Contrastive Representation for Data Filtering in Cross-Domain Offline Reinforcement Learning

Xiaoyu Wen [1] **Chenjia Bai** [2 3] **Kang Xu** [4] **Xudong Yu** [5] **Yang Zhang** [6] **Xuelong Li** [7] **Zhen Wang** [1]

Abstract

Cross-domain offline reinforcement learning leverages source domain data with diverse transition dynamics to alleviate the data requirement for the target domain. However, simply merging the data of two domains leads to performance degradation due to the dynamics mismatch. Existing methods address this problem by measuring the dynamics gap via domain classifiers while relying on the assumptions of the transferability of paired domains. In this paper, we propose a novel representation-based approach to measure the domain gap, where the representation is learned through a contrastive objective by sampling transitions from different domains. We show that such an objective recovers the mutual-information gap of transition functions in two domains without suffering from the unbounded issue of the dynamics gap in handling significantly different domains. Based on the representations, we introduce a data filtering algorithm that selectively shares transitions from the source domain according to the contrastive score functions. Empirical results on various tasks demonstrate that our method achieves superior performance, using only 10% of the target data to achieve 89.2% of the performance on 100% target dataset with state-of-the-art methods.

1. Introduction

Offline Reinforcement Learning (RL) (Lange et al., 2012; Fujimoto et al., 2019; Levine et al., 2020; Bai et al., 2022; Yang et al., 2022b; Bai et al., 2024) exhibits a distinctive advantage over online RL, leveraging previously collected offline data without requiring any additional online interactions. In real-world scenarios like robotic manipulation (Feng et al., 2023; Shi et al., 2024), autonomous driving (Zhang et al., 2023b), and healthcare (Fatemi et al., 2022), gathering a substantial offline dataset with good coverage of transitions for a specific environment is time-consuming and expensive (Alberti et al., 2020; Kuo et al., 2021; Walke et al., 2023). Nevertheless, the offline RL algorithms rely heavily on the data coverage of the offline dataset (Zhan et al., 2022; Deng et al., 2023), and the performance degenerates significantly if the amount of offline data decreases. To tackle this challenge for a specific target domain with scarce data, cross-domain offline RL leverages additional source domain data with dynamics shift to compensate for the (target) offline dataset (Liu et al., 2022; 2024). However, as we illustrated in Figure 1(a), simply combining the dataset from source and target domains induces a significant dynamics shift due to the dynamics mismatch, leading to policy divergence and poor performance (Yu et al., 2021; 2022). Therefore, how to appropriately incorporate source domain data to improve the data efficiency for learning in the target domain remains a challenge.

There are two key problems for cross-domain offline RL: how to *measure the domain gap* and how to *utilize the cross-domain data*. For the first problem, prior methods directly estimate the dynamics models with offline datasets or training domain discriminators to approximate the dynamics gap. Nevertheless, the dynamics model suffers from large extrapolation errors given limited target domain data, and domain discriminators fail to provide smooth measurement for the dynamics discrepancy. For example, the dynamics gap (i.e., $\log P_{\text{source}}/P_{\text{target}}[s'|s,a]$) can be unbounded when the two domains mismatch significantly (Xu et al., 2023). For the second problem, previous approaches modify the rewards according to the estimation of dynamics discrepancy (Liu et al., 2022) or employ pessimistic supported constraints for the source domain data (Liu et al., 2024). Despite these progresses, these methods typically experience rapid performance degradation when confronted with a larger dynamics gap, as shown in Figure 1(b).

In this paper, we propose a novel perspective to measure the domain gap via the mutual information (MI) of transitions.

[1]Northwestern Polytechnical University [2]Shanghai Artificial Intelligence Laboratory [3]Shenzhen Research Institute of Northwestern Polytechnical University [4]Fudan University [5]Harbin Institute of Technology [6]Tsinghua University [7]The Institute of Artificial Intelligence (TeleAI), China Telecom. Correspondence to: Chenjia Bai <baichenjia@pjlab.org.cn>, Zhen Wang <w-zhen@nwpu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

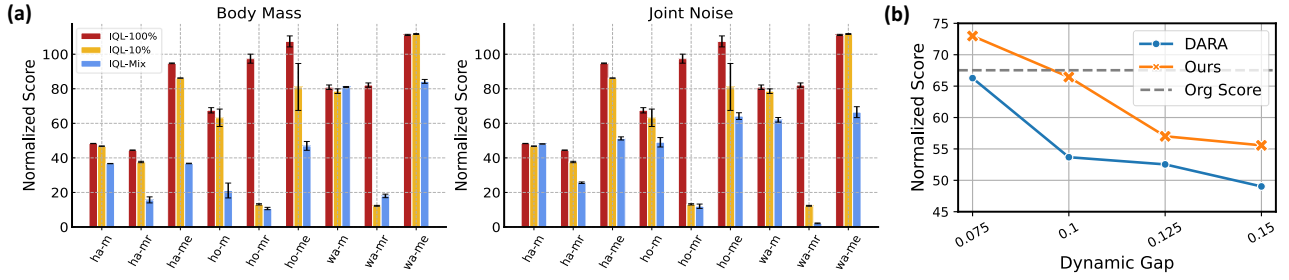

Figure 1. (a) Comparison of performance across five seeds in nine Mujoco tasks (ha: halfcheetah, ho: hopper, wa: walker2d, m: medium, mr: medium-replay, me: medium-expert) with IQL (Kostrikov et al., 2022). We set standard D4RL (Fu et al., 2020) as the target domain data. For the source domain, we modify environmental parameters, such as altering body mass or introducing joint noise, and then collect offline datasets in the modified environments. (IQL-100%: use 100% target data, IQL-10%: use reduced 10% target data, IQL-Mix: use 10% target data and 100% source data.) (b) Comparison of performance between our algorithm and DARA (Liu et al., 2022) with 100% source-domain dataset and 10% target-domain dataset in the Hopper-Medium-v2 when facing the increasing dynamics gap. Specifically, we simulate a process of increasing dynamics gaps by continuously increasing the head size in the Hopper-v2 environment. The x-axis is the head size of the Hopper-v2 (normal size is 0.05), and "Org Score" is the original performance of IQL when using 100% target data.

Specifically, we adopt the MI between the joint distribution of state-action pairs and the next states to capture the underlying dynamics of environmental transitions. Then, we use the MI gap between the source and target domains as a robust characterization of domain discrepancy when the data is shared from a significantly different source domain. In practice, such an MI gap can be estimated via a contrastive objective by using the positive samples from the target domain and the negative samples from the source domain. We employ the learned contrastive representation that captures the domain-distinguishable information as a data filter, which selectively shares the transitions from the source domain with small MI gaps to the target domain. Theoretical analysis shows that reducing the MI gap via data filtering reduces the performance bounds of two domains. Under mild assumptions, the proposed MI gap also recovers the expected dynamics gap without explicit dynamic estimators or domain discriminators.

We name the proposed method the Info-Gap Data Filtering (**IGDF**) algorithm. Empirically, we evaluate IGDF in various D4RL environments (Fu et al., 2020) with kinematic and morphology shifts (Liu et al., 2022; Xu et al., 2023), showcasing its superior performance compared to previous state-of-the-art algorithms. As an example of cross-domain offline RL in Figure 1(b), the MI gap used in IGDF is more robust than the dynamics gap in DARA (Liu et al., 2022), especially for shared domains with large dynamics gaps. Our code is available in this repository (`https://github.com/BattleWen/IGDF`).

2. Preliminaries

The RL problem is typically formulated as a Markov Decision Process (MDP), defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \hat{\rho}_0)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s'|s,a)$ is the transition dynamics, $r(s,a)$ is the reward function, $\gamma \in [0,1)$ is the discount factor, and $\hat{\rho}_0 : \mathcal{S} \to [0,1]$ is the initial state distribution.

In the offline RL setting, the agent does not interact with the environment and learns a policy from an offline dataset (Levine et al., 2020). Considering a target MDP $\mathcal{M}_{\text{tar}} = (\mathcal{S}, \mathcal{A}, P_{\text{tar}}, r, \gamma, \hat{\rho}_0)$ has limited dataset $\mathcal{D}_{\text{tar}}$. In cross-domain offline RL, we assume to access another offline dataset $\mathcal{D}_{\text{src}}$ collected on a source domain MDP $\mathcal{M}_{\text{src}} = (\mathcal{S}, \mathcal{A}, P_{\text{src}}, r, \gamma, \hat{\rho}_0)$. We assume that all of these MDPs share the same state space, action space, and reward function and only differ in the transition probabilities, i.e., $P_{\text{src}}(s'|s,a)$ and $P_{\text{tar}}(s'|s,a)$. The goal of cross-domain offline RL is to leverage the additional source-domain dataset $\mathcal{D}_{\text{src}}$ to relax the data requirements of the target domain. The policy is learned to maximize the expected return over the target environment $\mathcal{M}_{\text{tar}}$ using the static cross-domain offline data $\mathcal{D}_{\text{mix}} := \mathcal{D}_{\text{src}} \cup \mathcal{D}_{\text{tar}}$.

In the offline setting, we further define the empirical MDP that estimates the expectation of the transition function $P(s'|s,a)$ from the offline dataset. Formally, an empirical MDP estimated from $\mathcal{D}$ is $\widehat{\mathcal{M}} := (\mathcal{S}, \mathcal{A}, \hat{P}, r, \gamma, \hat{\rho}_0)$, where $\hat{P} = \max_{\hat{P}} \mathbb{E}_{s,a,s' \sim \mathcal{D}} \left[\log \hat{P}(s' \mid s,a)\right]$ is estimated by the maximum log-likelihood, and $\hat{P}(s' \mid s,a) = 0$ for all (s,a,s') not in dataset $\mathcal{D}$. Then the empirical MDPs for the source domain and target domain are defined as $\widehat{\mathcal{M}}_{\text{src}} = (\mathcal{S}, \mathcal{A}, \hat{P}_{\text{src}}, r, \gamma, \hat{\rho}_0)$ and $\widehat{\mathcal{M}}_{\text{tar}} = (\mathcal{S}, \mathcal{A}, \hat{P}_{\text{tar}}, r, \gamma, \hat{\rho}_0)$, respectively. We assume the two datasets follow the same behavior policy $\pi^b(a|s)$ (refer to Appendix F for more details). In source MDP, $\hat{\rho}_{\text{src}}(s)$ is the normalized probability that the policy π^b_{src} encounters state s, defined as

$\hat{\rho}_{\text{src}}(s) \triangleq (1-\gamma)\sum_{t=1}^{\infty}\gamma^t \hat{P}_{\text{src}}(s_t = s|\pi^b)$, and $\hat{\rho}_{\text{tar}}(s)$ for the target domain follows a similar formulation.

3. Methodology

In this section, we first introduce the proposed MI gap, which measures the domain gap in cross-domain offline RL. Then we give a contrastive objective to estimate such a gap with learned representations. Next, we give a data filtering method to leverage the source domain data based on the representations and score functions. Finally, we give the theoretical analysis for the proposed algorithm.

3.1. The MI Gap for Cross-Domain RL

In the following, we denote the information measure $I(\cdot;\cdot)$ as MI and $H(\cdot)$ as Shannon entropy. We use the uppercase letter (e.g., X) for random variables and the lowercase letter (e.g., x) for their realizations. We aim to adopt the MI term to capture the dynamics-relevant information about different domains. For a distribution over the transition tuple (s, a, s'), we use S, A, S' to stand for the corresponding random variables. We also use p to denote the joint distribution of these variables as well as their associated marginals. Then the MI between the state-action pair (S, A) and their future state S' is defined as

$$I([S, A]; S') = \mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[\log \frac{p(s, a, s')}{p(s, a)p(s')}\right], \quad (1)$$

where $p(s, a)$, $p(s')$ and $p(s, a, s')$ follow empirical distributions according to the offline dataset $\mathcal{D}$. For the source domain and target domain with different datasets (i.e., $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tar}}$), we denote the MI objective estimated in two domains as $I_{\text{src}}([S, A]; S')$ and $I_{\text{tar}}([S, A]; S')$, respectively. Then the MI gap between the two domains is defined as

$$\Delta I = I_{\text{tar}}([S, A]; S') - I_{\text{src}}([S, A]; S'), \quad (2)$$

where the two MI terms follow the different conditional and marginal probabilities. Specifically, we have

$$\begin{aligned} I_{\text{tar}}([S, A]; S') &= \mathbb{E}_{\mathcal{D}}\big[\log p(s, a, s')/[p(s, a)p(s')]\big] \\ &= \mathbb{E}_{\mathcal{D}}\big[\log \hat{P}_{\text{tar}}(s'|s, a)/\hat{\rho}_{\text{tar}}(s')\big], \end{aligned} \quad (3)$$

where $\hat{P}_{\text{tar}}(s'|s, a)$ is the empirical transition function in the target-domain dataset, and $\hat{\rho}_{\text{tar}}(s')$ denotes the normalized state distribution. We utilize maximum likelihood estimation in a given dataset $\mathcal{D}$ to fit $\hat{P}_{\text{tar}}(s' \mid s, a)$. If we denote the parameter of the empirical distribution $\hat{P}_{\text{tar}}$ by θ, then the empirical distribution can be obtained by $\hat{P}_{\text{tar}}(s' \mid s, a) = \text{argmax}_\theta \sum_{(s_i,a_i,s_i')\sim\mathcal{D}} \log \hat{P}_{\text{tar}}(s_i'|s_i, a_i; \theta)$. The expectation in Eq. (3) follows $(s, a, s') \sim \mathcal{D}$, where $\mathcal{D}$ is the actual dataset for sampling transitions. For example, $\mathcal{D} = \mathcal{D}_{\text{tar}}$ when the policy is trained with the target-domain dataset.

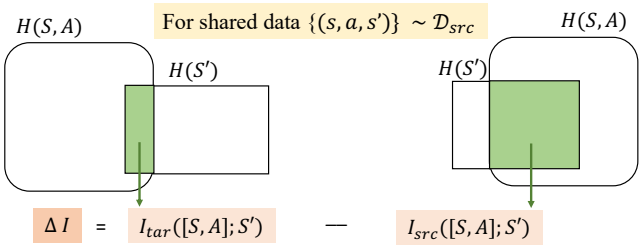

Figure 2. An illustration of the MI gap of data shared from $\mathcal{D}_{\text{src}}$.

In contrast, if the shared data from the source domain is used for training the target-domain policy, then we have $\mathcal{D} = \mathcal{D}_{\text{src}}$. The MI term I_{src} of the source domain follows a similar form as Eq. (3), but with $\hat{P}_{\text{src}}(s'|s, a)$ and $\hat{\rho}_{\text{src}}(s')$ that are estimated in the source-domain dataset.

When the two domains are significantly different, the proposed MI gap ΔI is more robust than the dynamics ratio (i.e., $\Delta P = \mathbb{E}_{\mathcal{D}_{\text{src}}}[\log \hat{P}_{\text{tar}}/\hat{P}_{\text{src}}]$) in cross-domain data sharing. Specifically, when the samples $\{(s, a, s')\}$ from $\mathcal{D}_{\text{src}}$ are shared to the target domain, the probability of $\hat{P}_{\text{tar}}(s'|s, a) \to 0$ since the two domains have very different transition functions, which makes $\Delta P \to -\infty$. In contrast, the ΔI term is lower-bounded by the state entropy of behavior policies, as $\Delta I \geq -I_{\text{src}}([S, A]; S') \geq -H(\hat{\rho}_{\text{src}}(s'))$. An illustration of the MI gap with significantly different domains is shown in Figure 2.

3.2. Contrastive Representation for the MI Gap

To estimate the MI gap in high-dimensional state space, a tractable variational estimator based on neural networks is required (Poole et al., 2019; van den Oord et al., 2018; Yang et al., 2023). We adopt contrastive learning to estimate the MI objective. A naive approach requires two independent estimators for I_{tar} and I_{src} separately. In contrast, we simplify this process by adopting a single contrastive objective to estimate ΔI directly. Specifically, we choose transitions $(s, a, s'_B) \sim \mathcal{D}_{\text{tar}}$ from the target domain as positive samples. The negative samples are obtained by first sampling a state-action pair $(s, a) \sim \mathcal{D}_{\text{tar}}$ and then sampling a negative state set S'^- from the source domain $\mathcal{D}_{\text{src}}$ independently. Then a negative sample is obtained by concatenating them together to form a tuple (s, a, s'_A), where $s'_A \in S'^-$. The contrastive objective can be expressed as

$$\begin{aligned} \mathcal{L}_{\text{NCE}} = -\mathbb{E}_{p(s,a,s'_B)}\mathbb{E}_{S'^-} \\ \left[\log \frac{h_2(s, a, s'_B)}{h_2(s, a, s'_B) + \sum_{s'_A\in S'^-} h_1(s, a, s'_A)}\right], \end{aligned} \quad (4)$$

where we use two score functions to measure the information density ratio which preserves the MI between (s, a) and s' for the source and target domains, respectively. Intuitively, the score functions assign scores representing an

exponential correlation between the state-action pair and the next state in the corresponding domains. Formally, we aim to approximate the information density of the target domain $h_2(s, a, s'_B) \propto \hat{P}_{\text{tar}}(s'_B|s, a)/\hat{\rho}_{\text{tar}}(s'_B)$ and source domain $h_1(s, a, s'_A) \propto \hat{P}_{\text{src}}(s'_A|s, a)/\hat{\rho}_{\text{src}}(s'_A)$, respectively (van den Oord et al., 2018).

The following theorem shows that the proposed contrastive objective serves as an approximate estimation for the MI gap with sufficient negative samples.

Theorem 3.1 (InfoNCE extension)**.** *The MI gap* $\Delta I = I_{\text{tar}}([S, A]; S') - I_{\text{src}}([S, A]; S')$ *can be lower bounded by the negative contrastive objective, as*

$$\Delta I \geq \log(K - 1) - \mathcal{L}_{\text{NCE}} \coloneqq I_{\text{NCE}}, \tag{5}$$

where $K - 1$ *is the number of negative samples from the source domain.*

The term I_{NCE} is an asymptotically tight lower bound for the MI gap, i.e., $\lim_{K\to\infty} I^K_{\text{NCE}}(X; Y) \to \Delta I(X; Y)$, which becomes tighter as K becomes larger (Poole et al., 2019; Guo et al., 2022). We refer to §A.1 for full derivations. We illustrate the contrastive learning process in Figure 3.

For optimizing the contrastive objective in Eq. (4), we adopt two score functions (i.e., h_1 and h_2) to estimate the information density of source and target domains, respectively. (i) $h_2(\cdot)$ only takes samples $\{(s, a, s'_B)\}$ from the target domain as inputs and assigns **high** scores to $h_2(s, a, s'_B)$, (ii) $h_1(\cdot)$ only assigns **low** scores to $h_1(s, a, s'_A)$ for samples $\{(s, a, s'_A)\}$, where $(s, a) \in \mathcal{D}_{\text{tar}}$ and $s'_A \in \mathcal{D}_{\text{src}}$. Interestingly, compared to training two independent contrastive estimators for I_{src} and I_{tar}, our objective in Eq. (4) has neither negative samples for score function h_2, nor positive samples for score function h_1, which provide an opportunity to further integrate the effects of h_1 and h_2 into a single score function h. The new score function h uses $\{(s, a, s'_B\}$ from the target domain as positive samples and constructed transitions $\{(s, a, s'_A)\}$ from two domains as negative samples. Then we have a simplified objective as

$$\widehat{\mathcal{L}}_{\text{NCE}} = -\mathbb{E}_{p(s,a,s'_B)}\mathbb{E}_{S'-}\left[\log \frac{h(s, a, s'_B)}{\sum_{s'_A \in S'-\cup s'_B} h(s, a, s'_A)}\right], \tag{6}$$

which serve as a simplified version of $\mathcal{L}_{\text{NCE}}$. The main reason of using a single score function is that we only share data from the source domain to the target domain (i.e., $\mathcal{D}_{\text{src}} \to \mathcal{D}_{\text{tar}}$) without a reverse data-sharing process. Thus, we do not require an independent score function to distinguish whether a transition comes from the source domain, but only required to distinguish whether a shared transition is similar to the data distribution of the target domain.

For implementation, we use two neural networks $\phi(s, a)$ and $\psi(s')$ to learn representations of state-action pairs and states only. Then we adopt a linear parameterization as

$$h(s, a, s') = \exp(\phi(s, a)^\top \psi(s')), \tag{7}$$

which resembles spectral decomposition and low-rank representation of transition functions (Uehara et al., 2022; Ren et al., 2023b;a), while we use $h(s, a, s')$ to approximate the information density. The representations are normalized to $\|\phi(\cdot)\|, \|\psi(\cdot)\| = 1$, which makes $h(\cdot) \in [1/e, e]$. In cross-domain data sharing, the score function h assigns high scores for source domain data that follows a similar information density (i.e., $\hat{P}_{\text{tar}}(s'_B|s, a)/\hat{\rho}_{\text{tar}}(s'_B)$) to the target domain data and assigns low scores for transitions that have significantly different distributions to the target domain.

Algorithm 1 IGDF: Info-Gap Data Filtering algorithm

Input: Source domain data $\mathcal{D}_{\text{src}}$, target domain data $\mathcal{D}_{\text{tar}}$
Initialize: Policy π, value function Q, encoders $\phi(s, a)$, $\psi(s')$, data filter ratio ξ, importance ratio α, batch size B

// Contrastive Representation Learning
Optimizing the contrastive objective in Eq. (6) to train the encoder networks $\phi(s, a)$ and $\psi(s')$
// Data Filtering algorithm
for each gradient step **do**
  Sample a batch $b_{\text{src}} := \{(s, a, r, s')\}^{\frac{B}{2\xi}}$ from $\mathcal{D}_{\text{src}}$
  Sample a batch $b_{\text{tar}} := \{(s, a, r, s')\}^{\frac{B}{2}}$ from $\mathcal{D}_{\text{tar}}$
  Sample the top-ξ samples from b_{src} ranked by $h(\cdot)$
  Combine top-ξ samples in b_{src} and all samples in b_{tar}
  Optimize the value function Q_θ via Eq. (8)
  Learn the policy $\pi(a|S)$ via offline RL algorithms
end for

3.3. Data Filtering via Contrastive Representation

Based on the representations and score function, we obtain a practical data filtering algorithm, termed IGDF (**I**nfo-**G**ap **D**ata **F**iltering), that leverages additional data with dynamics gap from the source domain to train a policy for the target MDP. Specifically, after training the encoder networks $\phi(\cdot)$ and $\psi(\cdot)$ by optimizing $\widehat{\mathcal{L}}_{\text{NCE}}$, we sample a batch of data $\{(s_A, a_A, s'_A)\}$ from $\mathcal{D}_{\text{src}}$ and rank the transitions according to the value of $\phi(s_A, a_A)^\top \psi(s'_A)$, then we extract the top ξ-quantile of batch samples for data sharing. The shared data is mixed with a batch of target domain data for offline RL training. The algorithmic description of IGDF is presented in Algorithm 1. In practice, data sharing is more convenient than modifying the reward function of shared data for pessimism, as it eliminates the need for meticulous adjustments to clip ranges and reward scaling ratios.

To further enhance the performance, we propose a variant of our method by weighting the Temporal-Difference (TD)-error of filtered data using the score function. Formally, we

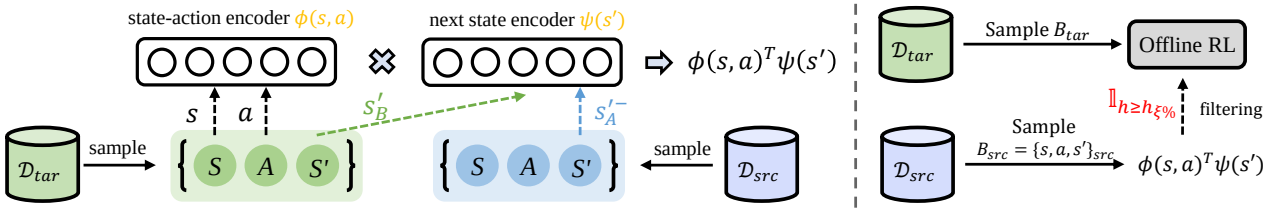

Figure 3. Illustration of our method. (a) We train two encoder networks using contrastive learning, treating target transitions as positive examples and constructed transitions as negative examples. (b) We tackle cross-domain offline RL by selectively sharing the source domain data with the score functions. The target data and the share data are used for offline RL algorithms to learn the policy.

train the value function as

$$\mathcal{L}_Q(\theta) = \frac{1}{2}\mathbb{E}_{(s,a,s')\sim\mathcal{D}_{\text{tar}}}\left[\left(Q_\theta - \mathcal{T}Q_\theta\right)^2\right] + \frac{1}{2}\alpha\cdot h(s,a,s')\mathbb{E}_{(s,a,s')\sim\mathcal{D}_{\text{src}}}\left[\omega(s,a,s')\left(Q_\theta - \mathcal{T}Q_\theta\right)^2\right], \tag{8}$$

where α is the importance coefficient for weighting the TD-error with the score function, and $\omega(s,a,s') := \mathbb{1}\left(h(s,a,s') > h_{\xi\%}\right)$ perform data filtering to extract samples with top ξ-quantile scores in the mini-batch sampled from the source domain. In Eq. (8), $\mathcal{T}Q_\theta$ is a general Bellman operator of the offline RL algorithms. It is also worth noting that IGDF can serve as an add-on module algorithm for arbitrary offline RL algorithms, and we select IQL as the base algorithm in experiments. The detailed procedure of IGDF+IQL is given in §B.

3.4. Theoretical Analysis

Connection to Dynamics Gap. The previous methods (Eysenbach et al., 2020; Liu et al., 2022) for cross-domain adaptation often adopt the dynamic ratio to measure the dynamics gap. In the following, we give a connection between the dynamics gap and the proposed MI gap with transitions from different domains.

Theorem 3.2. *For shared data from the source domain $\widehat{M}_{\text{src}}$, i.e., $(s,a,s')\in\mathcal{D}_{\text{src}}$, the relationship between the MI gap and dynamics gap is*

$$\Delta I = D_{\text{KL}}[\hat{\rho}_{\text{src}}(s')\|\hat{\rho}_{\text{tar}}(s')] - D_{\text{KL}}[\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)]. \tag{9}$$

In contrast, for data from the target domain $\widehat{M}_{\text{tar}}$, the relationship between the MI gap and dynamics gap is

$$\Delta I = D_{\text{KL}}[\hat{P}_{\text{tar}}(s'|s,a)\|\hat{P}_{\text{src}}(s'|s,a)] - D_{\text{KL}}[\hat{\rho}_{\text{tar}}(s')\|\hat{\rho}_{\text{src}}(s')]. \tag{10}$$

Then, the MI gap is bounded by

$$-H(\hat{\rho}_{\text{src}}(s')) \leq \Delta I \leq H(\hat{\rho}_{\text{tar}}(s')). \tag{11}$$

We give the detailed proof in §A.2. According to the Theorem 3.2, the decomposition of the MI gap also contains a KL-term to measure the dynamics gap. Nevertheless, the MI gap has an additional divergence term for state visitation distribution to regularize the dynamics gap. For example, as in Eq. (9), when the shared data from $\widehat{M}_{\text{src}}$ is significantly different from that of $\widehat{M}_{\text{tar}}$, the dynamics gap $-D_{\text{KL}}[\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)] \to -\infty$, while the state density ratio $D_{\text{KL}}[\hat{\rho}_{\text{src}}(s')\|\hat{\rho}_{\text{tar}}(s')] \to \infty$ counteracts this effect. Theoretically, we show the MI gap can be bounded by the entropy of state distribution, as in Eq. (11). As a result, the MI gap overcomes the drawback of the dynamics gap with large domain gaps (Liu et al., 2022) and provides a stable measurement for the domain gap.

Performance Guarantee. Built on the above analysis, we provide a theoretical guarantee for sharing the source domain data from $\widehat{M}_{\text{src}}$ to improve the performance of the true MDP M_{tar} in the target domain under the dynamic mismatch. Then we have the following performance bound for any policy π in cross-domain offline data sharing:

Theorem 3.3. *Under the setting of corss-domain offline RL, the performance difference of any policy π evaluated by the source domain $\widehat{M}_{\text{src}}$ and the true target MDP M_{tar} can be bounded as below,*

$$\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) \geq -\frac{\gamma R_{\max}}{(1-\gamma)^2}\Big\{2\mathbb{E}_{\hat{\rho}_{\text{tar}}}\Big[D_{\text{TV}}\left(P_{\text{tar}}\|\hat{P}_{\text{tar}}\right)\Big] + \sqrt{2D_{\text{KL}}\left(\hat{\rho}_{\text{src}}(s')\|\hat{\rho}_{\text{tar}}(s')\right) + 2|\Delta I|}\Big\}. \tag{12}$$

We give the detailed proof in §A.3. The first term $D_{\text{TV}}\left[P_{\text{tar}}\|\hat{P}_{\text{tar}}\right]$ of the divergence in Eq. (12) is caused by limited coverage of offline dataset and can be reduced by offline RL algorithms. The second divergence term includes the MI gap and the state distribution of the empirical MDP of source and target domains, which can be reduced by the proposed data filtering algorithms based on the MI gap. And we also provide additional details about a tight sub-optimality gap of IGDF in Appendix A.4.

4. Related Work

Dynamic adaptation in RL The problem of dynamic adaptation focuses on policy adaptation in domains with varying transition dynamics. Prior methods have proposed several design paradigms, including system identification (Fernando et al., 2013; Chebotar et al., 2019; Werbos, 1989; Wittenmark, 1995; Zhu et al., 2018) to capture the dynamics and visual properties of the real world, domain randomization methods (Peng et al., 2018; OpenAI et al., 2019; Tobin et al., 2017; Sadeghi & Levine, 2017; James et al., 2017) that introduce diversity by randomly altering simulation parameters, meta-RL (Finn et al., 2017; Clavera et al., 2019) that performs fast policy fine-tuning, and imitation learning (Chae et al., 2022; Kim et al., 2020) that learns expert policy. However, these methods require additional online interactions, offline historical transitions, or prior knowledge to select the parameters and the range of randomization. More recent works have explored the online dynamics adaptation given limited offline experiences from the target domain based on dynamics gap (Eysenbach et al., 2020) and value function (Xu et al., 2023), or via imperfect simulations from the source domain (Niu et al., 2022; 2023). In contrast, we explore cross-domain adaptation in a purely offline setting based on the MI of transitions.

Cross-domain offline RL Learning to act from a limited dataset without any possibility of improving exploration is a well-known challenge in offline RL (Wen et al., 2023; Zhang et al., 2023a). Cross-domain offline RL aims to leverage additional source domain data with dynamics shift to contribute to offline RL data efficiency. There are two preliminary problems: how to identify the dynamic discrepancy between source and target domain, and how to leverage source domain offline data. Prior works address the first problem by training two discriminators to evaluate a dynamics gap-related term (Eysenbach et al., 2020; Liu et al., 2022), employing a GAN-style discriminator (Xue et al., 2023), or directly estimating the dynamics models (Liu et al., 2024). However, the learned dynamics models suffer from large extrapolation errors given limited target domain data, and domain discriminators fail to provide reliable estimation when dynamics shifts are significant (Xu et al., 2023). To optimize the efficient reuse of source domain data, previous methods have explored various strategies, such as reward modification (Liu et al., 2022; Xue et al., 2023) and pessimistic supported constraints (Liu et al., 2024), but still encounter certain limitations. Their performance may degrade when confronted with a larger dynamics gap. In contrast, our method employs a representation-based approach to smoothly measure the dynamics gap, avoiding explicit estimation of transition probabilities. Moreover, we propose a score function-based data filtering method to selectively share source domain data, achieving comparable performance with fewer target domain data.

5. Experiments

In this section, we present empirical validations of our approach. We examine the effectiveness of our method in scenarios with various dynamics shifts. Furthermore, we provide ablation studies and qualitative analyses of our method.

5.1. Datasets and Baselines

To characterize the offline dynamics shift, we consider the *Halfcheetah*, *Hopper*, and *Walker2d* from the Gym-MuJoCo environment, using offline samples from D4RL as our target offline dataset. For the source dataset, we change the environment parameters by altering the XML file of the MuJoCo simulator following (Liu et al., 2022; Xu et al., 2023) and then collect the Medium, Medium-Replay, and Medium-Expert offline datasets in the changed environments following the same data collection procedure as in D4RL (refer to Appendix C for the details).

We compare our algorithms with three state-of-the-art baselines in the cross-domain offline RL setting: (i) DARA (Liu et al., 2022) trains a pair of binary classifiers $p(\text{target}|s,a,s')$ and $p(\text{target}|s,a)$ to evaluate dynamics gap-related transition probabilities. (ii) SRPO (Xue et al., 2023) gives a constrained optimization formulation that regards the state distribution as a regularizer. (iii) BOSA (Liu et al., 2024) proposes supported policy and value optimization, which explicitly regularizes the policy and value optimization with in-support transitions. Notably, both DARA and SRPO can be flexibly applied to a wide range of offline RL algorithms, whereas BOSA stands out as a comprehensive algorithm built upon the SPOT (Wu et al., 2022) implementation. More details are given in Appendix D.

5.2. Motivation Example

Question 1. *Is simply merging cross-domain data effective for cross-domain offline RL?*

To assess the efficacy of simply merging cross-domain data, we provide the results of different methods using single-domain offline data or cross-domain data. As shown in Table 1, we choose some typical model-based offline RL and model-free offline RL algorithms as backbones, including BCQ (Fujimoto et al., 2019), CQL (Kumar et al., 2020), MOPO (Yu et al., 2020), SPOT (Wu et al., 2022), and IQL (Kostrikov et al., 2022). In the single-domain setting, the numbers to the left of the arrow ($\rightarrow$) represent the scores trained on 100% D4RL data, and the numbers to the right of that represent the scores trained on only 10% D4RL data. In the left panel (single-domain setting), Average♠ represents the average performance change when the offline data is reduced (100% $\rightarrow$ 10%). In the right panel (cross-domain setting), Average♣ represents the average performance difference between the cross-domain results and the best results

Table 1. Results on the single-domain RL(100% D4RL → 10% D4RL) and cross-domain offline RL. We average our results over 5 seeds and for each seed, we compute the normalized average score using 10 episodes. And we take the results (single-domain setting with 100% D4RL) from their original papers. (ha: halfcheetah, ho: hopper, wa: walker2d, m: medium, mr: medium-replay, me: medium-expert.)

		single-domain setting (100% D4RL → 10% D4RL)					cross-domain setting (10% D4RL + source data)				
		BCQ	**MOPO**	**CQL**	**SPOT**	**IQL**	**BCQ**	**MOPO**	**CQL**	**SPOT**	**IQL**
body mass	ha-m	40.7 → 37.6	42.3 → 3.2	44.4 → 35.4	58.4 → 45.4	48.3 → 46.8	35.1	6.4	32.2	50.3	36.7
	ha-mr	38.2 → 1.1	53.1 → −0.1	46.2 → 0.6	52.2 → 9.8	44.5 → 37.6	40.1	10.2	3.3	37.6	15.7
	ha-me	64.7 → 37.3	63.5 → 4.2	62.4 → −3.3	86.9 → 46.2	94.7 → 86.2	26.4	8.9	12.9	33.8	36.8
	ho-m	54.5 → 37.1	28 → 4.1	58 → 43	86 → 62.5	67.5 → 63.2	25.7	5	44.9	85.96	21.1
	ho-mr	33.1 → 9.3	67.5 → 1	48.6 → 9.6	100.2 → 13.7	97.4 → 13.1	28.7	5.5	1.4	15.5	10.7
	ho-me	110.9 → 58	23.7 → 1.6	98.7 → 59.7	99.3 → 69	107.4 → 81.1	75.4	4.8	53.6	75.5	46.9
	wa-m	53.1 → 32.8	17.8 → 7	79.2 → 42.9	86.4 → 65.4	80.9 → 78.6	50.9	5.7	80	22.5	81
	wa-mr	15 → 6.9	39 → 5.1	26.7 → 4.6	91.6 → 18.6	82.2 → 12.3	14.9	3.1	0.8	16	18
	wa-me	57.5 → 32.5	44.6 → 5.3	111 → 49.5	112 → 84	111.2 → 111.7	55.2	5.5	63.5	14.3	84.3
joint noise	ha-m	40.7 → 37.6	42.3 → 3.2	44.4 → 35.4	58.4 → 45.4	48.3 → 46.8	40	3.5	40.7	50.1	48.1
	ha-mr	38.2 → 1.1	53.1 → −0.1	46.2 → 0.6	52.2 → 9.8	44.5 → 37.6	39.4	2.6	2	41	25.6
	ha-me	64.7 → 37.3	63.5 → 4.2	62.4 → −3.3	86.9 → 46.2	94.7 → 86.2	55.3	1.5	7.7	38.1	51.3
	ho-m	54.5 → 37.1	28 → 4.1	58 → 43	86 → 62.5	67.5 → 63.2	49	9.2	58	41.5	49
	ho-mr	33.1 → 9.3	67.5 → 1	48.6 → 9.6	100.2 → 13.7	97.4 → 13.1	23.8	2.3	2.6	23	12
	ho-me	110.9 → 58	23.7 → 1.6	98.7 → 59.7	99.3 → 69	107.4 → 81.1	96	6.1	73.4	52	64.2
	wa-m	53.1 → 32.8	17.8 → 7	79.2 → 42.9	86.4 → 65.4	80.9 → 78.6	44.9	7.8	73.2	38.8	62
	wa-mr	15 → 6.9	39 → 5.1	26.7 → 4.6	91.6 → 18.6	82.2 → 12.3	9.8	9.3	1.4	10.7	2.1
	wa-me	57.5 → 32.5	44.6 → 5.3	111 → 49.5	112 → 84	111.2 → 111.7	40.6	15.2	109.9	74.3	66.5
Average ♠		-48.3%	-88.7%	-61.5%	-47.1%	**-25.84%**					
Average ♣							-50.1%	-92.5%	-59.4%	-50.9%	-48.4%

among baselines that are trained with 100% D4RL. In each line, we bold the best score among baselines that are trained with 10% D4RL data, i.e., including the single-domain 10% D4RL setting and the cross-domain setting.

We observe that almost all offline RL methods experience a significant performance drop when the training data size is reduced from 100% D4RL to 10% D4RL. Additionally, incorporating additional source-domain data (i.e., simply merging cross-domain data) may lead to poor performance compared to using only target-domain data (10% D4RL). We argue that the primary reason is that the source offline data cannot guarantee that the same transition (state-action-next-state) can be achieved in the target environment.

5.3. Adaptation Performance in Cross-Domain RL

Question 2. *Can IGDF improve offline data efficiency and achieve better performance than prior methods?*

As shown in Table 1, across all offline RL approaches, we observe that IQL exhibits the least performance decline in the single-domain setting, with an average decrease of only 25%, and it exhibits the highest data efficiency. For the sake of fairness, we select IQL as the common backbone for IGDF and other baselines.

To systematically investigate the adaptation performance of IGDF, we design various dynamics shift scenarios, including kinematic shifts and morphology shifts. In our main experiment, we change the body mass of agents or introduce joint noise to the motion as our source domain environment. The empirical results are presented in Tables 2 and 4. We observe that IGDF achieves the highest summation of scores over 18 tasks compared to the baselines when utilizing 10% D4RL data. When compared to the best performance of the baselines with 100% D4RL data, IGDF exhibits the smallest performance degradation (-11.89% and -10.81%) among the baselines with 10% D4RL data. As shown in Eq. (5), benefiting from a substantial number of negative samples, IGDF can obtain a precise estimation of *MI gap* and consequently make significant progress in discerning whether the sampled source-domain data helps the training over the target-domain data. However, other baseline approaches suffer from limited target data, which exacerbates the under-fitting issue of domain discriminators. Therefore, IGDF outperforms other baselines and obtains SOTA results on 11 out of 18 tasks.

Question 3. *Can IGDF sustain stable performance when confronted with a larger dynamics gap?*

To assess the performance of IGDF under substantial dynamics shift, we conduct tests on broken thighs and morphology tasks, following the settings outlined in Xu et al. (2023). As shown in Table 3, DARA exhibits poor performance in this scenario when confronted with a larger dynamics gap. We attribute the failure of DARA to the potential unbounded issue of its estimation towards the dynamics gap based on likelihood probability. Similarly, SRPO can hardly show remarkable improvement on both two tasks compared to the results of training IQL with the mixed dataset. As the estimated dynamics gap in our IGDF (i.e., *MI gap*) is bounded by Eq.(11), it endows IGDF a consistently rational attitude to judge and utilize the whole source-domain dataset when

Table 2. Results on body mass shift and joint noise shift tasks across five seeds, where w/o Aug means training IQL with mixed datasets.

mass	w/o Aug	DARA	SRPO	**Ours**	joint	w/o Aug	DARA	SRPO	**Ours**
ha-m	36.69 ± 0.24	39.37 ± 0.11	47.02 ± 0.09	**47.21 ± 0.19**	ha-m	48.06 ± 0.09	**52.59 ± 1.48**	51.98 ± 0.31	50.40 ± 0.36
ha-mr	15.68 ± 4.28	35.90 ± 0.90	36.95 ± 0.84	**38.76 ± 0.88**	ha-mr	25.62 ± 1.09	**47.64 ± 0.44**	38.48 ± 1.071	39.11 ± 0.55
ha-me	36.78 ± 0.38	51.85 ± 2.02	**90.71 ± 1.05**	89.53 ± 2.72	ha-me	51.26 ± 2.23	83.40 ± 0.56	82.96 ± 1.48	**90.93 ± 3.21**
ho-m	21.12 ± 10.55	55.92 ± 4.91	55.89 ± 5.85	**63.78 ± 8.43**	ho-m	49.03 ± 6.67	**55.64 ± 3.71**	53.50 ± 6.61	54.04 ± 7.89
ho-mr	10.69 ± 1.68	27.71 ± 4.62	17.92 ± 5.86	**27.84 ± 9.36**	ho-mr	12.01 ± 3.08	53.04 ± 16.10	37.68 ± 15.34	**63.07 ± 27.96**
ho-me	46.95 ± 6.27	71.43 ± 7.70	96.03 ± 11.89	**96.88 ± 18.25**	ho-me	64.19 ± 4.76	67.98 ± 31.39	72.59 ± 34.23	**103.97 ± 7.68**
wa-m	81.05 ± 0.70	**86.77 ± 0.63**	83.49 ± 1.13	83.76 ± 0.14	wa-m	62.03 ± 3.29	73.6 ± 6.41	77.32 ± 4.326	**78.76 ± 2.74**
wa-mr	17.99 ± 2.51	**83.89 ± 0.61**	77.10 ± 2.73	79.19 ± 1.31	wa-mr	2.10 ± 0.54	56.52 ± 6.64	45.19 ± 8.68	**58.38 ± 10.55**
wa-me	84.28 ± 2.77	93.74 ± 0.25	108.88 ± 5.83	**112.10 ± 0.78**	wa-me	66.46 ± 7.75	119.25 ± 1.13	**120.33 ± 3.88**	116.19±5.76
Sum	351.21	546.56	613.98	**639.05**	**Sum**	380.75	609.66	580.03	**654.85**
Average	-51.79%	-23.51%	-15.37%	**-11.89%**	**Average**	-45.01%	-14.31%	-19.33%	**-10.81%**

Table 3. Results in broken thighs and morphology tasks across five seeds, where w/o Aug means training IQL with mixed datasets.

broken	w/o Aug	DARA	SRPO	Ours	morph	w/o Aug	DARA	SRPO	Ours
ha-m	45.70 ± 0.10	43.77 ± 0.31	45.58 ± 0.59	**46.48 ± 0.04**	ha-m	44.38 ± 0.17	37.49 ± 1.22	43.06 ± 1.44	**45.80 ± 0.43**
ha-mr	36.08 ± 1.23	32.99 ± 3.93	37.48 ± 1.33	**38.62 ± 0.68**	ha-mr	31.96 ± 5.07	1.73 ± 1.43	21.05 ± 11.92	**37.56 ± 1.60**
ha-me	81.67 ± 6.71	68.79 ± 12.00	85.06 ± 5.23	**89.76 ± 2.36**	ha-me	60.24 ± 3.58	35.57 ± 2.54	53.73 ± 8.57	**80.19 ± 8.23**
ho-m	53.63 ± 7.96	55.28 ± 6.87	58.83 ± 8.07	**63.15 ± 7.64**	ho-m	55.26 ± 4.48	54.12 ± 6.73	51.31 ± 7.20	**59.72 ± 3.64**
ho-mr	12.62 ± 1.29	12.06 ± 0.77	12.02 ± 0.61	**13.81 ± 2.30**	ho-mr	16.83 ± 4.25	13.86 ± 3.07	14.46 ± 2.18	**16.92 ± 0.62**
ho-me	60.25 ± 37.75	56.44 ± 31.04	50.71 ± 28.51	**95.95 ± 13.33**	ho-me	92.13 ± 9.74	63.55 ± 42.19	102.97 ± 6.60	**106.80 ± 5.00**
wa-m	63.47 ± 23.82	22.71 ± 19.57	78.95 ± 3.10	**79.19 ± 5.05**	wa-m	34.29 ± 10.77	11.29 ± 10.09	40.64 ± 13.54	**59.39 ± 24.87**
wa-mr	9.03 ± 6.02	14.20 ± 3.90	16.27 ± 3.63	**17.00 ± 3.52**	wa-mr	8.96 ± 6.68	1.35 ± 3.37	8.15 ± 5.53	**11.46 ± 4.58**
wa-me	96.33 ± 9.77	28.90 ± 12.79	**103.71 ± 5.20**	94.35 ± 10.59	wa-me	87.60 ± 10.02	13.64 ± 11.13	95.46 ± 26.70	**107.87 ± 3.60**
Sum	458.78	335.14	488.61	**538.31**	**Sum**	431.65	232.6	430.83	**525.71**
Average	-34.83%	-49.38%	-30.47%	**-24.63%**	**Average**	-39.52%	-66.62%	-41.57%	**-27.38%**

Table 4. Comparison on body mass shift and joint noise shift tasks, where DARA* denotes the best score among the offline RL baselines (BCQ, CQL, IQL, and MOPO) when using dynamics-aware reward modification (DARA).

mass	DARA*	BOSA	**Ours**	joint	DARA*	BOSA	**Ours**
ha-m	39.4	58.3 ± 2.5	47.21	ha-m	52.6	56.2 ± 0.27	50.40
ha-mr	35.9	37.2 ± 0.7	38.76	ha-mr	47.6	51.3 ± 1.1	39.11
ha-me	56.1	51.6 ± 0.1	89.53	ha-me	83.4	52.8 ± 0.45	90.93
ho-m	59.3	82.4 ± 2.1	63.78	ho-m	58.0	78 ± 7.3	54.04
ho-mr	34.1	39.7 ± 0.1	27.84	ho-mr	53.0	32.7 ± 1.3	63.07
ho-me	99.7	104.2 ± 0.5	96.88	ho-me	109.0	96.4 ± 0.5	103.97
wa-m	86.8	83 ± 2.9	83.76	wa-m	81.2	86.5±5.6	78.76
wa-mr	83.9	21.4 ± 2	79.19	wa-mr	56.5	38.2 ± 4.7	58.38
wa-me	93.3	86.5 ± 0.6	112.10	wa-me	119.3	85.8 ± 0.3	116.19
Sum	588.5	564.3	**639.05**	**Sum**	**660.6**	577.9	654.85

encountered with a larger dynamics gap. As a result, IGDF can deliver a more robust performance and even achieve the SOTA results on 17 out of 18 tasks.

5.4. Ablation Studies

Data Ratio Γ**.** We employ various ratios of target domain data ($\Gamma = 5\%, 10\%, 30\%, 50\%$) to investigate the algorithm's sensitivity to the amount of target domain data. The results shown in Figure 4 demonstrate that increasing the amount of target data generally improves the performance of both IQL and IGDF, which means the data amount is a critical factor for offline performance. Consistently, IGDF achieves better performance than IQL and SRPO across a wide range of target data, showing that IGDF can fully exploit the reusable source domain transitions to enhance the training efficiency concerning the target domain. It is worth noting that with the increase in the amount of target data, the previous source data selection ratio may not be suitable for the current situation. So we can adjust the proportion of the data selection ratio appropriately to avoid significant sampling errors.

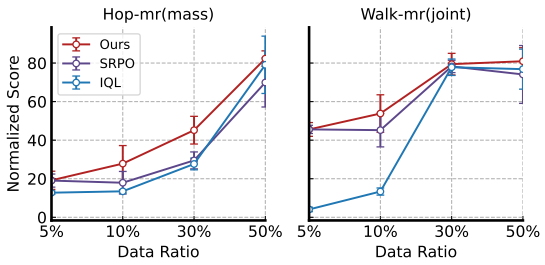

Figure 4. Sensitivity on the amount of target-domain data.

Importance Coefficient α We employ various importance coefficients ($\alpha = 0, 0.5, 1, 2$) to control the content of the importance weighting. Specifically, when alpha is equal to 0, we assign all filtered out source-domain data with the equal importance. The results shown in Figure 5 demonstrate that incorporating importance weighting at a measured level can enhance the performance and smoothness of the algorithm. Nevertheless, striking a suitable balance presents

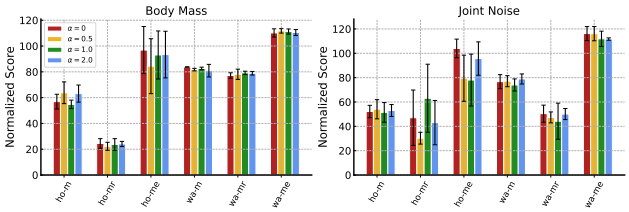

Figure 5. Sensitivity on the importance coefficient.

a challenge. A too-small importance coefficient may lead to performance degradation, while an excessively large coefficient may result in unstable training.

6. Conclusion

In this paper, we focus on the problem of leveraging source domain data with dynamics shifts for efficient RL training. Traditional methods often face performance degradation due to dynamics mismatch when merging cross-domain data. To address this issue, we propose a novel representation-based approach that measures the domain gap through a contrastive objective. The contrastive objective effectively captures the mutual-information gap of transition functions, providing a robust characterization of domain discrepancy without succumbing to unbounded issues. Practically, we present IGDF and the variant, serving as an add-on module for arbitrary offline RL algorithms. Empirical studies showcase the efficacy of our method outperforming previous methods, particularly in scenarios with significant dynamics gaps. One limitation of our approach is its exclusive emphasis on trajectories with similar transition probabilities while ignoring the quality of trajectories. Incorporating trajectory quality would add an interesting dimension for future exploration.

Acknowledgements

This work is supported by the National Science Foundation for Distinguished Young Scholarship of China (No. 62025602) and the National Natural Science Foundation of China (No. 62306242).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Achiam, J., Knight, E., and Abbeel, P. Towards characterizing divergence in deep q-learning. *CoRR*, abs/1903.08894, 2019.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Alberti, E., Tavera, A., Masone, C., and Caputo, B. Idda: A large-scale multi-domain dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5526–5533, 2020.

Bai, C., Wang, L., Yang, Z., Deng, Z.-H., Garg, A., Liu, P., and Wang, Z. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*, 2022.

Bai, C., Wang, L., Hao, J., Yang, Z., Zhao, B., Wang, Z., and Li, X. Pessimistic value iteration for multi-task data sharing in offline reinforcement learning. *Artificial Intelligence*, 326:104048, 2024.

Bose, A., Du, S. S., and Fazel, M. Offline multi-task transfer rl with representational penalization. *CoRR*, abs/2402.12570, 2024.

Chae, J., Han, S., Jung, W., Cho, M., Choi, S., and Sung, Y. Robust imitation learning against variations in environment dynamics. In *International Conference on Machine Learning*, pp. 2828–2852. PMLR, 2022.

Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., and Fox, D. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8973–8979. IEEE, 2019.

Clavera, I., Nagabandi, A., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019.

Deng, Z., Fu, Z., Wang, L., Yang, Z., Bai, C., Zhou, T., Wang, Z., and Jiang, J. False correlation reduction for offline reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Eysenbach, B., Asawa, S., Chaudhari, S., Salakhutdinov, R., and Levine, S. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.

Eysenbach, B., Zhang, T., Levine, S., and Salakhutdinov, R. R. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Eysenbach, B., Myers, V., Salakhutdinov, R., and Levine, S. Inference via interpolation: Contrastive representations provably enable planning and inference. *CoRR*, abs/2403.04082, 2024.

Fatemi, M., Wu, M., Petch, J., Nelson, W., Connolly, S. J., Benz, A., Carnicelli, A., and Ghassemi, M. Semi-markov offline reinforcement learning for healthcare. In *Conference on Health, Inference, and Learning*, pp. 119–137. PMLR, 2022.

Feng, Y., Hansen, N., Xiong, Z., Rajagopalan, C., and Wang, X. Finetuning offline world models in the real world. In *7th Annual Conference on Robot Learning*, 2023.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pp. 2960–2967, 2013.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Guo, Q., Chen, J., Wang, D., Yang, Y., Deng, X., Huang, J., Carin, L., Li, F., and Tao, C. Tight mutual information estimation with contrastive fenchel-legendre optimization. *Advances in Neural Information Processing Systems*, 35: 28319–28334, 2022.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Ishfaq, H., Nguyen-Tang, T., Feng, S., Arora, R., Wang, M., Yin, M., and Precup, D. Offline multitask representation learning for reinforcement learning. *CoRR*, abs/2403.11574, 2024.

James, S., Davison, A. J., and Johns, E. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In *Conference on Robot Learning*, pp. 334–343. PMLR, 2017.

Kim, K., Gu, Y., Song, J., Zhao, S., and Ermon, S. Domain adaptive imitation learning. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2020.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Kuo, N. I., Polizzotto, M. N., Finfer, S., Jorm, L., and Barbieri, S. Synthetic acute hypotension and sepsis datasets based on MIMIC-III and published as part of the health gym project. *CoRR*, abs/2112.03914, 2021.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., and Srinivas, A. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33: 19884–19895, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

Liu, J., Hongyin, Z., and Wang, D. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. In *International Conference on Learning Representations*, 2022.

Liu, J., Zhang, Z., Wei, Z., Zhuang, Z., Kang, Y., Gai, S., and Wang, D. Beyond OOD state actions: Supported cross-domain offline reinforcement learning. *the AAAI Conference on Artificial Intelligence*, 2024.

Luo, Y., Xu, H., Li, Y., Tian, Y., Darrell, T., and Ma, T. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*, 2018.

Niu, H., Qiu, Y., Li, M., Zhou, G., HU, J., Zhan, X., et al. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 36599–36612, 2022.

Niu, H., Ji, T., Liu, B., Zhao, H., Zhu, X., Zheng, J., Huang, P., Zhou, G., Hu, J., and Zhan, X. H2O+: an improved framework for hybrid offline-and-online RL with dynamics gaps. *CoRR*, abs/2309.12716, 2023.

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., and Zhang, L. Solving rubik's cube with a robot hand. *CoRR*, abs/1910.07113, 2019.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019.

Ren, T., Xiao, C., Zhang, T., Li, N., Wang, Z., Sanghavi, S., Schuurmans, D., and Dai, B. Latent variable representation for reinforcement learning. *CoRR*, abs/2212.08765, 2022a.

Ren, T., Zhang, T., Lee, L., Gonzalez, J. E., Schuurmans, D., and Dai, B. Spectral decomposition representation for reinforcement learning. *CoRR*, abs/2208.09515, 2022b.

Ren, T., Xiao, C., Zhang, T., Li, N., Wang, Z., sujay sanghavi, Schuurmans, D., and Dai, B. Latent variable representation for reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023a.

Ren, T., Zhang, T., Lee, L., Gonzalez, J. E., Schuurmans, D., and Dai, B. Spectral decomposition representation for reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Sadeghi, F. and Levine, S. Cad2rl: Real single-image flight without a single real image. In *Robotics: Science and Systems*, 2017.

Shi, J., Bai, C., He, H., Han, L., Wang, D., Zhao, B., Li, X., and Li, X. Robust quadrupedal locomotion via risk-averse policy learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *CoRR*, abs/2110.04652, 2021.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M., et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Wen, X., Yu, X., Yang, R., Bai, C., and Wang, Z. Towards robust offline-to-online reinforcement learning via uncertainty and smoothness. *CoRR*, abs/2309.16973, 2023.

Werbos, P. J. Neural networks for control and system identification. In *Proceedings of the 28th IEEE Conference on Decision and Control,*, pp. 260–265. IEEE, 1989.

Wittenmark, B. Adaptive dual control methods: An overview. *Adaptive Systems in Control and Signal Processing 1995*, pp. 67–72, 1995.

Wu, J., Wu, H., Qiu, Z., Wang, J., and Long, M. Supported policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31278–31291, 2022.

Xu, K., Bai, C., Ma, X., Wang, D., Zhao, B., Wang, Z., Li, X., and Li, W. Cross-domain policy adaptation via value-guided data filtering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Xue, Z., Cai, Q., Liu, S., Zheng, D., Jiang, P., Gai, K., and An, B. State regularized policy optimization on data with dynamics shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yang, G., Ajay, A., and Agrawal, P. Overcoming the spectral bias of neural value approximation. *CoRR*, abs/2206.04672, 2022a.

Yang, R., Bai, C., Ma, X., Wang, Z., Zhang, C., and Han, L. Rorl: Robust offline reinforcement learning via conservative smoothing. *Advances in neural information processing systems*, 35:23851–23866, 2022b.

Yang, R., Bai, C., Guo, H., Li, S., Zhao, B., Wang, Z., Liu, P., and Li, X. Behavior contrastive learning for unsupervised skill discovery. In *International Conference on Machine Learning*, pp. 39183–39204. PMLR, 2023.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. Conservative data sharing for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:11501–11516, 2021.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Finn, C., and Levine, S. How to leverage unlabeled data in offline reinforcement learning. In *International Conference on Machine Learning*, pp. 25611–25635. PMLR, 2022.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhang, J., Lyu, J., Ma, X., Yan, J., Yang, J., Wan, L., and Li, X. Uncertainty-driven trajectory truncation for data augmentation in offline reinforcement learning. In *ECAI 2023*, pp. 3018–3025. IOS Press, 2023a.

Zhang, L., Xiong, Y., Yang, Z., Casas, S., Hu, R., and Urtasun, R. Learning unsupervised world models for autonomous driving via discrete diffusion. *CoRR*, abs/2311.01017, 2023b.

Zhu, S., Kimmel, A., Bekris, K. E., and Boularias, A. Fast model identification via physics engines for data-efficient policy search. In *IJCAI*, pp. 3249–3256, 2018.

A. Theoretical Proof

A.1. Proof of Theorem 3.1

Theorem A.1. *The MI gap $\Delta I = I_{\text{tar}}([S, A]; S') - I_{\text{src}}([S, A]; S')$ can be lower bounded by the negative contrastive objective, as*

$$\Delta I \geq \log(K-1) - \mathcal{L}_{\text{NCE}} \coloneqq I_{\text{NCE}}, \tag{13}$$

where $K-1$ is the number of negative samples from the source domain.

Proof. For the standard Info-NCE (van den Oord et al., 2018), we adopt the score function $h = \exp(\phi(s,a)^\top \psi(s'))$ to approximate the information density ratio of $p(s'|s,a)$ and $p(s')$, which preserves the MI between (s,a) and s'. In Eq. (4), we use h_1 and h_2 to represent the information density ratio in source and target domains, respectively. Then the contrastive objective becomes

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E}_{p(s,a,s'_B)}\mathbb{E}_{S'-} \log \left[\frac{\frac{\hat{P}_{\text{tar}}(s'_B|s,a)}{\hat{\rho}_{\text{tar}}(s'_B)}}{\frac{\hat{P}_{\text{tar}}(s'_B|s,a)}{\hat{\rho}_{\text{tar}}(s'_B)} + \sum_{s'_A \in S'-} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)}} \right] \tag{14}$$

$$= \mathbb{E}_{p(s,a,s'_B)}\mathbb{E}_{S'-} \log \left[1 + \frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} \sum_{s'_A \in S'-} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \tag{15}$$

$$= \mathbb{E}_{p(s,a,s'_B)}\mathbb{E}_{S'-} \log \left[1 + (K-1)\frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} \frac{1}{K-1} \sum_{s'_A \in S'-} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \tag{16}$$

$$\geq \mathbb{E}_{p(s,a,s'_B)} \log \left[(K-1)\frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} \frac{1}{K-1} \sum_{s'_A \in S'-} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \tag{17}$$

$$= \mathbb{E}_{p(s,a,s'_B)} \log \left[\frac{1}{K-1} \sum_{s'_A \in S'-} (K-1)\frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right]. \tag{18}$$

Considering $\log$ is a convex function, we can derive the following from Eq. (18) according to Jensen's inequality, as

$$\mathcal{L}_{\text{NCE}} \geq \mathbb{E}_{p(s,a,s'_B)} \left[\frac{1}{K-1} \sum_{s'_A \in S'-} \log \left[(K-1)\frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \right]$$

$$= \mathbb{E}_{p(s,a,s'_B)} \left[\frac{1}{K-1} \sum_{s'_A \in S'-} \left[\log(K-1) + \log \frac{\hat{\rho}_{\text{tar}}(s'_B)}{p(s'_B|s,a)} + \log \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \right] \tag{19}$$

$$\approx \mathbb{E}_{p(s,a,s'_B)} \left[\log(K-1) + \log \frac{\hat{\rho}_{\text{tar}}(s'_B)}{\hat{P}_{\text{tar}}(s'_B|s,a)} + \mathbb{E}_{s'_A \in S'-} \log \frac{\hat{P}_{\text{src}}(s'_A|s,a)}{\hat{\rho}_{\text{src}}(s'_A)} \right] \tag{20}$$

$$= -I_{\text{tar}} + I_{\text{src}} + \log(K-1) \tag{21}$$

$$= -\Delta I + \log(K-1). \tag{22}$$

Then we have

$$\Delta I \geq \log(K-1) - \mathcal{L}_{\text{NCE}}. \tag{23}$$

□

A.2. Proof of Theorem 3.2

Theorem A.2. *For shared data from the source domain $\widehat{M}_{\text{src}}$, i.e., $(s,a,s') \in \mathcal{D}_{\text{src}}$, the relationship between the MI gap and dynamics gap is*

$$\Delta I = D_{\text{KL}}[\hat{\rho}_{\text{src}}(s') \| \hat{\rho}_{\text{tar}}(s')] - D_{\text{KL}}[\hat{P}_{\text{src}}(s'|s,a) \| \hat{P}_{\text{tar}}(s'|s,a)]. \tag{24}$$

In contrast, for data from the target domain $\widehat{M}_{\rm tar}$, the relationship between the MI gap and dynamics gap is

$$\Delta I = D_{\rm KL}[\hat{P}_{\rm tar}(s'|s,a)\|\hat{P}_{\rm src}(s'|s,a)] - D_{\rm KL}[\hat{\rho}_{\rm tar}(s')\|\hat{\rho}_{\rm src}(s')]. \tag{25}$$

Then, the MI gap is bounded by

$$-H(\hat{\rho}_{\rm src}(s')) \le \Delta I \le H(\hat{\rho}_{\rm tar}(s')). \tag{26}$$

Proof. For data shared from the source domain, we derive the MI gap as

$$\begin{aligned}\Delta I &= \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm src}}\left[\log\frac{\hat{P}_{\rm tar}(s'|s,a)}{\hat{\rho}_{\rm tar}(s')}\right] - \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm src}}\left[\log\frac{\hat{P}_{\rm src}(s'|s,a)}{\hat{\rho}_{\rm src}(s')}\right] \\ &= \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm src}}\left[-\log\frac{\hat{P}_{\rm src}(s'|s,a)}{\hat{P}_{\rm tar}(s'|s,a)} + \log\frac{\hat{\rho}_{\rm src}(s')}{\hat{\rho}_{\rm tar}(s')}\right] \\ &= -D_{\rm KL}\left[\hat{P}_{\rm src}(s'|s,a)\middle\|\hat{P}_{\rm tar}(s'|s,a)\right] + D_{\rm KL}\left[\hat{\rho}_{\rm src}(s')\middle\|\hat{\rho}_{\rm tar}(s')\right].\end{aligned} \tag{27}$$

Meanwhile, since data comes from the source domain, we have $\Delta I = I_{\rm tar} - I_{\rm src} \le 0$ since the information density of the source domain $i(s,a,s') = \hat{P}_{\rm src}(s'|s,a)/\hat{\rho}_{\rm src}$ is larger than that of the target domain, as we illustrated in Figure 2. Then we have

$$\Delta I \ge -I_{\rm src} = -H(S') + H(S'|S,A) \ge -H(\hat{\rho}_{\rm src}(s')). \tag{28}$$

For the data comes from the target domain, we derive the MI gap as

$$\begin{aligned}\Delta I &= \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm tar}}\left[\log\frac{\hat{P}_{\rm tar}(s'|s,a)}{\hat{\rho}_{\rm tar}(s')}\right] - \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm tar}}\left[\log\frac{\hat{P}_{\rm src}(s'|s,a)}{\hat{\rho}_{\rm src}(s')}\right] \\ &= \mathbb{E}_{s,a,s'\sim\mathcal{D}_{\rm tar}}\left[\log\frac{\hat{P}_{\rm tar}(s'|s,a)}{\hat{P}_{\rm src}(s'|s,a)} - \log\frac{\hat{\rho}_{\rm tar}(s')}{\hat{\rho}_{\rm src}(s')}\right] \\ &= D_{\rm KL}\left[\hat{P}_{\rm tar}(s'|s,a)\middle\|\hat{P}_{\rm src}(s'|s,a)\right] + D_{\rm KL}\left[\hat{\rho}_{\rm tar}(s')\middle\|\hat{\rho}_{\rm src}(s')\right].\end{aligned} \tag{29}$$

Similarly, since data comes from the target domain, we have $\Delta I = I_{\rm tar} - I_{\rm src} \ge 0$ since the information density of the target domain $i(s,a,s') = \hat{P}_{\rm tar}(s'|s,a)/\hat{\rho}_{\rm tar}$ is larger than that of the source domain. Then we have

$$\Delta I \le I_{\rm tar} = H(S') - H(S'|S,A) \le H(\hat{\rho}_{\rm src}(s')). \tag{30}$$

Combing the bounds in the source domain and target domain, we have

$$-H(\hat{\rho}_{\rm src}(s')) \le \Delta I \le H(\hat{\rho}_{\rm tar}(s')). \tag{31}$$

As a result, the proposed MI gap is bounded by the entropy of state distribution. The MI gap overcomes the drawback of the dynamics gap since the dynamics gap can be unbounded with a large domain gap.

□

A.3. Proof of Theorem 3.3

Theorem A.3. *Under the setting of cross-domain offline RL, the performance difference of any policy π evaluated by the source domain $\widehat{M}_{\rm src}$ and the true target MDP $M_{\rm tar}$ can be bounded as below,*

$$\eta_{M_{\rm tar}}(\pi) - \eta_{\widehat{M}_{\rm src}}(\pi) \ge -\frac{\gamma R_{\max}}{(1-\gamma)^2}\left\{2\mathbb{E}_{\hat{\rho}_{\rm tar}}\left[D_{\rm TV}\left(P_{\rm tar}\|\hat{P}_{\rm tar}\right)\right] + \sqrt{2D_{\rm KL}\left(\hat{\rho}_{\rm src}(s')\|\hat{\rho}_{\rm tar}(s')\right) + 2|\Delta I|}\right\} \tag{32}$$

Proof. For the performance bound $\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi)$ for any policy π, we can firstly convert the bound to the following form:

$$\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) = \underbrace{\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{tar}}}(\pi)}_{\text{(a)}} + \underbrace{\eta_{\widehat{M}_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi)}_{\text{(b)}}. \tag{33}$$

For term (a) in the RHS, we can obtain the performance bound based on the telescoping lemma (Luo et al., 2018):

$$\begin{aligned}
\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{tar}}}(\pi) &= \frac{\gamma}{1-\gamma}\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[\mathbb{E}_{P_{M_{\text{tar}}}}\left[V^{\pi}_{M_{\text{tar}}}(s')\right] - \mathbb{E}_{P_{\widehat{M}_{\text{tar}}}}\left[V^{\pi}_{M_{\text{tar}}}(s')\right]\right] \\
&= \frac{\gamma}{1-\gamma}\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[\sum_{s'}(P_{\text{tar}}(s'|s,a) - \hat{P}_{\text{tar}}(s'|s,a))V^{\pi}_{M_{\text{tar}}}(s')\right] \qquad (34) \\
&\geq -\frac{\gamma}{1-\gamma}\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[\sum_{s'}\left|P_{\text{tar}}(s'|s,a) - \hat{P}_{\text{tar}}(s'|s,a)\right|\frac{R_{\max}}{1-\gamma}\right] \\
&= -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[D_{\text{TV}}\left(P_{\text{tar}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)\right] \qquad (35)
\end{aligned}$$

Following a similar procedure, we can obtain the performance bound of term (b) in RHS of Eq. 33:

$$\begin{aligned}
\eta_{\widehat{M}_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) &\geq -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\mathbb{E}_{\hat{\rho}_{\text{src}}}\left[D_{\text{TV}}\left(\hat{P}_{\text{tar}}(s'|s,a)\|\hat{P}_{\text{src}}(s'|s,a)\right)\right] \\
&\geq -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\mathbb{E}_{\hat{\rho}_{\text{src}}}\left[\sqrt{\frac{1}{2}D_{\text{KL}}\left(\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)}\right], \qquad (36)
\end{aligned}$$

where the second inequality derives from Pinsker's inequality. Recalling the MI gap ΔI that we aim to bound for the source domain offline dataset D_{src}, we can build the connection between the MI gap and dynamics gap as follows:

$$\begin{aligned}
\Delta I &= I_{\text{tar}}\left[[S,A];S'\right] - I_{\text{src}}\left[[S,A];S'\right] \\
&= \mathbb{E}_{D_{\text{src}}}\left[\log\hat{P}_{\text{tar}}(s'|s,a) - \log\hat{\rho}_{\text{tar}}(s')\right] - \mathbb{E}_{D_{\text{src}}}\left[\log\hat{P}_{\text{src}}(s'|s,a) - \log\hat{\rho}_{\text{src}}(s')\right] \\
&= \mathbb{E}_{(s,a)\sim D_{\text{src}}}\left[\mathbb{E}_{s'\sim\hat{P}_{\text{src}}(s'|s,a)}\left[\log\frac{\hat{P}_{\text{tar}}(s'|s,a)}{\hat{P}_{\text{src}}(s'|s,a)}\right]\right] + \mathbb{E}_{s'\sim D_{\text{src}}}\left[\log\frac{\hat{\rho}_{\text{src}}(s')}{\hat{\rho}_{\text{tar}}(s')}\right] \\
&= -\mathbb{E}_{(s,a)\sim D_{\text{src}}}\left[D_{\text{KL}}\left(\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)\right] + D_{\text{KL}}\left(\hat{\rho}_{\text{src}}\|\hat{\rho}_{\text{tar}}\right).
\end{aligned}$$

Thus, we can formulate the dynamics gap considering the empirical MDPs as:

$$\mathbb{E}_{(s,a)\sim D_{\text{src}}}\left[D_{\text{KL}}\left(\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)\right] = D_{\text{KL}}\left(\hat{\rho}_{\text{src}}\|\hat{\rho}_{\text{tar}}\right) - \Delta I$$

Since the empirical MDP characterizes the distribution of the offline dataset (i.e., $\hat{P}(s'|s,a) = 0, \forall\,(s,a,s') \notin \mathcal{D}$), the state-action distribution conditioned on any policy π equals to that conditioned on the behavior policy π^b (i.e., $\hat{\rho}^{\pi}(s,a) = \hat{\rho}^{\pi^b}(s,a), \forall s,a$). Thus, we can further derive Eq. (36) to:

$$\begin{aligned}
\eta_{\widehat{M}_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) &\geq -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\mathbb{E}_{(s,a)\sim D_{\text{src}}}\left[\sqrt{\frac{1}{2}D_{\text{KL}}\left(\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)}\right] \\
&\geq -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{2}\mathbb{E}_{(s,a)\sim D_{\text{src}}}\left[D_{\text{KL}}\left(\hat{P}_{\text{src}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)\right]} \quad \text{(Jensen's inequality)} \\
&= -\frac{\sqrt{2}\gamma R_{\max}}{(1-\gamma)^2}\sqrt{D_{\text{KL}}\left(\hat{\rho}_{\text{src}}\|\hat{\rho}_{\text{tar}}\right) - \Delta I} \qquad (37)
\end{aligned}$$

Integrating Eq. (35) and Eq. (37), we can obtain the final performance bound:

$$\begin{aligned}
\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) &\geq -\frac{2\gamma R_{\max}}{(1-\gamma)^2}\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[D_{\text{TV}}\left(P_{\text{tar}}(s'|s,a)\|\hat{P}_{\text{tar}}(s'|s,a)\right)\right] - \frac{\sqrt{2}\gamma R_{\max}}{(1-\gamma)^2}\sqrt{D_{\text{KL}}\left(\hat{\rho}_{\text{src}}\|\hat{\rho}_{\text{tar}}\right) - \Delta I} \\
&= -\frac{\gamma R_{\max}}{(1-\gamma)^2}\left\{2\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[D_{\text{TV}}\left(P_{\text{tar}}\|\hat{P}_{\text{tar}}\right)\right] + \sqrt{2D_{\text{KL}}\left(\hat{\rho}_{\text{src}}(s')\|\hat{\rho}_{\text{tar}}(s')\right) - 2\Delta I}\right\} \qquad (38)
\end{aligned}$$

Meanwhile, according to Eq. (27), since $\Delta I \leq 0$ when we share the data from the source domain to the target domain, we can rewrite the performance bound as

$$\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) \geq -\frac{\gamma R_{\max}}{(1-\gamma)^2}\left\{2\mathbb{E}_{\hat{\rho}_{\text{tar}}}\left[D_{\text{TV}}\left(P_{\text{tar}} \| \hat{P}_{\text{tar}}\right)\right] + \sqrt{2D_{\text{KL}}\left(\hat{\rho}_{\text{src}}(s') \| \hat{\rho}_{\text{tar}}(s')\right) + 2|\Delta I|}\right\} \tag{39}$$

□

A.4. Sub-optimality gap of IGDF

According to Theorem 3.3, we have:

$$\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) \geq -\frac{\gamma R_{\max}}{(1-\gamma)^2} 2E_{\hat{\rho}_{\text{tar}}}\left[D_{\text{TV}}\left(P_{\text{tar}} || \hat{P}_{\text{tar}}\right)\right] + \sqrt{2D_{\text{KL}}\left(\hat{\rho}_{\text{src}}(s') || \hat{\rho}_{\text{tar}}(s')\right) + 2|\Delta I|}, \tag{40}$$

where $2D_{\text{TV}}\left(P_{\text{tar}} || \hat{P}_{\text{tar}}\right) = \sum_{s'} |\hat{P}_{\text{tar}}(s'|s,a) - P_{\text{tar}}(s'|s,a)| = ||\hat{P}_{\text{tar}}(s,a) - P_{\text{tar}}(s,a)||_1$.

by Hoeffding's inequality and union bound, the following inequalities hold with probability at least $1-\delta$:

$$\max_{s,a} ||\hat{P}_{\text{tar}}(s,a) - P_{\text{tar}}(s,a)||_1 \leq \max_{s,a} |\mathcal{S}| \cdot ||\hat{P}_{\text{tar}}(s,a) - P_{\text{tar}}(s,a)||_\infty \leq |\mathcal{S}| \sqrt{\frac{1}{2n} \ln \frac{4|\mathcal{S} \times \mathcal{A} \times \mathcal{S}|}{\delta}}, \tag{41}$$

where n is number of samples for each state action pair, $\hat{P}_{\text{tar}}$ as a matrix of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$.

Moreover, we can obtain a tighter analysis by proving an l_1 concentration bound for multinomial distribution directly: $\max_{s,a} ||\hat{P}_{\text{tar}}(s,a) - P_{\text{tar}}(s,a)||_1 \leq \sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{S} \times \mathcal{A}| \cdot 2^{|\mathcal{S}|}}{\delta}}$ (Refer to (Agarwal et al., 2019) for more details).

So we obtain the following conclusion:

$$\eta_{M_{\text{tar}}}(\pi) - \eta_{\widehat{M}_{\text{src}}}(\pi) \geq -\frac{\gamma R_{\max}}{(1-\gamma)^2} \underbrace{E_{\hat{\rho}_{\text{tar}}}\left[\sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{S} \times \mathcal{A}| \cdot 2^{|\mathcal{S}|}}{\delta}}\right]}_{(a)} + \sqrt{\underbrace{2D_{\text{KL}}\left(\hat{\rho}_{\text{src}}(s') || \hat{\rho}_{\text{tar}}(s')\right)}_{(b)} + \underbrace{2|\Delta I|}_{(c)}}. \tag{42}$$

For term (a), sampling more target-domain data allows us to obtain a more accurate estimate of $\hat{P}_{\text{tar}}$, thereby reducing the discrepancy between the true P_{tar} and the estimated $\hat{P}_{\text{tar}}$. For term (b), it is determined by the properties of the source and target domain datasets and cannot be optimized. Our focus lies on term (c), where by effectively selecting samples with smaller dynamics gaps, we can minimize ΔI and tighten the performance bound.

B. Algorithm Description

The pseudocode of IGDF+IQL is presented in Algorithm 2. We utilize IQL (Kostrikov et al., 2022) as our backbone.

C. Detailed Experiment Setting

C.1. Datasets

To generate environments with different transition functions, we design varying dynamics shift tasks based on three Mujoco benchmarks from Gym (HalfCheetah-v2, Hopper-v2, Walker2D-v2). These tasks encompass a range of modifications, such as adjusting the body mass (body mass shift), adding noises to joint (joint noise shift) of the agents, training with broken thighs and integrating morphological differences (refer to Table 6 and Figure 6 for the details). For each benchmark, we categorize these tasks into two variants: **kinematic shift tasks** and **morphology shift tasks**.

As shown in Table 5, in the *HalfCheetah*, *Hopper*, and *Walker2d* dynamics adaptation setting, we set D4RL datasets as our target domain. For the source domain, we change the environment parameters and then collect the source offline datasets in the changed environments. For body mass shift and joint noise shift, we follow the same setting of DARA, wherein 1) "Medium" offline data, generated by a trained policy with the "medium" level of performance in the source environment,

Algorithm 2 Info-Gap Data Filtering algorithm based on IQL

Input: Source domain offline dataset $\mathcal{D}_{\text{src}}$, target domain offline dataset $\mathcal{D}_{\text{tar}}$, mixed offline dataset $\mathcal{D}_{\text{mix}}$
Initialization: Policy network π_η, value network V_β, Q_θ, target Q network $Q_{\hat{\theta}}$, encoder networks $\phi(s,a)$, $\psi(s')$, data selection ratio ξ, batch size B, importance coefficient α

// Contrastive Representation Learning
Maximize the mutual information by training encoder networks $\phi(s,a)$, $\psi(s')$ via Eq. (6)
// TD Learning
for each gradient step **do**
  Sample $b_{\text{src}} := \{(s,a,r,s')\}_{\text{src}}^{\frac{B}{2\xi}}$ from $\mathcal{D}_{\text{src}}$
  Sample $b_{\text{tar}} := \{(s,a,r,s')\}_{\text{tar}}^{\frac{B}{2}}$ from $\mathcal{D}_{\text{tar}}$
  Sample the top-ξ samples from b_{src} ranked by $h(s_{\text{src}}, a_{\text{src}}, s'_{\text{src}}) := \exp(\phi(s_{\text{src}}, a_{\text{src}})^T \psi(s'_{\text{src}}))$ following:

$$\omega(s,a,s') := \mathbb{1}\left(h(s,a,s') > h_{\xi\%}\right)$$

  Optimize the V_β function following loss:

$$L_V(\beta) = \mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{mix}}}\left[L_2^\tau(Q_{\hat{\theta}}(s,a) - V_\beta(s))\right]$$

  Optimize the Q_θ function following loss:

$$\begin{aligned} L_Q(\theta) = &\frac{1}{2}\mathbb{E}_{(s,a,s')\sim\mathcal{D}_{\text{tar}}}\left[(r(s,a) + \gamma V_\beta(s')) - Q_\theta(s,a)\right]^2 \\ &+ \frac{1}{2}\alpha \cdot h(s,a,s')\mathbb{E}_{(s,a,s')\sim\mathcal{D}_{\text{src}}}\left[\omega(s,a,s')\left((r(s,a) + \gamma V_\beta(s')) - Q_\theta(s,a)\right)^2\right] \end{aligned}$$

  Update the target Q function:

$$\hat{\theta} \leftarrow (1-\mu)\hat{\theta} + \mu\theta$$

end for
// Policy Extractions (AWR)
for each gradient step **do**
  Optimize the policy network π_η following loss:

$$L_\pi(\eta) = \mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{mix}}}\left[\exp\lambda(Q_{\hat{\theta}} - V_\beta(s))\log\pi_\eta(a|s)\right]$$

end for

2) "Medium-Replay" offline data, consisting of recording all samples in the replay buffer observed during training until the policy reaches the "medium" level of performance, 3) "Medium-Expert" offline data, mixing equal amounts of expert demonstrations and "medium" data in the source environment. For broken thighs and morphology shift, we alter the XML file of the Mujoco simulator following VGDF (Xu et al., 2023), and then collect 1M replay transitions with SAC (Haarnoja et al., 2018) in every benchmark.

C.2. Kinematic Shift Tasks

Detailed modifications of the environments with kinematic shifts are shown below (for changing body mass and adding joint noise, see Table 6 for the details):

HalfCheetah - broken back thigh: We modify the rotation range of the joint on the thigh of the back leg from $[-0.52, 1.05]$ to $[-0.0052, 0.0105]$.

```
<joint axis="0 1 0" damping="6" name="bthigh" pos="0 0 0" range="-.0052 .0105" stiffness="
    240" type="hinge"/>
```

Hopper - broken joint: We modify the rotation range of the joint on the head from $[-150, 0]$ to $[-0.15, 0]$ and the joint on

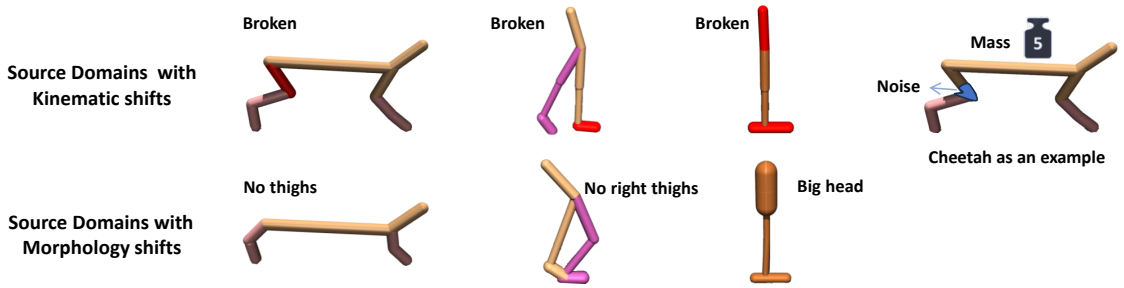

Figure 6. Illustration of all dynamics shift tasks, including kinematic shifts tasks (*Top*) and morphology shifts tasks (*Bottom*). For body mass shift and joint noise shift, we take halfcheetah as an example.

Table 5. Statistics for each task in our cross-domain offline setting.

Environment	Dynamics Shift	Task Name	Target Dataset	Source Dataset
HalfCheetah	Body Mass / Joint Noise	Medium	10^5 (D4RL)	10^6
		Medium-Replay	10100 (D4RL)	10^6
		Medium-Expert	2×10^5 (D4RL)	2×10^6
	Broken / Morphology	Medium	10^5 (D4RL)	10^6 (Replay)
		Medium-Replay	10100 (D4RL)	10^6 (Replay)
		Medium-Expert	2×10^5 (D4RL)	10^6 (Replay)
Hopper	Body Mass / Joint Noise	Medium	10^5 (D4RL)	10^6
		Medium-Replay	20092 (D4RL)	10^6
		Medium-Expert	2×10^5 (D4RL)	2×10^6
	Broken / Morphology	Medium	10^5 (D4RL)	10^6 (Replay)
		Medium-Replay	20092 (D4RL)	10^6 (Replay)
		Medium-Expert	2×10^5 (D4RL)	10^6 (Replay)
Walker2d	Body Mass / Joint Noise	Medium	10^5 (D4RL)	10^6
		Medium-Replay	10093 (D4RL)	10^6
		Medium-Expert	2×10^5 (D4RL)	2×10^6
	Broken / Morphology	Medium	10^5 (D4RL)	10^6 (Replay)
		Medium-Replay	10093 (D4RL)	10^6 (Replay)
		Medium-Expert	2×10^5 (D4RL)	10^6 (Replay)

Table 6. Dynamics shift for Halfcheetah, Hopper, Walker2d tasks. For the body mass shift, we change the mass of the body in the source MDP $\mathcal{M}_{\text{src}}$. For the joint noise shift, we add a noise (randomly sampling in [-0.05, +0.05]) to the actions when we collect the source offline data.

	Halfcheetah		Hopper		Walker	
	Body Mass shift	Joint noise shift	Body Mass shift	Joint noise shift	Body Mass shift	Joint noise shift
Source	mass[4]=0.5	action[-1]+noise	mass[-1]=2.5	action[-1]+noise	mass[-1]=1.47	action[-1]+noise
Target	mass[4]=1.0	action[-1]+0	mass[-1]=5.0	action[-1]+0	mass[-1]=2.94	action[-1]+0

foot from $[-45, 45]$ to $[-18, 18]$.

```
<joint axis="0 -1 0" name="thigh_joint" pos="0 0 1.05" range="-0.15 0" type="hinge"/>
```

```
<joint axis="0 -1 0" name="foot_joint" pos="0 0 0.1" range="-18 18" type="hinge"/>
```

Walker2d - broken right foot: We modify the rotation range of the joint on the foot of the right leg from $[-45, 45]$ to $[-0.45, 0.45]$.

```
<joint axis="0 -1 0" name="foot_joint" pos="0 0 0.1" range="-0.45 0.45" type="hinge"/>
```

C.3. Morphology Shift Tasks

Detailed modifications of the environments with morphology shifts are shown below:

HalfCheetah - no thighs: We modify the size of both thighs. Detailed modifications of the xml file are:

```
<geom fromto="0 0 0 -0.0001 0 -0.0001" name="bthigh" size="0.046" type="capsule"/>
<body name =" bshin" pos=" -0.0001 0 -0.0001 ">
```

```
<geom fromto="0 0 0 0.0001 0 0.0001" name="fthigh" size="0.046" type="capsule"/>
<body name="fshin" pos="0.0001 0 0.0001">
```

Hopper - big head: We modify the size of the head. Detailed modifications of the xml file are:

```
<geom friction="0.9" fromto="0 0 1.45 0 0 1.05" name="torso_geom" size="0.125" type="
    capsule"/>
```

Walker - no right thigh: We modify the size of thigh on the right leg. Detailed modifications of the xml file are:

```
<body name="thigh" pos="0 0 1.05">
  <joint axis="0 -1 0" name="thigh_joint" pos="0 0 1.05" range="-150 0" type="hinge"/>
  <geom friction="0.9" fromto="0 0 1.05 0 0 1.045" name ="thigh_geom" size ="0.05" type="
      capsule"/>
  <body name="leg" pos="0 0 0.35">
    <joint axis="0 -1 0" name="leg_joint" pos="0 0 1.045" range="-150 0" type="hinge"/>
    <geom friction="0.9" fromto="0 0 1.045 0 0 0.3" name="leg_geom" size="0.04" type="
        capsule"/>
    <body name="foot" pos="0.2 0 0">
        <joint axis="0 -1 0" name="foot_joint" pos="0 0 0.3" range="-45 45" type="hinge"/>
        <geom friction="0.9 " fromto="-0.0 0 0.3 0.2 0 0.3" name="foot_geom" size="0.06"
            type="capsule"/>
    </body>
  </body>
</body>
```

D. Implementation Details

D.1. Baselines

We select DARA, SRPO, BOSA as our baselines in cross-domain offline RL tasks and choose some typical offline RL including BCQ, CQL, MOPO, IQL, SPOT as our backbones. We adopt these offline RL of open source code implemented by CORL (github). We run all algorithms with the same five random seeds.

DARA. We follow the default configurations of the public implementation (openreview). A pair of binary classifiers $p(tar \mid s, a, s')$ and $p(tar \mid s, a)$ are learned to infer whether transitions come from the source or target domain. And the domain classifiers are trained by maximizing the cross-entropy losses:

$$J\left(\psi_{SAS}\right) := \mathbb{E}_{(s,a,s')\sim D_{tar}}\left[\log q_{\psi_{SAS}}\left(tar \mid s, a, s'\right)\right] + \mathbb{E}_{(s,a,s')\sim D_{src}}\left[\log\left(1 - q_{\psi_{SAS}}\left(tar \mid s, a, s'\right)\right)\right]$$
$$J\left(\psi_{SA}\right) := \mathbb{E}_{(s,a)\sim D_{\text{tar}}}\left[\log q_{\psi_{SA}}\left(tar \mid s, a\right)\right] + \mathbb{E}_{(s,a)\sim D_{src}}\left[\log\left(1 - q_{\psi_{SA}}\left(tar \mid s, a\right)\right)\right]$$

Applying Bayes' rule, a reward correction $\Delta r\left(s, a\right)$ is augmented to the original reward $r\left(s, a\right)$ of each source domain transition during training, i.e. $\tilde{r}\left(s, a\right) := r\left(s, a\right) + \Delta r\left(s, a\right)$. The reward correction is calculated by:

$$\Delta r\left(s, a\right) := \log \frac{\hat{P}_{\text{tar}}(s' \mid s, a)}{\hat{P}_{\text{src}}(s' \mid s, a)} = \log \frac{q_{\psi_{SAS}}\left(tar \mid s, a, s'\right)}{q_{\psi_{SAS}}\left(src \mid s, a, s'\right)} \frac{q_{\psi_{SA}}(src \mid s, a)}{q_{\psi_{SA}}(tar \mid s, a)}$$

In practical implementation, they also clip the above reward modification between -10 and 10.

SRPO. We implement it based on the pseudocode and default parameters provided in the paper (origin paper). SRPO samples a batch $\mathcal{D}_{\text{batch}}$ from $\mathcal{D}_{\text{off}}$ and $\mathcal{D}_{\text{rollout}}$ and rank them by state-values. Next, SRPO trains a GAN-style discriminator to selectively choose high state-value transitions as real data and low state-value transitions as fake data. For a one-step transition (s_{t+1}, r_t, s_t, a_t) in $\mathcal{D}_{\text{batch}}$, update r_t with $r_t + \lambda \frac{\mathcal{D}\delta(s_t)}{1-\mathcal{D}_\delta(s_t)}$.

BOSA. BOSA employs two support-constrained objectives to address the out-of-distribution issues which can greatly improve offline data efficiency in cross-domain offline RL setting. Although the code is not open-source, BOSA utilizes a

portion of the dataset that aligns with ours. Therefore, we directly compare our results with the scores reported in their paper(origin paper). The support optimization objectives are implemented by:

$$\max_{\pi_\theta} \mathcal{J}_{\mathcal{D}_{\text{mix}}}(\pi_\theta) := \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\text{mix}},\mathbf{a}\sim\pi_\theta(\mathbf{a}|\mathbf{s})}\left[Q_\phi(\mathbf{s},\mathbf{a})\right], \text{ s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\text{mix}}}\left[\log \hat{\pi}_{\beta_{\text{mix}}}(\pi_\theta(\mathbf{s}) \mid \mathbf{s})\right] > \epsilon_{\text{th}}$$

$$\min_{Q_\phi} \mathcal{L}_{\text{mix}}(Q_\phi) := \mathbb{E}_{\substack{(\mathbf{s},\mathbf{a},r,\mathbf{s}')\sim\mathcal{D}_{\text{mix}} \\ \mathbf{a}'\sim\pi_\theta(\mathbf{a}'|\mathbf{s}')}}\left[\delta(Q_\phi)\cdot\mathbb{1}\left(\hat{T}_{\text{target}}(\mathbf{s}' \mid \mathbf{s},\mathbf{a}) > \epsilon'_{\text{th}}\right)\right] + \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\text{source}}}\left[Q_\phi(\mathbf{s},\mathbf{a})\right]$$

D.2. Hyperparameters

The hyperparameters of our backbone offline RL remain unchanged and are fixed in all tasks following the original paper. We list the basic hyperparameters of our algorithm and baselines in Table 7.

Table 7. Hyper-parameters used for IQL, IGDF, DARA, and SRPO.

IQL hyper-parameter	**Value**
Hidden layers (Value and Policy)	2(ReLU)
Hidden units	256(MLP)
Optimizer	Adam
Batch size	256
Replay buffer capacity	2e6
Discount factor γ	0.99
Target network update rate	0.005
Inverse temperature β	3.0
Coefficient for asymmetric loss τ	0.7
V function learning rate	3e-4
Critic learning rate	3e-4
Actor learning rate	3e-4
IGDF hyper-parameter	**Value**
Representation dimension d	16 or 64
Contrastive encoder arch. $\phi(s,a)$	$\dim(S)+\dim(A)\to 256\to 256\to d$(MLP)
Contrastive encoder arch. $\psi(s)$	$\dim(S)\to 256\to 256\to d$(MLP)
Optimizer	Adam
Info learning rate	3e-4
Info batch size	128
Update number	7000
importance coefficient α	1.0
data selection ratio ξ	0.25 or 0.75
DARA hyper-parameter	**Value**
Classifier(s,a) arch. $f(s,a)$	$2\dim(S)+\dim(A)\to 256\to 256\to 256\to 2$(MLP with tanh)
Classifier(s,a,s') arch. $f(s,a,s')$	$\dim(S)+\dim(A)\to 256\to 256\to 256\to 2$(MLP with tanh)
Optimizer	RMSprop
Learning rate	3e-4
batch size	256
Update number	5000
Delta coefficient	0.1
SRPO hyper-parameter	**Value**
Hidden layers	2(ReLU)
Hidden units	256(MLP)
Optimizer	Adam
Learning rate	3e-4
Data selection ratio	0.5 or 0.2
Delta coefficient λ	0.1 or 0.3

E. Supplementary Experiments

E.1. Ablation Study

Data Selection Ratio ξ. As the dynamics gap between source and target domains vary in different task environments, the data selection ratio becomes particularly important. We employ different data selection ratio (25%, 50%, 75%, 100%) for our algorithms. Specifically, a ratio of 100% means that we directly learn from the mixed dataset with all source domain samples (*w/o Aug*). The results shown in Figure 7 demonstrate that different tasks have varying degrees of sensitivity to dynamics gap. As expected, when we set the data selection ratio to 100%, the performance of IGDF degrades dramatically. We observe that the *Halfcheetah* and *Hopper* environments are more suitable for smaller sampling ratios ($\xi = 25\%$), while the Walker2d environment is more suitable for a relatively large sampling ratio ($\xi = 75\%$). This underscores the importance of configuring the data selection ratio to achieve more robust performance when facing different dynamics gaps.

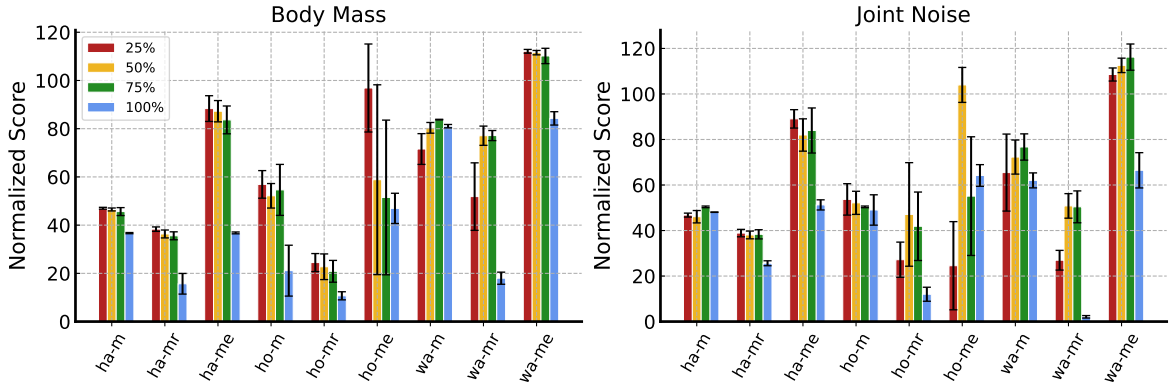

Figure 7. Sensitivity on data selection ratio.

Representation Dimension d Equipping RL algorithms with additional representation learning components has proven effective for task solving. We employ various representation dimensions ($d = 16, 32, 64$) for encoder networks. As illustrated in Figure 8, we observe that the representation dimension does not have a monotonic impact on algorithm performance (a larger representation dimension does not necessarily correlate with better performance in most experiments). Moreover, larger representation dimensions even may lead to information redundancy.

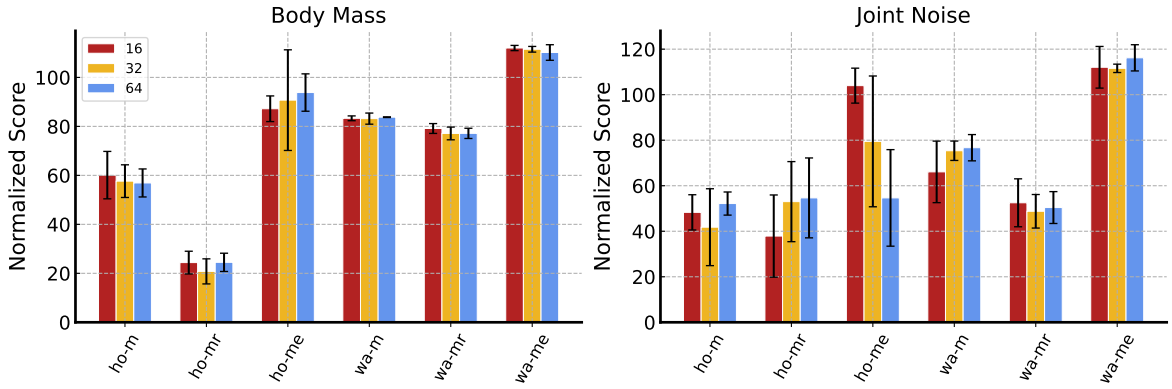

Figure 8. Sensitivity on the representation dimension.

Weight of TD loss h The weight of TD loss $h(s, a, s')$ serves as the measurement of the information density ratio, which has played an important role in further improving policy performance and training stability. It can distinguish the differences between the filtered data in a fine-grained manner by increasing the weight coefficients of samples with a smaller MI-Gap, thereby further improving learning efficiency. To assess the efficacy of using the weight $h(s, a, s')$, we perform an ablation analysis as shown in Table 8 to evaluate the performance of IGDF without this weight.

Table 8. Comparative performance of using weight h and without using it on body mass shift and joint noise shift tasks.

mass	w/o weight h	IGDF	joint	w/o weight h	IGDF
ha-m	47.01±0.38	**47.10±0.38**	ha-m	46.07±2.72	**47.84±0.76**
ha-mr	38.34±0.92	**38.76±0.88**	ha-mr	38.06±1.70	**39.11±0.55**
ha-me	88.34±5.34	**89.53±2.72**	ha-me	81.97±7.10	**90.93±3.21**
ho-m	56.90±5.72	**63.78±8.43**	ho-m	52.17±5.08	**54.04±7.89**
ho-mr	22.74±5.32	**27.84±9.36**	ho-mr	47.07±22.75	**63.07±27.96**
ho-me	**96.88±18.25**	93.82±7.63	ho-me	**103.97±7.68**	95.69±13.7
wa-m	**83.76±0.14**	82.60±1.02	wa-m	76.70±5.77	**78.76±2.74**
wa-mr	77.15±2.09	**79.19±1.31**	wa-mr	50.82±5.39	**58.38±10.55**
wa-me	110.17±3.18	**112.10±0.78**	wa-me	**116.19±5.76**	116.13±5.86
Sum	621.29	**634.72**	**Sum**	613.02	**643.95**
Average	-14.32%	**-12.35%**	**Average**	-16.55%	**-12.26%**

E.2. Additional Experiment Results

Reward modification variant To evaluate the efficacy of the reward modification variant in our algorithm, we compare the performance of IGDF with the reward modification variant. In the reward modification approach, a reward correction term $\Delta r\left(s_t, a_t\right)$ is added to the original reward $r\left(s_t, a_t\right)$ for each source domain transition. This results in the modified reward $\tilde{r}\left(s_t, a_t\right) := r\left(s_t, a_t\right) + \sigma \Delta r\left(s_t, a_t\right)$, where the reward correction is computed as $\phi(s_{\text{src}}, a_{\text{src}})^T \psi(s'_{\text{src}})$. As depicted in Table 9, our observations indicate that the data filtering method exhibits significant advantages.

Table 9. Comparative performance of IGDF and the reward modification variant on body mass shift and joint noise shift tasks.

mass	IGDF	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 2.0$	joint	IGDF	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 2.0$
ha-m	**47.10 ± 0.38**	46.09 ± 1.84	45.95 ± 1.77	46.07 ± 0.67	ha-m	**50.40 ± 0.36**	49.53 ± 1.24	46.83 ± 0.62	46.68 ± 0.18
ha-mr	**38.76 ± 0.88**	37.99 ± 0.68	37.25 ± 1.24	36.44 ± 0.89	ha-mr	39.11 ± 0.55	**39.36 ± 0.34**	38.61 ± 2.00	37.64 ± 1.30
ha-me	**89.53 ± 2.72**	88.83 ± 3.24	86.37 ± 1.84	85.46 ± 6.25	ha-me	**90.93 ± 3.21**	83.86 ± 1.29	84.70 ± 1.201	82.06 ± 5.60
ho-m	**63.78 ± 8.43**	62.02 ± 6.56	59.48 ± 7.92	53.16 ± 4.92	ho-m	**54.04 ± 7.89**	50.63 ± 3.66	51.28 ± 2.41	48.07 ± 4.40
ho-mr	**27.84 ± 9.36**	24.25 ± 2.88	23.60 ± 6.13	20.58 ± 3.22	ho-mr	**63.07 ± 27.96**	44.78 ± 14.78	41.74 ± 12.10	26.61 ± 5.99
ho-me	**96.88 ± 18.25**	53.30 ± 45.00	78.65 ± 30.59	93.31 ± 18.71	ho-me	**103.97 ± 7.68**	89.95 ± 23.13	101.64 ± 10.12	96.00 ± 15.55
wa-m	83.76 ± 0.14	**84.75 ± 0.28**	66.78 ± 8.90	78.33 ± 1.60	wa-m	**78.76 ± 2.74**	78.58 ± 5.13	74.99 ± 13.73	80.79 ± 3.31
wa-mr	**79.19 ± 1.31**	78.20 ± 1.47	78.81 ± 2.58	78.79 ± 1.50	wa-mr	**58.38 ± 10.55**	42.36 ± 9.15	47.80 ± 8.88	51.87 ± 10.01
wa-me	112.10 ± 0.78	110.55 ± 2.01	111.93 ± 0.66	111.81 ± 0.74	wa-me	116.19 ± 5.76	**121.53 ± 0.02**	119.34 ± 2.88	114.37 ± 2.56
Sum	**638.85**	585.98	588.82	603.95	**Sum**	**654.84**	600.58	606.912	584.09

Online learning with limited target-domain data In order to highlight the broader applicability of our work to another related line of research, we conducted additional experiments in offline-to-online settings compared with H2O (Niu et al., 2022). To assess the performance of H2O and IQL in online learning with limited offline data, we perform the online interactions with the source domain for 10^6 steps and use 10^5 target-domain transitions. For the sake of fairness, we select IQL as the backbone for IGDF and H2O. The comparison results are shown in Table 10.

Table 10. Comparative performance of IGDF and H2O on body mass shift and morphology shift tasks.

broken	H2O	IGDF	morph	H2O	IGDF
ha-m	5261 ± 76	5395 ± 32	ha-m	5246 ± 207	5351 ± 169
ha-mr	4505 ± 150	4469 ± 141	ha-mr	4631 ± 53	4512 ± 147
ha-me	8671 ± 840	9359 ± 553	ha-me	8807 ± 1442	9890 ± 874
ho-m	1643 ± 260	1771 ± 339	ho-m	1642 ± 107	1686 ± 240
ho-m	463 ± 56	616 ± 257	ho-mr	417 ± 39	431 ± 34
ho-me	1920 ± 1057	2676 ± 365	ho-m	1456 ± 572	1773 ± 1083
wa-m	3449 ± 237	3330 ± 528	wa-m	3254 ± 309	3226 ± 538
wa-mr	404 ± 219	493 ± 86	wa-mr	722 ± 182	630 ± 146
wa-me	4809 ± 130	4957 ± 99	wa-me	4247 ± 425	4919 ± 147
Sum	31125	33066	**Sum**	30422	32418

F. More discussions

Question 1: The inherent assumption of the behavior policy limits the applicability of the IGDF algorithm.

We recall the relationship between the MI gap and the dynamics gap in Equation (9):

$$\Delta I = \underbrace{D_{\text{KL}}[\hat{\rho}_{\text{src}}(s')||\hat{\rho}_{\text{tar}}(s')]}_{(a)} - \underbrace{D_{\text{KL}}[\hat{P}_{\text{src}}(s'|s,a)||\hat{P}_{\text{tar}}(s'|s,a)]}_{(b)},$$

when we use data shared from the source domain (i.e., $\mathcal{D}_{\text{src}}$) to estimate the MI gap. If the behavior policies of the two datasets are very different, the estimation of $\hat{\rho}_{\text{tar}}$ for $s' \sim \mathcal{D}_{\text{src}}$ can be difficult since the target-domain policy may never encounter similar states when interacting with the target domain, which makes $D_{\text{KL}}(\rho_{\text{tar}}(s')||\hat{\rho}_{\text{tar}}(s'))$ large, and the estimation of term (a) has a large bias. Similary, the estimation of $\hat{P}_{\text{tar}}(s'|s,a)$ for shared data $(s,a,s') \sim \mathcal{D}_{\text{src}}$ also contains large biases, which further increases the bias in estimating ΔI in data sharing.

Nevertheless, in our experiments, we find our method still achieves good results as long as there isn't a significant difference between the two behavior policies. Actually, even in data sharing between the same types of datasets (e.g., medium $\rightarrow$ medium), the behavior policies are not entirely the same since the (medium) policies are trained in environments with dynamics gap. A more apparent evidence is shown in Table 3. In the broken and morphology tasks, we use 10^5 D4RL transitions (medium, medium-replay, medium-expert) as our target-domain data and use 10^6 replay transitions with SAC in every benchmark. IGDF can deliver a more robust performance and even achieve the SOTA results on 17 out of 18 tasks. We believe this assumption holds validity: if the discrepancy between the behavior policies of the two datasets is too large, the source-domain data will become useless in data sharing for the target domain.

Question 2: Why use linear parametrization instead of directly learning $h(s,a,s')$ in an end-to-end manner?

We choose to use linear parameterization instead of directly learning the function $h(s,a,s')$ in an end-to-end manner for several reasons: 1) Intuitively, the score function $\phi(s,a)^\top\psi(s')$ measures whether the representation of state-action pair $\phi(s,a)$ aligns with the next state $\psi(s')$. It is easier to solve a task with linear parameterization given a good representation. In our work, the representation can be separately learned via contrastive learning, which achieves better quantification. 2) As illustrated in Figure 11 of the related research (Eysenbach et al., 2020), solely employing the (s,a,s') classifier to measure domain gaps significantly performs worse than simultaneously utilizing (s,a,s') and (s,a) classifiers. End-to-end learning shares a similar mechanism with solely learning the (s,a,s') classifier. 3) Given what prior work has shown about RL in the presence of function approximation and state aliasing (Achiam et al., 2019; Yang et al., 2022a), it is not surprising that end-to-end learning of representations is fragile (Laskin et al., 2020). RL algorithms require good representations to learn the value function and policy (Eysenbach et al., 2022). 4) A recent work (Eysenbach et al., 2024) also highlighted that representations learned via InfoNCE can effectively capture conditional probabilities between random variables x and y (akin to the conditional probability between (s,a) and s' in our context).

Question 3: The comparison with low-rank MDPs.

Although the low-rank MDP is a theoretical-grounded assumption (i.e., $P(s'|s,a) =< \phi(s,a), \psi(s') >$) that improves the sample complexity (Uehara et al., 2021), it can be hard to extend it to the cross-domain problem. As discussed in recent papers (Ren et al., 2022b;a) that adopt low-rank assumption to learn representations with neural networks in high-dimensional space, the representation is learned by maximizing the likelihood as $\arg\max_{\phi,\psi} \sum \log \phi(s_i,a_i)^\top\psi(s')$. Then the representation $\phi(s,a)$ and $\psi(s')$ will learn to regress the transition probability in this domain. In cross-domain adaptation, if $\phi(s,a)$ and $\psi(s')$ are learned specially adapted to function $P_{\text{src}}(s'|s,a)$ of the source domain, it can be hard to transfer $\phi(s,a)$ and $\psi(s')$ to the target domain since the transition probabilities of two domains are different. In contrast, the contrastive objective in our method is learned by both sampling positive sample and negative samples from both domains, which makes $h(s,a,s') = \exp(\phi(s,a)^\top\psi(s'))$ a score function to captures the domain-distinguishable information as a data filter. In our method, the learned representations $\phi(s,a)$ and $\psi(s')$ are not used for value/policy learning but only for data filtering.

Question 4: The comparison with offline multi-tasks transfer RL.

For the offline multi-task transfer problem studied in (Bose et al., 2024), the source and target tasks are assumed to have similar transition functions to make a core assumption (i.e., Assumption 1) that all tasks share a common representation $\phi_h^\star(s,a)$ holds. However, in offline cross-domain RL considered in our paper, the representations (i.e., $\phi_{\text{src}}^\star$ and $\phi_{\text{tar}}^\star$) can be very different since the transition functions are very different in cross-domain settings with large domain gaps, which makes

the error bound in representation transfer does not hold. Meanwhile, a pointwise linear span assumption (i.e., Assumption 2) is required in (Bose et al., 2024) to make the target transitions a linear combination of the source task dynamics. Similarly, (Ishfaq et al., 2024) also has assumptions about the shared representation $\phi^\star$, and the target task is assumed to be an ξ-approximated linear combination of T source tasks. Nevertheless, such an assumption may not hold when facing a large dynamics gap, as we studied in our paper. Empirically, our method is robust to domain gaps and significantly outperforms other methods on 17 out of 18 tasks with large dynamics gaps (see Table 3). Another difference between our setting and (Bose et al., 2024; Ishfaq et al., 2024) is that we do not adopt shared representation $\phi(s, a)$ for the shared domains, and the representation is only learned to capture the domain-distinguishable information as a data filter. As a result, we believe extending the theoretical results of (Bose et al., 2024; Ishfaq et al., 2024) to cross-domain offline RL requires additional efforts to relax the assumptions to allow source and target domains to have different optimal representations.