EgoBridge: Domain Adaptation for Generalizable Imitation from Egocentric Human Data

Anonymous Authors

Abstract:

Egocentric human experience data presents a vast resource for scaling up end-to-end imitation learning for robotic manipulation. However, significant domain gaps in visual appearance, sensor modalities, and kinematics between human and robot impede knowledge transfer. This paper presents EgoBridge, a unified co-training framework that explicitly aligns the policy latent spaces between human and robot data using domain adaptation. Through a measure of discrepancy on the joint policy latent features and actions based on Optimal Transport (OT), we learn observation representations that not only align between the human and robot domain but also preserve the action-relevant information critical for policy learning. EgoBridge achieves a significant absolute policy success rate improvement by 44% over human-augmented cross-embodiment baselines in three real-world single-arm and bimanual manipulation tasks. EgoBridge also generalizes to new objects, scenes, and tasks seen *only* in human data, where baselines fail entirely. Videos and additional information can be found at https://ego-bridge.github.io/

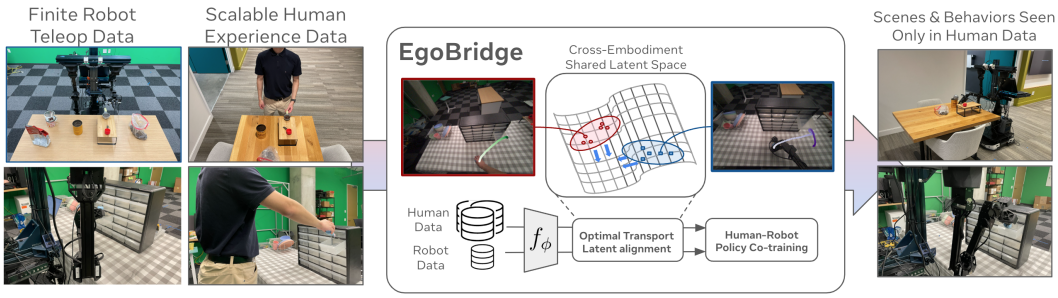

Figure 1: EgoBridge enables rich knowledge transfer from human to robot, based on our key hypothesis: aligned latent representations yield stronger transfer. Our algorithm, which adapts optimal transport, aligns behaviors which are similar across embodiments. This enables EgoBridge to generalize to objects, scenes and even motions demonstrated only in human data.

1 Introduction

Supervised imitation learning methods such as behavior cloning have emerged as a promising path to scaling robot performance across diverse objects, tasks, and environments. However, while large-scale models in vision and language have achieved remarkable generalization through Internet-sourced data, replicating this success in robotics remains challenging due to the labor-intensive nature of collecting teleoperated demonstrations. Deploying physical robots to many new environments to collect data with enough coverage and diversity is economically and practically intractable.

In this work, we aim to enable robots to learn from egocentric recordings of natural human behavior, collected by increasingly ubiquitous wearable devices (e.g., XR devices and smart glasses). Without a robot in the loop, such data is cheap and scalable to collect and captures natural human interactions with the world. More importantly, it reflects the *embodied human experience*, as it contains both observations (e.g., egocentric RGB images) and actions (e.g., hand motions). Unlike unstructured data sources such as Internet videos, the rich embodied information allows us to treat human data and

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

robot data as equal parts in a continuous spectrum of demonstration data and potentially learn from both with a unified learning framework.

However, the multitudes of *domain gaps* between human and robot pose significant challenges in designing such a framework. Human bodies and robots have different visual appearances. Even within a shared action space, kinematic differences can lead to behavior distribution shifts. Robots also have additional sensing modalities such as wrist cameras that are often missing from embodied human data. While recent works such as EgoMimic [1] have attempted to bridge the embodiment gaps with techniques such as visual masking, data normalization, and motion retargeting, such domain gaps still largely remain. More broadly, simply co-training from cross-domain data does not automatically yield effective knowledge transfer, as suggested by recent studies [2]. Such challenges prevent policies from scaling their performance primarily with human data.

We formalize the human-robot cross-embodiment learning problem as a *domain adaptation problem*, where human and robot data represent two labeled distributions with significant *covariate shifts* in observations due to embodiment gaps. Standard domain adaptation approaches often rely on global distribution alignment techniques such as adversarial training [3] and maximum mean discrepancy minimization [4]. However, they primarily address high-level tasks such as image classification and fail to preserve detailed action-relevant information—a critical requirement for robot learning where actions and observations are temporally correlated under compounding covariate shift.

To address these challenges, we propose **EgoBridge**, a novel domain adaptation approach that uses Optimal Transport (OT) to align latent representations from human and robot domains as part of the policy co-training objective. Unlike conventional domain alignment methods, our OT formulation explicitly exploits the inherent relationship between motion similarities in human and robot domains to form *pseudo-pairs* as supervision for the adaptation process. Concretely, we use the dynamic time warping (DTW) distance among human and robot motion trajectories to shape the OT ground cost. This encourages the transport map to find a minimal-cost coupling between human and robot data exhibiting similar behaviors. As such, EgoBridge aligns policy representations across domains via a differentiable OT loss (Sinkhorn distance), while preserving action-relevant information for policy learning. Importantly, we show that EgoBridge learns a shared latent representation that *generalizes* beyond the paired data. This enables the policy to learn behaviors observed *only within the human dataset*, effectively enabling the policy to scale primarily with human data.

We evaluate **EgoBridge** on both a reproducible simulation benchmark task and three challenging real-world manipulation tasks. Our results show that **EgoBridge** consistently improves policy success rates compared to human-augmented cross-embodiment baselines, for up to 44% absolute success rate improvement, and effectively transfers behaviors from diverse human demonstrations to robotic execution in tasks requiring spatial, visual, and task generalization.

2 Related Work

Supervised Imitation Learning. Supervised imitation learning (SIL), especially behavior cloning, learns policies from expert demonstrations and achieves strong results when trained on large datasets. [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Vision-Language-Action (VLA) models which integrate broad vision-language pretraining with action decoders, improve generalization but still require large labeled robot datasets for robust real-world performance [10, 8, 12, 11]. Consequently, our work leverages scalable human demonstrations alongside in-domain robot data to improve learning outcomes.

Learning from Human Data. Human data presents two main opportunities for robot learning: abundant unlabeled online videos and curated, labeled demonstrations [15, 16, 1]. Unlabeled web videos, though plentiful, require pseudo-labeling of actions via inverse dynamics models [17] or point tracking [18, 19, 20] for policy training, forming a basis for some foundation models [12], yet often still necessitating in-domain robot data. Alternatively, labeled human demonstrations can be co-trained with robot data as distinct embodiments [1, 21, 22], enhancing robustness and scene understanding. However, generalizing to novel behaviors observed only in human data remains

challenging. To address this limitation, we propose a novel learning framework for jointly aligning observation-action spaces across human and robot embodiments to improve generalization.

Domain Adaptation and Optimal Transport. Domain Adaptation (DA) aims to reduce reliance on target-specific data by leveraging labeled source domain data to bridge distribution gaps and improve performance on unlabeled target domains. In cross-embodiment learning, DA has been applied for shared dynamics modeling [23], unsupervised reward modeling [24], and high-level planning [25]. However, many DA methods primarily focus on global distribution alignment, which can neglect fine-grained action information crucial for transfer across robot embodiments. To address this, computer vision research introduced Optimal Transport (OT) as a loss function for DA to align both local and global distributions [26, 27, 28]. Building on these insights, we propose an action-aware DA approach using OT to learn shared representations across embodiments, thereby improving observation and behavior generalization.

3 Preliminaries and Problem Statement

3.1 Optimal Transport

Optimal Transport for Domain Adaptation. Optimal Transport (OT) offers a principled framework for comparing probability distributions by considering the geometry of their sample spaces. Given two distributions, μ_S (source) and μ_T (target), over a common metric space $\mathcal{X}$, and a cost function $\mathcal{C}(x^S, x^T)$ measuring the effort to move mass from $x^S \in \mathcal{X}$ to $x^T \in \mathcal{X}$, OT finds a probabilistic coupling $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ that minimizes the expected transport cost:

$$\gamma^* = \arg\min_{\gamma \in \Pi(\mu_S, \mu_T)} \mathbb{E}_{(x^S, x^T) \sim \gamma}[\mathcal{C}(x^S, x^T)],$$

where $\Pi(\mu_S, \mu_T)$ is the set of all joint distributions whose marginals are μ_S and μ_T. For discrete empirical distributions from N_S source samples $\{x_i^S\}$ and N_T target samples $\{x_j^T\}$, the cost matrix is $C_{ij} = \mathcal{C}(x_i^S, x_j^T)$, and the total cost is $\langle \gamma, C \rangle_F = \sum_{i,j} \gamma_{ij} C_{ij}$.

Differentiable Optimal Transport as a Loss Function. When used as a cost function to align representations, the standard OT problem is often regularized. The Sinkhorn algorithm [29] introduces an entropic regularization term to the OT objective, yielding a differentiable approximation T_ϵ^* to the optimal transport plan:

$$T_\epsilon^* = \arg\min_{T \in \Pi(\mu_S, \mu_T)} \mathbb{E}_{(x^S, x^T) \sim T}[\mathcal{C}(x^S, x^T)] - \epsilon H(T),$$

where $\epsilon > 0$ is the regularization strength and $H(T)$ is the entropy of the coupling. This regularization makes the problem strictly convex and efficiently solvable. The resulting regularized optimal transport cost, $\sum_{i,j} (T_\epsilon^*)_{ij} C_{ij}$, is differentiable with respect to the cost matrix C. This allows OT to serve as a loss function within deep learning frameworks, enabling the learning of feature encoders that map inputs to a space $\mathcal{X}$ where their distributions are aligned by minimizing this transport cost.

3.2 Human and Robot Data Sources

We conside egocentric human data ($\mathcal{D}_H$) and teleoperated robot data ($\mathcal{D}_R$). $\mathcal{D}_H = \{(o_t^H, a_t^H)\}_{t=1}^{N_H}$ consists of N_H egocentric human demonstrations, where $o_t^H \in \mathcal{O}^H$ are observations from wearable sensors (e.g., head-mounted cameras) and $a_t^H \in \mathcal{A}$ are human actions in a common action space (e.g., robot end-effector and human hand poses). This data is abundant and captures natural, diverse behaviors. Conversely, $\mathcal{D}_R = \{(o_t^R, a_t^R)\}_{t=1}^{N_R}$ comprises N_R robot experiences, typically from teleoperation, with $o_t^R \in \mathcal{O}^R$ being robot sensor observations (e.g., ego-centric/wrist cameras, joint states) and $a_t^R \in \mathcal{A}$ the robot actions. This data is often scarce. We describe how each data source is captured and processed in more detail in Sec. 4.3. We assume actions are in trajectory chunks, which is shown to improve prediction temporal consistency of the trained policies [5, 6].

3.3 Cross-Embodiment Imitation Learning: Challenges and Objectives

Our primary goal is to effectively learn from both limited robot demonstrations (D_R) and more abundant, diverse egocentric human demonstrations (D_H). We train a feature encoder $f_\phi : \mathcal{O}^H \cup$

$\mathcal{O}^R \to \mathcal{Z}$ to project observations from both human ($\mathcal{O}^H$) and robot ($\mathcal{O}^R$) into the shared latent space $\mathcal{Z}$. We jointly train a policy π_θ that maps these learned latent representations $z \in \mathcal{Z}$ to actions $a \in \mathcal{A}$.

Cross-Embodiment Co-Training. A popular approach [1, 21] involves training the policy end-to-end using a standard Behavior Cloning (BC) loss on the aggregated dataset:

$$\mathcal{L}_{\text{BC-cotrain}}(\phi, \theta) = \mathbb{E}_{(o,a)\sim D_H \cup D_R}[\mathcal{L}_{\text{BC}}(\pi_\theta(f_\phi(o)), a)],$$

To effectively learn from both data sources, a critical assumption is that a shared latent space $\mathcal{Z}$ would naturally emerge where the mapping from latent states to actions is domain-invariant, resulting in $P_R(a|f_\phi(o_R)) \approx P_H(a|f_\phi(o_H))$ for observations o_R and o_H from aligned underlying states.

Observation Covariate Shift. However, we argue that, without explicit mitigation, the induced marginal distributions over these latents, $\mu_H = P(f_\phi(\mathcal{O}^H))$ and $\mu_R = P(f_\phi(\mathcal{O}^R))$, will exhibit a significant *covariate shift* ($\mu_H \neq \mu_R$). This shift arises from inherent domain gaps in observations (e.g., differing visual appearances, viewpoints, sensor modalities like robot wrist cameras absent in human setups) and embodiment kinematics. We also empirically show that such co-trained representations often form disjoint latent clusters (Sec. 5.3). This covariate shift in the marginal latent distributions undermines the foundational assumption of consistent conditional action distributions across domains, thereby limiting effective knowledge transfer from human to robot.

Generalizable Cross-Embodiment Transfer. This motivates our method that aims at *joint domain adaptation* (Sec. 4.1), where we explicitly seek to align the latent representations from human and robot data while preserving action-relevant information. Successfully addressing this latent misalignment should enable two crucial levels of generalization: First, for tasks present in both $\mathcal{D}_H$ and $\mathcal{D}_R$, the system must achieve **observation generalization**. This involves effectively bridging visual and sensor gaps, which include appearance changes that do not affect behaviour. Second, and more ambitiously, the system should enable **behavior generalization** *(Beh. Gen.)* allowing the robot to perform tasks or handle novel situations (e.g. task variations) observed *only* in $\mathcal{D}_H$. This requires the learned encoder f_ϕ to generalize beyond scenarios with paired human and robot data which require motion information to be transferred, such as spatial variations in goal pose.

4 EgoBridge

EgoBridge is a co-training framework designed to effectively imitate embodied human demonstrations and robot demonstrations. It explicitly addresses the domain gap between human and robot experiences through an Optimal Transport (OT)-based domain adaptation mechanism integrated into the policy learning process. The core of EgoBridge lies in aligning the *joint distributions of latent policy features and corresponding actions* across the human and robot domains. The following sections detail this joint domain adaptation formulation (Sec. 4.1), the design of its OT cost function (Sec. 4.2), and the overall training process and system details (Sec. 4.3).

4.1 Joint Domain Adaptation via Optimal Transport

To address the latent covariate shift (Sec. 3.3) and generalizable cross-embodiment transfer, EgoBridge builds on Optimal Transport (OT, Sec. 3.1) to directly shape the shared feature encoder f_ϕ. Unlike standard domain adaptation techniques [30] that often aligns only the marginals $P(f_\phi(O))$, which can discard action-relevant information, EgoBridge optimizes f_ϕ to align the joint distributions of its output latent features and their corresponding actions, i.e., $P(f_\phi(O), A)$.

Given mini-batches of human data $\{(o_i^H, a_i^H)\}_{i=1}^{N_H}$ and robot data $\{(o_j^R, a_j^R)\}_{j=1}^{N_R}$, where a represents a temporally-extended action trajectory, we define an OT-based loss to guide the learning of f_ϕ. The differentiable Sinkhorn OT formulation [29] allows us to compute a loss based on the alignment of the empirical distributions of $(f_\phi(o^H), a^H)$ and $(f_\phi(o^R), a^R)$:

$$\mathcal{L}_{\text{OT-joint}}(\phi) = \sum_{i,j} (T_\epsilon^*)_{ij} \cdot \mathcal{C}\left((f_\phi(o_i^H), a_i^H), (f_\phi(o_j^R), a_j^R)\right).$$

Here, $(T_\epsilon^*)_{ij}$ is the optimal transport plan coupling the i-th human and robot (latent, action) pairs. The cost function $\mathcal{C}(\cdot, \cdot)$ measures the dissimilarity between these joint entities. Its design is crucial for capturing meaningful behavioral similarities across domains, which we detail in Sec. 4.2.

Minimizing $\mathcal{L}_{\text{OT-joint}}(\phi)$ directly influences the parameters ϕ of the encoder f_ϕ. The gradients from this loss encourage f_ϕ to produce latent features $f_\phi(o_i^H)$ and $f_\phi(o_j^R)$ that minimize the transport cost required to align them, especially when their associated actions a_i^H and a_j^R are behaviorally similar (as determined by $\mathcal{C}$). At each step a transport plan is computed which influences the feature encoder to couple the action pairs This iterative process shapes the latent space $\mathcal{Z}$ to be domain-invariant with respect to the joint observation-action manifold.

4.2 Designing OT Cost Function for Action-Aware Joint Adaptation

Our joint OT formulation (Section 4.1) relies on a cost function $\mathcal{C}((z^H, a^H), (z^R, a^R))$ to measure the dissimilarity between joint human and robot latent feature-action pairs. A critical challenge is designing this cost to be robust to inherent domain differences. Specifically, we aim to account for *temporal misalignments*, where human and robot often execute the same task at of different speed, e.g., humans might be 2-3 times faster than teleoperated robots, and *kinematic variations*, where even within a shared SE(3) end-effector action space and hand-eye alignment through an egocentric coordinate frame (Sec. 4.3), minor kinematic differences exist.

Dynamic Time Warping. To identify behaviorally similar action sequences while accounting for these differences, we propose to leverage Dynamic Time Warping (DTW) to guide the OT alignment. DTW [31] has been effective in prior work to compare time series data and trajectories. Formally, given two action sequences $\mathbf{a}^H = (a_1^H, \ldots, a_T^H)$ from human data and $\mathbf{a}^R = (a_1^R, \ldots, a_T^R)$ from robot data of identical length T, DTW finds an alignment path $\pi \subseteq \{1, \ldots, T\} \times \{1, \ldots, T\}$ that minimizes the cumulative distance:

$$\text{DTW}(\mathbf{a}^H, \mathbf{a}^R) = \min_{\pi \in \mathcal{A}(T)} \sum_{(i,j)\in\pi} \|a_i^H - a_j^R\|^2$$

where $\mathcal{A}(T)$ is the set of admissible monotonic alignments constrained to start at $(1, 1)$ and end at (T, T), while allowing small local shifts to account for temporal variations.

Soft Supervision. With DTW, we can identify highly correlated samples from both domains. However, directly utilizing the DTW cost is noisy and instead is a much stronger measure of relative pairing between human and robot samples. As such the DTW cost can be used to *pseudo-pair* D_H and D_R. On a mini-batch of size B of sampled ground-truth state-action pairs from D_S and D_T, we form a DTW cost matrix $A \in R^{B\times B}$. Here, $A_{i,j} = \text{DTW}(a_i^S, a_j^T)$. The row-wise minimum cost gives us the most behaviorally similar human pseudo-pair for each robot sample : $i^*(j) = \arg\min_i A_{ij}$.

Given the standard OT Euclidean distance cost $D_{ij} = \|f_\phi(o_i^H) - f_\phi(o_j^R)\|^2$, we define the joint cost $\tilde{\mathcal{C}}((f_\phi(o_i^H), a_i^H), (f_\phi(o_j^R), a_j^R))$ for L_{OT} as:

$$\tilde{C_{ij}} = \begin{cases} D_{ij} \cdot \lambda & \text{if } i = i^*(j) \\ D_{ij} & \text{otherwise} \end{cases}$$

where $0 < \lambda \ll 1$ is a small scalar. This cost function strongly incentivizes OT to match robot samples with their behaviorally closest human pseudo-pair (identified by DTW) by significantly reducing the cost for these pairs. For all other pairs, the cost is simply the distance in the latent feature space. This soft supervision from DTW guides the latent space alignment towards behaviorally relevant correspondences across embodiments.

4.3 Putting it all together: EgoBridge

With all the ingredients for joint distribution adaptation using OT with joint policy co-training, we present EgoBridge as a unified cross-embodiment imitation learning algorithm and describe its corresponding robot learning system and policy architecture.

Policy Co-Training with Joint Adaptation. EgoBridge jointly optimizes the feature encoder f_ϕ and π_θ, with the joint OT loss applied on the feature encoder and the BC co-training loss applied end-to-end through both components: $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{BC-cotrain}}(\phi, \theta) + \alpha\mathcal{L}_{\text{OT-joint}}$, with tunable weight α. Detail of the algorithm and hyperparameter choices are described in Appendix.

Egocentric Human Data. We largely follow EgoMimic [1] and leverage a wearable smart glass platform Meta Project Aria [32] as our main data collection platform. The platform allows us to collect *exteroceptive*, egocentric first person POV RGB images (I^H_{ego}), and *proprioceptive* data (q^H), cartesian pose for both arms $\in SE(3) \times SE(3)$. We take inspiration from EgoMimic to construct stable reference frames to form action sequences of cartesian pose (a^H) in the egocentric camera frame.

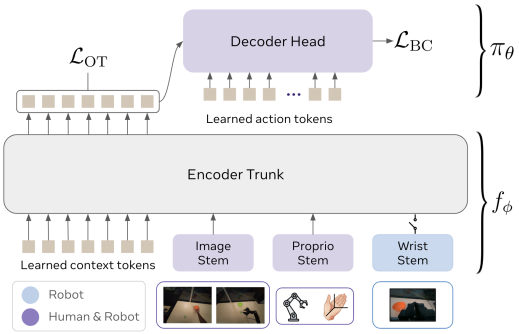

Figure 2: EgoBridge policy co-training with joint adaptation. The encoder f_ϕ consists of modality-specific input stems and the encoder trunk, while the policy π_θ consists of a shared multi-block transformer decoder. $\mathcal{L}_{\textbf{OT-joint}}$ optimizes the encoder while $\mathcal{L}_{\textbf{BC-cotrain}}$ optimizes the entire network.

Teleoperated Robot Data. We base our robot platform on the open-source Eve robot [1]. In particular, we leverage Aria glasses as the main egocentric perception sensor for the robot and mount it in a way that emulates the hand-eye configuration of a human adult (I^R_{ego}). This effectively mitigates the human-robot camera device gap, allowing us to specifically study the appearance, kinematic and behaviour gaps. The robot additionally provides RGB streams from its two RealSense D405 wrist cameras, I^R_{wrist}. The actions consist of a sequence of corresponding future end-effector poses, $a^R \in SE(3) \times SE(3)$.

Shared Policy Architecture. Inspired by recent cross-embodiment policy learning [33] and DETR-style architectures [34], our policy employs a shared transformer encoder "trunk" (f_ϕ) and a shared transformer policy decoder "head" (π_θ) (Fig. 2). We perform embodiment-specific gaussian normalization to the proprioception and actions. The encoder f_ϕ begins with *stems*—shallow networks that tokenize raw observations; notably, a *shared* vision stem processes main egocentric RGB images (I_{ego}) from both human and robot to enforce visual alignment, while separate stems handle robot wrist camera inputs (I_{wrist}). The subsequent multi-layer encoder trunk processes these concatenated tokens, along with M prepended learnable context tokens upon which the OT loss is applied. The multi-block decoder head then generates actions by attending to this encoded context, utilizing T learnable action tokens and injecting context through alternating self and cross-attention blocks.

5 Experiments

In this section, we aim to validate three core hypotheses. **H1**: EgoBridge improves co-training performance for scenarios present in both human and robot data. **H2**: EgoBridge enables generalization to scenarios only seen in human data. **H3**: EgoBridge learns a shared latent space where human and robot data are aligned in task-relevant manners. We validate the hypotheses through a standard simulation benchmark task (Sec. 5.1) and three complex real-world manipulation tasks (Sec. 5.2).

5.1 Simulation Evaluation

To facilitate reproducible study and eliminate the confounding factors in real robot systems, we study a well-explored planar pushing task [6], where the goal is to push a T-shaped object to a desired goal location. We emulate "human" (source) data through a blue circle pusher and "robot" (target) data through a salmon triangle pusher (Fig. 3), with lower floor friction. The differences aim to analogize the appearance and agent-environment dynamics gaps between human and robot data.

Source and Target Domain Data. In our "robot" target domain, we collect demos in the standard push-T setting, but in our "human" source domain, we alter the background color to purple and change the T configuration to be mirrored, requiring a new motion to slot into place 3. The change in background color is analogous to the human data containing new visual scenery, and the change in starting configuration is analogous to the human demonstrating new motions in their demonstrations.

Training and Evaluation. To eliminate factor from model design (Sec. 4.3), we choose a standard ResNet-UNet Diffusion Policy [6] and apply the OT-joint loss on the feature outputs of the ResNet

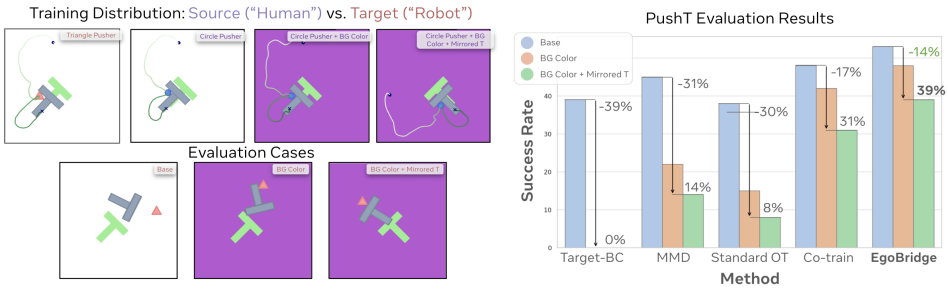

Figure 3: In the simulated Push-T experiments, we probe a toy version of visual and motion level generalization from human to robot. We have narrow target "robot" data represented by the triangle pusher on a white background, and diverse source "human" data represented by the circle pusher with changes in background color and T configuration. We test our "robot" on the diverse human scenarios, and find that EgoBridge outperforms traditional Domain Adaptation baselines.

Table 1: Real World Evaluation Results: In-Distribution and Generalization

Method	**Scoop Coffee (SR)**			**Drawer (SR)**			**Laundry**
	In-Dist.	Obj. Gen.	Scene+Obj Gen.	Total (Pts | SR)	Place Toy	Beh. Gen.	(Pts | SR)
Robot-only BC	33%	40%	7%	38 | 9%	28%	0%	38 | 28%
Co-train	53%	46%	0%	55 | 22%	42%	0%	41 | 33%
EgoMimic [1]	60%	53%	0%	49 | 14%	39%	0%	38 | 33%
MimicPlay [25]	33%	27%	0%	33 | 14%	22%	0%	32 | 28%
ATM [19]	47%	33%	0%	56 | 6%	17%	8%	35 | 28%
EgoBridge	**67%**	**60%**	**27%**	**77 | 47%**	**72%**	**33%**	**48 | 72%**

encoder. We perform standard action normalization and co-train the policy on both the triangle and circle pusher data. We evaluate the policy on 3 cases: Triangle in the standard setting, Triangle with purple background, Triangle with purple background and Triangle with purple background and flipped T. We evaluate a total of 100 fixed seeds across all the models and report the mean reward (max IoU with goal) and the success rate (reward ≥ 0.9).

Baselines. In the more controlled simulation settings, we choose to compare EgoBridge against conventional domain adaptation baselines. We choose **Maximum Mean Discrepancy (MMD)** [30] as an alternative domain adaptation loss to joint-OT on the feature encoder. We also test **Standard OT**, which performs marginal alignment instead of joint alignment. The **Co-train** baseline trains on evenly-sampled data from both domains without an alignment loss. Finally, **Target-only** is a control study which trains the policy only on the target (triangle) data.

5.2 Real World Evaluation

We evaluate EgoBridge on three challenging real-world manipulation tasks, as illustrated in Fig. 4.

Drawer: The robot interacts with a 6x4 drawer array, tasked to pick a toy, place it into a pre-opened drawer, and close it. Robot data (144 demonstrations) covers three of the four array quadrants, each quadrant being a 3x2 arrangement. Human data (1 hour) covers all four quadrants, providing demonstrations for motions into the fourth, robot-unseen quadrant. This setup specifically tests *behavior generalization* to drawer locations only seen in human data. Points (**Pts**) are awarded for successful completion of each stage and a trial is considered a success only if the robot completes all actions. Evaluation uses 48 trials (2 rollouts for each of the 24 drawers).

Scoop Coffee: The robot uses its left arm to scoop coffee beans with a spoon and empty them into a target. Robot data (50 demonstrations) involves a specific target (can) in one scene. Human data (2 hours) includes demonstrations with both the can and a new target (grinder), across two distinct scenes, one of which is novel to the robot. Target object positions are randomized (30x23 cm area). We evaluated *observation generalization* for: (1) the new grinder target, and (2) the new scene with the new target, that is, scooping to the grinder in the new scene, seen in human data only. Performance is measured by success rate over 15 rollouts across 5 distinct target locations.

Laundry: This is a bimanual task where the robot needs to fold the shirt in 50 x 22 cm range with a rotation range $\pm$ 30 degrees. The robot uses both arms to fold the right sleeve, the left sleeve, and

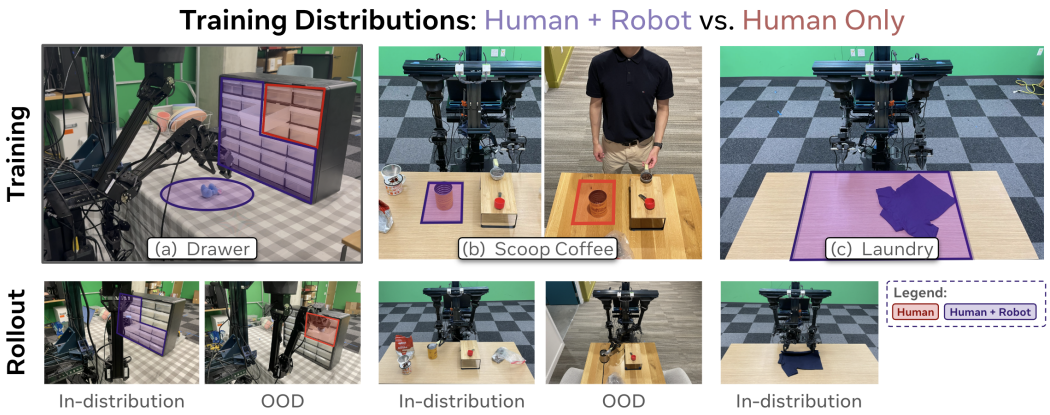

Figure 4: **Training Data and Evaluation Settings.** We show the distribution of human and robot training data (top) and evaluation setting, where in-distribution scenarios are in both human and robot data, while out-of-distribution (OOD) scenarios is seen only in human data.

then the final stage to fold the shirt in half. We award **Pts** for each successful stage and consider it success if all the individual stages are successful. We collect 2 hours of robot data which include 300 demonstrations across 3 shirts, and 2 hours of human data comprising approximately 700 demonstrations. We conduct 18 evaluations with diverse shirt initial placement and colors.

Baselines. We adopt the following baselines for real-world tasks. *Co-train:* Direct co-training of the robot and human data using BC loss, without any latent alignment. *EgoMimic: [1]* Co-training with explicit vision and action-space alignment using masking, shared end-effector pose head (human and robot) and a separate joint-space head for robot. *Mimicplay: [25]* A hierarchial policy with a latent high-level planner co-trained on human and robot data, and an action decoder fine-tuned on robot data. *Any-Point Trajectory Modeling (ATM): [19]* A hierarchical policy where the high-level planner is initially co-trained on 2D point tracks derived from both robot and human video data. These point tracksare obtained via Co-tracker [35]. Following this, high-level planner is frozen and an action decoder is fine-tuned specifically on robot data. *Target-only BC:* Trained only on robot data.

5.3 Results

EgoBridge improves in-domain task performance (H1). EgoBridge improves in-domain performance (H1), achieving 7–44% higher success rates than both human-augmented and robot-only baselines. While co-training, EgoMimic, and ATM also yield gains, EgoBridge consistently outperforms them, likely due to better-aligned latent representations that enhance cross-embodiment transfer.

EgoBridge enables generalization to objects and scenes only seen in human data (H2). While it is difficult to collect robot data across diverse scenes and objects, it is trivial to do for human data, so it's critical that we can transfer this knowledge from human to robot, inspiring our experimental setup. Specifically, in the *Scoop Coffee* task, our human data introduces a new coffee grinder, table, lighting and height variations, completely unseen to the robot. We find EgoBridge outperforms all baselines when tested on the new coffee grinder (7-33%). Further, most methods fail entirely when tested on the new grinder + scene, but EgoBridge retains a performance of 27%. We observe similar robustness trends in our simulated benchmark Fig. 3, where EgoBridge enables generalization in the push-T task to a new background and starting configuration, outperforming all baselines.

EgoBridge enables generalization to new behaviors only seen in human data (H2). In our most challenging setting, we seek to show that we can learn *entirely new* motions from human data alone. In the drawer task, the robot data covers 3/4 drawer quadrants, whereas the human data covers all 4 quadrants. We evaluate our policy's performance on these new drawers, and find that EgoBridge is able to generalize to these locations with a success rate of 33%, whereas most methods fail entirely (Tab. 1). While all the methods were exposed to the same human data, only EgoBridge was able to effectively transfer the human motion to a novel robot action. We attribute this success to the

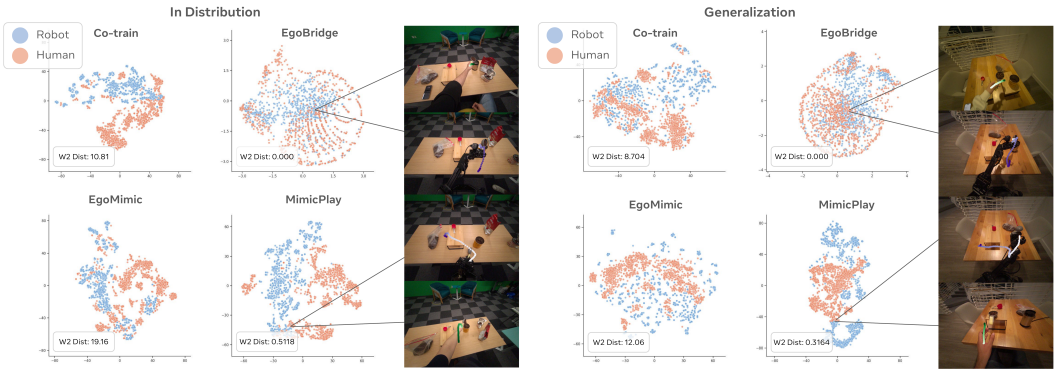

Figure 5: Visualization of TSNE plots on encoded features for EgoBridge and baselines, with the mean Wasserstein-2 distance and KNN pairs of aligned human-robot data visualized.

well aligned latent representations, which enables human to robot knowledge interpolation. We also observed a similar trend in the simulated benchmark where EgoBridge faces the lowest performance drop of 14% compared to all baselines in the *mirrored-T + background colour* reflected in Fig. 3.

EgoBridge learns a shared latent space that aligns human and robot data in a task relevant manner (H3). We hypothesize that an ideal latent space for transfer would jointly embed human and robot data into a space with high overlap and semantic interoperability. To probe this, we create a TSNE visualization of the action tokens from our transformer backbone. EgoBridge not only exhibits the highest latent overlap between human and robot as measured by Wasserstein distance as seen in Fig. 5, but also upon inspecting K-nearest neighbor pairs in latent space, exhibits the most semantically similar neighbors. For instance, we see the human and robot performing the same phase of a given task, whereas in baselines like MimicPlay that aligns marginals with KL-div, the semantic similarity is lacking. This result is highly correlated with the task success rates for in-distribution and generalization where baselines with poor alignment perform lower consistently across all evaluations.

Ablation. We ablate three key components of our method 1) replacing our DTW-based pairing metric to instead use simple MSE, 2) replacing the joint OT objective $\mathcal{L}_{\text{OT-joint}}$ with standard marginal alignment, and 3) removing any auxiliary alignment objectives (direct co-training). We find that replacing the cost function with MSE leads to the largest performance drop for in-distribution policy success rate from 47% to 17%, seen in Tab. 2, which emphasizes the importance of creating semantically similar pseudo-pairs. Ablating Joint-OT also shows a large performance drop in both in-distribution and generalization cases which emphasizes how naive marginal alignment cannot transfer knowledge from human data effectively. Ablating an auxiliary alignment loss also shows a significant performance drop for in-distribution success rate and leads to failure in generalization cases, emphasizing the need for joint distribution alignment.

Table 2: Ablation Results (Drawer)

Method	Drawer (SR)	Beh. Gen. (SR)
EgoBridge	**47%**	**33%**
MSE	14%	17%
Standard-OT	33%	17%
Co-train	22%	0%

6 Conclusion

We presented EgoBridge, a novel co-training framework designed to enable robots to learn effectively from egocentric human data by explicitly addressing domain gaps. By leveraging Optimal Transport on joint policy latent feature-action distributions, guided by Dynamic Time Warping cost on action trajectories, EgoBridge successfully aligns human and robot representations while preserving critical action-relevant information. Our experiments demonstrated significant improvements in real-world task success rates (up to 44% absolute gain) and, importantly, showed robust generalization to novel objects, scenes, and even tasks observed only in human demonstrations, where baselines often failed. Future work includes extending to multi-task settings, exploring alignment costs from language or foundation model features, and scaling to Internet-sourced human data without action labels.

References

[1] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221.

[2] A. Wei, A. Agarwal, B. Chen, R. Bosworth, N. Pfaff, and R. Tedrake. Empirical analysis of sim-and-real cotraining of diffusion policies for planar pushing from pixels. *arXiv preprint arXiv:2503.22634*, 2025.

[3] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation, 2017. URL https://arxiv.org/abs/1702.05464.

[4] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks, 2017. URL https://arxiv.org/abs/1605.06636.

[5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[6] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[9] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[10] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control, 2024. *URL https://arxiv. org/abs/2410.24164*, 2024.

[11] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025.

[12] NVIDIA, :, J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

[13] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[14] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[15] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild, 2022. URL https://arxiv.org/abs/2207.09450.

[16] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans, 2024. URL https://arxiv.org/abs/2406.10454.

[17] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, L. Liden, K. Lee, J. Gao, L. Zettlemoyer, D. Fox, and M. Seo. Latent action pretraining from videos, 2024. URL https://arxiv.org/abs/2410.11758.

[18] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

[19] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning, 2023.

[20] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning, 2025. URL https://arxiv.org/abs/2501.06994.

[21] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, G. Yang, J. Zhang, S. Yi, G. Shi, and X. Wang. Humanoid policy ˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[22] M. Lepert, J. Fang, and J. Bohg. Phantom: Training robots without robots using only human videos, 2025. URL https://arxiv.org/abs/2503.00779.

[23] S. J. Wang and A. M. Johnson. Domain adaptation using system invariant dynamics models. In A. Jadbabaie, J. Lygeros, G. J. Pappas, P. A.Parrilo, B. Recht, C. J. Tomlin, and M. N. Zeilinger, editors, *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 1130–1141. PMLR, 07 – 08 June 2021. URL https://proceedings.mlr.press/v144/wang21c.html.

[24] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos, 2020. URL https://arxiv.org/abs/1912.04443.

[25] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[26] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation, 2017. URL https://arxiv.org/abs/1705.08848.

[27] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation, 2016. URL https://arxiv.org/abs/1507.00504.

[28] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation, 2018. URL https://arxiv.org/abs/1803.10081.

[29] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. URL https://arxiv.org/abs/1306.0895.

[30] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola. A kernel method for the two-sample problem, 2008. URL https://arxiv.org/abs/0805.2368.

[31] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1): 43–49, 1978. URL https://doi.org/10.1109/TASSP.1978.1163055.

[32] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, C. Peng, C. Sweeney, C. Wilson, D. Barnes, D. DeTone, D. Caruso, D. Valleroy, D. Ginjupalli, D. Frost, E. Miller, E. Mueggler, E. Oleinik, F. Zhang, G. Somasundaram, G. Solaira, H. Lanaras, H. Howard-Jenkins, H. Tang, H. J. Kim, J. Rivera, J. Luo, J. Dong, J. Straub, K. Bailey, K. Eckenhoff, L. Ma, L. Pesqueira, M. Schwesinger, M. Monge, N. Yang, N. Charron, N. Raina, O. Parkhi, P. Borschowa, P. Moulon, P. Gupta, R. Mur-Artal, R. Pennington, S. Kulkarni, S. Miglani, S. Gondi, S. Solanki, S. Diener, S. Cheng, S. Green, S. Saarinen, S. Patra, T. Mourikis, T. Whelan, T. Singh, V. Balntas, V. Baiyya, W. Dreewes, X. Pan, Y. Lou, Y. Zhao, Y. Mansour, Y. Zou, Z. Lv, Z. Wang, M. Yan, C. Ren, R. D. Nardi, and R. Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

[33] L. Wang, X. Chen, J. Zhao, and K. He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers, 2024. URL https://arxiv.org/abs/2409.20537.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers, 2020. URL https://arxiv.org/abs/2005.12872.

[35] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker: It is better to track together, 2024. URL https://arxiv.org/abs/2307.07635.