Freeze the Policy, Infer the Goal: Cross-Domain Imitation with World Models

Xingyuan Zhang *
Technical University of Munich
Volkswagen AG
xingyuan.zhang@tum.de

Marvin Alles
Technical University of Munich
marvin.alles@tum.de

Patrick van der Smagt
Eötvös Loránd University Budapest
smagt@argmax.org

Philip Becker-Ehmck
Volkswagen AG
philip.becker-ehmck@volkswagen.de

Abstract

The paradigm of pretraining foundation models and subsequently finetuning them for downstream tasks has emerged as the prevailing approach in the decision-making community. Although imitation learning offers a powerful mechanism to adapt pretrained models using expert demonstrations, agents frequently struggle when data is collected under fundamentally different morphologies or viewpoints. This challenge, known as cross-domain imitation learning, typically requires complex algorithmic design or kinematic retargeting. In this paper, we propose Goal Inference Imitation Learning (GIIL), a framework that reframes cross-domain imitation as an efficient adaptation problem. First, we pretrain a goal-conditioned multi-task policy entirely on unlabelled, offline embodiment data using world models and self-supervised distance-based rewards. During the imitation phase, rather than updating the policy network, we freeze the policy and adapt the agent by inferring a continuous goal vector that best matches the expert demonstration. This inference is guided by an Optimal Transport reward evaluated within the world model's imagination. By decoupling the acquisition of embodiment-specific motor skills from the high-level intentions of the demonstrator, GIIL enables zero-shot transfer and lightweight finetuning without catastrophic forgetting. Empirical results on the DeepMind Control Suite demonstrate that GIIL outperforms baseline approaches across cross-embodiment and cross-viewpoint settings.

1 Introduction

The paradigm of pretraining large-scale foundation models and subsequently finetuning them for downstream tasks has revolutionized machine learning and is increasingly becoming the standard in decision-making [Bommasani et al., 2021]. In this context, Imitation Learning (IL) provides a natural mechanism to adapt these pretrained agents to new tasks using offline expert demonstrations. However, a critical bottleneck emerges when these demonstrations are collected under fundamentally different morphologies or viewpoints than those of the deploying agent. Replicating the ability to learn from such heterogeneous sources—known as *cross-domain imitation learning* (cross-domain IL)—is essential for scaling decision-making systems to make use of vast, uncurated offline datasets, such as training robotic agents directly from human videos.

*Corresponding author.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

Standard IL approaches often fail under cross-domain conditions due to severe domain shift. Because there is no direct mapping between the state and action spaces of different physical embodiments, gradient-based adaptation is inherently ill-posed. To overcome this, we take inspiration from research in neuroscience. The mirror neuron hypothesis [Rizzolatti and Craighero, 2004] suggests that humans understand and imitate the actions of others by internally simulating observed behaviors using their own motor systems, responding to high-level intentions rather than merely copying low-level kinematic details [Gallese et al., 1996]. We hypothesize that artificial agents can similarly bridge the cross-embodiment gap by strictly separating the acquisition of embodiment-specific motor skills from the inference of high-level task goals.

To implement this idea, we propose *Goal Inference Imitation Learning* (GIIL), a framework that reframes cross-domain transfer as an efficient offline adaptation problem. GIIL operates in two distinct phases. First, we use an unlabelled offline embodiment dataset to pretrain a world model and a goal-conditioned multi-task policy in a purely self-supervised manner. Second, when adapting to a demonstration from a different embodiment, we freeze this pretrained policy. Instead of updating the network weights, we cast imitation as an inference problem over a continuous goal space. By optimizing a goal vector using an Optimal Transport reward within the world model's imagination, GIIL performs lightweight adaptation.

We evaluate GIIL in simulated cross-embodiment and cross-viewpoint settings on the DeepMind Control Suite (DMC), transferring tasks between a Walker and a morphologically different Stickman robot. Our experiments show that goal inference not only preserves the pretrained motor skills but also outperforms baseline approaches — validating the core hypothesis that decoupling motor skills from task intent enables effective cross-domain transfer.

2 Problem Formulation

For cross-domain IL, there are two domains, which can be represented by two POMDPs – the *source* domain $\mathcal{M}^S = \{S^S, A^S, T^S, R^S, O^S, \Omega^S\}$ in which the expert demonstrator operates, and the *agent* domain $\mathcal{M}^A = \{S^A, A^A, T^A, R^A, O^A, \Omega^A\}$ in which the agent operates. To make imitation feasible, we assume that the task is shared across domains, i. e., $R^S \equiv R^A$. Here, the reward function is an abstract, domain-agnostic function that can be defined on the states of both domains, e. g., moving forward as fast as possible. We also assume that the observation space $O^S \equiv O^A$ is the same, e. g., both domains have visual observations, while the emission functions Ω^S and Ω^A are different, since they may render the presence of different robots. If only the emission functions are different but the underlying embodiment is the same, i. e., $\{S^S, A^S, T^S\} \equiv \{S^A, A^A, T^A\}$, the problem is normally referred to as third-person IL [Stadie et al., 2017] or cross-viewpoint IL as we will use in this paper. If the embodiment is also different, the problem becomes more challenging as there is no one-to-one mapping between the states and actions of the two domains; this is referred to as cross-embodiment IL [Fickinger et al., 2021, Zakka et al., 2022].

For simplicity, we will focus on the offline cases where the agent only has access to the *embodiment dataset* D_{body} containing trajectories $\{o_0^A, a_0^A, o_1^A, a_1^A \dots\}$ in the agent domain $\mathcal{M}^A$ that represent past experiences of interacting with the environment and the *demonstration dataset* D_{demo} containing a few expert trajectories $\{o_0^S, o_1^S, o_2^S \dots\}$ from the source domain $\mathcal{M}^S$ solving a certain task defined by R_{demo}. The goal of our agent is to learn a policy π from D_{demo} which can solve the task defined by R_{demo} as well as the expert π_{demo} who generated D_{demo}.

We restrict the agent to data from its own embodiment, as this reflects the realistic setting where cross-domain data is unavailable during training.

3 Related Work

Despite its importance, cross-domain IL has not been widely studied in the literature. This is likely due to the challenges brought by domain differences and the ill-posed nature of the problem. The majority of existing work focus on learning domain-invariant representations to bridge the gap between different domains. Stadie et al. [2017] pioneered this direction using the Adversarial Imitation Learning (AIL) framework with a domain discriminator and a confusion loss to enforce invariance. Following this, Cetin and Celiktutan [2020] proposed handling domain shifts in visual IL by enforcing cycle consistency across domains to extract domain-independent behavioural features.

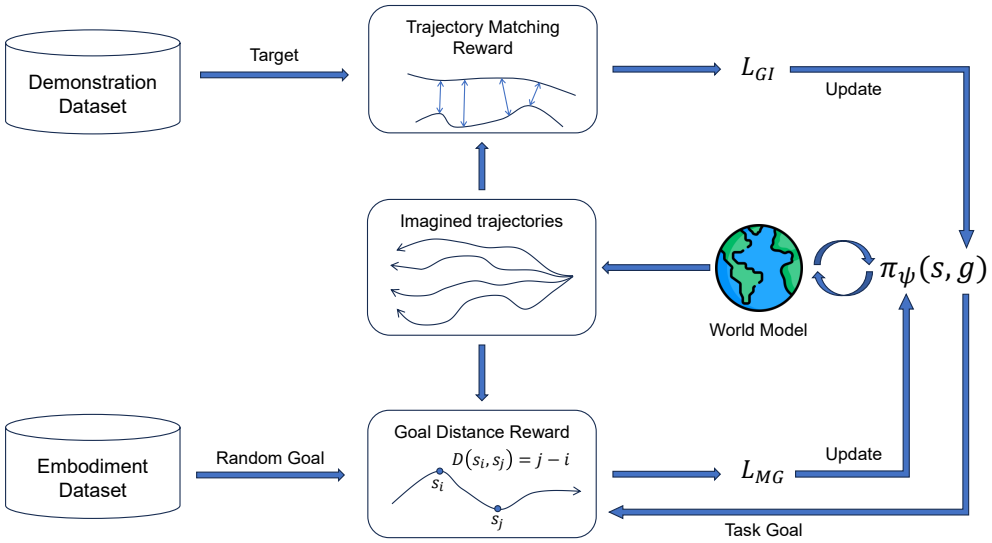

Figure 1: Overview of GIIL framework and its model-based implementation. The framework describes cross-domain imitation learning as a bi-level optimization problem. The parameter of the goal-conditioned multi-task policy ψ is trained on the random goals from the embodiment dataset and the current task goal with a distance-based goal-reaching reward formulated as L_{MG}. Simultaneously, the goal g for the task is inferred by minimising a trajectory-matching-based reward with the target trajectories from the demonstration dataset formulated as L_{GI}. Both of these objectives are optimized with the imagined trajectories from the world model. In this way, the cross-domain imitation learning problem is converted to a goal inference problem.

To further improve robustness, they introduce mutual information constraints to regularize the latent representation, ensuring that it only captures task-relevant information.

Moving beyond standard adversarial methods, recent work has explored alternative alignment metrics. Fickinger et al. [2021] use the Gromov-Wasserstein [Mémoli, 2011] distance to align state-action spaces between expert and agent without requiring paired data or proxy tasks. Similarly, Raychaudhuri et al. [2021] tackled the challenge of imitation from observations by learning state and action mappings to align unpaired demonstrations. More recently, Liu et al. [2023] proposed a generalized contextual imitation learning framework (CEIL) that employs a bi-level optimization objective with hindsight embeddings, enabling effective cross-domain transfer even with mismatched experts.

Besides general-purpose representation learning, domain-specific methods explicitly synthesize robot demonstrations from human data. Traditional approaches typically rely on inverse kinematics to retarget human motion to robot embodiments [Rakita et al., 2017]. Building on these foundations, MimicGen [Mandlekar et al., 2023] scales this approach by automatically generating extensive robot datasets from limited human examples using object-centric retargeting. Alternatively, to bridge the visual domain gap, methods such as AVID [Smith et al., 2020] leverage image-to-image translation models (e. g., CycleGAN [Zhu et al., 2017]) to transform human video demonstrations directly into synthesized robot observations, enabling the agent to learn from visually aligned data.

4 Methodology

In this section, we propose Goal Inference Imitation Learning (GIIL), a framework designed to tackle cross-domain IL and an instantiation of it using pretrained world models [Ha and Schmidhuber, 2018]. The fundamental challenge in cross-domain transfer is identifying exactly *what* information can be reliably transferred across severe morphological or visual gaps.

We argue that the essential transferable information is the *high-level intention (or goal)* of the demonstrator, rather than their low-level kinematic trajectory. Therefore, GIIL separates the extraction

of high-level intentions from low-level kinematic execution. To achieve this, the framework consists of two main phases: (1) learning a goal-conditioned multi-task policy in the agent's embodiment, and (2) inferring a goal vector from the demonstration trajectories to perform IL.

To be specific, in the first phase, we train a multi-task policy $\pi_\psi(a_t|s_t, g)$, where ψ represents the policy parameters, s_t is the current state, and g is a latent goal vector defining the current task. Once an optimal policy with parameters ψ^* is learned, the policy is frozen. It then serves as a dynamical constraint ensuring that any generated behavior adheres to physically viable motor skills natively supported by the agent's embodiment.

In the second phase, policy learning for a new demonstration is converted from updating network weights ψ to inferring the optimal goal vector g^*. Given a demonstration dataset D_{demo} from a different embodiment—where no direct state-action mapping exists—we find a goal vector g^* that minimizes the behavioral discrepancy between the agent and the demonstrator. We formulate this as a bi-level optimization problem:

$$g^* = \arg\min_g L_{\text{GI}}(\psi^*, g, D_{\text{demo}}) \; s.t. \; \psi^* = \arg\min_\psi L_{\text{MG}}(\psi, D_{\text{body}}), \tag{1}$$

where L_{GI} is the goal inference loss measuring the discrepancy agent's behavior under goal g against the demonstrations, and L_{MG} is the multi-task policy learning loss optimized over the agent's embodiment dataset D_{body}.

An overview of the GIIL framework and its model-based implementation is illustrated in Figure 1. The methodology is structured as follows: First, we briefly introduce the world model's components and training procedure. Second, we demonstrate how a pretrained video model is used to align the goal space within the world model's latent space. Finally, we instantiate L_{MG} using Reinforcement Learning (RL) within the world model's imagined rollouts, guided by a self-supervised distance reward, while Section 4 is defined L_{GI} through RL within the same imagined environment using an OT-based matching reward.

The World Model We model dynamics of the agent's embodiment using a variational latent world model [Karl et al., 2017, Hafner et al., 2019a,b, Becker-Ehmck et al., 2019, Klushyn et al., 2021]. Specifically, we adopt the Recurrent State Space Model (RSSM) [Hafner et al., 2019a] from the Dreamer series [Hafner et al., 2019b, 2020, 2025]. This architecture assumes observations o_t are governed by a Markovian hidden state s_t, using an encoder $z_t = f_\phi(o_t)$, a posterior $q_\phi(s_t \mid s_{t-1}, a_{t-1}, z_t)$, a prior $p_\theta(s_t \mid s_{t-1}, a_{t-1})$, and a decoder $p_\theta(o_t \mid s_t)$.

The generative parameters θ and the inference parameters ϕ are optimized jointly by maximizing the Evidence Lower Bound (ELBO) over trajectories τ sampled from the embodiment dataset $\mathcal{D}_{\text{body}}$:

$$\max_{\phi,\theta} \mathbb{E}_{\tau_{1:T}^{(o,a)} \sim \mathcal{D}_{\text{body}}, s_{1:T} \sim q_\phi} \left[\sum_{t=1}^{T} \Big(\underbrace{\log p_\theta(o_t \mid s_t)}_{\text{reconstruction loss}} - \underbrace{D_{\text{KL}}[q_\phi \parallel p_\theta]}_{\text{KL loss}} \Big) \right]. \tag{2}$$

To efficiently compute this objective, we apply the reparameterization trick [Kingma and Welling, 2024, Rezende et al., 2014], which allows us to autoregressively sample the inferred states from the observation and action sequences.

The Generalizable Goal Space Since the embodiment dataset contains no task labels, we define a generalizable goal space capable of representing diverse tasks. The goal must encode a sequence of observations, i. e., $g = f(o_{1:T})$, yet training directly on the agent's observations $o^A_{1:T}$ is unlikely to generalize to a different domain $o^S_{1:T}$. We therefore seek a representation that is transferable across embodiments.

Inspired by Mazzaglia et al. [2024], we use the pretrained InternVideo2 [Wang et al., 2024] video model, which maps a short sequence of frames to a unit feature vector $p_t = f(o_{t-k+1:t})$. Trained on large-scale video-language data with a contrastive objective, InternVideo2 captures high-level semantic content while remaining largely domain-agnostic. To further remove residual domain-specific information, we learn a goal encoder $q_{\phi^g}(s_t|p_t)$ that maps p_t into the world model's latent space, paired with a goal decoder $p_{\theta^g}(p_t|s_t)$. This side variational autoencoder [Kingma and Welling, 2024, Rezende et al., 2014] is trained by minimizing the negative ELBO (see Section A for

implementation details):

$$L_{\text{goal}}(\phi^g, \theta^g) = \mathbb{E}_{\tau_{1:T}^{(o,a,p)} \sim D_{\text{body}}, s_{1:T} \sim q_\phi} \left[D_{\text{KL}}[q_{\phi^g} \| \text{sg}(q_\phi)] - \mathbb{E}_{s_t^g \sim q_{\phi^g}} [\log p_{\theta^g}(p_t | s_t^g)] \right], \quad (3)$$

where the $\text{sg}(\cdot)$ stands for stop gradient operator.

Learning the Goal-conditioned Policy Since the embodiment dataset D_{body} lacks task labels and consistent behaviours across trajectories, supervised IL is infeasible. Instead, inspired by self-supervised skill discovery [Mendonca et al., 2021], we treat the multi-task policy $\pi_\psi(a_t|s_t, g)$ as a goal-reaching policy operating in the world model's latent space, trained to reach arbitrary latent states as goals without any external supervision. We measure progress toward a goal via a temporal distance $D^{\text{dis}}(s_i, s_j) = \frac{1}{T}(j - i)$ for states s_i, s_j with $j > i$ from the same trajectory [Ma et al., 2022], and learn a distance network $d_{\theta^{\text{dis}}}(s_i, s_j)$ to predict it (see Section B for details). We then define the goal-reaching reward as $r^{\text{goal}}(s_t, g) = -d_{\theta^{\text{dis}}}(s_t, g)$ and train the goal-conditioned policy following the actor-critic approach of Dreamer [Hafner et al., 2019b] on imagined rollouts:

$$L_{\text{MG}}(\psi, D_{\text{body}}) = -\mathbb{E}_{\hat{s}_{t:t+H} \sim D_{\text{body}} \circ q_\phi \circ p_\theta \circ \pi_\psi} \left[\sum_{k=1}^{H-1} V_{\eta^g}(\hat{s}_{t+k}, g) \right], \quad (4)$$

where $V_{\eta^g}(s_t, g)$ is a goal-conditioned value function trained with r^{goal} and TD(λ) targets [Hafner et al., 2019b].

Inferring the Goal from Demonstrations Given the pretrained goal-conditioned multi-task policy $\pi_{\psi^*}(a_t|s_t, g)$ and the goal encoder $q_{\phi^g}(s_t|p_t)$, we can infer the goal vector from demonstration trajectories zero-shot. Specifically, InternVideo2 encodes the demonstration sequence $o^S_{T-k+1:T}$ into a feature representation p_T, which the goal encoder then maps to a goal vector g that conditions the policy to execute the demonstrated task.

However, since both models are trained on the agent's embodiment dataset, the inferred goal vector may be suboptimal for demonstrations from a different embodiment. We therefore finetune the goal vector on the demonstration dataset using an Optimal Transport (OT)-based reward $r^{\text{ot}}(s_t)$ [Rupf et al., Papagiannis and Li, 2023, Haldar et al., 2022], which is more robust to domain differences than AIL-style alternatives (see Section C for details). The goal vector g is finetuned via the actor-critic loss:

$$L_{\text{GI}}(\psi^*, g, D_{\text{demo}}) = -\mathbb{E}_{\hat{s}_{t:t+H} \sim D_{\text{demo}} \circ q_\phi \circ p_\theta \circ \pi_{\psi^*}} \left[\sum_{t+1}^{t+H} V_{\eta^{\text{ot}}}(s_t) \right], \quad (5)$$

where $V_{\eta^{\text{ot}}}(s_t)$ is trained with $r^{\text{ot}}(s_t)$ and TD(λ) targets [Hafner et al., 2019b].

Since the initial goal vector may fall outside the support of the embodiment dataset, the pretrained policy π_{ψ^*} may no longer be optimal, violating the bi-level optimization assumption in Equation (1). We therefore jointly finetune the multi-task policy and the goal vector, using stop-gradients to ensure ψ is updated via L_{MG} and g via L_{GI} only.

5 Experiment

To validate the effectiveness of our method, we design experiments on the following environments. As cross-embodiment IL is expected to work when there exists some similarity between the agent and the demonstrator, our experiments mainly focus on two embodiments in the simulator. One is Walker from DMC [Tunyasuvunakool et al., 2020] which is a standard benchmark in continuous control, and the other is a simplified humanoid robot called Stickman from Mazzaglia et al. [2024]. Stickman has a similar morphology to Walker, but it has arms and a head, making it more human-like. Ideally, we would also like to test our method on the humanoid robot from DMC, but due to the complexity of the humanoid robot and the amount of time and computation required to train a good multi-task policy, we leave this for future works.

We will consider 4 tasks for both Walker and Stickman: stand, walk, run, jump. In order to generate the demonstration dataset, we will first train expert policies for each task in both embodiments using Dreamer [Hafner et al., 2019b]. Then we will collect demonstration trajectories from the trained

expert policies, with 10 trajectories for each task. In order to study the cross-domain IL in terms of changing of viewpoints, we also render the visual observations from different camera angles. The default camera in DMC is a side camera facing the 2D plane where the robot is moving. We add an additional diagonal camera that is placed behind the robot at a 45-degree angle. For the embodiment dataset, we directly use the dataset from Mazzaglia et al. [2024] as it provides a fair comparison with their method on their pretrained models. Each embodiment dataset contains around 2000 trajectories of replay buffer from Plan2Explore [Sekar et al., 2020] agent and 3×1000 trajectories from the replay buffer of a Dreamer agent trained on stand, walk, and run tasks. More information about the tasks and the viewpoints can be found in Section D.

In the experiments, we will mention different settings for the demonstration datasets. **SE** and **CE** stand for *Same Embodiment* and *Cross Embodiment*, respectively. **SV** and **CV** stand for *Same Viewpoint* and *Cross Viewpoint*, respectively.

We include four algorithms in our experiments for comparison:

- **GenRL** [Mazzaglia et al., 2024]: A finetuning method to derive policy from pretrain world models with multimodal prompts. It also uses the InternVideo2 model to extract features from the demonstration videos. But instead of training a goal-conditioned policy like our method, it maps the extracted feature to a short sequence of belief states and using a cosine distance reward to train the policy.
- **AIME-NoB** [Zhang et al., 2025]: A state-of-the-art model-based IL method. As in the cross-domain setting, the kinematic-based IL is not feasible, we turn-off the AIME loss in AIME-NoB. Note that AIME-NoB is the only method that needs further environment interactions in our experiments.
- **GIIL-ZS** (*GIIL-Zero-Shot*): Our proposed method using the zero-shot inferred goal.
- **GIIL-FT** (*GIIL-Finetune*): Our proposed method using the finetuned goal.

Besides AIME-NoB, all other methods use the last embedding of the first demonstration trajectory as the initial goal vector or prompt. Besides GIIL-ZS, all other methods are run with 3 random seeds and report the average performance. As GIIL-ZS does not involve any training, we run it with the last embedding of all 10 demonstration trajectories and report the average performance.

The results on Walker and Stickman tasks are shown in Table 1 and Table 2, respectively. The reported results are normalized against the random and expert performance of the task on the agent's embodiment. More detailed results on each task can be found in Section E.

As we can see from the results, AIME-NoB performs the best on the SE-SV setting, which is what the algorithm designed for. On the other hand, although it does not explicitly consider cross-domain changes, it is surprisingly not affected too much by the domain gap, as it still performs reasonably well on the three cross-domain settings. This is likely due to the similarity of the two embodiments, as both Walker and Stickman have very similar morphology and kinematics. It can also be attributed to task selection; for example, for the stand task, although we change the embodiment and viewpoint, the task is still very similar as both robots need to maintain an upright pose at the centre of the image. Moreover, we also find in the Stickman experiments that AIME-NoB learns to reduce degrees of freedom by always keeping the arms aligned with the body, which further reduces the domain gap between the two embodiments.

On the other hand, all three methods that use the pretrained InternVideo2 model are generally less affected by the domain gap, as evidenced by the difference between average performance with and without the SE-SV setting. Besides GIIL-ZS, both GIIL-FT and GenRL outperform AIME-NoB on average for the cross-domain settings, which shows the effectiveness of using the video representation model to extract domain-invariant features for cross-domain IL. With the multi-task policy pretrained, GIIL-ZS can already achieve decent performance without any finetuning. GIIL-FT improves further over GIIL-ZS by 23.13 on average for the cross-domain settings. This shows that finetuning the goal vector can help improve the imitation performance. GenRL baseline also performs reasonably well, and it slightly outperforms GIIL-FT on the Walker cross-domain tasks by 4.53 while being significantly outperformed by GIIL-FT on the Stickman cross-domain tasks by 20.68. Therefore, the overall performance of GIIL-FT is higher than GenRL by 8.08 on average for the cross-domain settings.

Table 1: Normalized score of each algorithms on Walker tasks with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	86.05 ± 4.05	84.80 ± 8.19	42.52 ± 3.93	68.10 ± 8.80
SE-CV	65.98 ± 7.48	51.46 ± 10.76	32.97 ± 3.51	64.13 ± 8.21
CE-SV	52.12 ± 8.73	64.91 ± 7.47	19.47 ± 2.81	60.76 ± 7.83
CE-CV	65.91 ± 7.49	28.92 ± 12.33	18.09 ± 2.43	45.53 ± 8.55
Average	67.52 ± 3.89	57.52 ± 5.63	28.27 ± 1.79	59.63 ± 4.23
Average w/o SE-SV	61.34 ± 4.58	48.43 ± 6.34	23.51 ± 1.80	56.81 ± 4.80

Table 2: Normalized score of each algorithms on Stickman tasks with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	55.30 ± 7.92	83.90 ± 7.26	68.27 ± 4.07	68.56 ± 8.67
SE-CV	63.58 ± 8.38	45.35 ± 9.77	61.26 ± 3.99	72.66 ± 7.77
CE-SV	47.37 ± 9.86	70.64 ± 7.90	54.89 ± 3.26	61.77 ± 8.57
CE-CV	20.44 ± 5.52	49.99 ± 9.87	38.37 ± 2.88	58.99 ± 8.92
Average	46.67 ± 4.56	62.47 ± 4.82	55.70 ± 1.98	65.49 ± 4.18
Average w/o SE-SV	43.79 ± 5.46	55.33 ± 5.49	51.51 ± 2.14	64.47 ± 4.83

However, the representation from InternVideo2 model is not optimal, and sometimes it fails to extract the right features that represent the task being performed in the video. In general, we observe that it is able to distinguish clear movement like run and stand, but it sometimes confuses similar vague movements like walk and jump. As we see from the visualisation of the learned policy, much of GIIL's poor performance comes from treating walk and jump as different standing poses without recognising the movement.

6 Discussion and Future Work

In this paper, we introduced *Goal Inference Imitation Learning* (GIIL), a framework that reframes cross-domain IL as an inference problem over a continuous goal space. By using a goal-conditioned multi-task policy as a behavioral constraint, our framework ensures that generated actions adhere to physically viable, learned motor skills, while simultaneously inferring the high-level semantic goals from offline demonstration trajectories. Our experiments, conducted across morphologically distinct simulated robots and varying viewpoints, validate the efficacy of framing cross-embodiment transfer as an inference and adaptation problem.

Despite these promising results, several limitations present exciting avenues for future research. Most notably, while this work focuses on adapting agents offline from demonstrations, the GIIL framework naturally extends to the paradigm of online adaptation. During online rollouts, the agent could iteratively refine its inferred goal vector, similar to the online framework in recent works like AIME-NoB [Zhang et al., 2025]. Furthermore, environment interaction would allow the agent to acquire entirely new motor skills that fall outside the distribution of the pretrained policy that could further improve adaptability [Mendonca et al., 2021]. We did not pursue this continuous refinement in the current study due to the computational overhead of running the heavy InternVideo2 model to process streaming online trajectories.

A further limitation is our reliance on an external, frozen video representation model to bridge the visual domain gap. Since off-the-shelf features are optimized for broad video-language alignment rather than fine-grained continuous control, they may not not capture the task-relevant structure needed for precise imitation. Ideally, future work would integrate native domain-invariant representations directly into the world model's objective, eliminating this dependency entirely.

Beyond this, an interesting direction is to view the continuous goal vector inferred by GIIL as a special case of latent action representations [Bruce et al., 2024, Alles et al., 2024, 2025], where the

latent variable encodes high-level intent rather than low-level motor commands — unifying these perspectives could enable world models that operate across the full spectrum from raw actions to high-level goals.

Finally, while we successfully demonstrated a multi-task policy setup, the diversity of the learned motor skills was inherently constrained by the relatively simple morphologies of the Walker and Stickman embodiments. Scaling this framework to more complex and high-dimensional embodiments—such as humanoids or dexterous manipulators—alongside a vastly broader distribution of offline tasks will be critical for developing truly generalizable cross-embodiment agents [Collaboration et al., 2025].

Acknowledgments and Disclosure of Funding

The authors thank Pietro Mazzaglia for valuable discussions regarding the transferability of InternVideo2 features, and for sharing the implementation details of the GenRL algorithm.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, August 2021. URL `http://arxiv.org/abs/2108.07258`.

Giacomo Rizzolatti and Laila Craighero. THE MIRROR-NEURON SYSTEM. 27:169–192, 2004. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.27.070203.144230. URL `https://www.annualreviews.org/content/journals/10.1146/annurev.neuro.27.070203.144230`.

Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. 119(2):593–609, 1996. ISSN 0006-8950. doi: 10.1093/brain/119.2.593. URL `https://doi.org/10.1093/brain/119.2.593`.

Bradly C. Stadie, Pieter Abbeel, and Ilya Sutskever. Third Person Imitation Learning. In *International Conference on Learning Representations*, February 2017. URL `https://openreview.net/forum?id=B16dGcqlx`.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-Domain Imitation Learning via Optimal Transport. In *International Conference on Learning Representations*, October 2021. URL `https://openreview.net/forum?id=xP3cPq2hQC`.

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*, pages 537–546. PMLR, 2022. URL `https://proceedings.mlr.press/v164/zakka22a.html`.

Edoardo Cetin and Oya Celiktutan. Domain-Robust Visual Imitation Learning with Mutual Information Constraints. In *International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=QubpWYfdNry.

Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. 11 (4):417–487, 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9093-5. URL https://doi.org/10.1007/s10208-011-9093-5.

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*, pages 8902–8912. PMLR, 2021. URL https://proceedings.mlr.press/v139/raychaudhuri21a.html.

Jinxin Liu, Li He, Yachen Kang, Zifeng Zhuang, Donglin Wang, and Huazhe Xu. CEIL: Generalized Contextual Imitation Learning. *Advances in Neural Information Processing Systems*, 36:75491–75516, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ee90fb9511b263f2ff971be9b374f9ee-Abstract-Conference.html.

Daniel Rakita, Bilge Mutlu, and Michael Gleicher. A Motion Retargeting Method for Effective Mimicry-Based Teleoperation of Robot Arms. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 361–370, 2017. URL https://ieeexplore.ieee.org/document/8534763.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations. In *Proceedings of The 7th Conference on Robot Learning*, pages 1820–1864. PMLR, 2023. URL https://proceedings.mlr.press/v229/mandlekar23a.html.

Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. AVID: Learning Multi-Stage Tasks via Pixel-Level Translation of Human Videos. In *Robotics: Science and Systems XVI*. Robotics: Science and Systems Foundation, 2020. ISBN 978-0-9923747-6-1. doi: 10.15607/RSS.2020.XVI.024. URL http://www.roboticsproceedings.org/rss16/p024.pdf.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251. IEEE, 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.244. URL http://ieeexplore.ieee.org/document/8237506/.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyTqHL5xg.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2555–2565. PMLR, May 2019a. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR 2020*, September 2019b. URL https://openreview.net/forum?id=S1lOTC4tDS.

Philip Becker-Ehmck, Jan Peters, and Patrick van der Smagt. Switching Linear Dynamics for Variational Bayes Filtering. In *Proceedings of the 36th International Conference on Machine Learning*, pages 553–562. PMLR, May 2019. URL https://proceedings.mlr.press/v97/becker-ehmck19a.html.

Alexej Klushyn, Richard Kurle, Maximilian Soelch, Botond Cseke, and Patrick van der Smagt. Latent Matters: Learning Deep State-Space Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 10234–10245. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/54b2b21af94108d83c2a909d5b0a6a50-Abstract.html.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*, September 2020. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR) 2014*, 2024. URL http://arxiv.org/abs/1312.6114.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286. PMLR, June 2014. URL https://proceedings.mlr.press/v32/rezende14.html.

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. GenRL: Multimodal-foundation world models for generalization in embodied agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=za9Jx8yqUA&referrer=%5Bthe%20profile%20of%20Sai%20Rajeswar%5D(%2Fprofile%3Fid%3D~Sai_Rajeswar2).

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding, August 2024. URL http://arxiv.org/abs/2403.15377.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and Achieving Goals via World Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 24379–24391. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cc4af25fa9d2d5c953496579b75f6f6c-Abstract.html.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=YJ7o2wetJ2.

Thomas Rupf, Marco Bagatella, Nico Gürtler, Jonas Frey, and Georg Martius. Zero-Shot Offline Imitation Learning via Optimal Transport. In *Forty-Second International Conference on Machine Learning*. URL https://openreview.net/forum?id=9hiq7LaV4G.

Georgios Papagiannis and Yunpeng Li. Imitation Learning with Sinkhorn Distances. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 116–131, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-26412-2. doi: 10.1007/978-3-031-26412-2_8.

Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and Match: Supercharging Imitation with Regularized Optimal Transport. In *6th Annual Conference on Robot Learning*, August 2022. URL https://openreview.net/forum?id=ZUtgUA0Fuwd.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. Dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, November 2020. ISSN 2665-9638. doi: 10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8583–8592. PMLR, November 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Xingyuan Zhang, Philip Becker-Ehmck, Patrick van der Smagt, and Maximilian Karl. Overcoming Knowledge Barriers: Online Imitation Learning from Visual Observation with Pretrained World Models. *Transactions on Machine Learning Research*, April 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=BaRD2Nfj41.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments, February 2024. URL http://arxiv.org/abs/2402.15391.

Marvin Alles, Philip Becker-Ehmck, Patrick van der Smagt, and Maximilian Karl. Constrained latent action policies for model-based offline reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 70381–70405. Curran Associates, Inc., 2024. doi: 10.52202/079017-2249. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/82389fbff376d1e8aec510916d50d054-Paper-Conference.pdf.

Marvin Alles, Xingyuan Zhang, Patrick van der Smagt, and Philip Becker-Ehmck. Latent action world models for control with unlabeled trajectories, 2025. URL https://arxiv.org/abs/2512.10016.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose

Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025. URL https://arxiv.org/abs/2310.08864.

Yucen Wang, Rui Yu, Shenghua Wan, Le Gan, and De-Chuan Zhan. FOUNDER: Grounding Foundation Models in World Models for Open-Ended Embodied Decision Making, July 2025. URL http://arxiv.org/abs/2507.12496.

Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative Adversarial Imitation from Observation, June 2019. URL http://arxiv.org/abs/1807.06158.

A Goal VAE Implementation Details

The side VAE is trained using the world model posterior q_ϕ as the prior, with the stop-gradient operator $\mathrm{sg}(\cdot)$ applied to prevent gradients from flowing into the world model parameters, ensuring the goal encoder and decoder do not interfere with world model training.

Since the RSSM latent state s_t comprises a deterministic component h_t and a stochastic component, only the latter is suitable for the Kullback-Leibler (KL) divergence term in Equation (3). Following Wang et al. [2025], we therefore replace the KL term for h_t with the negative log-likelihood $-\log q_{\phi^g}(h_t|p_t)$.

B Distance Predictor Details

To provide a meaningful training signal for goal reaching, we learn a distance network $d_{\theta^{\text{dis}}}(s_i, s_j)$ that predicts the temporal distance between two latent states. The temporal distance between states s_i, s_j with $j > i$ from the same trajectory is defined as

$$D^{\text{dis}}(s_i, s_j) = \frac{1}{T}(j - i), \tag{6}$$

where T is the maximum trajectory horizon. This provides a simple yet effective proxy for progress: states that are further apart in time are considered further apart in goal space, a notion previously used in control-centric representations [Ma et al., 2022]. The distance network is trained via

$$L_{\text{dis}}(\theta^d) = \mathbb{E}_{(o^{(i)}_{1:T}, a^{(i)}_{0:T-1})^N_{i=1} \sim D_{\text{body}}, (s^{(i)}_{1:T})^N_{i=1} \sim q_\phi} \Big[-\mathbb{E}_{i \neq j} || d_{\theta^{\text{dis}}}(s^{(i)}_t, s^{(j)}_{t'}) - 1||^2_2 + \mathbb{E}_{j>i} || d_{\theta^{\text{dis}}}(s^{(n)}_i, s^{(n)}_j) - D^{\text{dis}}(s^{(n)}_i, s^{(n)}_j) ||^2_2 \Big], \tag{7}$$

where the regression term fits the predicted distance to D^{dis}, and the contrastive term penalizes small predicted distances between states from *different* trajectories, encouraging the network to be trajectory-aware rather than purely state-dependent.

C OT Reward Details

Since the demonstration dataset may differ significantly in embodiment or viewpoint, AIL-style reward functions [Ho and Ermon, 2016, Torabi et al., 2019] are ill-suited for the cross-domain setting, as the discriminator can simply overfit to domain differences rather than task similarity. Instead, we define the reward $r^{\text{ot}}(s_t)$ using OT [Rupf et al., Papagiannis and Li, 2023, Haldar et al., 2022] with the cosine distance between InternVideo2 features as the cost function. As OT can measure distributional similarity even without overlap between distributions, it is more robust in the cross-domain setting [Fickinger et al., 2021]. We use the InternVideo2 feature space directly for the OT cost rather than the world model's latent space, as the pretrained features encode semantic information on the unit sphere and are thus better suited for cosine distance, whereas the world model's latent space has no such geometric constraints.

D Environment and Task Information

In Table 3 we summarise the task information which we used to normalized the returns to compute the scores. To be specific, the score is computed as

$$\text{Score} = \frac{R - R_{\text{min}}}{R_{\text{max}} - R_{\text{min}}}. \tag{8}$$

We also provide example views from the Walker and Stickman environments in Figure 2.

E Detailed results on each task

In this section, we provide the detailed results on each task for the experiments in Section 5.

Table 3: Task information for cross-domain imitation learning experiments.

Task	Min Return	Max Return
stand	150	970
walk	45	960
run	30	800
jump	120	650

Figure 2: Example views from the Walker and Stickman environments. From left to right, it is side view of Walker, diagonal view of Walker, side view of Stickman, and diagonal view of Stickman.

Table 4: Normalized score of each algorithms on Walker-stand task with different settings. All results are reported as mean $\pm$ standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	101.89 ± 0.12	101.88 ± 0.09	71.01 ± 5.75	100.57 ± 0.64
SE-CV	92.19 ± 0.65	101.30 ± 0.11	63.79 ± 4.14	97.37 ± 0.95
CE-SV	98.35 ± 1.15	101.27 ± 0.29	47.79 ± 3.19	49.45 ± 1.44
CE-CV	101.09 ± 0.31	96.02 ± 2.70	40.95 ± 3.14	49.90 ± 1.39
Average	98.38 ± 1.18	100.12 ± 0.92	55.89 ± 2.79	74.32 ± 7.46
Average w/o SE-SV	97.21 ± 1.37	99.53 ± 1.18	50.84 ± 2.65	65.57 ± 7.97

Table 5: Normalized score of each algorithms on Walker-walk task with different settings. All results are reported as mean $\pm$ standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	96.31 ± 1.14	99.36 ± 0.73	16.68 ± 3.77	25.70 ± 0.80
SE-CV	76.90 ± 4.05	16.13 ± 4.53	19.92 ± 4.32	27.31 ± 1.35
CE-SV	22.73 ± 7.26	61.01 ± 6.01	12.68 ± 1.19	92.11 ± 4.04
CE-CV	72.50 ± 5.99	18.74 ± 2.51	10.17 ± 1.47	36.70 ± 30.80
Average	67.11 ± 8.47	48.81 ± 10.45	14.86 ± 1.57	45.45 ± 10.56
Average w/o SE-SV	57.38 ± 9.17	31.96 ± 7.62	14.26 ± 1.70	52.04 ± 13.52

Table 6: Normalized score of each algorithms on Walker-run task with different settings. All results are reported as mean $\pm$ standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	76.04 ± 1.00	38.45 ± 6.37	55.92 ± 1.92	88.96 ± 1.03
SE-CV	25.85 ± 0.39	19.72 ± 2.45	12.36 ± 1.67	81.61 ± 0.86
CE-SV	39.81 ± 6.02	51.82 ± 7.90	6.40 ± 1.18	76.27 ± 4.66
CE-CV	34.96 ± 3.69	3.84 ± 6.63	4.24 ± 0.87	71.25 ± 7.35
Average	44.17 ± 5.95	28.45 ± 6.09	19.73 ± 3.45	79.52 ± 2.73
Average w/o SE-SV	33.54 ± 2.89	25.12 ± 7.69	7.67 ± 0.96	76.38 ± 2.93

Table 7: Normalized score of each algorithms on Walker-jump task with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	69.93 ± 1.10	99.49 ± 0.40	26.48 ± 1.89	57.19 ± 0.48
SE-CV	68.99 ± 1.18	68.68 ± 0.98	35.81 ± 1.68	50.24 ± 0.79
CE-SV	47.60 ± 1.25	45.53 ± 13.66	11.03 ± 1.68	25.21 ± 0.70
CE-CV	55.10 ± 2.60	-2.91 ± 12.82	17.01 ± 1.84	24.29 ± 0.54
Average	60.40 ± 2.93	52.70 ± 11.96	22.58 ± 1.73	39.23 ± 4.44
Average w/o SE-SV	57.23 ± 3.26	37.10 ± 11.85	21.29 ± 2.19	33.25 ± 4.26

Table 8: Normalized score of each algorithms on Stickman-stand task with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	97.26 ± 0.61	100.93 ± 0.09	96.96 ± 2.40	98.51 ± 0.51
SE-CV	100.30 ± 0.13	97.50 ± 1.52	97.59 ± 1.18	99.75 ± 0.17
CE-SV	97.88 ± 0.36	101.05 ± 0.18	66.48 ± 2.71	100.83 ± 0.70
CE-CV	6.51 ± 2.89	99.51 ± 0.67	62.68 ± 4.76	100.99 ± 0.28
Average	75.49 ± 12.03	99.75 ± 0.56	80.93 ± 3.01	100.02 ± 0.36
Average w/o SE-SV	68.23 ± 15.46	99.36 ± 0.70	75.58 ± 3.42	100.52 ± 0.30

Table 9: Normalized score of each algorithms on Stickman-walk task with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	45.67 ± 3.90	99.33 ± 0.69	59.98 ± 6.32	91.00 ± 0.38
SE-CV	80.11 ± 4.15	23.13 ± 2.80	59.38 ± 6.78	94.33 ± 0.18
CE-SV	12.40 ± 5.19	87.76 ± 3.32	46.77 ± 9.57	22.12 ± 2.70
CE-CV	-0.01 ± 0.48	21.72 ± 3.07	20.79 ± 1.98	17.57 ± 0.58
Average	34.54 ± 9.54	57.99 ± 10.85	46.73 ± 4.11	56.25 ± 11.01
Average w/o SE-SV	30.83 ± 12.60	44.21 ± 11.00	42.31 ± 4.85	44.67 ± 12.46

Table 10: Normalized score of each algorithms on Stickman-run task with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	26.11 ± 2.53	47.62 ± 13.70	80.65 ± 1.14	58.65 ± 0.54
SE-CV	35.17 ± 1.80	14.90 ± 3.85	51.70 ± 1.47	59.25 ± 0.43
CE-SV	26.75 ± 0.40	33.58 ± 3.00	69.71 ± 0.56	69.49 ± 2.03
CE-CV	30.10 ± 0.99	19.51 ± 0.84	37.93 ± 3.47	61.68 ± 0.61
Average	29.53 ± 1.29	28.90 ± 4.96	60.00 ± 2.80	62.27 ± 1.39
Average w/o SE-SV	30.67 ± 1.37	22.66 ± 3.15	53.11 ± 2.71	63.47 ± 1.67

Table 11: Normalized score of each algorithms on Stickman-jump task with different settings. All results are reported as mean ± standard error.

Setting	GenRL	AIME-NoB	GIIL-ZS	GIIL-FT
SE-SV	52.16 ± 0.32	87.74 ± 6.46	35.50 ± 2.01	26.07 ± 0.37
SE-CV	38.72 ± 1.23	45.87 ± 1.83	36.37 ± 0.89	37.30 ± 3.34
CE-SV	52.43 ± 0.96	60.14 ± 0.75	36.59 ± 0.77	54.64 ± 0.10
CE-CV	45.14 ± 1.68	59.22 ± 0.51	32.07 ± 0.80	55.73 ± 0.71
Average	47.11 ± 1.78	63.24 ± 4.81	35.14 ± 0.66	43.43 ± 3.81
Average w/o SE-SV	45.43 ± 2.09	55.08 ± 2.38	35.01 ± 0.60	49.22 ± 3.14