2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Latent Action Diffusion for Cross-Embodiment Manipulation

Erik Bauer[1,3] and Elvis Nava[1,2,3,4] and Robert K. Katzschmann[1,2,3,4]

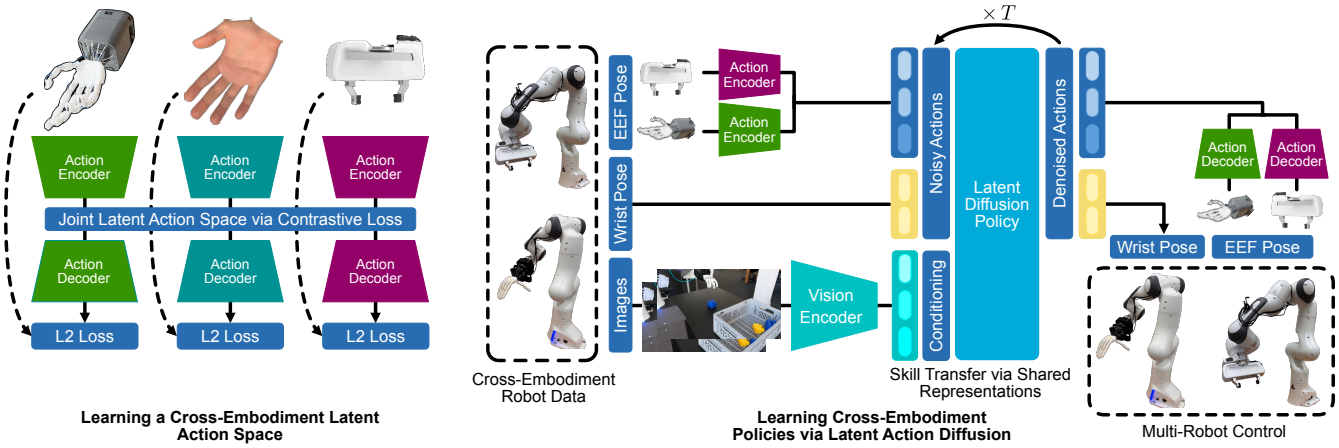

Fig. 1: Overview of our approach. **Left:** We construct a semantically aligned latent action space by training modality-specific encoders and decoders with a contrastive loss on retargeted end-effector (EEF) pose data from diverse end-effectors (dexterous hands, parallel gripper). **Right:** A single diffusion policy is trained in this shared latent space, enabling cross-embodiment policy learning, which in turn enables in multi-robot control with a single policy and skill transfer across embodiments.

***Abstract*— End-to-end learning is emerging as a powerful paradigm for robotic manipulation, but its effectiveness is limited by data scarcity and the heterogeneity of action spaces across robot embodiments. In particular, diverse action spaces across different end-effectors create barriers for cross-embodiment learning and skill transfer. We address this challenge through diffusion policies learned in a latent action space that unifies diverse end-effector actions. We first show that we can learn a semantically aligned latent action space for anthropomorphic robotic hands, a human hand, and a parallel jaw gripper using encoders trained with a contrastive loss. Second, we show that by using our proposed latent action space for co-training on manipulation data from different end-effectors, we can utilize a single policy for multi-robot control and obtain up to 25.3% improved manipulation success rates, indicating successful skill transfer despite a significant embodiment gap. Our approach using latent cross-embodiment policies presents a new method to unify different action spaces across embodiments, enabling efficient multi-robot control and data sharing across robot setups. This unified representation significantly reduces the need for extensive data collection for each new robot morphology, accelerates generalization across embodiments, and ultimately facilitates more scalable and efficient robotic learning.**

I. Introduction

End-to-end learning is a promising path toward creating adaptable, generalist robots. Scaling up both the data volume and diversity to match the desired model capabilities is extremely resource-intensive and expensive, and inevitably requires pooling together data from different robotic embodiments. However, efficiently learning from different embodiments remains a significant challenge, as observation and action spaces vary significantly across robots (the "embodiment gap").

Recent works on cross-embodiment learning have largely avoided explicitly addressing the problem of the embodiment gap in action spaces by only using data with a shared action space for pre-/co-training [1]–[3]. Other works showing pretraining on human manipulation datasets have relied on explicitly aligning the human action space to the robot action space [4]–[7]. In this work, instead of using an explicit action space, we introduce a learned latent action space which can encode diverse action spaces from different end-effectors into a unified, semantically aligned latent action space. To achieve semantic alignment within the latent action space, we utilize retargeting methods, which enable precise alignment of different end-effector action spaces. For policy learning with latent actions, we factorize policies into an embodiment-agnostic policy trained on latent actions and multiple embodiment-specific decoders that are trained separately. Our proposed framework combines the simplicity of training policies with aligned observation and action spaces while still enabling learning from diverse robotic embodiments.

In particular, we focus on embodiment transfer among single-arm robots with different end-effectors. For our experiments, we utilize the Faive robotic hand [8], the mimic hand [9] and a Franka parallel gripper. In two experiments,

[1]mimic robotics, Zurich, Switzerland
[2]ETH AI Center, ETH Zurich, Zurich, Switzerland
[3]Soft Robotics Lab, Dept. of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland
[4]Institute of Neuroinformatics, ETH Zurich and University of Zurich, Zurich, Switzerland
{erbauer, enava, rkk}@ethz.ch
Project page: https://mimicrobotics.github.io/lad/

pairing data from each dexterous hand with data from the Franka gripper utilizing our proposed framework (Fig. 1), we compare latent diffusion policies co-trained on cross-embodiment data with single-embodiment diffusion policies. We demonstrate that our methodology enables both cross-embodiment control with a single policy and facilitates positive skill transfer, with up to 25.3% performance (13.4% average) improvement compared to single-embodiment diffusion policies, despite a significant embodiment gap.

Our results indicate the potential of utilizing contrastive learning to bridge heterogeneous action spaces. As increasingly dexterous, human-like end-effectors become more common, our methodology provides a path forward for effectively sharing and reusing datasets across embodiments with diverse end-effectors through a unified latent action space.

A. Contributions

Our main contributions are:

1) We introduce a general framework that unifies diverse end-effector action spaces into a single, semantically aligned latent space using contrastive learning, enabling downstream learning across diverse robotic embodiments.
2) We show that factorizing diffusion policies into a latent, embodiment-agnostic policy and embodiment-specific action decoders enables multi-robot control across substantially different robot morphologies.
3) We demonstrate substantial real-world performance gains — up to 25.3% success rate improvement — from cross-embodiment co-training, and provide ablation studies validating our architecture for learning latent action spaces.

II. Related Works

Learning from cross-embodiment data is a promising path towards scaling up both the volume and diversity of training data for robot learning that has been considered from various angles in prior works[cite: 74].

a) Constrained Action Space: RT-1 [1] showed positive skill transfer by co-training on multi-robot datasets with the same action space. Building on the Open-X-Embodiment collaboration [10], multiple approaches have explored large-scale pretraining on more diverse robot data [2], [3], [11]. However, these approaches rely on constraining the action space for pretraining to a 7-dimensional space, effectively discarding any data with more complex action spaces, such as dexterous hands. In contrast, our approach utilizes a learned latent action space where the expressivity is a design choice detached from the physical constraints of any single embodiment.

b) One-Way Retargeting: Multiple approaches that focus on skill transfer from human video [4], [6], [7] propose retargeting human actions to the action space used by their robot. These approaches limit the utility of the resulting policy to the specific embodiment for which the retargeting was designed, as the policy predicts actions specific to that embodiment. Our methodology overcomes this bottleneck by learning a shared, semantically aligned latent action space that uses retargeting as a prior for alignment. This unified action representation supports any-to-any reconstruction across heterogeneous end-effectors (Fig. 3), enabling policies trained with latent actions to control multiple embodiments.

c) Alignment-Free Methods: Different approaches have investigated cross-embodiment learning without explicit alignment between different action spaces [12], [13]. However, skill transfer is largely driven by data scale for these architectures, as they lack architectural inductive biases to encourage transfer. The underlying assumption of large data volumes is often not satisfied for dexterous end-effectors with high-dimensional action spaces. Our proposed latent action space uses retargeting as an alignment prior that facilitates skill transfer between complex action spaces without requiring large data volumes.

Other methods propose video representations as embodiment-agnostic actions [14], [15], but these capture coarse motion primitives and require embodiment-specific finetuning before deployment. By learning a latent space directly from explicit action spaces instead of video, our approach encodes precise low-level actions, enabling deployment of the same model on different robots without further finetuning.

III. Methodology

Our proposed framework for cross-embodied policy learning consists of two parts: learning a latent action space (Fig. 2) and training latent policies. In this work, we focus on policy learning for single-arm robots with different end-effectors in a unified latent action space.

Our key insight is that learning aligned representations for different end-effector action spaces can be viewed as a multimodal representation learning problem. Based on this perspective, we design a pipeline for cross-embodied latent imitation learning comprised of the following steps: generating paired action data, learning encoders and decoders for the shared latent space, and latent policy learning.

A. Creating Aligned Action Pairs

Multimodal representation learning architectures for M modalities generally rely on tuples containing paired data of the form $\mathbf{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^M\right)$, where there is some form of cross-modal correspondence between the elements of each tuple. In multimodal learning, correspondences between data modalities are typically created through manual annotation (*e.g.*, image-caption pairs [16]) or created with modality-specific expert models (*e.g.*, creating depth pseudolabels from RGB images [17]).

In the context of robotics, we are looking for alignment functions between different action spaces which allow us to establish mappings in between the action spaces. We focus on aligning action spaces of different end-effectors (human hands, anthropomorphic robotic hands, parallel jaw gripper, ...). For this subproblem, retargeting functions from human hands to robotic end-effectors are a useful prior

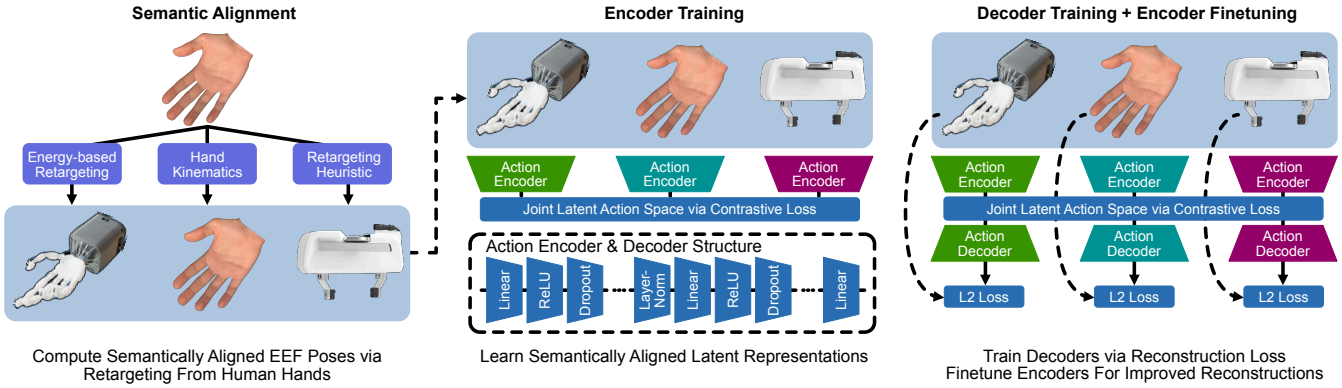

Fig. 2: The three-stage process for learning the cross-embodiment latent action space. Stage 1: Aligned end-effector (EEF) poses are generated by retargeting human hand poses to different robot end-effectors. Stage 2: Embodiment-specific encoders are trained to project these actions into a shared latent space using a contrastive loss. Stage 3: Decoders are trained to reconstruct the original poses from the latent space, and encoders are fine-tuned to improve reconstruction quality.

for alignment, as they typically already exist for different embodiments in order to teleoperate robots.

To construct tuples of paired end-effector poses, we proceed as follows:

$$\mathbf{x}_i = (x_i^H, f_H^{R_1}(x_i^H), \ldots, f_H^{R_M}(x_i^H)) \tag{1}$$

where $f_H^{R_j}$, $j \in \{1,\ldots,M\}$ are retargeting functions from human hands to the j-th robot embodiment.

1) Action Representations: For human hands, we derive a 189-dimensional pose representation θ_H using the local transformations in between the 21 joints according to the kinematic chain of the hand. To represent rotations, we utilize the continuous 6D rotation representation proposed by [18]. Poses for the Faive hand or mimic hand are represented as an 11- or 16-dimensional vector of joint angles θ_F or θ_M respectively. Poses of parallel jaw grippers are represented as normalized one-dimensional gripper width $\theta_P \in [0,1]$.

2) Retargeting: For retargeting, we follow the technique introduced by [19], which utilizes keyvectors for both the human and robot hand. The keyvectors $v_i^{\{H,M\}}(\theta_{\{H,M\}})$ are vectors from the palm to each fingertip and from each fingertip to all other fingertips and provide a unifying representation that can be defined for any hand with a notion of fingertips. To map from human hands to complex robotic hands such as the mimic hand, we can formulate retargeting functions as a minimum-energy solution to the squared keyvector difference of the human and the robot hand. By using the forward kinematics of each hand, we can determine the keyvectors as a function of its respective pose representation θ_H or θ_H. As a concrete example, to retarget from a human hand pose θ_H to a mimic hand pose θ_M, we can directly optimize over the θ_M with the differentiable objective shown in Equation (2). Each pair of keyvectors has a scaling factor s_i, which is used to compensate for different finger lengths. For all 15 keyvectors, scaling factors are determined through qualitative evaluation. The resulting retargeting function can be expressed as follows:

$$\theta_M(\theta_H) = \text{argmin}_{\theta_F} \sum_{i=1}^{15} \left|\left| v_i^H(\theta_H) - s_i v_i^F(\theta_M) \right|\right|_2^2 \tag{2}$$

For parallel jaw grippers, we take the minimum of all keyvectors originating at the thumb and normalize it by a standard gripper width W such that $\theta_P \in [0,1]$:

$$\theta_P(\theta_H) = \min_{\theta_P} \left(\min_i \frac{\left|\left| v_i^H(\theta_H) \right|\right|}{W}, 1 \right) \tag{3}$$

To add other robotic end-effectors to the learning scheme, it is only necessary to find a retargeting function from either human hands or another robotic end-effector to the newly added one.

B. Contrastive Latent Space Learning

For a shared latent action space, it is crucial that 1) for each modality, sufficient information is encoded such that we can precisely reconstruct end-effector poses and 2) the latent space has a coherent structure, meaning that the cross-modal alignment present in the model inputs during training is upheld in the learned latent space. To achieve both of these goals, we propose a two-step learning procedure: first, using batches with B aligned end-effector poses that were generated via retargeting, M modality-specific encoders $q_m, m \in 1\ldots M$ are trained that project actions x_m from each input modality into a shared latent space, where we utilize a pairwise InfoNCE loss [20] to ensure alignment within the batch:

$$\mathscr{L}_{\text{contrastive}} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \left(-\frac{1}{B} \sum_{n=1}^{B} \log \frac{\exp(q_i(x_i^n) \cdot q_j(x_j^n)/\tau)}{\sum_{k=1}^{B} \exp(q_i(x_i^n) \cdot q_j(x_j^k)/\tau)} \right) \tag{4}$$

where τ denotes the temperature. In the second stage, we train M modality-specific decoders $p_m, m \in 1\ldots M$, which

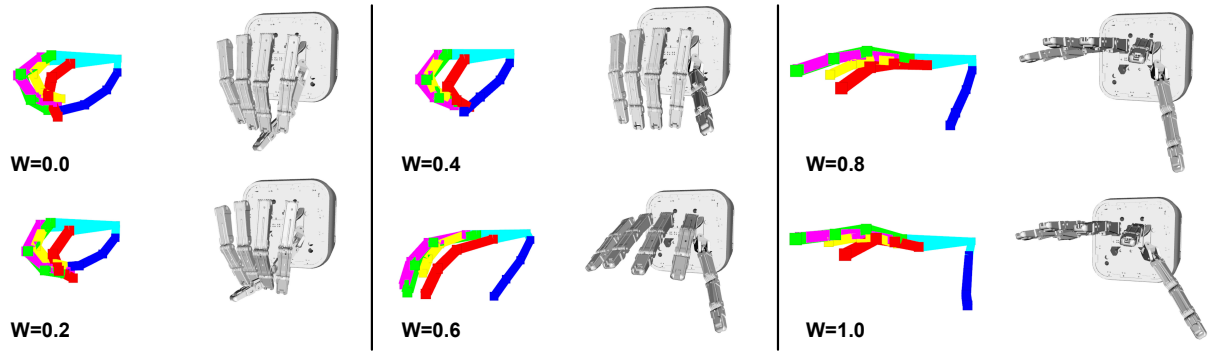

Fig. 3: Qualitative evaluation of the joint latent action space. We encode normalized gripper widths $W \in [0,1]$ (from closed to open) and perform cross-modal reconstruction by decoding them into human hand poses (colored lines on left) and poses for the Faive hand (grey model on right). Existing approaches using retargeting only allow for single-directional retargeting (i.e. human hands to robot hands), which is a limitation our latent action space overcomes. Any modality can be encoded and decoded to any other modality under the alignment constraints of the data.

learn to reconstruct ground truth actions $\hat{x}_i$ from their latent representations. Additionally, the encoders q_m are fine-tuned with a lower learning rate. The total loss $\mathscr{L}_{\text{total}}$ backpropagated through the encoders and decoders is a combination of a reconstruction loss $\mathscr{L}_{\text{recon}}$ and the previous contrastive loss $\mathscr{L}_{\text{contrastive}}$, where the hyperparameter λ can be used to control the trade-off in between alignment and self-reconstruction.

$$\mathscr{L}_{\text{recon}} = \frac{1}{M}\sum_{i=1}^{M}\sum_{n=1}^{B} ||p_i\left(q_i\left(x_i^n\right)\right) - \hat{x}_i^n||_2^2 \tag{5}$$

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{recon}} + \lambda \mathscr{L}_{\text{contrastive}} \tag{6}$$

We use a cross-reconstruction (CR) loss for validation. From modality i to j, given paired end-effector poses (x_i^n, x_j^n), the CR-Loss is:

$$\mathscr{L}_{\text{CR(i,j)}} = \frac{1}{B}\sum_{n=1}^{B} \left|\left|p_j\left(q_i\left(x_i^n\right)\right) - x_j^n\right|\right|_2^2 \tag{7}$$

We encode data from modality x_i^n, decode it to modality j and evaluate the result versus the paired ground truth data x_j^n. The self-reconstruction (SR) loss used for validation is $\mathscr{L}_{\text{SR(i)}} = \mathscr{L}_{\text{CR(i,i)}}$.

Action decoders and encoders are parameterized by standard multi-layer-perceptrons (MLPs) as encoders and decoders (Fig. 2). The input layer consists of a linear layer, ReLU layer and a dropout layer. For each hidden layer, we first have a normalization layer, then a linear layer, followed by a ReLU (rectified linear unit) activation and a dropout layer. After the hidden layers, a last linear layer is used to project the output to the desired dimension.

C. Policy Learning

With a learned latent action space, we employ Diffusion Policy [21] to map shared observations across datasets to latent end-effector actions of different embodiments and non-latent wrist poses of the robot arm as shown in Fig. 1. The encoders and decoders of the contrastive action model stay frozen for latent policy learning, with the diffusion objective being applied on the denoised latent end-effector actions and wrist poses. For inference, the suitable embodiment-specific decoder is used to obtain a decoded end-effector pose from the output of the latent policy.

We validate our method across two diffusion policy implementations: a transformer-based implementation [22] and a U-Net-based implementation [21]. The transformer-based policy is utilized in experiments with the Faive hand and Franka gripper, whereas the U-Net-based implementation is used for experiments with the mimic hand and the Franka gripper. Specifications for the models used in the experiments are given in Table I.

For co-training on differently sized datasets, we assign normalized weights w_j to all datasets. During training, we seek to combine samples from all datasets to fill batches with B samples in total. We sample per-dataset sub-batches with appropriately rounded sizes round$(\frac{B}{w_j})$, project the actions into the shared latent action space, normalize the sub-batches, and then concatenate them into a single batch for efficient training. Through this mechanism, the weight of each dataset approximately represents a sampling probability for each training step. In our experiments, all datasets have the same weights, such that the model is exposed to equal amounts of data from all embodiments.

TABLE I: Diffusion Policy Specifications

Hyperparameter	**Transformer**	**U-Net**
Parameter Count	~40M	~44M
Vision Encoder	Resnet18	ViT-S
Batch Size	300	256
Horizon	21 timesteps	48 timesteps
Diffusion Noise Schedule	Squared cosine	Squared cosine
β_{start}	0.0001	0.0001
β_{end}	0.02	0.02
Peak Learning Rate	0.0001	0.0001
Optimizer	AdamW	AdamW
Training Steps	90k	40k

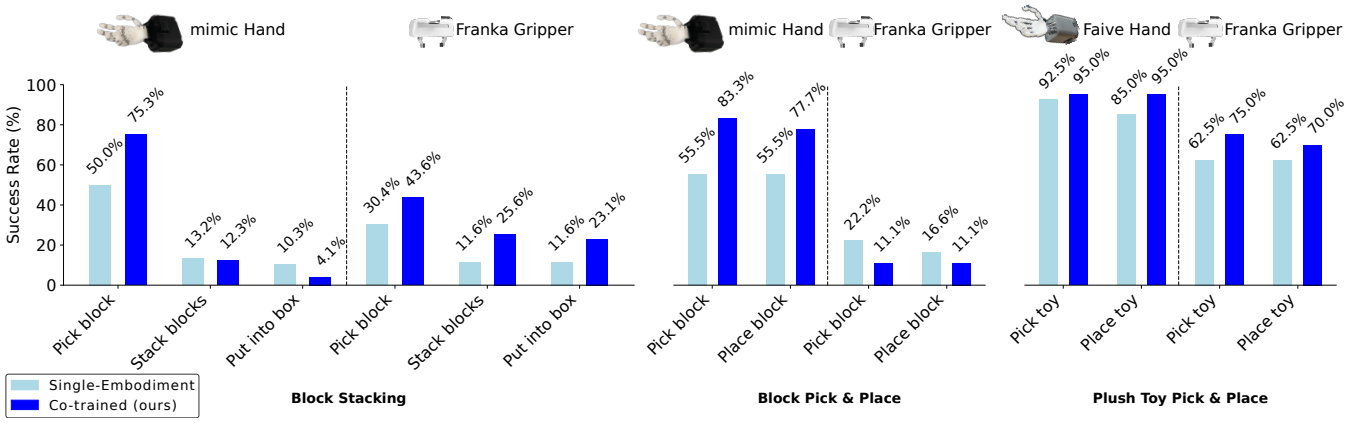

Fig. 4: Success rates for three different tasks comparing single-embodiment diffusion policies to cross-embodied latent diffusion policies trained on data from both embodiments for each task. Block stacking: 200 demos per embodiment, one external camera. Block pick and place: 200 demos per embodiment, one external camera + wrist camera for mimic hand, replaced by zero-padding for Franka gripper. Plush toy pick and place: 100 demos per embodiment, one external camera.

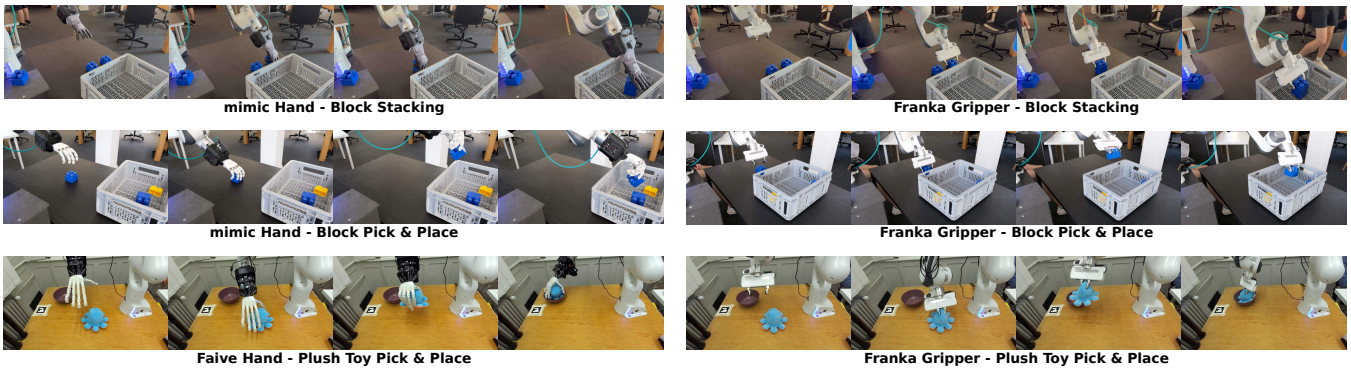

Fig. 5: Cross-embodiment policy rollouts for three tasks (block stacking, block pick and place, plush toy pick and place). For each task, both embodiments are controlled by the same cross-embodiment diffusion policy, demonstrating multi-robot control.

TABLE II: Contrastive Action Model Specifications

Hyperparameter	**Value**
Parameter Count (single encoder/decoder)	∼26k
Batch Size	16384
Learning Rate	0.00001
Finetuning Learning Rate	0.0001
Optimizer	AdamW
Weight Decay	0.001
Temperature Schedule	Exponential
Temperature Start→End	0.25→0.16
Latent Space Dimension	16
λ	0.1
MLP Hidden Dimensions	32, 128, 128, 32
Training Epochs (Encoders)	5000
Training Epochs (Decoders + finetune encoders)	10000

IV. Experimental Results and Discussion

We conducted experiments covering three different end-effectors and three tasks across two setups: one with the Faive hand and the Franka gripper and one setup with the mimic hand and the Franka gripper. For each end-effector in each setup, we compare single-embodiment policies with one cross-embodiment policy co-trained on data from all end-effectors. We train policies using contrastive action models that were trained for the end-effectors present in each task. In addition, we validate our design choices for our contrastive action model through an ablation study.

TABLE III: Ablation study on the impact of temperature annealing (TA) and finetuning (FT) on the contrastive action model, evaluating self- and cross-reconstruction losses.

Method	**SR-Loss**		**CR-Loss**	
	mimic	Franka	mimic→Franka	Franka→mimic
Full (ours)	**0.762**	3.7e-8	**0.002**	**214.20**
no TA	0.948	**1.5e-8**	0.007	286.64
no FT	44.76	2.6e-8	0.013	391.85
no FT&TA	49.765	2.1e-8	0.02	397.23

A. Ablation Study: Contrastive Action Model

To validate our design choices, we compare several versions of the contrastive action model (Table III). As metrics, we utilize self-reconstruction (SR) and cross-reconstruction

(CR) validation losses. The losses shown in the table are denormalized losses for the 16-dimensional action space of the mimic hand and the 1-dimensional action space of the Franka gripper. Due to the higher dimension and complexity of the larger action space, the losses for the mimic hand are larger than the losses for the Franka gripper.

We compare the full training pipeline for the contrastive model to three ablations. The ablation without temperature annealing keeps the temperature constant at the previous final value. The ablation without finetuning freezes the encoders in the second training step while the decoders are being trained. Both temperature annealing and finetuning the encoders reveal themselves to substantially improve both self- and cross-reconstruction metrics, with finetuning being the most important addition to the pipeline.

B. Experimental Setups

We evaluate our framework across three manipulation tasks using different combinations of end-effectors. For the Block Stacking task, we use the mimic hand and Franka gripper to pick a block, stack it on another, and place the pair into a box. For this task, we collected 200 demonstrations per embodiment using a single external camera for visual observations. Each policy is evaluated over 70 trials.

For the Block Pick & Place task, we use the mimic hand and the Franka gripper to place a plastic cube inside a box. We collected 200 demonstrations for each embodiment. In addition to an external camera and the robot arm pose, this setup uses a wrist camera for the mimic hand, which is replaced by zero-padding for the Franka gripper to test learning with asymmetric observations. Each policy is evaluated over 25 trials.

Finally, for the Plush Toy Pick & Place task, we use a Franka parallel gripper and a dexterous Faive hand to pick up a plush toy and place it into a bowl. For this setup, we collected 100 demonstrations for each end-effector and used a single external RGB camera for observations. Each policy is evaluated over 40 trials.

C. Results and Discussion

Our experiments demonstrate that co-training policies in a learned latent action space enable both multi-robot control from a single policy and significant performance gains through cross-embodiment skill transfer. We analyze the performance of our co-trained policies against single-embodiment baselines across three distinct manipulation tasks.

The multi-stage Block Stacking task showed the most pronounced skill transfer. For the initial coarse manipulation stage (”Pick block”), the co-trained policy yielded absolute success rate improvements of 25.3% for the mimic hand and 13% for the Franka gripper. More notably, the Franka gripper performance in the subsequent fine-grained stages (”Stack blocks” and ”Put into box”) improved significantly by 13% and 11%, respectively. This indicates that the gripper effectively learned more precise manipulation strategies from the dexterous hand data. The mimic hand saw a slight performance decrease in the final stage, likely due to the hand occluding the block from the camera after grasping it, making the final placement more difficult.

For the Block Pick & Place task, which featured asymmetric camera observations, the benefits of co-training were still evident for the more sensor-rich embodiment. The success rate of the mimic hand improved by 13% over its single-embodiment baseline. However, the Franka gripper’s performance decreased. We attribute this to the policy becoming reliant on the wrist camera view available only to the mimic hand, highlighting that skill transfer in the presence of asymmetric observations remains a significant challenge, despite the aligned action space.

In the Plush Toy Pick & Place task, the co-trained policy substantially outperformed the single-embodiment versions, achieving a 10% higher success rate for the Faive hand and a 7.5% improvement for the Franka gripper. This suggests that the policy learns a more robust, shared representation of the task, allowing skills learned from the dexterous hand (e.g., object handling) to benefit the simpler gripper, and vice versa.

In summary, our approach successfully unifies control across diverse robotic hardware and facilitates cross-embodiment skill transfer. The results confirm that policies learn helpful shared representations, leading to improved success rates in both coarse and fine-grained manipulation.

D. Limitations

a) Asymmetric Observations: Our policy struggles to achieve consistent performance on all embodiments if they have different sensor modalities (e.g., wrist camera for one robot only). This remains a fundamental challenge for cross-embodiment learning. Future work could explore vision encoder adaptation or learned observation alignment.

b) Latent Space Regularization: Our latent space is not explicitly regularized, which may lead to non-smooth regions that are harder to model for downstream policies. Integrating VAE-style priors or enforcing smoothness via time-contrastive losses could guarantee desirable latent space properties.

c) Dataset Scale Imbalance: We observe limited gains when adding large-scale datasets (e.g., BridgeV2 [23], DexYCB [24]) to our co-training mix. Training on datasets with highly imbalanced sizes and transferring skills despite significant environmental differences remains a fundamental challenge of cross-embodiment learning that could be explored via more sophisticated sampling strategies to determine dataset weights and visual representation learning that facilitates skill transfer given largely disjunct observation spaces.

V. Conclusion

Among current challenges in robotics, enabling effective skill transfer across diverse embodiments is of crucial importance to both maximize the volume and diversity of suitable training data and to ensure the reusability of training data throughout the life cycle of different end-effectors. To this

end, we frame cross-embodiment learning with different end-effectors as a multimodal representation learning problem and propose a two-stage pipeline to learn capable policies that can control multiple end-effectors. In real-world experiments with dexterous hands and a Franka parallel gripper, we demonstrate that through co-training on cross-embodiment data with our method for latent action spaces, we enable both multi-robot control and positive skill transfer across embodiments. In particular, the performance improvement of up to 25.3% (average: 13.4%) indicates that our method facilitates skill transfer between end-effectors with a large embodiment gap and underlines its potential for wider use across a broader range of robot morphologies. Future work includes expanding our method to a more diverse ecosystem of end-effectors and further investigating the behavior of skill transfer across different dataset sizes with more distinct visual differences.

ACKNOWLEDGMENT

We express our gratitude to Emanuele Palumbo for valuable discussions and insights on multimodal learning architectures. Furthermore, we are grateful for support from Chenyu Yang in setting up the initial version of srl_il and Davide Liconti in writing the control interface for the Franka gripper. Additionally, we thank the mimic team for support in realizing experiments with the mimic hand: Victor Montesinos, Jonas Pai, Benedek Forrai, Robert Malate, Philipp Wand, Stephan Polinski, and Norica Bacuieti.

We are grateful for grant funding from Innosuisse (122.679 SIP) and ESA BIC received by mimic robotics in connection to this project. We are also grateful to funding supporting academic co-authors: funding from the ETH AI Center, SNSF Project Grant MINT #200021_215489, Swiss Data Science Center, Amazon Research Award, Armasuisse.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *ArXiv*, vol. abs/2212.06817, 2022.

[2] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, 2023.

[3] Octo Model Team, D. Ghosh, H. Walke, *et al.*, *Octo: An open-source generalist robot policy*, 2024. arXiv: 2405.12213 [cs.RO].

[4] K. Shaw, S. Bahl, and D. Pathak, *Videodex: Learning dexterity from internet videos*, 2022. arXiv: 2212.04498 [cs.RO].

[5] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, *Pushing the limits of cross-embodiment learning for manipulation and navigation*, 2024. arXiv: 2402.19432 [cs.RO].

[6] 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. arXiv: 2410.24221 [cs.RO].

[7] 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*, 2023. arXiv: 2302.12422 [cs.RO].

[8] Y. Toshimitsu, B. Forrai, B. G. Cangan, U. Steger, M. Knecht, S. Weirich, and R. K. Katzschmann, *Getting the ball rolling: Learning a dexterous policy for a biomimetic tendon-driven hand with rolling contact joints*, 2023. arXiv: 2308.02453 [cs.RO].

[9] E. Nava, V. Montesinos, E. Bauer, *et al.*, *Mimic-one: A scalable model recipe for general purpose robot dexterity*, 2025. arXiv: 2506.11916 [cs.RO].

[10] E. Collaboration, A. O'Neill, A. Rehman, *et al.*, *Open x-embodiment: Robotic learning datasets and rt-x models*, 2024. arXiv: 2310.08864 [cs.RO].

[11] M. J. Kim, K. Pertsch, S. Karamcheti, *et al.*, *Openvla: An open-source vision-language-action model*, 2024. arXiv: 2406.09246 [cs.RO].

[12] K. Black, N. Brown, D. Driess, *et al.*, *π_0: A vision-language-action flow model for general robot control*, 2024. arXiv: 2410.24164 [cs.LG].

[13] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, *Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation*, 2024. arXiv: 2408.11812 [cs.RO].

[14] S. Ye, J. Jang, B. Jeon, *et al.*, *Latent action pretraining from videos*, 2024. arXiv: 2410.11758 [cs.RO].

[15] Y. Chen, Y. Ge, Y. Li, Y. Ge, M. Ding, Y. Shan, and X. Liu, "Moto: Latent motion token as the bridging language for robot manipulation," *arXiv preprint arXiv:2412.04445*, 2024.

[16] T. M. Sutter, I. Daunhawer, and J. E. Vogt, *Generalized multimodal elbo*, 2021. arXiv: 2105.02470 [cs.LG].

[17] D. Mizrahi, R. Bachmann, O. F. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir, *4m: Massively multimodal masked modeling*, 2023. arXiv: 2312.06647 [cs.CV].

[18] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, *On the continuity of rotation representations in neural networks*, 2020. arXiv: 1812.07035 [cs.LG].

[19] A. Sivakumar, K. Shaw, and D. Pathak, *Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube*, 2022. arXiv: 2202.10448 [cs.RO].

[20] A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*, 2019. arXiv: 1807.03748 [cs.LG].

[21] 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.

[22] A. Mandlekar, D. Xu, J. Wong, *et al.*, "What matters in learning from offline human demonstrations for robot manipulation," in *arXiv preprint arXiv:2108.03298*, 2021.

[23] H. Walke, K. Black, A. Lee, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[24] Y.-W. Chao, W. Yang, Y. Xiang, *et al.*, "DexYCB: A benchmark for capturing hand grasping of objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.