : For Gym-Mujoco tasks, a short-horizon model rollout is performed [19], [24]. The rollout horizon H is chosen from the set $\{1, 5\}$. We use $H = 1$ for Walker2d-v2 tasks, and $H = 5$ for Hopper-v2 and Halfcheetah-v2. For Maze2D tasks, the rollout horizon H is chosen from the set $\{5, 10, 100\}$. We use $H = 5$ for Maze2D-umaze task, $H = 10$ for Maze2D-medium task, and $H = 100$ for Maze2D-large task.

2) *Weight coefficient λ* : The parameter λ reflects different strengths of regularization. For all tasks, the parameter λ is chosen from the set $\{0.5, 5\}$. A larger λ is needed in narrower data distribution. A smaller is needed in diverse data distribution. For three Maze2D tasks, Gym-Mujoco medium-replay datasets and halfcheetah-medium datasets, λ is chosen as 0.5, while for the remaining datasets, λ is chosen as 5.

B. Experimental Results

Results on environment model prediction. This section first analyzes and visualizes the accuracy of trained dynamic

TABLE II

RESULTS FOR THE GYM-MUJOCO TASKS DURING THE LAST 5 ITERATIONS OF TRAINING AVERAGED OVER 5 SEEDS. $\pm$ CAPTURES THE STANDARD DEVIATION OVER SEEDS. BOLD INDICATES THE HIGHEST SCORE. HIGHLIGHTED NUMBERS INDICATE RESULTS WITHIN 2% OF THE MOST PERFORMANCE ALGORITHM.

Type	Environment	Ours	Model-based baseline					Model-free baseline		
		DOMAIN	ARMOR	RAMBO	COMBO	MOReL	MAPLE	MOAC	IQL	GORL
Medium	Halfcheetah	72.7 $\pm$ 4.7	54.2	77.6	54.2	42.1	63.5	54.3	47.4	51.7
	Hopper	70.3 $\pm$ 8.9	101.4	92.8	97.2	95.4	21.1	83.4	66.3	64.2
	Walker2d	81.5 $\pm$ 1.0	90.7	86.9	81.9	77.8	56.3	86.7	78.3	83.5
Medium Replay	Halfcheetah	67.8 $\pm$ 3.9	50.5	68.9	55.1	40.2	50.3	46.5	45.4	59.0
	Hopper	102.7 $\pm$ 2.3	97.1	96.6	89.5	93.6	87.5	98.0	94.7	74.7
	Walker2d	90.4 $\pm$ 3.6	85.6	85.0	56.0	49.8	76.7	90.1	73.9	79.3
Medium Expert	Halfcheetah	97.7 $\pm$ 0.7	93.5	93.7	90.0	53.3	63.5	87.2	86.7	88.2
	Hopper	109.7 $\pm$ 2.2	103.4	83.3	111.1	108.7	42.5	111.4	91.5	88.8
	Walker2d	110.1 $\pm$ 0.1	112.2	68.3	103.3	95.6	73.8	102.1	109.6	109.6
Average Score		89.2 $\pm$ 3.0	87.6	83.7	82.0	72.9	60.4	84.8	77.0	76.3

TABLE III

THE RESULTS OF D4RL TASK FOR DIFFERENT PARAMETERS. BOLDDED AND HIGHLIGHTED NUMBERS INDICATE THE BEST PERFORMANCE

Type	Environment	$\lambda = 0.5$		$\lambda = 5$	
		$H = 1$	$H = 5$	$H = 1$	$H = 5$
Medium	Halfcheetah	68.49	72.69	51.12	49.08
	Hopper	0.85	69.94	64.62	70.29
	Walker2d	37.44	78.57	81.46	-0.10
Medium Replay	Halfcheetah	54.40	67.75	50.00	46.18
	Hopper	84.35	102.68	90.53	100.6
	Walker2d	90.36	85.08	87.35	52.69
Medium Expert	Halfcheetah	33.24	50.31	87.82	97.74
	Hopper	1.84	55.50	101.1	109.67
	Walker2d	10.72	105.78	110.12	109.02

model prediction before answering the above questions. Fig. 4 gives the prediction results of cumulative error in rollout lengths 1,3,5,7 for different tasks and datasets. In this figure, the predicted errors of four distinct datasets for tasks halfcheetah, hopper and walker2d are arranged from left to right. The prediction error of the trained environmental model exhibits an increasing trend with the increment of steps across all datasets. Since the model exhibits the largest prediction error in three walker2d tasks, the rollout length is set to 1 in the walker2d task to ensure the stability of the agent.

However, single-step prediction is not conducive to the expansion of the OOD data by the environment model, because most of the predicted state-action data still falls within the region. As rollout length H increases, it is beneficial to explore the OOD region. Meanwhile, it is necessary to ensure the model data error is not too large. In Fig. 4, the prediction error of H from 1 to 7 of the environmental model does not change too much under Halfcheetah and Hopper tasks. Within the error range, the optimal H value should be determined by more experiments.

Results on the Gym-Mujoco tasks. To answer question 1, this section compares recent model-free algorithms, such as MOAC [40], IQL [37], and GORL [41], as well as model-based algorithms, including ARMOR [25], MAPLE [39], MOReL [20], COMBO [21], and RAMBO [24]. The performance of these algorithms is evaluated based on the cumulative return in the robot motion tasks within the MuJoCo environment. A higher score indicates better performance in terms of more stable and faster locomotion control. For comparison, the scores are normalized between 0 (random policy score) and 100 (expert policy score) [36]. The normalized score $\tilde{S}$ is computed by:

$$\tilde{S} = \frac{S - S_r}{S_e - S_r} \times 100.$$

where S_r , S_e and S are the expected return of a random policy, an expert policy, and the trained policy by offline RL algorithms, respectively. The random policy scores under Halfcheetah, Hopper, and Walker2d are -280.179, 1.629, and -20.272 respectively. The expert policy scores under Halfcheetah, Hopper, and Walker2d are 12135.0, 4592.3, and 3234.3 respectively.

Table II presents the normalized scores of different methods. The reported scores are the normalized scores of the learned policies during the last 5 iterations, averaged over 5 random seeds. Note that the average of the last 1 [21], [25], 5 and 10 [24] iterations are usually selected as final results. The number of the last iteration of different algorithms should be the same for fair comparison. For our algorithm, the results are basically consistent in these three cases. Here we choose the average of the last 5 iterations as the final result. The results for all methods are obtained from the original papers.

The table shows that the DOMAIN method achieves the best performance in three out of the nine datasets, comparable results in three out of the remaining six settings, and slightly poor performance in the three medium settings. It is difficult for model-based offline RL algorithms to perform well on all datasets due to the differences between various datasets.

TABLE IV
RESULTS FOR TASKS THAT REQUIRE GENERALIZATION DURING THE LAST 5 ITERATIONS OF TRAINING.

Environment	DOMAIN	COMBO	MoReL	MOPO	MBPO	CQL	SAC	BRAC-p	BEAR
Halfcheetah-jump	93.65 ± 3.02	45.02	28.26	34.61	26.19	8.23	−26.64	10.87	2.39

TABLE V
RESULTS FOR THE MAZE2D TASKS DURING THE LAST 5 ITERATIONS OF TRAINING AVERAGED OVER 5 SEEDS. ± CAPTURES THE STANDARD DEVIATION OVER SEEDS. BOLD INDICATES THE HIGHEST SCORE.

Type	Environment	Ours DOMAIN	Model-based baseline				Model-free baseline			
			ROMI-BCQ	COMBO	MoReL	MOPO	BEAR	CQL	BRAC	BAIL
Sparse	Maze2D-umaze	142.6 ± 7.7	139.5	76.4	−14.4	−14.8	65.7	18.9	−9.2	44.7
	Maze2D-medium	154.4 ± 12.0	82.4	68.5	−4.8	37.4	25.0	14.6	70.0	19.7
	Maze2D-large	152.1 ± 42.8	83.1	14.1	−2.3	−0.8	81.0	16.0	0.2	4.4
Average Score		149.7 ± 20.8	101.7	53.0	−7.2	7.3	57.2	16.5	20.3	22.9

TABLE VI
THE RESULTS OF MAZE2D TASKS FOR DIFFERENT PARAMETERS. BOLD AND HIGHLIGHTED NUMBERS INDICATE THE BEST PERFORMANCE

Type	Environment	$H = 5$	$H = 10$	$H = 50$	$H = 100$
Sparse	Maze2D-umaze	142.63	66.94	74.16	−15.82
	Maze2D-medium	107.03	154.39	126.33	114.15
	Maze2D-large	52.48	112.08	107.57	152.14

Therefore, most offline RL algorithms show the advantages of the algorithm by comparing the average values under different tasks. Overall, compared to other RL algorithms, the DOMAIN algorithm achieves the best average performance across multiple datasets. Corresponding learning curves are shown in Fig. 5.

Moreover, Table II shows that the performance of the “Medium-expert” dataset in three tasks is almost always superior to that of “Medium” and “Medium-replay” datasets for any algorithms. The difference between “Medium” and “Medium-replay” is that “Medium-replay” contains more data at different levels, which can facilitate agent exploration. However, the algorithm may not necessarily perform better than “Medium” under “Medium-replay” for different tasks. This mainly depends on the adaptability of the algorithm itself to diverse data and the coverage of the data.

Results on tasks that require generalization. To answer question 2, similar to [21], the halfcheetah-jump environment constructed by [19] is introduced, which requires the agent to solve a task different from the behavioral policy. In the original halfcheetah environment, the objective is to maximize the robot’s running speed, with the reward defined as $r(s, a) = \max\{v_x, 3\} - 0.1\|a\|_2^2$, where v_x denotes the robot’s velocity along the x -axis. In the halfcheetah-jump environment, the goal is to make the robot jump as high as possible, and the states of jumping highly are rarely seen in offline states. The rewards in the offline dataset are redefined as $r(s, a) = \max\{v_x, 3\} - 0.1\|a\|_2^2 + 15(z - \text{init } z)$, where z represents the robot’s position along the z -axis and $\text{init } z$

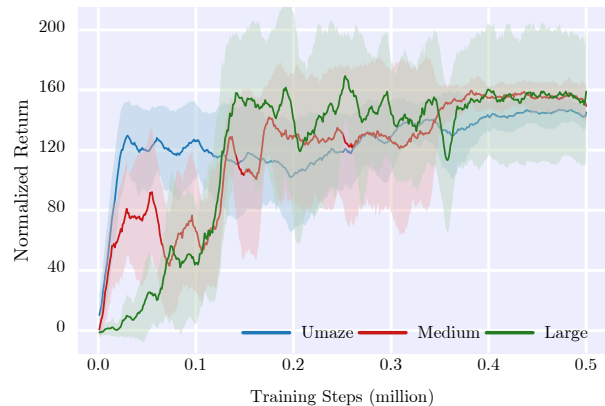


Fig. 6. Corresponding learning curves for Table V. The shaded area represents the standard variance of the return. The shaded area represents the standard variance of the return.

denotes the initial position along the z -axis.

The datasets used in both cases are collected by training the SAC algorithm for 1 million steps and are similar to the “Medium-expert” dataset in D4RL. Therefore, this dataset serves as the foundation dataset, and the rewards are appropriately modified to obtain the final dataset for training. Table IV presents the normalized scores of different algorithms in the halfcheetah-jump environment. The results for model-based algorithms are sourced from the work of [21], while the results for model-free algorithms (BEAR [9], BRAC [11], SAC [27], and CQL [13]) are obtained from [19]. It can be observed that the DOMAIN algorithm performs the best, demonstrating stronger generalization capabilities and robustness.

Results on the Maze2D tasks. The Maze2D tasks include three different maze layouts. The names of three tasks in the D4RL dataset are “Maze2d-umaze-sparse”, “Maze2d-medium-sparse” and “Maze2d-large-sparse”, respectively. These tasks are challenging for RL methods since the reward is sparse. The score calculation of Maze2D is the same as the Gym-Mujoco tasks. The random policy scores under Umaze, Medium, and

TABLE VII
THE PARAMETER CONFIGURATION OF OUR DESIGNED SIMPLE MDP.

Parameter	Designed Dynamic Model						Offline Dataset	
	a	$[-1, -0.6)$	$[-0.6, -0.2)$	$[-0.2, 0.2)$	$[0.2, 0.6)$	$[0.6, 1]$	Behavioral Policy	Data Size
Value	μ	$-a + 0.2$	$5(a + 0.4)^2 + 0.6$	$-5a^2 + 1$	$10(a - 0.4)^2 + 0.4$	$a + 0.2$	Random Policy	1000
	σ	0.06	0.04	0.02	0.04	0.06		

Large mazed layouts are 23.85, 13.13, and 6.7 respectively. The expert policy scores under Umaze, Medium, and Large mazed layouts are 161.86, 277.39, and 273.99 respectively.

Table V compares the normalized scores of DOMAIN with that of other current offline RL algorithms on Maze2D tasks, where model-based algorithms include ROMI [42], MOREL [20], COMBO [21] and MOPO [19], and model-free algorithms include BEAR [9], BRAC [11], BAIL [43] and CQL [13]. Apart from ROMI, the original papers of the above methods don't give the results on the Maze2D benchmark. Therefore, the results of the above method are sourced from the work of [42]. Table V shows that the performance of DOMAIN is much better than other methods, which also shows DOMAIN has advantages in sparse reward tasks. Corresponding learning curves are shown in Fig. 6.

Moreover, the methods DOMAIN and COMBO, which regularize the model data to incorporate conservatism, perform better than the methods MOPO and MOREL that use model uncertainty as a penalty for rewards on the Maze2D tasks. This shows the unreliability of estimating model uncertainty. DOMAIN performs much better than COMBO, which further verifies the advantage of the adaptive sampling distribution. The maximum rollout horizons of Umaze, Medium and Large are 300, 600 and 800 respectively. The tasks are becoming more and more complex. The parameter tuning of rollout length in Table VI indicates that the large rollout length is beneficial for the performance of long-horizon tasks.

C. Visualisation of DOMAIN Performance

Due to the high dimension of states and actions in the D4RL tasks, visualizing the performance is greatly challenging. Therefore, a simple example MDP is defined for illustrative purposes following [24]. For this MDP, both the state space $\mathcal{S}$ and action space $\mathcal{A} \in [-1, 1]$ are one-dimensional. The reward is defined as $R(s, a) = s$, where a larger value of state s leads to a higher reward. The dynamic transition model is given by $T(s' | s, a) = \mathcal{N}(\mu, \sigma)$ with the parameters shown in Table VII, indicating that the next state depends on the current action, independent of the initial state. To visualize the DOMAIN exploration in OOD region, we sample actions from $\mathcal{N}(\mu = 0, \sigma = 0.35)$, which means that the behavioral policy for the offline dataset is random policy.

Fig. 7(a) gives the collected offline data distribution for the above MDP, which illustrates that the actions do not cover all regions. Fig. 7(b) presents the prediction results of the trained dynamic model. It can be concluded that the dynamic model shows higher prediction accuracy in-distribution region (low uncertain area) but lower prediction accuracy in the OOD

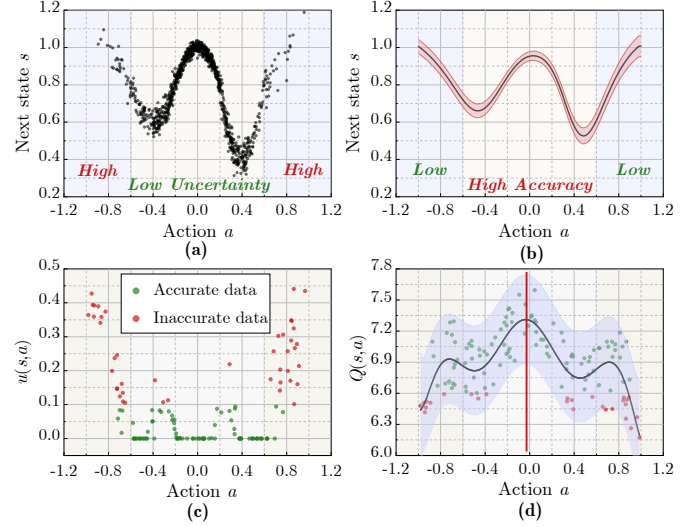


Fig. 7. The visualization results of the DOMAIN algorithm on designed MDP. (a) illustrates the distribution of collected offline data. The intermediate color region represents the low uncertainty area, while the region on both sides represents the high uncertainty area; (b) presents the prediction results of the trained dynamic model under the initial state $s = 0$. The red area represents the range of variance predicted by ensemble models; (c) depicts the error distribution of model data. The green color represents precise model data, while the red color represents imprecise model data; (d) illustrates the distribution of Q -values of model data. The black curve is the high-order fitting curve of the data. The red line indicates the optimal action, i.e. $a \approx 0$.

region (high uncertain area). Fig. 7(c) displays the model data error after 3000 iterations. It can be observed that the model data within the region exhibits relatively low errors, while the data outside the region shows larger errors. This result is consistent with the prediction results of the trained dynamic model. Fig. 7(d) shows the model data Q -value curve in state $s = 0$ after 3000 iterations. We can find that the DOMAIN algorithm does not universally underestimate all model data. Instead, it is more prone to underestimating inaccurate model data. Moreover, the value of the optimal action, i.e., $\arg \max_a Q(s, a)$, is approximately equal to zero, which aligns with the selection of the offline dataset. This is because the reward is solely dependent on the state, and the state value is maximized when $a \approx 0$ is in-distribution region. The consistency between model data and offline data in selecting the optimal actions contributes to enhancing the stability of agent training.

D. The Parameter Analysis

This part gives further analysis for weight λ and rollout

TABLE VIII

THE ANALYSIS RESULTS FOR PARAMETER H AND λ . BOLD INDICATES THE HIGHEST SCORE IN ABLATION EXPERIMENTS. HIGHLIGHTED NUMBERS INDICATE THE OPTIMAL RESULTS IN TABLE III.

Type	Environment	Analysis for weight λ					Analysis for rollout length H				
		H	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 5$	λ	$H = 1$	$H = 3$	$H = 5$	$H = 7$
Medium	Halfcheetah	5	70.1	72.7	65.1	49.1	0.5	68.5	72.5	72.7	74.5
	Hopper	5	13.6	69.9	24.3	70.3	5.0	64.6	81.6	70.3	68.1
	Walker2d	1	16.3	37.4	64.3	81.5	5.0	81.5	11.3	46.2	22.2
Medium Replay	Halfcheetah	5	68.3	67.8	64.4	46.2	0.5	54.4	65.9	67.8	65.0
	Hopper	5	13.8	102.7	77.2	100.6	0.5	84.4	103.8	102.7	103.7
	Walker2d	1	44.4	90.4	86.0	87.4	0.5	90.4	83.5	85.1	81.0
Medium Expert	Halfcheetah	5	56.7	50.3	100.9	97.7	5.0	87.8	98.0	97.7	97.6
	Hopper	5	10.2	55.5	104.2	109.7	5.0	101.1	109.3	109.7	111.2
	Walker2d	1	8.6	10.7	0.9	110.1	5.0	110.1	103.5	109.0	83.0

length H . Table VIII analyzes the impact of two parameter changes on the performance of the agent. Due to high training cost, here we fix one parameter and test the impact of changes in another parameter on agent performance.

The effect of weight λ . The parameter λ is used to regularization between Bellman error and Q value of data, and guarantee the lower bound of Q -function for inaccuracy model data. A larger λ is needed to guarantee the lower bound of inaccuracy model data in narrower data distribution. However, for diverse data distribution, too large λ may cause the lower bound to be too loose, thus reducing agent performance. Table VIII shows that in narrower “Medium” and “Medium-Expert” datasets, too small λ leads to unstable training, thus reducing the agent performance. For diverse “Medium-Replay” datasets, too large λ causes the high constraint, thus limiting the improvement of the agent.

The effect of rollout length H . When rollout length $H = 1$, the trained environment model has high prediction accuracy. As H increases, the accuracy of prediction decreases, but it is beneficial to explore the OOD region. If the model data error is too large, the algorithm will not be able to achieve a balance between accurate offline data and inaccurate model data, leading to unstable agent training. Therefore, it is necessary to ensure that the model data error is not too large. Table VIII shows that, as H increases, the agent performance shows a downward trend under the Walker2d task since the error of the trained environment is large. Under the Halfcheetah and Hopper tasks, as H increases, the agent performance does not show a downward trend, but the optimal H is different for different datasets.

VI. CONCLUSION

This paper proposes a mildly conservative model-based offline RL (DOMAIN) algorithm in this paper. By designing a model data adaptive sampling distribution, this algorithm can dynamically adjust the model data penalties, increasing the penalty for model data with larger errors while reducing the penalty for accurate model data, thereby reducing its conservatism. Moreover, DOMAIN avoids unreliable estimation

of model uncertainty. Theoretical analysis demonstrates that DOMAIN can guarantee safe policy improvements and Q -values learned by DOMAIN outside the region are lower bounds of the true Q -values. Experimental results on the D4RL benchmark show that DOMAIN outperforms previous RL algorithms on most datasets and achieves the best generalization performance in tasks that require adaptation to unseen behaviors.

Future works can focus on the below aspects: 1) Since incorporating conservatism during the training process constrains the improvement of agent performance, future efforts can focus on optimizing the environment model itself to further enhance the overall performance. 2) Since training the discriminator to judge the model data error has a certain error, future works can focus on finding a more accurate and stable way to measure the model data error. 3) Most RL algorithms need to tune parameters constantly for different datasets. The less parameter tuning RL methods need further studied in the future.

APPENDIX

A. Proof of Theorem 1

Before proof, we provide a notation clarification. Let $\mathcal{T}^\pi$ denote the ideal Bellman operator, and $\hat{\mathcal{T}}^\pi$ represents the actual Bellman operator. $\mathcal{M}$ refers to the true MDP, $\hat{\mathcal{M}}$ represents the learned MDP model, and $\bar{\mathcal{M}}$ denotes the empirical MDP, where “empirical” refers to operations based on samples. $\mathcal{T}_{\mathcal{M}}^\pi$, $T_{\mathcal{M}}$, and $r_{\mathcal{M}}$ denote the Bellman operator, state transition, and reward under the true MDP $\mathcal{M}$, respectively. $\mathcal{T}_{\hat{\mathcal{M}}}^\pi$, $T_{\hat{\mathcal{M}}}$, and $r_{\hat{\mathcal{M}}}$ represent the corresponding operators under the learned MDP $\hat{\mathcal{M}}$, while $\mathcal{T}_{\bar{\mathcal{M}}}^\pi$, $T_{\bar{\mathcal{M}}}$ and $r_{\bar{\mathcal{M}}}$ denote the operators under the empirical MDP $\bar{\mathcal{M}}$. Note that $\mathcal{T}^\pi = \mathcal{T}_{\mathcal{M}}^\pi$ and $\hat{\mathcal{T}}^\pi = (1 - f)\mathcal{T}_{\mathcal{M}}^\pi + f\mathcal{T}_{\hat{\mathcal{M}}}^\pi$.

Proof. $\mathcal{T}_{\hat{\mathcal{M}}}^\pi Q(s, a)$ is empirical Bellman operator, denoted as $\mathcal{T}_{\hat{\mathcal{M}}}^\pi Q(s, a) = r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s', a')$. $\mathcal{T}_{\mathcal{M}}^\pi$ is Bellman operator under learned MDP $\hat{\mathcal{M}}$, denoted as $\mathcal{T}_{\mathcal{M}}^\pi = r(s, a) +$

$\gamma \mathbb{E}_{s' \sim T_{\widehat{\mathcal{M}}}}[\max_{a' \in \mathcal{A}} Q(s', a')]$. Then, according to $\widehat{T}^\pi = (1 - f)\mathcal{T}_{\widehat{\mathcal{M}}}^\pi + f\mathcal{T}_{\widehat{\mathcal{M}}}^\pi$, we can derive:

$$\begin{aligned} \widehat{T}^\pi Q(s, a) &= (1 - f) \left\{ r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s', a') \right\} \\ &\quad + f \left\{ r(s, a) + \gamma \mathbb{E}_{s' \sim T_{\widehat{\mathcal{M}}}}[\max_{a' \in \mathcal{A}} Q(s', a')] \right\} \end{aligned} \quad (18)$$

Let Q_1 and Q_2 be two arbitrary Q -function, and $\|\cdot\|_\infty$ is L_∞ norm, then we can get the following equation:

$$\begin{aligned} &\|\widehat{T}^\pi Q_1 - \widehat{T}^\pi Q_2\|_\infty \\ &= (1 - f) \gamma \max_{s, a} \left| \max_{a' \in \mathcal{A}} Q_1(s', a') - \max_{a' \in \mathcal{A}} Q_2(s', a') \right| \\ &\quad + f \gamma \max_{s, a} \left| \mathbb{E}_{s' \sim T_{\widehat{\mathcal{M}}}}[\max_{a' \in \mathcal{A}} Q_1] - \mathbb{E}_{s' \sim T_{\widehat{\mathcal{M}}}}[\max_{a' \in \mathcal{A}} Q_2] \right| \\ &\leq (1 - f) \gamma \max_{s, a} \left| \max_{a' \in \mathcal{A}} Q_1(s', a') - \max_{a' \in \mathcal{A}} Q_2(s', a') \right| \\ &\quad + f \gamma \max_{s, a} \mathbb{E}_{s' \sim T_{\widehat{\mathcal{M}}}} \left| \max_{a' \in \mathcal{A}} Q_1 - \max_{a' \in \mathcal{A}} Q_2 \right| \\ &\leq (1 - f) \gamma \|Q_1 - Q_2\|_\infty + f \gamma \|Q_1 - Q_2\|_\infty \\ &= \gamma \|Q_1 - Q_2\|_\infty, \quad \text{where } \gamma \in (0, 1) \end{aligned} \quad (19)$$

According to the fixed point theorem, any initial Q function can converge to a unique fixed point Q^* by repeatedly applying the Bellman operator $\widehat{T}^\pi$. Thus, Theorem 1 is proven.

B. Proof of Theorem 2

The difference between empirical MDP $\widehat{\mathcal{M}}$ and actual MDP $\mathcal{M}$ lies in the sampling error. The sampling error comes from the inadequate dataset (it is difficult for the dataset to cover the entire data space). Therefore, if the dataset is large enough, the sampling error would be smaller and the gap between empirical MDP and actual MDP would become smaller. Therefore, we make the following assumptions following prior works [13], [21], [31], [34]:

Assumption 1: For $\forall s, a \in \mathcal{D}$, the following inequality holds with a probability greater than $1 - \delta$, where $\delta \in (0, 1)$:

$$\begin{aligned} |r_{\mathcal{M}}(s, a) - r_{\widehat{\mathcal{M}}}(s, a)| &< \frac{C_{r, \delta}}{\sqrt{|\mathcal{D}(s, a)|}} \\ |T_{\mathcal{M}}(s' | s, a) - T_{\widehat{\mathcal{M}}}(s' | s, a)|_1 &< \frac{C_{T, \delta}}{\sqrt{|\mathcal{D}(s, a)|}} \end{aligned} \quad (20)$$

where $|\mathcal{D}(s, a)|$ represents the cardinality (number of occurrences) of a specific state-action pair (s, a) in the dataset $\mathcal{D}$. For $(s, a) \notin \mathcal{D}$, we assume $|\mathcal{D}(s, a)|$ to be less than 1. The reward is bounded within a certain range, given by $|r_{\mathcal{M}}(s, a)| < R_{\max}$.

Lemma 1: For any policy π , the disparity between the actual Bellman operator $\mathcal{T}_{\mathcal{M}}^\pi$ and the Bellman operators $\mathcal{T}_{\widehat{\mathcal{M}}}^\pi$ under the empirical MDP, as well as the disparity between the actual Bellman operator $\mathcal{T}_{\mathcal{M}}^\pi$ and the Bellman operators $\mathcal{T}_{\widehat{\mathcal{M}}}^\pi$

under the learned MDP satisfy the following inequalities with high probability $1 - \delta$, respectively.

$$\begin{aligned} \left| \left(\mathcal{T}_{\mathcal{M}}^\pi \widehat{Q}^k \right) - \left(\mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k \right) \right| &< \frac{C_{r, T, \delta} R_{\max}}{(1 - \gamma) \sqrt{|\mathcal{D}(s, a)|}} \\ \left| \left(\mathcal{T}_{\mathcal{M}}^\pi \widehat{Q}^k \right) - \left(\mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k \right) \right| &< \varepsilon_r + D(T_{\mathcal{M}}, T_{\widehat{\mathcal{M}}}) \frac{2\gamma R_{\max}}{1 - \gamma} \end{aligned} \quad (21)$$

Proof. The difference between the Bellman operator $\mathcal{T}_{\widehat{\mathcal{M}}}^\pi$ under the empirical MDP and the actual Bellman operator $\mathcal{T}_{\mathcal{M}}^\pi$ is given by:

$$\begin{aligned} &\left| \left(\mathcal{T}_{\mathcal{M}}^\pi \widehat{Q}^k \right) - \left(\mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k \right) \right| \\ &\leq \left| r_{\mathcal{M}}(s, a) - r_{\widehat{\mathcal{M}}}(s, a) \right| + \gamma \left| \sum_{s'} (T_{\mathcal{M}}(s' | s, a) - T_{\widehat{\mathcal{M}}}(s' | s, a)) \mathbb{E}_{a' \sim \pi} [\widehat{Q}^k(s', a')] \right| \\ &\leq \frac{(1 - \gamma) C_{r, \delta} + \gamma C_{T, \delta} 2R_{\max}}{(1 - \gamma) \sqrt{|\mathcal{D}(s, a)|}} = \frac{C_{r, T, \delta} R_{\max}}{(1 - \gamma) \sqrt{|\mathcal{D}(s, a)|}} \end{aligned} \quad (22)$$

Here, $C_{r, T, \delta}$ is a function representing a combination of $C_{r, \delta}$ and $C_{T, \delta}$. Similarly, the difference between the Bellman operator $\mathcal{T}_{\widehat{\mathcal{M}}}^\pi$ under the learned MDP and the actual Bellman operator $\mathcal{T}_{\mathcal{M}}^\pi$ can be derived as:

$$\begin{aligned} &\left| \left(\mathcal{T}_{\mathcal{M}}^\pi \widehat{Q}^k \right) - \left(\mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k \right) \right| \\ &\leq \left| r_{\mathcal{M}}(s, a) - r_{\widehat{\mathcal{M}}}(s, a) \right| + \gamma \left| \sum_{s'} (T_{\mathcal{M}}(s' | s, a) - T_{\widehat{\mathcal{M}}}(s' | s, a)) \mathbb{E}_{a' \sim \pi} [\widehat{Q}^k(s', a')] \right| \\ &\leq \left| r_{\mathcal{M}}(s, a) - r_{\widehat{\mathcal{M}}}(s, a) \right| + D(T_{\mathcal{M}}, T_{\widehat{\mathcal{M}}}) \frac{2\gamma R_{\max}}{1 - \gamma} \end{aligned} \quad (23)$$

where $\varepsilon_r(s, a)$ represents the error between the learned reward and the actual reward, and $D(T_{\mathcal{M}}, T_{\widehat{\mathcal{M}}})$ denotes the discrepancy between the transitions of the learned model and the true model. Lemma 1 is thus proven.

The proof derivation of Theorem 2 is presented below. First, we introduce the following notation: Let $\mathcal{T}^\pi Q = r + \gamma P^\pi Q$, where $P^\pi Q(s, a) = \mathbb{E}_{s' \sim T(\cdot | s, a), a' \sim \pi(\cdot | s')} [Q(s', a')]$. Based on Eq. (14) and the Lemma 1, it can obtain:

$$\begin{aligned} \widehat{Q}^{k+1} &= (1 - f) \mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k + f \widehat{T}^\pi \widehat{Q}^k - \lambda \eta(s, a) \\ &\leq \mathcal{T}_{\widehat{\mathcal{M}}}^\pi \widehat{Q}^k + (1 - f) \underbrace{\frac{C_{r, T, \delta} R_{\max}}{(1 - \gamma) \sqrt{|\mathcal{D}(s, a)|}}}_{:= \Delta_1(s, a)} - \lambda \eta(s, a) \\ &\quad + f \underbrace{(|r_{\mathcal{M}} - r_{\widehat{\mathcal{M}}}| + D(T_{\mathcal{M}}, T_{\widehat{\mathcal{M}}}) \frac{2\gamma R_{\max}}{1 - \gamma})}_{:= \Delta_2(s, a)} \\ &= \mathcal{T}^\pi \widehat{Q}^k + \underbrace{(1 - f) \Delta_1(s, a) + f \Delta_2(s, a) - \lambda \eta(s, a)}_{:= \Delta(s, a)} \end{aligned} \quad (24)$$

By applying the fixed-point theorem, the above equation leads to:

$$\begin{aligned} \widehat{Q}^\pi &\leq \mathcal{T}^\pi \widehat{Q}^\pi + \Delta = r + \gamma P^\pi \widehat{Q}^\pi + \Delta \\ \Rightarrow \widehat{Q}^\pi &\leq (I - \gamma P^\pi)^{-1} (r + \Delta) \\ &= (I - \gamma P^\pi)^{-1} r + (I - \gamma P^\pi)^{-1} \Delta \end{aligned} \quad (25)$$

Since $\mathcal{T}^\pi Q := Q = r + \gamma P^\pi Q \Rightarrow Q = (I - \gamma P^\pi)^{-1} r$, the above equation can be further derived as:

$$\widehat{Q}^\pi(s, a) \leq Q^\pi(s, a) + (I - \gamma P^\pi)^{-1} \Delta(s, a) \quad (26)$$

Based on the value function $V^\pi(s) = \mathbb{E}_{a \sim \pi(a|s)} [Q^\pi(s, a)]$, it can derive:

$$\widehat{V}^\pi(s) \leq V^\pi(s) + (I - \gamma P^\pi)^{-1} \mathbb{E}_{a \sim \pi(a|s)} [\Delta(s, a)] \quad (27)$$

where $V^\pi(s)$ is the true value and $\widehat{V}^\pi(s)$ is the estimated value (the low bound). The lower bound value depends on the term $(I - \gamma P^\pi)^{-1} \mathbb{E}_{a \sim \pi(a|s)} [\Delta(s, a)]$, which also measures the gap between the true value and the lower bound. This gap is influenced by sampling error $\Delta_1(s, a)$, environment model error $\Delta_2(s, a)$ and the degree to which the data is far away from the offline region $\eta(s, a)$.

Note that the error of data from OOD region is likely to be greater than that from offline region, that is, $\omega(s, a)$ become larger and $d(s, a)$ become smaller for model data in OOD region, thereby satisfying $\Delta_3 > 0$. When the offline data is sufficient enough, the Δ_1 and Δ_2 approach zero. Therefore, the term $\Delta < 0$ holds with high probability, thus $(I - \gamma P^\pi)^{-1} \mathbb{E}_{a \sim \pi(a|s)} [\Delta(s, a)] < 0$ holds with high probability for OOD data.

Here, to ensure the lower bound, λ can be chosen to avoid overestimation. Moreover, since the expression $(I - \gamma P^\pi)^{-1}$ is positive semi-definite, as long as $\mathbb{E}_{a \sim \pi(a|s)} [\Delta(s, a)] < 0$, the choice of λ can be controlled by the following condition:

$$\begin{aligned} & \lambda \cdot \min_s \mathbb{E}_{a \sim \pi(a|s)} [\eta(s, a)] \\ & \geq \max_s \mathbb{E}_{a \sim \pi(a|s)} [(1-f)\Delta_1(s, a) + f\Delta_2(s, a)] \end{aligned} \quad (28)$$

Since $\Delta_1(s, a) > 0, \Delta_2(s, a) > 0, \lambda > 0$, in order for Eq. (28) to satisfy the condition, it is a prerequisite that the term $\min \mathbb{E}_{a \sim \pi(a|s)} [\eta(s, a)] > 0$ holds, leading to:

$$\begin{aligned} & \mathbb{E}_{a \sim \pi(a|s)} [\eta(s, a)] \\ & = \mathbb{E}_{a \sim \pi(a|s)} \left[\frac{\omega(s, a) - d(s, a)}{(1-f)d^{\pi_b}(s, a) + fd_{\widehat{\mathcal{M}}}^\pi(s, a)} \right] \\ & = \int_{\mathcal{A}} \frac{\omega(s, a) - d(s, a)}{(1-f)d^{\pi_b}(s, a) + fd_{\widehat{\mathcal{M}}}^\pi(s, a)} da \\ & \geq \frac{\int_{\mathcal{A}} \omega(s, a) da}{(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)} \\ & \quad - \frac{\int_{\mathcal{A}} d(s, a) da}{(1-f)d^{\pi_b}(s) \min_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)} \end{aligned} \quad (29)$$

Consequently, the condition for $\min \mathbb{E}_{a \sim \pi(a|s)} [\eta(s, a)] > 0$ to hold is derived as follows:

$$\begin{aligned} & \int_{\mathcal{A}} \omega(s, a) da \\ & > \frac{(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)}{(1-f)d^{\pi_b}(s) \min_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)} \int_{\mathcal{A}} d(s, a) da \\ & = \xi(s) \int_{\mathcal{A}} d(s, a) da \end{aligned} \quad (30)$$

Substituting Eq. (30) into (28), it can further derive:

$$\begin{aligned} & \lambda \cdot \min_s \frac{(\xi(s) - 1) \int_{\mathcal{A}} d(s, a) da}{(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)} \\ & \geq (1-f) \frac{C_{r,T,\delta} R_{\max}}{(1-\gamma) \min_{s,a} \{ \sqrt{|\mathcal{D}(s, a)|} \}} \\ & \quad + f \left[\max_{s,a} \{ |r_{\mathcal{M}} - r_{\widehat{\mathcal{M}}}| \} + \frac{2\gamma R_{\max}}{1-\gamma} \max_{s,a} \{ D(T_{\widehat{\mathcal{M}}}, T_{\mathcal{M}}) \} \right] \\ & \Rightarrow \lambda \geq \frac{(1-f) \frac{C_{r,T,\delta} R_{\max}}{(1-\gamma) \min_{s,a} \{ \sqrt{|\mathcal{D}|} \}}}{\min_s \left\{ \frac{(\xi(s)-1) \int_{\mathcal{A}} d(s, a) da}{[(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)]} \right\}} \\ & \quad + \frac{f[\max_{s,a} \{ |r_{\mathcal{M}} - r_{\widehat{\mathcal{M}}}| \} + \frac{2\gamma R_{\max}}{1-\gamma} \max_{s,a} \{ D_{TV}(T_{\mathcal{M}}, T_{\widehat{\mathcal{M}}}) \}]}{\min_s \left\{ \frac{(\xi(s)-1) \int_{\mathcal{A}} d(s, a) da}{[(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)]} \right\}} \end{aligned} \quad (31)$$

Thus, Theorem 2 is proven.

C. Proof of Theorem 3

Yu *et al.* [21] give expression: $\widehat{Q}^{k+1}(s, a) = (\widehat{\mathcal{T}}^\pi \widehat{Q}^k)(s, a) - \lambda \left[\frac{\rho(s, a) - d(s, a)}{(1-f)d(s, a) + f\rho(s, a)} \right]$. Assuming DOMAIN and COMBO methods have the same sampling and model errors, we have:

$$\begin{aligned} & \kappa_1 - \kappa_2 \\ & = \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)} [\widehat{Q}_1^\pi(s, a) - \widehat{Q}_2^\pi(s, a)] \\ & = \mathbb{E}_{s \sim d(s)} \left[\int_{\mathcal{A}} \frac{\pi(a|s) [\rho(s, a) - \omega(s, a)]}{(1-f)d^{\pi_b}(s) \pi_b(a|s) + fd_{\widehat{\mathcal{M}}}^\pi(s) \pi(a|s)} da \right] \\ & \geq \mathbb{E}_{s \sim d(s)} \left[\frac{\int_{\mathcal{A}} (\rho(s, a) - \omega(s, a)) da}{(1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s)} \right] \\ & \geq \frac{\mathbb{E}_{s \sim d(s)} [d_{\widehat{\mathcal{M}}}^\pi(s)] - \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)} [\omega(s, a)]}{\underbrace{\max_s \left\{ (1-f)d^{\pi_b}(s) \max_a \left\{ \frac{\pi_b(a|s)}{\pi(a|s)} \right\} + fd_{\widehat{\mathcal{M}}}^\pi(s) \right\}}_{>0}} \end{aligned} \quad (32)$$

where $d(s)$ represents the state distribution of the dataset, which can be an offline dataset, a model dataset, or the overall dataset. When $\mathbb{E}_{s \sim d(s)} [d_{\widehat{\mathcal{M}}}^\pi(s)] \geq \mathbb{E}_{s \sim d(s), a \sim \pi(a|s)} [\omega(s, a)]$, the average value function of the DOMAIN method on the dataset is greater than that of COMBO. In other words, under the aforementioned condition, DOMAIN is less conservative than COMBO. Thus, Theorem 3 is proven.

D. Proof of Theorem 4

RL aims to maximize the cumulative return, denoted as $\pi^* = \max_{\pi} J(\mathcal{M}, \pi) := \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^\pi(s,a)} [r(s, a)]$. Before proving Theorem 4, we present two lemmas.

Lemma 2: For any $\mathcal{M}$ and the experience $\overline{\mathcal{M}}$ generated by sampling according to the behavioral policy π_b , the performance of both $\mathcal{M}$ and $\overline{\mathcal{M}}$ under policy π satisfies:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\overline{\mathcal{M}}, \pi)| \\ & \leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^\pi(s,a)} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \end{aligned} \quad (33)$$

Proof. The proof follows a similar approach to Kumar *et al.* [13]. Firstly, we employ the triangle inequality to separate the reward return and the dynamic properties:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &= \frac{1}{1-\gamma} \left| \int_{\mathcal{S} \times \mathcal{A}} \left[d_{\widehat{\mathcal{M}}}^{\pi}(s) \pi(a|s) r_{\widehat{\mathcal{M}}} - d_{\mathcal{M}}^{\pi}(s) \pi(a|s) r_{\mathcal{M}} \right] ds da \right| \\ &\leq \frac{1}{1-\gamma} \left| \int_{\mathcal{S} \times \mathcal{A}} d_{\widehat{\mathcal{M}}}^{\pi}(s) \pi(a|s) [r_{\widehat{\mathcal{M}}}(s, a) - r_{\mathcal{M}}(s, a)] ds da \right| \\ &\quad + \frac{1}{1-\gamma} \left| \int_{\mathcal{S} \times \mathcal{A}} (d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)) \pi(a|s) r_{\mathcal{M}}(s, a) ds da \right| \end{aligned} \quad (34)$$

Based on Assumption 1, it can express the first term above as:

$$\left| \int_{\mathcal{S} \times \mathcal{A}} d_{\widehat{\mathcal{M}}}^{\pi} \pi [r_{\widehat{\mathcal{M}}} - r_{\mathcal{M}}] ds da \right| \leq \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} \left[\frac{C_{r,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \quad (35)$$

Following the proof of Kumar *et al.* [13] on the upper bound of $|d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)|$, we have:

$$\begin{aligned} & |d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)| \\ &\leq \int_{\mathcal{S} \times \mathcal{S}'} \left| \int_{\mathcal{A}} (T_{\widehat{\mathcal{M}}} - T_{\mathcal{M}}) \pi da \right| d_{\widehat{\mathcal{M}}}^{\pi}(s) ds ds' \end{aligned} \quad (36)$$

From Assumption 1, it can deduce that $|d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)| \leq \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} \left[\frac{C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right]$, resulting in an upper bound for the second term in Eq. (34):

$$\begin{aligned} & \left| \int_{\mathcal{S} \times \mathcal{A}} [d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)] \pi(a|s) r_{\mathcal{M}}(s, a) ds da \right| \\ &\leq \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} \left[\frac{R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \end{aligned} \quad (37)$$

Further derivation leads to:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \end{aligned} \quad (38)$$

Lemma 2 is thus proven.

Lemma 3: For any $\mathcal{M}$ and an MDP $\widehat{\mathcal{M}}$ learned on the offline dataset, their performance on policy π satisfies:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} [\varepsilon_r(s, a) + R_{\max} D(s, a)] \end{aligned} \quad (39)$$

Proof. The proof follows a similar process as in Lemma 2. First, it can deduce:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &\leq \frac{1}{1-\gamma} \left| \int_{\mathcal{S} \times \mathcal{A}} d_{\widehat{\mathcal{M}}}^{\pi}(s) \pi(a|s) [r_{\widehat{\mathcal{M}}}(s, a) - r_{\mathcal{M}}] ds da \right| \\ &\quad + \frac{1}{1-\gamma} \left| \int_{\mathcal{S} \times \mathcal{A}} (d_{\widehat{\mathcal{M}}}^{\pi}(s) - d_{\mathcal{M}}^{\pi}(s)) \pi(a|s) r_{\mathcal{M}} ds da \right| \end{aligned} \quad (40)$$

Let $|r_{\widehat{\mathcal{M}}}(s, a) - r_{\mathcal{M}}(s, a)| = \varepsilon_r(s, a)$ denote the reward prediction error, and $|(T_{\widehat{\mathcal{M}}}(s' | s, a) - T_{\mathcal{M}}(s' | s, a))|_1 =$

$D(s, a)$ denote the state transition prediction error. Consequently, it can obtain:

$$\begin{aligned} & |J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{\mathcal{M}}}^{\pi}} [\varepsilon_r(s, a) + R_{\max} D(s, a)] \end{aligned} \quad (41)$$

Lemma 3 is thus proven. Let $\widehat{Q}^{\pi}(s, a)$ denote the fixed point of Eq. (14). The policy update for the DOMAIN method is expressed as follows:

$$\pi^* = \max_{\pi} \left[J(\mathcal{M}_f, \pi) - \frac{\lambda}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}_f}^{\pi}} [\eta(s, a)] \right] \quad (42)$$

where $J(\mathcal{M}_f, \pi) = (1-f)J(\widehat{\mathcal{M}}, \pi) + fJ(\widehat{\mathcal{M}}, \pi)$, $\mathcal{M}_f := (1-f)\widehat{\mathcal{M}} + f\mathcal{M}$, and $\eta(s, a) = (\omega(s, a) - d(s, a))/d_{\beta}(s, a)$. Since π^* represents the optimal value in Eq. (42), and let $\varpi(\pi, f) = \mathbb{E}_{(s,a) \sim d_{\mathcal{M}_f}^{\pi}} [\eta(s, a)]$, we have:

$$J(\mathcal{M}_f, \pi^*) - \frac{\lambda \varpi(\pi^*, f)}{1-\gamma} \geq J(\mathcal{M}_f, \pi_b) - \frac{\lambda \varpi(\pi_b, f)}{1-\gamma} \quad (43)$$

In order to compare the performance of the learned policy π^* and the behavioral policy π_b on the actual MDP, Theorem 4 employs the triangle inequality to deduce:

$$\begin{aligned} & |J(\mathcal{M}_f, \pi) - J(\mathcal{M}, \pi)| \\ &= |(1-f)J(\widehat{\mathcal{M}}, \pi) + fJ(\widehat{\mathcal{M}}, \pi) - J(\mathcal{M}, \pi)| \\ &\leq (1-f)|J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| + f|J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \end{aligned} \quad (44)$$

By utilizing Lemma 2-3, it can infer:

$$\begin{aligned} & |J(\mathcal{M}_f, \pi) - J(\mathcal{M}, \pi)| \\ &\leq (1-f)|J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| + f|J(\mathcal{M}, \pi) - J(\widehat{\mathcal{M}}, \pi)| \\ &\leq \frac{1-f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \\ &\quad + \frac{f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi}} [\varepsilon_r(s, a) + R_{\max} D(s, a)] \end{aligned} \quad (45)$$

Substituting Eq. (45) into (43), it can further derive:

$$\begin{aligned} & J(\mathcal{M}, \pi^*) - J(\mathcal{M}, \pi_b) \\ &\geq \frac{\lambda}{1-\gamma} [\varpi(\pi^*, f) - \varpi(\pi_b, f)] \\ &\quad - \frac{1-f}{1-\gamma} \left\{ \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \right. \\ &\quad \left. + \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi_b}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s, a)|}} \right] \right\} \\ &\quad - \frac{f}{1-\gamma} \left\{ \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} [\varepsilon_r(s, a) + R_{\max} D(s, a)] \right. \\ &\quad \left. + \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi_b}} [\varepsilon_r(s, a) + R_{\max} D(s, a)] \right\} \end{aligned} \quad (46)$$

The behavioral policy π_b is obtained from the offline dataset $\mathcal{D}_{\text{offline}}$, resulting in a larger $\sqrt{|\mathcal{D}(s, a)|}$ under the distribution $d_{\mathcal{M}}^{\pi_b}$, indicating smaller sampling errors. Additionally, since the environment model is trained using $\mathcal{D}_{\text{offline}}$, the model error

is smaller under the distribution $d_{\mathcal{M}}^{\pi_b}$. Therefore, Eq. (46) can be further simplified as:

$$\begin{aligned}
& J(\mathcal{M}, \pi^*) - J(\mathcal{M}, \pi_b) \\
& \geq \frac{\lambda}{1-\gamma} [\varpi(\pi^*, f) - \varpi(\pi_b, f)] \\
& \quad - 2 \frac{1-f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} \left[\frac{C_{r,\delta} + R_{\max} C_{T,\delta}}{\sqrt{|\mathcal{D}(s,a)|}} \right] \\
& \quad - 2 \frac{f}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\mathcal{M}}^{\pi^*}} [\varepsilon_r(s,a) + R_{\max} D(s,a)]
\end{aligned} \tag{47}$$

Theorem 4 is proven.

REFERENCES

- [1] D. Kalashnikov *et al.*, “Scalable deep reinforcement learning for vision-based robotic manipulation,” in *Conf. Robot Learn.*, 2018, pp. 651-673.
- [2] Y. Wen, J. Si, A. Brandt, X. Gao, and H. H. Huang, “Online reinforcement learning control for the personalization of a robotic knee prosthesis,” *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2346-2356, 2019.
- [3] F. Yu *et al.*, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2636-2645.
- [4] A. Swaminathan and T. Joachims, “Batch learning from logged bandit feedback through counterfactual risk minimization,” *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1731-1755, 2015.
- [5] H. Mao, M. Schwarzkopf, S. B. Venkatakrishnan, Z. Meng, and M. Alizadeh, “Learning scheduling algorithms for data processing clusters,” in *Pro. ACM Special Interest Group Data Commun.*, 2019, pp. 270-288.
- [6] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, “Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications,” *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826-3839, 2020.
- [7] J. Chen, and W. Xu, “Policy Gradient From Demonstration and Curiosity,” *IEEE Trans. Cybern.*, vol. 53, no. 8, pp. 4923-4933, 2023.
- [8] S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline reinforcement learning: Tutorial, review, and perspectives on open problems,” 2020. [Online]. Available: arXiv:2005.01643.
- [9] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, “Stabilizing off-policy q-learning via bootstrapping error reduction,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11784-11794.
- [10] S. Fujimoto, D. Meger, and D. Precup, “Off-policy deep reinforcement learning without exploration,” in *Int. Conf. Mach. Learn.*, 2019, pp. 2052-2062.
- [11] Y. Wu, G. Tucker, and O. Nachum, “Behavior regularized offline reinforcement learning,” 2019. [Online]. Available: arXiv:1911.11361.
- [12] C. Zhang, S. Kuppannagari, and P. Viktor, “Brac+: Improved behavior regularized actor critic for offline reinforcement learning,” in *Asian Conf. Mach. Learn.*, 2021, pp. 204-219.
- [13] A. Kumar, A. Zhou, G. Tucker, and S. Levine, “Conservative q-learning for offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1179-1191.
- [14] S. Fujimoto and S. S. Gu, “A minimalist approach to offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 20132-20145.
- [15] J. Lyu, X. Ma, X. Li, and Z. Lu, “Mildly conservative Q-learning for offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1711-1724.
- [16] L. Kaiser *et al.*, “Model Based Reinforcement Learning for Atari,” in *Int. Conf. Learn. Represent.*, 2019.
- [17] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, “A survey on model-based reinforcement learning,” 2022. [Online]. Available: arXiv:2206.09328.
- [18] W.-C. Jiang, V. Narayanan, and J.-S. Li, “Model learning and knowledge sharing for cooperative multiagent systems in stochastic environment,” *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 5717-5727, 2020.
- [19] T. Yu *et al.*, “Mopo: Model-based offline policy optimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 14129-14142.
- [20] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, “Morel: Model-based offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21810-21823.
- [21] T. Yu, A. Kumar, R. Rafailov, A. Rajeswaran, S. Levine, and C. Finn, “Combo: Conservative offline model-based policy optimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 28954-28967.
- [22] O. Kroemer, S. Niekum, and G. Konidaris, “A review of robot learning for manipulation: Challenges, representations, and algorithms,” *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1395-1476, 2021.
- [23] M. Janner, J. Fu, M. Zhang, and S. Levine, “When to trust your model: Model-based policy optimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12519-12530.
- [24] M. Rigter, B. Lacerda, and N. Hawes, “Rambo-rl: Robust adversarial model-based offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16082-16097.
- [25] M. Bhardwaj, T. Xie, B. Boots, N. Jiang, and C.-A. Cheng, “Adversarial model for offline reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023.
- [26] R. SUTTON, “Policy gradient method for reinforcement learning with function approximation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057-1063.
- [27] T. Haarnoja *et al.*, “Soft actor-critic algorithms and applications,” 2018. [Online]. Available: arXiv:1812.05905.
- [28] H. Niu, Y. Qiu, M. Li, G. Zhou, J. HU, and X. Zhan, “When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 36599-36612.
- [29] H. Li, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-Q. Feng, and Z.-G. Hou, “CASOG: Conservative Actor-critic with SmOoth Gradient for Skill Learning in Robot-Assisted Intervention,” 2023. [Online]. Available: arXiv:2304.09632.
- [30] B. Eysenbach, S. Asawa, S. Chaudhari, S. Levine, and R. Salakhutdinov, “Off-dynamics reinforcement learning: Training for transfer with domain classifiers,” 2020. [Online]. Available: arXiv:2006.13916.
- [31] J. Li, C. Tang, M. Tomizuka, and W. Zhan, “Dealing with the unknown: Pessimistic offline reinforcement learning,” in *Conf. Robot Learn.*, 2022, pp. 1455-1464.
- [32] R. Laroche, P. Trichelair, and R. T. Des Combes, “Safe policy improvement with baseline bootstrapping,” in *Int. Conf. Mach. Learn.*, 2019, pp. 3652-3661.
- [33] M. Ghavamzadeh, M. Petrik, and Y. Chow, “Safe policy improvement by minimizing robust baseline regret,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2306-2314.
- [34] I. Osband and B. Van Roy, “Why is posterior sampling better than optimism for reinforcement learning?,” in *Int. Conf. Mach. Learn.*, 2017, pp. 2701-2710.
- [35] G. Brockman *et al.*, “Openai gym,” 2016. [Online]. Available: arXiv:1606.01540.
- [36] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, “D4rl: Datasets for deep data-driven reinforcement learning,” 2020, [Online]. Available: arXiv:2004.07219.
- [37] I. Kostrikov, A. Nair, and S. Levine, “Offline Reinforcement Learning with Implicit Q-Learning,” in *Int. Conf. Learn. Represent.*, 2021.
- [38] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094-2100.
- [39] X.-H. Chen *et al.*, “Offline Model-Based Adaptable Policy Learning for Decision-Making in Out-of-Support Regions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15260-15274, 2023.
- [40] L. Huang, B. Dong, J. Lu and W. Zhang, “Mild policy evaluation for offline Actor-Critic,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2023, doi: 10.1109/TNNLS.2023.3309906.
- [41] Q. Yang, S. Wang, Q. Zhang, G. Huang and S. Song, “Hundreds Guide Millions: Adaptive Offline Reinforcement Learning With Expert Guidance,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2023, doi: 10.1109/TNNLS.2023.3293508.
- [42] J. Wang, W. Li, H. Jiang, G. Zhu, S. Li, C. Zhang, “Offline reinforcement learning with reverse model-based imagination,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29420-29432.
- [43] X. Chen *et al.*, “Bail: Best-action imitation learning for batch deep reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18353-18363.