Cross-Embodiment Transfer via Behavior-Aligned Representations

Ajay Sridhar*[1], Jensen Gao*[1], Jonathan Yang[1], Jean Mercat[2], Suneel Belkhale[1], Dorsa Sadigh[1]

arXiv:2607.27549v1 [cs.RO] 30 Jul 2026

***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, achieving significant cross-embodiment transfer is often still challenging. In this work, we study the role of using behavior-aligned representations (e.g., object bounding boxes, language motions, end-effector traces of robot motion) in vision-language-action (VLA) models to promote cross-embodiment transfer. We hypothesize that by possessing invariances across embodiments while being predictive of robot actions, these representations can help unify large-scale cross-embodiment data to enhance transfer. To assess our hypothesis, we develop a simulation-based benchmark designed to assess transfer with diverse cross-embodiment data to new embodiments. Using this benchmark, we compare different representations and ways of incorporating them. We identify that end-effector traces can be particularly beneficial for transfer, representations are generally more useful with larger prior datasets, and can be used to benefit from action-free data. We also demonstrate that they can enhance sim-to-real cross-embodiment transfer, improving task completion progress of real robot policies pre-trained on simulation data by 28%. We provide videos of our evaluations at our website https://ajaysridhar.com/barx/.**

I. 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 [4], [5], with the hope of achieving more robust and capable policies. However, data collection efforts are usually specific to individual robot platforms. To further scale robot learning and benefit from as much data as possible, it becomes important to learn cross-embodiment policies that can leverage data from a wide variety of robot platforms and transfer abilities across embodiments and tasks [6]–[8]. Furthermore, cross-embodiment policies have the potential to aid transfer to entirely new embodiments, which is useful due to the constantly evolving nature of robot hardware.

However, despite the promise of cross-embodiment learning, prior work has seen mixed results in achieving successful cross-embodiment transfer for manipulation. For example, while there has been evidence that cross-embodiment data can improve various generalization axes [9] or help transfer skills across embodiments [6], often times performance does not significantly exceed training only on data specific to the target embodiment [6], [8]. This is most likely due to substantial variation between policy inputs (e.g., image observations) and outputs (e.g., action spaces) across embodiments, and not enough data coverage to overcome this.

*Equal contribution. [1]Stanford University, [2]Toyota Research Institute. Correspondence to: ajaysri@stanford.edu.

Prior works have improved cross-embodiment transfer through explicitly aligning observations [10]–[13] or action spaces [10], [14], [15]. 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 them challenging to scale. In order to better promote cross-embodiment transfer at scale, we need methods for aligning heterogeneous datasets that are more scalable and general.

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 [16] or 2D visual traces of robot motion [17], [18]. While prior works have shown the utility of these representations for robot learning, their role in promoting cross-embodiment transfer has not been studied in detail. We hypothesize these representations can offer an additional space for policy reasoning that is invariant across embodiments, enabling scalable transfer.

To test this hypothesis, we conduct an empirical study on how different behavior-aligned representations affect cross-embodiment transfer. In particular, we consider training vision-language-action (VLA) models to predict these representations jointly with actions, and vary different representations and ways of incorporating them. To facilitate our study, we develop **RoboCasa-X**, a simulated benchmark based on prior work [19], which involves pre-training policies using diverse, cross-embodiment data, and then assessing transfer to new embodiments with limited additional data.

Using our benchmark, we find that a variety of representations can enhance transfer, but traces of end-effector motion are generally the most impactful out of those we consider. Furthermore, we show that representations scale well with larger prior datasets, predicting representations during inference is not necessary for transfer, and representations can facilitate transfer from action-free datasets. Lastly, we evaluate on real robots, where we use data from **RoboCasa-X** for sim-to-real cross-embodiment transfer. We find that representations can improve task progress of policies pre-trained with simulation data by 28%. We provide videos of our evaluations at our website https://ajaysridhar.com/barx/.

II. 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.

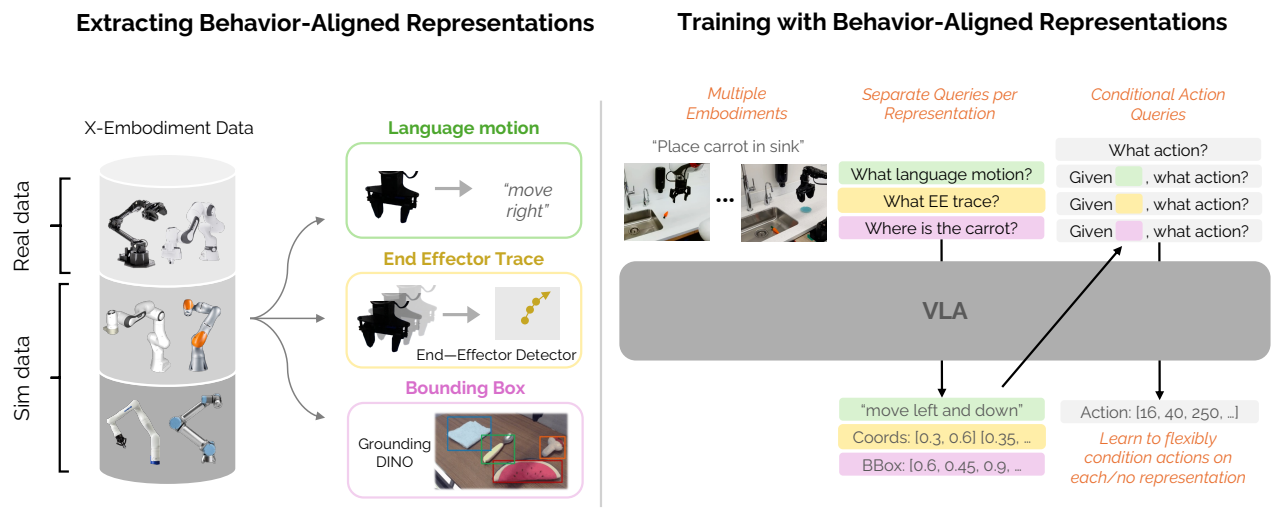

Fig. 1. **Left**: We consider cross-embodiment datasets of both simulated and real-world robots. We then extract behavior-aligned representations: *language motions* via heuristics and end effector pose deltas, *end effector* traces via segmentation models, and *bounding boxes* via Grounding DINO. **Right**: We train VLAs with different choices of representations and ways of incorporating them, and study how this impacts transfer from cross-embodiment datasets.

Cross-Embodiment Learning. Cross-embodiment learning has emerged as a popular paradigm for scaling robot manipulation data. 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 transfer [6], [20]–[23]. Others use a shared model backbone with specialized action heads or controllers for different embodiments [8], [10]. Other works align observations, such as by using wrist cameras, inpainting, or segmentation masks [10]–[14]. However, these approaches can be difficult to scale. Rather than explicitly aligning observations or actions, we study using *behavior-aligned* representations (e.g., language motions, end-effector traces), which we hypothesize can also help unify cross-embodiment data, but in a more implicit, scalable manner.

Vision-Language-Action Models. Vision-language-action (VLA) models have become an attractive policy class for generalist robot manipulation [21]–[27]. These policies involve fine-tuning vision-language models (VLMs) 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 as our policies due to these appealing qualities. Furthermore, because VLAs can also be trained to predict text [16], [28]–[30], 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 [9], [24], [27]. Other works use intermediate policy representations to guide action prediction. These representations include image points [31], [32], end effector traces [17], [18], [33] and poses [29], language motion descriptions [16], goal images [34]–[36], object bounding boxes [23], [27], [37], and segmentation masks [38]. These representations can also be combined in a chain-of-thought manner [30], [39], [40]. 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 with cross-embodiment data [18], [27], [32], but our work analyzes different representations, ways of incorporating them, and more carefully analyzes their effect on cross-embodiment transfer with diverse tasks.

III. Behavior-Aligned Representations for Cross-Embodiment Transfer

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 consider, our annotation process for obtaining them, and other details.

A. Formalization

We consider the language-conditioned cross-embodiment imitation learning (IL) setting. We define a robot embodiment space $\mathcal{R}$ 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 function ℓ 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 useful for predicting optimal behavior, and has invariance to the embodiments in $\mathcal{R}$, meaning that different embodiments should have similar representations for the same task. 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 one of K different representation spaces.

To leverage 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. For each training example, we sample a subset of representations, which we denote $\tilde{z} \subseteq \{z^{(1)}, \ldots, z^{(K)}\}$, using some predefined representation distribution p_{rep}, i.e., $\tilde{z}_i \sim p_{\text{rep}}(\tilde{z})$. Our total loss combines the behavior cloning objective with auxiliary supervision on the representations:

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

$$\ell_{\text{aux}} = \sum_{k=1}^{K} \lambda_k \ell_{\text{rep}}^{(k)}(\pi_\theta(\cdot \mid o, l), z^{(k)}).$$

Here, $\ell_{\text{rep}}^{(k)}$ is the loss for predicting the kth representation, weighted by λ_k. 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 to indicate which subset of representations it should predict. This conditioning scheme is similar to prior works that predict representations as forms of embodied reasoning before predicting actions [16], [30], [40].

We note that this approach can be applied to any dataset of language-annotated demonstrations, but we consider it specifically for cross-embodiment learning. We hypothesize that because these representations have invariance across embodiments and are useful for action prediction, they can aid cross-embodiment transfer through implicit alignment.

B. Choice of Representations

We describe the specific representations we study 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 have inherent invariance across different embodiments. We note that the best choice of representations may vary, depending on which have invariance to the robot platforms for a particular setting.

Bounding Boxes. We use bounding boxes for the current image observation of objects that the robot should manipulate [27], [30], [41]. To label our real-world datasets, we use an off-the-shelf pipeline from prior work [30] consisting of VLM scene descriptions and Grounding DINO [42].

Language Motions: We use language descriptions of the motion corresponding to the action the robot should execute next, such as "move left and down". We obtain this representation through a pipeline similar to prior work [16], based on thresholding changes in robot proprioceptive state.

End Effector Traces. We use traces of the robot end effector position during the future motion it should execute for its task [17], [18]. These traces consist of sequences of future 2D positions of where the end effector is in the image observation frame. We use a pre-trained model from prior work [18] to detect these positions from image observations.

C. Implementation Details

We use the MiniVLA architecture [25] for our policies, starting from its pre-trained VLM, but without any robot pre-training. We use single third-person camera views as our observation space. We set all loss weights $\lambda_k = 1$.

IV. Simulation Experiments

A. Benchmark Design

Although prior work has studied cross-embodiment learning for manipulation using diverse data in the real world [6], [8], [20], [21], to our knowledge, simulation benchmarks for this setting have not previously been considered. Therefore, we design a new simulation benchmark for cross-embodiment learning called **RoboCasa-X**, based on RoboCasa [19], a platform that supports realistic and diverse scenes and tasks. We consider variations of three tasks from RoboCasa (*PnP Counter to Sink*, *PnP Sink to Counter*, *Turn On Sink Faucet*), as well as a new task (*Flip Mug Upright*). These tasks capture a large amount of variation, including different kitchen layouts, textures, object types, and poses, making them challenging and suitable for studying transfer.

To create large-scale prior datasets, we use MimicGen [43] to generate data for three source robot embodiments: IIWA, Kinova3, and UR5e. The Kinova3 and UR5e use the Robotiq 2F-85 gripper, while the IIWA uses the Robotiq 2F-140. To simulate realistic diversity and promote sim-to-real transfer (as done later in Section V), we randomize the camera pose. We vary dataset size with 300 or 1000 demonstrations per task/embodiment, for a total of 900 or 3000 demonstrations per task. We refer to these datasets as **XP-900** and **XP-3K** (short for Cross Prior), respectively.

We use three additional embodiments as target robots: Panda and Jaco with the Robotiq 2F-85, and Panda with its original Franka Hand gripper (designated as Panda-OG). For each target robot, we collect 50 human demonstrations per task. We measure cross-embodiment transfer by pre-training on the prior robot datasets, and then transferring to each target robot by co-fine-tuning on the target demonstrations and prior data. Because 50 target demonstrations are not enough to capture the diversity in each task, policy performance is reflective of cross-embodiment transfer from the prior data.

In Fig. 2, we visualize initial observations for the source embodiments in the pre-training data for the task *PnP Counter to Sink*, and compare this to the initial observations for the target robots. Our other tasks are instantiated in similar scenes with similar types of variation.

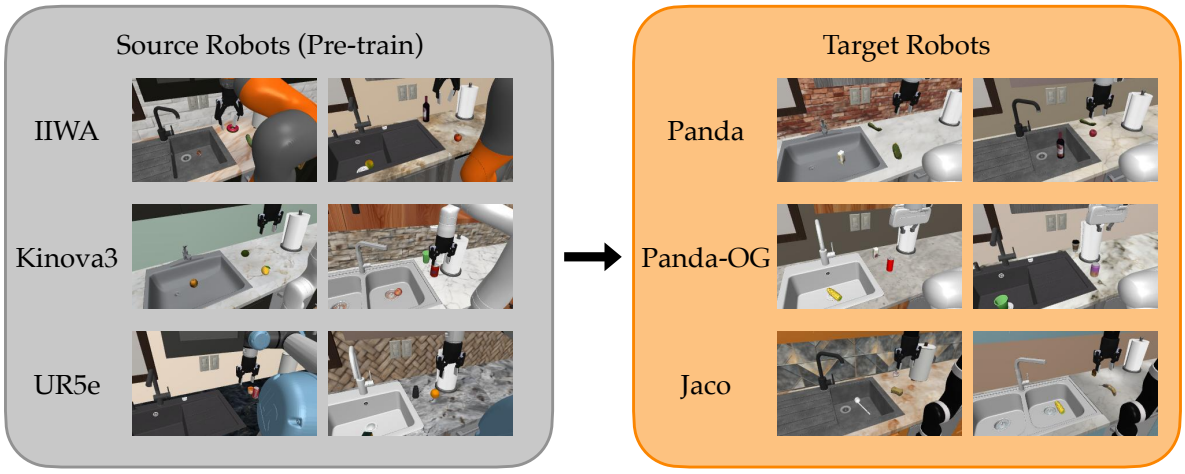

Fig. 2. We pre-train policies on diverse cross-embodiment data mixtures from three source robots (left), and then transfer these policies to different target robots using limited data (right). We show initial observations in our data for the task *PnP Counter to Sink*. Our tasks capture significant variation in kitchen layouts, textures, object types, and object poses.

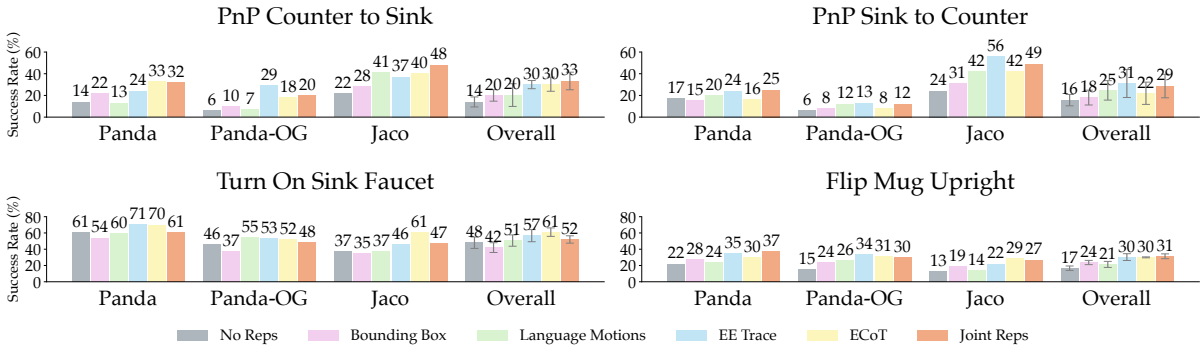

Fig. 3. We compare training with no representations, each representation separately, and combined using either **ECoT** or **Joint Reps**. Each representation improves transfer over using none, with end-effector traces helping the most. **Joint Reps** performs the best overall, except on Panda-OG, where end-effector traces alone are better.

Additional Details. Because we have access to privileged information in simulation, we obtain ground truth labels for bounding boxes and end effector traces, rather than the annotation methods described in Section III-B. For each evaluation, we conduct 100 rollouts with a fixed set of scene conditions. We report the highest success rate out of three checkpoints for each model. 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.

B. Experiments

In our experiments, we aim to address the following:

- **Q1 (Incorporating Representations)**: Which representations help most with transfer to new embodiments, and how should they be incorporated?
- **Q2 (Cross-Embodiment Scaling)**: Are representations more helpful with larger, cross-embodiment datasets than smaller, single-embodiment datasets?
- **Q3 (Representation Inference)**: Does predicting and conditioning on representations matter during inference?
- **Q4 (Action-Free Transfer)**: Can representations help with cross-embodiment transfer from action-free data?

Incorporating Representations. First, we analyze how each representation contributes to cross-embodiment transfer, and how to incorporate them during training. We consider the following approaches, which amount to different choices of representation distributions $p_{\text{rep}}(\tilde{z})$ in our framework:

1) **No Reps**: We train the model to only predict actions, without any representations.
2) **Single Rep**: We train the model to either predict a single representation and then actions, or only actions. We train separate models for each representation.
3) **ECoT** [30]: We train the model to either predict each representation and then actions as a sequential chain (bounding boxes $\rightarrow$ end-effector trace $\rightarrow$ language motion $\rightarrow$ actions), or only actions.
4) **Joint Reps**: Similar to **Single Rep**, except we train a single model with all representations (e.g., the model is trained to either predict one of multiple representations and then actions, or only actions).

For each approach, we use it to pre-train a model on **XP-900**. Then, separately for each target robot, we co-fine-tune it with the addition of target robot data. We do this procedure with separate models for each task, except for the two *PnP*

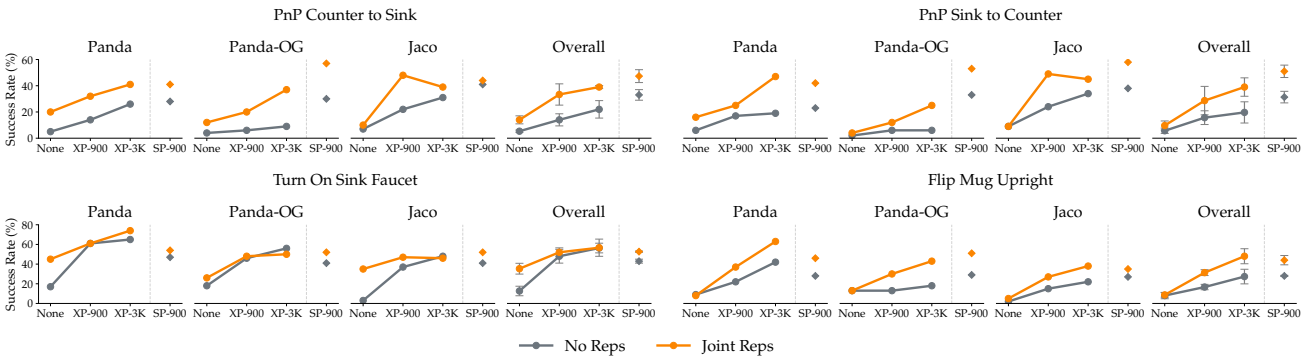

Fig. 4. We compare the effect of behavior-aligned representations when using different scales of prior data. Representations are generally more impactful when scaling up prior data (with the exception for the task *Turn On Sink Faucet*).

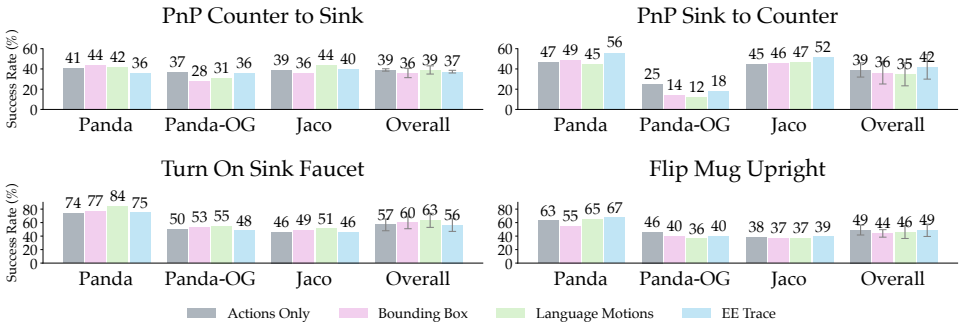

Fig. 5. We compare inference of **Joint Reps** models by either only predicting actions, or first predicting a representation and then actions. Predicting representations can help in some situations, but the impact is usually not substantial.

tasks, where we use the same model for both tasks. During inference for each model, we only predict actions without representations (which each model was also trained to do), as we found this to significantly speed up inference without significantly affecting performance.

In our results in Fig. 3, we find that using each representation individually improves transfer overall compared to using none. End-effector traces are generally the most helpful, followed by language motions, and lastly bounding boxes. When combining representations using either **ECoT** or **Joint Reps**, performance generally improves over each one in isolation for Panda and Jaco with the Robotiq 2F-85.

However, combining representations does not perform as well as only using end-effector traces for Panda-OG, whose end-effector is unseen in the prior data (shown in Fig. 2). We hypothesize this is because traces are particularly important for aligning data with more varied end-effectors. More generally, this suggests that the best choice of representations may depend on the differences between embodiments.

Cross-Embodiment Scaling. Next, we investigate if behavior-aligned representations are more beneficial with larger cross-embodiment datasets than smaller single-embodiment datasets. To do this, we compare **No Reps** and **Joint Reps**. We train using the same pre-training and co-fine-tuning procedure as before, but vary the prior dataset as either **XP-900** and **XP-3K**. We also compare to training on the target robot data from scratch with no prior (**None**), and using a prior dataset created in the same manner as **XP-900**, but with the same robot as the target (**SP-900**).

In Fig. 4, we find that behavior-aligned representations help when training only on target robot data, either with no prior or the same embodiment prior, corroborating prior work [18], [40]. However, for all of the tasks but one (*Turn On Sink Faucet*), the overall improvement when training with no prior (+5%) is smaller than when using **XP-900** (+15%) or **XP-3K** (+19%). This suggests that representations scale effectively with larger cross-embodiment datasets.

On Panda and Jaco, training on **XP-3K** with representations often matches or even exceeds training on the same embodiment prior (**SP-900**), suggesting that the degree of transfer is strong. However, this is not the case with Panda-OG, where performance lags significantly behind training on the same embodiment prior, suggesting that there is still room for improvement with larger embodiment gaps.

Unlike the other tasks, for *Turn On Sink Faucet*, representations have a larger impact when training only on target robot data (+22%) than when using **XP-900** (+4%) or **XP-3K** (+1%). We hypothesize this is because *Turn On Sink Faucet* has less variation than our other tasks, which involve manipulating a variety of objects in different poses, while there is not as much variation in the poses and types of faucets. Therefore, performance is saturated more easily, and prior data is not as critical.

Representation Inference. So far, we have evaluated all models by directly predicting actions, rather than first predicting representations and then actions. Concurrent

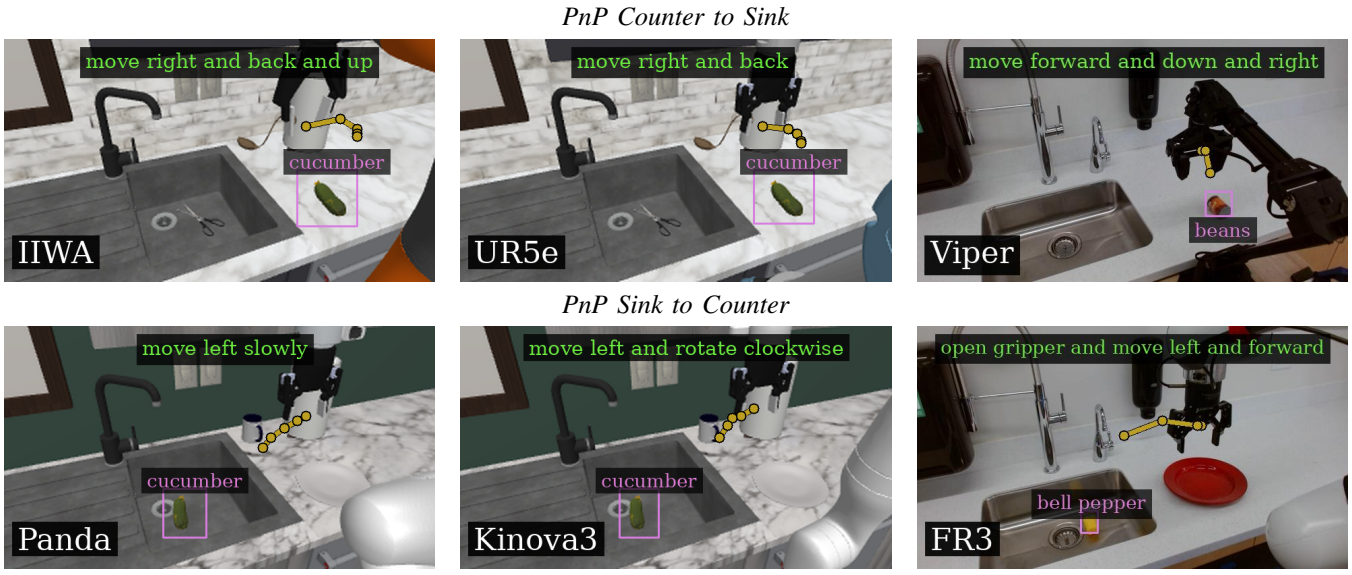

Fig. 6. We show examples of our real tasks (right) in comparison to their sim counterparts (left, middle). We also show actual predictions from **Joint Reps** models, which demonstrate similar predicted representations across different embodiments in unseen scenes.

work [40] has found that much of the benefit from representations can be obtained during training alone, which is desirable to reduce the latency introduced by predicting representations. We investigate this further in the cross-embodiment setting by evaluating **Joint Reps** models trained using **XP-3K** with each supported inference method (i.e., predicting actions only, or predicting one representation and then actions). In Fig. 5, we find that predicting representations can help slightly in some settings, although the overall effect is not substantial. This corroborates the findings from [40], and suggests that representations primarily enhance cross-embodiment transfer in an implicit manner.

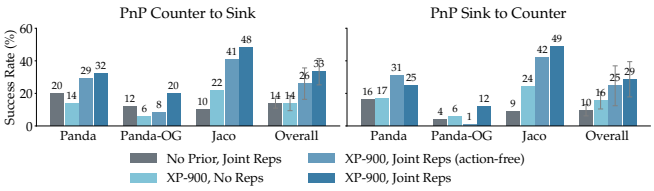

Fig. 7. We find that behavior-aligned representations can leverage action-free prior datasets to improve over learning with no prior dataset, as well as using the full prior dataset with actions but without representations. However, performance is slightly worse than using representations with the prior dataset with actions.

Action-Free Transfer. We conduct additional experiments to assess if representations can improve cross-embodiment transfer from action-free datasets. To do this, we pre-train a variant of our **Joint Reps** model for our two *PnP* tasks on **XP-900**, but without action prediction (only representations). Then, we co-fine-tune on target robot data, but with both representations and action prediction on the target robot data. This is similar to *reasoning pre-training* as proposed in [40]. Because our annotations for language motions are obtained using actions/proprioception, we do not include this representation when training on our action-free prior dataset.

In our results in Fig. 7, we find that action-free pre-training with representations does generally help, improving by 14% overall compared to learning from scratch with the target robot data, and by 11% compared to using the full prior dataset with actions, but without representations. There is a performance decrease compared to using the prior dataset with actions, but much of the performance is retained, except for with Panda-OG, where action-free pre-training does not help. This suggests that behavior-aligned representations can effectively induce transfer from action-free cross-embodiment data through the internal representations they induce, although this may depend on how similar the target embodiment is to those in the action-free data.

V. Real-World Experiments

To see if behavior-aligned representations also enhance real-world cross-embodiment learning, we consider two real target embodiments: the Franka Research 3 (FR3) and ViperX 300 S. We evaluate on variations of the *PnP* tasks in **RoboCasa-X**. We adapt our models pre-trained on **XP-3K** using 50 demonstrations per task. Our setting represents a similar paradigm for sim-to-real transfer as prior work [44], 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. 6.

Action Alignment. Unlike in our simulation experiments, there are now also significant differences in the action spaces between the source and target robots. Although all robots use delta end effector pose control, their control stacks differ significantly, including differences in control frequency, and blocking versus 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 to have less 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 (**XP-3K**), using **No Reps** and **Joint Reps**. Unlike our simulation experiments, for our models with prior data, we train a single model on both target embodiments, to allow for possible transfer. For each embodiment/task pair, we evaluate for 10-30 rollouts. We report task completion progress of each model.

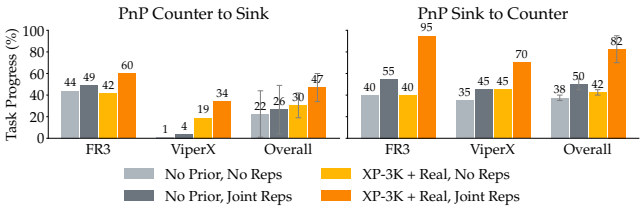

Fig. 8. We adapt models trained on cross-embodiment prior data from simulation to real-world target robots. Behavior-aligned representations significantly improve transfer from prior datasets despite gaps in visual observations and action spaces.

Results. In Fig. 8, we see that **Joint Reps** significantly improves task progress with cross-embodiment data (orange bars), with an overall increase of 28%. Furthermore, representations help more when training with the cross-embodiment data than when only using target robot data (+8%, gray bars), corroborating our simulation results. Without representations, the cross-embodiment data provides limited benefit (+7%) over using the target data alone, suggesting that the representations help enable cross-embodiment transfer that otherwise is difficult to achieve.

The benefit of behavior-aligned representations is more pronounced in our real experiments than in simulation. We suspect this is 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. 6, we show examples of how **Joint Reps** models can predict similar representations across different embodiments in unseen scenes.

VI. Conclusion

We study how behavior-aligned representations can enhance cross-embodiment transfer through implicit data alignment. These representations can easily be labeled with off-the-shelf models or simple transforms using action information. We demonstrate that representations consistently boost performance, with more significant gains with larger cross-embodiment datasets. Through our analysis using **RoboCasa-X**, we find that end-effector traces are the most impactful representation we consider, direct action prediction performs comparably to predicting and conditioning on representations, and representations can facilitate action-free cross-embodiment transfer. Finally, we demonstrate that representations can enhance sim-to-real cross-embodiment transfer on two real robot platforms. We believe that our work is an important step towards understanding positive transfer in cross-embodiment learning.

Limitations. Our work demonstrates the value of using multiple behavior-aligned representations when learning from cross-embodiment data. However, while we were able to achieve improved cross-embodiment transfer with our approach, our setting involves significant alignment between the different embodiments we consider, especially in simulation (e.g., each embodiment in **RoboCasa-X** has the same distribution of camera poses, scenes, and tasks). While we do this to isolate the problem of cross-embodiment transfer, this also causes the representations we study to be more aligned across embodiments. Future work can investigate achieving transfer in less structured settings with greater misalignment between embodiments, and study how different levels of misalignment influence the ability for representations to facilitate transfer.

Additionally, the behavior-aligned representations we consider 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 tasks and embodiments involved. We hope our framework serves as a strong starting point for exploring more behavior-aligned representations to facilitate cross-embodiment generalization.

Acknowledgments

This work was supported by ONR awards N00014-25-1-2479 and N00014-22-1-2293, and NSF award 1941722. Ajay Sridhar is supported by the NSF Graduate Research Fellowship. Toyota Research Institute ("TRI") provided computational resources to support this work. We thank Blake Wulfe, Kyle Hatch, Siddharth Karamcheti, and Sedrick Keh for helpful discussions.

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 (ICML)*, 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 (ICCV)*, 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 (CoRL)*, 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," in *International Conference on Robotics and Automation (ICRA)*, 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 *Conference on Robot Learning (CoRL)*, 2023.

[8] 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.

[9] J. Gao, S. Belkhale, S. Dasari, A. Balakrishna, D. Shah, and D. Sadigh, "A Taxonomy for Evaluating Generalist Robot Manipulation Policies," *IEEE Robotics and Automation Letters (RA-L)*, 2026.

[10] J. H. Yang, D. Sadigh, and C. Finn, "Polybot: Training One Policy Across Robots While Embracing Variability," in *Conference on Robot Learning (CoRL)*, 2023.

[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 (RSS)*, 2024.

[12] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg, "RoVi-Aug: Robot and Viewpoint Augmentation for Cross-Embodiment Robot Learning," in *Conference on Robot Learning (CoRL)*, 2024.

[13] M. Lepert, R. Doshi, and J. Bohg, "SHADOW: Leveraging Segmentation Masks for Cross-Embodiment Policy Transfer," in *Conference on Robot Learning (CoRL)*, 2024.

[14] 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," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[15] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[16] 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," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[17] 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," in *International Conference on Learning Representations (ICLR)*, 2024.

[18] 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," in *Conference on Robot Learning (CoRL)*, 2024.

[19] 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 *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[20] 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 (RSS)*, 2024.

[21] 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," in *Conference on Robot Learning (CoRL)*, 2024.

[22] 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," in *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[23] 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.

[24] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Conference on Robot Learning (CoRL)*, 2023.

[25] S. Belkhale and D. Sadigh, "MiniVLA: A Better VLA with a Smaller Footprint," 2024. [Online]. Available: https://github.com/Stanford-ILIAD/openvla-mini

[26] 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," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[27] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization," in *Conference on Robot Learning (CoRL)*, 2025.

[28] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, *et al.*, "RT-1: Robotics Transformer for Real-World Control at Scale," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[29] 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," in *International Conference on Robotics and Automation (ICRA)*, 2025.

[30] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic Control via Embodied Chain-of-Thought Reasoning," in *Conference on Robot Learning (CoRL)*, 2024.

[31] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg, "KITE: Keypoint-Conditioned Policies for Semantic Manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[32] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point Trajectory Modeling for Policy Learning," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[33] 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," in *Conference on Robot Learning (CoRL)*, 2023.

[34] 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," in *International Conference on Learning Representations (ICLR)*, 2024.

[35] 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 *Conference on Robot Learning (CoRL)*, 2024.

[36] 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," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[37] 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," in *European Conference on Computer Vision (ECCV)*, 2022.

[38] H. Huang, X. Chen, Y. Chen, H. Li, X. Han, Z. Wang, T. Wang, J. Pang, and Z. Zhao, "RoboGround: Robotic Manipulation with Grounded Vision-Language Priors," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[39] J. Clark, S. Mirchandani, D. Sadigh, and S. Belkhale, "Action-Free Reasoning for Policy Generalization," in *Conference on Robot Learning (CoRL)*, 2025.

[40] W. Chen, S. Belkhale, S. Mirchandani, O. Mees, D. Driess, K. Pertsch, and S. Levine, "Training Strategies for Efficient Embodied Reasoning," in *Conference on Robot Learning (CoRL)*, 2025.

[41] 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," in *Conference on Robot Learning (CoRL)*, 2023.

[42] 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," in *European Conference on Computer Vision (ECCV)*, 2024.

[43] 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 *Conference on Robot Learning (CoRL)*, 2023.

[44] 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)*, 2025.