DATA FUSION–ENHANCED DECISION TRANSFORMER FOR STABLE CROSS-DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain shifts present a significant challenge for decision transformer (DT) policies. Existing methods typically rely on a single simple filtering criterion to select source trajectory fragments and stitch them together. They match either state structure or action feasibility. However, the selected fragments still have poor stitchability: state structures can misalign, the return-to-go (RTG) becomes incomparable when the reward or horizon changes, and actions may jump at trajectory junctions. As a result, RTG tokens lose continuity, which compromises DT's inference ability. To tackle these challenges, we propose Data Fusion–Enhanced Decision Transformer (DFDT), a compact pipeline that restores stitchability. Particularly, DFDT fuses scarce target data with selectively trusted source fragments via a two-level filter, Maximum Mean Discrepancy (MMD) mismatch for state-structure alignment and Optimal Transport (OT) deviation for action feasibility. It then trains on a feasibility-weighted fusion distribution. Furthermore, DFDT replaces RTG tokens with advantage-conditioned tokens, which improves the continuity of the semantics in the token sequence. It also applies a Q-guided regularizer to suppress junction value and action jumps. Theoretically, we provide bounds that tie state value and policy performance gaps to MMD-mismatch and OT-deviation, and show that the bounds tighten as these two measures shrink. We show that DFDT improves return and stability over strong offline RL and sequence-model baselines across gravity, kinematic, and morphology shifts on D4RL-style control tasks, and further corroborate these gains with token-stitching and sequence-semantics stability analyses.

1 INTRODUCTION

Offline reinforcement learning (RL) promises to turn logged interactions into deployable policies without further environment access, improving safety and sample efficiency in costly or risky domains (Levine et al., 2020). Sequence modeling approaches such as Decision Transformer (DT) (Chen et al., 2021) recast RL as conditional sequence prediction and achieve strong results by conditioning actions on return-to-go (RTG) (Chen et al., 2021; Janner et al., 2021a). However, cross-domain deployment, when training and test dynamics differ, remains difficult (Wen et al., 2024): stitched fragments lose token continuity, state manifolds drift, actions become infeasible at junctions, and RTG becomes incomparable under reward and horizon changes. In practice, DT-style agents overfit source statistics and fail to generalize (Wu et al., 2023; Wang et al., 2023).

A pragmatic remedy is cross-domain offline RL that uses scarce target data to guide adaptation while still exploiting rich source logs (Xu et al., 2023; Liu et al., 2022). Many methods filter or reweight source data to bias learning toward transitions closer to the target dynamics, e.g., contrastive representation learning filtering (Wen et al., 2024), support-aware selection (Liu et al., 2024), stationary distribution regularization (Xue et al., 2023), or optimal-transport alignment (Lyu et al., 2025b). These strategies help value-based and actor–critic pipelines (Konda & Tsitsiklis, 1999) but do not act on sequence tokens that govern DT policies. In particular, they neither selectively trust fragments by token-level stitchability nor repair token semantics (e.g., discontinuous RTG) under reward and horizon shifts, leaving junction continuity unresolved. Supervised RL (RvS) variants (Emmons et al., 2022) inherit the same RTG fragility.

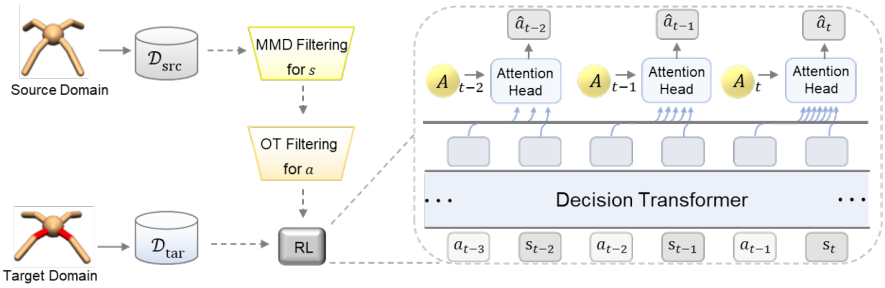

Figure 1: **An overview of our proposed framework.** Credible source fragments are selected by MMD state-alignment and OT action-feasibility filters, fused with scarce target data, and fed, together with advantage-conditioned A tokens, into a Decision Transformer whose attention heads predict stable actions $\hat{a}_t$ under cross-dynamics shifts.

Two classic tools are well-suited yet underused in DTs: kernel Maximum Mean Discrepancy (MMD) to test state-distribution similarity (Gretton et al., 2006), and Optimal Transport (OT) to measure geometry-aware shifts in joint state–action transitions (Villani et al., 2008; Peyré et al., 2019). We *unify* them into a DT-compatible pipeline that explicitly restores token-level stitchability.

We propose **Data Fusion–Enhanced Decision Transformer (DFDT)**, which fuses scarce target data with selectively trusted source fragments. DFDT (i) gates fragments via *MMD* (state-structure alignment) and assigns *OT*-derived credibility for action feasibility, (ii) trains critics on a feasibility-weighted fusion distribution, (iii) replaces brittle RTG tokens with *advantage-conditioned* tokens to stabilize conditioning across reward and horizon changes, and (iv) adds a lightweight Q-guided regularizer to suppress value and action jumps at stitch junctions. Theoretically, we formalize the reweighted fusion measure and prove weighted Bellman-error transfer results, including TD-residual shift, and the approximations of V and Q. Moreover, we prove that the target performance are controlled by two measurable "stitchability radii" (1) an MMD-based state-structure radius and (2) an OT transition radius.

We evaluate DFDT on D4RL continuous-control tasks (Fu et al., 2020) under gravity, kinematic, and morphology shifts with scarce target data. DFDT consistently improves return and stability over strong offline RL and DT-style baselines across diverse source–target pairs. Furthermore, we report token-stitching and sequence-semantics stability analyses—tracking action jumps, Q-value jumps, and local TD residuals at stitch junctions—which corroborate DFDT's stability and help explain its performance gains. More ablation results isolating the contributions of two-level filtering, advantage conditioning, and Q regularization are presented in Sec. G.2.

2 Preliminaries

Cross-domain Offline Reinforcement Learning. We consider two infinite-horizon Markov Decision Processes (MDPs), the *source* domain $\mathcal{M}_S := (\mathcal{S}, \mathcal{A}, P_S, r_S, \gamma, \rho_0)$ and the *target* domain $\mathcal{M}_T := (\mathcal{S}, \mathcal{A}, P_T, r_T, \gamma, \rho_0)$. The two domains share the same state space $\mathcal{S}$, action space $\mathcal{A}$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (bounded by $r_{\max}$), discount factor $\gamma \in [0, 1)$, and initial distribution ρ_0, but differ in their transition kernels $P_S \neq P_T$. For any MDP $\mathcal{M}$ and policy π, let the normalized discounted state and state-action occupancy measures be $d^\pi_{\mathcal{M}}(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P^\pi_{\mathcal{M}}(s\,|\,t)$ and $\nu^\pi_{\mathcal{M}}(s,a) := d^\pi_{\mathcal{M}}(s)\,\pi(a\,|\,s)$, and define the performance of π by $J_{\mathcal{M}}(\pi) := \mathbb{E}_{(s,a)\sim\nu^\pi_{\mathcal{M}}}[\,r(s,a)\,]$.

Let $\mathcal{D}_{\text{src}} = \{(s, a, r, s')\}$ be an offline dataset from $\mathcal{M}_S$ and $\mathcal{D}_{\text{tar}} = \{(s, a, r, s')\}$ a much smaller dataset from $\mathcal{M}_T$. We aim to learn a policy $\pi^\star$ that maximises $J_T(\pi)$ without online interaction with $\mathcal{M}_T$. The core challenge is the cross-domain shift $P_S \neq P_T$, which is especially harmful to DT policies that rely on token-level continuity in RTG, state, and action: source and target state manifolds misalign, RTG becomes incomparable under reward and horizon shifts, and actions feasible under P_S can be implausible under P_T, creating stitch discontinuities. This breaks local sequence structure, induces exposure bias, and destabilizes RTG conditioning; using only $\mathcal{D}_{\text{src}}$ extrapolates invalid next tokens under P_T, while $\mathcal{D}_{\text{tar}}$ alone lacks coverage. We therefore fuse $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tar}}$ to restore token continuity, aligning state structure, ensuring action feasibility, and replacing fragile

RTG with a stable conditioning signal, hence Transformer policies generalize across domains while controlling distributional mismatch.

Expectile Regression. For a response $Y \in \mathbb{R}$ and covariates $X \in \mathcal{X}$, the ζ-expectile regression function ($\zeta \in (0,1)$) is the map $m_\zeta : \mathcal{X} \to \mathbb{R}$ that minimizes the asymmetric least-squares (ALS) risk $\mathbb{E}\big[\rho_\zeta\big(Y - m_\zeta(X)\big)\big]$ with $\rho_\zeta(u) = \big|\zeta - \mathbf{1}\{u < 0\}\big|\, u^2$, yielding a unique minimizer under mild integrability by strict convexity. Equivalently, m_ζ satisfies the balance condition $\zeta\,\mathbb{E}[(Y - m_\zeta(X))_+ \mid X] = (1-\zeta)\,\mathbb{E}[(m_\zeta(X) - Y)_+ \mid X]$, $(z)_+ = \max\{z, 0\}$. Expectiles continuously interpolate tail emphasis: $\zeta = \frac{1}{2}$ recovers the conditional mean, while $\zeta \to 1$ (resp. $\to 0$) increases sensitivity to upper (resp. lower) tails; moreover, $m_\zeta(x)$ is nondecreasing in ζ and translation/scale equivariant, i.e., for $a \in \mathbb{R}$, $b > 0$, the expectile of $a + bY$ equals $a + b\, m_\zeta(x)$.

3 Cross-domain Policy Adaptation of Decision Transformer

We start by introducing a two-level data filtering and reweighted fusion framework in Sec. 3.1. Sec. 3.2 analyses the policy performance difference bound under the proposed data fusion framework. Moreover, we introduce our practical algorithm to fulfil the cross-domain DT-based policy adaptation algorithm in Sec. 3.3.

3.1 Two-level data filtering and reweighted fusion framework

In this section, to enable stable data filtering, we start by proposing a two-level trajectory fragment filtering technique and then formalize a feasibility-weighted fusion distribution.

3.1.1 Fragment Selection via MMD and Wasserstein Distance

To achieve more efficient and stable data augmentation of target domains using source datasets, we perform the fragment selection in two consecutive stages: (i) state transition structure similarity matching via Maximum Mean Discrepancy (MMD), and (ii) action credibility estimation via Wasserstein distance. Based on our proposed two-level data filtering mechanism, we can conduct a more precise selection of the trajectory segments while enabling suitable action weights.

MMD-based Fragment Selection. Let $z = f_\phi(s)$ be a shared encoder for extracting the feature information of states. For a source fragment $\tau^S = (s_1^S, a_1^S, r_1^S, \ldots, s_n^S, a_n^S, r_n^S)$ and a target fragment $\tau^T = (s_1^T, a_1^T, r_1^T, \ldots, s_m^T, a_m^T, r_m^T)$, we compute the RBF-kernel MMD in latent space:

$$\mathrm{MMD}_k^2(\tau^S, \tau^T) = \tfrac{1}{n^2}\sum_{i,j} k(z_i^S, z_j^S) + \tfrac{1}{m^2}\sum_{i,j} k(z_i^T, z_j^T) - \tfrac{2}{nm}\sum_{i,j} k(z_i^S, z_j^T), \tag{1}$$

where $z_i^S = f_\phi(s_i^S)$ and k is an RBF kernel. This score measures structural similarity between dynamics; we keep the top-$\xi\%$ source fragments with the smallest values to form a pseudo-target buffer $\mathcal{D}_{\mathrm{sim}}$. Then, we define the state-structure MMD distance of τ^S to the target dataset:

$$d^m(\tau^S) \;=\; \mathbb{E}_{\tau^T}\mathrm{MMD}_k\big(\{z_i^S = f_\phi(s_i^S)\}_{i=1}^{\ell},\ \{z_j^T = f_\phi(s_j^T)\}_j^{\ell}\big), \tag{2}$$

and the hard gate $I_m(\tau^S) := \mathbf{1}\big(d^m(\tau^S) \le q_\xi\big)$ that retains the top-$\xi\%$ most similar fragments. This gate removes fragments that induce large token discontinuities in the state space at stitch junctions.

OT-based Action Credibility. For each $\tau^S \in \mathcal{D}_{\mathrm{src}}$, we use the OT (Villani et al., 2008; Peyré et al., 2019) to evaluate how plausible its actions are under target dynamics. Define concatenations $v_t^S = s_t^S \oplus a_t^S \oplus r_t^S \oplus s_{t+1}^S$ and $v_t^T = s_t^T \oplus a_t^T \oplus r_t^T \oplus s_{t+1}^T$, with $v_t^S \sim \mathcal{D}_{\mathrm{src}}$, $v_t^T \sim \mathcal{D}_{\mathrm{tar}}$. Given a 1-Lipschitz cost function C and coupling μ, the Wasserstein distance is defined as

$$W_1 = \min_{\mu \in \mathcal{M}} \sum_{t=1}^{|\mathcal{D}_{\mathrm{src}}|} \sum_{t'=1}^{|\mathcal{D}_{\mathrm{tar}}|} C(v_t^S, v_{t'}^T)\, \mu_{t,t'}. \tag{3}$$

Suppose solving the optimization problem in Eq. 3 gives the OT μ^* (Kantorovich, 1942), we determine the deviation between a source domain data and the target domain dataset via:

$$d^w(u_t) = -\sum_{t'=1}^{|\mathcal{D}_{\mathrm{tar}}|} C(v_t^S, v_{t'}^T)\, \mu_{t,t'}^*, \tag{4}$$

which becomes larger when the source sample aligns well with the target behavior (i.e., lower transport cost), and smaller otherwise. The OT credibility score prioritises actions whose transport cost to the target behavior is small, improving action continuity across stitched tokens.

3.1.2 TWO-LEVEL REWEIGHTED DATA FUSION FRAMEWORK

This section formalises a two-level data filtering framework and introduces a reweighted data fusion distribution. We write triples $u = (s, a, s')$ and suppose the distributions of triples u_T and u_S are $\mathbb{P}_T$ and $\mathbb{P}_S$ induced form P_T and P_S, respectively. Then, the two-level reweighted data fusion framework is expressed as follows:

Definition 3.1 (Two-level reweighted data fusion framework)**.** For each triple $u = (s, a, s')$ from a gated source fragment, let $d^w(u)$ be the OT credibility score, and set the raw per-sample weight

$$w(u) \;:=\; I_m(\tau^S)\,\exp\big(\eta_w\, d^w(u)\big),$$

where $\eta_w > 0$ is the weight temperature coefficient. Normalize and clip batch-wise: $\tilde{w}(u) := \frac{w(u)}{\mathbb{E}_{u\sim\mathbb{P}_S}[w(u)]}$, Define the reweighted source distribution and data fusion distribution:

$$\mathbb{P}_S^w(u) \propto \tilde{w}(u)\,\mathbb{P}_S(u) \quad \text{and} \quad \mathbb{P}_{\text{mix}}^w = (1-\beta)\,\mathbb{P}_T \;+\; \beta\,\mathbb{P}_S^w,\;\; \beta \in [0,1].$$

Learning objective under reweighted data fusion distribution. We propose that training samples are drawn from $\mathbb{P}_{\text{mix}}^w$ and we minimize weighted expectile and TD losses for Q and V, respectively

$$\mathcal{L}_V = \mathbb{E}_{\mathbb{P}_{\text{mix}}^w}\Big[\rho_\zeta\big(r(s,a) + \gamma V(s') - V(s)\big)\Big], \quad \mathcal{L}_Q = \mathbb{E}_{\mathbb{P}_{\text{mix}}^w}\Big[\rho_{\frac{1}{2}}\big(r + \gamma V(s') - Q(s,a)\big)\Big].$$

We train the policy with a weighted DT objective under $\mathbb{P}_{\text{mix}}^w$ and a Q-regularized term:

$$\mathcal{L}_\pi = \mathcal{L}_{\text{DT}}^w - \alpha\,\mathbb{E}_{\mathbb{P}_{\text{mix}}^w}\big[Q(s,\pi(s))\big] \tag{5}$$

3.2 THEORETICAL INTERPRETATIONS

In this section, we theoretically unpack the performance difference between the optimal policy in the target domain and the learned one of the mixed dynamics $\mathbb{P}_{\text{mix}}$. The derived performance bound highlights the complementary nature of MMD fragment selection and OT credibility estimation: if only one criterion is used (MMD or OT), the other term remains. To obtain our theoretical results, we first define the estimation error for Q and V.

Definition 3.2 (Estimation errors $\varepsilon_V, \varepsilon_Q$)**.** Let the one–step TD residuals $\Delta_V(s,a,s') := r(s,a) + \gamma\,V(s') - V(s)$ and $\Delta_Q(s,a,s') := r(s,a) + \gamma\,V(s') - Q(s,a)$. We *define* the estimation errors $\varepsilon_V, \varepsilon_Q$ as the conditional residual bounds under $\mathbb{P}_{\text{mix}}^w$: $\varepsilon_V := \sup_{s\in\mathcal{S}} \Big|\mathbb{E}_{\mathbb{P}_{\text{mix}}^w}\big[\Delta_V(s,a,s') \mid s\big]\Big|$ and $\varepsilon_Q := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|\mathbb{E}_{\mathbb{P}_{\text{mix}}^w}\big[\Delta_Q(s,a,s') \,|\, s,a\big]\Big|$.

In the following assumption, we formally consider that the bounds of $\varepsilon_V, \varepsilon_Q$ are finite, which is reasonable due to the fitting ability of neural networks.

Assumption 3.1 (Finite estimation errors)**.** Due to finite samples and function class complexity, the fitted estimators satisfy finite errors, i.e., $\varepsilon_V, \varepsilon_Q < \infty$.

Given the fiber (preimage) of f_ϕ over $z \in \mathcal{Z}$ that is defined as the level set $f_\phi^{-1}(\{z\}) = \{\, s \in \mathcal{S} : f_\phi(s) = z \,\}$, we assume the value of V over the set $f_\phi^{-1}(\{z\})$ is almost constant.

Assumption 3.2 (Approximate fiber-constancy of V)**.** The value varies little within the fibers of f_ϕ: $\exists\, \varepsilon_H \geq 0, \sup\big\{|V(s) - V(\tilde{s})| : s, \tilde{s} \in \mathcal{S},\; f_\phi(s) = f_\phi(\tilde{s})\big\} \;\leq\; \varepsilon_H$.

We define stitchability radii to measure the distance between the mixed data and the target domain.

Definition 3.3 (Stitchability radii)**.** Let $\Delta_m := \sup_{\tau^S : I_m(\tau^S)=1} d^m(\tau^S)$ and $\Delta_w := W_1\big(\mathbb{P}_S^w,\, \mathbb{P}_T\big)$, which measures the residual state-structure MMD and the 1-Wasserstein distance between the reweighted source triples and the target domain.

Assumption 3.3 (Occupancy concentrability)**.** Let d_T^* and d_T^{mix} denote the normalized discounted state–occupancy measures of π_T^* and π_{mix} in the target MDP. We assume *absolute continuity* and a finite essential bound on the density ratio: $d_T^* \ll d_T^{\text{mix}}$ and $\kappa := \left\| \frac{\mathrm{d}d_T^*}{\mathrm{d}d_T^{\text{mix}}} \right\|_\infty < \infty$.

Theorem 3.1 (Performance bound under stitchability radii)**.** *Under Assumptions B.1–3.2, training with $\mathbb{P}_{\text{mix}}^w$ yields the estimators V and Q. V_T and Q_T are the state and state-action value functions learned from the target dataset. Let π_T^* and π_{mix} be any optimal policies learned from the target MDP $\mathbb{P}_T$ and mixed MDP $\mathbb{P}_{\text{mix}}^w$, respectively. Let $d_T^* \otimes \pi_T^*$ be the normalized discounted state–action occupancy of π_T^* under $\mathbb{P}_T$ and $\Delta_\pi := \mathbb{E}_{s\sim\mathbb{P}_T}\big[\|\pi_T^*(\cdot|s) - \pi_{\text{mix}}(\cdot|s)\|_1\big]$. Then, for some constants $C_1, C_2, C_3, C_H, C_\pi > 0$,*

$$\|V - V_T\|_{1,\mathbb{P}_T} \;\le\; \frac{C_1\,\beta\,(\Delta_m + \Delta_w) \;+\; 2\beta\,\varepsilon_H \;+\; \varepsilon_V}{1-\gamma}, \tag{6}$$

$$\|Q - Q_T\|_{1,\mathbb{P}_T} \;\le\; \frac{C_2\,\beta\,(\Delta_m + \Delta_w) \;+\; 2\beta\,\varepsilon_H \;+\; \varepsilon_Q}{1-\gamma}. \tag{7}$$

Moreover, by a performance difference bound,

$$J_T(\pi_T^*) - J_T(\pi_{\text{mix}}) \;\le\; \frac{C_3(1+\kappa)}{(1-\gamma)^2}\Big(\beta(\Delta_m + \Delta_w) \;+\; C_H\,\beta\,\varepsilon_H \;+\; \varepsilon_V\Big) \;+\; \frac{C_\pi}{(1-\gamma)^2}\,\Delta_\pi. \tag{8}$$

The proof can be found in Appendix E. The stitchability radii affect policy learning only via w in the optimization objectives and the sampling distribution $\mathbb{P}_{\text{mix}}^w$. If $w \equiv 1$ or $\beta = 0$, then $\Delta_m = \Delta_w = 0$ and the bounds reduce to pure target-domain learning. If only one criterion is used (MMD or OT), the other term remains, highlighting their complementarity.

3.3 Practical Algorithm

In this section, we implement a cross-domain DT-based policy adaptation algorithm utilizing our proposed two-level data filtering and reweighted fusion framework. The complete algorithm flow can be found in Algs. 1 and 2.

3.3.1 RTG Relabeling with Value Consistency for State Stitching

To stabilize token semantics across fragments and mitigate horizon and reward mismatch at junctions, we propose a reweighted advantage relabeling mechanism to treat the advantage function as a structural surrogate for RTG.

Cross-Domain RTG-Relabeling Q and V Function. We train a shared value function $V_\varphi(s)$ and Q-function $Q_\psi(s_t, a_t)$ using a combined dataset of target domain samples and the selected source fragments. Since selected fragments are structurally aligned with the target dynamics, the shared $V_\varphi(s)$ promotes a consistent reward structure across domains. Specifically, we estimate the value function $V_\varphi(s)$ based on a new reweighted expectile regression as follows:

$$\mathcal{L}_V = \mathbb{E}_{\tau^T}\left[\rho_\zeta(\Delta_V)\right] + \mathbb{E}_{\tau^S}\left[\exp(\eta_w d^w) I_m(\tau^S)\rho_\zeta(\Delta_V)\right], \tag{9}$$

and

$$\mathcal{L}_Q = \mathbb{E}_{\tau^T}\left[\rho_{\frac{1}{2}}(\Delta_Q)\right] + \mathbb{E}_{\tau^S}\left[\exp(\eta_w d^w) I_m(\tau^S)\rho_{\frac{1}{2}}(\Delta_Q)\right], \tag{10}$$

where $\rho_{\frac{1}{2}} := \rho_{\zeta=\frac{1}{2}}$ and $I_m(\tau^S) := \mathbf{1}\big(d^m(\tau^S) \le q_\xi\big)$ is the hard gate function introduced in Definition 3.1. The expectile function ρ_ζ is defined in Section 2. The definition of Δ_V and Δ_Q can be found in Definition 3.2.

Reweighted Advantage Policy Conditioning. For each state-action pair (s_i, a_i) in a selected fragment $\tau = (s_t, a_t, r_t, \ldots, s_{t+k}, a_{t+k}, r_{t+k})$, we generate a pseudo-return token as $A(s_i, a_i) = Q(s_i, a_i) - V(s_i)$. These reweighted advantage values replace the original (unavailable or inconsistent) RTG signals. The resulting transformer token sequence

$$(s_t, a_t, A_t, \ldots, s_{t+k}, a_{t+k}, A_{t+k}) \tag{11}$$

forms the input for policy training. This process enforces reward continuity and structure-aware return alignment, enabling stable cross-domain policy adaptation.

3.3.2 Reweighted Q-Regularized Transformer Training

To ensure continuity of action prediction and realise the stitching of cross-domain trajectory fragments, based on the cross-domain reweighted advantage conditioning, we adopt a Q-regularized training loss for the transformer policy with the reweighted data fusion distribution.

Critic Network Update. The optimization of the parameter ϕ of the critic network Q_ϕ is carried out by minimizing the following equation:

$$\mathcal{L}_Q(\phi; \mathcal{D}_{\text{tar}}) + \mathop{\mathbb{E}}_{\substack{\mathcal{T}\sim\mathcal{D}_{\text{src}} \\ \hat{a}_t\sim\pi_{\theta'}}} \left[I_m(\tau^S) \sum_{i=t-K+1}^{t-1} \left\| \exp(\eta_w d_i^w)\left(\hat{Q}_i - Q_{\phi_i}(s_i, a_i)\right) \right\|^2 \right], \tag{12}$$

where $\hat{Q}_i = \sum_{j=i}^{t-1} \gamma^{j-i} r_j + \gamma^{t-i} \min_{i=1,2} Q_{\phi_i'}(s_t, \hat{a}_t)$ is the multi-step TD estimation, and $\mathcal{L}_Q(\phi; \mathcal{D}_{tar})$ is the original Q-network loss computed with multi-step TD estimation and mean square error (MSE). We use the hard gate function I_m and OT weight $\exp(\eta_w d_i^w)$ to adjust the contribution of the source data in the loss function.

Cross-domain DT Loss. We use the conditional transformer policy DT that receives the relabeled sequences as input and predicts the next action at each step. The model is trained using a behavior cloning loss:

$$\mathcal{L}_{\text{DT}}^w = \mathcal{L}_{\text{DT}}(\theta; \mathcal{D}_{\text{tar}}) + \mathbb{E}_{\tau^S} \left[\frac{1}{K} I_m(\tau^S) \sum_{i=t-K+1}^{t} \exp(\eta_w d_i^w) \left(a_i - \pi(\tau^S)_i\right)^2 \right] \tag{13}$$

where $\mathcal{L}_{\text{DT}}(\theta; \mathcal{D}_{\text{tar}})$ denotes the original DT loss function computed with $\mathcal{D}_{\text{tar}}$.

Integrating Q-Value Regularization into the DT Loss Function. To bias the policy toward higher-value actions and improve stitching at the action level, we incorporate a Q-value regularizer:

$$\mathcal{L}_\pi = \mathcal{L}_{\text{DT}}^w - \alpha \cdot \mathbb{E}_{\tau\in\mathcal{D}_{\text{tar}}\cup\mathcal{D}_{\text{src}}} \left[\frac{I_m(\tau^S)}{K} \sum_{i=t-K+1}^{t} \exp(\eta_w d_i^w) Q_\phi(s, \pi(s)) \right], \tag{14}$$

where $Q(s, a)$ is trained using conservative Q-learning or fitted Q-iteration, and $\mathcal{L}_{\text{reg}}(\pi)$ is the regularization term of π.

4 Experiments

In this section, we evaluate DFDT under gravity, kinematic, and morphology shifts, centring on two research questions: (a) Does DFDT outperform strong prior baselines across gravity, kinematic, and morphology shifts and across source and target dataset qualities? (b) Can DFDT provide stable sequence semantics for policy adaptation?

4.1 Main Results

Tasks and Datasets. We evaluate policy adaptation under three dynamics shifts, *gravity*, *kinematics*, and *morphology*, on four MuJoCo tasks (HalfCheetah, Hopper, Walker2d, Ant) in OpenAI Gym Brockman et al. (2016). Gravity scales the magnitude of g; kinematics constrains joint ranges; morphology changes link dimensions. We adopt the configurations of Lyu et al. (2025b). The setting is cross-domain offline RL: abundant source data but scarce target data from shifted environments. Sources are D4RL "-v2" datasets (medium, medium-replay, medium-expert) (Fu et al., 2020); targets are the D4RL-style datasets of Lyu et al. (2025b) (medium / medium-expert / expert), each with 5,000 transitions. This low-data regime, known to challenge standard offline RL Liu et al. (2024); Wen et al. (2024); Lyu et al. (2025b), yields 108 tasks across the three shift families.

Baselines. We compare DFDT with strong offline RL: IQL (Kostrikov et al., 2022) (expectile value regression with advantage-weighted policy), and sequence-modelling baselines for cross-domain adaptation: DT (Chen et al., 2021) (return-to-go sequence model), QT (Hu et al., 2024) (value-aware DT), and a DADT variant (Kim et al., 2022) with dynamics-aware tokenisation but no filtering. We also include recent cross-domain methods: DARA (Liu et al., 2022), IGDF (Wen et al., 2024), and

Table 1: **Performance comparison of cross-domain offline RL algorithms under morphology shifts.** half = halfcheetah, hopp = hopper, walk = walker2d, m = medium, r = replay, e = expert. The 'Target' column indicates target-domain offline data quality. We report *normalized* target-domain performance (mean ± s.d.) across source qualities {medium, medium-replay, medium-expert} and target qualities {medium, medium-expert, expert}, averaged over five seeds; best per row is highlighted.

Source	Target	IQL	DARA	IGDF	OTDF	DT	QT	DADT	DFDT
half-m	medium	30.0	26.6	41.6	39.1	34.6	34.5	34.8	**42.5**±2.0
half-m	medium-expert	31.8	32.0	29.6	35.6	30.8	−1.3	36.5	**42.5**±1.9
half-m	expert	8.5	9.3	10.0	10.7	4.7	0.8	11.5	**69.0**±7.3
half-m-r	medium	30.8	35.6	28.0	40.0	30.3	31.1	30.2	**42.9**±2.0
half-m-r	medium-expert	12.9	16.9	12.0	34.4	19.4	24.6	25.7	**42.8**±0.6
half-m-r	expert	5.9	3.7	5.3	8.2	4.7	11.3	9.5	**53.0**±18.7
half-m-e	medium	41.5	40.3	40.9	41.4	34.9	22.2	36.4	**42.2**±3.0
half-m-e	medium-expert	25.8	30.6	26.2	35.1	36.5	20.7	37.1	**43.8**±0.5
half-m-e	expert	7.8	8.3	7.5	9.8	7.7	7.6	5.4	**73.7**±7.0
hopp-m	medium	13.5	13.5	13.4	11.0	12.1	10.1	11.4	**44.7**±16.5
hopp-m	medium-expert	13.4	13.6	13.3	12.6	13.2	13.2	13.1	**36.0**±20.5
hopp-m	expert	13.5	13.6	13.9	10.7	12.9	13.1	13.5	**56.1**±39.1
hopp-m-r	medium	10.8	10.2	12.0	8.7	13.3	13.1	14.4	**53.2**±21.5
hopp-m-r	medium-expert	11.6	10.4	8.2	9.7	12.4	15.6	12.2	**79.9**±13.0
hopp-m-r	expert	9.8	9.0	11.4	10.7	12.7	**15.7**	13.7	**15.7**±2.7
hopp-m-e	medium	12.6	13.0	12.7	7.9	11.8	9.9	11.9	**46.9**±27.4
hopp-m-e	medium-expert	14.1	13.8	13.3	9.6	11.8	12.6	10.7	**69.7**±27.3
hopp-m-e	expert	13.8	12.3	12.8	5.9	12.0	12.7	11.7	**86.5**±21.4
walk-m	medium	23.0	23.3	27.5	**50.5**	23.7	11.5	20.8	46.6±9.3
walk-m	medium-expert	21.5	22.2	20.7	**44.3**	22.4	29.0	25.3	27.2±10.8
walk-m	expert	20.3	17.3	15.8	55.3	15.6	23.8	28.3	**70.3**±22.1
walk-m-r	medium	11.3	10.9	13.4	37.4	12.3	30.1	28.3	**44.8**±5.0
walk-m-r	medium-expert	7.0	4.5	6.9	33.8	6.0	1.6	13.6	**40.6**±20.7
walk-m-r	expert	6.3	4.5	5.5	41.5	10.1	1.1	9.5	**86.3**±16.1
walk-m-e	medium	24.1	31.7	27.5	**49.9**	17.8	19.7	27.7	38.3±7.1
walk-m-e	medium-expert	27.0	23.3	25.3	**40.5**	14.3	24.2	25.2	28.4±6.3
walk-m-e	expert	22.4	25.2	24.7	45.7	10.2	21.8	26.7	**85.5**±9.9
ant-m	medium	38.7	41.3	40.9	39.4	37.9	38.6	42.5	**42.6**±0.6
ant-m	medium-expert	47.0	43.3	44.4	58.3	48.1	1.0	44.0	**75.4**±6.2
ant-m	expert	36.2	48.5	41.4	85.4	22.8	−1.0	23.7	**85.5**±11.8
ant-m-r	medium	38.2	38.9	39.7	41.2	17.5	25.0	37.8	**41.4**±1.3
ant-m-r	medium-expert	38.1	33.4	37.3	50.8	28.6	8.2	39.0	**78.3**±8.9
ant-m-r	expert	24.1	24.5	23.6	67.2	21.2	8.3	25.9	**75.0**±15.8
ant-m-e	medium	32.9	40.2	36.1	39.9	41.3	35.1	27.4	**42.0**±0.5
ant-m-e	medium-expert	35.7	36.5	30.7	65.7	57.3	12.8	43.1	**69.5**±11.4
ant-m-e	expert	36.1	34.6	35.2	**86.4**	37.9	12.3	31.1	81.9±6.9
Total Score		798.0	816.8	808.7	1274.3	760.8	570.6	859.6	**2000.7**

OTDF (Lyu et al., 2025b), covering reward reweighting, representation filtering, and OT-based data fusion. For all baselines, we use the authors' recommended hyperparameters and code, modifying only the dataset and environment identifiers.

Evaluation Protocol We adopt the cross-domain setup in Sec. 2, using abundant D4RL source logs (*medium*, *medium-replay*, *medium-expert*) and scarce target logs collected under gravity, kinematic, and morphology shifts. We report normalized target-domain returns (mean ± std.) over five seeds while sweeping all {source quality}×{target quality} pairs (3×3) across HalfCheetah, Hopper, Walker2d, and Ant. The computing method of normalized returns is described in Sec. All methods train offline on the prescribed source and target logs. More detailed experimental settings can be found in Sec. F.

Results. We train our method for 100k gradient updates with five random seeds and report normalized target-domain scores. Summary comparisons of DFDT against baselines under morphology and

Table 2: **Performance comparison of cross-domain offline RL algorithms under kinematic shifts.** Abbreviations are as in Table 1. We report *normalized* target-domain performance (mean ± s.d.) over five seeds; best per row is highlighted.

Source	Target	IQL	DARA	IGDF	OTDF	DT	QT	DADT	DFDT
half-m	medium	12.3	10.6	23.6	40.2	32.1	14.6	14.5	**41.2**±0.5
half-m	medium-expert	10.8	12.9	9.8	10.1	22.4	6.2	21.4	**40.8**±1.5
half-m	expert	12.6	12.1	12.8	8.7	13.9	5.0	15.8	**27.5**±5.0
half-m-r	medium	10.0	11.5	11.6	37.8	11.6	10.7	8.8	**40.8**±0.3
half-m-r	medium-expert	6.5	9.2	8.6	9.7	7.5	40.1	6.0	**41.4**±1.6
half-m-r	expert	13.6	14.8	13.9	7.2	2.7	19.2	5.7	**27.6**±7.4
half-m-e	medium	21.8	25.9	21.9	30.7	17.5	18.7	14.5	**41.2**±0.9
half-m-e	medium-expert	7.6	9.5	8.9	10.9	13.1	3.7	11.4	**35.5**±12.1
half-m-e	expert	9.1	10.4	10.7	3.2	19.5	10.3	19.4	**26.0**±14.2
hopp-m	medium	58.7	43.9	65.3	65.6	16.4	19.7	3.6	**66.5**±0.9
hopp-m	medium-expert	**68.5**	55.4	51.1	55.4	6.3	10.9	10.4	56.2±28.5
hopp-m	expert	79.9	83.7	**87.4**	35.0	3.5	7.8	3.5	57.6±32.7
hopp-m-r	medium	36.0	39.4	35.9	35.5	11.1	23.0	16.8	**63.1**±3.4
hopp-m-r	medium-expert	36.1	34.1	36.1	47.5	3.8	**54.0**	35.3	23.7±17.6
hopp-m-r	expert	36.0	36.1	36.1	49.9	9.8	19.9	6.7	**62.0**±20.7
hopp-m-e	medium	66.0	61.1	65.2	65.3	21.6	3.4	14.3	**66.8**±1.4
hopp-m-e	medium-expert	45.1	61.9	**62.9**	38.6	10.3	16.9	6.6	49.2±27.3
hopp-m-e	expert	44.9	**84.2**	52.8	29.9	18.7	10.9	15.5	68.1±16.8
walk-m	medium	34.3	35.2	41.9	49.6	31.6	26.9	27.3	**55.7**±11.0
walk-m	medium-expert	30.2	**51.9**	42.3	43.5	35.8	19.8	19.1	37.6±8.2
walk-m	expert	56.4	40.7	**60.4**	46.7	35.4	50.2	38.2	55.7±8.0
walk-m-r	medium	11.5	12.5	22.2	49.7	17.9	33.7	6.8	**54.2**±19.9
walk-m-r	medium-expert	9.7	11.2	7.6	**55.9**	24.2	49.8	28.1	31.3±9.7
walk-m-r	expert	7.7	7.4	7.5	51.9	18.4	3.1	18.0	**53.7**±6.9
walk-m-e	medium	41.8	38.1	41.2	44.6	38.6	5.6	**78.9**	60.1±4.9
walk-m-e	medium-expert	22.2	23.6	28.1	16.5	15.2	29.2	33.0	**51.4**±21.2
walk-m-e	expert	26.3	36.0	46.2	42.4	39.3	25.0	32.2	**56.8**±11.5
ant-m	medium	50.0	42.3	54.5	**55.4**	31.2	22.5	17.7	53.2±5.9
ant-m	medium-expert	57.8	54.1	54.5	**60.7**	13.0	7.9	13.5	60.3±5.4
ant-m	expert	59.6	54.2	49.4	**90.4**	7.0	7.0	11.7	88.7±8.9
ant-m-r	medium	43.7	42.0	41.4	**52.8**	31.1	22.4	30.3	51.7±4.6
ant-m-r	medium-expert	36.5	36.0	37.2	54.2	26.9	12.0	33.1	**62.8**±1.9
ant-m-r	expert	24.4	22.1	24.3	74.7	27.1	8.9	25.5	**89.9**±5.0
ant-m-e	medium	49.5	44.7	41.8	50.2	21.2	9.4	11.1	**52.2**±4.8
ant-m-e	medium-expert	37.2	33.3	41.5	48.8	16.5	10.8	13.6	**55.6**±3.2
ant-m-e	expert	18.7	17.8	14.4	78.4	7.2	8.0	11.7	**88.5**±10.3
Total Score		1193.0	1219.8	1271.0	1547.6	679.4	647.2	680	**1894.6**

kinematic shifts are given in Tables 1 and 2, respectively; results for the gravity shifts are deferred to Appendix G due to space limit.

For question (a): On both morphology and kinematic shifts, DFDT consistently surpasses sequence-modelling baselines (DT, QT, DADT) with one exception and frequently outperforms strong cross-domain offline RL methods (e.g., OTDF, DARA, IGDF) across all morphology and kinematic shift tasks. Notably, DFDT achieves higher normalized scores than all baselines on **31** out of **36** tasks under the morphology shifts and **23** out of **36** tasks under the kinematic shifts. In the few settings where a competing method attains the top score, DFDT typically ranks second with a small gap, indicating broad robustness rather than narrow wins. The total normalized score improves by **57.0%** under morphology shifts and **12.0%** under gravity shifts when using DFDT (both relative to the second-best baseline, OTDF; for reference, the gains vs. IQL are **150.7%** and **57.6%**, respectively), providing strong evidence for the method's effectiveness. Beyond these aggregates, the improvements are broad across environments (HalfCheetah, Hopper, Walker2d, Ant) rather than concentrated: for example, DFDT leads all **9** HalfCheetah configurations under both morphology and kinematic shifts, and its margins are most pronounced on expert-target datasets where sequence stitching yields especially high returns. Even in rows where another method briefly tops the table,

DFDT's mean typically sits within a few points while maintaining competitive seed-level stability, reinforcing that DFDT's gains reflect reliable cross-domain adaptation rather than isolated outliers.

4.2 Token-Stitching and Sequence-Semantics Stability Analyses

Token-stitching analyses setups. To directly probe sequence semantics at stitch junctions, we precompute the junction index set $\mathcal{J}$ for every relabeled training sequence (the boundary where two fragments are concatenated, or a source $\rightarrow$ target switch). At each training checkpoint, we evaluate three quantities on a fixed validation pool of such sequences: (i) the action jump $J_a = \mathbb{E}_{t^\star \in \mathcal{J}} \|\pi(s_{t^\star}) - \pi(s_{t^\star - 1})\|_2$; (ii) the Q-jump $J_Q = \mathbb{E}_{t^\star \in \mathcal{J}} |Q(s_{t^\star}, \pi(s_{t^\star})) - Q(s_{t^\star - 1}, \pi(s_{t^\star - 1}))|$; and (iii) the TD residual around junctions, computed as $\mathbb{E}_{t \in \mathcal{N}(t^\star)} |r_t + \gamma V(s_{t+1}) - V(s_t)|$, where $\mathcal{N}(t^\star) = \{\, t : |t - t^\star| \le w \,\} \cap \{1, \ldots, T-1\}$ denotes a small temporal neighborhood around the stitch junction index $t^\star$ with a fixed radius w (in our experiments $w = 2$, i.e., two steps before or after the junction). The TD residual is then averaged over this local window to reduce single–step noise at the boundary. Curves in Fig. 2 report moving means over checkpoints for DFDT, DADT, QT, and DT under the same backbone, budget, and data.

Results and answer to (b): DFDT exhibits uniformly lower-level variance on all three diagnostics throughout training, indicating smoother token transitions and better local Bellman consistency at stitch points. Concretely, its action-jump mean remains $\approx 0.06 \sim 0.09$ (vs. QT rising to $0.25 \sim 0.30$ and DT/DADT $\approx 0.10 \sim 0.16$); Q-jumps stay near $2 \sim 3$ (vs. QT often $15 \sim 35$ and DADT spikes > 20); and TD residuals remain around $3 \sim 6$ (vs. QT $15 \sim 30$, DT/DADT $8 \sim 20$). Beyond lower jump means, DFDT's trajectories show markedly fewer late-training spikes, suggesting that weighted advantage conditioning and two-level filtering suppress junction value and action discontinuities as learning progresses. **These trends directly support (b):** DFDT provides stable, value-consistent sequence semantics for policy adaptation, particularly where stitching is challenging; competing sequence models exhibit larger jumps and drift, reflecting unstable semantics across stitched tokens.

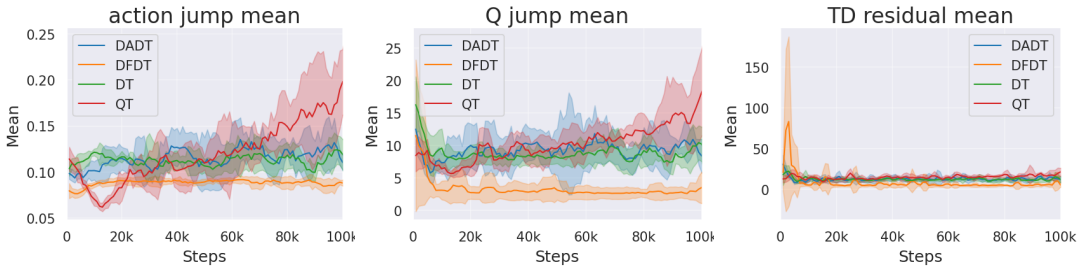

Figure 2: Mean action jump, Q-value jump, and TD error when evaluation.

5 Conclusion

We analyzed why Decision Transformer (DT) fails under cross-domain stitching—returns are incomparable, rewards shift, and feasible actions change—and proposed DFDT to restore token continuity. DFDT couples two-level fragment filtering (MMD state-structure gating and OT-based action credibility) with a reweighted fusion distribution and reweighted advantage conditioning, replacing brittle RTG with a value-consistent signal, plus a lightweight Q regularizer to suppress junction value jumps. Our theory bounds the target performance gap via stitchability radii and estimation errors, highlighting the complementarity of MMD and OT and clarifying that weights act through sampling. Empirically, across morphology, kinematic, and gravity shifts, DFDT attains the best aggregate scores, reduces action, Q-value, and TD jumps, and ablations confirm that both filtering and advantage relabeling are key to stable sequence semantics. Limitations include sensitivity to transport cost design and the risk of over-pruning when shifts are mostly scalar; adaptive, softer gates alleviate this by preserving borderline fragments. Future work will jointly learn task-adaptive costs with the encoder, align cross-domain rewards via value calibration, handle multi-source domain-unlabeled data, and combine DFDT with diffusion-style augmentation or light, safety-aware online fine-tuning.

REFERENCES

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. When does return-conditioned supervised learning work for offline reinforcement learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

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. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Shuo Cheng, Liqian Ma, Zhenyang Chen, Ajay Mandlekar, Caelan Garrett, and Danfei Xu. Generalizable domain adaptation for sim-and-real policy co-training. *arXiv preprint arXiv:2509.18631*, 2025. URL https://arxiv.org/abs/2509.18631.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. RvS: What is essential for offline rl via supervised learning? In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=S874XAIpkR-. arXiv:2112.10751.

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. URL https://arxiv.org/abs/2004.07219.

Scott Fujimoto and Shixiang Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Chen-Xiao Gao, Chenyang Wu, Mingjun Cao, Rui Kong, Zongzhang Zhang, and Yang Yu. Act: empowering decision transformer with dynamic programming via advantage conditioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12127–12135, 2024.

GitHub. Github copilot. https://docs.github.com/copilot, 2025. AI coding assistant; accessed 2025-09-25.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Shengchao Hu, Ziqing Fan, Chaoqin Huang, Li Shen, Ya Zhang, Yanfeng Wang, and Dacheng Tao. Q-value regularized transformer for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 19165–19181. PMLR, 2024.

Michael Janner, Qiyang Li, and Sergey Levine. Trajectory transformer: Off-policy reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a. URL https://arxiv.org/abs/2106.02039.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning (ICML)*, 2022.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

L. V. Kantorovich. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS*, 37:199–201, 1942. Introduces the Kantorovich relaxation of Monge.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Changyeon Kim, Junsu Kim, Younggyo Seo, Kimin Lee, Honglak Lee, and Jinwoo Shin. Dynamics-augmented decision transformer for offline dynamics generalization. In *3rd Offline RL Workshop: Offline RL as a"Launchpad"*, 2022.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8. arXiv:2110.06169.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465*, 2019b.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Guanghe Li, Yixiang Shan, Zhengbang Zhu, Ting Long, and Weinan Zhang. Diffstitch: Boosting offline reinforcement learning with diffusion-based trajectory stitching. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/li24bf.html.

Jinxin Liu, Hongyin Zhang, and Donglin Wang. Dara: Dynamics-aware reward augmentation in offline reinforcement learning. *arXiv preprint arXiv:2203.06662*, 2022.

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*, pp. 13945–13953, 2024.

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

Jiafei Lyu, Kang Xu, Jiacheng Xu, Mengbei Yan, Jing-Wen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, and Xiu Li. Odrl: A benchmark for off-dynamics reinforcement learning. In *NeurIPS Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2410.20750.

Jiafei Lyu, Mengbei Yan, Zhongjian Qiao, Runze Liu, Xiaoteng Ma, Deheng Ye, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=LRrbD8EZJl.

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 *The Thirteenth International Conference on Learning Representations*, 2025b.

A Maddukuri et al. Sim-and-real co-training: A simple recipe for vision-based robot manipulation. *arXiv preprint arXiv:2503.24361*, 2025. URL https://arxiv.org/abs/2503.24361.

Soichiro Nishimori, Xin-Qiang Cai, Johannes Ackermann, and Masashi Sugiyama. Offline reinforcement learning with domain-unlabeled data. *arXiv preprint arXiv:2404.07465*, 2024. URL https://arxiv.org/abs/2404.07465.

OpenAI. Chatgpt (gpt-5). https://chat.openai.com/, 2025. Large language model; Accessed: 2025-09-25.

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 (ICRA)*, 2018.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. In *International Conference on Learning Representations (ICLR)*, 2017.

Jürgen Schmidhuber. Reinforcement learning upside down: Don't predict rewards – just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Yuanfu Wang, Chao Yang, Ying Wen, Yu Liu, and Yu Qiao. Critic-guided decision transformer for offline reinforcement learning. *arXiv preprint arXiv:2312.13716*, 2023. URL https://arxiv.org/abs/2312.13716.

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. *arXiv preprint arXiv:2405.06192*, 2024.

Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. Elastic decision transformer. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3b3889d313ba9476c12c2d77ea66b24f-Paper-Conference.pdf.

Kang Xu, Chenjia Bai, Xiaoteng Ma, Dong Wang, Bin Zhao, Zhen Wang, Xuelong Li, and Wei Li. Cross-domain policy adaptation via value-guided data filtering. *Advances in Neural Information Processing Systems*, 36:73395–73421, 2023.

Zhenghai Xue, Qingpeng Cai, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. State regularized policy optimization on data with dynamics shift. *Advances in neural information processing systems*, 36:32926–32937, 2023.

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

Supplementary Material

Table of Contents

A Related work

Offline Reinforcement Learning. Offline RL (Yu et al., 2020) seeks to learn high-performing policies from fixed datasets without additional environment interaction, and thus must confront distributional shift and overestimation on out-of-distribution (OOD) actions (Levine et al., 2020; Kidambi et al., 2020). Constraint- or conservatism-based methods address this by penalising or avoiding unsupported actions (Lyu et al., 2022), including CQL's explicit value suppression for OOD actions (Kumar et al., 2020), IQL's advantage-weighted updates without importance sampling (Kostrikov et al., 2022), and TD3+BC's minimalist behavior-regularized regression (Fujimoto & Gu, 2021). Earlier behavior-constrained approaches, such as BCQ and BEAR, limit the learned policy's deviation from the behavior policy to reduce extrapolation error (Fujimoto et al., 2019; Kumar et al., 2019a); model-based variants (e.g., MOPO) leverage pessimistic rollouts to avoid compounding model bias (Yu et al., 2020). Benchmarks like D4RL standardise evaluation across tasks and dataset qualities, and have also catalysed analyses contrasting value-learning and supervised, return-conditioned paradigms (Fu et al., 2020; Brandfonbrener et al., 2022).

Cross-Domain Reinforcement Learning. Cross-domain RL studies transfer under mismatched dynamics, morphology, sensing, or rewards, where naively pooling data across domains induces value bias and out-of-distribution actions. Early robustness strategies, domain and dynamics randomisation for sim-to-real transfer and risk-averse ensemble training, remain foundational (Tobin et al., 2017; Peng et al., 2018; Rajeswaran et al., 2017). More principled distribution alignment narrows source–target gaps by matching states and transitions via kernel MMD or optimal transport (OT) with geometry-aware costs (Gretton et al., 2012; Courty et al., 2017; Villani et al., 2008; Peyré et al., 2019). From 2024–2025, several advances clarified data selection and evaluation in the offline setting: supported cross-domain offline RL formalised the problem and constraints (Liu et al., 2024); contrastive representation learning enabled domain-aware filtering without strong labels (Wen et al., 2024); the ODRL benchmark standardised off-dynamics evaluation across gravity, morphology, and other shifts (Lyu et al., 2024); and OTDF combined OT-based filtering with dataset constraints to bound target bias (Lyu et al., 2025b). In parallel, generative augmentation synthesised stitching transitions to connect suboptimal and optimal fragments (Li et al., 2024), domain-unlabeled formulations relaxed per-transition domain tags (Nishimori et al., 2024), and sim–real co-training explored domain-invariant alignment and unbalanced OT when simulation greatly exceeds scarce real data (Maddukuri et al., 2025; Cheng et al., 2025). Finally, sequence-modelling baselines increasingly replace brittle return-to-go conditioning with value- or advantage-aware signals, e.g., advantage-conditioned DT, critic-guided DT, and Q-regularized transformers that promote value-consistent stitching across domains (Gao et al., 2024; Wang et al., 2023; Hu et al., 2024).

Conditional Sequence Modelling for Decision Making. Viewing decision making as conditional sequence modelling enables reuse of powerful generative backbones originating from reward-/return-conditioned policies (Kumar et al., 2019b; Schmidhuber, 2019). Decision Transformer conditions on return-to-go (RTG) to autoregressively generate actions (Chen et al., 2021), Trajectory Transformer models trajectory tokens and performs planning-time search (Janner et al., 2021b), and Diffuser plans by denoising entire trajectories with diffusion models (Janner et al., 2022). However, RTG conditioning can be brittle under reward scaling or horizon mismatch and may degrade across domains; theory and empirical analyses clarify when return-conditioned supervised learning is reliable and where it fails (Brandfonbrener et al., 2022). This motivates replacing or augmenting RTG with value/advantage signals that better reflect local action quality—such as advantage-conditioned DT (Gao et al., 2024), critic-guided conditioning (Wang et al., 2023), and Q-regularized transformers (Hu et al., 2024). Building on this line, we adopt reweighted advantage conditioning with Q-regularization to mitigate value jumps at stitch junctions during cross-dynamics fusion and to stabilize token-level conditioning.

B Additional Lemma B.1 and its proof

Assumption B.1 (Boundedness and Lipschitz)**.** $|r_T| \le R_{\max}$. The state value function V and state-action value function Q are Lipschitz in u with constants L_V, L_Q under the given metric ρ.

Assumption B.2 (Encoder and kernel)**.** $\|f_\phi(s)\| \le B$, and the kernel k is bounded and induces an RKHS with unit-norm ball $\{h : \|h\|_{\mathcal{H}} \le 1\}$.

Assumption B.3 (Compact latent image and continuity)**.** The encoder $f_\phi : \mathcal{S} \to \mathcal{Z}$ is continuous and the latent image $K := \overline{f_\phi(\mathcal{S})} \subset \mathbb{R}^d$ is compact (e.g., via normalisation/clipping).

Assumption B.4 (Universal kernel on K)**.** The bounded kernel k is universal on K, i.e., the induced RKHS $\mathcal{H}$ is dense in $C(K)$ with respect to the uniform norm, where $C(K)$ is the set of all real-valued continuous functions on K.

Definition B.1 (Pushforward measure)**.** Let (X, Σ_X) and (Y, Σ_Y) be measurable spaces, $T : X \to Y$ measurable, and μ a measure on (X, Σ_X). The pushforward of μ by T, denoted $T_{\#}\mu$, is the measure on (Y, Σ_Y) defined by

$$(T_{\#}\mu)(B) = \mu\big(T^{-1}(B)\big) \quad \forall\, B \in \Sigma_Y,$$

equivalently, for integrable $g : Y \to \mathbb{R}$,

$$\int_Y g\, d(T_{\#}\mu) = \int_X g \circ T\, d\mu.$$

Definition B.2 (Kantorovich–Rubinstein Duality)**.** Let $(\mathcal{X}, d)$ be a metric space and let μ, ν be probability measures on $\mathcal{X}$ with finite first moments. The 1-Wasserstein distance is defined by the optimal transport ("primal") problem

$$W_1(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)\, \mathrm{d}\pi(x, y),$$

where $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν. The Kantorovich–Rubinstein (KR) duality states that

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \Big\{ \int_{\mathcal{X}} f\, \mathrm{d}\mu - \int_{\mathcal{X}} f\, \mathrm{d}\nu \Big\}, \qquad \|f\|_{\mathrm{Lip}} := \sup_{x \ne y} \frac{|f(x) - f(y)|}{d(x, y)}.$$

A directly usable inequality derived from the Kantorovich–Rubinstein Duality can be expressed as follows: For any L-Lipschitz function $g : \mathcal{X} \to \mathbb{R}$,

$$\Big|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]\Big| \;\le\; L\, W_1(\mu, \nu).$$

In particular, if g is 1-Lipschitz, then

$$\Big|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]\Big| \;\le\; W_1(\mu, \nu).$$

Lemma B.1 (Expectation deviation under the reweighted data fusion). *For any* 1*-Lipschitz* $g(u)$ *and any* h *with* $\|h\|_{\mathcal{H}} \le 1$*, we have* $\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} g - \mathbb{E}_{\mathbb{P}_T} g\right| \le \beta\,\Delta_w$ *and* $\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} h - \mathbb{E}_{\mathbb{P}_T} h\right| \le \beta\,\Delta_m$.

Proof. Let $\pi_z : \mathcal{U} \to \mathcal{Z}$ map $u = (s, a, s')$ to the latent state $z = f_\phi(s)$. Denote the pushforward marginals by $\mu_T := \pi_{z\#}\mathbb{P}_T$ and $\mu_S^w := \pi_{z\#}\mathbb{P}_S^w$. For each retained fragment τ^S (i.e., $I_m(\tau^S) = 1$), let μ_τ be its latent-state (empirical or normalized) distribution.

Step 1 (Lipschitz part via Kantorovich–Rubinstein duality). By linearity of expectation under the convex mixture, we have

$$\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} g = (1-\beta)\,\mathbb{E}_{\mathbb{P}_T} g + \beta\,\mathbb{E}_{\mathbb{P}_S^w} g \;\Rightarrow\; \mathbb{E}_{\mathbb{P}^w_{\text{mix}}} g - \mathbb{E}_{\mathbb{P}_T} g = \beta\big(\mathbb{E}_{\mathbb{P}_S^w} g - \mathbb{E}_{\mathbb{P}_T} g\big).$$

Taking absolute values and applying the Kantorovich–Rubinstein duality on $(\mathcal{U}, \rho)$,

$$\sup_{\text{Lip}(g)\le 1} \left|\mathbb{E}_{\mathbb{P}_S^w} g - \mathbb{E}_{\mathbb{P}_T} g\right| = W_1(\mathbb{P}_S^w, \mathbb{P}_T) = \Delta_w.$$

Hence, for any 1-Lipschitz g,

$$\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} g - \mathbb{E}_{\mathbb{P}_T} g\right| \le \beta\,\Delta_w.$$

Step 2 (RKHS/MMD part on latent states). For $h \in \mathcal{H}$ acting on z, expectations under triple distributions reduce to those under their latent pushforwards: $\mathbb{E}_{\mathbb{P}} h := \mathbb{E}_{z\sim\pi_{z\#}\mathbb{P}}[h(z)]$. As above,

$$\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} h - \mathbb{E}_{\mathbb{P}_T} h = \beta\big(\mathbb{E}_{\mathbb{P}_S^w} h - \mathbb{E}_{\mathbb{P}_T} h\big) = \beta\big(\mathbb{E}_{\mu_S^w} h - \mathbb{E}_{\mu_T} h\big).$$

Taking the supremum over the unit RKHS ball and using the kernel mean embedding characterisation of MMD,

$$\sup_{\|h\|_{\mathcal{H}}\le 1} \left|\mathbb{E}_{\mu_S^w} h - \mathbb{E}_{\mu_T} h\right| = \text{MMD}_k(\mu_S^w, \mu_T).$$

Therefore, for any $\|h\|_{\mathcal{H}} \le 1$,

$$\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} h - \mathbb{E}_{\mathbb{P}_T} h\right| \le \beta\,\text{MMD}_k(\mu_S^w, \mu_T).$$

Step 3 (Bounding $\text{MMD}_k(\mu_S^w, \mu_T)$ by Δ_m). Since $\mathbb{P}_S^w$ places mass only on retained fragments, its latent marginal is a convex combination $\mu_S^w = \sum_{\tau^S: I_m(\tau^S)=1} \alpha_\tau\,\mu_\tau$ with $\alpha_\tau \ge 0$, $\sum_\tau \alpha_\tau = 1$. As an IPM, MMD is convex in its first argument; thus,

$$\text{MMD}_k\Big(\sum_\tau \alpha_\tau \mu_\tau,\ \mu_T\Big) \;\le\; \sum_\tau \alpha_\tau\,\text{MMD}_k(\mu_\tau, \mu_T) \;\le\; \sup_{\tau: I_m(\tau)=1} \text{MMD}_k(\mu_\tau, \mu_T) = \Delta_m.$$

Hence $\text{MMD}_k(\mu_S^w, \mu_T) \le \Delta_m$, and therefore

$$\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} h - \mathbb{E}_{\mathbb{P}_T} h\right| \le \beta\,\Delta_m, \qquad \forall\, \|h\|_{\mathcal{H}} \le 1.$$

Combining the Lipschitz/Wasserstein bound (Step 1) and the RKHS/MMD bound (Steps 2–3) yields

$$\left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} g - \mathbb{E}_{\mathbb{P}_T} g\right| \le \beta\,\Delta_w, \qquad \left|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}} h - \mathbb{E}_{\mathbb{P}_T} h\right| \le \beta\,\Delta_m,$$

as claimed. □

C Additional Lemma C.2 and its proof

Definition C.1 (Polish space). A topological space (X, τ) is called *Polish* if it is

- **separable**: there exists a countable dense subset $D \subseteq X$, and
- **completely metrizable**: there exists a metric d that generates τ and under which (X, d) is complete.

Equivalently, a Polish space is a separable, complete metric space (up to homeomorphism). Typical examples: $\mathbb{R}^n$ with the Euclidean topology, any closed subset of a Polish space, and countable products of Polish spaces.

Definition C.2 (Quotient map and induced map)**.** Let $f_\phi : \mathcal{S} \to \mathcal{Z}$ be a continuous map. Define an equivalence relation on $\mathcal{S}$ by $s \sim \tilde{s} \iff f_\phi(s) = f_\phi(\tilde{s})$, and let $q : \mathcal{S} \to \mathcal{S}/\!\sim$ be the canonical quotient map $q(s) = [s]$. Write $K := f_\phi(\mathcal{S}) \subseteq \mathcal{Z}$ for the image (with the subspace topology).

There is a unique map

$$\bar{f} : \mathcal{S}/\!\sim \longrightarrow K, \qquad \bar{f}([s]) = f_\phi(s),$$

such that $f_\phi = \bar{f} \circ q$. The map $\bar{f}$ is a bijection. Equipping $\mathcal{S}/\!\sim$ with the quotient topology (induced by q), $\bar{f}$ is a homeomorphism *iff* f_ϕ is a quotient map onto K (equivalently, the subspace topology on K agrees with the quotient topology via f_ϕ). In this case, we (canonically) identify K with the quotient $\mathcal{S}/\!\sim$ and call q the quotient map associated with f_ϕ.

Lemma C.1 (Continuity of the quotient map and induced factor)**.** *Let $(\mathcal{S}, \tau_\mathcal{S})$ be a topological space and $\sim$ an equivalence relation on $\mathcal{S}$. Equip $\mathcal{S}/\!\sim$ with the quotient topology*

$$U \subseteq \mathcal{S}/\!\sim \text{ is open} \iff q^{-1}(U) \in \tau_\mathcal{S},$$

where $q : \mathcal{S} \to \mathcal{S}/\!\sim$, $q(s) = [s]$, is the canonical projection. Then q is continuous.

Moreover, let $f_\phi : \mathcal{S} \to \mathcal{Z}$ be continuous and define $s \sim \tilde{s} \iff f_\phi(s) = f_\phi(\tilde{s})$. Writing $K := f_\phi(\mathcal{S}) \subseteq \mathcal{Z}$ with the subspace topology, there exists a unique map

$$\bar{f} : \mathcal{S}/\!\sim \longrightarrow K, \qquad \bar{f}([s]) = f_\phi(s),$$

such that $f_\phi = \bar{f} \circ q$, and $\bar{f}$ is continuous.

Proof. By the definition of the quotient topology, for every open $U \subseteq \mathcal{S}/\!\sim$ we have $q^{-1}(U) \in \tau_\mathcal{S}$, hence q is continuous.

For the second part, the definition of $\bar{f}$ is well-posed because $s \sim \tilde{s}$ implies $f_\phi(s) = f_\phi(\tilde{s})$. Uniqueness follows from $f_\phi = \bar{f} \circ q$. To prove continuity of $\bar{f}$, let $O \subseteq K$ be open (in the subspace topology). Then

$$q^{-1}\big(\bar{f}^{-1}(O)\big) = \{\, s \in \mathcal{S} : \bar{f}(q(s)) \in O \,\} = \{\, s \in \mathcal{S} : f_\phi(s) \in O \,\} = f_\phi^{-1}(O),$$

which is open in $\mathcal{S}$ since f_ϕ is continuous. By the quotient definition, this implies $\bar{f}^{-1}(O)$ is open in $\mathcal{S}/\!\sim$, i.e., $\bar{f}$ is continuous. □

Lemma C.2 (Approximate value-in-RKHS from universality and latent sufficiency)**.** *Under Assumptions B.1, B.2, 3.1, B.3, B.4, and 3.2, for every $\eta > 0$ there exists $h_V \in \mathcal{H}$ such that*

$$\sup_{s \in \mathcal{S}} |V(s) - h_V(f_\phi(s))| \;\le\; \varepsilon_H + \eta.$$

Proof. By Assumption B.1, V is Lipschitz in $u = (s, a, s')$, hence in s; thus V is continuous on $\mathcal{S}$. By Assumption B.3, f_ϕ is continuous and $K = \overline{f_\phi(\mathcal{S})}$ is compact; hence the quotient map induces a continuous function on K up to the fibre variation. Define a (measurable) section $\sigma : K \to \mathcal{S}$ with $f_\phi(\sigma(z)) = z$ (e.g., choose any representative in each fibre) and set $\widetilde{V}(z) := V(\sigma(z))$. For any $s \in \mathcal{S}$ with $z = f_\phi(s)$,

$$|V(s) - \widetilde{V}(f_\phi(s))| = |V(s) - V(\sigma(z))| \;\le\; \varepsilon_H$$

by Assumption 3.2. Therefore,

$$\sup_{s \in \mathcal{S}} |V(s) - \widetilde{V}(f_\phi(s))| \;\le\; \varepsilon_H.$$

Now $\widetilde{V} \in C(K)$ because V and f_ϕ are continuous and K is compact, where $C(K)$ is the set of all real-valued continuous functions on K. By universality (Assumption B.4), for any $\eta > 0$ there exists $h_V \in \mathcal{H}$ such that $\sup_{z \in K} |\widetilde{V}(z) - h_V(z)| \le \eta$. Combining the two displays gives

$$\sup_{s \in \mathcal{S}} |V(s) - h_V(f_\phi(s))| \;\le\; \sup_{s} |V(s) - \widetilde{V}(f_\phi(s))| + \sup_{z \in K} |\widetilde{V}(z) - h_V(z)| \;\le\; \varepsilon_H + \eta,$$

as claimed. The RKHS norm bound $\|h_V\|_\mathcal{H} \le C_V(\eta)$ follows from standard RKHS approximation estimates and can be absorbed into constants elsewhere. □

Remark C.1. If f_ϕ is *value-sufficient* (i.e., $V(s) = \widetilde{V}(f_\phi(s))$ exactly), then $\varepsilon_H = 0$. In practice, ε_H can be made small by training f_ϕ to preserve value-relevant information (e.g., adding an auxiliary head $s \mapsto V(s)$ or contrastive/value-aware objectives) and by normalizing the latent K to be compact.

D ADDITIONAL LEMMA D.1 AND ITS PROOF

Lemma D.1 (Weighted Bellman error transfer)**.** *Let $y = r(s,a) + \gamma V(s')$. There exist constants $R_1, R_2, R_3 > 0$ (depending on $R_{\max}, L_V$ and the encoder/kernel bounds) such that*

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}}[y - V(s)] - \mathbb{E}_{\mathbb{P}_T}[y - V(s)]\right| \;\le\; \beta\,(R_1\,\Delta_m + R_2\,\Delta_w) \;+\; 2\beta\,\varepsilon_H, \tag{15}$$

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}}[y - Q(s,a)] - \mathbb{E}_{\mathbb{P}_T}[y - Q(s,a)]\right| \;\le\; \beta\,(R_1\,\Delta_m + R_3\,\Delta_w) \;+\; 2\beta\,\varepsilon_H. \tag{16}$$

Proof. **Step 1 (Bounding the $V(s)$ term).** Write the one-step TD residual as

$$R(u) \;=\; y - V(s) \;=\; r_T(s,a) + \gamma\,V(s') - V(s), \qquad u = (s,a,s').$$

By linearity of expectation under the mixture,

$$\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}} R - \mathbb{E}_{\mathbb{P}_T} R = \beta\big(\mathbb{E}_{\mathbb{P}^{\,w}_S} R - \mathbb{E}_{\mathbb{P}_T} R\big).$$

Hence

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}} R - \mathbb{E}_{\mathbb{P}_T} R\right| \le \beta\Big(\underbrace{\left|\mathbb{E}_{\mathbb{P}^{\,w}_S} y - \mathbb{E}_{\mathbb{P}_T} y\right|}_{\text{(I)}} + \underbrace{\left|\mathbb{E}_{\mathbb{P}^{\,w}_S} V(s) - \mathbb{E}_{\mathbb{P}_T} V(s)\right|}_{\text{II}}\Big). \tag{17}$$

Term I: Bounding the $(r_T + \gamma V)$ term via the W_1 distance. According to Assumption B.1, r_T is L_r-Lipschitz in (s,a) (or bounded by $R_{\max}$ and L_r finite) and V is L_V-Lipschitz in s' under the given metric ρ on triples $u = (s,a,s')$. Then the function

$$f(u) \;=\; r_T(s,a) + \gamma V(s')$$

is L_f-Lipschitz with $L_f \le L_r + \gamma L_V$. By the Kantorovich–Rubinstein duality,

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_S} f - \mathbb{E}_{\mathbb{P}_T} f\right| \le L_f\, W_1(\mathbb{P}^{\,w}_S, \mathbb{P}_T) = L_f\,\Delta_w.$$

Absorb L_g into a constant $R_2 > 0$ to obtain

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_S}[r_T(s,a) + \gamma V(s')] - \mathbb{E}_{\mathbb{P}_T}[r_T(s,a) + \gamma V(s')]\right| \le R_2\,\Delta_w. \tag{18}$$

Term II: Bounding the $V(s)$ term via MMD. Let $z = f_\phi(s)$ be the latent state and let $\pi_z(u) = z$. Denote the latent pushforwards $\mu_T = \pi_{z\#}\mathbb{P}_T$ and $\mu^{\,w}_S = \pi_{z\#}\mathbb{P}^{\,w}_S$. Then by Lemma B.1 and C.2

$$\begin{aligned}\left|\mathbb{E}_{\mathbb{P}^{\,w}_S} V(s) - \mathbb{E}_{\mathbb{P}_T} V(s)\right| &\le \left|\mathbb{E}_{z\sim\mu^{\,w}_S} h_V(z) - \mathbb{E}_{z\sim\mu_T} h_V(z)\right| + 2\varepsilon_H \\ &\le \|h_V\|_{\mathcal{H}}\,\mathrm{MMD}_k(\mu^{\,w}_S, \mu_T) + 2\varepsilon_H \\ &\le C_V\,\mathrm{MMD}_k(\mu^{\,w}_S, \mu_T) + 2\varepsilon_H.\end{aligned}$$

Since $\mathbb{P}^{\,w}_S$ is supported on retained fragments and $\mu^{\,w}_S$ is their convex combination, MMD convexity yields $\mathrm{MMD}_k(\mu^{\,w}_S, \mu_T) \le \Delta_m$, hence

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_S} V(s) - \mathbb{E}_{\mathbb{P}_T} V(s)\right| \le C_V\,\Delta_m + 2\varepsilon_H. \tag{19}$$

Let $R_1 := C_V$ and plug Eq. (18) and Eq. (19) into Eq. (17) gives

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}} R - \mathbb{E}_{\mathbb{P}_T} R\right| \le \beta\,(R_1\,\Delta_m + R_2\,\Delta_w + 2\varepsilon_H), \tag{20}$$

which is the desired bound about V.

Step 2 (Bounding the $Q(s,a)$ term). We want to bound the distributional shift of the Q–residual $y(u) - Q(s,a)$ between $\mathbb{P}^{\,w}_{\text{mix}}$ and $\mathbb{P}_T$. Introduce and subtract $V(s)$:

$$y - Q = \underbrace{\big(y - V(s)\big)}_{\text{(I)}} \;+\; \underbrace{\big(V(s) - Q(s,a)\big)}_{\text{(II)}}.$$

Hence, by the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}}[y - Q] - \mathbb{E}_{\mathbb{P}_T}[y - Q]\right| \le\; &\underbrace{\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}}[y - V(s)] - \mathbb{E}_{\mathbb{P}_T}[y - V(s)]\right|}_{\text{Term (I)}} \\ &+ \underbrace{\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\text{mix}}}[V(s) - Q(s,a)] - \mathbb{E}_{\mathbb{P}_T}[V(s) - Q(s,a)]\right|}_{\text{Term (II)}}.\end{aligned}$$

Term (I). This is exactly the value–residual transfer term handled in Step 1.

Term (II). Let $g(u) := V(s) - Q(s,a)$. According to Assumption B.1, g is Lipschitz in u under the metric ρ, so that $\mathrm{Lip}(g) \le L_V + L_Q$. By the Kantorovich–Rubinstein duality,

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[g] - \mathbb{E}_{\mathbb{P}_T}[g]\right| \;\le\; \mathrm{Lip}(g)\, W_1(\mathbb{P}^{\,w}_{\mathrm{mix}}, \mathbb{P}_T) \;=\; \beta\, \widetilde{R}_2\, \Delta_w, \tag{II}$$

with $\widetilde{R}_2 := L_V + L_Q$ absorbed into constants.

Combining (I) and (II) and taking a supremum over (s,a) (or dropping the conditioning) yields

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[y - Q] - \mathbb{E}_{\mathbb{P}_T}[y - Q]\right| \;\le\; \beta\,(R_1\,\Delta_m + (R_2 + \widetilde{R}_2)\,\Delta_w) \;+\; 2\beta\,\varepsilon_H.$$

Renaming $R_3 \leftarrow R_2 + \widetilde{R}_2$ gives the stated form

$$\left|\mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[y - Q(s,a)] - \mathbb{E}_{\mathbb{P}_T}[y - Q(s,a)]\right| \;\le\; \beta\,(R_1\,\Delta_m + R_3\,\Delta_w) \;+\; 2\beta\,\varepsilon_H.$$

□

Corollary D.1 (Weighted Bellman transfer for π_{mix})**.** *Let $y = r_T(s,a) + \gamma V(s')$ as in Lemma D.1. For any measurable $\varphi : \mathcal{S} \to \mathbb{R}$ with $\|\varphi\|_\infty \le 1$, there exist constants $R_1, R_2 > 0$ (depending only on $R_{\max}, L_V$ and the encoder/kernel bounds, absorbed into the same symbols) such that*

$$\left|\; \mathbb{E}_{(s,a,s')\sim\mathbb{P}_T^{\pi_{\mathrm{mix}}}}\big[\varphi(s)\,y\big] \;-\; \mathbb{E}_{(s,a,s')\sim\mathbb{P}^{\,w}_{\mathrm{mix}}}\big[\varphi(s)\,y\big] \;\right| \;\le\; \beta\,(R_1\Delta_m + R_2\Delta_w) \;+\; 2\beta\,\varepsilon_H.$$

Proof sketch. *Apply Lemma D.1 to $\psi(s,a,s') := \varphi(s)\,y(s,a,s')$ (bounded by $\|\varphi\|_\infty \le 1$) and note that the target-side law is $\mathbb{P}_T^{\pi_{\mathrm{mix}}}$.* □

E Proof of Theorem 3.1

Definition E.1 (Occupancy–weighted L_1 norms)**.** For any policy π, let d_T^π be the normalized discounted state–occupancy on $\mathcal{S}$ and $d_T^\pi \otimes \pi$ the corresponding state–action occupancy on $\mathcal{S} \times \mathcal{A}$. Define

$$\|f\|_{1,d_T^\pi} := \mathbb{E}_{s\sim d_T^\pi}[\,|f(s)|\,], \qquad \|g\|_{1,d_T^\pi\otimes\pi} := \mathbb{E}_{\substack{s\sim d_T^\pi \\ a\sim\pi(\cdot|s)}}[\,|g(s,a)|\,].$$

Definition E.2 (Radon–Nikodym derivative)**.** Let $(\Omega, \mathcal{F})$ be a measurable space and let ν, μ be σ-finite measures with $\nu \ll \mu$ (i.e., $\mu(A) = 0 \Rightarrow \nu(A) = 0$ for all $A \in \mathcal{F}$). The *Radon–Nikodym derivative* of ν with respect to μ is the (a.e.-unique) measurable function $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} : \Omega \to [0,\infty]$ such that

$$\nu(A) \;=\; \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu \qquad \text{for all } A \in \mathcal{F}.$$

Equivalently, for any μ-integrable g,

$$\int g\,\mathrm{d}\nu \;=\; \int g\,\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu.$$

Lemma E.1 (Radon–Nikodym)**.** *If ν, μ are σ-finite and $\nu \ll \mu$, then the derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ exists and is unique μ-almost everywhere.*

Lemma E.2 (Performance–difference lemma (discounted MDP) Schulman et al. (2015); Kakade & Langford (2002))**.** *Consider a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with $\gamma \in [0,1)$ and an initial-state distribution ρ on $\mathcal{S}$. For any policies π, π', define*

$$V^\pi(s) \;:=\; \mathbb{E}\Big[\sum_{t\ge0}\gamma^t r(s_t,a_t)\;\Big|\; s_0 = s,\; a_t \sim \pi(\cdot|s_t),\; s_{t+1}\sim P(\cdot|s_t,a_t)\Big],$$

$$Q^\pi(s,a) \;:=\; r(s,a) + \gamma\,\mathbb{E}[V^\pi(s') \mid s,a],$$

and the advantage $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$. Let the (discounted) performance be $J(\pi) := \mathbb{E}_{s_0\sim\rho}[V^\pi(s_0)]$. Define the normalized discounted state-occupancy of π':

$$d_\rho^{\pi'}(s) \;:=\; (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\Pr(s_t = s \mid s_0\sim\rho,\; \pi').$$

Then

$$J(\pi') - J(\pi) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\rho^{\pi'},\; a\sim\pi'(\cdot|s)}\big[\,A^{\pi}(s,a)\,\big].$$

In particular, if π' is deterministic, this reduces to

$$J(\pi') - J(\pi) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\rho^{\pi'}}\big[\,A^{\pi}\big(s,\pi'(s)\big)\,\big].$$

Lemma E.3 (Bellman operator is a γ-contraction)**.** *Let $\mathcal{B}(\mathcal{S})$ denote the bounded real-valued functions on $\mathcal{S}$ equipped with the norm $\|\cdot\|$. Fix $\gamma \in [0,1)$ and a target-domain Markov kernel $P(s' \mid s,a)$. Let $\pi(a \mid s)$ be any (possibly stochastic) conditional law of actions given s. Define the target Bellman operator*

$$(\mathcal{T}f)(s) := \mathbb{E}_{\substack{a\sim\pi(\cdot|s)\\ s'\sim P(\cdot|s,a)}}\big[r_T(s,a) + \gamma f(s')\big].$$

Then, for all $f, g \in \mathcal{B}(\mathcal{S})$,

$$\|\mathcal{T}f - \mathcal{T}g\| \le \gamma\|f-g\|.$$

Consequently, $\mathcal{T}$ is a γ-contraction and has a unique fixed point $V^* \in \mathcal{B}(\mathcal{S})$.

Proof. For any $s \in \mathcal{S}$,

$$(\mathcal{T}f - \mathcal{T}g)(s) = \gamma\,\mathbb{E}[f(s') - g(s') \mid s]\,.$$

Hence $\big|(\mathcal{T}f - \mathcal{T}g)(s)\big| \le \gamma\,\|f-g\|$, and taking the supremum over s gives $\|\mathcal{T}f - \mathcal{T}g\| \le \gamma\|f-g\|$. By the Banach fixed-point theorem, $\mathcal{T}$ admits a unique fixed point V^*. □

Assumption E.1 (Occupancy-to-sampling concentrability)**.** There exists a constant $\chi \in (0,\infty)$ such that for every stationary policy π, the normalized discounted occupancy satisfies $d_T^\pi \ll \mathbb{P}_T$ and

$$\left\|\frac{\mathrm{d}d_T^\pi}{\mathrm{d}\mathbb{P}_T}\right\|_\infty \;\le\; \chi.$$

All such density-ratio constants are absorbed into numerical constants below.

Lemma E.4 (Policy mismatch for evaluation residuals)**.** *Let $B_V := R_{\max} + \gamma\|V\|_\infty$. Then for any two stationary policies π_1, π_2,*

$$\|\mathcal{T}_T^{\pi_1}V - \mathcal{T}_T^{\pi_2}V\|_{1,\mathbb{P}_T} \;\le\; B_V \cdot \mathbb{E}_{s\sim\mathbb{P}_T}\big[\|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1\big].$$

In particular, with $\Delta_\pi := \mathbb{E}_{s\sim\mathbb{P}_T}\big[\|\pi_T^*(\cdot|s) - \pi_{\mathrm{mix}}(\cdot|s)\|_1\big]$,

$$\|\mathcal{T}_T^{\pi_T^*}V - \mathcal{T}_T^{\pi_{\mathrm{mix}}}V\|_{1,\mathbb{P}_T} \;\le\; B_V\,\Delta_\pi.$$

Proof. *For each s, $|\mathbb{E}_{a\sim\pi_1}[g_s(a)] - \mathbb{E}_{a\sim\pi_2}[g_s(a)]| \le \|g_s\|_\infty\,\|\pi_1(\cdot|s) - \pi_2(\cdot|s)\|_1$, where $g_s(a) := \mathbb{E}[r_T(s,a) + \gamma V(s') \mid s,a]$ satisfies $\|g_s\|_\infty \le B_V$. Average over $s \sim \mathbb{P}_T$.* □

Theorem 3.1 (Performance bound under stitchability radii)**.** *Under Assumptions B.1–3.2, training with $\mathbb{P}_{\mathrm{mix}}^w$ yields the estimators V and Q. V_T and Q_T are the state and state-action value functions learned from the target dataset. Let π_T^* and π_{mix} be any optimal policies learned from the target MDP $\mathbb{P}_T$ and mixed MDP $\mathbb{P}_{\mathrm{mix}}^w$, respectively. Let $d_T^* \otimes \pi_T^*$ be the normalized discounted state–action occupancy of π_T^* under $\mathbb{P}_T$ and $\Delta_\pi := \mathbb{E}_{s\sim\mathbb{P}_T}\big[\|\pi_T^*(\cdot|s) - \pi_{\mathrm{mix}}(\cdot|s)\|_1\big]$. Then, for some constants $C_1, C_2, C_3, C_H, C_\pi > 0$,*

$$\|V - V_T\|_{1,\mathbb{P}_T} \;\le\; \frac{C_1\,\beta\,(\Delta_m + \Delta_w) \;+\; 2\beta\,\varepsilon_H \;+\; \varepsilon_V}{1-\gamma}, \tag{6}$$

$$\|Q - Q_T\|_{1,\mathbb{P}_T} \;\le\; \frac{C_2\,\beta\,(\Delta_m + \Delta_w) \;+\; 2\beta\,\varepsilon_H \;+\; \varepsilon_Q}{1-\gamma}. \tag{7}$$

Moreover, by a performance difference bound,

$$J_T(\pi_T^*) - J_T(\pi_{\mathrm{mix}}) \;\le\; \frac{C_3(1+\kappa)}{(1-\gamma)^2}\Big(\beta(\Delta_m + \Delta_w) \;+\; C_H\,\beta\,\varepsilon_H \;+\; \varepsilon_V\Big) \;+\; \frac{C_\pi}{(1-\gamma)^2}\,\Delta_\pi. \tag{8}$$

Proof. **Step 1 (One-step residual under the target domain).** Define the (conditional) TD residuals

$$\delta_V(s) := \mathbb{E}_{\mathbb{P}_T}\big[r_T(s,a)+\gamma V(s')-V(s)\,\big|\,s\big], \quad \delta_Q(s,a) := \mathbb{E}_{\mathbb{P}_T}\big[r_T(s,a)+\gamma V(s')-Q(s,a)\,\big|\,s,a\big].$$

Add and subtract the mixed-distribution residuals and apply the triangle inequality:

$$\|\delta_V\|_{1,\mathbb{P}_T} \le \underbrace{\Big|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-V(s)]\Big|}_{\le \varepsilon_V \text{ by Assumption 3.1}} + \Big|\mathbb{E}_{\mathbb{P}_T}[y-V(s)] - \mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-V(s)]\Big|,$$

where $y = r_T(s,a) + \gamma V(s')$. Taking $\sup_s$ and using that conditional deviations are bounded by unconditional ones,

$$\|\delta_V\|_{1,\mathbb{P}_T} \le \varepsilon_V + \Big|\mathbb{E}_{\mathbb{P}_T}[y-V(s)] - \mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-V(s)]\Big|. \tag{21}$$

An identical argument yields

$$\|\delta_Q\|_{1,\mathbb{P}_T} \le \varepsilon_Q + \Big|\mathbb{E}_{\mathbb{P}_T}[y-Q(s,a)] - \mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-Q(s,a)]\Big|. \tag{22}$$

Step 2 (Stitchability transfer). By Lemma D.1, there exist constants $R_1, R_2 > 0$ such that

$$\Big|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-V(s)] - \mathbb{E}_{\mathbb{P}_T}[y-V(s)]\Big| \le \beta\,(R_1\,\Delta_m + R_2\,\Delta_w) + 2\beta\,\varepsilon_H, \tag{23}$$

and

$$\Big|\mathbb{E}_{\mathbb{P}^w_{\text{mix}}}[y-Q(s,a)] - \mathbb{E}_{\mathbb{P}_T}[y-Q(s,a)]\Big| \le \beta\,(R_1\,\Delta_m + R_3\,\Delta_w) + 2\beta\,\varepsilon_H. \tag{24}$$

Plugging Eq. (23) into Eq. (21) and Eq. (24) into Eq. (22) gives

$$\|\delta_V\|_{1,\mathbb{P}_T} \le \varepsilon_V + \beta\,(R_1\,\Delta_m + R_2\,\Delta_w) + 2\beta\,\varepsilon_H, \tag{25}$$
$$\|\delta_Q\|_{1,\mathbb{P}_T} \le \varepsilon_Q + \beta\,(R_1\,\Delta_m + R_3\,\Delta_w) + 2\beta\,\varepsilon_H. \tag{26}$$

Step 3 (Contraction to fixed-point errors). Let $\mathcal{T}_V$ be the (target) Bellman operator associated with y, i.e., $(\mathcal{T}_V f)(s) := \mathbb{E}_{\mathbb{P}_T}[r_T(s,a) + \gamma f(s') \,|\, s]$. By Lemma E.3, T_V is a γ-contraction in $\|\cdot\|_{1,\mathbb{P}_T}$ with unique fixed point V_T. Note that $\delta_V = \mathcal{T}_V V - V$ pointwise, hence

$$\|V - V_T\|_{1,\mathbb{P}_T} \le \frac{\|\mathcal{T}_V V - V\|_{1,\mathbb{P}_T}}{1-\gamma} = \frac{\|\delta_V\|_{1,\mathbb{P}_T}}{1-\gamma}.$$

Using Eq. (25) produces the inequality in Eq. (6) with $C_1 := \max\{R_1, R_2\}$ (absorbing constants).

For Q, define the (evaluation) Bellman operator $(\mathcal{T}_Q f)(s,a) := \mathbb{E}_{\mathbb{P}_T}[r_T(s,a) + \gamma V(s') \,|\, s,a]$, for any f, g,

$$(\mathcal{T}_Q f)(s,a) = r_T(s,a) + \gamma\,\mathbb{E}[V(s') \mid s,a] \quad \text{does not depend on } f \text{ at all,}$$

hence $(\mathcal{T}_Q f) - (\mathcal{T}_Q g) \equiv 0$ and $\|\mathcal{T}_Q f - \mathcal{T}_Q g\|_{1,\mathbb{P}_T} = 0 \le \gamma \|f - g\|_{1,\mathbb{P}_T}$. Therefore $\mathcal{T}_Q$ is (trivially) a γ-contraction with fixed point Q_T for the target problem tied to V.[1] Since $\delta_Q = \mathcal{T}_Q Q - Q$, we obtain

$$\|Q - Q_T\|_{1,\mathbb{P}_T} \le \frac{\|\delta_Q\|_{1,\mathbb{P}_T}}{1-\gamma} \le \frac{\varepsilon_Q + \beta\,(R_1\,\Delta_m + R_3\,\Delta_w) + 2\beta\,\varepsilon_H}{1-\gamma},$$

which yields the second inequality in Eq. (7) (renaming the constant to C_2).

Step 4 (Performance bound). By the performance–difference lemma in occupancy form,

$$J_T(\pi_T^*) - J_T(\pi_{\text{mix}}) \le \frac{1}{1-\gamma}\,\|V_T^* - V_T^{\text{mix}}\|_{1,d_T^*} \le \frac{1}{1-\gamma}\Big(\|V_T^* - V\|_{1,d_T^*} + \|V - V_T^{\text{mix}}\|_{1,d_T^*}\Big). \tag{27}$$

For the second term, change measure to d_T^{mix} by Assumption 3.3:

$$\|V - V_T^{\text{mix}}\|_{1,d_T^*} \le \kappa\,\|V - V_T^{\text{mix}}\|_{1,d_T^{\text{mix}}}, \qquad \kappa := \Big\|\tfrac{\mathrm{d}d_T^*}{\mathrm{d}d_T^{\text{mix}}}\Big\|_\infty.$$

[1] Any standard control/evaluation choice can be used so long as the associated Bellman operator is a γ-contraction with a unique fixed point; the constants absorb the specific choice.

Next, relate value gaps to evaluation residuals along the corresponding policy (standard residual-to-value inequality):

$$\|V_T^{\pi} - V\|_{1,d_T^{\pi}} \;\le\; \frac{1}{1-\gamma}\,\|\mathcal{T}_T^{\pi} V - V\|_{1,d_T^{\pi}}, \qquad \pi \in \{\pi_T^*, \pi_{\mathrm{mix}}\}.$$

Therefore, from Eq. (27),

$$J_T(\pi_T^*) - J_T(\pi_{\mathrm{mix}}) \;\le\; \frac{1}{(1-\gamma)^2}\Big(\|\mathcal{T}_T^{\pi_T^*} V - V\|_{1,d_T^*} \;+\; \kappa\,\|\mathcal{T}_T^{\pi_{\mathrm{mix}}} V - V\|_{1,d_T^{\mathrm{mix}}}\Big). \tag{28}$$

Use Assumption E.1 to transfer both norms to $\mathbb{P}_T$:

$$\|\mathcal{T}_T^{\pi} V - V\|_{1,d_T^{\pi}} \;\le\; \chi\,\|\mathcal{T}_T^{\pi} V - V\|_{1,\mathbb{P}_T}, \qquad \pi \in \{\pi_T^*, \pi_{\mathrm{mix}}\}.$$

Fix $\pi \in \{\pi_T^*, \pi_{\mathrm{mix}}\}$ and decompose (now with m_{mix} *defined* as the Bellman operator under π_{mix}):

$$\|\mathcal{T}_T^{\pi} V - V\|_{1,\mathbb{P}_T} \le \underbrace{\|\mathcal{T}_T^{\pi} V - \mathcal{T}_T^{\pi_{\mathrm{mix}}} V\|_{1,\mathbb{P}_T}}_{\text{policy mismatch}} + \underbrace{\|\mathcal{T}_T^{\pi_{\mathrm{mix}}} V - \mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[\,y \mid s]\|_{1,\mathbb{P}_T}}_{\text{distribution shift}} + \underbrace{\|\mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[\,y \mid s] - V(s)\|_{1,\mathbb{P}_T}}_{\text{estimation}}, \tag{29}$$

where $y = r_T(s,a) + \gamma V(s')$. The *policy-mismatch* term is bounded by Lemma E.4:

$$\|\mathcal{T}_T^{\pi} V - \mathcal{T}_T^{\pi_{\mathrm{mix}}} V\|_{1,\mathbb{P}_T} \;\le\; B_V\,\Delta_\pi, \qquad B_V := R_{\max} + \gamma\|V\|_\infty,\;\; \Delta_\pi := \mathbb{E}_{s\sim\mathbb{P}_T}\big[\|\pi_T^*(\cdot|s) - \pi_{\mathrm{mix}}(\cdot|s)\|_1\big].$$

(For $\pi = \pi_{\mathrm{mix}}$ this term is 0.) The *distribution-shift* term is controlled by Corollary D.1 via the L^1 duality with bounded test functions:

$$\|\mathcal{T}_T^{\pi_{\mathrm{mix}}} V - \mathbb{E}_{\mathbb{P}^{\,w}_{\mathrm{mix}}}[\,y \mid s]\|_{1,\mathbb{P}_T} \;\le\; \beta\,(R_1\Delta_m + R_2\Delta_w) + 2\beta\,\varepsilon_H.$$

For the *estimation* term, Assumption 3.1 gives an ε_V bound under $\mathbb{P}^{\,w}_{\mathrm{mix}}$; changing the measuring law to $\mathbb{P}_T$ introduces only a bounded multiplicative factor (absorbed into constants). Hence, uniformly in π,

$$\|\mathcal{T}_T^{\pi} V - V\|_{1,\mathbb{P}_T} \;\le\; B_V\,\Delta_\pi \;+\; C\Big(\beta(\Delta_m + \Delta_w) + \beta\,\varepsilon_H + \varepsilon_V\Big).$$

Putting this into Eq. (28) and absorbing the multiplicative constants (including χ and those from Lemma D.1) into C_3, C_H, C_π, we obtain

$$J_T(\pi_T^*) - J_T(\pi_{\mathrm{mix}}) \;\le\; \frac{C_3(1+\kappa)}{(1-\gamma)^2}\Big(\beta(\Delta_m + \Delta_w) \;+\; C_H\,\beta\,\varepsilon_H \;+\; \varepsilon_V\Big) \;+\; \frac{C_\pi}{(1-\gamma)^2}\,\Delta_\pi.$$

If one additionally assumes a mild *policy proximity* condition (e.g., $\Delta_\pi \;\le\; C_s\,(\Delta_m + \Delta_w)$ as a stitchability consequence), the Δ_π term can be absorbed into the existing $\beta(\Delta_m + \Delta_w)$ term, recovering the original shape of Eq. (8).

This completes the proof. □

F Algorithm details of DFDT

Computing method of normalized scores. Because raw returns are not directly comparable across environments, we follow D4RL (Fu et al., 2020) and report the Normalized Score (NS):

$$\mathrm{NS} \;=\; \frac{\hat{J} - \hat{J}_{\text{rand}}}{\hat{J}_{\text{exp}} - \hat{J}_{\text{rand}}} \times 100, \tag{30}$$

where $\hat{J}$ is the empirical return of the learned policy, $\hat{J}_{\text{exp}}$ is the expert policy's empirical return, and $\hat{J}_{\text{rand}}$ is the empirical return of a random policy. By construction, $\mathrm{NS} = 100$ corresponds to expert-level performance and $\mathrm{NS} = 0$ corresponds to random performance. See Appendix C.1 of Lyu et al. (2025b) for dataset details about $\hat{J}_{\text{rand}}$ and $\hat{J}_{\text{exp}}$.

Algorithm 1 DFDT Training

Require: Source domain dataset $\mathcal{D}_{\text{src}}$, target domain dataset $\mathcal{D}_{\text{tar}}$, batch size N, sequence length K, data filtering proportion ξ, target update rate η_{exp}

Initialize DT policy π_ϕ, critic networks Q_ϕ, target critic networks $Q_{\phi'}$, state value networks V_φ and state-action value network Q_ψ for computing weighted advantage-conditioned tokens, and command network C_ω

// Offline cost computation

Pre-compute the MMD distance $\{d^m\}^{|\mathcal{D}_{\text{src}}|}$ using Eq. (2) and optimal transport distance $\{d^w\}^{|\mathcal{D}_{\text{src}}|}$ using Eq. (4)

Use the distance information $\{d^m\}^{|\mathcal{D}_{\text{src}}|}$ and $\{d^w\}^{|\mathcal{D}_{\text{src}}|}$ to augment the source dataset $\mathcal{D}_{\text{src}}$ and get $\widehat{\mathcal{D}}_{\text{src}} = \{(s_t, a_t, r_t, s'_t, \text{timesteps}, \text{masks}, d^m_t, d^w_t)\}$

// Compute weighted advantage-conditioned tokens

Train the state and state-action value networks using Eq. (10) and Eq. (9), respectively

Compute the advantage A of each state-action pair in $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tar}}$

// Train the command network

Train the command network C_ω using the advantage information A and MSE loss.

// Main training loop

for $i = 1, 2, \ldots$ **do**

  Sample mini-batch $b_{\text{src}} := \{(s, a, r, s', \text{timestep}, \text{mask}, d^m, d^w)\}$ with size $\frac{N}{2}$ from $\widehat{\mathcal{D}}_{\text{src}}$

  Sample mini-batch $b_{\text{tar}} := \{(s, a, r, s', \text{timesteps}, \text{masks})\}$ with size $\frac{N}{2}$ from $\widehat{\mathcal{D}}_{\text{tar}}$

  Normalize the deviations d^w via Eq. (31) to obtain normalized deviations $\hat{d}^w$

  // Two-level data filtering

  Rank the deviations of the sampled source domain data according to d^m and admit the top $\xi\%$ of them

  Compute the weights for the remaining source domain data via $\exp(\eta^w \hat{d}^w)$

  Optimize the state-action value function Q_ϕ on $b_{\text{src}} \cup b_{\text{tar}}$ via:

$$\mathcal{L}_Q(\phi; \mathcal{D}_{\text{tar}}) + \mathop{\mathbb{E}}_{\substack{\mathcal{T}\sim\mathcal{D}_{\text{src}} \\ \hat{a}_t\sim\pi_{\theta'}}} \left[I_m(\tau^S) \sum_{i=t-K+1}^{t-1} \left\| \exp(\eta_w d^w_i) \big(\hat{Q}_i - Q_{\phi_i}(s_i, a_i)\big) \right\|^2 \right].$$

  Update the target network via $\phi' \leftarrow \eta_{\text{exp}}\phi + (1 - \eta_{\text{exp}})\phi'$

  // Policy adaptation

  Optimize the policy π on $b_{\text{src}} \cup b_{\text{tar}}$ using the Q-guided loss function:

$$\mathcal{L}_\pi = \mathcal{L}^w_{\text{DT}} - \alpha \cdot \mathbb{E}_{\tau\in\mathcal{D}_{\text{tar}}\cup\mathcal{D}_{\text{src}}} \left[\frac{I_m(\tau^S)}{K} \sum_{i=t-K+1}^{t} \exp(\eta_w d^w_i) Q_\phi(s, \pi(s)) \right].$$

end for

Normalisation of OT-based deviations. To make the OT-derived deviations d^w_i numerically stable across tasks and batches, we apply a min–max normalisation that shifts the range to $[-1, 0]$:

$$\hat{d}^w_i = \frac{d^w_i - \max_{j\in\mathcal{D}_{\text{src}}} d^w_j}{\max_{j\in\mathcal{D}_{\text{src}}} d^w_j - \min_{j\in\mathcal{D}_{\text{src}}} d^w_j}. \tag{31}$$

This mapping guarantees $\hat{d}^w_i \in [-1, 0]$, hence the exponential weights

$$w_i \;:=\; \exp(\eta_w\, \hat{d}^w_i) \;\in\; \left[e^{-\eta_w},\, 1\right] \tag{32}$$

are bounded, preventing gradient explosion while still down-weighting OT-distant (less feasible) source fragments. Practically, Eq. (31) makes weighting scale-free across domains and robust to outliers in d^w. We use w_i both in critic fitting and in weighted DT losses on the source batch (see blue terms in Alg. 1).

Command network C_ω trained via expectile regression. The command network C_ω produces a *value-consistent command token* that serves as an RTG replacement during inference. Concretely,

Algorithm 2 DFDT Inference

Require: Trained DT policy π_ϕ, trained command network C_ω, sequence length K, (optional) normalization stats (μ_A, σ_A) from training, environment $\mathcal{M}_T$

// No critics or OT/MMD are needed at test time. We only use C_ω to produce command tokens and π_ϕ to act.

Initialise circular buffers for the last K tokens:

$$\mathsf{S} \leftarrow [], \quad \mathsf{A} \leftarrow [], \quad \mathsf{C} \leftarrow [], \quad \mathsf{T} \leftarrow [], \quad \mathsf{M} \leftarrow []$$

Reset environment; receive initial state s_1 and set $t \leftarrow 1$

while episode not terminal **do**

  // Compute command token from the current state

  $c_t^{\text{raw}} \leftarrow C_\omega(s_t)$

  if training used standardized advantages (cf. Eq. (34)) **then** $c_t \leftarrow c_t^{\text{raw}}$ **else** $c_t \leftarrow \dfrac{c_t^{\text{raw}} - \mu_A}{\sigma_A + \varepsilon}$ **end if**

  // Update rolling context (pad left with zeros and mask invalid tokens)

  Append s_t to S, c_t to C, t to T, and 1 to M; keep only the last K entries of each

  Let $\mathsf{S}_{t-K+1:t}$, $\mathsf{C}_{t-K+1:t}$, $\mathsf{T}_{t-K+1:t}$, $\mathsf{M}_{t-K+1:t}$ be the length-K sequences after left-padding with zeros;

  Define $\mathsf{A}_{t-K+1:t-1}$ as the last $K-1$ actions (left-padded with zeros); if $t = 1$ then $\mathsf{A}_{t-K+1:t-1}$ is all zeros and the first mask entries in M are 0

  // Policy inference with command-conditioned tokens

  $a_t \leftarrow \pi_\phi\big(\mathsf{S}_{t-K+1:t},\ \mathsf{A}_{t-K+1:t-1},\ \mathsf{C}_{t-K+1:t},\ \mathsf{T}_{t-K+1:t},\ \mathsf{M}_{t-K+1:t}\big)$

  Execute a_t in $\mathcal{M}_T$; observe (r_t, s_{t+1})

  Append a_t to A (keep last $K-1$); set $t \leftarrow t+1$ and $s_t \leftarrow s_t$

end while

return trajectory $\tau = \{(s_t, a_t, r_t)\}_{t=1}^{T}$

we first form per-token advantages from the auxiliary value estimators:

$$A_i \;:=\; Q_\psi(s_i, a_i) \;-\; V_\varphi(s_i), \tag{33}$$

optionally standardized within $\mathcal{D}_{\text{src}}$ to improve numerical stability:

$$\tilde{A}_i \;:=\; \frac{A_i - \mu_A}{\sigma_A + \varepsilon}, \tag{34}$$

with (μ_A, σ_A) the mean and std of $\{A_i\}$. We then train C_ω to predict a high-expectile summary of the advantage distribution conditioned on the current state, using the asymmetric least-squares (ALS) loss from expectile regression:

$$\mathcal{L}_C(\omega) \;=\; \mathbb{E}_{(s_i,a_i)\sim\mathcal{D}_{\text{tar}}\cup\mathcal{D}_{\text{src}}}\Big[\,\big|\zeta - \mathbf{1}\{\tilde{A}_i - C_\omega(s_i) < 0\}\big|\,\big(\tilde{A}_i - C_\omega(s_i)\big)^2 \cdot \big(I_m(\tau^S)\, w_i\big)\Big], \tag{35}$$

where $\zeta \in (0.5, 1)$ (e.g., $0.7 \sim 0.9$) emphasises the upper tail of advantages to encode optimistic but value-grounded commands, $I_m(\tau^S)$ gates source fragments based on MMD distances, and w_i is the OT-based weight from Eq. (32). The ALS penalty $\rho_\zeta(u) = |\zeta - \mathbf{1}\{u < 0\}|u^2$ makes over-predicting low-advantage states costly while being tolerant to under-predicting high-advantage noise, improving robustness to cross-domain reward and horizon shifts.

Usage of C_ω at inference. At test time, we compute $c_t = C_\omega(s_t)$ and feed it as the conditioning token to the DT in place of RTG. Because c_t summarises the state-conditional advantage landscape learned from mixed (but filtered & reweighted) data, it supplies a reward- and horizon-agnostic guidance signal that remains consistent under cross-domain shifts, stabilizing token-level conditioning and mitigating stitching artifacts.

Hyperparameter overview. Table 3 summarises a compact Transformer backbone for DFDT (multi-head attention with moderate depth, width, and context length), trained with standard optimization and stabilisation choices (Adam, dropout, ReLU, soft target updates, and a fixed discount).

Method-specific settings for DFDT include pretrained critics and a command network and an OT-based filtering module (cosine cost) coupled with a fixed fragment-filtering ratio and a source–target mixing coefficient. We adopt asymmetric batch sizes to emphasise target-domain learning while still leveraging filtered source fragments. The defaults were chosen from small grids over context length and mixing strength and found robust across seeds and tasks.

Table 3: Default hyperparameter setup for DFDT.

Hyperparameter	**Value**
Number of layers	4
Number of attention heads	4
Embedding dimension	256
Context length K	$\{5, 10, 20\}$
Dropout	0.1
Learning rate	3×10^{-4}
Optimizer	Adam (Kingma & Ba, 2015)
Discount factor	0.99
Nonlinearity	ReLU
Target update rate	5×10^{-3}
Pretrained Q network hidden size	256
Pretrained V network hidden size	256
Command network hidden size	256
Number of sampled latent variables M	10
Standard deviation of Gaussian distribution	$\sqrt{0.1}$
OT Cost function	cosine
Data filtering ratio $\xi\%$	25%
Policy coefficient β	$\{0.5, 0.6, 0.7\}$
Source domain Batch size	64
Target domain Batch size	128

G Wider experimental results

G.1 Experimental results under gravity shifts

We further report comprehensive results for gravity shifts in Table 4. DFDT attains the best mean performance on **20** out of **36** tasks and achieves the highest total normalized score of **1300.2**, exceeding IQL by **57.6**% (1300.2 vs. 825.0), the second–best approach OTDF by **12.0**% (1300.2 vs. 1160.7), and the strong sequence baseline QT by **27.4**% (1300.2 vs. 1020.3). Breaking down by environment family, DFDT dominates *Hopper* (wins **8** out of **9**) and *Ant* (wins **9** out of **9**), remains competitive on *Walker2d* (wins **2** out of **9**), while *HalfCheetah* is largely led by QT. Notably, DFDT delivers large margins in challenging settings such as *hopp-m-e/expert* ($\mathbf{75.4} \pm 19.0$) and *ant-m/medium* ($\mathbf{61.0} \pm 8.7$), reflecting robust cross-dynamics stitching. Overall, these results corroborate DFDT's offline policy adaptation strength under gravity shifts, complementing its competitiveness on the remaining tasks.

G.2 Extra ablation experiments

Ablation setup. To test whether DFDT stabilizes sequence semantics, we ablate its two key components while keeping the backbone, training budget, and evaluation protocol fixed. (i) *Two-level filtering*: we replace the full MMD + OT gate with MMD-only, OT-only, or None (no filtering). (ii) *Advantage relabeling*: we replace DFDT's weighted advantage tokens with either a critic-only target learned from the mixed data (no advantage tokens) or an advantage variant that computes ACT-style advantages without any filtering when estimating advantages (Gao et al., 2024). We report normalized scores on four representative tasks spanning gravity, morphology, and kinematics (Table 5). Because the MMD+OT filtering leaves noncontiguous, cross-dynamics fragments with

Table 4: **Performance comparison of cross-domain offline RL algorithms given gravity shifts.** The meanings of each abbreviation are the same as those listed in Table 1. We **bold** and highlight the best cell.

Source	Target	IQL	DARA	IGDF	OTDF	DT	QT	DADT	DFDT
half-m	medium	39.6	**41.2**	36.6	40.7	28.4	40.2	36.6	7.3±4.3
half-m	medium-expert	39.6	40.7	38.7	28.6	45.1	**62.1**	34.7	7.8±2.4
half-m	expert	42.4	39.8	39.6	36.1	41.8	**49.1**	45.7	13.8±11.7
half-m-r	medium	20.1	17.6	14.4	21.5	18.3	**51.6**	25.3	5.9±2.5
half-m-r	medium-expert	17.2	20.2	10.0	14.7	17.2	2.1	**27.1**	5.7±2.4
half-m-r	expert	20.7	22.4	15.3	11.4	7.8	2.5	**23.6**	17.9±10.0
half-m-e	medium	38.6	37.8	37.7	39.5	35.1	**69.3**	44.0	5.6±2.6
half-m-e	medium-expert	39.6	39.4	40.7	32.4	38.2	**67.0**	32.0	6.0±2.9
half-m-e	expert	43.4	45.3	41.1	26.5	40.7	**68.5**	37.8	21.9±8.3
hopp-m	medium	11.2	17.3	15.3	32.4	19.7	16.1	12.8	**48.4**±16.3
hopp-m	medium-expert	14.7	15.4	15.1	24.2	11.6	12.8	11.6	**56.7**±23.3
hopp-m	expert	12.5	19.3	14.8	**33.7**	11.0	12.3	12.7	22.7±11.2
hopp-m-r	medium	13.9	10.7	15.3	31.1	14.2	19.9	22.6	**58.8**±27.5
hopp-m-r	medium-expert	13.3	12.5	15.4	24.2	13.7	22.3	16.6	**66.4**±17.7
hopp-m-r	expert	11.0	14.3	16.1	31.0	19.6	18.7	21.5	**42.4**±16.6
hopp-m-e	medium	19.1	18.5	22.3	26.4	13.0	14.3	11.6	**54.0**±21.6
hopp-m-e	medium-expert	16.8	16.0	16.6	28.3	13.6	14.4	11.7	**39.2**±27.8
hopp-m-e	expert	20.9	23.9	26.0	44.9	13.1	14.0	13.2	**75.4**±19.0
walk-m	medium	28.1	28.4	22.1	36.6	36.2	29.5	37.4	**43.1**±7.2
walk-m	medium-expert	35.7	30.7	35.4	44.8	38.2	**45.2**	29.1	21.5±4.5
walk-m	expert	37.3	36.0	36.2	44.0	46.4	44.0	**54.0**	22.6±5.7
walk-m-r	medium	14.6	14.1	11.6	32.7	28.6	18.9	24.8	**44.1**±2.9
walk-m-r	medium-expert	15.3	15.9	13.9	**31.6**	26.9	20.0	29.8	22.7±7.0
walk-m-r	expert	15.8	15.7	15.2	**31.3**	28.0	28.6	20.1	26.7±11.8
walk-m-e	medium	39.9	41.6	33.8	30.2	42.5	**56.7**	45.5	41.4±3.2
walk-m-e	medium-expert	49.1	45.8	44.7	53.3	39.4	**55.8**	30.6	23.6±5.1
walk-m-e	expert	40.4	56.4	45.3	**61.1**	39.6	47.4	34.5	23.6±8.9
ant-m	medium	10.2	9.4	11.3	45.1	22.0	15.3	12.4	**61.0**±8.7
ant-m	medium-expert	9.4	10.0	9.4	33.9	17.7	14.1	14.0	**52.8**±15.7
ant-m	expert	10.2	9.8	9.7	33.2	18.9	15.7	13.7	**58.3**±5.8
ant-m-r	medium	18.9	21.7	19.6	29.6	18.8	13.9	21.4	**66.9**±8.5
ant-m-r	medium-expert	19.1	18.3	20.3	25.4	13.9	13.6	18.5	**44.9**±5.5
ant-m-r	expert	18.5	20.0	18.8	24.5	14.6	10.6	17.7	**38.8**±11.1
ant-m-e	medium	9.8	8.1	8.9	18.6	11.3	11.6	20.6	**63.9**±2.1
ant-m-e	medium-expert	9.0	6.4	7.2	34.0	18.0	12.2	15.2	**45.7**±18.0
ant-m-e	expert	9.1	10.4	9.2	23.2	11.6	10.0	15.3	**41.2**±11.9
Total Score		825.0	851.0	803.6	1160.7	874.7	1020.3	895.7	**1300.2**

inconsistent horizons and reward scales, traditional RTG is ill-defined on the filtered source data and thus not comparable with DEDT.

Answer to (b): Does DFDT provide stable sequence semantics for policy adaptation? Yes. Ablations that isolate two–level filtering from reweighted advantage relabeling show that under *kinematic* shifts, removing either sequence-aware conditioning or the MMD+OT gate yields clear drops; both are needed to keep actions feasible and stitch junctions smooth. For *morphology*, all variants cluster tightly, indicating semantics are already coherent and DFDT preserves this stability. Under *gravity*, occasional wins by "no filtering" suggest over-pruning; a softer retention works better there. Overall, DFDT provides stable, value-consistent token semantics for adaptation-use the full model on challenging dynamics, and a softened gate in gravity-dominant regimes.

H LLM USAGE

The authors acknowledge that LLMs were used in the following scenarios:

Table 5: Ablation of two-level filtering and advantage relabeling.

			Filtering				Relabeling		
Source	Target	Shift	DFDT	MMD	OT	None	DFDT	Value	Advantage
ant-m-e	expert	gravity	46.5	48.2	41.4	**52.0**	46.5	**47.4**	44.6
half-m-r	medium	morph	**43.6**	43.0	**43.6**	**43.6**	43.6	**43.8**	43.2
hop-m-r	medium	kinematic	64.6	**64.7**	61.2	61.4	**64.6**	54.4	50.7
walk-m	medium	gravity	39.9	42.2	44.5	**49.4**	39.9	38.6	**40.2**

- Ablation experiments: ChatGPT (OpenAI, 2025) provided the suggestion to use action jumps, Q-value jumps, and TD residuals to show the stable, value-consistent sequence semantics of DFDT. The code generated by the LLM was adapted by the authors to be included in the code.
- Code generation: Various models accessed through GitHub Copilot (GitHub, 2025) were used to write sections of the code base, including documentation and utilities for conducting experiments. Visualization scripts, including the code for producing Figure 2. Copilot's autocomplete feature was also used throughout the code base for general assistance.
- Paper writing: Templates for tables and algorithms were generated using ChatGPT, based on screenshots of previous work such as figures from (Lyu et al., 2025a). The values inside these tables have been reviewed to ensure that they are correct. Furthermore, we use ChatGPT to polish the language in the paper.