WHERE TO REUSE, WHAT TO PROPAGATE: TRANSITION RECONSTRUCTION FOR CROSS-DOMAIN OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain offline reinforcement learning leverages abundant source data to alleviate target-domain data scarcity, but dynamics mismatch often induces severe negative transfer. While existing methods attempt to mitigate this by selectively reusing target-compatible source transitions, they create an inherent trade-off between preserving valuable state-action coverage and avoiding dynamics-induced bias. Our analysis shows that this bottleneck stems from treating transitions as indivisible units, which strictly couples where experience is reused with which original outcome is propagated. To break this inherent coupling, we propose shifting the paradigm from transition selection to transition reconstruction, thereby enlarging the admissible transition family. Specifically, our framework, T-RAFT, isolates target-supported source conditions via unbalanced optimal transport (UOT), and amortizes these cross-domain correspondences into a continuous outcome translator using conditional Flow Matching (FM). This decoupled mechanism allows the agent to inherit broad source coverage while replacing biased next states with target-compatible dynamics. Empirical evaluations demonstrate that T-RAFT effectively mitigates the coverage–bias dilemma and achieves competitive performance across diverse dynamics-shift benchmarks.

1 INTRODUCTION

Offline reinforcement learning (RL) enables policy learning entirely from previously collected experience, making it attractive when online interaction is costly, unsafe, or unavailable (Levine et al., 2020; Prudencio et al., 2023). Its performance, however, is fundamentally constrained by the coverage and quality of the available dataset. Cross-domain offline RL alleviates this limitation by leveraging abundant source-domain data together with scarce target-domain experience (Liu et al., 2022; Lyu et al., 2025). The central challenge is dynamics mismatch: directly aggregating source and target data introduces transitions whose next-state outcomes are inconsistent with the target environment, leading to biased Bellman targets and potentially severe negative transfer (Eysenbach et al., 2021). Thus, the key question is not simply how to exploit more source data, but how to convert dynamically shifted source experience into target-compatible supervision.

Based on how source-domain information is utilized, existing methods can be divided into three categories: 1) selective reuse methods filter and reweight source transitions according to their estimated compatibility with the target domain (Liu et al., 2024; Wen et al., 2024; Lyu et al., 2025); 2) information correction methods modify rewards or trajectory returns according to estimated dynamics discrepancy (Liu et al., 2022; Wang et al., 2026); 3) shared information extraction methods transfer domain-invariant structures such as stationary-state information or latent transition mechanisms (Xue et al., 2023; Guo et al., 2026). Despite different designs, most approaches rely on selective source reuse to balance two competing goals: preserving useful state-action coverage and avoiding dynamics-induced negative transfer. Reusing more source data can preserve broader coverage, but it also exposes learning to outcomes generated by mismatched source dynamics. Conservative reuse reduces this risk, but may discard valuable experience when target data are scarce. This persistent tension raises a more basic question:

Can we preserve source-domain coverage without inheriting its mismatched dynamics outcomes?

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.

To understand whether this tension is intrinsic, we analyze the transition family reachable when source transitions are treated as fixed samples that can only be selected or reweighted. Our analysis shows that condition-level reweighting can alter only the state-action marginal while leaving the source conditional dynamics unchanged, resulting in an irreducible residual mismatch even under oracle condition alignment. More general transition-level filtering offers greater flexibility, but the resulting conditional distribution remains confined to the support of the source dynamics and therefore cannot represent target outcomes absent from that support. These results reveal that filtering can determine where to reuse source experience, but cannot independently determine what to propagate, motivating an enlarged transition family based on transition reconstruction.

To break this restriction, our key idea is to enlarge what a reusable source transition is allowed to become. Specifically, we decouple two distinct degrees of freedom: where source experience should be reused and what outcome should be propagated at each retained condition. This shifts cross-domain reuse from transition selection to transition reconstruction. Building on this principle, we propose **T-RAFT** (**T**ransition **R**econstruction via **A**mortized **F**low and unbalanced optimal **T**ransport). UOT instantiates the first degree of freedom by identifying target-supported source conditions and establishing local cross-domain correspondences, while conditional Flow Matching instantiates the second by translating source outcomes toward target-compatible dynamics. This design directly mirrors our theoretical decomposition of reconstruction error: UOT reduces the condition discrepancy governing where source experience is retained, while Flow Matching addresses the outcome discrepancy that remains irreducible under filtering alone. Experimental results on diverse dynamics shifts show that T-RAFT achieves competitive performance across a broad range of settings. Visualization results further show that the reconstructed source outcomes move consistently toward the target distribution, supporting the intended effect of transition reconstruction.

2 Preliminaries

2.1 Cross-Domain Offline Reinforcement Learning

A reinforcement learning problem can be modeled as a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (Sutton & Barto, 1998). In the cross-domain offline setting, the agent learns entirely from static datasets collected from a source-domain $\mathcal{M}_{\text{src}}$ and a target domain $\mathcal{M}_{\text{tar}}$. These domains share the state space $\mathcal{S}$, action space $\mathcal{A}$, reward function R, and discount factor γ, but exhibit misaligned transition dynamics, i.e., $P_{\text{src}}(s' \mid s, a) \neq P_{\text{tar}}(s' \mid s, a)$ (Zhang et al., 2026b). The agent has access to a massive source dataset $\mathcal{D}_{\text{src}}$ and a severely limited target dataset $\mathcal{D}_{\text{tar}}$, where $|\mathcal{D}_{\text{src}}| \gg |\mathcal{D}_{\text{tar}}|$ and $|\cdot|$ denote the size of dataset. The core challenge is to leverage the broad state-action coverage of $\mathcal{D}_{\text{src}}$ to maximize target-domain returns. Naively aggregating the two datasets injects dynamically inconsistent next states into Bellman backups, precipitating severe negative transfer.

2.2 Optimal Transport and Marginal Relaxation

Let $\mu = \sum_{i=1}^{n_\mu} \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{n_\nu} \beta_j \delta_{z_j}$ be discrete probability measures with mass vectors $\boldsymbol{\alpha} \in \Delta^{n_\mu}$ and $\boldsymbol{\beta} \in \Delta^{n_\nu}$, and let $C_{ij} = c(x_i, z_j)$ denote the pairwise transport cost. Standard optimal transport (OT) (Cuturi, 2013; Panaretos & Zemel, 2019) seeks the minimum-cost coupling:

$$\min_{\mathbf{\Gamma} \geq 0} \langle \mathbf{\Gamma}, \mathbf{C} \rangle, \quad \text{s.t. } \mathbf{\Gamma}\mathbf{1} = \boldsymbol{\alpha},\ \mathbf{\Gamma}^\top \mathbf{1} = \boldsymbol{\beta}. \tag{1}$$

However, exact marginal constraints require all probability mass to be transported, which can produce undesirable matches when two distributions contain weakly overlapping regions. Unbalanced optimal transport (Chizat et al., 2018) relaxes these constraints by penalizing marginal preservation:

$$\min_{\mathbf{\Gamma} \geq 0} \langle \mathbf{\Gamma}, \mathbf{C} \rangle + \lambda_{\text{src}} D_{\text{KL}}^{g}(\mathbf{\Gamma}\mathbf{1} \| \boldsymbol{\alpha}) + \lambda_{\text{tar}} D_{\text{KL}}^{g}(\mathbf{\Gamma}^\top \mathbf{1} \| \boldsymbol{\beta}), \tag{2}$$

where $\lambda_{\text{src}}, \lambda_{\text{tar}} > 0$ control the strength of marginal relaxation. By allowing mass to be partially discarded or created, UOT supports flexible partial matching and is therefore better suited to distributions with unequal support or substantial outlier mass (Chapel et al., 2020).

2.3 Flow Matching

Flow Matching (FM) (Lipman et al., 2023) is a generative modeling framework that learns a continuous transport between distributions through a time-dependent velocity field $v_\theta : \mathbb{R}^{d_x} \times [0,1] \to \mathbb{R}^{d_x}$. The induced flow ψ_θ follows the ordinary differential equation (ODE):

$$\frac{d}{dt}\psi_\theta(x_0, t) = v_\theta(\psi_\theta(x_0, t), t)\,, \qquad \psi_\theta(x_0, 0) = x_0. \tag{3}$$

Given endpoint pairs $(x_0, x_1) \sim q(x_0, x_1)$, conditional FM commonly uses the linear path $x_t = (1-t)x_0 + tx_1$ (Albergo & Vanden-Eijnden, 2023) and learns its velocity by:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{(x_0,x_1)\sim q,\, t\sim\mathcal{U}[0,1]}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|_2^2\right]. \tag{4}$$

This regression objective provides a simulation-free training procedure for continuous flows, avoiding ODE integration during optimization. Once trained, the learned vector field defines a continuous transport map that can generalize beyond the finite endpoint pairs used for training.

3 Why Is Filtering Alone Insufficient?

Prior filtering-based methods determine which source transitions contribute to target-domain learning, but they do not necessarily correct the conditional dynamics associated with retained state-action conditions. We formalize this limitation below. Let $x = (s, a)$ denote a state-action condition and $y = s'$ its outcome. For domain $d \in \{\mathrm{src}, \mathrm{tar}\}$, write $P_d(dx, dy) = \mu_d(dx)P_d(dy \mid x)$, where μ_d is the condition marginal and $P_d(\cdot \mid x)$ is the conditional dynamics. A condition-level reweighting $w(x)$ induces $P_w(dx, dy) = \mu_w(dx)P_{\mathrm{src}}(dy \mid x)$, where $\mu_w(dx) \propto w(x)\mu_{\mathrm{src}}(dx)$ changes the condition marginal while leaving the source conditional dynamics unchanged.

Theorem 3.1 (Irreducible conditional dynamics mismatch)**.** *For any condition-level reweighting w,*

$$D_{\mathrm{TV}}(P_{\mathrm{tar}}, P_w) \geq \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[D_{\mathrm{TV}}\big(P_{\mathrm{tar}}(\cdot \mid x), P_{\mathrm{src}}(\cdot \mid x)\big)\right] - D_{\mathrm{TV}}(\mu_{\mathrm{tar}}, \mu_w). \tag{5}$$

In particular, under oracle condition alignment $\mu_w = \mu_{\mathrm{tar}}$,

$$D_{\mathrm{TV}}(P_{\mathrm{tar}}, P_w) = \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[D_{\mathrm{TV}}\big(P_{\mathrm{tar}}(\cdot \mid x), P_{\mathrm{src}}(\cdot \mid x)\big)\right]. \tag{6}$$

Theorem 3.1 shows that even perfect condition alignment cannot remove the residual conditional dynamics mismatch. The limitation therefore lies not only in how accurately transitions are weighted, but in the restricted transition family reachable by reweighting: μ_w may change, whereas $P_{\mathrm{src}}(\cdot \mid x)$ remains fixed. More general transition-level filtering can reshape this conditional distribution, but only within the support of the source dynamics, i.e., $P_h(\cdot \mid x) \ll P_{\mathrm{src}}(\cdot \mid x)$. Hence, eliminating the residual mismatch generally requires an additional degree of freedom that can modify the outcome associated with a retained condition, motivating the enlarged transition family introduced next.

4 Proposed Method

Building on the limitation identified above, in this section we enlarge the admissible transition family to introduce the missing degree of freedom for outcome correction. We first characterize the reconstruction error of this enlarged family by decomposing it into condition-alignment and outcome-correction terms, which motivate the two components of our method. We then instantiate these two objectives with UOT and conditional Flow Matching, respectively, and further establish how their optimization errors jointly control the empirical reconstruction error. Finally, we formalize how the UOT alignment and Flow Matching objectives jointly control the resulting empirical reconstruction error in the following proposition. The overall framework is illustrated in Figure 1.

4.1 Enlarging the Admissible Transition Family

Section 3 shows that condition-level reweighting is restricted to $P_w(dx, dy) = \mu_w(dx)P_{\mathrm{src}}(dy \mid x)$, where only the condition marginal can change. We therefore enlarge this family by additionally allowing the outcome distribution to be transformed while preserving the state-action condition.

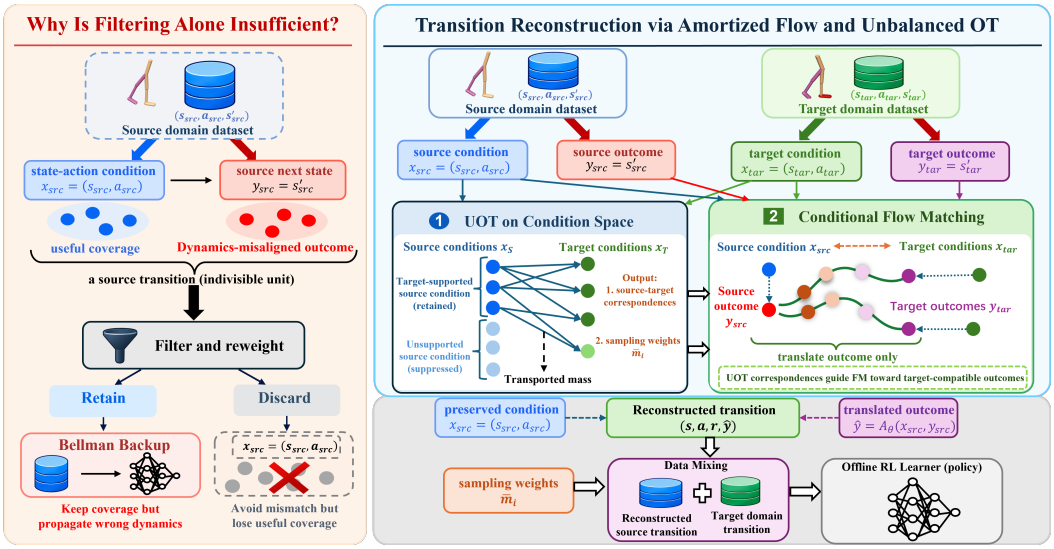

Figure 1: Overview of T-RAFT. **Left:** Filtering treats each source transition as an indivisible unit, thereby forcing a trade-off between coverage and dynamics compatibility. **Right:** T-RAFT decouples where source experience is reused from what outcome is propagated. UOT first aligns source and target conditions, after which conditional Flow Matching translates the source outcome while preserving its state-action condition. The resulting target-compatible reconstructed transitions are combined with target-domain data for downstream offline RL.

Definition 4.1 (Condition-preserving outcome transformation)**.** *Let $\mathcal{Z}$ be an optional latent space with base distribution ρ, and let $A_\theta : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$ be a measurable outcome transformation with $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and $\mathcal{Y} = \mathcal{S}$. For a source transition (x, y), we retain the condition x and construct $\bar{y} = A_\theta(x, y, \zeta)$, where $\zeta \sim \rho$. This induces the translated conditional kernel $K_\theta(\cdot \mid x) = \big(A_\theta(x, \cdot, \cdot)\big)_{\#}\big(P_{\mathrm{src}}(\cdot \mid x) \otimes \rho\big)$ with corresponding enlarged transition family*

$$\mathcal{P}_{\mathrm{aug}} = \left\{ \widetilde{P}_{w,\theta}(dx, d\bar{y}) = \mu_w(dx) K_\theta(d\bar{y} \mid x) \right\}, \tag{7}$$

where $(\cdot)_{\#}(\cdot)$ denotes the pushforward operator.

We next characterize the reconstruction error within $\mathcal{P}_{\mathrm{aug}}$. Let $W_1^{\mathcal{X}}$ and $W_1^{\mathcal{Y}}$ denote the 1-Wasserstein distances on the condition and outcome spaces, and let W_1 denote their joint counterpart under the additive product metric. We have

Proposition 4.1 (Wasserstein condition-outcome decomposition)**.** *Suppose that the translated conditional kernel K_θ is L_K-Lipschitz with respect to the condition under $W_1^{\mathcal{Y}}$, and denote $C_K := 1 + L_K$. Then, for any $\widetilde{P}_{w,\theta} \in \mathcal{P}_{\mathrm{aug}}$, the target-transition reconstruction error is bounded by*

$$W_1\Big(P_{\mathrm{tar}}, \widetilde{P}_{w,\theta}\Big) \le C_K \underbrace{W_1^{\mathcal{X}}(\mu_{\mathrm{tar}}, \mu_w)}_{\text{(a): condition discrepancy}} + \underbrace{\mathbb{E}_{x \sim \mu_{\mathrm{tar}}}\left[W_1^{\mathcal{Y}}(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x))\right]}_{\text{(b): outcome discrepancy}}. \tag{8}$$

In particular, if $\mu_w = \mu_{\mathrm{tar}}$ and $K_\theta(\cdot \mid x) = P_{\mathrm{tar}}(\cdot \mid x)$ for μ_{tar}-almost every x, then $\widetilde{P}_{w,\theta} = P_{\mathrm{tar}}$. The basic assumption and detailed derivation are given in Appendix C.

Remark 4.1. *Proposition 4.1 formalizes the two degrees of freedom required for reconstruction: term (a) determines where source mass is retained, whereas term (b) determines what outcome is produced at retained conditions. Setting $K_\theta = P_{\mathrm{src}}$ recovers the reweighting family, leaving term (b) as the Wasserstein analogue of the residual conditional mismatch exposed in Theorem 3.1.*

4.2 Target-Supported Condition Alignment via UOT

Proposition 4.1 identifies the condition marginal μ_w as the degree of freedom that determines where source experience is retained. We therefore seek an empirical source marginal that preserves condi-

tions supported by the target data while suppressing those without plausible target counterparts. Let $\{x_i^{\mathrm{src}}\}_{i=1}^{N_{\mathrm{src}}}$ and $\{x_j^{\mathrm{tar}}\}_{j=1}^{N_{\mathrm{tar}}}$ denote the source and target conditions, with empirical marginals

$$\widehat{\mu}_{\mathrm{src}} = \sum_{i=1}^{N_{\mathrm{src}}} \alpha_i \delta_{x_i^{\mathrm{src}}}, \qquad \widehat{\mu}_{\mathrm{tar}} = \sum_{j=1}^{N_{\mathrm{tar}}} \beta_j \delta_{x_j^{\mathrm{tar}}}, \tag{9}$$

with $\alpha \in \Delta^{N_{\mathrm{src}}}$ and $\beta \in \Delta^{N_{\mathrm{tar}}}$. This selective objective naturally calls for unbalanced rather than balanced optimal transport, as standard OT enforces $\Gamma \mathbf{1} = \alpha$, which forces every source condition to transport its full mass even when no plausible target counterpart exists. UOT instead relaxes the marginal constraints, allowing costly source mass to be removed rather than forcibly matched. We therefore solve Equation (2) on the condition space with $C_{ij} = \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2^2$:

$$\Gamma^\star = \arg\min_{\Gamma \geq 0} \langle \Gamma, C \rangle + \lambda_{\mathrm{src}} D_{\mathrm{KL}}^g(\Gamma \mathbf{1} \| \alpha) + \lambda_{\mathrm{tar}} D_{\mathrm{KL}}^g(\Gamma^\top \mathbf{1} \| \beta). \tag{10}$$

The transported source mass directly instantiates the desired condition marginal, where conditions well supported by the target data retain larger transported mass, whereas unsupported conditions are suppressed. We therefore define

$$\widehat{\mu}_w = \sum_{i=1}^{N_{\mathrm{src}}} \bar{m}_i \, \delta_{x_i^{\mathrm{src}}}, \tag{11}$$

where $m_i := (\Gamma^\star \mathbf{1})_i$, and $\bar{m}_i := m_i / \sum_{k=1}^{N_{\mathrm{src}}} m_k$. Equation (11) is the empirical realization of μ_w in Proposition 4.1. It instantiates the condition degree of freedom by determining how much source mass should be retained under target support. Beyond its transported marginal, the optimal coupling $\Gamma^\star$ also records pairwise source-target correspondences. We next exploit this additional structure to determine how the outcomes associated with retained source conditions should be corrected.

4.3 Coupling-Guided Outcome Translation via Flow Matching

Section 4.2 instantiates the condition degree of freedom through $\widehat{\mu}_w$, determining where source experience should be retained. However, the retained transitions still follow the source conditional dynamics $P_{\mathrm{src}}(\cdot \mid x)$. We next instantiate the remaining outcome degree of freedom by learning a translated kernel $K_\theta(\cdot \mid x)$ from the pairwise correspondences encoded by the UOT coupling.

Specifically, for each retained source condition x_i^{src} with $m_i > 0$, we first define

$$\eta_{ij} := \frac{\Gamma_{ij}^\star}{m_i}, \qquad \sum_{j=1}^{N_{\mathrm{tar}}} \eta_{ij} = 1. \tag{12}$$

The weights η_{ij} specify a conditional correspondence from x_i^{src} to target conditions x_j^{tar}. Since UOT favors low-cost condition matches, the associated target outcomes y_j^{tar} provide local empirical supervision for correcting the source outcome y_i^{src}.

However, these correspondences are discrete and sample-specific, whereas Proposition 4.1 requires a reusable conditional kernel $K_\theta(\cdot \mid x)$ for any source transition. We therefore use conditional Flow Matching to amortize these UOT-induced correspondences into a continuous outcome transformation. Specifically, for a sampled pair $i \sim \bar{m}$, $j \sim \eta_i$, with a time step $t \sim \mathcal{U}[0,1]$, we construct $y_t = (1-t) y_i^{\mathrm{src}} + t y_j^{\mathrm{tar}}$ and train a condition-dependent velocity field $v_\theta(y, t; x)$ via

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\left[\left\| v_\theta(y_t, t; x_i^{\mathrm{src}}) - \left(y_j^{\mathrm{tar}} - y_i^{\mathrm{src}}\right) \right\|_2^2 \right]. \tag{13}$$

Crucially, the condition x_i^{src} is used only to condition the velocity field, and the flow itself evolves exclusively in the outcome space. We can produce a translated source outcome by integrating $\frac{du_t}{dt} = v_\theta(u_t, t; x_i^{\mathrm{src}})$ with $u_0 = y_i^{\mathrm{src}}$, which we denote by $\bar{y}_i = A_\theta(x_i^{\mathrm{src}}, y_i^{\mathrm{src}})$. This condition-dependent velocity field thereby realizes the conditional kernel $K_\theta(\cdot \mid x)$ introduced in Section 4.1.

Finally, we formalize how both the UOT alignment and Flow Matching objectives jointly control the resulting empirical reconstruction error and get the following proposition.

Algorithm 1 Transition Reconstruction via Amortized Flow and unbalanced optimal Transport

// Stage 1: Target-Supported Condition Alignment
Construct the condition cost $C_{ij} = \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2^2$.
Solve the condition-space UOT problem in Equation (10) to obtain $\Gamma^\star$.
Compute $m_i = (\Gamma^\star \mathbf{1})_i$ and construct $\widehat{\mu}_w$ using $\bar{m}_i$ as in Equation (11).
Compute the conditional correspondences η_{ij} according to Equation (12).
// Stage 2: Coupling-Guided Outcome Translation
Train v_θ by sampling $i \sim \bar{m}$, $j \sim \eta_i$, and optimizing Equation (13).
Initialize $\widetilde{\mathcal{D}}_{\mathrm{src}} \leftarrow \varnothing$.
for each $\tau_i^{\mathrm{src}} = (s_i^{\mathrm{src}}, a_i^{\mathrm{src}}, r_i^{\mathrm{src}}, y_i^{\mathrm{src}}) \in \mathcal{D}_{\mathrm{src}}$ **do**
  Integrate the learned flow from y_i^{src} to obtain $\bar{y}_i = A_\theta(x_i^{\mathrm{src}}, y_i^{\mathrm{src}})$.
  Add $\widetilde{\tau}_i^{\mathrm{src}} = (s_i^{\mathrm{src}}, a_i^{\mathrm{src}}, r_i^{\mathrm{src}}, \bar{y}_i)$ to $\widetilde{\mathcal{D}}_{\mathrm{src}}$.
end for
// Stage 3: Offline Policy Learning
repeat
  Sample $\mathcal{B}_{\mathrm{tar}} \subset \mathcal{D}_{\mathrm{tar}}$ uniformly, $\widetilde{\mathcal{B}}_{\mathrm{src}} \subset \widetilde{\mathcal{D}}_{\mathrm{src}}$ with $i \sim \bar{m}$, and set $\mathcal{B}_{\mathrm{mix}} = \mathcal{B}_{\mathrm{tar}} \cup \widetilde{\mathcal{B}}_{\mathrm{src}}$.
  Update V_ψ, Q_ξ, and π_ω using Equation (16), (17), and (18), respectively.
until convergence

Proposition 4.2 (UOT–Flow Matching reconstruction bound)**.** *Let $M := \|\Gamma^\star\|_1 > 0$, and let $\widehat{P}_\theta$ and $\widehat{P}_{\mathrm{tar}}$ denote the empirical reconstructed and target transition distributions, respectively. Under the regularity conditions stated in Appendix D, we have*

$$W_1\left(\widehat{P}_\theta, \widehat{P}_{\mathrm{tar}}\right) \le \underbrace{\sqrt{\frac{\langle \Gamma^\star, C\rangle}{M}}}_{\text{(a): condition alignment}} + \underbrace{C_{\mathrm{FM}}\sqrt{\mathcal{L}_{\mathrm{FM}}(\theta)}}_{\text{(b): outcome correction}} + \underbrace{C_{\mathrm{tar}}\sqrt{\frac{D_{\mathrm{KL}}^g((\Gamma^\star)^\top \mathbf{1} \,\|\, \beta)}{M}}}_{\text{(c): marginal relaxation}}. \tag{14}$$

where $C_{\mathrm{FM}}, C_{\mathrm{tar}} < \infty$ are regularity constants specified in the appendix.

Remark 4.2. *Proposition 4.2 provides the empirical realization of the two-error decomposition in Proposition 4.1. The condition component is controlled by the UOT transport geometry together with the error introduced by relaxing the target marginal, while the outcome component is controlled by the Flow Matching error. Hence, the same coupling $\Gamma^\star$ determines where source experience is reused and supplies the supervision from which Flow Matching learns what outcome should replace the original source dynamics.*

4.4 Target-Compatible Transition Reconstruction and Offline RL

We finally instantiate the reconstructed distribution $\widetilde{P}_{w,\theta}$ within an offline RL learner. For each source transition $\tau_i^{\mathrm{src}} = (s_i^{\mathrm{src}}, a_i^{\mathrm{src}}, r_i^{\mathrm{src}}, y_i^{\mathrm{src}})$, we integrate the learned conditional flow from y_i^{src} and obtain $\bar{y}_i = A_\theta(x_i^{\mathrm{src}}, y_i^{\mathrm{src}})$, where $x_i^{\mathrm{src}} = (s_i^{\mathrm{src}}, a_i^{\mathrm{src}})$. This yields

$$\widetilde{\tau}_i^{\mathrm{src}} = \left(s_i^{\mathrm{src}}, a_i^{\mathrm{src}}, r_i^{\mathrm{src}}, \bar{y}_i\right), \tag{15}$$

where the condition is preserved and only the next-state outcome is translated. During policy learning, reconstructed source transitions are sampled according to $\bar{m}_i$, thereby realizing the target-supported marginal $\widehat{\mu}_w$ derived in Section 4.2.

As our reconstruction module is compatible with general offline RL algorithms, we instantiate it with the IQL (Kostrikov et al., 2022) framework. At each update, we sample a target minibatch $\mathcal{B}_{\mathrm{tar}} \subset \mathcal{D}_{\mathrm{tar}}$ uniformly and a reconstructed source minibatch $\widetilde{\mathcal{B}}_{\mathrm{src}}$ according to $i \sim \bar{m}$, and define $\mathcal{B}_{\mathrm{mix}} = \mathcal{B}_{\mathrm{tar}} \cup \widetilde{\mathcal{B}}_{\mathrm{src}}$. Then the value function is learned by expectile regression:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{B}_{\mathrm{mix}}}\left[\ell_{\tau_{\mathrm{IQL}}}\left(Q_{\bar{\xi}}(s,a) - V_\psi(s)\right)\right], \quad \text{where} \quad \ell_{\tau_{\mathrm{IQL}}}(u) = |\tau_{\mathrm{IQL}} - \mathbf{1}\{u<0\}|\, u^2. \tag{16}$$

The critic is then trained using the original successor state for target transitions and the translated successor state for reconstructed source transitions:

$$\mathcal{L}_Q(\xi) = \mathbb{E}_{\mathcal{B}_{\mathrm{tar}}}\left[\left(Q_\xi(s,a) - r - \gamma V_\psi(y)\right)^2\right] + \mathbb{E}_{\widetilde{\mathcal{B}}_{\mathrm{src}}}\left[\left(Q_\xi(s_i^{\mathrm{src}}, a_i^{\mathrm{src}}) - r_i^{\mathrm{src}} - \gamma V_\psi(\bar{y}_i)\right)^2\right]. \tag{17}$$

Table 1: Performance comparison on MuJoCo kinematic-shift tasks. We report normalized target-domain scores; values with uncertainty are mean±standard deviation over five seeds. Source abbreviations use the form Env-Type, where m, me, and mr denote medium, medium-expert, and medium-replay, respectively. The target column denotes target-dataset quality.

Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	DmC	ROOT	T-RAFT
Ant-m	medium	50.0±5.6	42.3±7.6	20.9±2.6	50.5±6.7	54.5±1.3	55.4±0.0	62.1±0.6	61.9±0.3	**62.9±0.5**
Ant-m	med-expert	57.8±7.2	54.1±3.8	31.7±7.0	54.9±1.3	54.5±4.6	60.7±3.6	68.9±1.0	67.4±1.1	**72.8±1.8**
Ant-m	expert	59.6±18.5	54.2±11.3	45.4±8.6	45.5±9.3	49.4±14.6	90.4±4.8	92.1±3.5	**97.2±0.8**	92.2±6.2
Ant-me	medium	49.5±4.1	44.7±4.3	19.0±8.0	41.3±8.1	41.8±8.8	50.2±4.3	60.6±1.3	60.5±1.5	**61.2±1.1**
Ant-me	med-expert	37.2±2.0	33.3±7.0	6.4±2.5	32.8±8.0	41.5±4.9	48.8±2.7	60.4±3.7	65.4±2.8	**65.5±2.5**
Ant-me	expert	18.7±8.1	17.8±23.6	14.5±9.0	35.2±15.5	14.4±22.9	78.4±12.2	76.0±4.1	**90.2±7.0**	76.7±6.5
Ant-mr	medium	43.7±4.6	42.0±5.4	19.0±1.8	45.3±5.1	41.4±5.0	52.8±4.4	61.9±0.5	60.8±1.2	**62.5±1.1**
Ant-mr	med-expert	36.5±5.9	36.0±6.7	19.1±1.6	36.2±6.6	37.2±4.7	54.2±5.2	58.8±3.6	**63.8±3.8**	62.6±1.4
Ant-mr	expert	24.4±4.8	22.1±0.4	19.5±0.8	27.1±3.7	24.3±2.8	**74.7±10.5**	43.8±2.6	55.2±7.7	40.8±4.2
Half-m	medium	12.3±1.2	10.6±1.2	8.3±1.2	16.8±4.2	23.6±5.7	40.2±0.0	38.5±1.4	**40.9±0.5**	39.6±3.8
Half-m	med-expert	10.8±1.9	12.9±2.8	8.7±1.3	10.3±2.7	9.8±2.4	10.1±4.0	19.1±1.0	**29.2±4.4**	20.2±3.7
Half-m	expert	12.6±1.7	12.1±1.0	10.8±1.7	12.2±0.9	12.8±0.7	8.7±2.0	13.1±0.8	12.4±1.0	**13.2±0.7**
Half-me	medium	21.8±6.5	25.9±7.4	30.0±4.3	17.2±3.3	21.9±6.5	30.7±9.6	38.4±1.4	40.4±0.7	**40.9±0.6**
Half-me	med-expert	7.6±1.4	9.5±4.2	6.8±2.9	9.6±2.4	8.9±3.3	10.9±4.2	24.1±4.6	**29.0±4.5**	25.4±6.7
Half-me	expert	9.1±2.4	10.4±1.3	4.9±3.2	11.2±1.0	10.7±1.4	3.2±0.6	13.4±2.0	10.7±1.8	**13.9±2.4**
Half-mr	medium	10.0±5.4	11.5±4.9	7.5±3.1	10.2±3.7	11.6±4.6	**37.8±2.1**	19.5±1.8	30.4±4.9	30.2±4.3
Half-mr	med-expert	6.5±3.1	9.2±4.7	6.6±1.7	9.5±1.8	8.6±2.3	9.7±2.0	**11.4±2.1**	10.1±1.7	10.6±1.4
Half-mr	expert	13.6±1.4	14.8±2.0	10.4±4.9	14.8±2.2	13.9±2.2	7.2±1.4	15.6±2.9	**21.4±3.6**	19.2±2.7
Hopp-m	medium	58.7±8.4	43.9±15.2	12.3±6.6	65.4±1.5	65.3±1.4	65.6±1.9	**69.8±2.3**	68.2±0.5	68.2±0.6
Hopp-m	med-expert	68.5±12.4	55.4±16.9	15.6±10.8	43.9±30.8	51.1±18.5	55.4±25.1	78.2±5.1	**90.3±13.7**	84.7±9.2
Hopp-m	expert	79.9±35.5	83.7±19.6	14.8±5.5	53.1±39.8	87.4±25.4	35.0±19.4	59.8±21.8	99.4±2.0	**99.5±0.6**
Hopp-me	medium	66.0±0.5	61.1±4.0	35.0±20.1	64.6±2.6	65.2±1.5	65.3±2.4	**69.6±1.3**	68.2±0.9	68.3±1.2
Hopp-me	med-expert	45.1±15.7	61.9±16.9	13.9±4.9	54.7±17.0	62.9±15.6	38.6±15.9	75.5±9.6	86.8±9.3	**91.8±10.9**
Hopp-me	expert	44.9±19.8	84.2±21.1	12.0±4.3	57.6±40.6	52.8±19.7	29.9±11.3	64.5±24.2	**92.3±19.7**	81.8±25.4
Hopp-mr	medium	36.0±0.1	39.4±7.2	3.2±2.6	36.1±0.2	35.9±2.4	35.5±12.2	**64.8±2.4**	54.9±7.7	60.1±7.4
Hopp-mr	med-expert	36.1±0.1	34.1±3.6	4.4±2.8	36.0±0.1	36.1±0.1	47.5±14.6	**69.7±7.5**	60.4±9.8	69.2±11.6
Hopp-mr	expert	36.0±0.1	36.1±0.2	3.7±2.5	36.1±0.1	36.1±0.3	49.9±30.5	**69.9±18.0**	56.8±15.3	59.0±18.3
Walk-m	medium	34.3±9.8	35.2±22.5	14.3±11.2	39.0±6.7	41.9±11.2	49.6±18.0	63.2±4.2	**64.3±6.3**	62.5±10.6
Walk-m	med-expert	30.2±12.5	51.9±11.5	13.6±7.7	38.6±6.5	42.3±19.3	43.5±16.4	53.5±7.0	70.8±11.6	**75.3±8.7**
Walk-m	expert	56.4±18.2	40.7±14.4	15.3±2.5	57.3±12.2	60.4±17.5	46.7±13.6	70.5±12.0	**90.5±8.9**	86.3±5.7
Walk-me	medium	41.8±8.8	38.1±14.4	21.4±8.3	36.9±4.3	41.2±13.0	44.6±6.0	59.4±6.8	60.9±5.2	**64.5±2.5**
Walk-me	med-expert	22.2±8.7	23.6±8.1	15.9±4.1	23.2±7.9	28.1±4.0	16.5±7.2	53.2±7.3	53.5±5.0	**75.4±11.7**
Walk-me	expert	26.3±10.4	36.0±9.2	18.5±3.6	40.9±9.6	46.2±19.4	42.4±9.1	69.2±7.0	75.8±6.7	**83.5±7.6**
Walk-mr	medium	11.5±7.1	12.5±4.3	1.9±2.1	14.3±3.1	22.2±5.2	49.7±9.7	**52.9±8.4**	52.6±1.9	49.1±4.9
Walk-mr	med-expert	9.7±3.8	11.2±5.0	4.6±3.0	4.2±5.1	7.6±4.9	**55.9±17.1**	36.4±5.4	35.5±7.3	45.6±12.2
Walk-mr	expert	7.7±4.8	7.4±2.4	3.6±1.5	13.2±8.5	7.5±2.1	**51.9±7.9**	44.4±8.5	45.4±15.9	48.5±9.6

Thus, UOT determines which source conditions participate in value learning through the sampling distribution $\bar{m}$, whereas Flow Matching replaces the source-dynamics continuation value $V_\psi(y_i^{\mathrm{src}})$ with the translated continuation value $V_\psi(\bar{y}_i)$. Finally, the policy is optimized using:

$$\mathcal{L}_\pi(\omega) = -\mathbb{E}_{(s,a)\sim\mathcal{B}_{\mathrm{mix}}}\left[\exp\left(\beta_{\mathrm{IQL}}\,\mathrm{Adv}_{\xi,\psi}(s,a)\right)\log\pi_\omega(a\mid s)\right], \tag{18}$$

where $\mathrm{Adv}_{\xi,\psi}(s,a) := Q_\xi(s,a) - V_\psi(s)$ is the advantage function and $\beta_{\mathrm{IQL}} > 0$ controls the strength of advantage weighting. The reconstructed dynamics therefore influence policy learning through the critic and its induced advantages, while preserving the state-action coverage of reusable source data. The complete pipeline of our algorithm is summarized in Algorithm 1.

5 Experiments

In this section, we evaluate T-RAFT by answering three questions: (Q1) Does T-RAFT improve target-domain policy learning? (Q2) When does the method remain effective under increasingly large dynamics shifts? (Q3) How sensitive is T-RAFT to its key parameters λ_{src} and λ_{tar}?

5.1 Main Results

Experimental setup. We evaluate T-RAFT on both locomotion and dexterous-manipulation benchmarks with controlled dynamics shifts. For MuJoCo locomotion, we follow the experimental protocol of OTDF (Lyu et al., 2025) and consider four environments—Ant, HalfCheetah, Hopper, and Walker2d—under three representative types of dynamics mismatch: gravity, kinematic, and morphology shifts. We use the D4RL (Fu et al., 2020) medium, medium-replay, and medium-expert datasets as source-domain data, whereas the target datasets are collected in the modified environments with medium, medium-expert, and expert behavior quality. Following the limited-target-data

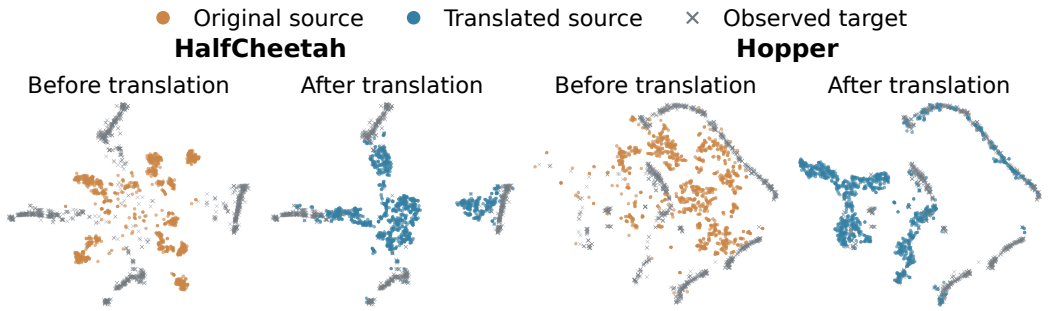

Figure 2: Visualization results on source transition outcomes before and after translation. HalfCheetah (left pair) and Hopper (right pair) under kinematic shifts. Orange and blue points denote original and translated source increments, respectively.

Table 2: Results on HalfCheetah under gravity shift with different levels. We report normalized scores as mean±standard deviation. Baseline results are reproduced from Guo et al. (2026) over five seeds. **Bold** and underlined values indicate the best and second-best results, respectively.

Gravity scale	IQL	DARA	BOSA	REAG	SRPO	MOBODY	T-RAFT
0.1×	9.62 ± 4.27	12.90 ± 1.01	9.31 ± 1.94	16.14 ± 0.66	$\mathbf{32.94 \pm 1.65}$	14.18 ± 1.06	$\underline{27.56 \pm 3.17}$
0.5×	44.23 ± 2.93	46.11 ± 1.93	43.96 ± 5.68	40.50 ± 1.58	41.99 ± 1.63	$\underline{47.18 \pm 1.23}$	$\mathbf{49.86 \pm 0.41}$
2.0×	31.34 ± 1.68	31.85 ± 1.31	27.86 ± 0.94	33.28 ± 3.16	32.24 ± 1.97	$\mathbf{41.60 \pm 7.35}$	$\underline{34.70 \pm 0.89}$
5.0×	44.00 ± 23.13	27.67 ± 17.01	17.95 ± 11.97	71.31 ± 2.80	-2.33 ± 0.69	$\mathbf{83.05 \pm 1.21}$	$\underline{73.53 \pm 0.44}$

setting, each target dataset contains only five trajectories, corresponding to approximately $5{,}000$ transitions. We compare T-RAFT against IQL (Kostrikov et al., 2022), DARA (Liu et al., 2022), BOSA (Liu et al., 2024), SRPO (Xue et al., 2023), IGDF (Wen et al., 2024), OTDF (Lyu et al., 2025), DmC (Le Pham Van et al., 2025), and ROOT (Chen et al., 2026), and report normalized returns over five random seeds.

To further assess whether T-RAFT extends beyond locomotion, we additionally evaluate it on the Adroit Pen and Door manipulation tasks, including both kinematic and morphology shifts at medium and hard mismatch levels. Detailed dataset construction, dynamics modifications, and evaluation protocols are provided in Appendix F and wider experimental results are shown in Appendix G.

Performance Analysis As shown in Table 1, T-RAFT achieves the highest mean return in 14 of 36 configurations, with leading results appearing across all four environments. The gains are particularly pronounced when source coverage is useful but its associated dynamics are mismatched. For example, with medium-expert Walker2d source data, T-RAFT exceeds the strongest baseline by 21.9 and 7.7 normalized-score points on the medium-expert and expert target datasets, respectively.

Figure 2 provides complementary evidence for this mechanism. We jointly embed original source, translated source, and observed target state increments $\Delta s = s' - s$ for HalfCheetah and Hopper. After translation, source outcomes move visibly toward the target distribution. More quantitatively, the local outcome discrepancy to the same condition-nearest target samples decreases for 83.7% of HalfCheetah and 91.6% of Hopper source transitions, with lower median discrepancies in both cases. These results indicate that reconstructing source transitions can retain useful decision coverage without being restricted to their original source-domain outcomes.

5.2 Robustness to Large Dynamics Shifts

Experimental setup. We further examine whether T-RAFT remains reliable as the source-target dynamics discrepancy becomes substantial. Following the experimental setup introduced in ODRL (Lyu et al., 2024b), we consider the HalfCheetah environment under gravity shifts, which keeps the source dynamics fixed and scales the target gravity by $\{0.1, 0.5, 2.0, 5.0\}$ with medium-quality source and target datasets. These settings perturb the dynamics in both directions and include severe shifts far from the source-domain.

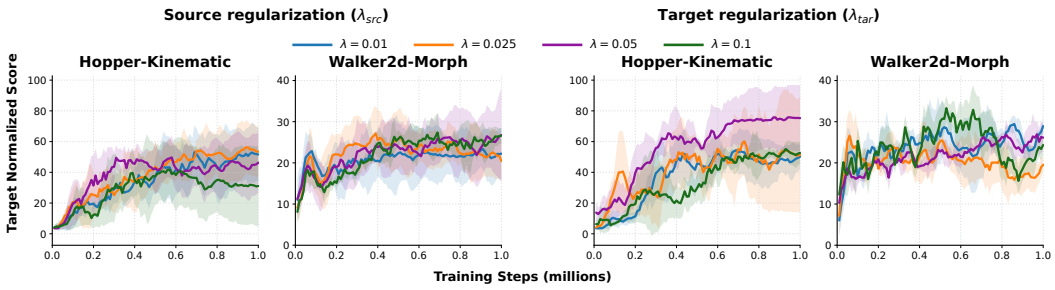

Figure 3: Sensitivity to the two coefficients. The left two panels vary λ_{src}, while the right two vary λ_{tar}, on Hopper-Kinematic and Walker2d-Morph. The other coefficient is fixed to 0.05. Curves show mean target normalized scores with standard-deviation bands over training.

Performance Analysis As shown in Table 2, T-RAFT remains competitive across all tested shifts rather than collapsing as the mismatch increases. It achieves the best result at $0.5\times$ gravity, and second-best results at the other three levels. This direction-dependent behavior indicates that reconstruction difficulty is not determined solely by the magnitude of the physical perturbation. More importantly, the absence of severe degradation under large shifts suggests that separating target-supported conditions from their source-domain outcomes prevents T-RAFT from directly propagating increasingly incompatible source dynamics.

5.3 Sensitivity to Key Parameters

We study the sensitivity of T-RAFT to the two key coefficients, λ_{src} and λ_{tar}. For each coefficient, we vary its value over $\{0.01, 0.025, 0.05, 0.1\}$. We conduct the analysis on Hopper-Kinematic and Walker2d-Morph, which represent two distinct dynamics-shift settings.

Figure 3 shows that T-RAFT remains effective across a reasonably broad range of both coefficients. Varying λ_{src} mainly changes the learning trajectory, while the final performance remains relatively close on Walker2d-Morph and competitive across most settings on Hopper-Kinematic. The target-side coefficient can have a stronger effect: on Hopper-Kinematic, $\lambda_{\mathrm{tar}} = 0.05$ yields a clear performance advantage, whereas the curves remain more closely grouped on Walker2d-Morph despite larger training variance. Overall, the default choice $\lambda_{\mathrm{src}} = \lambda_{\mathrm{tar}} = 0.05$ provides consistently competitive performance across the evaluated tasks without task-specific tuning.

6 Conclusion and Limitations

In this paper, we revisit source-data reuse in cross-domain offline RL and show that the commonly observed trade-off between preserving state-action coverage and avoiding dynamics-induced bias is structural to filtering-based reuse. Even under ideal condition alignment, filtering cannot in general eliminate the residual mismatch in conditional dynamics. Motivated by this observation, we propose T-RAFT, which enlarges the admissible transition family by decoupling where source experience is reused from what outcome is propagated. T-RAFT uses unbalanced optimal transport to identify target-supported source conditions and conditional Flow Matching to reconstruct their next-state outcomes toward target-compatible dynamics. Experiments across MuJoCo locomotion and Adroit manipulation tasks demonstrate that this reconstruction-based view provides an effective alternative to selective source reuse under diverse dynamics shifts.

Our framework also has several limitations. First, reliable outcome reconstruction requires sufficient local target support. Under severe support mismatch or extremely scarce target data, the learned correspondences and translated outcomes may become inaccurate. Second, our evaluation is restricted to simulated control benchmarks, and extending transition reconstruction to more complex or real-world dynamics shifts remains an important direction for future work.

Reproducibility Statement

We provide complete theoretical derivations, algorithmic details, implementation settings, hyperparameters, compute infrastructure, and experimental protocols in the main paper and appendices. All reported results are obtained under the specified settings and averaged over multiple random seeds to facilitate reproducibility.

AI Use Statement

Generative AI tools were used only as writing assistants to improve the clarity, grammar, organization, and presentation of the manuscript. They were not used as an autonomous source for literature search, experimental results, or unverified scientific claims. All technical content, citations, theoretical derivations, experimental results, and conclusions were independently checked and finalized by the authors, who take full responsibility for the content of this paper.

References

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proceedings of the International Conference on Learning Representations*, 2023.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447, 2021.

Jongseong Chae, Jongeui Park, Yongjae Shin, Gyeongmin Kim, Seungyul Han, and Youngchul Sung. Flow actor-critic for offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2026.

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 2903–2913, 2020.

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 *IEEE International Conference on Robotics and Automation*, pp. 8973–8979, 2019.

Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. Offline reinforcement learning via high-fidelity generative behavior modeling. In *Proceedings of the International Conference on Learning Representations*, 2023a.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021.

Wenyu Chen, Yujia Zhang, Wei Guo, Linli Ma, Yanbo Wang, Pinle Qin, and Jianchao Zeng. Transport or discard: Robust unbalanced optimal transport for cross-domain policy adaptation. In *Proceedings of the International Conference on Machine Learning*, 2026.

Xingyu Chen, Jiayi Xie, Zhijian Xu, Ruixun Liu, Shuai Yang, Zeyang Liu, Lipeng Wan, and Xuguang Lan. State revisit and re-explore: Bridging sim-to-real gaps in offline-and-online reinforcement learning with an imperfect simulator. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 8724–8732, 2025.

Xiong-Hui Chen, Yang Yu, Zhengmao Zhu, ZhiHua Yu, Chen Zhenjun, Chenghe Wang, Yinan Wu, Rong-Jun Qin, Hongqiu Wu, Ruijin Ding, and Huang Fangsheng. Adversarial counterfactual environment model learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 70654–70706, 2023b.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Ignasi Clavera, Anusha Nagabandi, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 26, 2013.

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

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.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the International Conference on Machine Learning*, pp. 2052–2062, 2019.

Kaiyang Guo, Shao Yunfeng, and Yanhui Geng. Model-based offline reinforcement learning with pessimism-modulated dynamics belief. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 449–461, 2022.

Yihong Guo, Yu Yang, Pan Xu, and Anqi Liu. MOBODY: Model-based off-dynamics offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2026.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning*, pp. 1861–1870, 2018.

Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 64896–64917, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 1273–1286, 2021.

Zhengyao Jiang, Tianjun Zhang, Michael Janner, Yueying Li, Tim Rocktäschel, Edward Grefenstette, and Yuandong Tian. Efficient planning in a compact latent action space. In *Proceedings of the International Conference on Learning Representations*, 2023.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823, 2020.

Lingkai Kong, Haichuan Wang, Tonghan Wang, Guojun Xiong, and Milind Tambe. Composite flow matching for reinforcement learning with shifted-dynamics data. In *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Linh Le Pham Van, Hung The Tran, and Sunil Gupta. Policy learning for off-dynamics rl with deficient support. In *International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1093–1100, 2024.

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 *European Conference on Artificial Intelligence*, volume 413, pp. 2331–2338, 2025.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations*, 2023.

Jinxin Liu, Zhang Hongyin, and Donglin Wang. DARA: Dynamics-aware reward augmentation in offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

Jinxin Liu, Li He, Yachen Kang, Zifeng Zhuang, Donglin Wang, and Huazhe Xu. Ceil: Generalized contextual imitation learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 75491–75516, 2023a.

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.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the International Conference on Learning Representations*, 2023b.

Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 1711–1724, 2022.

Jiafei Lyu, Chenjia Bai, Jingwen Yang, Zongqing Lu, and Xiu Li. Cross-domain policy adaptation by capturing representation mismatch. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 33638–33663, 2024a.

Jiafei Lyu, Kang Xu, Jiacheng Xu, Jing-Wen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, Xiu Li, et al. Odrl: A benchmark for off-dynamics reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 59859–59911, 2024b.

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 *Proceedings of the International Conference on Learning Representations*, 2025.

Alexander Nikulin, Vladislav Kurenkov, Denis Tarasov, and Sergey Kolesnikov. Anti-exploration by random network distillation. In *Proceedings of the International Conference on Machine Learning*, pp. 26228–26244, 2023.

Haoyi Niu, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan, et al. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 36599–36612, 2022.

Haoyi Niu, Jianming Hu, Guyue Zhou, and Xianyuan Zhan. A comprehensive survey of cross-domain policy transfer for embodied agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 8197–8206, 2024.

Haoyi Niu, Tianying Ji, Bingqi Liu, Haocheng Zhao, Xiangyu Zhu, Jianying Zheng, Pengfei Huang, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. H2o+: An improved framework for hybrid offline-and-online rl with dynamics gaps. In *IEEE International Conference on Robotics and Automation*, pp. 1421–1428, 2025.

Haoyi Niu, Qimao Chen, Tenglong Liu, Jianxiong Li, Guyue Zhou, Yi Zhang, Jianming Hu, and Xianyuan Zhan. xted: Cross-domain adaptation via diffusion-based trajectory editing. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 3229–3231, 2026.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation*, pp. 3803–3810, 2018.

Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257, 2023.

Zhongjian Qiao, Jiafei Lyu, Kechen Jiao, Qi Liu, and Xiu Li. Sumo: Search-based uncertainty estimation for model-based offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20033–20041, 2025.

Zhongjian Qiao, Jiafei Lyu, Chenjia Bai, Peisong Wang, Siyang Gao, and Shuang Qiu. Unifying value alignment and assignment in cross-domain offline reinforcement learning with heterogeneous datasets. In *Proceedings of the International Conference on Machine Learning*, 2026a.

Zhongjian Qiao, Jiafei Lyu, Boxiang Lyu, Yao Shu, Siyang Gao, and Shuang Qiu. Model-based offline RL via robust value-aware model learning with implicitly differentiable adaptive weighting. In *Proceedings of the International Conference on Learning Representations*, 2026b.

Zhongjian Qiao, Rui Yang, Jiafei Lyu, Chenjia Bai, Xiu Li, Zhuoran Yang, Siyang Gao, and Shuang Qiu. Cross-domain offline policy adaptation with dynamics-and value-aligned data filtering. *arXiv preprint arXiv:2005.01643*, 2026c.

Zhongjian Qiao, Rui Yang, Jiafei Lyu, Xiu Li, Zhongxiang Dai, Zhuoran Yang, Siyang Gao, and Shuang Qiu. Dual-robust cross-domain offline reinforcement learning against dynamics shifts. In *Proceedings of the International Conference on Learning Representations*, 2026d.

Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu. Policy regularization with dataset constraint for offline reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 28701–28717, 2023.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 16082–16097, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations*, 2021.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 11592–11620, 2023.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ruhan Wang, Yu Yang, Zhishuai Liu, Dongruo Zhou, and Pan Xu. Return augmented decision transformer for off-dynamics reinforcement learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

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 International Conference on Machine Learning*, volume 235, pp. 52720–52743, 2024.

Jialong Wu, Haixu Wu, Zihan Qiu, Jianmin Wang, and Mingsheng Long. Supported policy optimization for offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 31278–31291, 2022.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M. Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 11319–11328, 2021.

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 *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 32926–32937, 2023.

Junghyuk Yeom, Yonghyeon Jo, Jeongmo Kim, Sanghyeon Lee, and Seungyul Han. Exclusively penalized q-learning for offline reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 113405–113435, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142, 2020.

Yujia Zhang, Lin Li, Wei Wei, Jianguo Wu, Yi Ma, and Jiye Liang. Efficient offline reinforcement learning via peer-influenced constraint. In *Proceedings of the International Conference on Learning Representations*, 2026a.

Yujia Zhang, Lin Li, Jianguo Wu, Ting Guo, Wei Wei, and Jiye Liang. Trajectory alignment via diffusion models in cross-domain offline reinforcement learning. *Frontiers of Computer Science*, 2026b.

Hongling Zheng, Li Shen, Yong Luo, Deheng Ye, Shuhan Xu, Bo Du, Jialie Shen, and Dacheng Tao. Value-guided decision transformer: A unified reinforcement learning framework for online and offline settings. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 38, pp. 71655–71684, 2025.

A Related Work

Offline Reinforcement Learning. Offline reinforcement learning (RL) aims to learn effective policies solely from a fixed dataset without further environment interaction (Levine et al., 2020). A central difficulty is distributional shift: value estimates or learned policies may exploit out-of-distribution actions that are poorly supported by the offline data. Existing approaches can be broadly grouped into model-free, model-based, and trajectory-level methods. Model-free methods typically mitigate this issue either through value pessimism, which penalizes or regularizes value estimates for unsupported actions (Kumar et al., 2020; An et al., 2021; Wu et al., 2021; Lyu et al., 2022; Nikulin et al., 2023; Yeom et al., 2024), or through implicit or explicit policy constraints that restrict policy improvement toward the behavior-data support (Fujimoto et al., 2019; Fujimoto & Gu, 2021; Wu et al., 2022; Tarasov et al., 2023; Ran et al., 2023; Zhang et al., 2026a). Model-based methods instead learn an offline dynamics model and control model exploitation through uncertainty-aware pessimism, conservative rollout, or model-generated data augmentation (Yu et al., 2020; Kidambi et al., 2020; Rigter et al., 2022; Guo et al., 2022; Chen et al., 2023b; Qiao et al., 2025; 2026b). A complementary line of work formulates offline RL at the trajectory level, casting policy learning as sequence modeling or return-conditioned supervised learning rather than explicit step-wise policy improvement (Chen et al., 2021; Janner et al., 2021; Jiang et al., 2023; Zheng et al., 2025). While these methods have substantially advanced single-domain offline RL, they generally assume that the available dataset is generated under the same dynamics as the deployment environment.

Cross-Domain Reinforcement Learning. Cross-domain reinforcement learning studies policy transfer across environments with mismatched observations, embodiments, or transition dynamics (Niu et al., 2024). Earlier studies on dynamics adaptation mainly rely on system identification (Chebotar et al., 2019), domain randomization (Peng et al., 2018), meta-RL (Clavera et al., 2019), or imitation-based transfer (Liu et al., 2023a), often assuming additional interaction or prior knowledge about the target environment. More generally, existing methods can be categorized by whether the source and target domains are available online or only through fixed datasets.

In the online-online setting, interaction is available in both domains, allowing agents to directly identify dynamics differences and adapt policies through additional experience (Le Pham Van et al., 2024; Lyu et al., 2024a). The online-offline setting assumes an interactive source-domain, typically a simulator, together with a fixed target dataset, and representative approaches exploit the online source environment while regularizing learning toward limited target-domain data (Niu et al., 2022; 2025; Chen et al., 2025). Conversely, offline-online methods start from previously collected source data and permit limited interaction with the target environment, thereby adapting an offline-pretrained policy using a small number of target rollouts (Kong et al., 2025). The most restrictive offline-offline setting assumes that both domains are represented only by static datasets. Existing methods address the resulting dynamics mismatch through reward correction, support constraints, representation learning, or nonparametric discrepancy measures (Qiao et al., 2026c; Chen et al., 2026; Qiao et al., 2026a). More recently, DROCO (Qiao et al., 2026d) further studies dual robustness in this setting to handle both train-time source-target mismatch and test-time dynamics perturbations.

Generative Models for Cross-Domain Reinforcement Learning. Generative models have recently become an important tool for offline RL due to their ability to represent complex and multi-modal distributions. Diffusion models (Ho et al., 2020; Song et al., 2021) and Flow Matching (Liu et al., 2023b; Tong et al., 2024), in particular, have been explored for expressive policy parameterization (Wang et al., 2023; Chae et al., 2026), trajectory generation (He et al., 2023), planning (Chen et al., 2023a), and offline data augmentation, providing an alternative to conventional unimodal policies and one-step dynamics models. More recently, generative modeling has been extended to cross-domain settings to directly adapt mismatched experience at the data level. xTED (Niu et al., 2026) uses a target-trained diffusion model to edit complete source trajectories through partial noising and denoising, thereby aligning their states, actions, rewards, and dynamics with target-domain characteristics. DmC (Le Pham Van et al., 2025) instead estimates source-target proximity with a non-parametric k-NN score and uses it to guide a diffusion model toward generating source samples that are more compatible with the target domain. In a hybrid offline-to-online setting, COMPFLOW (Kong et al., 2025) employs composite Flow Matching to adapt an offline transition model toward the online dynamics and uses the induced Wasserstein distance to quantify the dynamics gap.

B Proofs of Theorem 3.1

B.1 Auxiliary Results

For a finite signed measure σ, let $\|\sigma\|_{\mathrm{var}}$ denote its total variation norm, such that $D_{\mathrm{TV}}(P,Q) = \frac{1}{2}\|P-Q\|_{\mathrm{var}}$ for probability measures P and Q.

Lemma B.1 (Shared-marginal decomposition). *Let μ be a probability measure on $\mathcal{X}$ and let $K(\cdot \mid x)$ and $L(\cdot \mid x)$ be probability kernels on $\mathcal{Y}$. Define $P(dx,dy) = \mu(dx)K(dy \mid x)$, and $Q(dx,dy) = \mu(dx)L(dy \mid x)$. Then*

$$D_{\mathrm{TV}}(P,Q) = \mathbb{E}_{x\sim\mu}\left[D_{\mathrm{TV}}\big(K(\cdot \mid x), L(\cdot \mid x)\big)\right]. \tag{19}$$

Proof.

$$\begin{aligned} D_{\mathrm{TV}}(P,Q) &= \frac{1}{2}\|P-Q\|_{\mathrm{var}} \\ &= \frac{1}{2}\int_{\mathcal{X}} \|K(\cdot \mid x) - L(\cdot \mid x)\|_{\mathrm{var}}\, \mu(dx) \\ &= \int_{\mathcal{X}} D_{\mathrm{TV}}\big(K(\cdot \mid x), L(\cdot \mid x)\big)\mu(dx). \end{aligned} \tag{20}$$

□

Lemma B.2 (Shared-kernel identity). *Let μ and ν be probability measures on $\mathcal{X}$ and $K(\cdot \mid x)$ be a probability kernel on $\mathcal{Y}$. Define $P(dx,dy) = \mu(dx)K(dy \mid x)$, and $Q(dx,dy) = \nu(dx)K(dy \mid x)$. Then*

$$D_{\mathrm{TV}}(P,Q) = D_{\mathrm{TV}}(\mu,\nu). \tag{21}$$

Proof. For $A \in \mathcal{X}\otimes\mathcal{Y}$, let $A_x := \{y : (x,y) \in A\}$. Then

$$\begin{aligned} D_{\mathrm{TV}}(P,Q) &= \sup_A \left|\int_{\mathcal{X}} K(A_x \mid x)(\mu-\nu)(dx)\right| \\ &\le D_{\mathrm{TV}}(\mu,\nu), \end{aligned} \tag{22}$$

and

$$\begin{aligned} D_{\mathrm{TV}}(P,Q) &\ge \sup_{B\in\mathcal{X}} |P(B\times\mathcal{Y}) - Q(B\times\mathcal{Y})| \\ &= \sup_{B\in\mathcal{X}} |\mu(B) - \nu(B)| \\ &= D_{\mathrm{TV}}(\mu,\nu). \end{aligned} \tag{23}$$

Combining Equation (22) and Equation (23) gives Equation (21). □

B.2 Proof of Theorem 3.1

Proof. Define $\bar{P}(dx,dy) := \mu_{\mathrm{tar}}(dx)P_{\mathrm{src}}(dy \mid x)$. Then

$$\begin{aligned} D_{\mathrm{TV}}(P_{\mathrm{tar}},P_w) &\ge D_{\mathrm{TV}}(P_{\mathrm{tar}},\bar{P}) - D_{\mathrm{TV}}(\bar{P},P_w) \\ &= \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[D_{\mathrm{TV}}\big(P_{\mathrm{tar}}(\cdot \mid x), P_{\mathrm{src}}(\cdot \mid x)\big)\right] - D_{\mathrm{TV}}(\mu_{\mathrm{tar}},\mu_w), \end{aligned} \tag{24}$$

where the second equality follows from Lemmas B.1 and B.2. Moreover, when $\mu_w = \mu_{\mathrm{tar}}$ we have:

$$D_{\mathrm{TV}}(P_{\mathrm{tar}},P_w) = \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[D_{\mathrm{TV}}\big(P_{\mathrm{tar}}(\cdot \mid x), P_{\mathrm{src}}(\cdot \mid x)\big)\right]. \tag{25}$$

□

Remark B.1. *The intermediate distribution $\bar{P}$ isolates the two sources of joint transition mismatch: its condition marginal is identical to that of the target distribution, while its conditional dynamics remain those of the source-domain. Equation (24) therefore separates the residual conditional*

dynamics mismatch from the error due to imperfect condition alignment. In particular, Equation (25) shows that even oracle condition-level reweighting cannot eliminate the former unless $P_{\mathrm{src}}(\cdot \mid x) = P_{\mathrm{tar}}(\cdot \mid x)$ for μ_{tar}-almost every x.

More generally, for a nonnegative transition-level filter $h(x, y)$, the induced conditional distribution satisfies

$$P_h(dy \mid x) = \frac{h(x,y)P_{\mathrm{src}}(dy \mid x)}{\int h(x,\tilde{y})P_{\mathrm{src}}(d\tilde{y} \mid x)} \ll P_{\mathrm{src}}(dy \mid x), \tag{26}$$

whenever the denominator is positive. Hence, transition-level filtering can redistribute probability mass over source-generated outcomes, but cannot create target outcomes outside the conditional support of the source dynamics. This motivates introducing an additional degree of freedom that explicitly transforms the outcome associated with a retained condition.

C Proof of Proposition 4.1

C.1 Basic Assumptions

Assumption C.1 (Regularity of the transition space and translated kernel)**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be Polish metric spaces, and equip $\mathcal{X} \times \mathcal{Y}$ with the additive product metric*

$$d\big((x,y),(x',y')\big) := d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y'). \tag{27}$$

All distributions considered below have finite first moments. Moreover, the translated conditional kernel K_θ satisfies

$$W_1^{\mathcal{Y}}\big(K_\theta(\cdot \mid x), K_\theta(\cdot \mid x')\big) \le L_K d_{\mathcal{X}}(x,x'), \qquad \forall x, x' \in \mathcal{X}, \tag{28}$$

for some $L_K < \infty$.

C.2 Proof of Proposition 4.1

Proof. Define $P_\theta^\circ(dx, dy) := \mu_{\mathrm{tar}}(dx)K_\theta(dy \mid x)$. Then

$$W_1\big(P_{\mathrm{tar}}, \widetilde{P}_{w,\theta}\big) \le W_1\big(P_{\mathrm{tar}}, P_\theta^\circ\big) + W_1\big(P_\theta^\circ, \widetilde{P}_{w,\theta}\big). \tag{29}$$

For $\varepsilon > 0$, let $\gamma_x^\varepsilon \in \Pi\big(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x)\big)$ be a coupling that satisfies

$$\int_{\mathcal{Y}\times\mathcal{Y}} d_{\mathcal{Y}}(y,\bar{y})\gamma_x^\varepsilon(dy, d\bar{y}) \le W_1^{\mathcal{Y}}\big(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x)\big) + \varepsilon. \tag{30}$$

Hence

$$\begin{aligned} W_1\big(P_{\mathrm{tar}}, P_\theta^\circ\big) &\le \int_{\mathcal{X}} \int_{\mathcal{Y}\times\mathcal{Y}} d_{\mathcal{Y}}(y,\bar{y})\gamma_x^\varepsilon(dy, d\bar{y})\mu_{\mathrm{tar}}(dx) \\ &\le \int_{\mathcal{X}} W_1^{\mathcal{Y}}\big(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x)\big)\mu_{\mathrm{tar}}(dx) + \varepsilon. \end{aligned} \tag{31}$$

Letting $\varepsilon \to 0$ gives

$$W_1\big(P_{\mathrm{tar}}, P_\theta^\circ\big) \le \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[W_1^{\mathcal{Y}}\big(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x)\big)\right]. \tag{32}$$

For any $\pi \in \Pi(\mu_{\mathrm{tar}}, \mu_w)$ and $\varepsilon > 0$, let

$$\gamma_{x,x'}^\varepsilon \in \Pi\big(K_\theta(\cdot \mid x), K_\theta(\cdot \mid x')\big) \tag{33}$$

satisfy

$$\int_{\mathcal{Y}\times\mathcal{Y}} d_{\mathcal{Y}}(y,y')\gamma_{x,x'}^\varepsilon(dy, dy') \le W_1^{\mathcal{Y}}\big(K_\theta(\cdot \mid x), K_\theta(\cdot \mid x')\big) + \varepsilon. \tag{34}$$

Thus

$$\begin{aligned} W_1\big(P_\theta^\circ, \widetilde{P}_{w,\theta}\big) &\le \int_{\mathcal{X}\times\mathcal{X}} \Big[d_{\mathcal{X}}(x,x') + W_1^{\mathcal{Y}}\big(K_\theta(\cdot \mid x), K_\theta(\cdot \mid x')\big) + \varepsilon\Big]\pi(dx, dx') \\ &\le (1 + L_K)\int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')\pi(dx, dx') + \varepsilon. \end{aligned} \tag{35}$$

Taking the infimum over $\pi \in \Pi(\mu_{\mathrm{tar}}, \mu_w)$ and letting $\varepsilon \to 0$ yields

$$W_1\big(P^\circ_\theta, \widetilde{P}_{w,\theta}\big) \le (1 + L_K) W_1^{\mathcal{X}}\big(\mu_{\mathrm{tar}}, \mu_w\big). \tag{36}$$

Combining Equation (29), (32), and (36), gives

$$W_1\big(P_{\mathrm{tar}}, \widetilde{P}_{w,\theta}\big) \le C_K W_1^{\mathcal{X}}\big(\mu_{\mathrm{tar}}, \mu_w\big) + \mathbb{E}_{x\sim\mu_{\mathrm{tar}}}\left[W_1^{\mathcal{Y}}\big(P_{\mathrm{tar}}(\cdot \mid x), K_\theta(\cdot \mid x)\big)\right]. \tag{37}$$

where $C_K := 1 + L_K$.

□

Remark C.1. *Proposition 4.1 separates the joint transition-reconstruction error into two sources. The first is the discrepancy between the retained condition marginal μ_w and the target marginal μ_{tar}, where the factor $1 + L_K$ accounts for the additional change in translated outcomes caused by moving between nearby conditions. The second is the residual discrepancy between the translated kernel $K_\theta(\cdot \mid x)$ and the target conditional dynamics $P_{\mathrm{tar}}(\cdot \mid x)$ at a fixed condition. Hence, μ_w and K_θ constitute two distinct degrees of freedom: the former determines* where *source experience is retained, while the latter determines* what outcome *is propagated there. If $K_\theta = P_{\mathrm{src}}$, the enlarged family reduces to condition-level reweighting and the second term becomes the Wasserstein counterpart of the irreducible conditional mismatch in Theorem 3.1.*

D Proof of Proposition 4.2

D.1 Basic Assumptions

Assumption D.1 (Metric and transport cost)**.** *The condition and outcome spaces are Euclidean, and the joint transition space is equipped with*

$$d\big((x,y),(x',y')\big) = \|x - x'\|_2 + \|y - y'\|_2. \tag{38}$$

The UOT cost satisfies $C_{ij} = \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2^2$.

Assumption D.2 (Velocity-field regularity)**.** *For every condition x, the ODE induced by $v_\theta(\cdot, t; x)$ admits a unique solution on $[0,1]$, and there exists $L_v < \infty$ such that*

$$\|v_\theta(y,t;x) - v_\theta(\tilde{y},t;x)\|_2 \le L_v \|y - \tilde{y}\|_2 \tag{39}$$

for all $y, \tilde{y}$, $t \in [0,1]$, and x.

Assumption D.3 (Target empirical support)**.** *Let $\boldsymbol{\beta} \in \Delta^{N_{\mathrm{tar}}}$ with $\beta_j > 0$. The target empirical transition support has finite diameter*

$$D_{\mathrm{tar}} := \max_{j,k} d\big((x_j^{\mathrm{tar}}, y_j^{\mathrm{tar}}), (x_k^{\mathrm{tar}}, y_k^{\mathrm{tar}})\big) < \infty. \tag{40}$$

D.2 Proof of Proposition 4.2

Proof. Let $m := \Gamma^\star \mathbf{1}$, $n := (\Gamma^\star)^\top \mathbf{1}$, $M := \|\Gamma^\star\|_1 > 0$, and $\bar{m}_i := \frac{m_i}{M}$, $\bar{n}_j := \frac{n_j}{M}$, $q_{ij} := \frac{\Gamma^\star_{ij}}{M}$ with $\sum_j q_{ij} = \bar{m}_i$, $\sum_i q_{ij} = \bar{n}_j$, $\sum_{i,j} q_{ij} = 1$. Further denote $\widehat{P}_\theta = \sum_i \bar{m}_i \delta_{(x_i^{\mathrm{src}}, \bar{y}_i)}$, $\widehat{P}^\Gamma_{\mathrm{tar}} = \sum_j \bar{n}_j \delta_{(x_j^{\mathrm{tar}}, y_j^{\mathrm{tar}})}$, and $\widehat{P}_{\mathrm{tar}} = \sum_j \beta_j \delta_{(x_j^{\mathrm{tar}}, y_j^{\mathrm{tar}})}$. Then for each (i,j), let

$$\Delta_{ij} := y_j^{\mathrm{tar}} - y_i^{\mathrm{src}}, \qquad y_t^{ij} := (1-t) y_i^{\mathrm{src}} + t y_j^{\mathrm{tar}}. \tag{41}$$

Let $\dot{u}_i(t) = v_\theta(u_i(t), t; x_i^{\mathrm{src}})$, where $u_i(0) = y_i^{\mathrm{src}}$, $\bar{y}_i = u_i(1)$, and $e_{ij}(t) := u_i(t) - y_t^{ij}$. Then

$$\begin{aligned} \dot{e}_{ij}(t) &= v_\theta(u_i(t), t; x_i^{\mathrm{src}}) - \Delta_{ij} \\ &= v_\theta(u_i(t), t; x_i^{\mathrm{src}}) - v_\theta(y_t^{ij}, t; x_i^{\mathrm{src}}) + v_\theta(y_t^{ij}, t; x_i^{\mathrm{src}}) - \Delta_{ij}, \end{aligned} \tag{42}$$

and hence

$$\begin{aligned} \|e_{ij}(1)\|_2 &\le \int_0^1 e^{L_v(1-t)} \left\| v_\theta(y_t^{ij}, t; x_i^{\mathrm{src}}) - \Delta_{ij} \right\|_2 dt \\ &\le \left(\int_0^1 e^{2L_v(1-t)} dt \right)^{1/2} \left(\int_0^1 \left\| v_\theta(y_t^{ij}, t; x_i^{\mathrm{src}}) - \Delta_{ij} \right\|_2^2 dt \right)^{1/2}. \end{aligned} \tag{43}$$

Further define

$$\kappa(L_v) := \begin{cases} \dfrac{e^{2L_v}-1}{2L_v}, & L_v > 0, \\ 1, & L_v = 0, \end{cases} \qquad C_{\mathrm{FM}} := \sqrt{\kappa(L_v)}. \tag{44}$$

Then

$$\begin{aligned} \sum_{i,j} q_{ij} \|\bar{y}_i - y_j^{\mathrm{tar}}\|_2^2 &= \sum_{i,j} q_{ij} \|e_{ij}(1)\|_2^2 \\ &\le \kappa(L_v) \sum_{i,j} q_{ij} \int_0^1 \left\| v_\theta(y_t^{ij}, t; x_i^{\mathrm{src}}) - \Delta_{ij} \right\|_2^2 dt \\ &= \kappa(L_v) \mathcal{L}_{\mathrm{FM}}(\theta). \end{aligned} \tag{45}$$

Therefore,

$$\begin{aligned} \sum_{i,j} q_{ij} \|\bar{y}_i - y_j^{\mathrm{tar}}\|_2 &\le \left(\sum_{i,j} q_{ij} \|\bar{y}_i - y_j^{\mathrm{tar}}\|_2^2 \right)^{1/2} \\ &\le C_{\mathrm{FM}} \sqrt{\mathcal{L}_{\mathrm{FM}}(\theta)}. \end{aligned} \tag{46}$$

Moreover,

$$\begin{aligned} \sum_{i,j} q_{ij} \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2 &\le \left(\sum_{i,j} q_{ij} \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2^2 \right)^{1/2} \\ &= \sqrt{\frac{\langle \Gamma^\star, C \rangle}{M}}. \end{aligned} \tag{47}$$

Since q is a coupling between $\widehat{P}_\theta$ and $\widehat{P}_{\mathrm{tar}}^{\Gamma}$,

$$\begin{aligned} W_1\left(\widehat{P}_\theta, \widehat{P}_{\mathrm{tar}}^{\Gamma}\right) &\le \sum_{i,j} q_{ij} d\big((x_i^{\mathrm{src}}, \bar{y}_i), (x_j^{\mathrm{tar}}, y_j^{\mathrm{tar}})\big) \\ &= \sum_{i,j} q_{ij} \|x_i^{\mathrm{src}} - x_j^{\mathrm{tar}}\|_2 + \sum_{i,j} q_{ij} \|\bar{y}_i - y_j^{\mathrm{tar}}\|_2 \\ &\le \sqrt{\frac{\langle \Gamma^\star, C \rangle}{M}} + C_{\mathrm{FM}} \sqrt{\mathcal{L}_{\mathrm{FM}}(\theta)}. \end{aligned} \tag{48}$$

Furthermore,

$$\begin{aligned} W_1\left(\widehat{P}_{\mathrm{tar}}^{\Gamma}, \widehat{P}_{\mathrm{tar}}\right) &\le D_{\mathrm{tar}} D_{\mathrm{TV}}(\bar{n}, \beta) \\ &\le D_{\mathrm{tar}} \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\bar{n} \| \beta)}. \end{aligned} \tag{49}$$

Using $n = M\bar{n}$ and $\sum_j \beta_j = 1$,

$$\begin{aligned} D_{\mathrm{KL}}^g(n \| \beta) &= \sum_j \left[n_j \log \frac{n_j}{\beta_j} - n_j + \beta_j \right] \\ &= M \sum_j \bar{n}_j \log \frac{\bar{n}_j}{\beta_j} + M \log M - M + 1 \\ &= M D_{\mathrm{KL}}(\bar{n} \| \beta) + M \log M - M + 1. \end{aligned} \tag{50}$$

Since $M \log M - M + 1 \ge 0$, we obtain

$$D_{\mathrm{KL}}(\bar{n} \| \beta) \le \frac{D_{\mathrm{KL}}^g(n \| \beta)}{M}. \tag{51}$$

Thus,

$$W_1\left(\widehat{P}^{\Gamma}_{\mathrm{tar}}, \widehat{P}_{\mathrm{tar}}\right) \le \frac{D_{\mathrm{tar}}}{\sqrt{2}}\sqrt{\frac{D^g_{\mathrm{KL}}(n\|\beta)}{M}}. \tag{52}$$

Finally, let $C_{\mathrm{tar}} := \frac{D_{\mathrm{tar}}}{\sqrt{2}}$. Then,

$$\begin{aligned} W_1\left(\widehat{P}_{\theta}, \widehat{P}_{\mathrm{tar}}\right) &\le W_1\left(\widehat{P}_{\theta}, \widehat{P}^{\Gamma}_{\mathrm{tar}}\right) + W_1\left(\widehat{P}^{\Gamma}_{\mathrm{tar}}, \widehat{P}_{\mathrm{tar}}\right) \\ &\le \sqrt{\frac{\langle \Gamma^\star, C\rangle}{M}} + C_{\mathrm{FM}}\sqrt{\mathcal{L}_{\mathrm{FM}}(\theta)} + C_{\mathrm{tar}}\sqrt{\frac{D^g_{\mathrm{KL}}\left((\Gamma^\star)^\top \mathbf{1} \,\|\, \beta\right)}{M}}, \end{aligned} \tag{53}$$

which gives Equation (14). □

Remark D.1. *Proposition 4.2 gives an empirical counterpart of the condition–outcome decomposition in Proposition 4.1. The first term measures the geometry of the source-target condition alignment induced by the normalized UOT coupling, while the second converts the Flow Matching regression error into the resulting endpoint translation error. The third term accounts for the discrepancy between the target marginal induced by the relaxed UOT coupling and the empirical target distribution. Hence, jointly reducing the UOT alignment cost, target-marginal relaxation, and Flow Matching loss controls the discrepancy between the reconstructed and target transition distributions. The source-side generalized KL term does not appear explicitly in the bound, but influences $\Gamma^\star$, the retained source marginal, and the transported mass M, thereby regulating source-mass removal.*

E Implementation Details

In this part, we describe the implementation details of our proposed T-RAFT algorithm. Furthermore, we list the hyperparameter setup used for it.

E.1 Algorithm Details of T-RAFT

Implementation of UOT. The UOT alignment is computed once before Flow Matching and policy optimization, since both the source and target datasets are fixed in the offline setting. The resulting transport statistics are cached and shared across all five policy-learning seeds of the same source-target dataset pair, avoiding repeated transport computation.

Before constructing the transport problem, we jointly standardize the source and target states and actions using their empirical means and standard deviations. We represent each condition by $c = [\widetilde{s}, \sqrt{\lambda_a}\widetilde{a}]$, where $\widetilde{s}$ and $\widetilde{a}$ denote the standardized state and action, respectively, and set $\lambda_a = 1$ in all experiments. To avoid explicitly constructing the full $N_{\mathrm{src}} \times N_{\mathrm{tar}}$ cost matrix, which can be prohibitively large for D4RL datasets, we use a local sparse approximation. Specifically, the source dataset is processed in blocks of 4096 transitions, and for each block we use at most 4096 target transitions as reference samples. A k-d tree is built in the normalized condition space, and each source condition is connected only to its $k_{\mathrm{NN}} = 32$ nearest target conditions. The UOT problem is then solved over this sparse support.

We use the entropy-regularized Sinkhorn solver implemented in OTT-JAX. The entropic regularization coefficient is set to $\epsilon = 0.01$, while the source- and target-side marginal relaxation coefficients are set to $\lambda_{\mathrm{src}} = \lambda_{\mathrm{tar}} = 0.05$ by default. Sinkhorn iterations terminate when the residual falls below 10^{-5} or after 1000 iterations. For memory efficiency, entries outside the local support are masked with a sufficiently large cost, and the computation is performed independently for each source block.

After solving UOT, we cache two quantities required by the subsequent stages. First, the row marginal $m_i = (\Gamma^\star \mathbf{1})_i$ records the transported mass associated with each source condition and is normalized when constructing the source reuse distribution. Second, we retain the largest coupling entry for each source condition and rescale the retained entry to preserve its row mass. This sparse representation substantially reduces storage while retaining the UOT-induced source mass and local source-target correspondence used for Flow Matching. All UOT preprocessing is performed with a fixed preprocessing seed and is reused across RL seeds.

Table 3: Hyperparameter setup for T-RAFT.

Hyperparameter	**Value**
Shared / Offline RL	
Actor network	$(256, 256)$
Critic network	$(256, 256)$
Value network	$(256, 256)$
Learning rate	3×10^{-4}
Optimizer	Adam
Discount factor γ	0.99
Target update rate	5×10^{-3}
Source-domain batch size	128
Target-domain batch size	128
Expectile parameter τ_{IQL}	0.7
Inverse temperature β_{IQL}	3.0
Policy-training steps	10^6
Evaluation interval	10^4
Evaluation episodes	10
Number of seeds	5
UOT	
Condition representation	(s, a)
State/action normalization	joint standardization
Action-distance coefficient λ_a	1.0
Entropic regularization ϵ	0.01
Source relaxation λ_{src}	0.05
Target relaxation λ_{tar}	0.05
Source block size	4096
Target reference size	4096
Mask cost	50
Maximum Sinkhorn iterations	1000
Sinkhorn tolerance	10^{-5}
UOT solver	OTT-JAX Sinkhorn
Flow Matching	
Transport variable	$\Delta s = s' - s$
Velocity network depth	4
Hidden dimension	512
Nonlinearity	SiLU
Optimizer	AdamW
Learning rate	3×10^{-4}
Training steps	100000
Batch size	1024
Time sampling	$\mathcal{U}[0, 1]$
Gradient clipping	10
ODE solver	RK4
ODE integration steps	50
Translation batch size	4096
Preprocessing seed	0

Implementation of Flow Matching. For numerical stability, we implement outcome translation in the one-step residual-dynamics space rather than directly transporting absolute next states. For each transition, we use $\Delta s = s' - s$.

Since the source state s is fixed throughout translation, a translated residual $\bar{\Delta} s$ uniquely determines the reconstructed next state as $\bar{s}' = s + \bar{\Delta} s$ and therefore realizes the same condition-preserving transformation described in Section 4.1.

We use the state, action, and residual normalization statistics computed during UOT preprocessing. Source-target endpoint pairs are sampled directly from the sparse transport cache with probability proportional to their transported mass. For a sampled pair (i, j), we form the normalized source and

target residuals $\widetilde{\Delta}s_i^{\mathrm{src}}$ and $\widetilde{\Delta}s_j^{\mathrm{tar}}$, sample $t \sim \mathcal{U}[0,1]$, and construct the linear interpolation

$$\Delta s_t = (1-t)\widetilde{\Delta}s_i^{\mathrm{src}} + t\widetilde{\Delta}s_j^{\mathrm{tar}}. \tag{54}$$

The conditional velocity field takes $[\Delta s_t, t, \widetilde{s}_i^{\mathrm{src}}, \widetilde{a}_i^{\mathrm{src}}]$ as input and predicts a velocity in the normalized residual space.

We parameterize the velocity field by a four-layer MLP with hidden width 512 and SiLU activations. The model is optimized with AdamW at a learning rate of 3×10^{-4} for 10^5 gradient steps, using a batch size of 1024. We clip the gradient norm to 10 during training. Importantly, endpoint pairs are sampled according to the UOT pair masses rather than uniformly, so the Flow Matching objective preserves the correspondence structure produced by the condition-alignment stage.

After training, the translator is fixed and applied to the source dataset offline. Starting from the normalized source residual, we integrate the learned ODE from $t = 0$ to $t = 1$ using a fourth-order Runge–Kutta solver with 50 uniform integration steps. The terminal residual is transformed back to the original state scale and combined with the original source state:

$$\bar{s}_i' = s_i^{\mathrm{src}} + \bar{\Delta}s_i. \tag{55}$$

Translation is performed in batches of 4096 transitions. The resulting reconstructed dataset $\widetilde{\mathcal{D}}_{\mathrm{src}} = \{(s_i^{\mathrm{src}}, a_i^{\mathrm{src}}, r_i^{\mathrm{src}}, \bar{s}_i')\}$ is cached before offline RL training; therefore, neither UOT nor ODE integration is repeated during policy optimization.

E.2 Hyperparameter Setup

Unless otherwise specified, we use the same T-RAFT hyperparameters across all environments, dynamics shifts, and dataset qualities. The hyperparameter configuration is summarized in Table 3. For the IQL backbone, we follow the same architecture and optimization configuration across all experiments. The UOT and Flow Matching preprocessing is shared across the five RL seeds for each source-target dataset pair.

F Detailed Experimental Setup

In this section, we provide detailed experimental setups that are omitted from the main text due to the page limit. We mainly present the experimental setups for the MuJoCo and Adroit environments used in Section 5.1. Furthermore, we introduce the evaluation protocol used to compute normalized scores.

F.1 MuJoCo Locomotion Benchmarks

Source datasets. For the MuJoCo experiments, we follow the cross-domain offline RL benchmark adopted by OTDF (Lyu et al., 2025). We consider four continuous-control environments: HalfCheetah, Hopper, Walker2d, and Ant. The source-domain corresponds to the original MuJoCo dynamics, and the source datasets are taken from the D4RL "-v2" benchmark. For each environment, we consider three source-data qualities: medium, medium-replay, and medium-expert. The medium dataset is collected by an early-stopped SAC policy (Haarnoja et al., 2018), the medium-replay dataset contains transitions from the replay buffer accumulated during SAC training, and the medium-expert dataset combines medium- and expert-level experience.

Target dynamics shifts. The target domains are constructed by modifying the physical dynamics of the corresponding MuJoCo environments while keeping the state and action spaces unchanged. We consider three complementary types of dynamics shift. For the gravity shift, the magnitude of gravitational acceleration is reduced to one half of its source-domain value while preserving its direction. For the kinematic shift, selected robot joints are constrained to substantially narrower rotation ranges, simulating partially impaired or broken joints. For the morphology shift, the geometry of selected body parts or limbs is modified. The affected joints or body components depend on the environment, following the benchmark configuration used in OTDF.

Target datasets. Target-domain datasets are collected directly in the corresponding shifted environments following the D4RL-style data collection protocol. An SAC agent is trained in each target environment, and checkpoints at different stages of training are used to construct medium, medium-expert, and expert target datasets. The expert dataset is generated from the final policy checkpoint, whereas the medium dataset uses an intermediate policy attaining approximately one-third to one-half of the expert performance. To strictly control the target-data budget, each target dataset contains only five trajectories, corresponding to approximately $5{,}000$ transitions. For the medium-expert setting, we preserve the same data budget by combining two medium trajectories and three expert trajectories rather than concatenating the full medium and expert datasets.

F.2 Adroit Manipulation Benchmarks

To evaluate whether T-RAFT remains effective beyond locomotion, we further consider the dexterous-manipulation benchmarks introduced in ODRL (Lyu et al., 2024b). Specifically, we evaluate on the Adroit Pen and Door tasks and construct target domains with two types of dynamics mismatch: kinematic and morphology shifts. Each shift type is evaluated at two difficulty levels, denoted as medium (M) and hard (H), resulting in eight Adroit configurations in total.

Dynamics shifts. The kinematic shift simulates impaired hand articulation by restricting the rotation range of selected hand joints, thereby producing a broken-joint dynamics mismatch between the source and target domains. The morphology shift modifies the physical geometry of the robotic hand by shrinking selected fingers. These modifications change the transition dynamics without altering the state or action spaces. The medium and hard settings correspond to two different magnitudes of the same physical perturbation, following the environment configurations of the ODRL (Lyu et al., 2024b) benchmark.

Offline datasets. Following MOBODY (Guo et al., 2026), we use the medium-level offline dataset collected in the unmodified Adroit environment as the source-domain dataset. The behavior policy is obtained from SAC and is selected to achieve approximately 50% of expert performance. The source dataset contains approximately one million transitions. For each shifted target environment, trajectories are collected using the corresponding behavior policy until approximately $5{,}000$ target transitions are obtained. Thus, the target dataset remains substantially smaller than the source dataset, preserving the limited-target-data regime considered throughout the paper.

F.3 Evaluation Protocol

We evaluate all methods in the corresponding target environment and report the normalized target-domain score

$$\mathrm{NS} = \frac{J - J_{\mathrm{random}}}{J_{\mathrm{expert}} - J_{\mathrm{random}}} \times 100, \tag{56}$$

where J denotes the evaluation return of the learned policy, and J_{random} and J_{expert} are obtained from a random policy and an expert policy, respectively, under the same target dynamics. We therefore use target-domain reference returns rather than the original D4RL normalization constants. All methods are trained for 10^6 gradient updates and evaluated over five random seeds.

G Wider Experimental Results

This section complements the main-text results with broader evaluations under morphology and gravity shifts on MuJoCo, followed by experiments on the Adroit dexterous-manipulation tasks. For the MuJoCo experiments, we preserve the same source-target dataset-quality combinations and evaluation protocol as in the main text. All scores are normalized target-domain returns reported as mean$\pm$standard deviation over five random seeds. We further evaluate T-RAFT on Adroit Pen and Door under kinematic and morphology perturbations to examine whether the proposed transition-reconstruction mechanism extends beyond locomotion.

Table 4: Performance comparison on MuJoCo under morphology shifts. We report normalized target-domain scores as mean±standard deviation over five seeds. Source abbreviations use *m*, *me*, and *mr* for medium, medium-expert, and medium-replay, respectively.

Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	ROOT	T-RAFT
Half-m	medium	30.0 ± 1.6	26.6 ± 3.3	19.3 ± 3.5	41.3 ± 0.4	41.6 ± 0.5	39.1 ± 2.3	$\mathbf{43.0 \pm 0.1}$	42.5 ± 0.2
Half-m	med-expert	31.8 ± 1.1	32.0 ± 0.7	33.6 ± 1.1	30.7 ± 0.8	29.6 ± 2.2	$\mathbf{35.6 \pm 0.7}$	35.5 ± 1.1	34.2 ± 1.6
Half-m	expert	8.5 ± 1.0	9.3 ± 1.6	7.9 ± 0.8	8.6 ± 0.9	10.0 ± 0.8	10.7 ± 1.2	11.4 ± 0.5	$\mathbf{11.9 \pm 1.1}$
Half-mr	medium	30.8 ± 4.4	35.6 ± 0.7	35.0 ± 4.6	32.0 ± 1.4	28.0 ± 2.0	40.0 ± 1.2	$\mathbf{41.3 \pm 0.2}$	38.3 ± 0.6
Half-mr	med-expert	12.9 ± 2.2	16.9 ± 4.1	19.9 ± 5.5	12.4 ± 1.6	12.0 ± 3.7	$\mathbf{34.4 \pm 0.7}$	30.4 ± 1.7	28.1 ± 1.4
Half-mr	expert	5.9 ± 1.7	3.7 ± 2.7	2.4 ± 1.9	6.2 ± 1.4	5.3 ± 2.3	8.2 ± 2.7	$\mathbf{8.7 \pm 2.0}$	8.2 ± 0.9
Half-me	medium	41.5 ± 0.1	40.3 ± 1.2	41.3 ± 0.3	41.3 ± 0.4	40.9 ± 0.4	41.4 ± 0.3	$\mathbf{42.7 \pm 0.2}$	41.2 ± 0.4
Half-me	med-expert	25.8 ± 2.0	30.6 ± 2.8	32.1 ± 0.8	27.2 ± 0.8	26.2 ± 1.8	$\mathbf{35.1 \pm 0.6}$	33.0 ± 0.5	34.2 ± 0.4
Half-me	expert	7.8 ± 1.3	8.3 ± 1.3	9.1 ± 0.8	7.8 ± 0.9	7.5 ± 0.9	9.8 ± 1.0	10.2 ± 1.9	$\mathbf{11.1 \pm 1.2}$
Hopp-m	medium	13.5 ± 0.2	13.5 ± 0.4	13.2 ± 0.3	13.4 ± 0.1	13.4 ± 0.2	11.0 ± 0.9	13.5 ± 0.6	$\mathbf{14.7 \pm 0.9}$
Hopp-m	med-expert	13.4 ± 0.1	13.6 ± 0.2	11.2 ± 4.6	13.3 ± 0.2	13.3 ± 0.4	12.6 ± 0.8	$\mathbf{14.0 \pm 0.3}$	13.5 ± 0.3
Hopp-m	expert	13.5 ± 0.2	13.6 ± 0.3	13.3 ± 0.4	13.6 ± 0.2	13.9 ± 0.1	10.7 ± 4.7	13.7 ± 0.5	$\mathbf{24.2 \pm 9.0}$
Hopp-mr	medium	10.8 ± 1.1	10.2 ± 1.0	1.2 ± 0.0	10.7 ± 1.6	12.0 ± 4.4	8.7 ± 2.8	17.3 ± 5.9	$\mathbf{31.1 \pm 2.4}$
Hopp-mr	med-expert	11.6 ± 1.6	10.4 ± 0.9	1.3 ± 0.2	10.4 ± 1.2	8.2 ± 2.8	9.7 ± 2.7	$\mathbf{19.1 \pm 7.5}$	14.9 ± 2.1
Hopp-mr	expert	9.8 ± 0.5	9.0 ± 0.3	1.3 ± 0.1	10.4 ± 1.4	11.4 ± 1.5	10.7 ± 2.4	14.5 ± 5.2	$\mathbf{17.5 \pm 8.8}$
Hopp-me	medium	12.6 ± 1.4	13.0 ± 0.5	15.7 ± 7.2	14.0 ± 2.3	12.7 ± 0.8	7.9 ± 3.2	$\mathbf{20.4 \pm 5.6}$	15.2 ± 0.4
Hopp-me	med-expert	14.1 ± 1.3	13.8 ± 0.6	12.0 ± 1.4	13.5 ± 0.3	13.3 ± 1.2	9.6 ± 3.5	17.4 ± 1.6	$\mathbf{20.7 \pm 6.8}$
Hopp-me	expert	13.8 ± 0.5	12.3 ± 1.8	10.5 ± 5.0	14.7 ± 2.3	12.8 ± 0.9	5.9 ± 4.0	14.9 ± 1.1	$\mathbf{16.6 \pm 1.5}$
Walk-m	medium	23.0 ± 4.7	23.3 ± 3.3	6.2 ± 2.9	24.7 ± 1.7	27.5 ± 9.5	50.5 ± 5.8	$\mathbf{57.3 \pm 4.1}$	54.4 ± 3.8
Walk-m	med-expert	21.5 ± 8.6	22.2 ± 7.6	7.2 ± 2.9	18.7 ± 7.3	20.7 ± 5.9	44.3 ± 23.8	48.1 ± 9.2	$\mathbf{49.3 \pm 6.2}$
Walk-m	expert	20.3 ± 2.8	17.3 ± 3.4	15.8 ± 8.7	21.1 ± 7.2	15.8 ± 4.5	55.3 ± 8.3	$\mathbf{58.3 \pm 10.7}$	54.6 ± 4.2
Walk-mr	medium	11.3 ± 3.0	10.9 ± 4.6	5.4 ± 4.0	10.4 ± 4.8	13.4 ± 7.2	37.4 ± 5.1	$\mathbf{40.5 \pm 3.9}$	39.3 ± 4.4
Walk-mr	med-expert	7.0 ± 1.5	4.5 ± 1.1	4.0 ± 2.2	4.9 ± 1.7	6.9 ± 2.2	$\mathbf{33.8 \pm 6.9}$	24.5 ± 3.1	23.0 ± 1.0
Walk-mr	expert	6.3 ± 0.9	4.5 ± 1.1	3.8 ± 3.4	5.5 ± 0.9	5.5 ± 2.2	$\mathbf{41.5 \pm 6.8}$	15.5 ± 1.4	25.7 ± 5.3
Walk-me	medium	24.1 ± 7.4	31.7 ± 6.6	18.7 ± 6.5	29.9 ± 4.7	27.5 ± 2.3	49.9 ± 4.6	$\mathbf{52.7 \pm 2.8}$	49.6 ± 4.2
Walk-me	med-expert	27.0 ± 5.5	23.3 ± 5.5	11.1 ± 0.9	22.9 ± 3.8	25.3 ± 6.4	40.5 ± 11.0	$\mathbf{49.6 \pm 5.8}$	38.1 ± 2.9
Walk-me	expert	22.4 ± 3.3	25.2 ± 5.7	9.9 ± 3.9	18.7 ± 5.7	24.7 ± 2.4	45.7 ± 6.9	$\mathbf{58.8 \pm 9.1}$	50.8 ± 16.9
Ant-m	medium	38.7 ± 3.8	41.3 ± 1.8	18.2 ± 1.9	40.6 ± 2.1	40.9 ± 1.7	39.4 ± 1.7	44.0 ± 0.3	$\mathbf{44.3 \pm 0.6}$
Ant-m	med-expert	47.0 ± 5.1	43.3 ± 2.0	45.3 ± 7.0	47.2 ± 4.3	44.4 ± 1.7	$\mathbf{58.3 \pm 8.9}$	55.3 ± 3.2	53.5 ± 2.8
Ant-m	expert	36.2 ± 3.5	48.5 ± 4.2	72.2 ± 10.5	42.2 ± 9.9	41.4 ± 4.2	$\mathbf{85.4 \pm 4.4}$	67.7 ± 9.1	64.7 ± 6.5
Ant-mr	medium	38.2 ± 2.9	38.9 ± 2.7	20.2 ± 3.7	38.3 ± 1.9	39.7 ± 1.2	41.2 ± 0.9	43.4 ± 0.3	$\mathbf{43.8 \pm 0.7}$
Ant-mr	med-expert	38.1 ± 3.5	33.4 ± 5.5	15.2 ± 1.6	35.0 ± 5.7	37.3 ± 2.4	50.8 ± 4.5	52.3 ± 2.3	$\mathbf{52.6 \pm 2.9}$
Ant-mr	expert	24.1 ± 1.9	24.5 ± 2.6	16.0 ± 1.7	22.7 ± 3.0	23.6 ± 1.4	$\mathbf{67.2 \pm 7.5}$	44.5 ± 4.9	33.3 ± 3.4
Ant-me	medium	32.9 ± 5.1	40.2 ± 1.5	28.1 ± 5.6	35.9 ± 2.5	36.1 ± 4.4	39.9 ± 2.9	44.0 ± 0.3	$\mathbf{44.7 \pm 0.5}$
Ant-me	med-expert	35.7 ± 3.9	36.5 ± 8.7	14.8 ± 15.9	24.5 ± 15.7	30.7 ± 10.8	$\mathbf{65.7 \pm 4.5}$	58.8 ± 3.9	63.9 ± 4.6
Ant-me	expert	36.1 ± 8.5	34.6 ± 5.8	53.9 ± 5.0	38.4 ± 9.4	35.2 ± 6.6	$\mathbf{86.4 \pm 2.2}$	67.3 ± 9.9	61.0 ± 7.2

G.1 Results under Morphology Shifts

Table 4 reports the complete results under morphology shifts. T-RAFT achieves the highest mean score in 13 of the 36 configurations, with improvements spanning all four locomotion environments. The gains are particularly pronounced on Hopper, where T-RAFT substantially improves several settings in which the source and target behavior qualities differ. For example, with medium source and expert target data, T-RAFT reaches 24.2 ± 9.0, compared with the strongest baseline score of 13.9 ± 0.1. With medium-replay source and medium target data, it further improves the best baseline from 17.3 ± 5.9 to 31.1 ± 2.4.

The results on the remaining environments are more heterogeneous. T-RAFT achieves the strongest performance on several Ant configurations and improves Walker2d with medium source and medium-expert target data, whereas OTDF or ROOT remains stronger in a number of HalfCheetah and Walker2d settings. This variation is consistent with the intended role of reconstruction: translating the outcome of a reusable source condition can be beneficial when useful source coverage coexists with substantial dynamics mismatch, but its advantage depends on how reliably the scarce target data identify the corresponding target outcomes.

G.2 Results under Gravity Shifts

Table 5 presents the complete gravity-shift results. T-RAFT attains the highest mean score in 10 of the 36 configurations. Its strongest results occur on HalfCheetah, Hopper, and Walker2d. On Hopper with medium source data, for example, T-RAFT improves the strongest baseline from 52.5 ± 5.5 to 53.9 ± 7.3 for the medium-expert target and from 55.4 ± 13.9 to 66.3 ± 9.1 for the expert target. A similarly large gain appears for Hopper with medium-replay source and expert target data, where T-RAFT reaches 47.5 ± 9.9 compared with 34.0 ± 5.3 for the strongest baseline.

Table 5: Performance comparison on MuJoCo under gravity shifts. We report normalized target-domain scores as mean±standard deviation over five seeds. Source abbreviations follow Table 4. The highest mean in each row is highlighted in bold.

Source	Target	IQL	DARA	BOSA	SRPO	IGDF	OTDF	DmC	ROOT	T-RAFT
Ant-m	medium	10.2 ± 1.8	9.4 ± 0.9	12.4 ± 2.0	11.7 ± 1.0	11.3 ± 1.3	45.1 ± 12.4	$\mathbf{56.9 \pm 2.2}$	27.0 ± 3.0	34.7 ± 3.1
Ant-m	med-expert	9.4 ± 1.2	10.0 ± 0.9	11.6 ± 1.3	10.2 ± 1.2	9.4 ± 1.4	33.9 ± 5.4	$\mathbf{47.5 \pm 3.9}$	17.4 ± 2.4	23.2 ± 3.9
Ant-m	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	$\mathbf{36.1 \pm 7.8}$	20.7 ± 1.9	24.8 ± 2.2
Ant-me	medium	9.8 ± 2.4	8.1 ± 1.8	8.1 ± 3.0	8.4 ± 2.1	8.9 ± 1.5	18.6 ± 11.9	$\mathbf{55.7 \pm 8.1}$	28.4 ± 6.0	34.8 ± 6.9
Ant-me	med-expert	9.0 ± 0.8	6.4 ± 1.4	6.2 ± 1.5	6.1 ± 3.5	7.2 ± 2.9	34.0 ± 9.4	$\mathbf{46.3 \pm 4.9}$	17.6 ± 0.8	18.8 ± 2.7
Ant-me	expert	9.1 ± 2.6	10.4 ± 2.9	4.2 ± 3.9	8.8 ± 1.0	9.2 ± 1.5	23.2 ± 2.9	$\mathbf{42.7 \pm 13.0}$	20.9 ± 4.2	18.8 ± 0.9
Ant-mr	medium	18.9 ± 2.6	21.7 ± 2.1	13.9 ± 1.5	18.7 ± 1.7	19.6 ± 1.0	29.6 ± 10.7	$\mathbf{41.8 \pm 4.7}$	31.4 ± 1.1	28.7 ± 2.1
Ant-mr	med-expert	19.1 ± 3.0	18.3 ± 2.1	15.9 ± 2.7	18.7 ± 1.8	20.3 ± 1.6	25.4 ± 2.1	27.6 ± 0.1	$\mathbf{28.5 \pm 1.8}$	26.1 ± 0.8
Ant-mr	expert	18.5 ± 0.9	20.0 ± 1.3	14.5 ± 1.7	19.9 ± 2.1	18.8 ± 2.1	24.5 ± 2.8	28.0 ± 0.3	$\mathbf{28.3 \pm 1.5}$	26.6 ± 1.9
Half-m	medium	39.6 ± 3.3	41.2 ± 3.9	38.9 ± 4.0	36.9 ± 4.5	36.6 ± 5.5	40.7 ± 7.7	48.0 ± 0.6	47.7 ± 0.9	$\mathbf{49.9 \pm 1.5}$
Half-m	med-expert	39.6 ± 3.7	40.7 ± 2.8	40.4 ± 3.0	40.7 ± 2.3	38.7 ± 6.2	28.6 ± 3.2	$\mathbf{48.9 \pm 0.7}$	48.6 ± 0.8	48.4 ± 1.0
Half-m	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	$\mathbf{48.8 \pm 1.0}$	48.4 ± 1.1	41.0 ± 2.9
Half-me	medium	38.6 ± 6.0	37.8 ± 3.3	41.8 ± 5.1	42.5 ± 2.3	37.7 ± 7.3	39.5 ± 3.5	49.9 ± 1.4	49.0 ± 1.7	$\mathbf{50.5 \pm 0.9}$
Half-me	med-expert	39.6 ± 3.0	39.4 ± 4.4	38.7 ± 3.7	43.3 ± 2.7	40.7 ± 3.2	32.4 ± 5.5	48.1 ± 1.9	48.3 ± 1.3	$\mathbf{50.4 \pm 1.4}$
Half-me	expert	43.4 ± 0.9	45.3 ± 1.3	39.9 ± 2.7	43.3 ± 3.0	41.1 ± 4.1	26.5 ± 9.1	$\mathbf{51.0 \pm 0.8}$	48.9 ± 0.8	48.0 ± 4.0
Half-mr	medium	20.1 ± 5.0	17.6 ± 6.2	20.0 ± 4.9	17.5 ± 5.2	14.4 ± 2.2	21.5 ± 6.5	32.2 ± 0.8	$\mathbf{35.7 \pm 2.1}$	33.9 ± 4.5
Half-mr	med-expert	17.2 ± 1.6	20.2 ± 5.2	16.7 ± 4.2	16.3 ± 1.7	10.0 ± 2.5	14.7 ± 4.1	30.3 ± 5.1	$\mathbf{34.0 \pm 1.6}$	26.7 ± 4.8
Half-mr	expert	20.7 ± 5.5	22.4 ± 1.7	15.4 ± 4.2	23.1 ± 4.0	15.3 ± 3.7	11.4 ± 1.9	32.7 ± 2.6	$\mathbf{34.9 \pm 3.4}$	24.1 ± 6.6
Hopp-m	medium	11.2 ± 1.1	17.3 ± 3.8	15.2 ± 3.3	12.4 ± 1.0	15.3 ± 3.5	32.4 ± 8.0	30.6 ± 5.5	$\mathbf{50.3 \pm 5.4}$	46.6 ± 10.9
Hopp-m	med-expert	14.7 ± 3.6	15.4 ± 2.5	21.1 ± 9.3	14.2 ± 1.8	15.1 ± 3.6	24.2 ± 3.6	35.7 ± 1.4	52.5 ± 5.5	$\mathbf{53.9 \pm 7.3}$
Hopp-m	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	51.3 ± 10.6	55.4 ± 13.9	$\mathbf{66.3 \pm 9.1}$
Hopp-me	medium	19.1 ± 6.6	18.5 ± 12.3	15.9 ± 5.9	19.7 ± 8.5	22.3 ± 5.4	26.4 ± 10.1	52.2 ± 6.6	$\mathbf{65.9 \pm 12.3}$	30.5 ± 5.3
Hopp-me	med-expert	16.8 ± 2.7	16.0 ± 6.1	17.3 ± 2.5	15.8 ± 3.3	16.6 ± 7.7	28.3 ± 6.7	$\mathbf{50.4 \pm 11.3}$	49.8 ± 7.9	39.7 ± 3.0
Hopp-me	expert	20.9 ± 4.1	23.9 ± 14.8	23.2 ± 7.9	21.4 ± 1.9	26.0 ± 9.2	44.9 ± 10.6	52.8 ± 10.3	$\mathbf{60.3 \pm 10.6}$	50.3 ± 6.7
Hopp-mr	medium	13.9 ± 2.9	10.7 ± 4.3	3.3 ± 1.9	14.0 ± 2.6	15.3 ± 4.4	31.1 ± 13.4	35.4 ± 0.7	$\mathbf{35.8 \pm 2.3}$	31.6 ± 0.3
Hopp-mr	med-expert	13.3 ± 6.3	12.5 ± 5.6	4.6 ± 1.7	14.4 ± 4.2	15.4 ± 5.5	24.2 ± 6.1	$\mathbf{41.0 \pm 2.0}$	40.4 ± 3.5	35.1 ± 2.8
Hopp-mr	expert	11.0 ± 2.6	14.3 ± 6.0	3.2 ± 0.8	16.4 ± 5.0	16.1 ± 4.0	31.0 ± 9.8	29.9 ± 7.8	34.0 ± 5.3	$\mathbf{47.5 \pm 9.9}$
Walk-m	medium	28.1 ± 12.9	28.4 ± 13.7	38.0 ± 11.2	21.4 ± 7.0	22.1 ± 8.4	36.6 ± 2.3	52.5 ± 2.0	$\mathbf{54.8 \pm 3.0}$	54.7 ± 3.4
Walk-m	med-expert	35.7 ± 4.7	30.7 ± 9.7	40.9 ± 7.2	34.0 ± 9.9	35.4 ± 9.1	44.8 ± 7.5	59.2 ± 2.7	60.1 ± 2.5	$\mathbf{65.3 \pm 2.4}$
Walk-m	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	63.8 ± 2.7	63.0 ± 2.7	$\mathbf{63.9 \pm 3.9}$
Walk-me	medium	39.9 ± 13.1	41.6 ± 13.0	32.3 ± 7.2	46.4 ± 3.5	33.8 ± 3.1	30.2 ± 9.8	57.5 ± 3.3	58.5 ± 6.5	$\mathbf{63.0 \pm 3.1}$
Walk-me	med-expert	49.1 ± 6.9	45.8 ± 9.4	40.1 ± 4.5	36.4 ± 3.4	44.7 ± 2.9	53.3 ± 7.1	$\mathbf{67.8 \pm 4.0}$	62.5 ± 2.8	65.6 ± 4.6
Walk-me	expert	40.4 ± 11.9	56.4 ± 3.5	43.7 ± 4.4	45.8 ± 8.0	45.3 ± 10.4	61.1 ± 3.4	67.1 ± 4.8	65.7 ± 3.3	$\mathbf{67.3 \pm 2.8}$
Walk-mr	medium	14.6 ± 2.5	14.1 ± 6.1	7.6 ± 5.8	17.9 ± 3.8	11.6 ± 4.6	32.7 ± 7.0	$\mathbf{42.2 \pm 5.9}$	40.3 ± 4.5	40.7 ± 7.4
Walk-mr	med-expert	15.3 ± 1.9	15.9 ± 5.8	4.8 ± 5.8	15.3 ± 4.5	13.9 ± 6.5	31.6 ± 6.1	31.5 ± 5.2	$\mathbf{41.1 \pm 6.5}$	27.7 ± 4.4
Walk-mr	expert	15.8 ± 7.2	15.7 ± 4.5	7.1 ± 4.6	13.7 ± 8.1	15.2 ± 5.3	31.3 ± 5.3	39.3 ± 6.2	$\mathbf{40.9 \pm 2.0}$	34.3 ± 5.7

On Walker2d, T-RAFT achieves the highest mean in several medium and medium-expert source settings, including 65.3 ± 2.4 and 63.9 ± 3.9 with medium source data and medium-expert/expert targets. In contrast, DmC is consistently stronger on the Ant configurations, while ROOT performs better on several medium-replay and Hopper medium-expert source settings. The gravity results therefore indicate that reconstruction is not uniformly advantageous for every source-target distribution pair, but can yield substantial gains when target-compatible outcome correction complements useful source coverage.

G.3 Results on Adroit Manipulation Tasks

We further evaluate T-RAFT on the Adroit Pen and Door manipulation tasks, following the perturbation protocol of MOBODY. The target dynamics contain either a kinematic shift, implemented as a broken-joint perturbation, or a morphology shift, implemented by shrinking the fingers. Each perturbation is evaluated at medium (M) and hard (H) severity. We compare against IQL, TD3-BC, DARA, BOSA, MOPO, REAG, SRPO, and MOBODY. Baseline results are reproduced from the corresponding MOBODY benchmark, while T-RAFT results are reported over five random seeds.

As shown in Table 6, T-RAFT obtains the highest mean score in six of the eight configurations. The improvement is particularly consistent on Door, where T-RAFT performs best under both perturbation types and at both severity levels. Under broken-joint shifts, it reaches 53.82 ± 5.93 and 74.71 ± 4.31 at medium and hard severity, improving over the strongest baseline by 14.56 and 13.10 normalized-score points, respectively. Under shrunk-finger morphology shifts, it achieves 76.19 ± 2.18 and 77.13 ± 5.28, corresponding to improvements of 12.52 and 8.49 points.

T-RAFT also performs strongly on Pen under medium perturbations, obtaining 77.00 ± 8.58 for broken joint and 23.77 ± 9.89 for shrunk fingers. Its advantage is less consistent under the hard Pen settings: it remains close to REAG under the hard kinematic shift but falls below MOBODY under the hard morphology shift. The relatively large variance in several Pen configurations also suggests that these dexterous-manipulation tasks are substantially more sensitive to the particular target trajectories available. Nevertheless, the consistent improvements on Door and the two medium

Table 6: Normalized target-domain scores on Adroit manipulation tasks. Broken joint and shrunk fingers denote kinematic and morphology shifts, respectively; M/H denote medium/hard shift severity. T-RAFT results are mean±standard deviation over five seeds, while baseline statistics are reproduced from the MOBODY benchmark. The highest mean in each row is highlighted in bold.

Env.	Shift	Level	BOSA	IQL	TD3-BC	DARA	MOPO	REAG	SRPO	MOBODY	T-RAFT
Pen	Broken joint	M	30.63 ± 9.01	24.34 ± 15.49	6.86 ± 6.63	38.60 ± 3.44	37.99 ± 7.46	62.36 ± 2.60	23.41 ± 2.39	37.67 ± 4.54	$\mathbf{77.00 \pm 8.58}$
Pen	Broken joint	H	7.18 ± 2.02	7.74 ± 3.48	1.31 ± 1.29	9.41 ± 6.06	8.14 ± 2.92	$\mathbf{26.30 \pm 6.49}$	8.10 ± 5.75	13.73 ± 6.32	25.30 ± 12.23
Pen	Shrunk fingers	M	10.72 ± 6.65	13.75 ± 4.91	2.20 ± 1.71	8.72 ± 3.12	3.48 ± 0.85	18.06 ± 3.52	7.60 ± 3.73	16.48 ± 10.46	$\mathbf{23.77 \pm 9.89}$
Pen	Shrunk fingers	H	11.78 ± 6.57	32.16 ± 1.14	9.12 ± 9.03	22.17 ± 3.90	28.89 ± 2.48	18.38 ± 4.59	8.76 ± 6.82	$\mathbf{37.80 \pm 1.18}$	26.95 ± 13.72
Door	Broken joint	M	25.42 ± 22.04	37.43 ± 12.76	-0.23 ± 0.01	20.18 ± 5.29	27.90 ± 7.92	38.86 ± 7.63	0.67 ± 1.25	39.26 ± 3.72	$\mathbf{53.82 \pm 5.93}$
Door	Broken joint	H	30.64 ± 26.87	56.02 ± 7.74	-0.12 ± 0.02	58.22 ± 9.91	57.45 ± 9.58	60.22 ± 5.23	0.56 ± 1.25	61.61 ± 9.84	$\mathbf{74.71 \pm 4.31}$
Door	Shrunk fingers	M	41.59 ± 5.95	60.74 ± 12.83	-0.19 ± 0.01	50.32 ± 4.78	52.02 ± 1.74	61.45 ± 0.68	1.44 ± 0.98	63.67 ± 9.52	$\mathbf{76.19 \pm 2.18}$
Door	Shrunk fingers	H	26.97 ± 8.62	68.64 ± 8.34	-0.20 ± 0.02	44.22 ± 7.19	67.06 ± 1.96	62.66 ± 1.60	0.76 ± 0.62	62.88 ± 5.25	$\mathbf{77.13 \pm 5.28}$

Pen settings indicate that the transition-reconstruction mechanism extends beyond locomotion and remains effective under qualitatively different physical perturbations.

G.4 Compute Infrastructure

All experiments are conducted on a local workstation equipped with an Intel Core i9-13900K CPU, one NVIDIA RTX A6000 GPU with 48GB GPU memory, and 32GB system memory. Unless otherwise specified, each training run is executed on a single GPU. The reported wall-clock time in Appendix G.5 is measured under the same hardware configuration.

Table 7: Compute infrastructure.

CPU	GPU	Memory
Intel Core i9-13900K	RTX A6000 $\times$ 1	32GB

G.5 Time Cost

We further compare the wall-clock training time of T-RAFT with representative cross-domain offline RL baselines under the same experimental setting. As reported in Table 8, T-RAFT requires approximately 2h1m for a complete run. The additional computation mainly comes from three preprocessing stages. First, T-RAFT solves the UOT problem to obtain the transported source mass and source-target correspondences. Second, it pretrains the conditional Flow Matching model using the UOT-induced pairs. Third, the learned flow is integrated once to translate the source next-state outcomes and construct the reconstructed source dataset. These operations are performed before policy optimization, and the resulting UOT statistics and reconstructed transitions are cached and reused throughout downstream offline RL training. Therefore, the additional cost does not scale with the number of policy gradient updates.

Table 8: Wall-clock time comparison. We report the time required by each method under the same experimental setting. h=hour(s), m=minute(s).

Method	IGDF	OTDF	DROCO	T-RAFT
Time Cost	1h11m	1h40m	1h25m	2h1m

Although T-RAFT introduces additional preprocessing compared with methods that only filter or reweight source transitions, its overall computational cost remains moderate: the complete pipeline takes 2h1m on a single RTX A6000 GPU under our experimental setting.