TEMPORALLY DIRECTED GROMOV–WASSERSTEIN IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain imitation learning studies how to transfer expert behavior between agents with different embodiments. Gromov–Wasserstein matching enables comparison across different state-action spaces without learning an explicit mapping, but its global coupling makes each imitation reward depend on the complete trajectory and ignores temporal ordering. We propose Temporally Directed Gromov–Wasserstein Imitation Learning (GWτdIL), which replaces full pairwise geometry with temporally ordered comparisons of local transition magnitude and directional variation. This produces per-transition imitation rewards from a single expert demonstration while preserving cross-domain structural comparison. For equal-length trajectories with uniform weights, the resulting empirical objective can be computed in linear time. We establish sufficient conditions for policy-optimality preservation and show that task and embodiment matching impose complementary constraints on the solution space. Experiments across domains with different state dimensions and morphologies show that GWτdIL consistently learns target-directed behavior and outperforms prior cross-domain imitation methods under comparable assumptions. Ablations isolate the roles of temporal ordering, directional information, and task–embodiment matching.

1 INTRODUCTION

Reinforcement learning (RL) has achieved impressive results across diverse domains (Vinyals et al., 2019; Schrittwieser et al., 2020). However, applying RL often requires careful reward design and extensive tuning. Imitation learning (IL) addresses this difficulty by learning directly from expert demonstrations (Ho & Ermon, 2016; Dadashi et al., 2020). Conventional IL typically assumes that the expert and learner share the same embodiment and state-action spaces, requiring demonstrations from the learner domain. This assumption rarely holds in realistic scenarios. In particular, this restrictive setting seems merely to shift the problem from designing rewards to obtaining suitable expert demonstrations. Since expert demonstrations from the same domain of learner are usually hard to obtain, relaxing this assumption and allowing agents to learn from experts with different morphologies and dynamics would broaden the possibilities for training artificial agents.

Learning from expert demonstrations collected in a different domain is commonly referred to as cross-domain imitation learning (CDIL) (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018; Kim et al., 2020; Raychaudhuri et al., 2021). A common strategy is to learn a mapping between the different domains through proxy tasks. Designing these tasks requires prior knowledge of both domains, which limits transfer to previously unseen experts or new robots. Franzmeyer et al. (2022) retain task-relevant expert information while discarding domain-specific features. This highlights an inherent ambiguity in CDIL whereby matching task progress and reproducing the expert's movement structure may not identify the same learner policy. Fickinger et al. (2022) avoid learning an explicit cross-domain mapping by comparing within-domain pairwise distances through the Gromov–Wasserstein distance. While very promising, its per-step rewards depend on the complete trajectory and the global transport coupling, complicating local credit assignment and increasing computational cost. More broadly, CDIL is generally an under-defined problem and matching task outcomes and reproducing the expert's movement patterns may result in different learner behaviors. These observations raise a common question: what information is actually needed to guide cross-domain imitation learning?

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

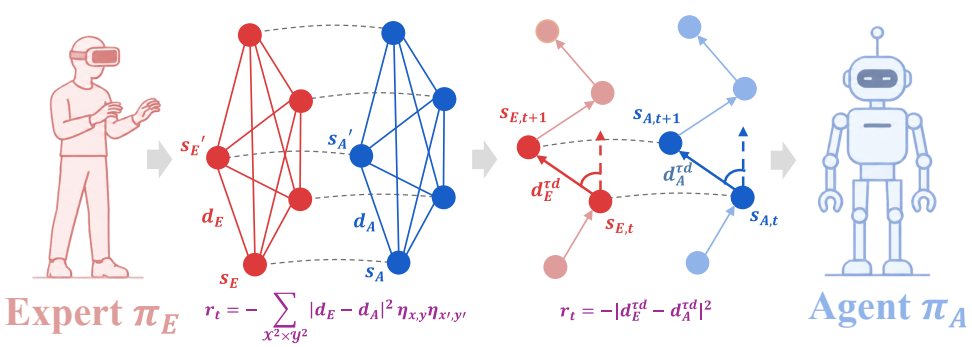


Figure 1: Overview of GWτdIL. GWIL compares global pairwise geometry over complete trajectories, whereas GWτdIL compares temporally aligned local transition magnitude and directional variation, yielding per-transition imitation rewards across different state-action spaces.

To answer this question, we propose Temporally Directed Gromov–Wasserstein Imitation Learning (GWτdIL), which replaces full pairwise geometry with temporally ordered comparisons of local transition magnitude and directional variation. These comparisons yield per-step imitation rewards from a single expert demonstration, without proxy tasks or a learned cross-domain mapping. For equal-length, uniformly weighted trajectories, the empirical comparison takes linear time. We formally characterize the scenario where GWτdIL preserves policy optimality and characterize how task and embodiment matching jointly constrain the solution space. Our experiments show that GWτdIL can learn optimal behaviors with a single demonstration while other methods learn slower or fail. And we conduct several ablation studies to examine how directional information, temporal order, and task–embodiment weighting shape the behavior of learner.

2 Related Work

Imitation learning. An early and still widely used IL approach, Behavioral Cloning, learns a policy directly from expert demonstrations through supervised learning (Pomerleau, 1988; 1991). Another key approach inverse Reinforcement Learning recovers a reward under which the expert behavior is optimal (Ng & Russell, 2000; Abbeel & Ng, 2004) to train agents via RL. Ho & Ermon (2016) instead formulate imitation as occupancy-measure matching and train a discriminator to distinguish expert from learner state-action samples. Moreover, PWIL (Dadashi et al., 2020) and SIL (Papagiannis & Li, 2020) replace this adversarial objective with optimal transport between expert and learner occupancy measures to alleviate the training instability. Recent work further investigates what metric and representation should be used for such distribution matching. PW-DICE matches state occupancies using a primal Wasserstein objective with a contrastively learned ground metric (Yan et al., 2024). LWAIL instead performs Wasserstein matching in a dynamics-aware latent space learned from state-only data (Yang et al., 2026). These methods all assume a shared or directly comparable state-action spaces, making local-level credit directly available. Our work extends this line of work by relaxing this assumption, taking the OT view of imitation beyond directly comparable spaces and into the cross-domain setting.

Cross-domain imitation learning. A common strategy for CDIL is to first make the expert and learner spaces comparable. Earlier methods achieve this through handcrafted correspondences, paired and time-aligned demonstrations, or proxy tasks available in both domains (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018; Kim et al., 2020; Raychaudhuri et al., 2021). XIRL uses temporal cycle consistency across demonstrations from multiple embodiments to learn an embedding of task progress, from which a distance-to-goal reward is defined (Zakka et al., 2022). UDIL asks what information matters for CDIL and selects task-relevant expert-state dimensions using a mutual information criterion and learns a learner encoder through adversarial transition matching in this reduced expert space (Franzmeyer et al., 2022). More recent work further explores which information should be retained for cross-embodiment transfer. XSkill learns cross-embodiment skill prototypes from human and robot videos (Xu et al., 2023), while X-Sim uses object motion extracted

from human demonstrations as an embodiment-independent reward for policy learning in simulation (Dan et al., 2025). These approaches differ in their supervision requirements but enable transfer by learning or selecting information that can be compared across embodiments. GWIL takes a different view. It compares within-domain pairwise distances through the Gromov–Wasserstein distance, avoiding an explicit cross-domain mapping or shared latent space (Fickinger et al., 2022). However, its global coupling makes per-step rewards depend on complete trajectories. Our work extends GWIL from global geometric matching to temporally directed local structural matching, allowing imitation rewards to be computed directly from corresponding transitions.

3 PRELIMINARIES

Metric Markov Decision Process: We describe an infinite time horizon Markov decision process (MDP) as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, p_0, \gamma,\ T)$, where $\mathcal{S}$ is the state spaces, $\mathcal{A}$ is the action spaces, R is the reward function, P is the transition function, p_0 is the initial-state distribution and γ is the discount factor (Sutton & Barto, 2018). We equip MDPs with a distance $d : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$(Fickinger et al., 2022) and refer to $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, p_0, \gamma, T, d)$ as a metric MDP.

Gromov–Wasserstein Distance: Let $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y})$ be two metric-measure spaces, where $\mathcal{X}$ and $\mathcal{Y}$ are point sets, $d_\mathcal{X}$ and $d_\mathcal{Y}$ are distance and $\mu_\mathcal{X}$ and $\mu_\mathcal{Y}$ are measures on their respective spaces. The Gromov–Wasserstein (GW) distance (Mémoli, 2011; Peyré et al., 2016; Vayer, 2020) between metric measure spaces is defined by

$$GW^2\left((\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X}), (\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y})\right) = \inf_{\eta \in \Pi(\mu_\mathcal{X}, \mu_\mathcal{Y})} \sum_{\mathcal{X}^2 \times \mathcal{Y}^2} \left|d_\mathcal{X}(x, x') - d_\mathcal{Y}(y, y')\right|^2 \eta_{x,y} \eta_{x',y'}, \quad (1)$$

where $\Pi(\mu_\mathcal{X}, \mu_\mathcal{Y}) = \left\{\eta \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{Y}|} \;\middle|\; \sum_y \eta_{x,y} = \mu_\mathcal{X}(x),\ \sum_x \eta_{x,y} = \mu_\mathcal{Y}(y)\right\}$ is the couplings matrix between the two measures. GW compares the pairwise intrinsic distances to compare the global structure of the metric measure spaces with different dimension. Figure 1 illustrates the alignment in the case of the metric measure space $(S_E \times A_E, d_E, \rho_{\pi_E})$ and $(S_A \times A_A, d_A, \rho_{\pi_A})$ between expert policy and learner policy.

GWτ Distance: Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$ have uniform empirical measures μ_X and μ_Y on their sample indices. Taking $r_X = x_1$ and $r_Y = y_1$, define $d_X^\tau(r_X, x_i) = \sum_{k=1}^{i-1} \|x_{k+1} - x_k\|_2$, and similarly for Y. The GWτ Distance between two trajectory is defined by,

$$GW_\tau^2(X, Y) = \inf_{\eta \in \Pi(\mu_X, \mu_Y)} \sum_{\mathcal{X} \times \mathcal{Y}} \left|d_X^\tau\left(r_X, x_i\right) - d_Y^\tau\left(r_Y, y_j\right)\right|^2 \eta_{ij}. \quad (2)$$

GWτ compares ordered trajectories through their intrinsic distances from their starting points (Kravtsova et al., 2023), and can be seen as a one-dimensional Wasserstein distance between distributions of cumulative path lengths.

4 TEMPORALLY DIRECTED GROMOV–WASSERSTEIN IMITATION LEARNING

4.1 COMPARING POLICIES UNDER GWτD EQUIVALENCE

For a stationary policy π acting on a metric MDP, the normalized discounted occupancy measure is $\rho_\pi(s, a) = (1 - \gamma)\pi(a \mid s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi)$. Policies from arbitrarily different MDPs can be compared in terms of their occupancy measures (Fickinger et al., 2022).

For empirical trajectory comparison, we use uniform weights over finite rollout indices rather than approximate the discounted occupancy measure. For $D \in \{E, A\}$, write $z_{D,t} = (s_{D,t}, a_{D,t})$, and let $\mathcal{X}_D$ denote the set of temporally indexed samples equipped with the uniform measure μ_D. Repeated visits to the same state-action pair are treated as distinct samples. We reserve ρ_π for discounted occupancy analysis and use μ_D for trajectory comparison.

We therefore define GWτd distance which denotes the comparison of the specified rollouts of the two policies. All proofs are provided in Appendix A.

Definition 1 (Temporally directed Gromov-Wasserstein distance between trajectories). Let $\tau_D = (z_{D,1}, \ldots, z_{D,T})$ be a rollout of policy π_D, where $r_D = z_{D,1}$ is the root point and $\mu_D = \frac{1}{T}\sum_{t=1}^{T}\delta_i$ is the measure. We define a temporal structural function $d_D^{\tau d}$ incorporating transition magnitude and projected directional variation. Let $\left(\mathcal{X}_E, d_E^{\tau d}, \mu_E\right)$ and $\left(\mathcal{X}_A, d_A^{\tau d}, \mu_A\right)$ be two metric-measure spaces and the temporally directed Gromov-Wasserstein distance is defined as

$$GW_{\tau d}^2\left(\pi_E, \pi_A\right) = \min_{\eta \in \Pi(\mu_E, \mu_A)} \sum_{\mathcal{X}_E \times \mathcal{X}_A} \left|d_E^{\tau d}\left(r_E, z_E\right) - d_A^{\tau d}\left(r_A, z_A\right)\right|^2 \eta_{z_E, z_A}, \tag{3}$$

where

$$d_D^{\tau d}\left(r_D, z_{D,i}\right) = \sum_{k=1}^{i-1}\left(\left\|z_{D,k+1} - z_{D,k}\right\| + \left\|\mathbf{1}_D^{\top}\left(z_{D,k+1} - z_{D,k}\right)\right\|\right), \tag{4}$$

where $\mathbf{1}_D$ is the all-ones vector in the domain' s sample space. This compares the distributions of temporally directed structural distances from the two reference points.

We now define policy isomorphisms that preserve this local structure and time progression.

Definition 2 (Temporally directed isomorphic policies). Two policies π_E and π_A are temporally directed isomorphic if there exist bijections $\phi_E : \mathcal{X}_E \to \mathcal{Z}$ and $\phi_A : \mathcal{X}_A \to \mathcal{Z}$ with $\mathcal{Z} \subseteq \mathcal{Z}_{mag} \times \mathcal{Z}_{dir}$, where $\mathcal{Z}_{mag}$ and $\mathcal{Z}_{dir}$ represent transition magnitude and directional information, respectively, such that the correspondence $T = \phi_A^{-1} \circ \phi_E : \mathcal{X}_E \to \mathcal{X}_A$ satisfies

$$T\left(r_E\right) = r_A \tag{5}$$

$$d_E^{\tau d}\left(z_E, z_E'\right) = d_A^{\tau d}\left(T\left(z_E\right), T\left(z_z'\right)\right) \tag{6}$$

$$\rho_{\pi_A}\left(z_A\right) = \rho_{\pi_E}\left(T^{-1}\left(z_A\right)\right). \tag{7}$$

Since ϕ_E and ϕ_A are bijections onto the same space, T is a bijection with inverse $T^{-1} = \phi_E^{-1} \circ \phi_A$. In other words, ρ_{π_A} is the push-forward measure $T_{\#}\rho_{\pi_E}$ and T is an isometry between $\left(S_E \times A_E, d_E^{\tau d}, \rho_{\pi_E}\right)$ and $\left(S_A \times A_A, d_A^{\tau d}, \rho_{\pi_A}\right)$

We first express GWτd as a Wasserstein distance between the two occupancy measures pushforward by $d_D^{\tau d}$. This provides a direct way to compute the distance easily and establish its metric properties.

Proposition 1. For the GWτd distance between π_E and π_A, there exists

$$GW_{\tau d}^q\left(\pi_E, \pi_A\right) = W^q\left(\left(d_E^{\tau d}\left(r_E, \cdot\right)\right)_{\#}\mu_E\ ,\ \left(d_A^{\tau d}\left(r_A, \cdot\right)\right)_{\#}\mu_A\right). \tag{8}$$

That is to say, for empirical trajectories with N samples each and uniform weights, this distance can be computed in $\mathcal{O}(N)$ time, avoiding the quadratic coupling optimization of standard GW.

Proposition 2. GWτd defines a pseudo-metric on the collection of all temporally directed isomorphic classes of policies.

Theorem 1. Let $\phi : \left(S_E, d_E^S\right) \to \left(S_A, d_A^S\right)$ and $\psi : \left(A_E, d_E^A\right) \to \left(A_A, d_A^S\right)$ be two isometries satisfied the reference-point and distance conditions in definition 2. For all $\left(s_E, a_E, s_E'\right) \in S_E \times A_E \times S_E$, there exists:

$$R_E\left(s_E, a_E\right) = R_A\left(\phi\left(s_E\right), \psi\left(a_E\right)\right) \tag{9}$$

$$P_{E\, s_E, a_E}\left(s_E'\right) = P_{A\, \phi(s_E), \psi(a_E)}\left(\phi\left(s_E'\right)\right) \tag{10}$$

$$p_E\left(s_E\right) = p_A\left(\phi\left(s_E\right)\right). \tag{11}$$

Let π_E^* be an optimal policy in M_E. If $\pi_{GW_{\tau d}}$ minimizes $GW_{\tau d}(\pi_E^*, \pi_{GW_{\tau d}})$, then $\pi_{GW_{\tau d}}$ is temporally directed isomorphic to an optimal target policy in M_A.

Computing $GW_{\tau d}$ does not require knowing the isometries assumed in Theorem 1. The next theorem shows how task and embodiment matching impose complementary constraints on imitation.

Theorem 2. Let $D_T(\pi_A) = GW^q_{\tau d,T}(\pi_E, \pi_A)$ and $D_J(\pi_A) = GW^q_{\tau d,J}(\pi_E, \pi_A)$. For $0 < \lambda < 1$, define

$$D_\lambda(\pi_A) = \lambda D_T(\pi_A) + (1-\lambda) D_J(\pi_A). \tag{12}$$

Denote the exact-matching sets by $\Pi_T^\star = \{\pi_A : D_T(\pi_A) = 0\}$ and $\Pi_J^\star = \{\pi_A : D_J(\pi_A) = 0\}$. If $\Pi_T^\star \cap \Pi_J^\star \neq \varnothing$, then we have $\arg\min_{\pi_A} D_\lambda(\pi_A) = \Pi_T^\star \cap \Pi_J^\star$.

Moreover, neither exact-matching condition implies the other in general. Hence, if both set differences are nonempty, then $\Pi_T^\star \cap \Pi_J^\star \subsetneq \Pi_T^\star$ and $\Pi_T^\star \cap \Pi_J^\star \subsetneq \Pi_J^\star$.

Remark 1: Theorem 2 characterizes the complementary roles of task and embodiment matching. Task matching constrains task progress, while embodiment matching constrains the motion used to achieve it. When simultaneous exact matching is feasible, their combination restricts the solution set to $\Pi_T^\star \cap \Pi_J^\star$, while the problem may remain under-defined within this intersection. Our experiments examine the contribution of each component and their combination.

4.2 From Global Alignment to Local Credit

We construct local imitation rewards using the temporal correspondence established in Proposition 1. In this section, we set $q = 2$ and compare state observations.

Definition 3 (Local temporally directed imitation reward). Given two equal-length trajectories with uniform weights from policies π_E and π_A, the per-transition temporally directed Gromov-Wasserstein reward is defined as $r_{GW_{\tau d}} : (s_A, a_A, s'_A) \to \mathbb{R}$ given by

$$r_{GW_{\tau d}}(z_{A,t}) = -\frac{1}{\rho_\pi(z_A)} \left| d_E^{\tau d}(z_{E,t}, z_{E,t+1}) - d_A^{\tau d}(z_{A,t}, z_{A,t+1}) \right|^2 \tag{13}$$

Under the exact-matching setting considered above, we obtain the following equivalence between the global objective and its local reward construction.

Proposition 3. If π_A minimizes $GW_{\tau d}(\pi_E, \pi_A)$, then π_A is an optimal policy for the local reward $r_{GW_{\tau d}}$ as defined in Definition 3.

Following Theorem 2, we evaluate the task and embodiment components separately. With $s_D = (x_D, q_D)$, the grouped empirical objective is

$$\begin{aligned} GW^2_{\tau d}(\tau_E, \tau_A) = \min_{\theta \in \Theta^{T_E \times T_A}} \sum_{\substack{1 \le t \le T_E \\ 1 \le t \le T_A}} \Big[& \\ & \big(\lambda |d_{E\tau d}(x_{E,t}, x_{E,t+1}) - d_{A\tau d}(x_{A,t}, x_{A,t+1})|\big)^2 \\ & + (1-\lambda) |d_{E\tau d}(q_{E,t}, q_{E,t+1}) - d_{A\tau d}(q_{A,t}, q_{A,t+1})|^2) \Big] \end{aligned} \tag{14}$$

We construct its local reward surrogate by comparing the corresponding one-step distances:

$$\begin{aligned} r_{GW_{\tau d}}(s_{A,t}, s_{A,t+1}) = -\big(\lambda |d_{E\tau d}(x_{E,t}, x_{E,t+1}) - d_{A\tau d}(x_{A,t}, x_{A,t+1})|\big)^2 \\ - (1-\lambda) |d_{E\tau d}(q_{E,t}, q_{E,t+1}) - d_{A\tau d}(q_{A,t}, q_{A,t+1})|^2) \end{aligned} \tag{15}$$

For $0 < \lambda < 1$, both discrepancies must vanish to attain zero reward, so Proposition 3 applies componentwise. The index t specifies the reference transition, and the reward can be evaluated as

Algorithm 1 GWτdIL from a single expert demonstration.

Inputs: an expert demonstration τ_E, task weight λ, episode horizon T
Initialize the episode-step index $t = 0$, imitation agent's policy π_θ and value estimates V_θ
Compute the expert's transition features from τ_E
while $t < T$ **do**
Collect transitions (s_A, a_A, s'_A) of learner at episode step t using π_θ
Compute local imitate pseudo-reward $r_{GW_{\tau d}}$ using Eq.(15)
Add the transitions and $r_{GW_{\tau d}}$ to replay buffer
Update π_θ and V_θ with SAC
end while

soon as the learner's next state is observed. Algorithm 1 uses these transition rewards to train the policy from a single expert demonstration.

Remark 2. Our construction retains GWIL's metric-measure view of policy comparison, but replaces global pairwise geometry and trajectory-level coupling with temporally aligned local transition structure. Each imitation reward is therefore attributable to the current learner transition and can be evaluated once the next state is observed, rather than from a complete trajectory. Including the reference index t in the state makes this local reward Markovian.

Computing complexity. For equal-length trajectories with T uniformly weighted samples, expert transition distances are precomputed in $O(Td)$ time and $O(T)$ storage, where d is the state dimension. Each learner reward requires $O(d)$ operations, giving $O(T)$ total reward-computation time.

Optimizing the pseudo-rewards. We train the learner policy with soft actor-critic (SAC; Haarnoja et al., 2018), replacing the environment reward with $r_{GW_{\tau d}}$. Algorithm 1 summarizes the training procedure using a single expert demonstration.

5 Experiments

We evaluate GWτdIL on continuous-control transfer tasks to address three questions:

1, *Can* GWτdIL *efficiently transfer behavior across different state-action dimensions and morphologies?* (Sections 5.1–5.2)

2, *How do temporal order and directional information affect trajectory comparison and policy learning?* (Section 5.3)

3, *How does the balance between task and embodiment matching affect the learned behavior?* (Section 5.3)

To answer these three questions, we use continuous control tasks implemented in DeepMind Control Suite (Tassa et al., 2018) and MuJoCo (Todorov et al., 2012). We include videos of learned policies on our project site. Unless otherwise specified, curves report the mean and shaded regions indicate ± 1 standard deviation across 5 random seeds.

5.1 Cross-domain imitation learning of Pendulum-to-CartPole

We first evaluate GWτdIL on Pendulum-to-CartPole transfer, following Fickinger et al. (2022), where the domains have different state dimensions but an apparent behavioral mapping. Without task–embodiment decomposition, the CartPole learner is trained from a single Pendulum demonstration using only GWτdIL rewards.

As shown in Figure 2, GWτdIL is able to recover the swing-up behavior and reach a native CartPole return of about 600 within 0.4M environment steps, while UDIL remains below 500 at 1M steps. Moreover, GWIL assigns a higher normalized reward to a failed rollout ($r_{GW} = -0.007471$, $r_{native} = 0.321$,) than to a successful one ($r_{GW} = -0.009268$, $r_{native} = 709.5$,), illustrating a mismatch between global structural similarity and task performance. These results show that tem-

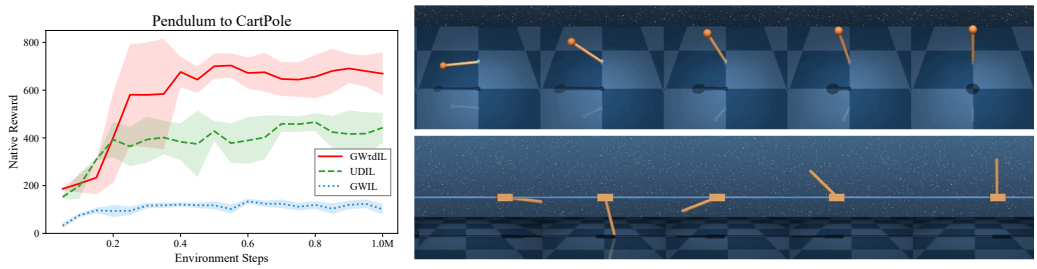


Figure 2: Pendulum-to-CartPole transfer from a single expert demonstration. Left: native CartPole return during training. Right: the Pendulum expert demonstration (top) and a learned CartPole rollout (bottom).

porally directed structural matching alone can indeed support transfer across unequal-dimensional domains, even beyond the isomorphic setting of Theorem 1.

5.2 Cross-domain imitation learning of robot control

To test transfer across larger morphological differences, we next evaluate GWτdIL on six directed transfers among Hopper, Walker, and HalfCheetah, which have been widely used to evaluate imitation and cross-domain imitation learning methods (Ho & Ermon, 2016; Gangwani & Peng, 2020; Raychaudhuri et al., 2021; Franzmeyer et al., 2022).

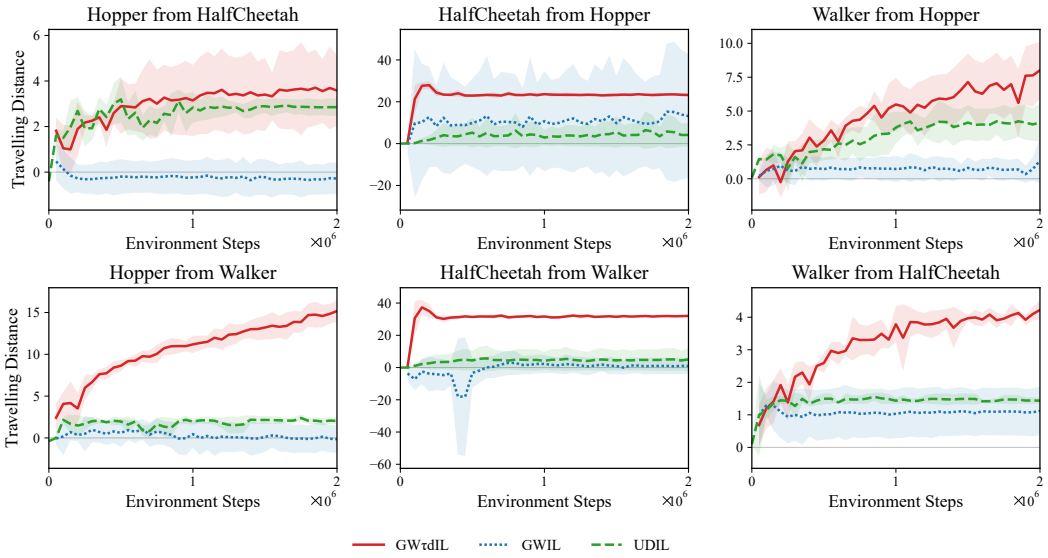


Figure 3: Cross-domain locomotion transfer among Hopper, Walker, and HalfCheetah. The six panels show all directed transfers, with curves reporting signed displacement in the target direction.

GWτdIL learns from a single expert demonstration using the task–embodiment decomposition described in Section 4.1. We use the same decomposition rule that treating joint-related states as the embodiment component and assigning the remaining states to the task component, without using environment rewards for dimension selection. Following Franzmeyer et al. (2022), we compare against GWIL and UDIL using displacement in the target direction, while the learner is trained only with imitation rewards.

As shown in Figure 3, GWτdIL achieves the highest final mean traveling distance on all six transfers. On Hopper-from-Walker, it even reaches up to 74.2% of the expert displacement. In contrast, GWIL is highly stochastic across random seeds, as reported by Choi et al. (2023) and also noted by

its authors (Fickinger et al., 2022). UDIL is more stable but is sensitive to the expert task representation, showing trends consistent with Franzmeyer et al. (2022). These results show that GWτdIL with task–embodiment decomposition can consistently learn target-directed motion across larger morphological and state-space differences.

5.3 Ablation Studies

We now isolate the effects of directional information, temporal ordering, and task–embodiment weighting on GWτdIL, focusing on Hopper-from-Walker and the more morphologically distinct Walker-from-HalfCheetah.

Omitting directional information. We remove the directional term to test whether temporal transition magnitudes alone recover motion direction. Figure 4 shows that the resulting GWτIL achieves near-zero displacement for Hopper-from-Walker and negative displacement in the other three tasks, while GWτdIL remains positive throughout. These results show that temporal magnitude matching alone does not reliably preserve forward motion. Rollouts are provided in Appendix C.1.

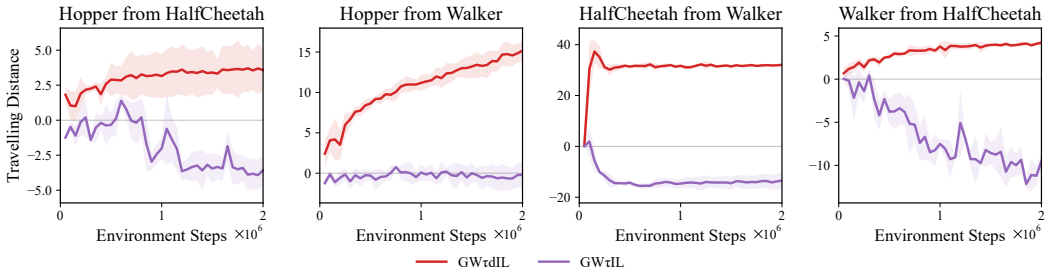


Figure 4: Directional ablation on four cross-embodiment transfers. GWτIL removes the directional term from GWτdIL, and curves report signed displacement in the target direction. Lines denote the mean and shaded regions ± 1 standard deviation across 3 random seeds.

Distinguishing temporal order. We nest compare accelerating and decelerating HalfCheetah trajectories against an accelerating reference. Across 30 rollouts per condition, GW assigns nearly identical discrepancy scores to acceleration and deceleration. In contrast, GWτ and GWτd assign local discrepancies to deceleration that are 3.42 and 2.53 times those for acceleration, respectively. This shows that temporal progression is necessary to distinguish motions with the same direction. Full results are provided in Appendix C.

Varying task and embodiment weights. Finally, we vary the task and embodiment weights to examine how their balance affects task performance and movement patterns, keeping the remaining training procedure fixed. Figure 5 shows that embodiment matching alone yields near-zero or negative mean displacement in both tasks. For Walker-from-HalfCheetah, a task weight of 0.1 is already sufficient for forward progress. Hopper learning from Walker is more sensitive to the balance: performance improves as the task weight increases from 0.1 to 0.7, while removing embodiment matching reduces it, with 0.7/0.3 giving the best final performance among the tested settings.



Figure 5: Effect of task–embodiment weighting on Hopper-from-Walker and Walker-from-HalfCheetah transfer. Labels denote task/embodiment weights, and curves report signed displacement in the target direction.

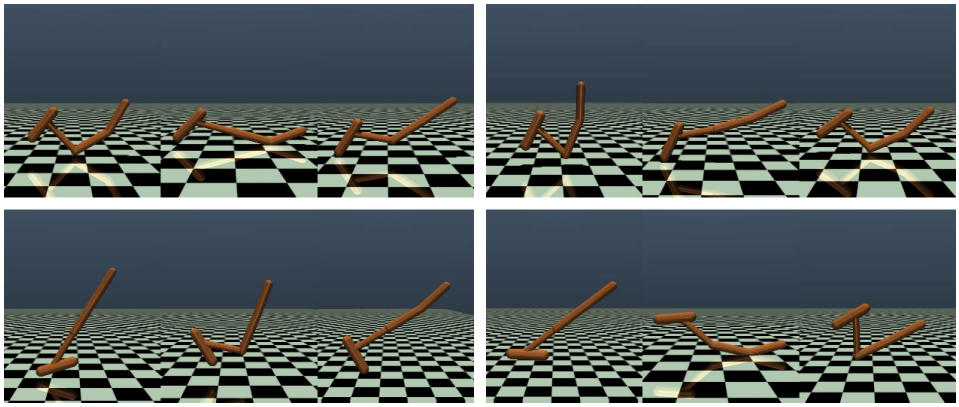


Figure 6: Hopper-from-Walker rollouts under task/embodiment weights 0.1/0.9, 0.3/0.7, 0.7/0.3, and 1/0, ordered from top left to bottom right. The resulting locomotion strategies differ across weightings.

Figure 6 shows corresponding changes in behavior. With a task weight of 0.1, Hopper moves in a crouched, swimming-like motion; at 0.7, it extends its body and pushes off with its foot. With task matching alone, it again moves close to the ground. These results suggest that task matching drives forward progress, while embodiment matching improves performance and constrains how that progress is achieved, consistent with Theorem 2.

6 Conclusion

We introduce GWτdIL, a novel cross-domain imitation learning method that enables an agent to learn from a single demonstration across different embodiments. By comparing local transition magnitude and directional variation in temporal order, GWτdIL constructs per-step imitation rewards without requiring an explicit cross-domain mapping. The method performs consistently across domains with different state dimensions and morphologies while remaining simple and computationally efficient. Theoretically, GWτdIL recovers optimal behavior only up to isometries, and although task and embodiment matching further constrain the solution space, they do not eliminate the inherent under-defined of cross-domain imitation. Consequently, structural matching alone may remain insufficient in complex, high-dimensional domains. In future work, we will investigate GWτdIL for exploration in extremely sparse-reward environments and extend the method to trajectories of unequal lengths.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript readability, reference formatting, manuscript formatting, software configuration, and routine code-related tasks. All AI-assisted content was independently reviewed and verified by the authors. The core ideas, methodological development, theoretical results, experiments, and scientific conclusions were developed and finalized by the authors.

ETHICS STATEMENT

This work uses simulated continuous-control environments and involves no human participants, animal subjects, personally identifiable information, or sensitive datasets. All experiments are conducted in DeepMind Control Suite or Gymnasium MuJoCo.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in Appendix A. Appendix B documents the environments, expert demonstrations, evaluation protocols, network architectures, training hyperparameters, and baseline implementations, while Appendix C provides additional ablations and trajectory-comparison experiments.

We include videos of learned policies on our anonymous project site (`https://anonymous.4open.science/api/repo/gwtdil-video-supplement-source-2E12/file/index.html`) and release our code in an anonymous code repository (`https://anonymous.4open.science/r/gwtdil-7C4C/README.md`).

REFERENCES

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004. doi: 10.1145/1015330.1015430.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TtYSU29zgR.

Prithwish Dan, Kushal Kedia, Angela Chao, Edward Duan, Maximus Adrian Pace, Wei-Chiu Ma, and Sanjiban Choudhury. X-Sim: Cross-embodiment learning via real-to-sim-to-real. In *Proceedings of the 9th Conference on Robot Learning*, volume 305, pp. 816–833, 2025. URL https://proceedings.mlr.press/v305/dan25a.html.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nDITHIIblzh.

Tim Franzmeyer, Philip H. S. Torr, and João F. Henriques. Learn what matters: Cross-domain imitation learning with task-relevant embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26283–26294, 2022. doi: 10.52202/068431-1906.

Tanmay Gangwani and Jian Peng. State-only imitation with transition dynamics mismatch. *arXiv preprint arXiv:2002.11879*, 2020. URL https://arxiv.org/abs/2002.11879.

Abhishek Gupta, Coline Devin, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. *arXiv preprint arXiv:1703.02949*, 2017. URL https://arxiv.org/abs/1703.02949.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th*

International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Kuno Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Domain adaptive imitation learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5286–5295, 2020. URL https://proceedings.mlr.press/v119/kim20c.html.

Natalia Kravtsova, Reginald L. McGee, and Adriana T. Dawes. Scalable gromov–wasserstein based comparison of biological time series. *Bulletin of Mathematical Biology*, 85(8):77, 2023. doi: 10.1007/s11538-023-01175-y.

YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation*, pp. 1118–1125, 2018. doi: 10.1109/ICRA.2018.8462901.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. doi: 10.1007/s10208-011-9093-5.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 663–670, 2000.

Georgios Papagiannis and Yunpeng Li. Imitation learning with sinkhorn distances. *arXiv preprint arXiv:2008.09167*, 2020. URL https://arxiv.org/abs/2008.09167.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2664–2672, 2016. URL https://proceedings.mlr.press/v48/peyre16.html.

Dean A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, volume 1, pp. 305–313, 1988. URL https://proceedings.neurips.cc/paper/1988/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html.

Dean A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991. doi: 10.1162/neco.1991.3.1.88.

Dripta S. Raychaudhuri, Sujoy Paul, Jeroen Vanbaar, and Amit K. Roy-Chowdhury. Cross-domain imitation from observations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8902–8912, 2021. URL https://proceedings.mlr.press/v139/raychaudhuri21a.html.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation*, pp. 1134–1141, 2018.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018. URL https://arxiv.org/abs/1801.00690.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Titouan Vayer. *A Contribution to Optimal Transport on Incomparable Spaces*. PhD thesis, Université Bretagne Sud, 2020. URL https://theses.fr/2020LORIS576.

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

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Çağlar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019. doi: 10.1038/s41586-019-1724-z.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. XSkill: Cross-embodiment skill discovery. In *Proceedings of the 7th Conference on Robot Learning*, volume 229, pp. 3536–3555, 2023. URL https://proceedings.mlr.press/v229/xu23a.html.

Kai Yan, Alexander G. Schwing, and Yu-Xiong Wang. Offline imitation from observation via primal wasserstein state occupancy matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 55814–55838, 2024. URL https://proceedings.mlr.press/v235/yan24e.html.

Siqi Yang, Kai Yan, Alexander G. Schwing, and Yu-Xiong Wang. Latent wasserstein adversarial imitation learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xfnR9ok8HJ.

Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. XIRL: Cross-embodiment inverse reinforcement learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 537–546, 2022. URL https://proceedings.mlr.press/v164/zakka22a.html.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proposition 1. For the GWτd distance between π_E and π_A, there exists

$$GW_{\tau d}^{q}\left(\pi_E, \pi_A\right) = W^q\left(\left(d_E^{\tau d}\left(r_E, \cdot\right)\right)_{\#}\mu_E\ ,\ \left(d_A^{\tau d}\left(r_A, \cdot\right)\right)_{\#}\mu_A\right).$$

That is to say, for empirical trajectories with N samples each and uniform weights, this distance can be computed in $\mathcal{O}(N)$ time, avoiding the quadratic coupling optimization of standard GW.

Proof. By the pushforward argument of Kravtsova et al. (2023) (Proposition 1(a)), for equal-length trajectories with uniform weights, the cumulative distances from root are already ordered, so the one-dimensional optimal transport formula gives by

$$GW^{\,q}_{\tau d}\left(\pi_E, \pi_A\right) = \frac{1}{N}\sum_{i=1}^{N}\left|d_E^{\tau d}\left(r_E, z_{E,i}\right) - d_A^{\tau d}\left(r_A, z_{A,i}\right)\right|^q.$$

Computing the cumulative distances and evaluating this sum both require $\mathcal{O}(N)$ operations, with no additional sorting. □

A.2 PROOF OF PROPOSITION 2

Proposition 2. GWτd defines a pseudo-metric on the collection of all temporally directed isomorphic classes of policies.

Proof. By Definition 2, temporally directed isomorphic policies induce the same structural pushforward measure, so $GW_{\tau d}$ is well-defined on the corresponding equivalence classes. Proposition 1 expresses $GW_{\tau d}$ as a Wasserstein distance between these pushforward measures. Nonnegativity, symmetry, and the triangle inequality therefore follow directly from the Wasserstein distance (Villani, 2009). □

A.3 PROOF OF THEOREM 1

Theorem 1. Let $\phi : \left(S_E, d_E^S\right) \to \left(S_A, d_A^S\right)$ and $\psi : \left(A_E, d_E^A\right) \to \left(A_A, d_A^A\right)$ be two bijective isometries satisfied the reference-point and distance conditions in definition 2. For all $(s_E, a_E, s'_E) \in S_E \times A_E \times S_E$, the following conditions hold:

$$R_E\left(s_E, a_E\right) = R_A\left(\phi\left(s_E\right), \psi\left(a_E\right)\right)$$

$$P_{E\,s_E, a_E}\left(s'_E\right) = P_{A\,\phi(s_E), \psi(a_E)}\left(\phi\left(s'_E\right)\right)$$

$$p_E\left(s_E\right) = p_A\left(\phi\left(s_E\right)\right).$$

Let π_E^* be an optimal policy in M_E. If $\pi_{GW_{\tau d}}$ minimizes $GW_{\tau d}\left(\pi_E^*, \pi_{GW_{\tau d}}\right)$, then $\pi_{GW_{\tau d}}$ is temporally directed isomorphic to an optimal target policy in M_A.

Proof. For readability, we give the proof for finite state-action spaces, with an analogous measure-theoretic extension. We proceed in four steps. We first construct a feasible learner occupancy measure by pushing forward the expert occupancy through (ϕ, ψ), then show that the induced policy is optimal and temporally directed isomorphic to the expert. Finally, Proposition 2 implies that any GWτd minimizer lies in the same temporally directed isomorphism class.

(a) Feasibility. Consider the occupancy measure $\rho_A^\star\left(s_A, a_A\right) = \rho_{\pi_E^\star}\left(\phi^{-1}\left(s_A\right), \psi^{-1}\left(a_A\right)\right)$,since $\rho_{\pi_E^\star}$ satisfies the discounted flow constraints, the transition and initial-distribution conditions give

$$\begin{aligned}\sum_{a_A} \rho_A^\star(s_A, a_A) &= \sum_{a_E} \rho_{\pi_E^\star}\left(\phi^{-1}(s_A), a_E\right) \\ &= (1-\gamma) p_E\left(\phi^{-1}(s_A)\right) \\ &+ \gamma \sum_{s'_E, a'_E} P_E\left(\phi^{-1}(s_A) \mid s'_E, a'_E\right) \rho_{\pi_E^\star}(s'_E, a'_E) \\ &= (1-\gamma) p_A(s_A) \\ &+ \gamma \sum_{s'_A, a'_A} P_A(s_A \mid s'_A, a'_A) \rho_A^\star(s'_A, a'_A).\end{aligned}$$

Thus, $\rho_A^\star$ is feasible in M_A, and there exists a policy $\pi_A^\star$ with occupancy measure $\rho_A^\star$.

(b) Optimality. For any policy π_A, reversing the argument in (a) shows that

$$(s_E, a_E) \longmapsto \rho_{\pi_A}(\phi(s_E), \psi(a_E))$$

is a feasible occupancy measure in M_E. Reward preservation and the optimality of $\pi_E^\star$ imply

$$\begin{aligned}\sum_{s_A, a_A} \rho_{\pi_A}(s_A, a_A) R_A(s_A, a_A) &= \sum_{s_E, a_E} \rho_{\pi_A}(\phi(s_E), \psi(a_E)) R_E(s_E, a_E) \\ &\leq \sum_{s_E, a_E} \rho_{\pi_E^\star}(s_E, a_E) R_E(s_E, a_E) \\ &= \sum_{s_A, a_A} \rho_A^\star(s_A, a_A) R_A(s_A, a_A).\end{aligned}$$

These quantities equal $(1-\gamma)$ times the corresponding discounted returns. Therefore, $\pi_A^\star$ is optimal in M_A.

(c) Structural isomorphism. By assumption, the state and action correspondences preserve the reference points and satisfy

$$d_E^{\tau d}((s_E, a_E), (s'_E, a'_E)) = d_A^{\tau d}((\phi(s_E), \psi(a_E)), (\phi(s'_E), \psi(a'_E))).$$

The definition of $\rho_A^\star$ also preserves the occupancy measure under this correspondence. Hence, $\pi_E^\star$ and $\pi_A^\star$ are temporally directed isomorphic by Definition 2.

(d) GWτd minimization. By (c) and Proposition 2,

$$GW_{\tau d}(\pi_E^\star, \pi_A^\star) = 0.$$

Since $\pi_{GW_{\tau d}}$ minimizes the same nonnegative distance, it also attains zero and is therefore isomorphic to $\pi_E^\star$. By transitivity, it is temporally directed isomorphic to $\pi_A^\star$, which is optimal in M_A. □

A.4 Proof of Theorem 2

Theorem 2. Let $D_T(\pi_A) = GW_{\tau d, T}^q(\pi_E, \pi_A)$ and $D_J(\pi_A) = GW_{\tau d, J}^q(\pi_E, \pi_A)$. For $0 < \lambda < 1$, define

$$D_\lambda(\pi_A) = \lambda D_T(\pi_A) + (1-\lambda) D_J(\pi_A). \tag{12}$$

Denote the exact-matching sets by $\Pi_T^\star = \{\pi_A : D_T(\pi_A) = 0\}$ and $\Pi_J^\star = \{\pi_A : D_J(\pi_A) = 0\}$. If $\Pi_T^\star \cap \Pi_J^\star \neq \varnothing$, then we have $\arg\min_{\pi_A} D_\lambda(\pi_A) = \Pi_T^\star \cap \Pi_J^\star$.

Moreover, neither exact-matching condition implies the other in general. Hence, if both set differences are nonempty, then $\Pi_T^\star \cap \Pi_J^\star \subsetneq \Pi_T^\star$ and $\Pi_T^\star \cap \Pi_J^\star \subsetneq \Pi_J^\star$.

Proof. Since $D_T, D_J \geq 0$ and $0 < \lambda < 1$,

$$D_\lambda(\pi_A) = 0 \iff D_T(\pi_A) = D_J(\pi_A) = 0.$$

If $\Pi_T^\star \cap \Pi_J^\star \neq \varnothing$, the minimum of D_λ is therefore zero, and

$$\arg\min_{\pi_A} D_\lambda(\pi_A) = \Pi_T^\star \cap \Pi_J^\star.$$

The strict inclusions follow whenever $\Pi_T^\star \setminus \Pi_J^\star \neq \varnothing$ and $\Pi_J^\star \setminus \Pi_T^\star \neq \varnothing$. □

A.5 Proof of Proposition 3.

Proposition 3. For equal-length trajectories with uniform weights, if π_A minimizes $GW_{\tau d}(\pi_E, \pi_A)$, then π_A is an optimal policy for the local reward $r_{GW_{\tau d}}$ as defined in Definition 3.

Proof. Suppose that π_A minimizes $GW_{\tau d}(\pi_E, \pi_A)$, then by Definition 1, *for equal-length and uniformly weighted comparisons, we have*

$$-GW_{\tau d}^2(\pi_E, \pi_A) = -\frac{1}{N}\sum_{i=1}^{N}\left|d_E^{\tau d}(r_E, z_{E,i}) - d_A^{\tau d}(r_A, z_{A,i})\right|^2$$

$$-\frac{1}{N}\sum_{t=1}\rho_{\pi_A}(z_A)\frac{1}{\rho_\pi(z_A)}\left|d_E^{\tau d}(z_{E,t}, z_{E,t+1}) - d_A^{\tau d}(z_{A,t}, z_{A,t+1})\right|^2$$

$$= \frac{1}{N}\sum_{t=1}\rho_{\pi_A}(z_A) r_{GW_{\tau d}}(z_A).$$

Under the exact-matching assumption stated above, Proposition 1 implies that exact matching is feasible within the policy class under consideration, for equal-length and uniformly weighted comparisons, $GW_{\tau d}(\pi_E, \pi_A) = 0$ if and only if $d_A^{\tau d}(r_A, z_{A,i}) = d_E^{\tau d}(r_E, z_{E,i})$ for every i. Taking consecutive differences shows that equal cumulative distances imply equal one-step distances. Conversely, summing equal one-step distances from the two roots recovers equal cumulative distances. Hence, $GW_{\tau d}(\pi_E, \pi_A) = 0$ if and only if $d_E^{\tau d}(z_{E,t}, z_{E,t+1}) = d_A^{\tau d}(z_{A,t}, z_{A,t+1})$ for every transition t.

The negatives of the squared local discrepancies are nonpositive and attain their common maximum of zero exactly under this condition.

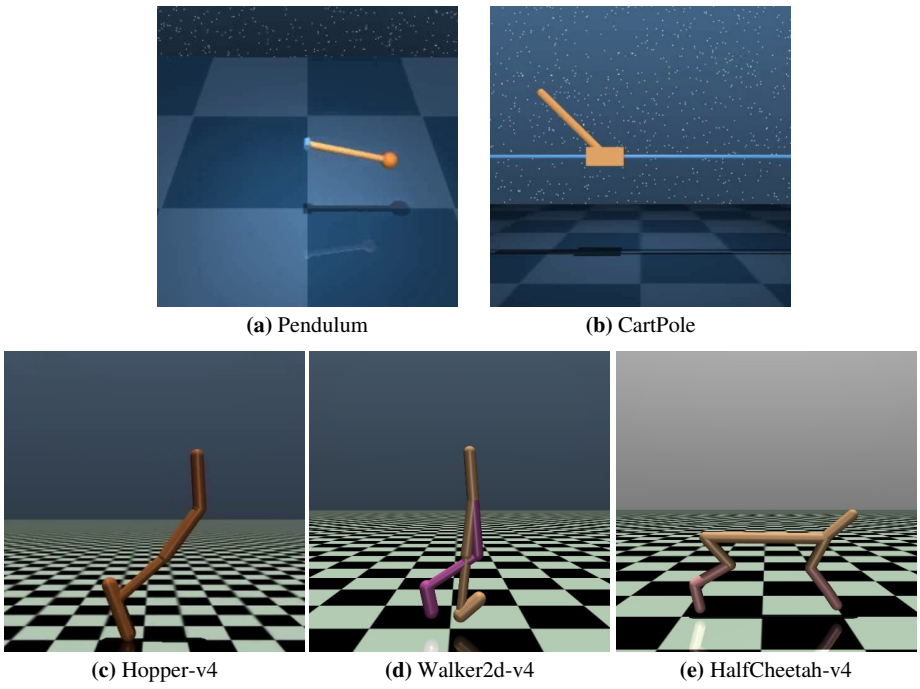
$$\arg\max_{\pi_A}(-GW_{\tau d}^2(\pi_E, \pi_A)) = \arg\max_{\pi_A}\frac{1}{N}\sum_{t=1}\rho_{\pi_A}(z_A)\, r_{GW_{\tau d}}(z_A).$$

Whenever exact matching is feasible, they therefore preserve the global objective's optimal policies. Thus, π_A is an optimal policy for reward $r_{GW_{\tau d}}$.

B Experimental Details

B.1 Environments and Experts Trajectory

We evaluate Pendulum-to-CartPole transfer in the DeepMind Control Suite and six directed transfers among Hopper-v4, Walker2d-v4, and HalfCheetah-v4 in Gymnasium MuJoCo. Figure B.1 shows an illustration of those environments.



(a) Pendulum **(b)** CartPole

(c) Hopper-v4 **(d)** Walker2d-v4 **(e)** HalfCheetah-v4

Figure B.1: Environments used in our experiments. Top, left to right: Pendulum and CartPole from the DeepMind Control Suite. Bottom, left to right: Hopper-v4, Walker2d-v4, and HalfCheetah-v4 from Gymnasium MuJoCo.

Hopper. The agent controls a single-legged robot to hop forward in the x–z plane. Hopper-v4 has an 11-dimensional state describing torso height, joint angles, and velocities, and three joint-torque actions. Its native reward combines forward velocity and a healthy-state bonus, minus a control cost.

HalfCheetah. The agent controls a cheetah-shaped robot to run forward in the x–z plane. HalfCheetah-v4 has a 17-dimensional state of positions and velocities and six joint-torque actions. Its native reward is forward velocity minus a control cost.

Walker2d. The agent controls a bipedal robot to walk forward while remaining upright. Walker2d-v4 has a 17-dimensional state of positions and velocities and six joint-torque actions. Its native reward combines forward velocity and a healthy-state bonus, minus a control cost.

Pendulum. The agent applies a single torque to swing a pendulum upright. Its three-dimensional state contains the cosine and sine of the pole angle and its angular velocity. The native reward is one when the pole is within 8° of upright and zero otherwise.

CartPole. The agent applies a single horizontal force to move a cart and swing its pole upright. Its five-dimensional state contains cart position and velocity, the cosine and sine of the pole angle, and angular velocity. The native reward combines uprightness and cart centering with preferences for small control inputs and low angular velocity.

Following Franzmeyer et al. (2022), we report signed displacement in the target direction for locomotion, measured as the final minus initial horizontal torso position. For swing-up, we report the sum of native rewards. These are evaluation measures; learners are trained using imitation rewards. We train SAC experts with native locomotion rewards, evaluate deterministic rollouts, and select one successful, high-return trajectory of 1,000 steps per source environment. Imitation uses only its states. Table B.1 reports the selected demonstrations; CartPole serves as a learner domain in our experiments.

Table B.1: Selected expert demonstrations used for imitation. Each trajectory contains 1,000 steps. For locomotion, the task score is horizontal displacement; for Pendulum, it is native return.

Expert	Native return	Evaluation score
Hopper	3,556.27	20.46 m
HalfCheetah	12,740.79	654.98 m
Walker2d	5,100.20	32.82 m
Pendulum	882.00	882.00

Table B.2: Main training hyperparameters for GWτdIL.

Hyperparameter	Value
Critic learning rate	0.0001
Actor learning rate	0.0001
Gradient steps	2 gradient steps / 4 environment steps
Batch size	256
Critic Hidden dimensions	[256, 256]
Discount factor γ	0.99
Target interpolation coefficient τ	0.005
Replay capacity	1000000
Optimizer	Adam
Episode horizon	1,000 steps
Q value Aggregation	min
Evaluation interval / episodes	50,000 steps / 10 episodes

B.2 Network Architecture and Hyperparameters

In this section, we report the main hyperparameters used in our experiments. Table B.2 summarizes the training configuration of GWτdIL. GWIL and UDIL follow the policy optimization settings of their original implementations, using SAC and a DAC-style TD3 optimizer, respectively; GWτdIL also uses SAC, matching GWIL. Table B.3 lists the main hyperparameters used for UDIL. Unless otherwise specified, the same settings are used across experiments.

B.3 Baseline

We compare against GWIL and UDIL because both address cross-domain imitation without proxy-task demonstrations and support the state-only setting considered here. GWIL provides a global structural-matching baseline, whereas UDIL learns through a task-relevant expert representation.

For GWIL, we adapt the authors' official implementation to our experimental setting. For UDIL, we implement the published method following the reported architecture, hyperparameters, and task-relevant state selections. Training configurations and implementation details are summarized in Table B.3.

GWIL. Fickinger et al. (2022), Cross-Domain Imitation Learning via Optimal Transport, ICLR 2022. Our local implementation adapts the authors' code: `https://github.com/facebookresearch/gwil`.

UDIL. Franzmeyer et al. (2022), Learn What Matters: Cross-Domain Imitation Learning with Task-Relevant Embeddings, NeurIPS 2022.

C More Ablations

C.1 Omitting Directional Information

In this part, we supplement the directional ablation in Section 5.3. Without the directional term, GWτIL produces repeated body motion but fails to recover reliable forward locomotion. Agents

Table B.3: Main training hyperparameters for UDIL.

Hyperparameter	Value
Discriminator hidden dimensions	[256, 256]
Discriminator learning rate	0.001
Actor learning rate	Initially 0.001
Critic learning rate	0.001
Batch size	100
Actor hidden layers	[400, 300]
Discount factor γ	0.99
Target interpolation coefficient τ	0.005
Replay capacity	1000000
Optimizer	Adam
Episode horizon	1,000 steps
Exploration / target-policy noise	0.1 / 0.2
Q value Aggregation	min
Evaluation interval / episodes	50,000 steps / 10 episodes

move backward or largely remain in place, consistent with Figure 4. These behaviors show why matching transition magnitudes alone is insufficient to recover the intended movement direction.

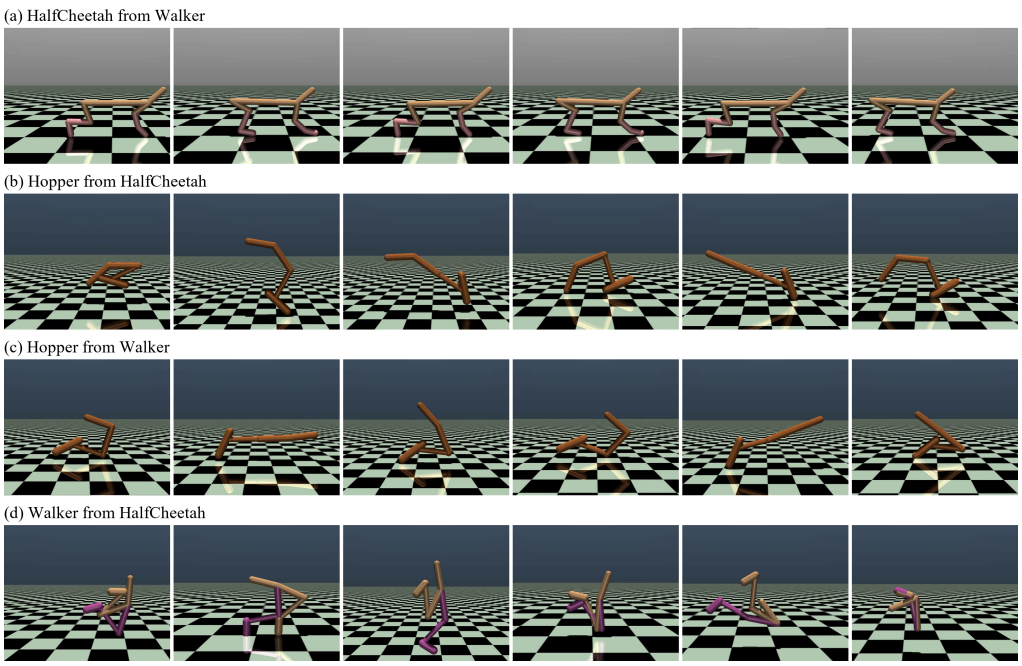


Figure C.1: GWτIL rollouts without directional matching. Each row shows one transfer task, with frames ordered from left to right. All rollouts use seed 0 and equal task–embodiment weights.

C.2 Distinguishing Temporal Order

In this part, we supplement Section 5.3 with the trajectory-generation details. A frozen HalfCheetah expert with velocity feedback generates 1,000-step rollouts at 20 Hz after a 5 s warm-up. Target speeds increase from 2 to 6 m/s or decrease from 6 to 2 m/s; deceleration is generated through the dynamics. We select 30 pairs matched by displacement and speed distribution before scoring against a fixed accelerating reference. Mean displacements are 200.18 m and 200.07 m, respectively.

Table C.1: Discrepancy to a fixed accelerating HalfCheetah reference for accelerating and decelerating test rollouts. Values are mean ± sample standard deviation over 30 rollouts per condition.

Method	Accelerating test	Decelerating test
GW	0.00979962 ± 0.00001176	0.00980474 ± 0.00002499
GWτ	0.00836198 ± 0.00047149	0.02863608 ± 0.00039556
GWτd	0.01318801 ± 0.00071682	0.03342322 ± 0.00045701

We can see that GW assigns nearly equal scores to both conditions, whereas GWτ and GWτd rank acceleration closer in all 30 pairs. This supports preserving temporal progression when comparing motions with the same travel direction.

C.3 Imitation from more demonstrations.

We provide UDIL with 20 expert trajectories to more closely match the setting of Franzmeyer et al. (2022), while GWτdIL continues to use a single demonstration. Figure C.2 shows an improvement for Walker-from-HalfCheetah, but little benefit in the other two tasks. This task-dependent effect is consistent with the original report of UDIL that additional demonstrations need not improve every transfer. GWτdIL retains higher final mean performance in all three tasks, showing that the observed gap persists when UDIL receives more demonstrations.

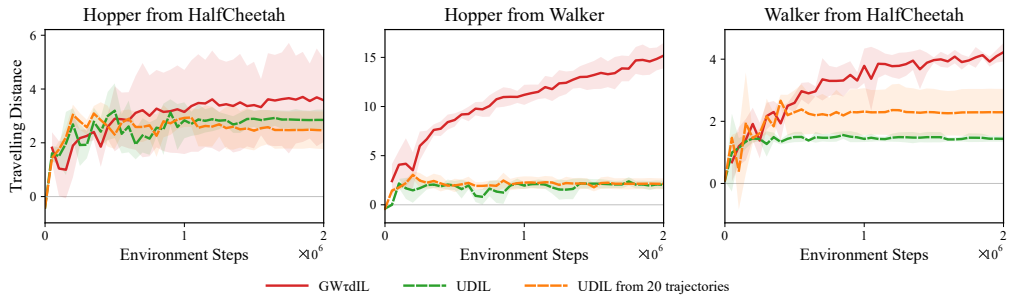


Figure C.2: Effect of the number of expert demonstrations provided to UDIL. GWτdIL uses one demonstration in all cases. Curves report signed displacement in the target direction; lines denote the mean and shaded regions ±1 standard deviation across 5 GWτdIL seeds and 3 UDIL seeds.

C.4 GWIL with task and embodiment groups

We test whether task–embodiment grouping alone improves global GW matching. Grouped GWIL computes a separate trajectory-level coupling for each group and combines their rewards with equal weights. Same as GWτdIL, embodiment features are torso/joint angles and angular velocities, and remaining states are used as task features.

Figure C.3 shows that grouping does not improve GWIL's final mean performance in these tasks, while GWτdIL remains ahead. Under the recorded settings, grouping alone is insufficient. This supports the motivation in Section 4.2 for simplifying the comparison structure and assigning local rewards.

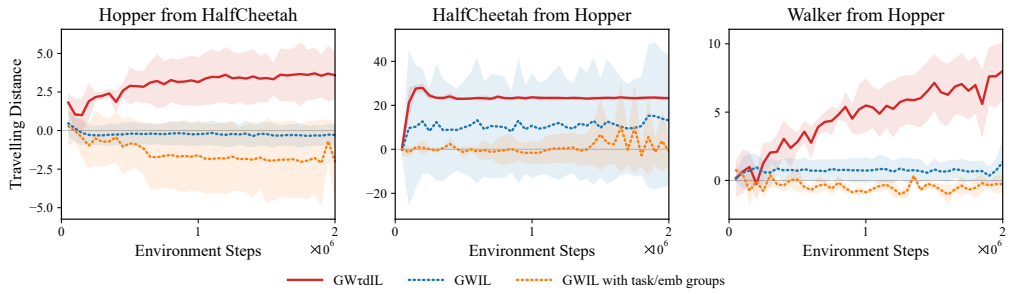


Figure C.3: Effect of task–embodiment grouping on GWIL. We compare full-state GWIL, grouped GWIL, and GWτdIL; the grouped methods use equal task–embodiment weights. Curves report signed displacement in the target direction; lines denote the mean and shaded regions ± 1 standard deviation across 5 random seeds.

D Computational Resources

All experiments were conducted on Ubuntu 20.04 servers equipped with NVIDIA A100 40 GB GPUs, 64 CPU cores, and 512 GB system memory. Independent random seeds and experimental configurations were executed in parallel when resources were available. Specifically, for 2 million environment steps, GWτdIL training takes approximately 3-3.5 hours, while GWIL training requires about 10-12 hours and UDIL takes about 13-15 hours.