Cross-Embodiment Generalization via Behavior-Aligned Representations

Anonymous Author(s)
Affiliation
Address
email

Abstract: Recent progress in large-scale imitation learning for robot manipulation has been driven by leveraging datasets across a wide range of robot embodiments. However, while prior work on cross-embodiment learning has shown signs of positive transfer, achieving significant transfer is often still challenging. In this work, we propose using behavior-aligned representations (e.g., object bounding boxes, language motions, 2D end effector traces) as a scalable means of unifying diverse data across embodiments to enhance transfer. Our approach, BARX, incorporates these representations by simply training vision-language-action (VLA) models to predict them in tandem with actions as a single sequence. Our experiments show that BARX can significantly improve transfer from diverse cross-embodiment datasets to new embodiments. When learning from such datasets in simulation, BARX improves success rate by 13.8% when evaluated zero-shot on unseen embodiments, and by 16.0% when adapting with limited data. BARX is also able to enhance sim-to-real cross-embodiment transfer, improving task completion progress of real robot policies pre-trained on simulation data by 28.3%.

Keywords: Cross-Embodiment Learning, Imitation Learning, Manipulation

1 Introduction

Modern machine learning has shown that general-purpose models trained on large, diverse datasets outperform specialized models trained on narrow, domain-specific data [1–3]. This has inspired researchers in robot learning to scale robot datasets collected on a single robot platform [4, 5], with the hope of achieving more robust and capable policies via imitation learning. However, to scale our datasets even further, it is important to learn cross-embodiment policies that can transfer skills and demonstrate robustness across embodiments and tasks [6–8]. Furthermore, learning cross-embodiment policies has the potential to aid transfer to entirely new embodiments, which is useful due to the constantly evolving nature of robot platforms.

Despite the promise of large-scale cross-embodiment learning, prior work has seen mixed results in achieving positive transfer across embodiments. For example, while there has been evidence that cross-embodiment data can improve various axes of generalization [9], or transfer skills from one embodiment to another [6], often times performance does not significantly exceed the performance when training only on data specific to the target embodiment [6, 8]. This is most likely due to significant variation between policy inputs (e.g., image observations) and outputs (e.g., action spaces) across embodiments. Prior works have improved cross-embodiment transfer through explicitly aligning observations [10–12] or action spaces [10, 13, 14]. However, these works have various limitations, such as requiring significant manual effort (e.g., aligning camera poses), or knowledge of the deployment-time robot during training, which can make scaling to large datasets challenging. In order to better promote cross-embodiment transfer with large-scale datasets, we need methods for aligning heterogeneous datasets that are more scalable and general.

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

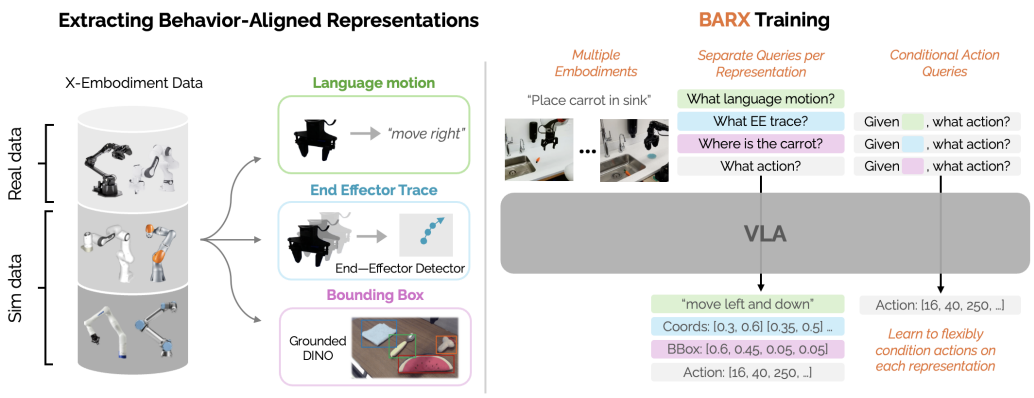

Figure 1: **Left**: We consider cross-embodiment datasets of both simulated and real-world robots. We then extract behavior-aligned representations: *language motions* via heuristics using end effector pose deltas, *end effector* traces via segmentation models, and *bounding boxes* via Grounding DINO. **Right**: BARX trains VLAs to predict actions with or without first predicting representations. This enhances transfer from cross-embodiment datasets, without needing representations during inference.

Our insight is that while observation and action spaces are often difficult to explicitly align, we can instead *implicitly* align them. In particular, we can connect data from different embodiments through *behavior-aligned* representations – representations that contain information relevant for action prediction – such as language descriptions [15] or 2D visual traces of robot motion [16, 17]. While these representations have shown utility for robot learning, their role for cross-embodiment learning has not been studied in detail. We hypothesize these representations can offer an additional space for policy reasoning that is invariant across embodiments, enabling a more scalable transfer mechanism.

In this work, we propose **B**ehavior-**A**ligned **R**epresentations for **X**-Embodiment Learning (BARX), a simple and scalable approach for improving cross-embodiment learning for robot manipulation. We train vision-language-action (VLA) models to predict multiple behavior-aligned representations (object bounding boxes, language motions, and 2D end effector traces), to predict low-level actions conditioned on these representations, and to directly predict just actions. At inference time, the policy is able to either first predict these representations and then actions, or predict actions directly.

To evaluate our approach, we develop **RoboCasa-X**, a simulation-based manipulation benchmark for assessing transfer with diverse, cross-embodiment data to new embodiments. On this benchmark, BARX improves success rate by 13.8% on zero-shot generalization to new embodiments, and 16.0% when adapting to new embodiments with limited data. We conduct ablations to support our choice of representations in BARX, and show that predicting representations during inference are not necessary for improved performance. We additionally evaluate on real-world robots, where we leverage data from **RoboCasa-X** for sim-to-real cross-embodiment transfer. We find that BARX improves task completion progress of real robot policies pre-trained with simulation data by 28.3%. See our website for more results and videos: `https://bar-x-anon.github.io`.

2 Related Work

In this section, we review prior work on cross-embodiment learning, vision-language-action models, and using auxiliary objectives and representations for robot learning.

Cross-Embodiment Learning. Cross-embodiment learning has emerged as a popular paradigm for scaling data for robot manipulation. However, differences between embodiments in their observations and actions can pose a significant challenge for transfer. Some efforts to leverage cross-embodiment data train a single model with minimal data alignment, which can achieve some degree of positive transfer [6, 18–21]. Others utilize a shared model backbone with specialized action heads or controllers for different embodiments [8, 10]. Other methods align observation spaces, such as by using wrist-mounted cameras or segmentation masks [10–13]. However, these approaches can

be difficult to scale. In this work, rather than explicitly modifying observation or action spaces, we use *behavior-aligned* representations (e.g., language motions, end effector traces), which we hypothesize can also help unify cross-embodiment data, but in a more scalable manner.

Vision-Language-Action Models. Vision-language-action (VLA) models have become an attractive policy class for generalist robot manipulation [19–25]. These policies involve fine-tuning vision-language models to predict robot actions, such that the policy can inherit knowledge from internet-scale vision-language data to enhance generalization. Their high-capacity architectures also allow them to effectively ingest large and heterogeneous robot datasets. In our work, we use VLAs for our policies due to these appealing qualities. Furthermore, because VLAs can also be trained to predict text [15, 26–28], we use this as a simple way of incorporating behavior-aligned representations.

Auxiliary Representations for Robot Learning. Several works have used auxiliary training objectives to improve policy performance and robustness. Co-training VLA policies on visual question answering (VQA) data has been shown to benefit generalization to novel task semantics and scene visuals [22, 25, 29]. Other works use intermediate policy representations to help guide action prediction. These representations include keypoints [30], end effector traces [16, 17, 31] and poses [27], language motion descriptions [15], goal images [32–34], and object bounding boxes [21, 25, 35]. These representations can also be combined in a chain-of-thought manner [28, 36]. While these works demonstrate the general utility of these representations, we focus on studying their effect on cross-embodiment transfer. Some works have used these representations when learning with cross-embodiment data [17, 20], but our work considers the contributions of multiple representations, and more carefully analyzes their effect on cross-embodiment transfer with diverse tasks.

3 Behavior-Aligned Representations for X-Embodiment Learning

In this section, we first overview our framework for learning policies with behavior-aligned representations. Next, we discuss our instantiation of this framework, including the specific representations we use, our approaches to scalably extract labels for learning them, and other details.

3.1 BARX Formalization

We consider the language-conditioned cross-embodiment imitation learning (IL) setting. We define a robot embodiment space R where $r \in \mathcal{R}$ designates a specific embodiment, which has an associated observation space $\mathcal{O}^r$ and action space $\mathcal{A}^r$. These spaces may have some relation/alignment across robots, but we do not assume this. Our goal is to learn a policy $\pi_\theta(a \mid o, l)$ that maps observations $o \in \mathcal{O}^r$ and language instructions $l \in \mathcal{L}$ to actions $a \in \mathcal{A}^r$. We assume expert data of the form $\mathcal{D} = \{(o_i^r, a_i^r, l_i)\}_{i=1}^N$, where r designates which embodiment a given observation or action is from. In behavior cloning, the policy is trained to minimize a supervised loss to imitate expert actions:

$$\mathcal{L}_{\text{BC}}(\theta) = \mathbb{E}_{(o,a,l)\sim\mathcal{D}}\left[\ell(\pi_\theta(\cdot \mid o, l), a)\right].$$

In addition, we assume access to a set of *behavior-aligned* representations. These can be any quantity that can be computed from $\mathcal{D}$, contains information that can be useful for predicting optimal behavior, and is invariant to the robot embodiments in $\mathcal{R}$, regardless of observation or action spaces. Formally, each observation o_i is annotated with a tuple of representations $z_i = (z_i^{(1)}, \ldots, z_i^{(K)})$, where each $z_i^{(k)} \in \mathcal{Z}^{(k)}$ belongs to a distinct representation space.

To utilize these representations, we modify the policy π_θ to condition on subsets of them when predicting actions. We also train π_θ to predict these representations. At each training step, we sample a subset of representations, which we denote $\tilde{z} \subseteq \{z^{(1)}, \ldots, z^{(K)}\}$, using some pre-defined representation distribution p_{rep}, i.e., $\tilde{z}_i \sim p_{\text{rep}}(\tilde{z})$. This encourages the policy to adaptively rely on different behavior-aligned representations when predicting actions. Our total loss combines the behavior cloning objective with auxiliary supervision on the behavior-aligned representations:

$$\mathcal{L}_{\text{total}}(\theta) = \mathbb{E}_{(o,z,a,l)\sim\mathcal{D},\tilde{z}\sim p_{\text{rep}}(z)}\left[\ell(\pi_\theta(\cdot \mid o, l, \tilde{z}), a) + \sum_{k=1}^{K} \ell_{\text{rep}}^{(k)}(\pi_\theta(\cdot \mid o, l), z^{(k)})\right].$$

Here, $\ell_{\text{rep}}^{(k)}$ corresponds to the loss for predicting the representations. In practice, we instantiate π_θ as a VLA. Then, our formulation amounts to autoregressively predicting a subset of behavior-aligned representations as text, and then actions, as a single sequence. The model receives different text prompts depending on which subset of representations it should predict. This conditioning scheme is similar to prior work [15, 28], but allows for variation on which behavior-aligned representations are used with the policy. For instance, during inference the policy can be prompted to predict any set of representations from the distribution $p_{\text{rep}}(\tilde{z})$, rather than all representations.

We note that this method can be applied to any dataset of language-annotated expert demonstrations, but we seek to explore its utility specifically for cross-embodiment learning. We hypothesize that because these representations are invariant across embodiments and are useful for action prediction, they can help facilitate cross-embodiment transfer through implicit alignment.

3.2 Choice of Representations

Here, we describe the specific behavior-aligned representations we incorporate in BARX and our methods for annotating them. Drawing from prior work, we choose a set of representations that are simple to annotate, informative for predicting actions, and are naturally invariant across various embodiments. We note that one may choose a given set of representations depending on whether they are invariant to the robots present in a given setting. We provide more details in the Appendix.

Bounding Boxes. We use bounding boxes for objects of interest that the robot must manipulate [25, 28, 37]. To label our real-world datasets, we use an off-the-shelf pipeline from prior work [28] consisting of VLM scene descriptions and Grounding DINO [38].

End Effector Traces. We use traces of the robot end effector position during the future motion it should execute for its task [16, 17]. These traces consist of sequences of future 2D positions of the where the end effector is in the image observation frame. We use a pre-trained model from prior work [17] to detect these positions from image observations.

Language Motions: We use language descriptions of the motion corresponding the action the robot should execute next, such as "move left and down". We obtain this representation through a pipeline similar to prior work [15], based on thresholding changes in robot proprioceptive state.

3.3 Additional Details

We use the MiniVLA architecture for our policies [23]. We use single third-person camera views as our observation space for all robots. For simplicity, we choose our training distribution of representation $p_{\text{rep}}(\tilde{z})$ to be uniform over each singleton set of one representations $\{z^{(i)}\}$, and the empty set of no representations. In other words, we train our VLA to either predict a single representation and then actions, or actions directly. We found this was enough for our models to effectively incorporate our representations, while also allowing for inference without needing to predict representations. This can be desirable because predicting representations can add considerable latency during inference [28]. We provide additional training details in the Appendix.

4 Experiments

In our experiments, we aim to address the following questions:

- **Q1 (New Embodiments)**: Can BARX help transfer to new embodiments with limited/no data?
- **Q2 (Cross-Embodiment Scaling)**: Is BARX more helpful with larger, cross-embodiment datasets than smaller, single-embodiment datasets?
- **Q3 (Rep Training)**: Which behavior-aligned representations are important for training in BARX?
- **Q4 (Rep Inference)**: Does predicting and conditioning on behavior-aligned representations matter during inference for BARX?

PnP Counter to Sink

PnP Sink to Counter

Figure 2: Example initial scenes from our RoboCasa tasks with the Panda robot. These tasks capture a significant degree of variation, including different kitchen layouts, textures, object types, and object poses.

4.1 Simulation Experiments

4.1.1 Benchmark Design

While prior work has studied cross-embodiment learning for manipulation using diverse data in the real world [6, 8, 18, 19], to our knowledge there are no existing simulation benchmarks for this setting. Therefore, we design a new simulation benchmark for cross-embodiment learning called **RoboCasa-X**, based on RoboCasa [39], a platform that supports realistic and diverse kitchen tasks and scenes. We consider variations of the tasks *PnP Sink to Counter* and *PnP Counter to Sink*. These tasks capture a large amount of variation, including different kitchen layouts, textures, object types, and object poses. Examples of this task variation are shown in Fig. 2. This variation makes these tasks challenging and suitable for studying transfer with diverse data.

To create diverse, cross-embodiment prior datasets, we collect 50 human demonstrations for each task, and then use MimicGen [40] with this data to produce additional data for three robot embodiments: Kinova3, UR5e, and IIWA. To simulate realistic diversity and promote sim-to-real transfer (as done later in Section 4.2), we randomize the camera pose in the generated data. We treat mixtures of this data for all tasks and embodiments as our prior datasets. We vary dataset size with either 300 demonstrations per task/embodiment combination (1800 total) or 1000 per task/embodiment (9000 total). We refer to these datasets as **X-Prior-300** and **X-Prior-1000**, respectively.

We consider two additional embodiments designated as target robots: Panda and Jaco. For each target robot, we collect 50 human demonstrations per task with a single camera view. In our benchmark, we measure cross-embodiment transfer by aiming to train policies on the prior datasets that perform well on the target robots, either zero-shot, or with the addition of the target robot data.

Experiment Details. Because we have access to privileged information in simulation, we obtain ground truth labels for bounding boxes and end effector traces, rather than using the annotation methods described in Section 4.2. Unless otherwise stated, we evaluate BARX policies by only predicting actions. For each evaluation, we conduct 100 rollouts with a fixed set of scene conditions. We evaluate using the same camera view for all robots (same as in the target robot data). All of our simulated robots share the same action space (delta Cartesian end effector pose control) and controller. We provide more experiment details in the Appendix.

4.1.2 Experiments

Zero-Shot Generalization. First, we evaluate if BARX can enhance zero-shot generalization to new robot embodiments. We compare two models: one trained only to predict actions (No Reps), and one trained with BARX. We pre-train models on **X-Prior-1000**, and evaluate them on the unseen target robots without any additional data. We also evaluate on the embodiments seen during training (Kinova3, UR5e, and IIWA).

In our results in Fig. 3, we find that although both models degrade on the unseen robots, BARX is able to achieve much better zero-shot generalization, improving overall success rate by 13.8% (**Q1**). BARX is also able to retain a greater fraction of its in-distribution performance, suggesting that its improved zero-shot generalization is not solely due to better performance on seen robots.

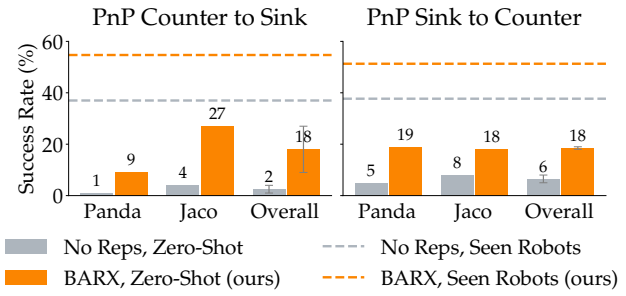

Figure 3: BARX significantly improves zero-shot generalization to unseen robot embodiments.

Adaptation to New Embodiments. Next, we consider adaptation to new robots with limited data. We pre-train on **X-Prior-300** and **X-Prior-1000**, and then co-fine-tune separate models for each target robot using its data for both tasks. We also train models from scratch on target robot data.

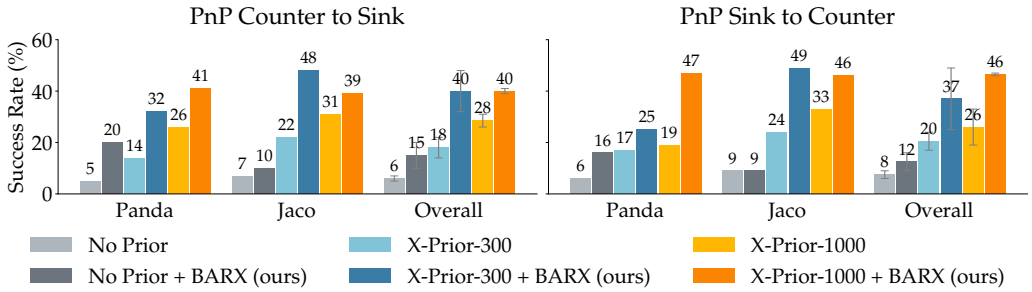

Figure 4: We adapt models trained on diverse cross-embodiment prior data to new target robots. BARX significantly improves transfer from prior datasets, scales with prior dataset size, and is more helpful with cross-embodiment prior data than when learning from only target robot data.

In Fig. 4, we find that BARX significantly improves transfer from cross-embodiment data, improving overall success rate by 19.3% when using **X-Prior-300**, and 16.0% when scaling up to **X-Prior-1000** (**Q1**). There is mild positive scaling of BARX with more prior data (+4.8%). The benefit of BARX decreases slightly with more prior data, indicating that larger datasets can make cross-embodiment learning easier without representations, although BARX still significantly helps.

We also evaluate BARX when only learning from target robot data (dark gray bars in Fig. 4). We find that it also helps in this setting, but by a reduced margin of 7.0%, suggesting that BARX is more effective with larger, cross-embodiment datasets (**Q2**). We hypothesize BARX is particularly beneficial in the cross-embodiment setting because it can help unify data across different embodiments with common representations. We note that these results are when all robots share the same action space, suggesting that BARX has benefits beyond action alignment.

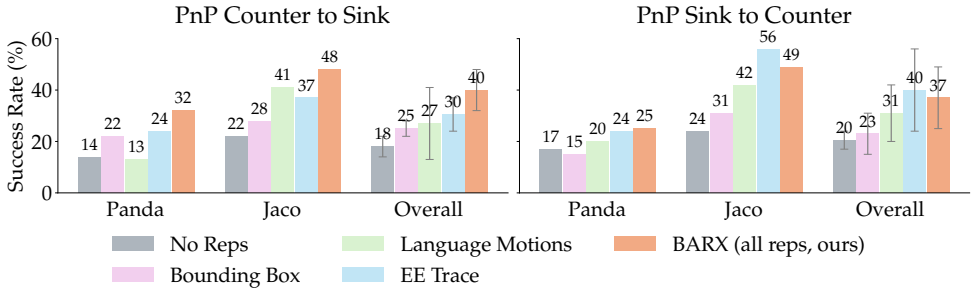

Figure 5: We compare training using each representation in BARX separately, training with all representations, and training with none. Each representation improves over using none, with 2D end-effector traces helping the most, but using all representations performs the best overall.

Ablating Training Representations. Here, we consider the contribution of each individual representation used in BARX. We pre-train additional models on **X-Prior-300** using each representation

in isolation, and then co-fine-tune them using the target robot data with the same representation. We present our results in Fig. 5. We find that training on each representation in isolation generally improves over none, supporting each of their use. End effector traces generally help the most, sometimes even outperforming all representations together. However, combining all representations does the best overall, motivating their joint use in BARX rather than only end effector traces (**Q3**).

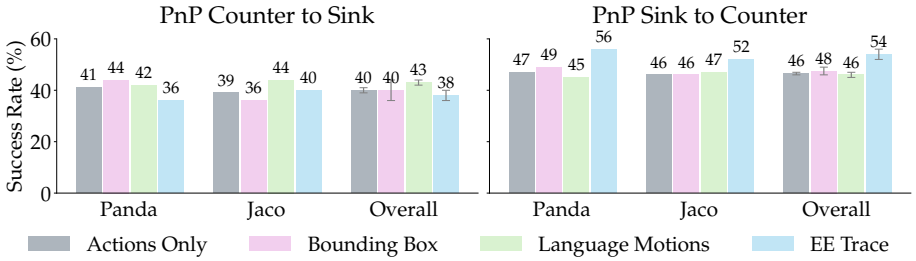

Figure 6: We compare inference of BARX models by either only predicting actions, or first predicting a representation and then actions. None of the representations significantly affect performance.

Predicting Representations. In our results so far, we evaluate BARX models by directly predicting actions during inference, rather than first predicting representations and then the action [15, 17, 27, 28]. We do this because we found that conditioning actions on representations did not significantly affect performance. We show this in Fig. 6, where we evaluate the adapted BARX models pre-trained on **X-Prior-1000** using each possible intermediate representation as policy conditioning during inference (i.e., actions only, or predicting one representation and then actions). We find that each inference method performs comparably, with the possible exception of end-effector traces, which improves by 7.5% on *PnP Sink to Counter*, but does not significantly affect *PnP Counter to Sink*. These results suggest that BARX primarily helps in an implicit manner (i.e., learning good internal representations during training) and can improve policies without explicitly predicting representations, which is desirable to avoid additional inference latency (**Q4**).

4.2 Real-World Experiments

To see if BARX also enhances real-world cross-embodiment learning, we consider two real target embodiments: the Franka Research 3 (FR3) and ViperX 300 S. We evaluate our method on variations of the tasks in **RoboCasa-X**. We adapt our models pre-trained on **X-Prior-1000** using 50 human demonstrations per task, similar to in Section 4.1.2. Prior work has shown that RoboCasa data can help real-world tasks [41]. Our setting represents a similar paradigm for sim-to-real transfer, but also considers cross-embodiment transfer, as the real target robots are not in the simulation data. We loosely align the camera view and environment layout from simulation, but there remain significant domain gaps. We compare examples of our real tasks to their simulation counterparts in Fig. 7.

Action Alignment. Besides differences in visual observations, there also now exist significant differences in the action spaces between the source and target robots, unlike our previous simulation experiments. Although the simulated and real robots all use delta end effector pose control, the control stacks for each robot differ significantly. For example, they differ in control frequency: the ViperX uses 5 Hz, the FR3 uses 10 Hz, and the simulated robots use 20 Hz. Additionally, data was collected on the ViperX with blocking control, while data collected for **RoboCasa-X** and the FR3 used non-blocking control. This makes transfer in this setting especially challenging.

Experimental Setup. For practicality of evaluation, we modify the real-world instantiations of the **RoboCasa-X** tasks by reducing some variation. For instance, we only use one kitchen, and reduce the number of object types to four for *PnP Counter to Sink* and one for *PnP Sink to Counter*. Like in our simulation experiments, we train models with and without cross-embodiment prior data (**X-Prior-1000**), and with and without BARX. Unlike our simulation experiments, for our models with prior data, we train a single model on both target embodiments, to allow for transfer between each real embodiment. For each embodiment/task pair, we evaluate for 10-30 rollouts. We report task completion progress of each model. We provide more experiment details in the Appendix.

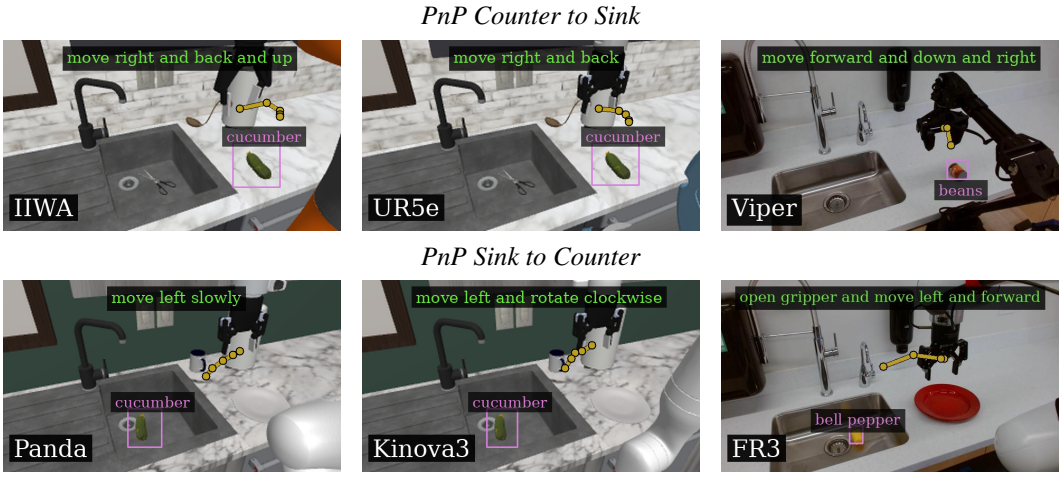

Figure 7: We show examples of our real tasks (right) in comparison to their sim counterparts (left, middle). We also show actual predictions from BARX models, which demonstrate similar predicted representations across different embodiments in unseen scenes.

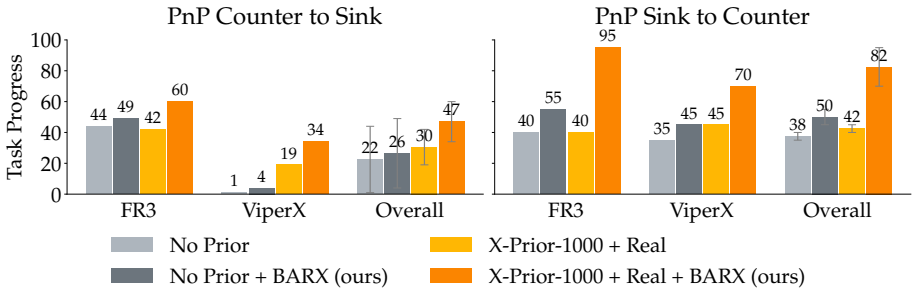

Figure 8: We adapt models trained on cross-embodiment prior data from simulation to real-world target robots. BARX significantly improves transfer from prior datasets despite gaps in visual observations and action space.

Results. In Fig. 8, we see that BARX significantly improves task progress with cross-embodiment data (orange bars), with an overall increase of 28.5% (**Q1**). Furthermore, it helps more with cross-embodiment data than single embodiment (+8.3%, gray bars), suggesting that BARX scales with embodiment and data diversity (**Q2**). When training without BARX, cross-embodiment data often does not provide any gain over the single-embodiment model, with an overall improvement of only 6.5%. However, with BARX the benefit from cross-embodiment data becomes much larger (+26.5%), suggesting that it helps leverage cross-embodiment data through implicit alignment. We suspect the benefit of BARX is more pronounced in our real experiments due to the larger domain gap in observations and actions between the simulated and real robots, which makes cross-embodiment transfer without representations more challenging. In Fig. 7, we show examples of how BARX models can predict similar representations across different embodiments in unseen scenes.

5 Conclusion

We introduce BARX, a method for cross-embodiment learning that leverages behavior-aligned representations to more effectively learn from cross-embodiment data. Behavior-aligned representations can easily be labeled with off-the-shelf models or simple transforms in the actions. We demonstrate–both in simulation and the real-world–that BARX consistently boosts performance in both single and cross-embodiment settings. Furthermore, BARX yields much larger gains in the cross-embodiment setting, suggesting that these representations become more important for learning as the diversity of embodiments scales. We also find that using multiple behavior-aligned representations during training outperforms using any one representation. During inference, we find that direct action prediction performs as well as conditioning on any one intermediate representation, enabling us to reduce inference time compared to prior work. We believe that BARX is an important step towards strong positive transfer in cross-embodiment learning.

Limitations. BARX demonstrates the importance of using multiple behavior-aligned representations when learning from cross-embodiment data. However, it can be time consuming to tune and run labeling methods for each representation on existing robotics datasets, especially as they scale in size and embodiment diversity. Future work might develop more robust labeling techniques. Additionally, the behavior-aligned representations used in BARX are not exhaustive, and these representations favor object-centric manipulation tasks. We hypothesize that other behavior-aligned representations can be incorporated that may be more beneficial depending on the task. We hope our framework serves as a strong starting point for exploring more behavior-aligned representations to facilitate cross-embodiment generalization.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[4] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[5] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[6] A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. 2024.

[7] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. ViNT: A foundation model for visual navigation. In *7th Annual Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2306.14846.

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

[9] J. Gao, S. Belkhale, S. Dasari, A. Balakrishna, D. Shah, and D. Sadigh. A taxonomy for evaluating generalist robot policies. 2025.

[10] J. H. Yang, D. Sadigh, and C. Finn. Polybot: Training One Policy Across Robots While Embracing Variability. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=HEIRj51lcS.

[11] 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 *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[12] M. Lepert, R. Doshi, and J. Bohg. SHADOW: Leveraging Segmentation Masks for Cross-Embodiment Policy Transfer. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=MyyZZAPgpy.

[13] 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. URL https://arxiv.org/abs/2402.19432.

[14] 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, 2025. URL https://arxiv.org/abs/2501.10105.

[15] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh. Rt-h: Action hierarchies using language, 2024. URL https://arxiv.org/abs/2403.01823.

[16] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, P. Sundaresan, P. Xu, H. Su, K. Hausman, C. Finn, Q. Vuong, and T. Xiao. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches, 2023.

[17] D. Niu, Y. Sharma, G. Biamby, J. Quenum, Y. Bai, B. Shi, T. Darrell, and R. Herzig. Llarva: Vision-action instruction tuning enhances robot learning, 2024.

[18] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. 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 *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

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

[20] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[21] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[22] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[23] S. Belkhale and D. Sadigh. Minivla: A better vla with a smaller footprint, 2024. URL https://github.com/Stanford-ILIAD/openvla-mini.

[24] A. Szot, B. Mazoure, O. Attia, A. Timofeev, H. Agrawal, D. Hjelm, Z. Gan, Z. Kira, and A. Toshev. From multimodal llms to generalist embodied agents: Methods and lessons. *arXiv preprint arXiv:2412.08442*, 2024.

[25] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

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

[27] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation, 2024. URL https://arxiv.org/abs/2411.02704.

[28] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[29] J. Gao, S. Belkhale, S. Dasari, A. Balakrishna, D. Shah, and D. Sadigh. A taxonomy for evaluating generalist robot policies. *arXiv preprint arXiv:2503.01238*, 2025.

[30] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg. KITE: Keypoint-conditioned policies for semantic manipulation. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=veGdf4L4Xz.

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

[32] K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

[33] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman, et al. Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches. In *8th Annual Conference on Robot Learning*, 2024.

[34] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, et al. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. *arXiv preprint arXiv:2503.22020*, 2025.

[35] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. URL https://arxiv.org/abs/2205.06230.

[36] J. Clark, S. Mirchandani, D. Sadigh, and S. Belkhale. Action-free reasoning for policy generalization. *arXiv preprint arXiv:2502.03729*, 2025.

[37] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, C. Finn, and K. Hausman. Open-world object manipulation using pre-trained vision-language models, 2023. URL https://arxiv.org/abs/2303.00905.

[38] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

[39] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

[40] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox. MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations. In *7th Annual Conference on Robot Learning*, 2023.

[41] A. Maddukuri, Z. Jiang, L. Y. Chen, S. Nasiriany, Y. Xie, Y. Fang, W. Huang, Z. Wang, Z. Xu, N. Chernyadev, S. Reed, K. Goldberg, A. Mandlekar, L. Fan, and Y. Zhu. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, 2025.