UniLatent: Cross-Embodiment Transfer via Latent Observation Alignment

Anonymous Authors
Submitted to ICRA Workshop

***Abstract*—Effective robot learning requires diverse and realistic data, yet collecting such data is expensive and often embodiment-specific. As a result, existing datasets are fragmented across embodiments. While prior work has explored cross-embodiment transfer, it remains challenging due to the large visual gap between robots, especially under third-person viewpoints. We propose UniLatent, a cross-embodiment transfer framework based on observation alignment. UniLatent renders motion-aligned views of different robots in simulation and aligns their visual encoders so that embodiment-specific observations map into a shared latent space. Policies trained in this latent space transfer efficiently across robots without explicit pixel-level translation. Across simulation and real-world experiments, UniLatent outperforms pixel-translation baselines by over 30% on average and enables effective few-shot real-world transfer.**

***Index Terms*—cross-embodiment transfer, visual representation learning, imitation learning, manipulation.**

I. Introduction

Large-scale and diverse data is a key driver of recent progress in robot learning [1, 2], and recent efforts have produced several sizable manipulation datasets [3–5]. However, existing data is highly imbalanced across embodiments and concentrated on a small set of popular robots, while many other arms and grippers remain underrepresented. Moreover, much of existing data is collected via teleoperation, making it inherently tied to specific robot embodiments. Meanwhile, real-world robot data collection is costly and embodiment-specific, since demonstrations depend on the robot's morphology, appearance, end-effector geometry, and kinematics. As hardware continues to diversify, recollecting data for each new embodiment does not scale, motivating methods that reuse existing demonstrations—particularly teleoperated data—and generalize to unseen robots.

A key challenge in cross-embodiment transfer is the visual gap between robots with different morphologies, which often leads to significant performance degradation when deploying a vision-based policy on a new robot [6, 7]. This issue is particularly pronounced under third-person viewpoints, where the robot occupies a large portion of the image. Prior work addresses this gap through image editing or generative translation, modifying demonstration frames via masking [8], inpainting [6], or diffusion [7] to match the target robot's visual domain. However, such pixel-space translation is inherently brittle: errors in segmentation, inpainting, or generation introduce systematic artifacts, and the resulting images often form a separate synthetic domain that policies may overfit to.

To address this challenge, we propose UniLatent, a cross-embodiment transfer pipeline that maps observations from different robots into a shared latent space, enabling a policy trained on source data to transfer to a target robot in a few-shot manner (Fig. 1). The core idea is to learn embodiment-invariant visual features using simulation-rendered paired data. Specifically, we render large-scale paired images in which source and target robots execute matched actions under the same camera, while randomizing backgrounds, lighting, and scene factors. We instantiate a pretrained vision model as a fixed target encoder and train a source encoder to match its representation by aligning embeddings of paired images. After alignment, source demonstrations are projected into the shared latent space and used to train a latent-space policy, which is deployed on the target robot using the target encoder, avoiding pixel-level translation entirely.

Contributions.: (1) A simulation-based pipeline that renders large-scale paired observations for different robots executing matched actions, providing reliable supervision for aligning visual encoders across embodiments. (2) A latent-space policy learning framework that achieves $\sim$30% higher success on average than pixel-space translation baselines in simulation zero-shot transfer, and enables effective real-world few-shot transfer with only 10 target demonstrations.

Related work.: Prior visual domain adaptation for robotics, including domain randomization [9], photometric editing [10–12], and generative augmentation [13–15], largely assumes a *fixed* embodiment and attributes domain gaps to environmental factors. A second line of work learns *robot-generalist* policies through multi-robot co-training [16–19], shared action abstractions [20, 21], or embodiment conditioning [22], typically requiring large-scale multi-robot datasets. Most closely related are cross-embodiment transfer methods that bridge the visual gap via test-time cross-painting (Mirage [6]), diffusion-based augmentation (RoVi-Aug [7]), or segmentation masking (Shadow [8]). These rely on accurate image-editing components and can introduce artifacts that corrupt the policy input. In contrast, UniLatent operates in representation space and avoids pixel-level editing at both training and deployment.

II. Method: UniLatent

A. Problem Setup

We study *robot-to-robot cross-embodiment imitation*: demonstrations collected on a source robot $R_{\mathcal{S}}$ are used to learn a policy for a target robot $R_{\mathcal{T}}$ under minimal target-robot

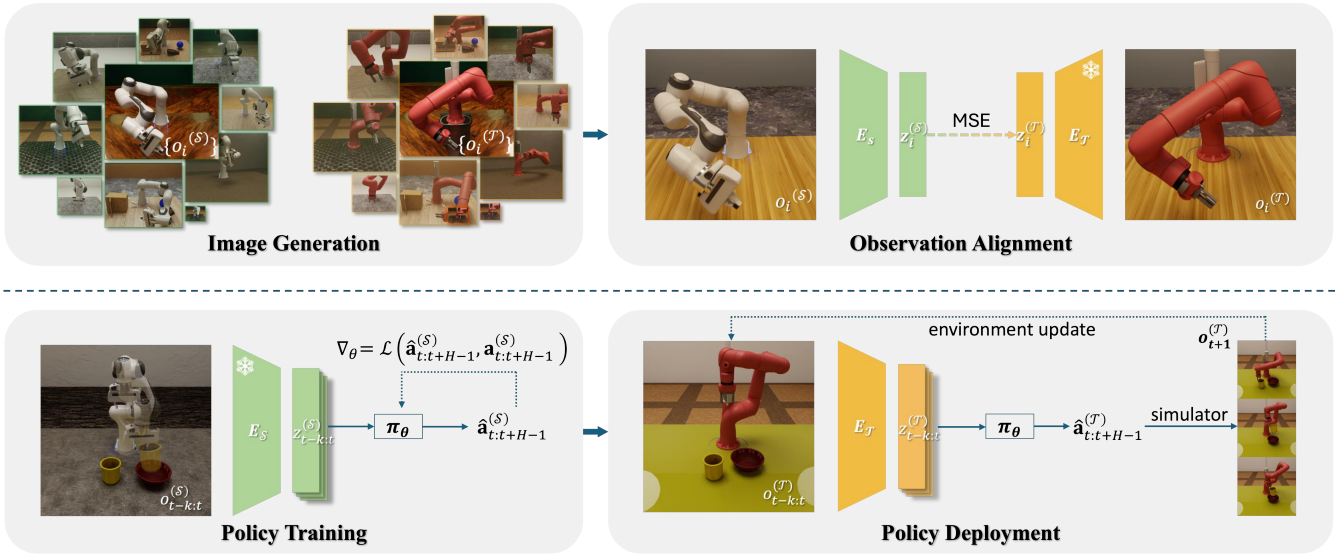

Fig. 1: **Pipeline overview of UniLatent.** We render paired (source, target) observations $\{(o_i^{(\mathcal{S})}, o_i^{(\mathcal{T})})\}$ for matched end-effector poses and use them to align encoders by training $E_{\mathcal{S}}$ to match a fixed $E_{\mathcal{T}}$ in a shared latent space. A latent policy π_θ is then trained on source demonstrations and deployed on the target robot by encoding target observations with $E_{\mathcal{T}}$ and calling π_θ.

data. Each robot has RGB observations $o^{(R)}$ and end-effector actions $a^{(R)}$. We assume access to M source demonstrations $\{\tau_i\}_{i=1}^{M}$, where $\tau = \{(o_t^{\mathcal{S}}, a_t^{\mathcal{S}})\}_{t=1}^{T}$, under a fixed camera. We focus on single-arm manipulators with parallel-jaw grippers that share an end-effector control interface. Robots may differ in geometry, kinematics, and appearance. We do *not* assume paired or time-synchronized cross-robot trajectories.

B. Pipeline Overview

UniLatent has four stages (Fig. 1): (i) generate paired observations $\{(o_i^{\mathcal{S}}, o_i^{\mathcal{T}})\}$ in simulation; (ii) align visual encoders into a shared latent space; (iii) train a latent policy π_θ on encoded source demonstrations; (iv) deploy π_θ on $R_{\mathcal{T}}$ by encoding target observations with the target encoder.

C. Paired Image Generation

We render large-scale paired frames in simulation, where $R_{\mathcal{S}}$ and $R_{\mathcal{T}}$ execute matched actions under identical camera setups. Our environment is built in IsaacLab [23] on top of RoboVerse [24]. We sample over 10k end-effector poses $\{p_i\}$ covering task-relevant regions of the reachable workspace, solve inverse kinematics for both robots, and render images under identical scenes. We apply domain randomization over backgrounds, lighting, and object assets to obtain N paired RGB images $\{(o_i^{\mathcal{S}}, o_i^{\mathcal{T}})\}_{i=1}^{N}$; the same randomized parameters are applied to both images within each pair, preserving action-level correspondence.

D. Observation Alignment

We initialize both $E_{\mathcal{S}}$ and $E_{\mathcal{T}}$ from the same DINOv2 model [25], which provides strong visual features for robotics [26, 27]. We keep $E_{\mathcal{T}}$ frozen to anchor the latent space, and train $E_{\mathcal{S}}$ with an MSE loss on paired embeddings:

$$\mathcal{L} = \left\| E_{\mathcal{S}}(o_i^{\mathcal{S}}) - E_{\mathcal{T}}(o_i^{\mathcal{T}}) \right\|_2^2. \tag{1}$$

Anchoring to a fixed $E_{\mathcal{T}}$ ensures a consistent test-time representation for the target robot, prevents representation drift, and lets $E_{\mathcal{S}}$ adapt to embodiment-specific appearance differences.

E. Latent-Space Policy Training and Deployment

After alignment, we project source demonstrations into the shared latent space using $E_{\mathcal{S}}$ and train a policy conditioned on latent observation histories. For each timestep t, we form a window $\mathbf{z}_{t-k:t}^{\mathcal{S}} = \{E_{\mathcal{S}}(o_{t-k}^{\mathcal{S}}), \ldots, E_{\mathcal{S}}(o_t^{\mathcal{S}})\}$ and predict the next action chunk $\mathbf{a}_{t:t+H-1}^{\mathcal{S}}$ [28]. We instantiate π_θ as a conditional flow-matching model [29, 30]: sampling $\mathbf{a}_0 \sim \mathcal{N}(0, I)$, forming an interpolation $\mathbf{a}_u$ between $\mathbf{a}_0$ and the target $\mathbf{a}_1$ for $u \in [0, 1]$, and predicting a velocity field trained with

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big[\|v_\theta(\mathbf{a}_u, u \mid \mathbf{z}_{t-k:t}^{\mathcal{S}}) - v^\star(\mathbf{a}_u, u)\|_2^2\big]. \tag{2}$$

At deployment, we encode $o_t^{\mathcal{T}}$ with the fixed $E_{\mathcal{T}}$ to obtain $z_t^{\mathcal{T}}$, query π_θ on $z_{t-k:t}^{\mathcal{T}}$, and integrate the learned flow from $u{=}0$ to $u{=}1$ with 10 Euler steps to produce actions executed by a robot-specific controller. UniLatent is largely insensitive to the policy architecture. We found that using Diffusion Policy [31] variants yield similar trends.

III. Experiments

We evaluate UniLatent in simulation for zero-shot cross-embodiment transfer, and in the real world for few-shot transfer.

A. Simulation Experiments

Setup.: We build three manipulation tasks on the RoboVerse [24] platform with Isaac Sim: (1) *Close Box Lid* (rotate a box lid to close it), (2) *Stack Cup into Bowl* (pinch a thin cup and insert it into a bowl), and (3) *Kick Football* (push a small ball into a wire goal with gripper closed). See Fig. 2. We use Franka as the source and transfer to xArm

Task & Setup	Close Box Lid			Stack Cup into Bowl			Kick Football		
	Franka	xArm	Sawyer	Franka	xArm	Sawyer	Franka	xArm	Sawyer
Direct Policy Transfer	97%	2%	5%	80%	0%	0%	90%	1%	1%
UniLatent w/o observation alignment	97%	40%	16%	89%	3%	0%	93%	11%	1%
RoVi-Aug*	97%	93%	54%	89%	15%	5%	93%	60%	40%
UniLatent (Ours)	97%	**93%**	**81%**	89%	**65%**	**51%**	93%	**89%**	**77%**

TABLE I: **Zero-shot cross-embodiment transfer in simulation.** We train on Franka demonstrations and evaluate zero-shot transfer to xArm and Sawyer across three manipulation tasks. Each entry is the success rate over 100 rollouts. UniLatent consistently outperforms direct transfer, a generalist-encoder ablation, and the pixel-space translation baseline.

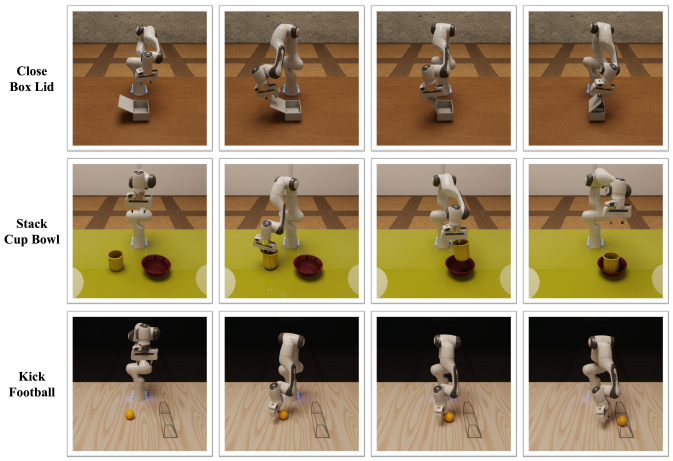

Fig. 2: **Simulation tasks.** Example frames from our simulation experiments. Each row shows representative frames of a Franka demonstration under a fixed third-person camera.

and Sawyer. For each task, we collect $M = 300$ teleoperated demonstrations on Franka using DexCap [32], filter failed rollouts, and render ray-traced RGB from a fixed third-person camera with randomized initial object poses. We evaluate with 100 rollouts per task under an automatic success checker, with object poses randomized independently from training.

Implementation details.: We render $N \approx 100$k paired images per target embodiment and fine-tune with an additional 10k task-specific paired frames per task. The policy is a conditional UNet [33] flow-matching model with observation history k=3 and action horizon H=8. The aligned $E_{\mathcal{S}}$ is kept frozen during policy training.

Baselines and ablations.: We compare against: (1) *Direct Policy Transfer* using an MLP-based scratch encoder trained end-to-end on source demonstrations and deployed on the target without alignment; (2) *UniLatent w/o alignment*, which uses a single shared DINOv2 encoder on both sides; and (3) *RoVi-Aug**, a strong pixel-space translation baseline we construct by decoding the aligned source latents back to the target domain using a pretrained Representation Autoencoder [34], $o' = D_{\mathcal{T}}(E_{\mathcal{S}}(o^{\mathcal{S}}))$, and training a pixel-space policy on $\{(o', a^{\mathcal{S}})\}$ (see Fig. 3 for decoded examples). We found this decoding-based translation more stable than RoVi-Aug's image-to-image diffusion pipeline in our setting, making it a strong proxy for RoVi-Aug [7].

Results.: Table I summarizes the results. Across tasks and target embodiments, UniLatent consistently outperforms

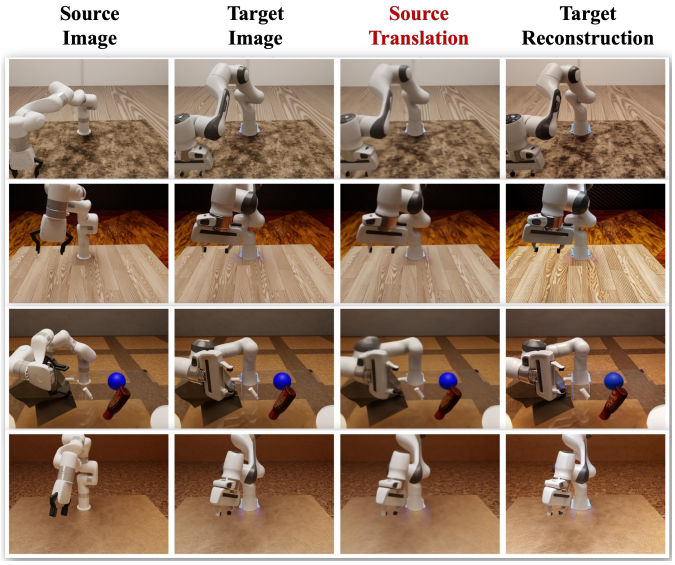

Fig. 3: **Decoding-based translation for the RoVi-Aug* baseline.** Columns: (1) source observation $o^{\mathcal{S}}$, (2) paired target observation $o^{\mathcal{T}}$, (3) decoded translation $o'{=}D_{\mathcal{T}}(E_{\mathcal{S}}(o^{\mathcal{S}}))$, and (4) target reconstruction $D_{\mathcal{T}}(E_{\mathcal{T}}(o^{\mathcal{T}}))$. Column 3 shows the training inputs to the pixel-space baseline.

all ablations, reaching up to 93% success after transfer with only a 4–38% absolute drop from the in-domain Franka result. Direct transfer and a shared DINOv2 encoder perform poorly, highlighting the severity of embodiment-induced visual shift and the need for explicit observation alignment. Pixel-space translation (RoVi-Aug*) partially closes the gap but remains substantially worse than UniLatent, especially on Franka→Sawyer where the embodiment difference is largest. Qualitatively, decoded frames are brittle (Fig. 3), exhibiting blurring and temporal jitter that degrade the policy input; operating in latent space avoids these artifacts. The advantage of UniLatent grows with embodiment disparity: while methods perform similarly on Franka→xArm for *Close Box Lid*, UniLatent achieves marked improvement on Franka→Sawyer, which exhibits larger visual and kinematic differences.

Setting	Close Drawer	Close Box Lid	Wipe Table
Source-only (xArm)	10/10	8/10	10/10
Target-only (few-shot, Franka)	7/10	2/10	2/10
RoVi-Aug pretrain + few-shot	9/10	9/10	5/10
Cross-emb pretrain + few-shot (Ours)	**10/10**	**9/10**	**8/10**

TABLE II: **Few-shot transfer in the real world.** UniLatent pretraining substantially outperforms training from scratch, and matches or exceeds RoVi-Aug [7] pretraining.

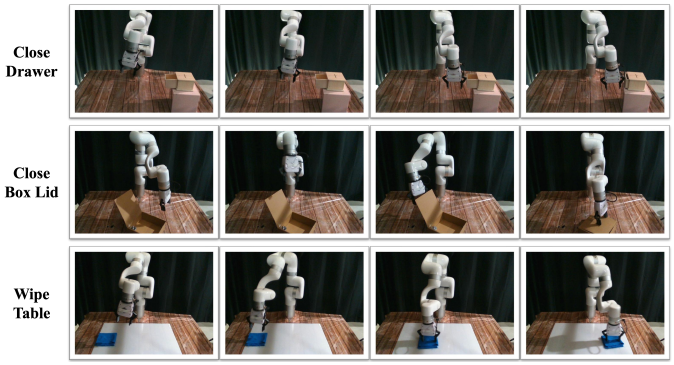

Fig. 4: **Real-world tasks.** Example frames from *Close Drawer*, *Close Box Lid*, and *Wipe Table*, collected via teleoperation on xArm under a fixed third-person camera.

B. Real-World Experiments

Setup.: We transfer from xArm (source) to Franka (target) on three tabletop tasks (Fig. 4): *Close Drawer*, *Close Box Lid*, and *Wipe Table*. We collect 150 source demonstrations on xArm for *Close Drawer* and 61 each for the other two, at 10 Hz with a RealSense D435i camera at 640×480, keeping the camera extrinsics fixed across embodiments. We additionally collect $M_{\mathcal{T}} = 10$ demonstrations on Franka per task for few-shot fine-tuning, which mitigates residual sim-to-real and hardware discrepancies. We evaluate with 10 rollouts per task.

Implementation details.: We render $N \approx 60$k path-traced paired frames and train a single task-agnostic $E_{\mathcal{S}}$ for all real tasks. Policy architecture and hyperparameters match the simulation setup (k=3, H=8, conditional UNet, 10 Hz control).

Results.: Table II shows that RoVi-Aug pretraining improves over training from scratch on the target robot in several tasks, indicating that pixel-space translation can provide useful transfer signal in real-world settings. However, UniLatent consistently achieves stronger performance across tasks, with the largest gains on tasks that involve more complex or contact-rich interactions (e.g., *Wipe Table*), where accurate preservation of geometry and gripper pose is critical.

We attribute this gap to the brittleness of pixel-space translation under real-world conditions: errors from segmentation, generation, or occlusion can introduce inconsistencies in robot appearance and geometry, which directly corrupt the visual inputs consumed by the policy. In contrast, UniLatent aligns observations in a compact latent space and executes the policy on target observations without explicit translation, reducing sensitivity to such artifacts.

Overall, these results suggest that representation learning provides a more robust pathway for cross-embodiment transfer in the real world than relying on pixel-space translation.

C. Multi-source transfer.

UniLatent extends to *multiple* source embodiments by aligning each source encoder to the same frozen target encoder and training a shared latent policy on the union of source demonstrations. On *Close Box Lid* (xArm target, Table III), combining 100 Sawyer and 100 Franka demonstrations achieves 91% success, on par with using 200 Franka demonstrations alone and substantially outperforming 200 Sawyer demonstrations.

This suggests that the shared latent space effectively integrates heterogeneous source data without requiring a unified encoder or specialized multi-robot architecture.

Sawyer (100)	Franka (100)	Sawyer (200)	Franka (200)	**Both (200)**
64	63	84	92	**91**

TABLE III: **Multi-source transfer.** Success rate (%) on *Close Box Lid* when transferring to xArm with source demonstrations from Sawyer only, Franka only, or both. Numbers in parentheses denote the total source demonstrations used.

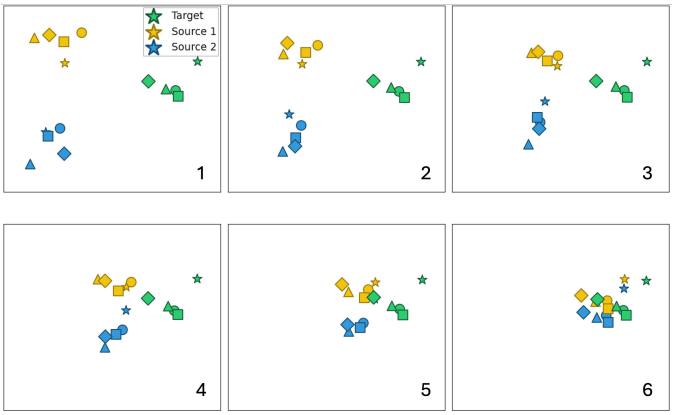

Fig. 5: **t-SNE of latent embeddings during alignment.** Each point is a latent embedding of a rendered image; colors indicate embodiments and marker shapes indicate paired samples. Across training iterations (1→6), embeddings from different robots become increasingly overlapping, and paired images converge to nearby points in latent space.

D. Alignment analysis.

To better understand what observation alignment learns, we visualize the latent space using t-SNE [35] on embeddings of rendered images from three robots (Fig. 5). Before alignment, embeddings form separated clusters by embodiment, reflecting the large visual gap. As training progresses, the clusters move closer and increasingly overlap, and paired images converge to nearby points in latent space. This confirms that the source encoder converges to a target-consistent representation and supports our design of anchoring the shared space to a fixed target encoder: because $E_{\mathcal{T}}$ is never updated, its embedding geometry remains a stable target for $E_{\mathcal{S}}$ to match, and alignment transfers directly to the test-time encoder.

IV. Conclusions and Limitations

UniLatent achieves strong cross-embodiment transfe by aligning visual encoders into a shared latent space rather than translating pixels. We focus on parallel-jaw arms; extending to multi-finger hands and embodiments with substantially different workspaces is left to future work.

REFERENCES

[1] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, and M. Schwager, "Foundation models in robotics: Applications, challenges, and the future," *arXiv preprint arXiv:2312.07843*, 2023.

[2] Y. Zhong, F. Bai, S. Cai, X. Huang, Z. Chen, X. Zhang, Y. Wang, S. Guo, T. Guan, K. N. Lui, Z. Qi, Y. Liang, Y. Chen, and Y. Yang, "A survey on vision-language-action models: An action tokenization perspective," *arXiv preprint arXiv:2507.01925*, 2025.

[3] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, M. Z. Irshad, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, V. Guizilini, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[4] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, V. Guizilini, D. A. Herrera, M. Heo, K. Hsu, J. Hu, M. Z. Irshad, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems (RSS)*, 2024.

[5] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, X. He, X. Huang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[6] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," in *Robotics: Science and Systems (RSS)*, 2024.

[7] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg, "Roviaug: Robot and viewpoint augmentation for cross-embodiment robot learning," in *Conference on Robot Learning (CoRL)*, 2024.

[8] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for cross-embodiment policy transfer," in *Conference on Robot Learning (CoRL)*, 2024.

[9] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[10] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[11] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, D. M, J. Peralta, B. Ichter, K. Hausman, and F. Xia, "Scaling robot learning with semantically imagined experience," in *Robotics: Science and Systems (RSS)*, 2023.

[12] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-invariant policy learning via zero-shot novel view synthesis," in *Conference on Robot Learning (CoRL)*, 2024.

[13] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar, "Semantically controllable augmentations for generalizable robot learning," in *International Journal of Robotics Research (IJRR)*, 2024.

[14] C. Yuan, S. Joshi, S. Zhu, H. Su, H. Zhao, and Y. Gao, "Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[15] J. Chen, I.-C. A. Liu, G. Sukhatme, and D. Seita, "ROPA: Synthetic robot pose generation for RGB-D bimanual data augmentation," *arXiv preprint arXiv:2509.19454*, 2025.

[16] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak,

T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," in *Robotics: Science and Systems (RSS)*, 2025.

[17] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems (RSS)*, 2024.

[18] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *Conference on Robot Learning (CoRL)*, 2024.

[19] S. Kareer, K. Pertsch, J. Darpinian, J. Hoffman, D. Xu, S. Levine, C. Finn, and S. Nair, "Emergence of human to robot transfer in vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2512.22414

[20] G. Salhotra, I.-C. A. Liu, and G. Sukhatme, "Learning robot manipulation from cross-morphology demonstration," in *Conference on Robot Learning (CoRL)*, 2023.

[21] J. Zheng, J. Li, D. Liu, Y. Zheng, Z. Wang, Z. Ou, Y. Liu, J. Liu, Y.-Q. Zhang, and X. Zhan, "Universal actions for enhanced embodied foundation models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[22] T. Chen, A. Murali, A. Gupta, and J. Malik, "Hardware conditioned policies for multi-robot transfer learning," in *Neural Information Processing Systems (NeurIPS)*, 2018.

[23] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Muñoz, X. Yao, R. Zurbrügg, N. Rudin, L. Wawrzyniak, M. Rakhsha, A. Denzler, E. Heiden, A. Borovicka, O. Ahmed, I. Akinola, A. Anwar, M. T. Carlson, J. Y. Feng, A. Garg, R. Gasoto, L. Gulich, Y. Guo, M. Gussert, A. Hansen, M. Kulkarni, C. Li, W. Liu, V. Makoviychuk, G. Malczyk, H. Mazhar, M. Moghani, A. Murali, M. Noseworthy, A. Poddubny, N. Ratliff, W. Rehberg, C. Schwarke, R. Singh, J. L. Smith, B. Tang, R. Thaker, M. Trepte, K. V. Wyk, F. Yu, A. Millane, V. Ramasamy, R. Steiner, S. Subramanian, C. Volk, C. Chen, N. Jawale, A. V. Kuruttukulam, M. A. Lin, A. Mandlekar, K. Patzwaldt, J. Welsh, H. Zhao, F. Anes, J.-F. Lafleche, N. Moënne-Loccoz, S. Park, R. Stepinski, D. V. Gelder, C. Amevor, J. Carius, J. Chang, A. H. Chen, P. de Heras Ciechomski, G. Daviet, M. Mohajerani, J. von Muralt, V. Reutskyy, M. Sauter, S. Schirm, E. L. Shi, P. Terdiman, K. Vilella, T. Widmer, G. Yeoman, T. Chen, S. Grizan, C. Li, L. Li, C. Smith, R. Wiltz, K. Alexis, Y. Chang, D. Chu, L. J. Fan, F. Farshidian, A. Handa, S. Huang, M. Hutter, Y. Narang, S. Pouya, S. Sheng, Y. Zhu, M. Macklin, A. Moravanszky, P. Reist, Y. Guo, D. Hoeller, and G. State, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.

[24] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, Y. Liang, D. Goetting, C. Xu, H. Chen, Y. Qian, Y. Geng, J. Mao, W. Wan, M. Zhang, J. Lyu, S. Zhao, J. Zhang, J. Zhang, C. Zhao, H. Lu, Y. Ding, R. Gong, Y. Wang, Y. Kuang, R. Wu, B. Jia, C. Sferrazza, H. Dong, S. Huang, Y. Wang, J. Malik, and P. Abbeel, "Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning," in *Robotics: Science and Systems (RSS)*, 2025.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2024.

[26] Z. Dong, Y. Liu, Y. Li, H. Zhao, and J. Hao, "Conditioning matters: Training diffusion policies is faster than you think," 2025. [Online]. Available: https://arxiv.org/abs/2505.11123

[27] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," 2025. [Online]. Available: https://arxiv.org/abs/2411.04983

[28] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[29] Q. Liu, "Rectified flow: A marginal preserving approach to optimal transport," *arXiv preprint arXiv:2209.14577*, 2022.

[30] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[31] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

[32] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," in *Robotics: Science and Systems (RSS)*, 2024.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015.

[34] B. Zheng, N. Ma, S. Tong, and S. Xie, "Diffusion transformers with representation autoencoders," 2025.

[35] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research (JMLR)*, vol. 9, no. 86, pp. 2579–2605, 2008.