RIFT: REWARD IMPROVEMENT FIELD TRANSFER FOR CROSS-DOMAIN OFFLINE RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from limited target data is a central challenge in cross-domain offline reinforcement learning. Although source data can provide additional experience, differences in dynamics can make these transitions unsuitable for the target domain. We propose Reward Improvement Field Transfer (RIFT), a method that instead uses source data to learn how behavior changes from low reward to high reward. RIFT encodes source and target transitions in a shared latent space, pairs low- and high-reward source transitions using optimal transport, and fits an affine reward-improvement field to their displacements. We apply this transformation to target transitions and decode the results to augment the target dataset. Our key assumption is that related domains may differ in their dynamics while still sharing a similar reward-improvement structure. We bound synthesis error under shared improvement geometry, identify conditions favoring target anchoring over source-data transplant, and derive a conditional connection to downstream return. Experiments on locomotion tasks with changes in gravity, morphology, and friction show that RIFT improves aggregate performance over the existing baselines.

1 INTRODUCTION

Learning a good policy often requires a large amount of experience. In practice, however, collecting that experience in the exact environment where the policy will be used can be expensive or difficult. A robot may have millions of transitions from simulation but only a small amount of data from the real world (Tobin et al., 2017; Chebotar et al., 2019). A system may also change after data have already been collected: the robot's body, surface friction, payload, or other environment dynamics may be different. In these cases, we would like to reuse the large amount of existing data rather than collect a new dataset from scratch.

Cross-domain offline reinforcement learning (RL) studies this setting (Liu et al., 2022; Lyu et al., 2025). We have a large dataset from a source domain and a much smaller dataset from a related target domain, and we want to learn a policy for the target without collecting new experience. The difficulty is that source data cannot always be used directly. Even when the two domains share the same task and reward-scoring rule, differences in dynamics mean that the same state and action can produce different next states (Peng et al., 2018). Simply mixing the datasets can therefore teach the policy transitions that are incorrect for the target environment.

At the same time, the source dataset still contains useful information. In a large dataset, we can observe many examples of both poor and successful behavior. These examples can tell us not only what states occur, but also how behavior changes as reward improves. A small target dataset may contain too few high-reward examples to learn this structure reliably.

This leads to the main idea of our work: instead of transferring source transitions directly, can we transfer *how behavior improves*?

We propose **R**eward **I**mprovement **F**ield **T**ransfer (**RIFT**). RIFT first maps source and target transitions into a shared latent space. In the source domain, it matches low-reward transitions with high-reward transitions and measures the change between them. These changes provide examples of how a transition can move toward a better outcome. RIFT learns a simple transformation from these examples and then applies it to low-reward target transitions to create additional target-domain data (see Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

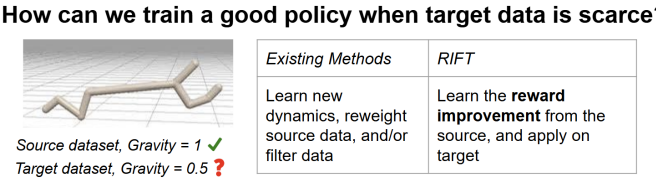

Figure 1: Motivation behind RIFT. A scarce dataset makes it difficult to learn in the target domain. How can we instead use existing structures from the source dataset to transfer what we learn?

The key assumption is that related domains may have different dynamics while still sharing a similar pattern of improvement. For example, the exact motion of an agent may change under different gravity, friction, or body structure, while the changes associated with better task performance may still be related. RIFT is designed to capture this shared structure rather than requiring individual source transitions to match the target domain.

Our contributions are: (1) we introduce RIFT, which transfers the change from low- to high-reward behavior across domains and uses it to generate additional target-domain data; (2) we bound target-anchored synthesis error, give conditions for greater accuracy than source-high transplant, and connect synthesis error to a conditional downstream return guarantee; and (3) we evaluate RIFT on cross-domain locomotion tasks with changes in gravity, morphology, and friction, using approximately one million source transitions and only $5{,}000$ target transitions, and show improved aggregate performance across several source- and target-data settings.

2 Related Work

Cross-Domain reinforcement learning. Reinforcement-learning policies often need to adapt across changes in agent embodiment, observations, action spaces, or environment dynamics. Prior work has studied these shifts through domain randomization, system identification, meta-RL, imitation learning, and domain adaptation (Gamrian & Goldberg, 2019; Hejna et al., 2020; Eysenbach et al., 2021; Fickinger et al., 2022). More recent policy-adaptation methods explicitly address dynamics mismatch through learned representations, data selection, or reward modification (Niu et al., 2022; Xu et al., 2023; Guo et al., 2024). In the fully offline setting, recent methods modify rewards, constrain learning to target-supported regions, filter source data, or correct source transitions (Liu et al., 2022; 2024; Xue et al., 2023; Wen et al., 2024; Lyu et al., 2025; Yan et al., 2026; Van et al., 2025b). We build on this literature so that RIFT uses the source data to learn reward-improvement structure that can be applied to limited target data.

Representations under domain shift. Domain adaptation and generalization often seek representations that retain task-relevant structure across domains. Marginal feature alignment can be insufficient when the relationship between features and outcomes changes (Zhao et al., 2019). We borrow ideas from conditional-invariance methods, which align label- or outcome-conditioned distributions across domains (Li et al., 2018; Tachet des Combes et al., 2020). Recent work has extended this idea beyond classification; for example, COD (Yang et al., 2024) studies conditional invariant representations for regression. More broadly, work on latent representation learning has shown that variational regularization, structured priors, and latent-space objectives can shape the organization and usefulness of learned representations (Higgins et al., 2017; Shi et al., 2025). RIFT uses a shared latent representation organized by reward class. This encourages common low- and high-reward structure across source and target contexts to support transfer of the learned improvement field.

Reward-structured learning and data augmentation. Recent work shows reward and preference information can provide structure beyond a single scalar training target. In RLHF, pairwise preferences are used to train reward models that guide policy optimization (Ouyang et al., 2022; Gao et al., 2023; Rafailov et al., 2023). When preference data are limited, LENS (Tao et al., 2025) synthesizes additional training examples in a learned VAE latent space. Reward structure has also been used to improve offline RL data. MetaDiffuser (Ni et al., 2023) uses return and dynamics

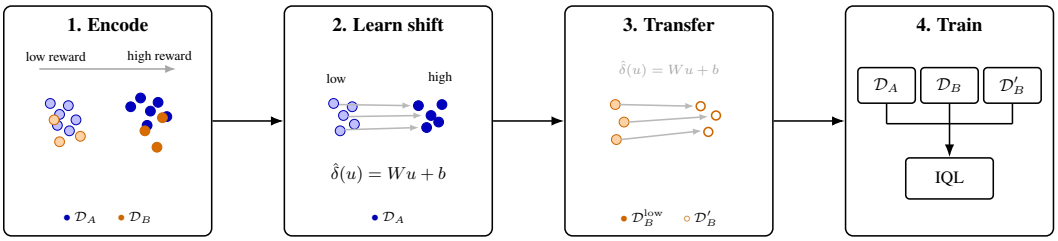

Figure 2: Overview of RIFT. The method encodes source and target data, learns a reward-improvement field from source transitions, transfers it to target anchors, and augments downstream IQL training.

guidance for trajectory generation, DiffStitch (Li et al., 2024) connects low- and high-reward trajectory segments, GTA (Lee et al., 2024) generates higher-return trajectories, and GODA (Huang et al., 2025) uses return-oriented goal conditioning for data augmentation. RIFT combines these ideas in the cross-domain setting. It learns the transformation between low- and high-reward source distributions and uses this transformation to augment scarce target data.

3 Preliminaries

A reinforcement learning problem is formulated as a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, P is the transition kernel, r is the expected one-step reward, $\gamma \in [0, 1)$ is the discount factor, and ρ is the initial-state law. The objective is to learn a policy π that maximizes the expected discounted return

$$J_{\mathcal{M}}(\pi) = \mathbb{E}_{S_0 \sim \rho, \pi}\left[\sum_{t=0}^{T-1} \gamma^t R_t\right],$$

where $T = \infty$ for continuing tasks; in episodic tasks the final transition's reward is included.

We consider two contexts: $\mathcal{M}_A = (\mathcal{S}, \mathcal{A}, P_A, r_A, \gamma, \rho)$ and $\mathcal{M}_B = (\mathcal{S}, \mathcal{A}, P_B, r_B, \gamma, \rho)$ with known termination functions. The contexts share the same state space, action space, and task objective, but may differ in dynamics and induced state–action rewards. A common rule for scoring realized transitions does not need to give the same reward at a fixed (s, a) when the dynamics change. We are given fixed offline datasets $\mathcal{D}_A$ and $\mathcal{D}_B$, where $|\mathcal{D}_B| \ll |\mathcal{D}_A|$. Our objective is to improve policy performance in the under-sampled context B by using

$$\mathcal{D}_{\text{mix}} = \mathcal{D}_A \cup \mathcal{D}_B,$$

without further interaction with either environment.

4 Methodology

We propose Reward Improvement Field Transfer (RIFT), a data-augmentation method for cross-context offline reinforcement learning. RIFT learns a reward-improving transformation from the well-sampled source context A to generate structured additional transitions for the undersampled target context B (see Figure 2).

4.1 Algorithm

First, we construct a shared latent representation of transitions from $\mathcal{D}_{\text{mix}} = \mathcal{D}_A \cup \mathcal{D}_B$. To organize this representation by reward, we label the bottom and top reward quantiles within each context:

$$y(x) = \begin{cases} 0, & r_c \le Q_c(q), \\ 1, & r_c \ge Q_c(1-q), \end{cases} \qquad c \in \{A, B\},$$

where Q_c is the reward quantile function and $q \in (0, 1/2)$ is a fixed hyperparameter. Smaller q weakly increases tail-reward separation but leaves fewer observations for field fitting and target anchoring.

We then train a shared variational autoencoder (VAE) (Kingma & Welling, 2014) on transitions $x = (s_t, a_t, s_{t+1})$. To encourage a shared reward direction across contexts, we apply a regularization term on the transitions with low- and high- rewards. Beyond the normal prior at $N(0, I)$, we assign low- and high-reward transitions an additional Gaussian prior on one latent coordinate z_0:

$$p(z_0 \mid y = 0) = \mathcal{N}(-c_q, s_q^2), \qquad p(z_0 \mid y = 1) = \mathcal{N}(c_q, s_q^2).$$

Their means and variance match those of the lower and upper q-tails of a standard normal, so both are determined by the existing quantile parameter q. Training combines reconstruction loss with KL regularization toward the corresponding priors. We use the encoder posterior mean

$$z = e_\vartheta(s_t, a_t, s_{t+1})$$

as the latent representation, while the decoder g_ϑ maps latent samples back to complete transitions. Applying the same class-conditioned prior in both contexts encourages a common low-to-high reward direction along z_0, supporting the transfer of transformations learned in A to B.

We then use the well-sampled source context A to estimate a transformation from low- to high-reward transitions. We first lift each source point to $u = (z, r)$ by appending its observed reward and normalize to account for differences in location and scale between contexts:

$$\tilde{u} = \boldsymbol{\mu}_B + \boldsymbol{\sigma}_B \cdot \frac{u - \boldsymbol{\mu}_A}{\boldsymbol{\sigma}_A},$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ are the mean and standard deviation over all samples in context c. Mini-batch optimal transport (Villani, 2009) then matches normalized low- and high-reward points under a squared Euclidean cost on the latent component.

Using the matched pairs, we compute the reward-improving displacements $\delta_i = \tilde{v}_i - \tilde{u}_i$ where $\tilde{v}_i$ is the high-reward point and $\tilde{u}_i$ is the low reward point and fit an affine displacement model by ordinary least squares:

$$\widehat{\delta}(\tilde{u}_i) = \widehat{\boldsymbol{W}}\tilde{u}_i + \widehat{\boldsymbol{b}}$$

where

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{b}}) = \arg\min_{\boldsymbol{W}, \boldsymbol{b}} \sum_i \|\delta_i - (\boldsymbol{W}\tilde{u}_i + \boldsymbol{b})\|_2^2 .$$

Then, for each low-reward target anchor u^B, we generate the synthetic point $\widetilde{u}^B = u^B + \widehat{\delta}(u^B)$. Its latent component is decoded into a transition, while its reward component is used as the synthetic reward. The terminal flag is computed from the decoded next state using the known termination function, and samples with terminal initial states are discarded.

This structured model is motivated by the scarcity of target-context data. In particular, $\mathcal{D}_B$ may be too small to construct a dynamics model or support a flexible neural transition generator. We thus simplify the problem to use optimal transport to provide explicit low-to-high reward structure from the abundant context A to target context B.

For evaluation, we train a downstream offline RL policy using implicit Q-learning (IQL) (Kostrikov et al., 2022) and evaluate it in context B. We standardize observations using means and standard deviations. Each minibatch draws a fraction $1 - \lambda$ of its transitions uniformly from $\mathcal{D}_A$ and the remaining fraction λ uniformly from $\mathcal{D}_B \cup \mathcal{D}'_B$, where $\mathcal{D}'_B$ contains RIFT's synthetic target transitions. The weight λ balances the additional coverage and transition information available in the source data against the reward and dynamics mismatch they introduce. The realized synthetic fraction within the target pool is $\theta = |\mathcal{D}'_B|/(|\mathcal{D}_B| + |\mathcal{D}'_B|)$.

4.2 Theoretical Analysis

RIFT is motivated by the hypothesis that, for related domains, the change associated with reward improvement can transfer even when the underlying transition distributions differ. We formalize this idea in the shared latent space, bound synthesis error, and compare target anchoring with directly decoding source data. All proofs and the conditional downstream guarantee are in Appendix A.

Setting. Let η_c^y be the transition–reward law in context c and immediate-reward class y, and let $\mu_c^y = E_\# \eta_c^y$, where $E(x,r) = (e_\vartheta(x), r)$ and $D(z,r) = (g_\vartheta(z), r)$. Write $\bar{\mu}_A^y = S_\# \mu_A^y$ for source points in target coordinates, using the normalization of Section 4.1. All norms and Wasserstein distances use Euclidean coordinates, and all displayed laws have finite second moments. Fix the representation, normalization, and fitted field. A population coupling $\pi_A \in \Pi(\bar{\mu}_A^0, \bar{\mu}_A^1)$ matches joint points by squared distance between their latent blocks. Define its population affine fit by

$$\delta_{\mathrm{aff}}(u) = Wu + b \in \arg\min_{\delta \text{ affine}} \mathbb{E}_{(X,Y)\sim\pi_A} \|Y - X - \delta(X)\|^2, \qquad F = \mathrm{id} + \widehat{\delta}, \quad \nu = (D \circ F)_\# \mu_B^0.$$

Fix one minimizer if necessary, such that ν is a population synthesis law. A latent transport map or exact reward decoding is not needed for the following result.

Assumption 1 (Shared class geometry)**.** *There is a common invertible affine map $h(u) = Hu + t$ such that*

$$W_2(\mu_B^y, h_\# \bar{\mu}_A^y) \le \rho_y, \qquad y = 0, 1.$$

It represents the residual cross-context geometry after normalization. The VAE representation and reward-conditioned regularization are designed to promote this correspondence.

Synthesis error. Set $\Gamma_h(u) = \delta_{\mathrm{aff}}(u) - H\delta_{\mathrm{aff}}(h^{-1}(u))$ and define

$$\varepsilon_{\mathrm{aff}} = \left(\mathbb{E}_{\pi_A} \|Y - X - \delta_{\mathrm{aff}}(X)\|^2\right)^{1/2}, \quad \varepsilon_{\mathrm{est}} = \|\widehat{\delta} - \delta_{\mathrm{aff}}\|_{L^2(\mu_B^0)},$$
$$\varepsilon_{\mathrm{conj}} = \|\Gamma_h\|_{L^2(\mu_B^0)}, \qquad \varepsilon_{\mathrm{dec}} = \left(\mathbb{E}_{(x,r)\sim\eta_B^1} \|g_\vartheta(e_\vartheta(x)) - x\|^2\right)^{1/2}.$$

These measure source affine approximation, estimation on target anchors, context incompatibility, and target-high reconstruction, respectively. Write $L_h = \|H(I+W)H^{-1}\|$ and $L_D = \max(1, L_g)$, where g_ϑ is L_g-Lipschitz. Finally, define the population transfer-error budget as

$$\varepsilon_{\mathrm{tr}} := \varepsilon_{\mathrm{conj}} + \|H\|\varepsilon_{\mathrm{aff}} + L_h\rho_0 + \rho_1.$$

It bounds the error from applying the population source affine field to target anchors, before estimation and decoding.

Theorem 1 (Accuracy of transferred synthesis)**.** *Under Assumption 1, the fixed-object and finite-moment setting above, and a Lipschitz decoder,*

$$W_2(\nu, \eta_B^1) \le \varepsilon_{\mathrm{syn}} := L_D\left[\varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{tr}}\right] + \varepsilon_{\mathrm{dec}}. \tag{1}$$

Equivalently, a generated point can be coupled to a genuine target-high point with root-mean-square distance at most $\varepsilon_{\mathrm{syn}}$.

For a residual shift $h(u) = u + t$, the incompatibility is $\varepsilon_{\mathrm{conj}} = \|Wt\|$: a large common shift does not need to cause large transfer error when the improvement field is nearly constant.

Proposition 1 (Comparison with source-high transplant)**.** *Suppose the encoder is L_e-Lipschitz, put $L_E = \max(1, L_e)$, and let $\varepsilon_{\mathrm{cyc}} = (\mathbb{E}_{(z,r)\sim\bar{\mu}_A^1} \|e_\vartheta(g_\vartheta(z)) - z\|^2)^{1/2}$. With finite second moments,*

$$W_2(D_\# \bar{\mu}_A^1, \eta_B^1) \ge \frac{W_2(\bar{\mu}_A^1, \mu_B^1) - \varepsilon_{\mathrm{cyc}}}{L_E}. \tag{2}$$

If this lower bound exceeds $\varepsilon_{\mathrm{syn}}$ under Theorem 1's assumptions, RIFT is strictly closer to η_B^1 than transplant is. Under Assumption 1 with $h(u) = u + t$, transplant's lower bound is at least $(\|t\| - \rho_1 - \varepsilon_{\mathrm{cyc}})/L_E$, whereas RIFT's field-incompatibility contribution is $L_D\|Wt\|$.

From source fitting to synthesis accuracy. Theorem 1 separates field-estimation error, the population transfer-error budget, and decoding error. Appendix A.3 controls this estimation error under additional regularity and sampling conditions. For the population coupling and sample-size condition specified there, the fitted field satisfies

$$\varepsilon_{\mathrm{est}} \le \overline{\varepsilon}_{\mathrm{est}}(J) := \kappa_M \left[B_k + C_2\sqrt{\frac{\log(2/\eta)}{J}}\right]$$

with probability at least $1 - \eta$, for fixed batch size k and confidence parameter $\eta \in (0, 1)$. Here J is the number of independent source batches, B_k bounds the bias from matching batches of k points

per class, and κ_M accounts for the change from source prediction error to error on target anchors, while C_2 depends on source boundedness and moment conditioning.

Substituting this estimate into Theorem 1 gives the corresponding high-probability synthesis bound:

$$W_2(\nu, \eta_B^1) \leq L_D\left[\overline{\varepsilon}_{\mathrm{est}}(J) + \varepsilon_{\mathrm{tr}}\right] + \varepsilon_{\mathrm{dec}}. \tag{3}$$

Under the conditions of Proposition 1, if the right-hand side of equation 3 is smaller than the right-hand side of equation 2, then RIFT's population synthetic law is strictly closer to the genuine target high-reward law than source-high transplant, with probability at least $1 - \eta$. Thus the comparison applies to an estimated improvement field, not only to an ideal population field.

For fixed matching-batch size k, increasing the number J of independent source batches reduces the $J^{-1/2}$ sampling term in $\overline{\varepsilon}_{\mathrm{est}}(J)$, but not its minibatch-bias contribution $\kappa_M B_k$, the population transfer-error budget $\varepsilon_{\mathrm{tr}}$, or the decoding error $\varepsilon_{\mathrm{dec}}$. Suppose the combined minibatch-bias, transfer, and decoding contribution $L_D(\kappa_M B_k + \varepsilon_{\mathrm{tr}}) + \varepsilon_{\mathrm{dec}}$ lies below the transplant lower bound in equation 2. For fixed k and η, sufficiently many independent source batches then make RIFT's population synthesis error strictly smaller than transplant's with probability at least $1 - \eta$. The relevant distinction is therefore between transferring an accurately learned improvement rule and transferring source points themselves; the source and target class distributions need not be close before the context transformation.

Sampling interpretation. Theorem 1 describes RIFT's pre-filtering synthetic distribution $\nu = (D \circ F)_{\#}\mu_B^0$, rather than a particular finite dataset. If $\varepsilon_{\mathrm{syn}} = 0$, this distribution equals η_B^1. In this ideal case, applying the fixed RIFT generator to independent draws from the target low-reward distribution produces independent draws from the target high-reward distribution. Even in this ideal setting, a finite generated dataset still has sampling error, quantified in Appendix A.2.

Conditional downstream consequence. Theorem A.1 in Appendix A.4 connects equation 1 to return. Under target and surrogate regularity, occupancy domination, and a surrogate learning-error allowance $\varepsilon_{\mathrm{learn}}$, it gives

$$J_B(\widehat{\pi}) \geq J_B(\pi_c) - \varepsilon_{\mathrm{learn}} - \mathcal{R}_{\lambda,\theta}(\widehat{\pi}) - \mathcal{R}_{\lambda,\theta}(\pi_c).$$

The explicit budget $\mathcal{R}$ includes synthesis error, source reward and dynamics mismatch, filtering, and surrogate fitting; the episodic version also accounts for terminal flags. The guarantee is conditional on the stated surrogate and coverage assumptions.

5 Experiments

We evaluate RIFT on the ODRL cross-domain offline RL benchmark (Lyu et al., 2024a). We study whether RIFT improves policy performance when target-domain data are scarce, and construct ablation studies on key hyperparameters to further understand the components of RIFT. Please check Appendix C for hyperparameter details and Appendix D for more experimental results.

5.1 Setup

Tasks and evaluation protocol. We evaluate HalfCheetah, Hopper, Walker2d, and Ant under the gravity, morphology, and friction shifts provided by ODRL, using the same benchmark environments as OTDF (Lyu et al., 2025) and STC (Yan et al., 2026). The gravity setting multiplies target gravity by 0.5, while the friction setting multiplies the target friction coefficients by 0.5. For morphology, we use ODRL's medium morphology shift, which modifies one body component for each task. The source dataset $\mathcal{D}_A$ is the full ODRL source dataset, containing approximately one million transitions. The target dataset $\mathcal{D}_B$ contains $5{,}000$ transitions and is held fixed across policy seeds. Dataset qualities follow the D4RL convention (Fu et al., 2020): medium (`med`), medium-replay (`med-r`), medium-expert (`med-e`), and expert. Please see Appendix B for further dataset details.

Baselines. We compare against several established cross-domain offline RL methods. IQL (Kostrikov et al., 2022) serves as the standard offline RL baseline trained on pooled source and target data. DARA (Liu et al., 2022) adjusts source rewards using domain-discrimination signals,

while BOSA (Liu et al., 2024) restricts policy and value learning to supported regions. SRPO (Xue et al., 2023) incorporates stationary-state information into reward shaping, and IGDF (Wen et al., 2024) uses a learned contrastive criterion to decide which source transitions to transfer. OTDF (Lyu et al., 2025) uses optimal transport to score and filter source transitions based on target compatibility. STC (Yan et al., 2026) instead learns corrections that map source transitions toward the target domain and removes corrections judged unreliable before downstream policy training. Published baseline results in the source- and target-quality comparisons are taken from STC under the same ODRL benchmark protocol, with 1M gradient steps and five random seeds. Further implementation details are in Appendix C.

5.2 Main Results

Performance across source qualities. Table 1 compares RIFT with prior cross-domain offline RL methods while fixing the target dataset to $5,000$ medium-quality transitions and varying source-data quality. Across the 24 gravity and morphology settings, RIFT achieves the highest mean score in 16 settings and the highest aggregate normalized score of 1318.7. This is a 26.2% improvement over STC (1045.3), the strongest competing method in aggregate. The gains are particularly pronounced on HalfCheetah and Hopper, especially under morphology shifts, where RIFT remains strong across medium, medium-replay, and medium-expert source datasets. In contrast, the advantage is less consistent on Walker2d and Ant, where STC, OTDF, or another baseline obtains the highest score in several settings.

Several earlier adaptation methods, including DARA, BOSA, SRPO, and IGDF, remain relatively close to the pooled-IQL baseline in many settings. OTDF and STC provide substantially stronger comparisons, suggesting that explicitly accounting for cross-domain transition mismatch is important when only a small target dataset is available. RIFT further improves aggregate performance by using the abundant source dataset to learn how transitions change with reward improvement.

Performance across target qualities. Table 5 instead fixes the source dataset to medium quality and varies the quality of the limited target dataset. RIFT obtains the highest mean score in 13 of the 20 reported settings and again achieves the highest aggregate normalized score, 940.9, compared with 820.9 for the second highest method, STC. This corresponds to a 14.6% aggregate improvement over STC.

RIFT performs particularly well when the target data are medium quality, where augmentation can provide additional high-reward examples from a small target dataset. However, its advantage tends to decrease as target quality increases, as we can see by comparing the results from Table 1 and Table 5. This behavior is consistent with the motivation for RIFT: when useful target-domain behavior is already well represented in the observed dataset, the potential benefit of constructing additional reward-improving transitions is reduced. Overall, the two evaluations show that RIFT remains effective across changes in both source and target dataset quality, while its largest gains occur in settings where the scarce target data leaves meaningful room for augmentation.

5.3 Ablations

We investigate the augmentation strategy and reward-quantile threshold across eight medium-to-medium gravity and morphology settings. All experiments follow the main evaluation protocol, except for the component being varied. We denote the baseline by IQL*, which follows the same selection protocol as RIFT. This distinguishes it from the published IQL results used in our main comparisons.

Augmentation strategy. Table 2 compares RIFT against Mixup (Zhang et al., 2018) and SMOTE (Chawla et al., 2002), applied either to the target dataset alone or to the combined source and target datasets. Mixup and SMOTE generate 1,000 synthetic transitions; RIFT generates up to 1,000 after filtering. RIFT achieves the highest aggregate score of 451.3, compared with 394.5 for the strongest SMOTE variant. Although generic augmentation performs better on some individual tasks, RIFT's aggregate gains suggest that transferring a structured reward-improvement field is more effective than interpolation alone in these settings.

Table 1: Performance under gravity and morphology shifts when the target dataset contains medium-quality transitions and source-data quality varies. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

Type	Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	STC	**RIFT (ours)**
Gravity	half-med	med	39.6 ± 3.3	41.2 ± 3.9	38.9 ± 4.0	36.9 ± 4.5	36.6 ± 5.5	40.7 ± 7.7	42.4 ± 5.3	$\mathbf{47.3 \pm 6.6}$
	half-med-r	med	20.1 ± 5.0	17.6 ± 6.2	20.0 ± 4.9	17.5 ± 5.2	14.4 ± 2.2	21.5 ± 6.5	26.7 ± 2.2	$\mathbf{31.4 \pm 5.3}$
	half-med-e	med	38.6 ± 6.0	37.8 ± 3.3	41.8 ± 5.1	42.5 ± 2.3	37.7 ± 7.3	39.5 ± 3.5	39.2 ± 4.2	$\mathbf{49.6 \pm 5.3}$
	hopp-med	med	11.2 ± 1.1	17.3 ± 3.8	15.2 ± 3.3	12.4 ± 1.0	15.3 ± 3.5	32.4 ± 8.0	43.4 ± 6.1	$\mathbf{60.4 \pm 7.3}$
	hopp-med-r	med	13.9 ± 2.9	10.7 ± 4.3	3.3 ± 1.9	14.0 ± 2.6	15.3 ± 4.4	31.1 ± 13.4	36.8 ± 17.8	$\mathbf{61.3 \pm 4.3}$
	hopp-med-e	med	19.1 ± 6.6	18.5 ± 12.3	15.9 ± 5.9	19.7 ± 8.5	22.3 ± 5.4	26.4 ± 10.1	45.3 ± 7.5	$\mathbf{65.2 \pm 6.5}$
	walk-med	med	28.1 ± 12.9	28.4 ± 13.7	38.0 ± 11.2	21.4 ± 7.0	22.1 ± 8.4	36.6 ± 2.3	41.6 ± 4.0	$\mathbf{46.5 \pm 5.5}$
	walk-med-r	med	14.6 ± 2.5	14.1 ± 6.1	7.6 ± 5.8	17.9 ± 3.8	11.6 ± 4.6	32.7 ± 7.0	29.0 ± 1.9	$\mathbf{33.0 \pm 8.8}$
	walk-med-e	med	39.9 ± 13.1	41.6 ± 13.0	32.3 ± 7.2	$\mathbf{46.4 \pm 3.5}$	33.8 ± 3.1	30.2 ± 9.8	34.9 ± 9.1	40.3 ± 7.6
	ant-med	med	10.2 ± 1.8	9.4 ± 0.9	12.4 ± 2.0	11.7 ± 1.0	11.3 ± 1.3	$\mathbf{45.1 \pm 12.4}$	42.6 ± 8.4	40.8 ± 10.3
	ant-med-r	med	18.9 ± 2.6	21.7 ± 2.1	13.9 ± 1.5	18.7 ± 1.7	19.6 ± 1.0	29.6 ± 10.7	40.9 ± 5.6	$\mathbf{46.7 \pm 3.5}$
	ant-med-e	med	9.8 ± 2.4	8.1 ± 1.8	8.1 ± 3.0	8.4 ± 2.1	8.9 ± 1.5	18.6 ± 11.9	$\mathbf{39.2 \pm 9.2}$	38.1 ± 7.0
Morph	half-med	med	24.5 ± 2.4	21.0 ± 3.9	24.2 ± 5.6	18.1 ± 1.8	23.7 ± 3.4	21.1 ± 7.6	19.5 ± 2.2	$\mathbf{57.9 \pm 4.4}$
	half-med-r	med	11.0 ± 1.2	9.5 ± 2.3	4.7 ± 2.9	8.9 ± 1.2	9.2 ± 0.6	6.5 ± 1.4	13.0 ± 5.3	$\mathbf{49.2 \pm 4.6}$
	half-med-e	med	21.1 ± 2.8	19.2 ± 2.2	23.2 ± 3.9	21.1 ± 1.9	18.6 ± 1.3	20.8 ± 2.5	16.8 ± 8.8	$\mathbf{60.5 \pm 3.1}$
	hopp-med	med	15.9 ± 6.8	17.8 ± 10.1	12.8 ± 0.1	21.7 ± 7.7	25.3 ± 9.7	16.4 ± 7.1	43.1 ± 23.9	$\mathbf{69.4 \pm 13.2}$
	hopp-med-r	med	12.9 ± 0.3	12.8 ± 0.1	2.0 ± 1.2	12.4 ± 0.7	12.5 ± 1.7	13.3 ± 0.1	22.9 ± 6.1	$\mathbf{61.8 \pm 10.6}$
	hopp-med-e	med	14.9 ± 3.1	11.1 ± 5.6	14.4 ± 1.8	16.6 ± 1.9	18.3 ± 7.5	25.4 ± 9.4	53.4 ± 20.8	$\mathbf{65.7 \pm 10.4}$
	walk-med	med	31.5 ± 8.6	35.0 ± 10.8	26.7 ± 6.6	38.6 ± 5.1	38.5 ± 8.4	42.5 ± 3.1	$\mathbf{56.7 \pm 8.1}$	52.8 ± 8.2
	walk-med-r	med	41.5 ± 3.0	38.5 ± 7.9	15.3 ± 7.6	36.0 ± 4.4	24.2 ± 8.6	17.9 ± 13.4	$\mathbf{63.1 \pm 8.0}$	46.9 ± 6.7
	walk-med-e	med	32.8 ± 4.3	41.4 ± 10.9	45.1 ± 13.7	39.8 ± 14.3	37.9 ± 4.2	55.3 ± 2.2	$\mathbf{62.1 \pm 8.1}$	61.3 ± 7.0
	ant-med	med	71.4 ± 2.4	71.5 ± 6.8	54.8 ± 13.2	72.8 ± 2.2	71.8 ± 2.7	75.1 ± 2.3	$\mathbf{77.2 \pm 2.8}$	76.2 ± 2.4
	ant-med-r	med	65.9 ± 5.5	62.3 ± 8.2	15.2 ± 2.3	59.3 ± 4.0	65.0 ± 5.3	63.1 ± 5.9	76.2 ± 2.6	$\mathbf{78.1 \pm 0.4}$
	ant-med-e	med	70.2 ± 7.3	64.3 ± 5.8	64.0 ± 6.0	68.5 ± 4.4	66.8 ± 11.0	76.4 ± 1.9	$\mathbf{79.3 \pm 0.2}$	78.3 ± 0.3
Total (displayed means)			677.6	670.8	549.8	681.3	660.7	818.2	1045.3	**1318.7**

Table 2: Comparison of augmentation strategies across eight medium-to-medium gravity and morphology settings. Mixup and SMOTE generate 1,000 synthetic transitions using either pooled source and target data ($A \cup B$) or target data alone (B). RIFT generates up to 1,000 synthetic transitions. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

			Mixup		SMOTE		
Shift	Task	IQL*	$A \cup B$	B	$A \cup B$	B	RIFT
Gravity	Ant	29.8 ± 4.9	29.5 ± 1.9	26.1 ± 6.9	28.2 ± 7.5	34.4 ± 7.2	$\mathbf{40.8 \pm 10.3}$
	HalfCheetah	42.0 ± 2.3	42.9 ± 2.1	43.3 ± 2.0	45.2 ± 4.7	45.4 ± 4.9	$\mathbf{47.3 \pm 6.6}$
	Hopper	52.7 ± 12.5	49.1 ± 12.1	44.1 ± 12.8	$\mathbf{61.0 \pm 6.6}$	60.8 ± 8.6	60.4 ± 7.3
	Walker2d	30.6 ± 4.3	43.9 ± 6.1	36.5 ± 13.6	45.1 ± 6.6	40.5 ± 5.5	$\mathbf{46.5 \pm 5.5}$
Morphology	Ant	76.9 ± 0.3	74.1 ± 2.8	76.2 ± 0.7	$\mathbf{77.3 \pm 0.5}$	$\mathbf{77.3 \pm 0.3}$	76.2 ± 2.4
	HalfCheetah	22.8 ± 2.4	18.8 ± 1.5	17.3 ± 1.8	18.8 ± 2.5	20.2 ± 4.8	$\mathbf{57.9 \pm 4.4}$
	Hopper	55.5 ± 6.4	59.9 ± 13.9	46.6 ± 10.3	54.3 ± 12.1	60.2 ± 11.2	$\mathbf{69.4 \pm 13.2}$
	Walker2d	49.2 ± 5.1	54.1 ± 8.6	$\mathbf{59.0 \pm 9.4}$	54.3 ± 7.8	55.7 ± 10.9	52.8 ± 8.2
Sum		359.5	372.3	349.1	384.2	394.5	**451.3**

Reward quantile. Table 6 examines the effect of the reward quantile q. Our default $q = 0.2$ achieves the highest aggregate score of 451.3, although the preferred value varies across tasks. This is consistent with the tradeoff described in Section 4.1. The endpoint $q = 0.5$ is also evaluated, while the formal analysis uses $q < 0.5$.

VAE design. Table 7 compares a VAE trained on uniformly pooled data (base), one with balanced source–target sampling (bal), and our reward-regularized balanced variant (bal + reg). Their aggregate scores are 394.6, 432.7, and 451.3, respectively. Our full variant achieves the highest or a comparable score across all eight settings, suggesting that reward regularization can improve latent-space structure.

5.4 Latent Space Representation

We provide empirical diagnostics of the shared representation. Figure 3 visualizes the shared VAE representations for HalfCheetah and Walker2d under gravity shifts. With balanced sampling and reward regularization, low- and high-reward transitions from both domains exhibit a common di-

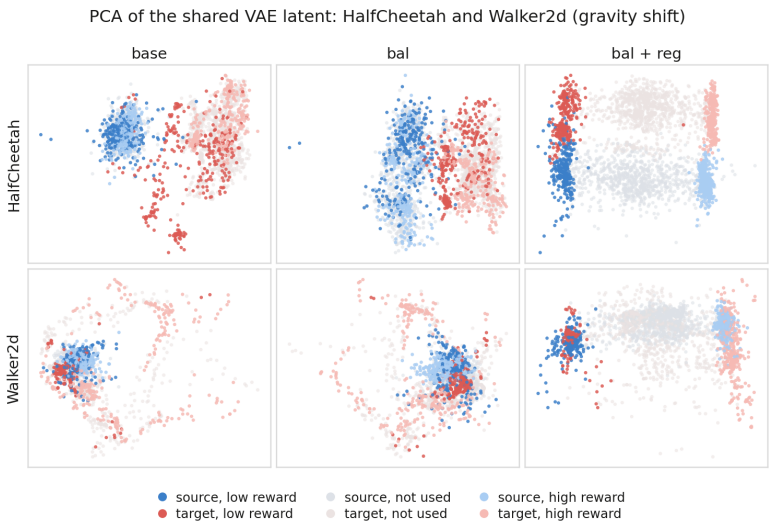

Figure 3: PCA projections of shared VAE representations for HalfCheetah and Walker2d under gravity shifts. Colors distinguish source and target transitions and their reward classes. The regularized representation exhibits a similar low-to-high reward direction across domains.

rection of separation. This is consistent with the motivation for shared improvement geometry, but a PCA projection does not capture the full-dimensional alignment or field-incompatibility terms in Theorem 1. Table 8 provides reconstruction quality for the same environments. Adding reward regularization preserves high reconstruction quality, explaining over 98% of the variance in both domains for both tasks. The variance-explained metric is therefore only a reconstruction diagnostic and is distinct from the target-high RMS error appearing in Theorem 1.

6 Conclusions

We introduced Reward Improvement Field Transfer (RIFT), a method for cross-domain offline RL that uses abundant source data to learn how behavior changes from low reward to high reward, then transfers this improvement structure to a data-scarce target domain. Rather than directly reusing source transitions, RIFT learns a shared latent representation, estimates low-to-high reward transformations with optimal transport, and applies them to target transitions to generate additional training data. Our theoretical analysis bounds synthesis error, identifies when target anchoring is more accurate than source-high transplant, and gives a conditional connection to target return. Across locomotion tasks with changes in gravity, morphology, and friction, RIFT improves aggregate performance over existing cross-domain offline RL methods, with the largest gains when target data are limited and leave room for useful augmentation. These results suggest that, even when source and target dynamics differ, the structure of *how behavior improves* can provide useful information for transfer.

There are also limitations to RIFT. The method assumes that source and target domains share a similar reward-improvement structure, which may not hold under larger domain shifts. RIFT also uses a simple affine transformation, which may be too restrictive when the change from low- to high-reward behavior is more complex. The analysis conditions on the representation and uses additional regularity and sampling assumptions; it does not establish an IQL-specific learning rate or a positive return gain from accurate synthesis alone.

AI use statement

In this work, we used generative AI tools for assistance with writing, code development, extended literature review, and mathematical formalization and verification. For writing, we used generative AI for clarification, LaTeX formatting, and expanding author-written outlines into prose. For the

theoretical analysis, the underlying mathematical ideas and reasoning were developed by the authors, while generative AI was used to assist in formalizing arguments, writing portions of proofs, and checking formulations and consistency. We did not use generative AI to develop the core research ideas, hypotheses, or initial problem scoping and literature review; generative AI was used subsequently to help identify additional related work.

We have reviewed all AI-assisted work. In particular, we checked LLM-assisted code for consistency with the intended algorithms and implementation specifications, and independently reviewed all AI-assisted mathematical arguments and cited literature. We take responsibility for the final content of this work, including all text, claims, code, proofs, and other artifacts produced with the aid of generative AI.

Ethics statement

This work does not introduce any obvious ethical concerns beyond the standard concerns for machine learning systems.

Reproducibility statement

We provide our implementation as supplementary material along with the hyperparameters and experimental details required to reproduce our results. The appendix documents all configurations that we used for RIFT in Appendix C. Our experiments build on the publicly available ODRL benchmark and OTDF codebases, and we follow their benchmark environments and experimental conventions where applicable. For the theoretical results, Section 4.2 states the synthesis assumptions and main results; Appendix A gives all proofs and the additional assumptions for estimation and conditional return guarantees.

References

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. 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.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2257–2269, 2020.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In *International Conference on Learning Representations*, 2021.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: Asymptotic and gradient properties. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141. PMLR, 2020. URL `https://proceedings.mlr.press/v108/fatras20a.html`.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *International Conference on Learning Representations*, 2022.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Shani Gamrian and Yoav Goldberg. Transfer learning for related reinforcement learning tasks via image-to-image translation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2063–2072. PMLR, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Yihong Guo, Yixuan Wang, Yuanyuan Shi, Pan Xu, and Anqi Liu. Off-dynamics reinforcement learning via domain adaptation and reward augmented imitation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4331.

Donald Hejna, Lerrel Pinto, and Pieter Abbeel. Hierarchically decoupled imitation for morphological transfer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4159–4171. PMLR, 2020.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

Xingshuai Huang, Di Wu, and Benoit Boulet. Goal-conditioned data augmentation for offline reinforcement learning. *Transactions on Machine Learning Research*, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

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

Jaewoo Lee, Sujin Yun, Taeyoung Yun, and Jinkyoo Park. GTA: Generative trajectory augmentation with guidance for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1808.

Guanghe Li, Yixiang Shan, Zhengbang Zhu, Ting Long, and Weinan Zhang. DiffStitch: Boosting offline reinforcement learning with diffusion-based trajectory stitching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28597–28609. PMLR, 2024.

Wenbo Li and Ricardo H. Nochetto. Quantitative stability and error estimates for optimal transport plans. *IMA Journal of Numerical Analysis*, 41(3):1941–1965, 2021. doi: 10.1093/imanum/draa045.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11682.

Jinxin Liu, Hongyin Zhang, and Donglin Wang. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=9SDQB3b68K`.

Jinxin Liu, Ziqi Zhang, Zhenyu Wei, Zifeng Zhuang, Yachen Kang, Sibo Gai, and Donglin Wang. Beyond OOD state actions: Supported cross-domain offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13945–13953, 2024. doi: 10.1609/aaai.v38i12.29302.

Jiafei Lyu, Kang Xu, Jiacheng Xu, Mengbei Yan, Jingwen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, and Xiu Li. ODRL: A benchmark for off-dynamics reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1912.

Jiafei Lyu, Kang Xu, Jiacheng Xu, Mengbei Yan, Jingwen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, and Xiu Li. ODRL: Off-dynamics reinforcement learning benchmark code, 2024b. URL `https://github.com/OffDynamicsRL/off-dynamics-rl`. Official implementation of the ODRL benchmark.

Jiafei Lyu, Mengbei Yan, Zhongjian Qiao, Runze Liu, Xiaoteng Ma, Deheng Ye, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In *International Conference on Learning Representations*, 2025.

Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135, 2024. doi: 10.1214/23-AAP1986.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, and Zhixuan Liang. MetaDiffuser: Diffusion model as conditional planner for offline meta-RL. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26087–26105. PMLR, 2023.

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *arXiv preprint arXiv:2206.13464*, 2022.

Dirk Ormoneit and Śaunak Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2–3): 161–178, 2002.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2011.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 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. doi: 10.1109/ICRA.2018.8460528.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xinxing Shi, Xiaoyu Jiang, and Mauricio A. Álvarez. Neighbour-driven gaussian process variational autoencoders for scalable structured latent modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54934–54959. PMLR, 2025.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Leitian Tao, Xuefeng Du, and Sharon Li. Limited preference data? learning better reward model with latent space synthesis. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 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.

Linh Le Pham Van, Minh Hoang Nguyen, Duc Kieu, Hung Le, Hung The Tran, and Sunil Gupta. DmC: Official code and experimental logs, 2025a. URL `https://github.com/linhlpv/DmC`. Experimental logs available under `offdynamics/logs`.

Linh Le Pham Van, Minh Hoang Nguyen, Duc Kieu, Hung Le, Hung The Tran, and Sunil Gupta. DmC: Nearest neighbor guidance diffusion model for offline cross-domain reinforcement learning. In *ECAI 2025*, 2025b. doi: 10.3233/FAIA251077.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Xiaoyu Wen, Chenjia Bai, Kang Xu, Xudong Yu, Yang Zhang, Xuelong Li, and Zhen Wang. Contrastive representation for data filtering in cross-domain offline reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52720–52743. PMLR, 2024. URL https://proceedings.mlr.press/v235/wen24a.html.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 550–559. PMLR, 2020. URL https://proceedings.mlr.press/v124/xie20a.html.

Kang Xu, Chenjia Bai, Xiaoteng Ma, Dong Wang, Bin Zhao, Zhen Wang, Xuelong Li, and Wei Li. Cross-domain policy adaptation via value-guided data filtering. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3210.

Zhenghai Xue, Qingpeng Cai, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. State regularized policy optimization on data with dynamics shift. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1427.

Mengbei Yan, Jiafei Lyu, Shengjie Sun, Zhongjian Qiao, Jingwen Yang, Zichuan Lin, Deheng Ye, and Xiu Li. Cross-domain offline policy adaptation via selective transition correction. *Transactions on Machine Learning Research*, 2026.

Hao-Ran Yang, Chuan-Xian Ren, and You-Wei Luo. COD: Learning conditional invariant representation for domain adaptation regression. In *European Conference on Computer Vision*, pp. 108–125, 2024.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019.

A Theoretical Results and Proofs

A.1 Formal setting and conventions

A labeled transition is $\xi = (x, r) \in \Xi := \mathbb{R}^{d_x} \times \mathbb{R}$ with $x = (s, a, s')$, and η_c^y denotes its population law in reward class $y \in \{0, 1\}$ of context $c \in \{A, B\}$. For $q \in (0, 1/2)$, classes are the bottom and top q fractions of rewards. At atoms on a threshold, independent uniform ranks break ties so that each class has probability q. This convention defines the population laws used in the analysis; it does not assume that fitting an empirical quantile introduces no error.

Write $\mathcal{U} = \mathbb{R}^{d_z} \times \mathbb{R}$ and $d_u = d_z + 1$. The lift $E(x, r) = (e_\vartheta(x), r)$ and decoder $D(z, r) = (g_\vartheta(z), r)$ give $\mu_c^y = E_\# \eta_c^y$. The fixed normalization $S = S_B^{-1} \circ S_A$ uses the coordinatewise means and positive standard deviations of each full context distribution, including all reward classes; empirically, these statistics use all samples from that context. Thus $\bar{\mu}_A^y = S_\# \mu_A^y$. Squared latent matching is computed in these target coordinates, not before an additional anisotropic rescaling.

The Euclidean norm is used on both joint latent–reward and decoded transition–reward coordinates; the same symbol on matrices means the operator norm, with $\|\cdot\|_F$ the Frobenius norm. Write $\|f\|_{L^2(\mu)} = (\int \|f\|^2 d\mu)^{1/2}$, and let W_2 denote quadratic Wasserstein distance. All laws entering these distances have finite second moments. Standard coupling and pushforward properties are used below (Villani, 2009).

The source coupling in Section 4.2 can be any fixed minimizer of

$$\inf_{\pi \in \Pi(\bar{\mu}_A^0, \bar{\mu}_A^1)} \mathbb{E}_{(X,Y)\sim\pi} \|Z_Y - Z_X\|^2,$$

where Z_X, Z_Y are the latent blocks. Its endpoints remain joint latent–reward points. The affine fit and residual therefore include the reward displacement. If the affine minimizer is not unique, all subsequent quantities refer to one fixed representative. Synthesis uses only the coupling's marginals, not its optimality, determinism, or an absolutely continuous class law.

The theorem, Theorem 1, conditions on the learned representation, normalization, coupling, and field. The sampling and estimation results below use separate conditional-i.i.d. assumptions; they do not provide a generalization guarantee for jointly learned quantities.

A.2 Synthesis, transplant, and sampling

Proof of Theorem 1. Define the affine comparison map

$$G_h = h \circ (\mathrm{id} + \delta_{\mathrm{aff}}) \circ h^{-1}, \qquad \mathrm{Lip}(G_h) = \|H(I + W)H^{-1}\| = L_h.$$

Then

$$F - G_h = (\widehat{\delta} - \delta_{\mathrm{aff}}) + \Gamma_h, \qquad \|F - G_h\|_{L^2(\mu_B^0)} \le \varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{conj}}.$$

For $(X, Y) \sim \pi_A$, the pair $(G_h(h(X)), h(Y))$ couples $(G_h)_\# h_\# \bar{\mu}_A^0$ and $h_\# \bar{\mu}_A^1$, with difference $H(X + \delta_{\mathrm{aff}}(X) - Y)$. Its root-mean-square cost is at most $\|H\|\varepsilon_{\mathrm{aff}}$. The triangle inequality and Lipschitz pushforward bound therefore give

$$\begin{aligned} W_2(F_\# \mu_B^0, \mu_B^1) &\le \|F - G_h\|_{L^2(\mu_B^0)} + L_h W_2(\mu_B^0, h_\# \bar{\mu}_A^0) + \|H\|\varepsilon_{\mathrm{aff}} + \rho_1 \\ &\le \varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{conj}} + \|H\|\varepsilon_{\mathrm{aff}} + L_h \rho_0 + \rho_1. \end{aligned} \tag{A.1}$$

The nonlinear or nondeterministic affine-approximation error is paid only under the source coupling. The lifted decoder is $L_D = \max(1, L_g)$-Lipschitz. Coupling $D(E(\xi))$ to $\xi \sim \eta_B^1$ gives $W_2(D_\# \mu_B^1, \eta_B^1) \le \varepsilon_{\mathrm{dec}}$. Hence

$$W_2(\nu, \eta_B^1) \le L_D W_2(F_\# \mu_B^0, \mu_B^1) + \varepsilon_{\mathrm{dec}} \le \varepsilon_{\mathrm{syn}}.$$

An optimal coupling of the decoded laws gives the claimed root-mean-square interpretation. □

Context interpretation. For $h(u) = Hu + t$,

$$\Gamma_h(u) = (W - HWH^{-1})u + HWH^{-1}t + (I - H)b.$$

For a shift this equals Wt. For a rotation $H = O$, $t = 0$, and a constant field $W = 0$, it equals $(I - O)b$. Thus a constant improvement is invariant to a common translation but does not need to be invariant to a rotation. The result does not assume that normalization removes every residual class-relative or cross-coordinate change.

Proof of Proposition 1. The encoder lift is $L_E = \max(1, L_e)$-Lipschitz, so

$$W_2(D_{\#}\bar{\mu}_A^1, \eta_B^1) \geq L_E^{-1} W_2((E \circ D)_{\#}\bar{\mu}_A^1, \mu_B^1).$$

Since $(E \circ D)(z, r) = (e_\vartheta(g_\vartheta(z)), r)$, coupling by the same source-high point gives $W_2((E \circ D)_{\#}\bar{\mu}_A^1, \bar{\mu}_A^1) \leq \varepsilon_{\mathrm{cyc}}$. The reverse triangle inequality proves equation 2. Under Assumption 1, a second reverse triangle inequality yields

$$W_2(\bar{\mu}_A^1, \mu_B^1) \geq W_2(\bar{\mu}_A^1, h_{\#}\bar{\mu}_A^1) - \rho_1.$$

For a shift, $(X, X + t)$ is a coupling with cost $\|t\|$, while every coupling satisfies $\mathbb{E}\|X - Y\|^2 \geq \|\mathbb{E}(X - Y)\|^2 = \|t\|^2$. Thus $W_2(\bar{\mu}_A^1, h_{\#}\bar{\mu}_A^1) = \|t\|$. Combining the lower bound with Theorem 1 gives the strict comparison. □

Finite synthetic samples. Under the conditional-i.i.d. anchor assumption of Appendix A.1, let $\widehat{\xi}_i = D(F(U_i))$, $i = 1, \ldots, m$, be the actual generated points and $\widehat{\nu}_m$ their empirical law. On an enlarged probability space they admit partners $\xi_i^\star \sim \eta_B^1$ such that the pairs $(\widehat{\xi}_i, \xi_i^\star)$ are i.i.d. and $\mathbb{E}\|\widehat{\xi}_i - \xi_i^\star\|^2 \leq \varepsilon_{\mathrm{syn}}^2$. With $\widehat{\nu}_m^\star$ their partners' empirical law,

$$\mathbb{E}W_2^2(\widehat{\nu}_m, \widehat{\nu}_m^\star) \leq \varepsilon_{\mathrm{syn}}^2, \qquad \mathbb{E}W_2(\widehat{\nu}_m, \eta_B^1) \leq \varepsilon_{\mathrm{syn}} + \mathbb{E}W_2(\widehat{\nu}_m^\star, \eta_B^1). \tag{A.2}$$

Proof. Disintegrate an optimal coupling of (ν, η_B^1) at each actual $\widehat{\xi}_i$, using independent auxiliary draws across indices to obtain $\xi_i^\star$. Euclidean spaces admit these conditional laws. The pairs have the claimed laws and squared cost. Matching pairs proves the first inequality; the triangle inequality and Jensen prove the second. □

If $\varepsilon_{\mathrm{syn}} = 0$, then $\nu = \eta_B^1$. For a measurable region $\mathcal{O}$ fixed under the conditioning, at least one of the m independent generated points hits $\mathcal{O}$ with probability $1 - (1 - \eta_B^1(\mathcal{O}))^m$. Thus synthesis can fill empirical coverage gaps without expanding the support of the target conditional law. These statements concern candidates before rejection and condition on the retained anchor sample.

A.3 Estimating the field with latent-only minibatch matching

This section analyzes the matching cost used by RIFT under an independent-batch sampling model. The additional regularity and sampling conditions below are needed only for this estimation result, not for Theorem 1.

Population conditions. Suppose the joint source classes are supported in $\{\|u\| \leq R_u\}$, where $R_u > 0$. Put $\widetilde{U} = (U, 1)$ and define

$$M_A = \mathbb{E}_{\bar{\mu}_A^0}[\widetilde{U}\widetilde{U}^\top] \succ 0, \qquad M_B = \mathbb{E}_{\mu_B^0}[\widetilde{U}\widetilde{U}^\top],$$
$$a_* = \lambda_{\min}(M_A), \qquad \kappa_M = \|M_B^{1/2} M_A^{-1/2}\|.$$

Let $P_y = \bar{\mu}_{A,z}^y$ denote the source latent marginal for class y. Suppose that $T_z = \nabla\varphi_z$ pushes P_0 to P_1, where φ_z is globally convex and L_T-smooth, with $L_T > 0$. On the normalized source high class, assume

$$R = \psi_A(Z) \quad \text{almost surely,}$$

where ψ_A is L_ψ-Lipschitz on its support.

Reference coupling. For this estimation analysis, specialize the population source coupling used in Theorem 1 to

$$\mathcal{T}_A(z, r) = \big(T_z(z), \psi_A(T_z(z))\big), \qquad \pi_A = (\mathrm{id}, \mathcal{T}_A)_{\#}\bar{\mu}_A^0.$$

The lift $\mathcal{T}_A$ is needed only $\bar{\mu}_A^0$-almost surely.

Estimation experiment. Conditionally on the fixed representation and normalization, draw N_A independent points from each source class, with the two samples independent. Divide each sample into

$$J = \frac{N_A}{k}$$

disjoint batches of size k, where $k \geq 1$ divides N_A. Within each pair of batches, solve the latent assignment problem exactly, using a fixed measurable tie rule and independent batchwise randomness when the matching procedure is randomized. Fit $\widehat{\delta}$ by ordinary least squares to all resulting joint latent–reward displacements.

Minibatch-bias quantity. Let $\widehat{P}_k^y$ denote the empirical latent law of k independent draws from P_y. Define the expected excess matching cost and its associated bias certificate by

$$\mathrm{excess}_k := \mathbb{E}W_2^2(\widehat{P}_k^0, \widehat{P}_k^1) - W_2^2(P_0, P_1) \geq 0, \qquad B_k := \sqrt{(1 + L_\psi^2) L_T \, \mathrm{excess}_k}. \tag{A.3}$$

Proposition A.1 (Finite-sample field estimation)**.** *Under the setup above, there exist constants C_1, C_2, depending only on R_u and a_*, such that for every $\eta \in (0,1)$ and*

$$J \geq C_1 \log(2/\eta),$$

with probability at least $1 - \eta$,

$$\varepsilon_{\mathrm{est}} \leq \kappa_M \left[B_k + C_2 \sqrt{\frac{\log(2/\eta)}{J}} \right]. \tag{A.4}$$

Scope. The smooth latent transport, source-high reward graph, positive-design condition, and independent-batch experiment are sufficient conditions for Proposition A.1; they are not required by Theorem 1. The smooth transport is assumed and need not exist for arbitrary singular class laws. The condition $M_A \succ 0$ can also fail, for example when the source low-class reward is an affine function of its latent coordinates. In such cases, Theorem 1 still applies to a fixed affine field and its actual estimation error, but Proposition A.1 does not provide a certificate for that error.

Lemma A.1 (Same-marginal excess-cost control)**.** *Let P, Q have finite second moments. Suppose $T = \nabla\varphi$ pushes P to Q, where φ is convex and globally L-smooth, $L > 0$. For every $\pi \in \Pi(P, Q)$,*

$$\mathbb{E}_\pi \|Y - T(X)\|^2 \leq L \left[\mathbb{E}_\pi \|X - Y\|^2 - W_2^2(P, Q) \right]. \tag{A.5}$$

Proof. This is the transport-plan estimate of Li & Nochetto (2021, Proposition 3.1); the following gap argument gives the stated finite-moment version. For the ordinary conjugate, evaluating at $w = x + (y - T(x))/L$ and using smoothness gives

$$\varphi(x) + \varphi^*(y) - \langle x, y \rangle \geq \frac{\|y - T(x)\|^2}{2L}.$$

Global smoothness bounds $|\varphi(x)|$ quadratically and $\|T(x)\|$ linearly. Fenchel equality at $T(X)$ therefore makes φ^* integrable under $Q = T_\# P$. The gap is nonnegative and zero on $(\mathrm{id}, T)_\# P$, so this coupling minimizes quadratic cost. With fixed marginals, the integrated gap is half the excess cost in equation A.5. Integration proves the claim. □

Proof of Proposition A.1. Let Π_k be the law of a uniformly selected matched joint pair from one batch. A permutation uses every point once, so $\Pi_k \in \Pi(\bar{\mu}_A^0, \bar{\mu}_A^1)$; its latent cost is $\mathbb{E}W_2^2(\widehat{P}_k^0, \widehat{P}_k^1)$. This proves nonnegativity in equation A.3. Let δ_k be the population affine fit under Π_k. Since $M_B \preceq \kappa_M^2 M_A$, every affine f satisfies $\|f\|_{L^2(\mu_B^0)} \leq \kappa_M \|f\|_{L^2(\bar{\mu}_A^0)}$.

Bias. Put $a = \delta_k - \delta_{\mathrm{aff}}$. The normal equations and common input marginal give

$$\|a\|_{L^2(\bar{\mu}_A^0)}^2 = \mathbb{E}_{\Pi_k} \langle Y - \mathcal{T}_A(X), a(X) \rangle \leq \left(\mathbb{E}_{\Pi_k} \|Y - \mathcal{T}_A(X)\|^2 \right)^{1/2} \|a\|_{L^2(\bar{\mu}_A^0)}.$$

Both high endpoints lie on the source reward graph, so

$$\|Y - \mathcal{T}_A(X)\|^2 \leq (1 + L_\psi^2) \|Z_Y - T_z(Z_X)\|^2.$$

Apply Lemma A.1 to the latent marginal of the mean population coupling Π_k, not to an individual empirical coupling with different marginals. It follows that $\|\delta_k - \delta_{\mathrm{aff}}\|_{L^2(\bar{\mu}_A^0)} \le B_k$.

Sampling. Write $\delta_k(u) = \Theta_k \widetilde{u}$, $\widehat{\delta}(u) = \widehat{\Theta}\widetilde{u}$, and $R_1 = \sqrt{R_u^2 + 1}$. The normal equations give $\Theta_k = \mathbb{E}_{\Pi_k}[(Y - X)\widetilde{X}^\top] M_A^{-1}$, so $\|\Theta_k\| \le 2R_u R_1 / a_*$. For batch j set

$$M_j = \frac{1}{k}\sum_{i \in j} \widetilde{X}_i \widetilde{X}_i^\top, \qquad Z_j = \frac{1}{k}\sum_{i \in j} (Y_i - X_i - \Theta_k \widetilde{X}_i)\widetilde{X}_i^\top.$$

These are independent across batches, with $\mathbb{E}M_j = M_A$, $\mathbb{E}Z_j = 0$, $\|M_j\|_F \le R_1^2$, and $\|Z_j\|_F \le B := 2R_u R_1(1 + R_1^2/a_*)$. For independent matrices A_j with $\|A_j\|_F \le B_0$, independence and Jensen bound the expected centered-average Frobenius norm by $B_0/\sqrt{J}$. McDiarmid's inequality (McDiarmid, 1989), with per-batch sensitivity $2B_0/J$, yields

$$\left\| \frac{1}{J}\sum_j (A_j - \mathbb{E}A_j) \right\|_F \le \frac{B_0}{\sqrt{J}} \left(1 + \sqrt{2\log(1/\zeta)}\right)$$

with probability at least $1 - \zeta$. Apply this to M_j and Z_j with $\zeta = \eta/2$ each. Put $\widehat{M}_A = J^{-1}\sum_j M_j$ and $\widehat{Z} = J^{-1}\sum_j Z_j$. For $J \ge C_1 \log(2/\eta)$ the event gives $\|\widehat{M}_A - M_A\| \le a_*/2$, hence $\widehat{M}_A \succeq M_A/2$ and $\|\widehat{M}_A^{-1}\| \le 2/a_*$. In particular,

$$\|\widehat{M}_A^{-1} M_A^{1/2}\|^2 = \|\widehat{M}_A^{-1} M_A \widehat{M}_A^{-1}\| \le 2\|\widehat{M}_A^{-1}\| \le 4/a_*.$$

The empirical normal equations now give

$$\widehat{\Theta} - \Theta_k = \widehat{Z}\widehat{M}_A^{-1}, \qquad \|\widehat{\delta} - \delta_k\|_{L^2(\bar{\mu}_A^0)} = \|\widehat{Z}\widehat{M}_A^{-1} M_A^{1/2}\|_F \le \frac{2}{\sqrt{a_*}}\|\widehat{Z}\|_F \le C_2\sqrt{\frac{\log(2/\eta)}{J}}.$$

For $c_* = 1/\sqrt{\log 2} + \sqrt{2}$, valid choices are $C_1 = 4c_*^2 R_1^4/a_*^2$ and $C_2 = 2c_* B/\sqrt{a_*}$. Combining bias, sampling, and the affine source-to-target norm comparison proves equation A.4. □

Rate and scope. The expected absolute quadratic-cost error of Chizat et al. (2020, Theorem 2), rescaled from diameter one to a radius-R_u ball, implies

$$\mathrm{excess}_k \le C_{d_z} R_u^2 r_{d_z}(k), \qquad r_d(k) = \begin{cases} k^{-1/2}, & d < 4, \\ k^{-1/2}\log(1+k), & d = 4, \\ k^{-2/d}, & d > 4. \end{cases} \tag{A.6}$$

The expected excess is nonnegative and bounded by that absolute error. See Manole & Niles-Weed (2024) for the broader smooth-cost analysis. Thus the high-dimensional bias certificate is of order k^{-1/d_z}, while the sampling term uses $J = N_A/k$ independent batches. Bias does not need to vanish as N_A grows at fixed k (Fatras et al., 2020). Radius and moment conditioning also carry dimension dependence; the bound is not dimension-free.

Note that proposition A.1 is an idealized independent-batch result. Its assumptions—including the reward graph, fixed representation, and positive design condition—are sufficient conditions for the bound rather than empirical diagnostics of the reported experiments.

A.4 Conditional target-return guarantee

The following result connects synthesis accuracy to a specified fitted or analytical surrogate.

(R1) Target and data laws. The target has discount $\gamma \in [0, 1)$, initial law ρ, bounded reward $r_B(s, a)$, and deterministic dynamics $S' = f_B(S, A)$. The reward and dynamics are respectively L_r- and L_f-Lipschitz on a domain containing the real, candidate synthetic, and comparison coordinates. Let η_A, η_B be the full-context transition–reward laws underlying η_c^y. Genuine target tuples obey $R = r_B(S, A)$ and $S' = f_B(S, A)$. Source rewards are bounded but does not need to equal $r_B(S, A)$, and source dynamics may differ.

(R2) Filtering and joint mixture. Fix a measurable acceptance event K with $p_{\mathrm{keep}} = \nu(K) > 0$, and let $\nu_K = \nu(\cdot \mid K)$. Retained actions are feasible at their states. Without filtering take $K = \Xi$ and $p_{\mathrm{keep}} = 1$. For fixed mixture weights $\lambda, \theta \in [0, 1]$, let

$$\eta_{\lambda,\theta} = (1-\lambda)\eta_A + \lambda(1-\theta)\eta_B + \lambda\theta\nu_K, \qquad d_{\lambda,\theta} = \text{its state–action marginal.} \tag{A.7}$$

For uniform sampling within the empirical target pool, the corresponding synthetic fraction is $\theta = m_{\mathrm{keep}}/(n_B + m_{\mathrm{keep}})$. The population training backup is

$$(\mathcal{B}_{\lambda,\theta}V)(s,a) = \mathbb{E}_{\eta_{\lambda,\theta}}[R + \gamma V(S') \mid S = s, A = a]. \tag{A.8}$$

It conditions the joint mixture: component weights does not need to remain constant after conditioning on (s, a).

(R3) Surrogate regularity and fitting. Let $\widehat{\mathcal{M}}$ be a fitted or analytical surrogate with the same spaces, discount, and initial law, bounded rewards, and state–action backup $\widehat{\mathcal{B}}$. Conditional laws, backups, policies, and values have fixed measurable definitions at all coordinates used below; extensions beyond observed support are part of the chosen model. For $\pi \in \{\widehat{\pi}, \pi_c\}$, its actual policy value $\widehat{V}^\pi = V^\pi_{\widehat{\mathcal{M}}}$ satisfies

$$\|\widehat{V}^\pi\|_\infty \le V_{\max}, \qquad \mathrm{Lip}(\widehat{V}^\pi) \le L_V.$$

The bounds also apply on the extensions used in the backups. Define

$$\varepsilon_{\mathrm{fit}}(\pi) = \mathbb{E}_{d_{\lambda,\theta}}\left|(\widehat{\mathcal{B}} - \mathcal{B}_{\lambda,\theta})\widehat{V}^\pi(S,A)\right|. \tag{A.9}$$

(R4) Occupancy domination. For both policies, let

$$d^\pi_B = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^\pi_B((S_t, A_t) \in \cdot), \qquad d^\pi_B \le C_\pi d_{\lambda,\theta} \tag{A.10}$$

for a finite C_π. All displayed error expectations are finite. This is a population coverage condition; in continuous spaces a finite atomic empirical law cannot generally replace $d_{\lambda,\theta}$.

Continuing error budget. Define source discrepancies in the decoded coordinates and reward units used by the policy learner:

$$\varepsilon_{A,r} = \mathbb{E}_{\eta_A}|R - r_B(S,A)|, \qquad \varepsilon_{A,f} = \mathbb{E}_{\eta_A}\|S' - f_B(S,A)\|.$$

Write

$$K_{\mathrm{syn}} = \sqrt{1 + (\gamma L_V)^2 + (L_r + \gamma L_V L_f)^2}, \tag{A.11}$$

from the reward, next-state, and state–action blocks, and set

$$\mathcal{E}_{\mathrm{data}} = (1-\lambda)(\varepsilon_{A,r} + \gamma L_V \varepsilon_{A,f}) + \lambda\theta K_{\mathrm{syn}}\frac{\varepsilon_{\mathrm{syn}}}{\sqrt{p_{\mathrm{keep}}}}, \qquad \mathcal{R}_{\lambda,\theta}(\pi) = \frac{C_\pi}{1-\gamma}[\varepsilon_{\mathrm{fit}}(\pi) + \mathcal{E}_{\mathrm{data}}]. \tag{A.12}$$

Here $\varepsilon_{\mathrm{syn}}$ is exactly equation 1. Source immediate-reward error has no factor γ; it vanishes under shared state–action rewards.

A.4.1 Episodic definitions

For terminating tasks keep (R1)–(R4), with the following replacements. Let $\tau_B(s') \in \{0, 1\}$ be the measurable target termination function, $\mathrm{term}_B(s,a) = \tau_B(f_B(s,a))$, and $\mathcal{T}_B = \{s' : \tau_B(s') = 1\}$. Genuine transitions have their own context's correct flag d; synthetic transitions use $d = \tau_B(s')$. The initial law is supported on target nonterminal states and the final transition's reward is included. The boundary of $\mathcal{T}_B$ is in the ambient Euclidean state space.

Attach the flags to the same mixture laws in equation A.7. Replace the training backup by

$$(\mathcal{B}_{\lambda,\theta}V)(s,a) = \mathbb{E}_{\eta_{\lambda,\theta}}[R + \gamma(1-d)V(S') \mid S = s, A = a].$$

The surrogate includes its own continuation flag, with (R3) still required. Hard termination does not need to preserve value regularity; Appendix A.5 supplies a concrete sufficient surrogate class. If T is the target termination time, use the survival occupancy

$$d_B^\pi = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr_B^\pi((S_t, A_t) \in \cdot,\ t < T)$$

in (R4), and use flagged backups in $\varepsilon_{\mathrm{fit}}$.

Put $\varepsilon_{A,d} = \mathbb{E}_{\eta_A}|d - \mathrm{term}_B(S,A)|$ and $b_B(t) = \eta_B^1(\mathrm{dist}(S', \partial\mathcal{T}_B) \le t)$. For $\varepsilon > 0$ and $p > 0$, define

$$\Omega_B(\varepsilon, p) = \min\left\{1, \frac{1}{p}\inf_{t>0}\left[b_B(t) + \frac{\max\{1, L_f^2\}\varepsilon^2}{t^2}\right]\right\}. \tag{A.13}$$

Set $\Omega_B(0,p) = 0$; distance to an empty boundary is $+\infty$. For fixed p, the positive-error limit is $\min\{1, \eta_B^1(S' \in \partial\mathcal{T}_B)/p\}$, so it tends to zero when the boundary has zero target-high mass. Exact partners still have identical flags at zero error, including on the boundary.

In the episodic version, add

$$\mathcal{E}_{\mathrm{term}} = \gamma V_{\max}\left[(1-\lambda)\varepsilon_{A,d} + \lambda\theta\Omega_B(\varepsilon_{\mathrm{syn}}, p_{\mathrm{keep}})\right] \tag{A.14}$$

to $\mathcal{E}_{\mathrm{data}}$ in equation A.12, and use the resulting $\mathcal{R}_{\lambda,\theta}$. For continuing tasks all flags and $\mathcal{E}_{\mathrm{term}}$ are zero and $T = \infty$.

Theorem A.1 (Synthesis-to-return comparison)**.** *Assume (R1)–(R4), using the continuing or episodic definitions, and the decoded coupling bound in Theorem 1. Then for $\pi \in \{\widehat{\pi}, \pi_c\}$,*

$$|J_B(\pi) - J_{\widehat{\mathcal{M}}}(\pi)| \le \mathcal{R}_{\lambda,\theta}(\pi). \tag{A.15}$$

If $J_{\widehat{\mathcal{M}}}(\widehat{\pi}) \ge J_{\widehat{\mathcal{M}}}(\pi_c) - \varepsilon_{\mathrm{learn}}$ for $\varepsilon_{\mathrm{learn}} \ge 0$, then

$$J_B(\widehat{\pi}) \ge J_B(\pi_c) - \varepsilon_{\mathrm{learn}} - \mathcal{R}_{\lambda,\theta}(\widehat{\pi}) - \mathcal{R}_{\lambda,\theta}(\pi_c). \tag{A.16}$$

The result holds for each realized surrogate and policies satisfying these conditions; it requires neither comparator optimality nor nested policy classes.

Subtracting any baseline return $J_B(\pi_0)$ gives the corresponding improvement comparison without a coverage assumption on π_0. Strict gain follows only if the comparator's advantage over that baseline exceeds the error budget. No positive comparator advantage, IQL-specific rate, or data-driven certificate of these population quantities is supplied by the theorem.

Lemma A.2 (Uniform training-law consistency)**.** *Under (R1)–(R2) and the decoded coupling bound, every V with $\|V\|_\infty \le V_{\max}$ and $\mathrm{Lip}(V) \le L_V$ satisfies*

$$\mathbb{E}_{\eta_{\lambda,\theta}}|b_V| \le \mathcal{E}_{\mathrm{data}}, \quad b_V(\xi, d) = r + \gamma(1-d)V(s') - (\mathcal{B}_B V)(s,a),$$

where $(\mathcal{B}_B V)(s,a) = r_B(s,a) + \gamma(1 - \mathrm{term}_B(s,a))V(f_B(s,a))$. Use zero flags and the continuing budget in continuing tasks.

Proof. For any flagged transition, adding and subtracting $(1-d)V(f_B(s,a))$ gives

$$|b_V(\xi,d)| \le |r - r_B(s,a)| + \gamma L_V\|s' - f_B(s,a)\| + \gamma V_{\max}|d - \mathrm{term}_B(s,a)|. \tag{A.17}$$

Couple $\widehat{\xi} \sim \nu$ to $\xi^\star \sim \eta_B^1$ as in Theorem 1, and put $Z = \|\widehat{\xi} - \xi^\star\|$, so $\mathbb{E}Z^2 \le \varepsilon_{\mathrm{syn}}^2$. The phantom is target-consistent. Lipschitz continuity and Cauchy–Schwarz on the reward, next-state, and state–action blocks bound the first two terms by $K_{\mathrm{syn}}Z$. Conditioning on acceptance gives mean at most

$$K_{\mathrm{syn}}\frac{\mathbb{E}[Z\mathbf{1}_K]}{p_{\mathrm{keep}}} \le K_{\mathrm{syn}}\frac{\varepsilon_{\mathrm{syn}}}{\sqrt{p_{\mathrm{keep}}}}.$$

For termination put $M_\star = \mathrm{dist}(s^{\star\prime}, \partial\mathcal{T}_B)$ and $a = \max(1, L_f)$. Both $\widehat{s}'$ and $f_B(\widehat{s},\widehat{a})$ lie within aZ of $s^{\star\prime}$. If $M_\star > aZ$, that ball avoids the boundary, so their flags agree. Splitting on $Z > t/a$ and using Markov's inequality gives, for each $t > 0$,

$$p_{\mathrm{keep}}\mathbb{E}_{\nu_K}|d - \mathrm{term}_B(S,A)| \le b_B(t) + a^2\varepsilon_{\mathrm{syn}}^2/t^2.$$

Taking the infimum and capping the conditional probability at one yields equation A.13 for positive error. For zero error the partners coincide almost surely, so flags agree even on the boundary. Thus

$$\mathbb{E}_{\nu_K}|b_V| \le K_{\mathrm{syn}}\frac{\varepsilon_{\mathrm{syn}}}{\sqrt{p_{\mathrm{keep}}}} + \gamma V_{\max}\Omega_B(\varepsilon_{\mathrm{syn}}, p_{\mathrm{keep}}).$$

On target data $b_V = 0$; on source data equation A.17 gives $\mathbb{E}_{\eta_A}|b_V| \le \varepsilon_{A,r} + \gamma L_V \varepsilon_{A,f} + \gamma V_{\max}\varepsilon_{A,d}$. Expanding the mixture proves the claim. All flag terms vanish in the continuing case. This bounds retained-sample consistency, not the filtered law's Wasserstein distance. □

Proof of Theorem A.1. Covered Bellman error. For either policy write $e_\pi = (\widehat{\mathcal{B}} - \mathcal{B}_B)\widehat{V}^\pi$. Conditional Jensen, Lemma A.2, the fitting definition, and occupancy domination give

$$\mathbb{E}_{d_B^\pi}|e_\pi| \le C_\pi \mathbb{E}_{d_{\lambda,\theta}}|e_\pi| \le C_\pi[\varepsilon_{\mathrm{fit}}(\pi) + \mathcal{E}_{\mathrm{data}}]. \tag{A.18}$$

The consistency estimate is uniform over bounded Lipschitz V, so no independence of the learned surrogate is needed.

Trajectory telescoping. We use the same-policy Bellman-residual identity (e.g., Xie & Jiang, 2020, Lemma 1), retaining survival indicators for episodic tasks. Let $I_t = \mathbf{1}_{\{t<T\}}$ along a target trajectory, extended arbitrarily after termination, and let $r_t = r_B(S_t, A_t)$. Then $I_t(1-\mathrm{term}_B(S_t, A_t)) = I_{t+1}$. The surrogate policy Bellman equation, averaged over the action at each live state, yields

$$\sum_{t=0}^{n-1}\gamma^t \mathbb{E}_B^\pi[I_t e_\pi(S_t, A_t)] = J_{\widehat{\mathcal{M}}}(\pi) - \mathbb{E}_B^\pi \sum_{t=0}^{n-1}\gamma^t I_t r_t - \gamma^n \mathbb{E}_B^\pi[I_n \widehat{V}^\pi(S_n)].$$

The final term is bounded by $\gamma^n V_{\max}$. Bounded rewards and values allow $n \to \infty$, giving

$$J_{\widehat{\mathcal{M}}}(\pi) - J_B(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d_B^\pi} e_\pi.$$

Together with equation A.18, this proves equation A.15. The learned-policy condition then gives

$$J_B(\widehat{\pi}) \ge J_{\widehat{\mathcal{M}}}(\widehat{\pi}) - \mathcal{R}_{\lambda,\theta}(\widehat{\pi}) \ge J_B(\pi_c) - \varepsilon_{\mathrm{learn}} - \mathcal{R}_{\lambda,\theta}(\widehat{\pi}) - \mathcal{R}_{\lambda,\theta}(\pi_c),$$

which is equation A.16. Setting $T = \infty$ and all flags to zero proves the continuing case with the same argument. □

Interpretation. Filtering and termination enter through the displayed consistency penalties, while source reward and dynamics mismatch can remain even with perfect synthesis. The theorem leaves the source-data effects on coverage, fitting, and optimization explicit, so it does not select an optimal λ. The quantities $\varepsilon_{\mathrm{fit}}$ and $\varepsilon_{\mathrm{learn}}$ are theoretical error terms rather than observed training losses.

A.5 A REGULAR ANALYTICAL SURROGATE

Hard continuation flags can make exact surrogate values discontinuous even at nonterminal states. The correction Ω_B controls flag disagreement, not those value jumps. The following construction verifies (R3) for a common analytical reference. It is a finite-outcome model of the kernel-interpolation type (Ormoneit & Sen, 2002).

Fix outcomes (y_i, r_i, d_i), $i = 1, \dots, n_*$, with $|r_i| \le R_*$, $R_* > 0$, $d_i \in \{0, 1\}$, and y_i in the state domain, including required extensions. Choose nonnegative smooth weights with

$$\sum_i w_i(u) = 1, \qquad \sum_i |w_i(u) - w_i(v)| \le L_w\|u - v\|, \quad L_w < \infty.$$

At (s, a), select i with probability $w_i(s, a)$, receive r_i, and stop if $d_i = 1$; otherwise continue at y_i. The resulting stochastic surrogate has the same discount and initial law, and

$$(\widehat{\mathcal{B}}V)(s, a) = \sum_i w_i(s, a)[r_i + \gamma(1 - d_i)V(y_i)].$$

Stopping uses the sampled flag, not an additional hard stopping rule imposed by the target terminal set. Both continuation probability and the weighted continuation outcomes vary smoothly.

Lemma A.3 (Value regularity of the interpolated surrogate)**.** *For any deterministic policy with* $\mathrm{Lip}(\pi) \le L_\pi$ *on this domain, the exact surrogate value satisfies*

$$\|\widehat{V}^\pi\|_\infty \le V_* := \frac{R_*}{1-\gamma}, \qquad \mathrm{Lip}(\widehat{V}^\pi) \le V_* L_w \sqrt{1+L_\pi^2}.$$

No contraction of physical dynamics in the Euclidean state metric is assumed.

Proof. The policy Bellman operator is a sup-norm contraction with modulus at most γ, and bounded rewards give V_*. At its fixed point, $\widehat{V}^\pi(s) = \sum_i w_i(s,\pi(s)) b_i^\pi$, where $b_i^\pi = r_i + \gamma(1-d_i)\widehat{V}^\pi(y_i)$ is independent of the current state and $|b_i^\pi| \le R_* + \gamma V_* = V_*$. Therefore

$$|\widehat{V}^\pi(s) - \widehat{V}^\pi(t)| \le V_* \sum_i |w_i(s,\pi(s)) - w_i(t,\pi(t))| \le V_* L_w \sqrt{1+L_\pi^2}\|s-t\|.$$

□

For deterministic Lipschitz $\widehat{\pi}, \pi_c$, one common surrogate satisfies (R3) with $V_{\max} = V_*$ and $L_V = V_* L_w \max_{\pi\in\{\widehat{\pi},\pi_c\}} \sqrt{1+L_\pi^2}$. The result holds for every realized outcome set and weight function satisfying these conditions, including fitted choices.

Fitting cost and scope. For any measurable conditional coupling Γ of training outcomes (R, S', d) and surrogate outcomes $(\widetilde{R}, \widetilde{S}', \widetilde{d})$ at the same (S, A), with expectation also over $d_{\lambda,\theta}$, conditional Jensen and the add-and-subtract argument in equation A.17 give

$$\varepsilon_{\mathrm{fit}}(\pi) \le \mathbb{E}_\Gamma \left[|R - \widetilde{R}| + \gamma L_V \|S' - \widetilde{S}'\| + \gamma V_{\max} |d - \widetilde{d}| \right].$$

Smoothing discrepancy enters $\varepsilon_{\mathrm{fit}}$. The target-flag and filtering penalties, including Ω_B, still apply. Steeper weights can enlarge L_V, while broader averaging can enlarge fitting error; neither is asserted small. Coverage and the learning comparison must hold for this same surrogate. This construction gives a sufficient condition for value regularity for the surrogate class considered here.

B Environment Setting

We construct the same environment settings as the ODRL benchmark (Lyu et al., 2024a) and use the ODRL offline datasets directly. This matches the benchmark protocol used by recent cross-domain offline RL methods, including OTDF, STC, and DmC (Lyu et al., 2025; Yan et al., 2026; Van et al., 2025b), allowing for comparison under the same settings.

B.1 Source and Target Datasets

We consider four MuJoCo locomotion tasks: Ant, HalfCheetah, Hopper, and Walker2d. The source datasets are the corresponding D4RL datasets distributed with ODRL. We consider medium (`med`), medium-replay (`med-r`), and medium-expert (`med-e`) source datasets. The full source dataset is used in each experiment and contains approximately one million transitions, depending on the D4RL dataset (Fu et al., 2020).

For the target domain, we use the limited offline datasets provided by ODRL. ODRL provides medium and expert target datasets containing $5,000$ transitions. To construct the medium-expert dataset, we follow the OTDF and STC protocol of combining two whole trajectories from the medium dataset with three from the expert dataset, to get a total of $5{,}000$ transitions. We use the provided target datasets directly rather than recollecting trajectories. For each experimental setting, the same target dataset is held fixed across all five policy training seeds.

B.2 Target Dynamics Shifts

The source environment uses the default MuJoCo dynamics. The target environment modifies these dynamics according to the ODRL benchmark. The descriptions below follow its released XML assets (Lyu et al., 2024b). We study gravity, friction, and morphology shifts.

Gravity shift. For the gravity shift, the magnitude of the gravitational acceleration is reduced by a factor of 0.5 while preserving its direction. In the MuJoCo XML, this corresponds to changing

```
<option gravity="0 0 -9.81" ... />
```

to

```
<option gravity="0 0 -4.905" ... />
```

for the target environment.

Friction shift. ODRL halves explicitly specified geometry friction values. In Hopper, the torso, thigh, and leg entries change from `friction="0.9"` to `friction="0.45"`, and the foot entry changes from 2.0 to 1.0. The other tasks halve their explicit tuples and geometry-specific overrides; omitted coefficients retain inherited defaults.

Morphology shift. For the morphology shift, we use ODRL's medium-level change to one body part of each robot. Unlike the gravity and friction shifts, this alters model geometry rather than gravity or friction coefficients:

- **HalfCheetah (thigh):** the back and front thigh capsules are shortened from approximate half-lengths 0.145 and 0.133 to 0.057 and 0.053, with the shin and foot bodies repositioned accordingly.
- **Hopper (foot):** the foot capsule shrinks from radius 0.06 and half-length 0.195 to 0.036 and 0.117, and the foot body moves downward by 0.10.
- **Walker2d (leg):** on both legs, the thigh capsule lengthens from half-length 0.225 to 0.35 and the shin shortens from 0.25 to 0.125, with the foot bodies moved downward by 0.10.
- **Ant (legs):** the four ankle capsules are shortened from approximate half-length 0.283 to 0.141 by halving their endpoint offsets; the ankle-body origins are unchanged.

For example, Hopper's global-coordinate XML changes the foot capsule from

```
<geom name="foot_geom" size="0.06"
      fromto="-0.13 0 0.1 0.26 0 0.1" ... />
```

to

```
<geom name="foot_geom" size="0.036"
      fromto="-0.078 0 0.1 0.156 0 0.1" ... />
```

in the target environment.

B.3 Evaluation

Policies are evaluated only in the corresponding target environment. Following ODRL (Lyu et al., 2024a;b), the evaluation statistic G is the mean undiscounted episodic return, distinct from the discounted $J_{\mathcal{M}}$ used in the theory. The D4RL-style normalized score is

$$\mathrm{NS} = 100 \cdot \frac{G - G_{\mathrm{random}}}{G_{\mathrm{expert}} - G_{\mathrm{random}}},$$

where G_{random} and G_{expert} are the benchmark reference scores for the corresponding task and shift. Unless otherwise stated, reported results use the final policy checkpoint and average over five policy seeds. The target dataset is fixed across those seeds, so the reported variation is over policy training rather than independently sampled target datasets.

Table 3: IQL reference results across eight medium-to-medium gravity and morphology settings. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

		IQL (reported by STC)		IQL (ours)		
Shift	Environment	$\mathcal{D}_A \cup \mathcal{D}_B$	$\mathcal{D}_B$	$\mathcal{D}_A \cup \mathcal{D}_B$	$\mathcal{D}_B$	**RIFT**
Gravity	Ant	10.2 ± 1.8	32.0 ± 7.6	10.1 ± 1.4	29.8 ± 4.9	$\mathbf{40.8 \pm 10.3}$
	HalfCheetah	39.6 ± 3.3	0.6 ± 0.5	42.0 ± 2.3	2.6 ± 0.6	$\mathbf{47.3 \pm 6.6}$
	Hopper	11.2 ± 1.1	46.2 ± 3.7	14.1 ± 4.8	52.7 ± 12.5	$\mathbf{60.4 \pm 7.3}$
	Walker2d	28.1 ± 12.9	28.5 ± 7.4	30.6 ± 4.3	26.7 ± 4.5	$\mathbf{46.5 \pm 5.5}$
Morphology	Ant	71.4 ± 2.4	76.8 ± 2.8	65.5 ± 2.0	$\mathbf{76.9 \pm 0.3}$	76.2 ± 2.4
	HalfCheetah	24.5 ± 2.4	12.3 ± 3.7	22.8 ± 2.4	19.0 ± 6.6	$\mathbf{57.9 \pm 4.4}$
	Hopper	15.9 ± 6.8	42.5 ± 17.3	18.1 ± 7.1	55.5 ± 6.4	$\mathbf{69.4 \pm 13.2}$
	Walker2d	31.5 ± 8.6	43.1 ± 11.5	30.8 ± 9.2	49.2 ± 5.1	$\mathbf{52.8 \pm 8.2}$
Total (displayed means)		232.4	282.0	234.0	312.4	**451.3**

C Algorithmic Implementation

We follow the implementation settings used in the ODRL benchmark along with their corresponding baseline configurations. We first verify our IQL implementation against the reference results reported by STC, and then list the hyperparameters used for all methods.

C.1 IQL Replication

Table 3 compares our IQL implementation with the IQL results reported by STC. We reproduce both the pooled source–target setting and the target-only setting on the eight medium-to-medium gravity and morphology tasks. The pooled aggregate is close to STC (234.0 vs. 232.4), while the target-only aggregate is higher (312.4 vs. 282.0); individual task scores differ.

C.2 Hyperparameter Setup

We include the baseline algorithms and hyperparameter settings as reported in STC (Yan et al., 2026) and DmC (Van et al., 2025b).

IQL. IQL (Kostrikov et al., 2022) learns a value function using expectile regression,

$$\mathcal{L}_V = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau(Q(s,a) - V(s))\right],$$

where $L_2^\tau(u) = |\tau - \mathbf{1}_{\{u<0\}}|u^2$. Its Q-function uses temporal-difference regression,

$$\mathcal{L}_Q = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\left[\left(r + \gamma(1-d)V(s') - Q(s,a)\right)^2\right].$$

Here d denotes environmental termination in the displayed discounted backup, not time-limit truncation by itself. The policy is trained by advantage-weighted behavior cloning,

$$\mathcal{L}_\pi = -\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp(\beta_{\mathrm{IQL}}(Q(s,a) - V(s)))\log\pi(a \mid s)\right].$$

DARA. DARA (Liu et al., 2022) trains domain classifiers to estimate the dynamics discrepancy between source and target transitions. The resulting discrepancy Δr is used to modify the source reward as

$$\hat{r} = r - \alpha\Delta r,$$

with the reward penalty clipped to $[-10, 10]$. The corrected source transitions are then used for downstream IQL training.

BOSA. BOSA (Liu et al., 2024) models target-domain support using a CVAE and restricts value learning to transitions supported by the estimated target dynamics. Its critic objective takes the form

$$\mathcal{L}_{\mathrm{BOSA}} = \mathbb{E}_{\mathcal{D}_{\mathrm{mix}}}\left[\delta(Q)\,\mathbf{1}\left(\log\hat{P}_{\mathrm{tar}}(s' \mid s,a) > \epsilon'_{\mathrm{th}}\right)\right] + \mathbb{E}_{(s,a)\sim\mathcal{D}_A}[Q(s,a)],$$

where $\delta(Q)$ denotes squared TD error, and ϵ'_{th} is a log-likelihood threshold. The dynamics ensemble uses its minimum log-likelihood for this test. BOSA additionally constrains policy updates to behavior support in the pooled data (Lyu et al., 2024b).

IGDF. IGDF (Wen et al., 2024) learns a contrastive representation that assigns each transition a score $h(s, a, s')$ reflecting source–target dynamics mismatch. Target samples are always used for value learning, while source samples are selected according to this score:

$$\mathcal{L}_Q = \frac{1}{2}\mathbb{E}_{\mathcal{D}_B}\left[(Q - \mathcal{T}Q)^2\right] + \frac{\alpha}{2}\mathbb{E}_{\mathcal{D}_A}\left[\mathbf{1}\big(h > h_{\xi\%}\big)\,(Q - \mathcal{T}Q)^2\right].$$

The threshold $h_{\xi\%}$ determines the retained fraction of source transitions.

SRPO. SRPO (Xue et al., 2023) estimates stationary-state information using a discriminator trained from high- and low-value transitions. The discriminator is then used to modify the reward,

$$\hat{r}_{\text{SRPO}} = r + \lambda \frac{D_\delta(s)}{1 - D_\delta(s)},$$

and policy learning proceeds using the modified rewards.

OTDF. OTDF (Lyu et al., 2025) uses optimal transport to measure the dynamics discrepancy between source and target transitions. Given the optimal coupling $\mu^\star$, it assigns each source transition u_t the deviation score

$$d(u_t) = -\sum_{t'} C(u_t, u_{t'})\,\mu^\star_{t,t'}.$$

In its practical implementation, OTDF computes the coupling using entropy-regularized optimal transport with the Sinkhorn algorithm and normalizes the resulting deviation scores. Source transitions are then selected according to the top-ξ scores and weighted in the TD objective according to their target-domain compatibility. OTDF additionally trains a CVAE on the target dataset and uses it to regularize the learned policy toward the support of the target behavior policy.

STC. STC (Yan et al., 2026) mitigates the dynamics gap by correcting source transitions toward the target domain. It first pretrains an inverse policy model, forward dynamics model, and reward model on the target dataset for $50{,}000$ steps. For each source transition, STC corrects the action using the inverse policy model and estimates the corrected reward using a first-order Taylor approximation. The corrected transition is retained only when its estimated target-domain dynamics discrepancy satisfies the correction criterion controlled by threshold λ. The value function is then trained with TD regression, while the policy uses a Q-weighted behavior-cloning objective with coefficient β.

DmC. DmC (Van et al., 2025b) estimates source–target dynamics proximity using a k-nearest-neighbor score over transition tuples (s, a, s'). The resulting score is used both to select and weight source transitions during value learning and to condition a diffusion model that generates additional source samples closer to the target domain. DmC further applies a target-policy regularization term through a conditional VAE to reduce bias toward the source dataset. We include the DmC comparison in Appendix D.4.

IQL*. IQL* uses the same IQL implementation and hyperparameters as above, but follows the same data-mixing protocol used by RIFT, modifying the batch target-to-source ratio.

RIFT. We train a shared VAE using balanced source–target minibatches and add the additional reward-conditioned regularization from Section 4.1 to the loss:

$$\mathcal{L}_{\text{class}} = \eta\,\mathbb{E}_{c\in\{A,B\}}\mathbb{E}_{y\in\{0,1\}}\mathbb{E}_{x\sim\mathcal{D}_c^y}\left[\mathrm{KL}\big(q_\vartheta(z_0 \mid x)\,\|\,\mathcal{N}\big((2y-1)c_q, s_q^2\big)\big)\right].$$

Here q_ϑ is the encoder posterior and η weights the class penalty. We use mini-batch optimal transport with squared Euclidean cost to pair low- and high-reward source samples. Their joint latent–reward displacements are fit with an affine model using ordinary least squares; matching itself uses only squared latent distance. The learned field is then applied to low-reward target anchors to generate synthetic transitions. We similarly modify the batch target-to-source ratio.

Specifications. We summarize the hyperparameter settings for all methods in Table 4. For IQL, DARA, and BOSA, we use the fixed configurations reported by STC (Yan et al., 2026). For SRPO, the reward coefficient is selected from $\{0.1, 0.3\}$. For IGDF, the data selection ratio is fixed to 75%, while the representation dimension is selected from $\{16, 64\}$. For OTDF, we follow the reported implementation settings, using a fixed data selection ratio of 80% and a task-dependent policy coefficient in $\{0.1, 0.5\}$. For STC, the reward gradient coefficient is fixed to 0.5, while the correction threshold is selected from $\{1.0, 5.0\}$ and the Q-weighted loss coefficient from $\{0.5, 5.0\}$, with the best configuration reported for each environment. For RIFT, we fix the representation and field-fitting hyperparameters across tasks and evaluate two downstream training configurations, $\lambda \in \{0.5, 1\}$.

Table 4: Reported hyperparameter settings: shared IQL settings and method-specific configurations.

Method	Hyperparameter	Value
Shared IQL backbone		
	Actor network	$(256, 256)$
	Critic network	$(256, 256)$
	Learning rate	3×10^{-4}
	Optimizer	Adam
	Discount factor	0.99
	Nonlinearity	ReLU
	Target update rate	5×10^{-3}
	Source-domain batch size	128
	Target-domain batch size	128
	Temperature coefficient	0.2
	Maximum log std	2
	Minimum log std	-20
	Inverse temperature β_{IQL}	3.0
	Expectile τ	0.7
DARA	Classifier network	$(256, 256)$
	Reward penalty coefficient	0.1
BOSA	Policy regularization coefficient	0.1
	Transition coefficient	0.1
	Threshold parameters ϵ, ϵ'	$\log(0.01)$
	Value weight	0.1
	CVAE ensemble size	5
SRPO	Discriminator network	$(256, 256)$
	Data selection ratio	0.5
	Reward coefficient	$\{0.1, 0.3\}$
IGDF	Representation dimension	$\{16, 64\}$
	Contrastive encoder network	$(256, 256)$
	Encoder pretraining steps	$7{,}000$
	Importance coefficient	1.0
	Data selection ratio	75%
OTDF	CVAE training steps	$10{,}000$
	CVAE learning rate	10^{-3}
	Number of sampled latent variables	10
	Gaussian standard deviation	$\sqrt{0.1}$
	Cost function	cosine
	Data selection ratio	80%
	Policy coefficient	$\{0.1, 0.5\}$
STC	Correction threshold λ	$\{1.0, 5.0\}$
	Reward gradient coefficient α	0.5
	Q-weighted loss coefficient β	$\{0.5, 5.0\}$
DmC	k-nearest neighbor	5
	Data selection ratio	50%
	Policy regularization coefficient	0.1
	Guided-sampling score threshold	90%
IQL*	Data-mixing coefficient λ	$\{0.5, 1\}$
RIFT	Reward quantile q	0.2
	VAE latent dimension	16
	VAE hidden layers	$(256, 256)$
	VAE learning rate / batch / steps	10^{-3} / 256 / 100,000
	OT mini-batch size	2,000 per side, 8 problems
	Data-mixing coefficient λ	$\{0.5, 1\}$

D Additional Results

We provide the detailed tables summarized in the main text and additional results.

D.1 Target-Quality Results

Table 5 gives the complete target-quality comparison summarized in Section 5.2.

Table 5: Performance under gravity and morphology shifts when the source dataset is fixed to medium quality and target-data quality varies. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

Type	Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	STC	**RIFT (ours)**
Gravity	half-med	med	39.6 ± 3.3	41.2 ± 3.9	38.9 ± 4.0	36.9 ± 4.5	36.6 ± 5.5	40.7 ± 7.7	42.4 ± 5.3	$\mathbf{47.3 \pm 6.6}$
	half-med	med-e	39.6 ± 3.7	40.7 ± 2.8	40.4 ± 3.0	40.7 ± 2.3	38.7 ± 6.2	28.6 ± 3.2	37.3 ± 1.5	$\mathbf{46.2 \pm 3.0}$
	half-med	expert	42.4 ± 3.8	39.8 ± 4.4	40.5 ± 3.9	39.4 ± 1.6	39.6 ± 4.6	36.1 ± 5.3	43.1 ± 5.7	$\mathbf{45.5 \pm 2.6}$
	hopp-med	med	11.2 ± 1.1	17.3 ± 3.8	15.2 ± 3.3	12.4 ± 1.0	15.3 ± 3.5	32.4 ± 8.0	43.4 ± 6.1	$\mathbf{60.4 \pm 7.3}$
	hopp-med	med-e	14.7 ± 3.6	15.4 ± 2.5	21.1 ± 9.3	14.2 ± 1.8	15.1 ± 3.6	24.2 ± 3.6	23.4 ± 6.4	$\mathbf{30.5 \pm 17.2}$
	hopp-med	expert	12.5 ± 1.6	19.3 ± 10.5	12.7 ± 1.7	11.8 ± 0.9	14.8 ± 4.0	33.7 ± 7.8	37.1 ± 14.1	$\mathbf{48.2 \pm 20.8}$
	walk-med	med	28.1 ± 12.9	28.4 ± 13.7	38.0 ± 11.2	21.4 ± 7.0	22.1 ± 8.4	36.6 ± 2.3	41.6 ± 4.0	$\mathbf{46.5 \pm 5.5}$
	walk-med	med-e	35.7 ± 4.7	30.7 ± 9.7	40.9 ± 7.2	34.0 ± 9.9	35.4 ± 9.1	44.8 ± 7.5	36.9 ± 6.5	$\mathbf{48.8 \pm 8.5}$
	walk-med	expert	37.3 ± 8.0	36.0 ± 7.0	41.3 ± 8.6	39.5 ± 3.8	36.2 ± 13.6	44.0 ± 4.0	$\mathbf{49.9 \pm 10.1}$	42.6 ± 7.6
	ant-med	med	10.2 ± 1.8	9.4 ± 0.9	12.4 ± 2.0	11.7 ± 1.0	11.3 ± 1.3	$\mathbf{45.1 \pm 12.4}$	42.6 ± 8.4	40.8 ± 10.3
	ant-med	med-e	9.4 ± 1.2	10.0 ± 0.9	11.6 ± 1.3	10.2 ± 1.2	9.4 ± 1.4	$\mathbf{33.9 \pm 5.4}$	23.2 ± 6.1	25.0 ± 11.7
	ant-med	expert	10.2 ± 0.3	9.8 ± 0.6	11.8 ± 0.4	9.5 ± 0.6	9.7 ± 1.6	33.2 ± 9.0	$\mathbf{35.6 \pm 5.6}$	15.9 ± 4.7
Morph	half-med	med	24.5 ± 2.4	21.0 ± 3.9	24.2 ± 5.6	18.1 ± 1.8	23.7 ± 3.4	21.1 ± 7.6	19.5 ± 2.2	$\mathbf{57.9 \pm 4.4}$
	half-med	expert	11.1 ± 1.5	11.3 ± 0.6	9.1 ± 1.7	10.1 ± 1.6	10.1 ± 0.8	7.4 ± 2.3	7.7 ± 1.2	$\mathbf{12.3 \pm 1.6}$
	hopp-med	med	15.9 ± 6.8	17.8 ± 10.1	12.8 ± 0.1	21.7 ± 7.7	25.3 ± 9.7	16.4 ± 7.1	43.1 ± 23.9	$\mathbf{69.4 \pm 13.2}$
	hopp-med	expert	15.5 ± 4.1	13.0 ± 1.4	4.5 ± 4.0	10.4 ± 3.1	13.4 ± 2.0	14.0 ± 4.0	53.2 ± 50.3	$\mathbf{62.4 \pm 40.9}$
	walk-med	med	31.5 ± 8.6	35.0 ± 10.8	26.7 ± 6.6	38.6 ± 5.1	38.5 ± 8.4	42.5 ± 3.1	$\mathbf{56.7 \pm 8.1}$	52.8 ± 8.2
	walk-med	expert	37.0 ± 6.0	43.7 ± 5.5	31.3 ± 8.6	43.8 ± 8.4	38.5 ± 4.3	$\mathbf{49.9 \pm 5.4}$	32.9 ± 7.3	33.2 ± 18.4
	ant-med	med	71.4 ± 2.4	71.5 ± 6.8	54.8 ± 13.2	72.8 ± 2.2	71.8 ± 2.7	75.1 ± 2.3	$\mathbf{77.2 \pm 2.8}$	76.2 ± 2.4
	ant-med	expert	63.9 ± 8.1	68.9 ± 7.6	47.5 ± 11.9	59.0 ± 4.7	66.1 ± 4.0	63.8 ± 6.4	74.1 ± 21.2	$\mathbf{79.0 \pm 12.9}$
Total (displayed means)			561.7	580.2	535.7	556.2	571.6	723.5	820.9	**940.9**

D.2 Detailed Ablations and Reconstruction

Tables 6 and 7 report the detailed quantile and VAE-design results. Table 8 gives the reconstruction diagnostic discussed in Section 5.4.

Table 6: Comparison of reward quantiles across eight medium-to-medium gravity and morphology settings. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

		Reward quantile q					
Shift	Task	0 (no augmentation)	0.1	0.2 (ours)	0.3	0.4	0.5
Gravity	Ant	29.8 ± 4.9	34.5 ± 11.9	$\mathbf{40.8 \pm 10.3}$	36.8 ± 14.1	37.8 ± 4.4	31.5 ± 6.9
	HalfCheetah	42.0 ± 2.3	49.4 ± 2.7	47.3 ± 6.6	49.5 ± 3.6	48.3 ± 3.9	$\mathbf{51.4 \pm 2.6}$
	Hopper	52.7 ± 12.5	$\mathbf{62.0 \pm 14.0}$	60.4 ± 7.3	60.2 ± 8.6	58.7 ± 6.5	41.9 ± 12.9
	Walker2d	30.6 ± 4.3	34.7 ± 3.2	$\mathbf{46.5 \pm 5.5}$	42.4 ± 7.9	41.4 ± 10.0	40.8 ± 10.9
Morphology	Ant	76.9 ± 0.3	77.1 ± 0.2	76.2 ± 2.4	$\mathbf{78.0 \pm 0.3}$	$\mathbf{78.0 \pm 0.1}$	76.8 ± 1.3
	HalfCheetah	22.8 ± 2.4	54.5 ± 5.0	57.9 ± 4.4	$\mathbf{58.0 \pm 3.8}$	51.9 ± 2.3	50.9 ± 6.1
	Hopper	55.5 ± 6.4	67.7 ± 8.2	$\mathbf{69.4 \pm 13.2}$	53.3 ± 16.4	65.7 ± 14.6	66.3 ± 10.9
	Walker2d	49.2 ± 5.1	42.9 ± 15.7	$\mathbf{52.8 \pm 8.2}$	47.3 ± 18.6	51.3 ± 8.6	47.1 ± 4.6
Sum		359.5	422.8	**451.3**	425.5	433.1	406.7

Table 7: Comparison of VAE design across eight medium-to-medium gravity and morphology settings. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

		Shared VAE		
Shift	Task	base	bal	bal + reg (ours)
Gravity	Ant	35.8 ± 6.2	31.5 ± 13.9	$\mathbf{40.8 \pm 10.3}$
	HalfCheetah	$\mathbf{48.3 \pm 5.0}$	47.6 ± 4.5	47.3 ± 6.6
	Hopper	39.2 ± 19.6	$\mathbf{60.4 \pm 7.8}$	$\mathbf{60.4 \pm 7.3}$
	Walker2d	37.3 ± 6.5	42.1 ± 8.2	$\mathbf{46.5 \pm 5.5}$
Morphology	Ant	$\mathbf{78.1 \pm 0.2}$	77.9 ± 0.2	76.2 ± 2.4
	HalfCheetah	$\mathbf{58.3 \pm 2.8}$	57.3 ± 5.1	57.9 ± 4.4
	Hopper	64.0 ± 10.7	68.2 ± 12.0	$\mathbf{69.4 \pm 13.2}$
	Walker2d	33.6 ± 12.3	47.7 ± 9.5	$\mathbf{52.8 \pm 8.2}$
Sum		394.6	432.7	**451.3**

Table 8: Reconstruction variance explained (%) by the three shared-VAE variants under gravity shifts. Results are mean ± standard deviation.

Task	VAE	Source	Target
HalfCheetah	base	99.17 ± 0.03	96.41 ± 0.23
	bal	98.77 ± 0.04	99.38 ± 0.06
	bal + reg	98.52 ± 0.03	99.30 ± 0.06
Walker2d	base	99.36 ± 0.03	98.32 ± 0.09
	bal	98.95 ± 0.02	99.65 ± 0.03
	bal + reg	98.80 ± 0.02	99.62 ± 0.03

D.3 Friction Results

Table 9 reports results under friction shifts. RIFT obtains the highest aggregate score across the twelve settings, although STC or other baselines perform better on several individual tasks.

D.4 Comparison with DmC

We additionally compare against DmC (Van et al., 2025b), whose official implementation is available in (Van et al., 2025a), on the gravity-shift settings reported in its paper. Table 10 reports best-over-training scores: the maximum normalized score over evaluation checkpoints for each seed, followed by the mean and population standard deviation across seeds. This statistic selects checkpoints using target returns, unlike the final-checkpoint statistic in the main experiments, so we present this comparison separately.

Table 9: Performance under friction shifts when the source dataset is fixed to medium quality and target-data quality varies. Values report the mean ± standard deviation over five seeds, with the highest mean bolded.

Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	STC	**RIFT (ours)**
halfcheetah-med	med	69.7 ± 1.3	66.2 ± 4.1	68.5 ± 1.2	66.7 ± 2.1	66.1 ± 1.9	66.8 ± 0.6	67.4 ± 1.0	$\mathbf{72.4 \pm 2.6}$
halfcheetah-med	med-e	66.8 ± 0.6	60.4 ± 11.9	69.2 ± 0.6	66.8 ± 3.9	68.6 ± 0.6	61.3 ± 5.4	69.4 ± 1.0	$\mathbf{71.4 \pm 1.4}$
halfcheetah-med	expert	68.2 ± 0.4	67.9 ± 0.5	69.7 ± 1.3	67.1 ± 2.3	68.2 ± 2.5	68.3 ± 0.6	67.3 ± 0.6	$\mathbf{70.5 \pm 3.2}$
hopper-med	med	24.5 ± 1.7	24.2 ± 2.2	25.4 ± 1.9	26.5 ± 2.0	27.2 ± 4.3	29.3 ± 4.6	27.4 ± 2.0	$\mathbf{38.8 \pm 2.9}$
hopper-med	med-e	$\mathbf{26.4 \pm 2.5}$	21.7 ± 5.6	23.0 ± 1.7	24.3 ± 3.7	25.7 ± 2.3	$\mathbf{26.4 \pm 2.4}$	23.4 ± 5.1	23.8 ± 11.0
hopper-med	expert	21.5 ± 1.1	24.6 ± 2.5	25.7 ± 0.8	21.1 ± 2.7	23.6 ± 3.8	25.9 ± 1.1	$\mathbf{41.6 \pm 26.1}$	37.8 ± 15.2
walker2d-med	med	72.6 ± 6.2	73.9 ± 11.7	72.1 ± 5.2	67.5 ± 6.0	70.3 ± 3.6	70.4 ± 6.6	73.6 ± 8.6	$\mathbf{74.1 \pm 6.2}$
walker2d-med	med-e	59.0 ± 1.3	61.5 ± 20.2	42.8 ± 17.5	57.4 ± 7.9	59.1 ± 5.4	59.9 ± 10.2	$\mathbf{71.2 \pm 5.8}$	69.7 ± 10.0
walker2d-med	expert	52.5 ± 3.6	60.0 ± 11.7	51.5 ± 12.9	52.3 ± 12.7	54.9 ± 6.1	49.6 ± 14.6	$\mathbf{67.9 \pm 6.6}$	64.1 ± 11.9
ant-med	med	58.2 ± 2.3	58.7 ± 2.0	57.6 ± 4.0	56.9 ± 2.5	55.2 ± 3.2	58.3 ± 0.2	$\mathbf{60.2 \pm 3.1}$	58.5 ± 0.6
ant-med	med-e	59.3 ± 2.5	58.3 ± 1.0	$\mathbf{60.5 \pm 0.3}$	58.4 ± 0.4	57.8 ± 0.5	58.1 ± 0.4	45.7 ± 10.5	58.6 ± 0.4
ant-med	expert	58.2 ± 0.3	58.3 ± 0.5	59.3 ± 1.8	57.2 ± 2.0	58.3 ± 0.2	56.5 ± 2.3	60.6 ± 1.8	$\mathbf{66.0 \pm 16.7}$
Total (displayed means)		636.9	635.7	625.3	622.2	635.0	630.8	675.7	**705.7**

Table 10: Best-over-training performance under gravity shifts when the target dataset is fixed to medium quality and source-data quality varies. Values report the mean ± population standard deviation over five seeds, with the highest mean bolded.

Source	Target	DmC	**RIFT (ours)**
half-med	med	48.0 ± 0.6	$\mathbf{55.0 \pm 1.0}$
half-med-r	med	32.2 ± 0.8	$\mathbf{43.6 \pm 1.9}$
half-med-e	med	49.9 ± 1.4	$\mathbf{56.6 \pm 0.6}$
hopp-med	med	30.6 ± 5.5	$\mathbf{74.4 \pm 5.2}$
hopp-med-r	med	35.4 ± 0.7	$\mathbf{71.9 \pm 4.3}$
hopp-med-e	med	52.2 ± 6.6	$\mathbf{75.0 \pm 2.1}$
walk-med	med	52.5 ± 2.0	$\mathbf{56.2 \pm 3.0}$
walk-med-r	med	42.2 ± 5.9	$\mathbf{53.9 \pm 4.3}$
walk-med-e	med	57.5 ± 3.3	$\mathbf{68.0 \pm 5.9}$
ant-med	med	56.9 ± 2.2	$\mathbf{64.0 \pm 5.6}$
ant-med-r	med	41.8 ± 4.7	$\mathbf{67.8 \pm 4.3}$
ant-med-e	med	55.7 ± 8.1	$\mathbf{59.9 \pm 4.0}$
Total (displayed means)		554.9	**746.3**

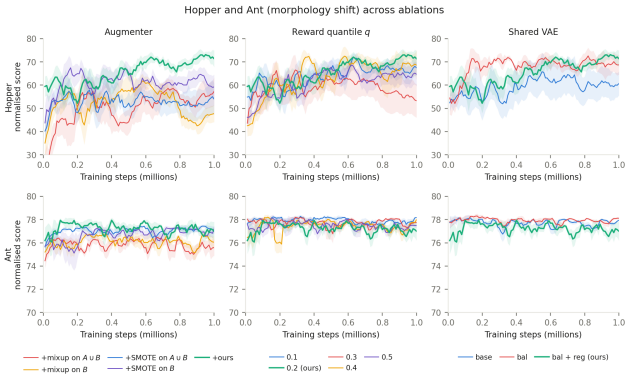

Figure 4: Learning curves for augmentation, reward-quantile, and VAE-design ablations under morphology shifts. Differences are larger on Hopper than on Ant.

D.5 Learning Curves

Figure 4 shows learning curves for the ablations under morphology shifts. The effect of the design choices is task dependent: the variants separate more clearly on Hopper, whereas their performance remains closer on Ant.

D.6 Computing Infrastructure

We list the computing infrastructure used to run all experiments in this paper in Table 11. All experiments are CPU-only. Each IQL training run took around 2 hours on a training core, the main bottleneck for compute.

Table 11: Computing infrastructure used to run all experiments in this paper.

Component	Specification
CPU	AMD EPYC 9474F / 9654
Cores	up to 96 in parallel
Memory	4 GB per run