A DATA-CENTRIC ALIGNMENT STUDY FOR CROSS-EMBODIMENT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-embodiment learning offers a compelling way to scale robot learning by transferring knowledge from other source embodiments to a target embodiment. However, a shared end-effector (EEF) format does not guarantee data compatibility, as embodiment-specific differences can exist even for similar manipulation behaviors. We take a data-centric perspective, studying how explicit data alignment affects this transfer with an otherwise unchanged policy. Specifically, we consider four aspects, including frame coordinate, action mapping across arm configurations, Cartesian states, and observation viewpoints. We evaluate the cross-embodiment transfer through a carefully designed setting around source-only tasks tested on the target embodiment. In real-world skill transfer, data alignment alone achieves 90.5% success, compared with 34.8% for naive data mixing. Further analyses show the benefits of shared tasks across embodiment, indicate that alignment benefits extend beyond a particular embodiment, and provide insights into the roles of different alignment components, as well as the choice of strategies within each component. These findings identify cross-embodiment data compatibility as an important consideration in combining heterogeneous robot data and provide a practical data-centric baseline for cross-embodiment transfer. *Real-world videos are in supplementary materials.*

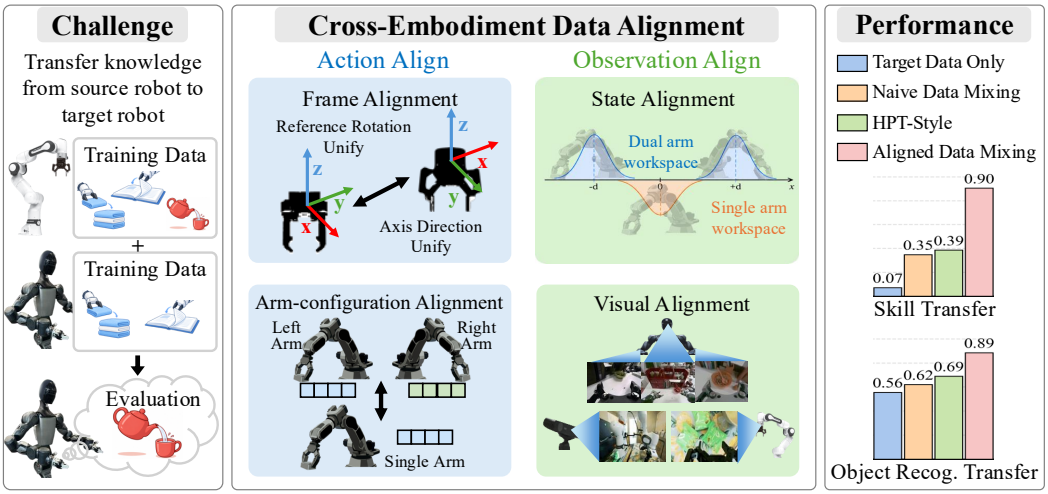


Figure 1: **Overview of data-centric cross-embodiment alignment.** Directly mixing data from different embodiments can introduce mismatches in action and observation, hindering the cross embodiment transfer. We study four aspects of data alignment: frame and arm-configuration alignment for actions, and state and visual alignment for observations. Our experiments examine how these alignment choices affect the transfer of source-only skills and objects to the target embodiment.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Scaling robot learning requires large and diverse demonstration data (Lin et al., 2024; O'Neill et al., 2024; Khazatsky et al., 2024), yet demonstrations are distributed unevenly across embodiments. A task may be available in one robot's dataset while absent from another's. Cross-embodiment learning offers a compelling way that allows a target robot to acquire capabilities from demonstrations collected elsewhere (Zha et al., 2026; Kim et al., 2024; Black et al., 2024; Intelligence et al., 2026). Its value therefore extends beyond increasing training samples. The central question is whether data from other source robots can help the target robot execute tasks unavailable in its data. Answering this requires understanding both what information source data provide, but also how compatible their action and observation representations are with the target embodiment.

A common data format, e.g. end-effector (EEF) action, is an important starting point, but substantial differences can remain within that format. It can use different coordinate-axis directions and tool-orientation references. Action slots can correspond to different numbers of arms, and camera slots can provide different views of the ongoing manipulation. Proprioceptive inputs can also expose workspace coordinates that vary across embodiments. Existing approaches address such heterogeneity through shared representations or embodiment-specific model components (Wang et al., 2024; Bjorck et al., 2025; Team et al., 2024; Bu et al., 2025; Ye et al., 2024; Zheng et al., 2025a;b). We investigate another design choice, asking how much can the usefulness of source demonstrations change when their action and observation conventions are explicitly aligned before joint training.

We study cross-embodiment data compatibility through a set of data alignments, leaving the policy architecture unchanged. Action alignment provides basic representation compatibility, reconciling the coordinate and orientation and mapping source action configurations to the target interface. Based on this compatibility, observation alignment reduces the influence of embodiment-specific state on the learned policy, and assigns available camera inputs according to their observation viewpoints. These operations address different sources of mismatch and have different properties.

We evaluate cross-embodiment transfer performance using skills or objects available only in the source-embodiment data (source-only) and evaluate whether the policy can execute or recognize them on the target embodiment, as shown in Figure 1. The overall real-world study pools data from ten embodiments and evaluates transfer from the other nine to Moz-1 (Team, 2026). We show that data-centric alignment alone can achieve strong cross-embodiment transfer without introducing model-level designs, with further analyses of which tasks benefit more from alignment and qualitative failure cases in the appendix. A controlled simulation study shows the benefits of shared tasks across embodiments in facilitating transfer. Finally, Franka-Moz-1 analyses and ablations indicate that alignment benefits extend beyond a particular embodiment, show distinct benefits of action and observation alignments, and provide insights into the choice of strategies within each alignment component. Together, these studies highlight cross-embodiment data compatibility as an important consideration and provide a practical data-centric baseline for cross-embodiment transfer.

2 RELATED WORKS

Large robotic policies are often trained on demonstrations from different embodiments (Intelligence et al., 2026; Bjorck et al., 2025; Brohan et al., 2022; Wu et al., 2026; Lin et al., 2025; Intelligence et al., 2025; Yu et al., 2025; Liu et al., 2026b; Lin et al., 2024; Liu et al., 2026a), making data compatibility important for cross-embodiment learning (Zha et al., 2026; Chen et al., 2026; Intelligence et al., 2026). Joint-angle actions are difficult to share because robots differ in joint numbers and meanings (Zhao et al., 2023; Wang et al., 2026). End-effector actions offer unified dimensionality, but can still use different coordinate axes and rotational references (Intelligence et al., 2026; Chi et al., 2023; Ganapathi et al., 2022; Team et al., 2024). Delta actions reduce dependence on global coordinates, yet their local frame conventions can still differ (Xie et al., 2025). Therefore, the same physical motion can carry inconsistent information across embodiments

Existing methods often address this problem by introducing explicit embodiment-specific action heads (Wang et al., 2024; Bjorck et al., 2025; Team et al., 2024) or learning shared latent action spaces (Bu et al., 2025; Ye et al., 2024; Zheng et al., 2025a;b; Chen et al., 2025b). While these model-centric designs can effectively improve flexibility across diverse embodiments, data-level alignment across heterogeneous demonstrations remains less systematically studied.

3 CROSS-EMBODIMENT DATA ALIGNMENT

3.1 PROBLEM FORMULATION

Let $\mathcal{D}_r$ denote demonstrations from a target embodiment r, and let $\{\mathcal{D}_s\}$ denote demonstrations from source embodiments $\{s\}$. Each training example contains an observation o_t, a task instruction l, and an action chunk $a_{t:t+H-1}$. Observations include camera images and proprioceptive states. The policy predicts end-effector and gripper actions through the target robot's control interface. We focus on transferring tasks available in source embodiment's data $\{\mathcal{D}_s\}$ but absent from target data $\mathcal{D}_r$. And the target data can still contain demonstrations of other tasks, allowing us to examine how shared tasks between source and target embodiments affect transfer to unseen tasks.

We align source observations o_s and actions a_s with the target embodiment r using data sample d_s,

$$\tilde{o}_s = \Phi_{s\to r}(d_s), \qquad \tilde{a}_s = \Psi_{s\to r}(d_s), \tag{1}$$

where $\Phi_{s\to r}$ denotes the observation transformation and $\Psi_{s\to r}$ denotes the action transformation. The transformed source data are then jointly trained with the target data using the policy's standard training objective. These transformations only modify the data presented to the shared policy and introduce no additional learnable policy modules.

Our alignment is organized around two requirements. Action records should follow compatible physical conventions, and observations should provide an appropriate context for interpreting those actions. These requirements motivate four concrete components: **frame alignment**, **arm-configuration alignment**, **state alignment**, and **visual-viewpoint alignment**. The arm-configuration and visual-viewpoint alignment designs primarily focus on transfer between single-arm and bimanual embodiments in this study. These components serve different roles, where frame and arm-configuration alignment primarily make action information meaningful under the target interface, whereas state and visual alignment reduce mismatches in the observations provided to the policy. Accordingly, some components reconcile the semantics of existing signals, while others remove or reorganize embodiment-dependent information.

3.2 FRAME ALIGNMENT

Even robot datasets represent actions using end-effector poses, their frame conventions can differ across embodiments (O'Neill et al., 2024). A particularly subtle mismatch is the end-effector zero orientation, i.e., the physical orientation represented by a zero rotation $[0, 0, 0]$. Aligning the world axes does not ensure that this orientation is consistent across embodiments. We therefore distinguish and align two factors separately, including the world-coordinate axes in which an orientation is expressed and the end-effector zero orientation used as its reference. Let $R^s \in \mathrm{SO}(3)$ denote an end-effector orientation from source embodiment s. We convert it to the target convention r as

$$\widetilde{R}^{\,r} = R^{\mathrm{axis}}_{s\to r}\, R^s\, R^{\mathrm{zero}}_{s\to r}. \tag{2}$$

where $R^{\mathrm{axis}}_{s\to r}$ aligns the world-coordinate axes and $R^{\mathrm{zero}}_{s\to r}$ aligns the end-effector zero orientation. Both transformations are fixed for each source–target embodiment pair. Importantly, matching the world-axis directions alone is insufficient whenever the same numerical orientation corresponds to different physical end-effector orientations across embodiments.

After determining the world-axis convention, we obtain the remaining zero-orientation offset from a pair of corresponding reference orientations. Let $R^s_\star$ and $R^r_\star$ describe the same physical end-effector orientation under the source and target conventions, respectively. We compute

$$R^{\mathrm{zero}}_{s\to r} = \left(R^{\mathrm{axis}}_{s\to r} R^s_\star\right)^\top R^r_\star. \tag{3}$$

We obtain this correspondence by selecting a clear physical end-effector orientation and expressing it under both conventions. The resulting offset is fixed and reused across all trajectories.

The same world-axis conversion applies to translational quantities. For a displacement vector,

$$\widetilde{\Delta p}^{\,r} = R^{\mathrm{axis}}_{s\to r} \Delta p^s. \tag{4}$$

For policies using relative orientation actions, we first convert the corresponding poses and then construct the relative action according to the policy's original action-encoding convention. This pose-level alignment is independent of the downstream action representation.

The EEF poses in all datasets we use follow right-handed world-coordinate conventions, although their axis directions and the physical orientations represented by the initial rotations are not consistent. We convert them to a frame aligned with Moz-1 and provide the required axis $R^{\text{axis}}_{s\to r}$ and zero-orientation $R^{\text{zero}}_{s\to r}$ offsets in Euler-angle form in Appendix A. Since our policy uses relative EEF actions (Chi et al., 2024), the aligned actions specify local motion increments, reducing dependence on embodiment-specific workspace reachability.

3.3 Arm-Configuration Alignment

Equal action dimensionality does not establish a meaningful correspondence between single-arm and bimanual demonstrations. We map arm configurations according to how the source behavior can be represented through the target interface.

For bimanual-to-single-arm transfer, we compare arm activity within each action chunk and retain the arm with the larger action magnitude. Its action is assigned to the target's working-arm slots. For single-arm-to-bimanual transfer, we place the demonstrated action in a working-arm slot and pad the other arm with an idle action. These transformations are applied to source training examples; the deployed policy predicts actions through the target robot's native interface.

This mapping is particularly relevant to bimanual embodiments containing single-arm behavior. It does not establish equivalence between arbitrary bimanual and single-arm tasks. A slowly moving arm can still provide necessary support or maintain contact. We therefore evaluate arm mapping against alternative assignments and interpret its benefits within the tested manipulation tasks.

3.4 State Alignment

Absolute Cartesian states introduce an additional source of cross-embodiment mismatch, as their numerical values are tied to the choice of workspace origin. Consequently, the same Cartesian state value can correspond to different physical configurations across embodiments when their coordinate origins differ. Recent work has shown that removing proprioceptive states can improve cross-embodiment adaptation and spatial generalization, although relying purely on visual observations can become less reliable when task-relevant visual coverage is insufficient (Zhao et al., 2025).

Motivated by these observations, we mask only the Cartesian position components of the EEF state rather than discarding proprioception entirely. Since the rotational reference has been explicitly aligned, we retain it in the state. In Section 4.4.2, we compare this choice with retaining the original state and shifting the Cartesian origin, allowing us to test whether removing absolute positional information provides benefits beyond simple recentering. Note that masking the Cartesian components is applied to both source- and target-embodiment data.

3.5 Visual Alignment

Camera inputs with the same slot index can provide different information across robots. A wrist view may observe the interaction region, while a third-person or inactive-arm view may provide scene context. We assign source images to target input slots according to these observation roles.

When converting bimanual demonstrations to a single-arm interface, the selected working arm supplies the wrist view, while the other arm's camera can provide scene context. In the reverse direction, an available third-person view can occupy the observation slot associated with the padded idle arm. This preserves access to context without duplicating the working-arm image.

Visual-role alignment reorganizes available observations. It does not perform geometric view synthesis or recover objects that were never observed. The camera mappings are specified for each configuration and remain explicit parts of the data interface.

3.6 Representative Instantiation: Moz-1 and Franka

To make the above alignment operations concrete, we instantiate them between Moz-1 and Franka, as shown in Figure 2. We choose this pair because it represents a deliberately challenging embodiment gap between a human-like bimanual and a single-arm robot. For Franka, we use the DROID dataset (Khazatsky et al., 2024) and follow the widely used visual setup (Intelligence et al., 2025),

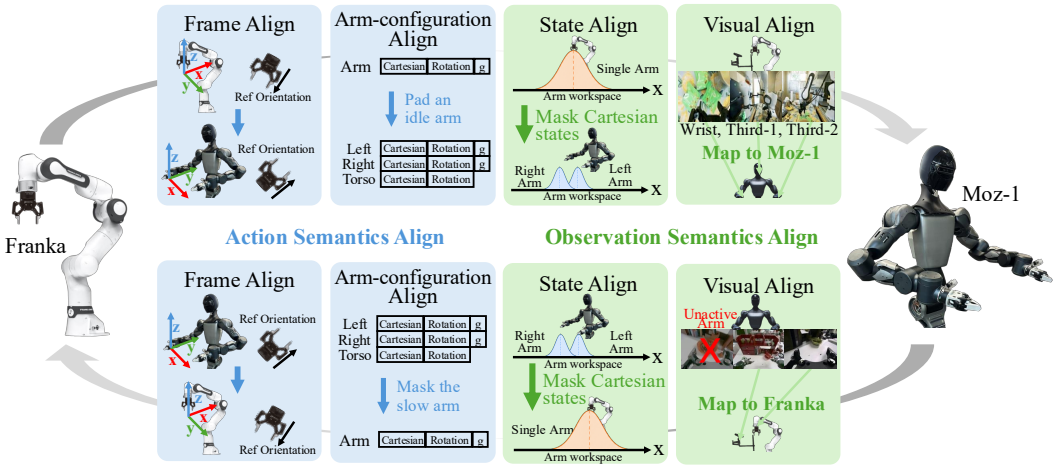


Figure 2: **Overview of our data alignment in a representative human-like bimanual and single-arm setting.** Given a target embodiment, we align source-embodiment demonstrations to the target data semantics from both action and observation perspectives. Action representations are aligned through frame alignment and arm-configuration alignment, while the observation is aligned through state alignment and visual alignment.

where one wrist-camera view together with one of the two available third-person views are input into the policy. These two embodiments differ simultaneously in arm configuration, workspace and state coordinates, end-effector frame conventions, and camera viewpoints. This makes the pair a useful setting for exposing the data-compatibility issues addressed by the four alignment components above. It also serves as the controlled setting for our later bidirectional ablations.

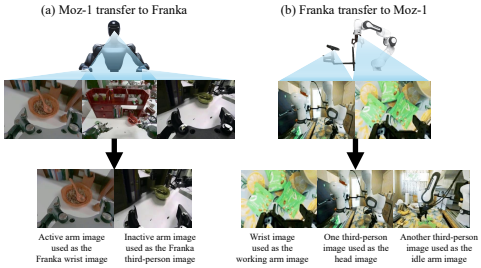


Figure 3: Camera-view mappings for Franka-to-Moz-1 and Moz-1-to-Franka transfer.

Franka to Moz-1. When Moz-1 is the target embodiment, we convert Franka poses and actions to the Moz-1 frame convention and place the single-arm action in a working-arm slot, while padding the other arm with an idle action. The Franka wrist observation is mapped to the working-arm wrist view, while its third-person observations are mapped to the inactive-arm wrist view and the overhead view. We again retain the aligned rotational state while masking Cartesian position states.

Moz-1 to Franka. When Franka is the target embodiment, we first convert Moz-1 EEF poses and actions to Franka frame. For predominantly single-arm behavior in the bimanual demonstrations, we retain the more active arm within each action chunk and map its action to the Franka control interface. Its wrist observation is mapped to the Franka wrist view. Since the inactive arm typically faces the manipulation workspace, we randomly map its wrist observation and the overhead observation to Franka's third-person camera slot. For proprioception, we retain the aligned rotational state and mask the Cartesian position state.

4 Experiments

Our experiments investigate when and how data from different embodiments can support task transfer. We ask four questions. **First,** can heterogeneous data provide knowledge absent from the target data, and how does alignment affect transfer? **Second,** does transferring source-only tasks require shared tasks across embodiments? **Third,** are the benefits of alignment specific to Moz-1 as the target embodiment? **Fourth,** how do different choices of each alignment affect transfer performance?



Figure 4: **Examples of evaluation skills and objects.** We evaluate cross-embodiment transfer from two perspectives: object-recognition transfer and skill-capability transfer. For each target embodiment, we select source-embodiment-only skills and objects to test whether the policy can transfer skill capabilities and object knowledge from the source demonstrations.

4.1 Setups

4.1.1 Real-World Embodiments and Data

Our real-world study covers 10 embodiments with diverse robot configurations. We use seven embodiments from RoboCOIN (Wu et al., 2025), including 3 tabletop bimanual and 4 human-like bimanual robots, together with the Franka single-arm robot from DROID (Khazatsky et al., 2024), the bimanual Frankas from RoboMind (Wu et al., 2024), and the human-like bimanual Moz-1 (Team, 2026). The corresponding datasets contain approximately 700, 400, 500, and 600 hours of demonstrations, respectively. For the single Franka from DROID, following the widely used π-family setting (Intelligence et al., 2025), the policy takes one wrist-camera view together with one of the two available third-person views. In our overall evaluation, Moz-1 serves as the target embodiment, and we study transfer from the other nine source embodiments to Moz-1.

4.1.2 Real-World Evaluation Tasks and Metrics

We evaluate cross-embodiment transfer from 2 perspectives, skill-capability and object-recognition. For skill-capability transfer evaluation, we select 27 skills, where 6 from the Moz-1 dataset and 21 from other embodiments that are absent from the Moz-1 data. We perform 10 rollouts per skill. Performance on the 21 source-only skills measures skill-transfer performance, while performance on the 6 Moz-1 skills measures skill-indomain performance.

For object-recognition evaluation, we select 25 objects, where 5 from the Moz-1 dataset and 20 source-only objects. We organize the objects into groups of five, with the 20 source-only objects form four groups, while the 5 Moz-1 objects form one in-domain group. For each group, the policy is instructed to grasp a specified object, with 10 rollouts per object (success is determined by the initial motion toward the object and an attempt to grasp it, regardless of the subsequent grasp completion). Performance on the two sets measures transfer and in-domain performance, respectively.

Example evaluation skills and objects are shown in Figure 4. The complete list of evaluation skills and objects, together with their source embodiments is provided in Appendix B.

4.1.3 Model

We use Spirit-v1.5 (Team, 2026), detailed implement details are provided in Appendix C. The pretraining data of Spirit-v1.5 do not include DROID, RoboMind, or RoboCOIN, as also reflected by the low transfer performance of "*Target Data Only*". For the model-centric comparison, we adapt an HPT-style (Wang et al., 2024) approach to Spirit-v1.5. We keep the VLM backbone shared, as all embodiments use three observation slots with broadly corresponding viewpoints, where Franka's third-person view is mapped to the fixed scene-observation slot. Moreover, maintaining separate visual backbones for many embodiments is impractical for a large pretrained VLA. We therefore apply HPT-style embodiment-specific modeling only to the state and action interfaces.

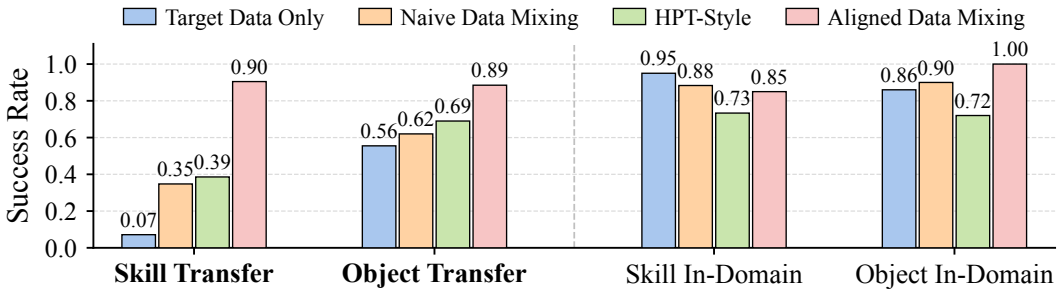


Figure 5: **Overall real-world cross-embodiment transfer results from 9 diverse source embodiments to Moz-1.** Transfer evaluates source-only skills and objects absent from the Moz-1 data, while in-domain evaluates those already present in the Moz-1 data.

We exclude camera-frame methods, as they require accurate camera extrinsics that are unavailable for some of the datasets used in our study. And we do not extensively compare against model-centric baselines, as data-centric and model-centric approaches are not inherently in conflict; this study instead focuses on investigating how data-level alignment affects cross-embodiment transfer.

4.1.4 Controlled Simulation

We also conduct evaluations on a controlled simulation benchmark. We use the "Clean" setting of RoboTwin (Chen et al., 2025a) with Agilex ALOHA as the target and bimanual ARX5 as the source. Five of the 50 skills are excluded from the target data, and source data provide these skills. The list of 5 evaluated skills is provided in Appendix B. Each skill is evaluated with 50 rollouts. This setting allows us to examine both whether source-only skills can transfer to the target embodiment and how the presence of shared skills across embodiments affects such transfer.

4.2 Overal Transfer Performance

We compare four training settings. "*Target Data Only*" uses only Moz-1 data. "*Naive Data Mixing*" combines data without alignment (the missing arm dimensions from Droid are zero-padded after normalization, and the three Franka camera views are assigned to the three input slots in Droid official order). "*HPT-Style*" introduces embodiment-specific state encoders and action encoders and decoders. "*Aligned Data Mixing*" applies all four alignment strategies described in Section 3.

As shown in Figure 5, cross-embodiment training largely preserves in-domain performance, with some degradation compared with "*Target Data Only*". The differences are much larger on skill transfer. "*Target Data Only*" achieves 7.1% success, "*Naive Data Mixing*" 34.8%, "*HPT-Style*" 38.6%, and "*Aligned Data Mixing*" 90.5%. And object recognition shows a similar pattern. These results suggest three observations. **First,** mixed training can retain strong performance on target-native skills. **Second,** source data can support partial skill transfer, with substantially stronger transfer after explicit alignment. **Third,** object-recognition transfer appears easier than skill transfer in this evaluation. **Furthermore,** Appendix E reports task-level results, showing broad transfer improvements (20/21 skills) over "*Naive Data Mixing*". It also analyzes the different sensitivities to alignment of easier skills such as pick-and-place and more challenging skills such as bimanual manipulation. Appendix D further provides a qualitative analysis of some failure cases.

4.3 Controlled Evaluation on RoboTwin

We further study skill transfer in the controlled setting of RoboTwin (Chen et al., 2025a). We use bimanual Agilex ALOHA as the target embodiment and bimanual ARX5 as the source, and select five of the 50 skills for evaluation. We compare four controlled training settings: "*Target Only*," using target-only data (excluding the five selected skills); "*Naive Data Mixing*," additionally using the full source-embodiment data without alignment; "*Aligned Mixing*," aligning the full source-data; and "*Aligned Mixing w/o Shared Skills*," where the aligned source-data are restricted to the five selected skills, explicitly removing shared tasks across the two embodiments from source-embodiment data.

Table 1: Skill transfer evaluation on 5 selected skills from RoboTwin.

Setting	Success
Target Only	3.6%
Naive Data Mixing	26.0%
Aligned Data Mixing	**60.0%**
Aligned Mixing w/o Shared	41.6%
In-Domain	64.4%

As shown in Table 1, "*Target Only*" achieves 3.6% success, and "*Naive Data Mixing*" increases to 26.0%, while "*Aligned Data Mixing*" reaches 60.0%, suggesting that the alignment clearly affects the skill transfer. For reference, "*In-Domain*", trained with target-embodiment demonstrations retaining the selected skills, achieves 64.4% success, showing comparable results with "*Aligned Data Mixing*". When shared skills are removed from the source data, "*Aligned Mixing w/o Shared Skills*" achieves 41.6% success. This shows that shared skills are not necessary for transfer, but suggests a potential benefit from shared tasks across embodiments.

4.4 Analysis and Ablations

In this section, we further conduct in-depth analyses and ablations using Moz-1 and Franka, with their embodiment-specific skills and objects. We focus on them because they present a challenging and representative gap between a human-like bimanual and fixed-base tabletop single-arm robot.

4.4.1 Are the Benefits of Alignment Specific to Moz-1 as the Target Embodiment?

To examine whether the benefits are specific to Moz-1 as the target embodiment, we separately train two models for Moz-1-to-Franka and Franka-to-Moz-1 transfer, based on the alignment strategy in Figure 2. We also conduct ablations by progressively adding the four alignment components in order, as some later alignments rely on the preceding ones to establish compatible data information.

Table 2: Transfer analysis and ablations between Moz-1 and Franka. Gray cells indicate transfer performance on skill-capability and object recognition.

Transfer direction	Source data	Frame align	Arm-config align	Visual align	State align	Moz-1 tasks Skill	Moz-1 tasks Object	Franka tasks Skill	Franka tasks Object
Moz-1 → Franka	✗	✗	✗	✗	✗	31.7%	28.0%	55.0%	82.0%
	✓	✗	✗	✗	✗	36.7%	68.0%	48.3%	86.0%
	✓	✓	✗	✗	✗	38.3%	76.0%	53.3%	82.0%
	✓	✓	✓	✗	✗	48.3%	86.0%	66.7%	90.0%
	✓	✓	✓	✓	✗	58.3%	86.0%	81.7%	86.0%
	✓	✓	✓	✓	✓	76.7%	92.0%	96.7%	90.0%
Franka → Moz-1	✗	✗	✗	✗	✗	73.3%	72.0%	11.7%	28.0%
	✓	✗	✗	✗	✗	70.0%	76.0%	30.0%	40.0%
	✓	✓	✗	✗	✗	76.7%	80.0%	51.7%	72.0%
	✓	✓	✓	✗	✗	81.7%	86.0%	63.3%	80.0%
	✓	✓	✓	✓	✗	83.3%	90.0%	75.0%	86.0%
	✓	✓	✓	✓	✓	93.3%	90.0%	85.0%	80.0%

Table 2 shows that the benefit of alignment is observed in both transfer directions. With all four alignments, skill-transfer success reaches 76.7% for Moz-1-to-Franka and 85.0% for Franka-to-Moz-1, compared with 36.7% and 30.0% under naive data mixing. Object-recognition transfer shows a similar overall trend. These results suggest that the observed benefit is not specific to Moz-1 as the target embodiment, with similar benefits observed when Franka serves as the target.

The ablations further show how transfer changes as the alignment components are introduced. Skill-transfer performance generally increases as alignments are applied, while object recognition benefits particularly from the earlier alignment steps. Notably, the ablations distinguish action-interface compatibility from the choice of conditioning observations. Action alignment establishes the basic compatibility, while observation alignment remains important beyond this basis and provides further great gains.

4.4.2 How Do Different Choices of Each Alignment Affect Transfer Performance?

We further analyze the design choices within each alignment component. Since frame alignment is determined by the action coordinate conventions, we focus on arm-configuration, visual, and state alignment, comparing different choices for each to understand their effects on transfer.

Arm-configuration Alignment. We compare different arm-configuration alignment strategies while using frame alignment and disabling visual and state alignment.

Table 3: Arm-configuration alignment strategies. Gray cells indicate transfer performance.

Transfer direction	Arm-configuration align strategy	Moz-1 tasks		Franka tasks	
		Skill	Object	Skill	Object
Moz-1 → Franka	✗	38.3%	76.0%	53.3%	82.0%
	Randomly mask an arm	41.7%	82.0%	41.7%	78.0%
	Mask the slower arm	48.3%	86.0%	66.7%	90.0%
Franka → Moz-1	✗	76.7%	80.0%	51.7%	72.0%
	Direct copy the arm	78.3%	78.0%	45.0%	72.0%
	Concatenate 2 arms	70.0%	82.0%	40.0%	56.0%
	Pad an idle arm	81.7%	86.0%	63.3%	80.0%

As shown in Table 3, selecting the slower arm for masking shows the strongest transfer performance when reducing bimanual demonstrations to a single-arm interface, reaching 48.3% skill-transfer success. When expanding single-arm demonstrations to a bimanual interface, padding an idle arm performs best, reaching 63.3%. In comparison, random masking, copying, or concatenating arm actions is less effective. These results favor preserving the more active behavior while avoiding artificial motion in the other arm.

Visual Alignment. For visual alignment, the choice of how source camera observations are mapped to target camera slots affects transfer performance. Mapping the inactive-arm or third-person view to the corresponding target camera slot achieves 58.3% and 75.0% skill-transfer success, respectively, outperforming masking or directly copying camera observations.

Table 4: Visual-alignment strategies. Gray cells indicate transfer performance.

Transfer direction	Alignment strategy	Moz-1 tasks		Franka tasks	
		Skill	Object	Skill	Object
Moz-1 → Franka	✗	48.3%	86.0%	66.7%	90.0%
	Mask head cam	26.7%	54.0%	28.3%	56.0%
	Inactive-arm cam as head cam	58.3%	86.0%	81.7%	86.0%
Franka → Moz-1	✗	81.7%	86.0%	63.3%	80.0%
	Copy arm cam	80.0%	84.0%	55.0%	78.0%
	Third-person cam as idle-arm cam	83.3%	90.0%	75.0%	86.0%

Table 5: State-alignment strategies. Gray cells indicate transfer performance.

Transfer direction	Alignment strategy	Moz-1 tasks		Franka tasks	
		Skill	Object	Skill	Object
Moz-1 → Franka	✗	58.3%	86.0%	81.7%	86.0%
	Align state origin	61.7%	90.0%	61.7%	80.0%
	Mask Cartesian state	76.7%	92.0%	96.7%	90.0%
Franka → Moz-1	✗	83.3%	90.0%	75.0%	86.0%
	Align state origin	91.7%	92.0%	78.3%	80.0%
	Mask Cartesian state	93.3%	90.0%	85.0%	80.0%

State Alignment. Masking the Cartesian component in states gives the strongest skill-transfer results, reaching 76.7% and 85.0%, while aligning only the state origin provides smaller gains. This suggests that removing embodiment-dependent Cartesian information can be more effective than reconciling coordinate origins alone in cross-embodiment transfer.

5 Conclusion

We study cross-embodiment transfer from a data-centric perspective, focusing on how the compatibility of heterogeneous demonstrations affects their usefulness to a target robot. We investigate four kinds of alignment, including frame, arm-configuration alignment, state and visual alignment. We evaluate the cross-embodiment transfer performance by testing source-only tasks on the target embodiment. Across real-world experiments, controlled simulation, and detailed analyses, we show that these alignments enable stronger transfer of skills and object knowledge absent from the target data. Controlled experiments further show that shared tasks across embodiments can facilitate the transfer. Finally, our ablations show that alignment benefits are not limited to a particular target embodiment, analyze the distinct benefits of different alignments, and compare concrete strategies for each alignment, identifying suitable choices under our evaluated settings. Together, these findings highlight cross-embodiment data alignment as an important consideration in reusing heterogeneous robot data and provide a practical data-centric baseline for cross-embodiment task transfer.

Limitations. Our study is currently limited to aligning observations already available in existing datasets; it does not synthesize potentially more suitable visual observations when the required viewpoints are unavailable. In addition, our Franka-Moz-1 transfer evaluation currently trains separate models for the two transfer directions. Our future work will explore more unified models that support flexible transfer across embodiments within a single model.

AI USE STATEMENT

Large Language Model tools were used to assist with language editing and revision of the manuscript. All AI-assisted contents were critically reviewed by the authors. The experimental results were obtained from author-conducted experiments, and the final methodological choices, analyses, interpretations, and claims were determined and verified by the authors. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This work studies cross-embodiment transfer using existing robot datasets and robot demonstrations collected for research purposes. Our experiments involve no human subjects or personally sensitive information. We use the datasets in accordance with their intended research use and licenses.

REPRODUCIBILITY STATEMENT

We build on the official Spirit-v1.5 codebase and use its standard training pipeline for policy training and fine-tuning, see Appendix C. To facilitate reproduction of our data alignment and evaluation, we report the frame-offset parameters required to transform each evaluated embodiment to the Moz-1 convention. We also provide the complete lists of skills and objects used in the real-world evaluations and the five skills selected for the controlled RoboTwin evaluation. Together, these details specify the data transformations, training setup, and evaluation sets used in our experiments.

REFERENCES

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi 0$: A vision-language-action flow model for general robot control. *ArXiv*, abs/2410.24164, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, and et al. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *ArXiv*, abs/2505.06111, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025a.

William Chen, Jagdeep Singh Bhatia, Catherine Glossop, Nikhil Mathihalli, Ria Doshi, Andy Tang, Danny Driess, Karl Pertsch, and Sergey Levine. Steerable vision-language-action policies for embodied reasoning and hierarchical control. *ArXiv*, abs/2602.13193, 2026.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-x: Enhancing latent action modeling in vision-language-action models. *ArXiv*, abs/2507.23682, 2025b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

Aditya Ganapathi, Peter R. Florence, Jacob Varley, Kaylee Burns, Ken Goldberg, and Andy Zeng. Implicit kinematic policies: Unifying joint and cartesian action spaces in end-to-end robot learning. *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2656–2662, 2022.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, and et al. π0.5: a vision-language-action model with open-world generalization. *ArXiv*, abs/2504.16054, 2025.

Physical Intelligence, Bo Ai, Ali Abdalmagid Amin, Raichelle J. Aniceto, Ashwin Balakrishna, Gregor Balke, Kevin Black, George Bokinsky, and et al. π0.7: a steerable generalist robotic foundation model with emergent capabilities. *ArXiv*, abs/2604.15483, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *ArXiv*, abs/2410.18647, 2024.

Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *ArXiv*, abs/2505.11917, 2025.

Songming Liu, Bangguo Li, Kai Ma, Lingxuan Wu, Hengkai Tan, Ouyang Xiao, Hang Su, and Jun Zhu. Rdt2: Exploring the scaling limit of umi data towards zero-shot cross-embodiment generalization. *ArXiv*, abs/2602.03310, 2026a.

Yufeng Liu, Hang Yu, Juntu Zhao, Bocheng Li, Di Zhang, Ming-Zhe Li, Wenxuan Wu, Yingdong Hu, Junyuan Xie, Junliang Guo, Dequan Wang, and Yang Gao. Learning native continuation for action chunking flow policies. *ArXiv*, abs/2602.12978, 2026b.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Spirit AI Team. Spirit-v1.5: Clean data is the enemy of great robot foundation models. *Spirit AI Blog*, 2026. https://www.spirit-ai.com/en/blog/spirit-v1-5.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *ArXiv*, abs/2409.20537, 2024.

Shihefeng Wang, Kangchen Lv, Mingrui Yu, and Xiang Li. Embodisteer: Steering embodiment-agnostic visuomotor policies with joint-space guidance for zero-shot cross-embodiment deployment. *arXiv preprint arXiv:2606.12965*, 2026.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fan Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yi Ren, Kejia Zhang, Hui Yu, Jing Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Ze Jun Wang, Ran Cheng, Yong-Lu Li, Yong-Zhi Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic vla foundation model. *ArXiv*, abs/2601.18692, 2026.

Sicheng Xie, Lingchen Meng, Zijie Diao, Haidong Cao, Zhiying Du, Shuyuan Tu, Jiaqi Leng, Qiuyue Wang, Mingsheng Li, Shuai Bai, et al. Unify robot actions in camera frame. *arXiv preprint arXiv:2511.17001*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Lidén, Kimin Lee, Jianfeng Gao, Luke S. Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. *ArXiv*, abs/2410.11758, 2024.

Hang Yu, Juntu Zhao, Yufeng Liu, Kaiyu Li, Chengwei Ma, Di Zhang, Yingdong Hu, Guang Chen, Junyuan Xie, Junliang Guo, Junqiao Zhao, and Yang Gao. Point what you mean: Visually grounded instruction policy. *ArXiv*, abs/2512.18933, 2025.

Lihan Zha, Asher Hancock, Mingtong Zhang, Tenny Yin, Yixuan Huang, Dhruv Shah, Allen Z. Ren, and Anirudha Majumdar. Lap: Language-action pre-training enables zero-shot cross-embodiment transfer. *ArXiv*, abs/2602.10556, 2026.

Juntu Zhao, Wenbo Lu, Di Zhang, Yufeng Liu, Yushen Liang, Tianluo Zhang, Yifeng Cao, Junyuan Xie, Yingdong Hu, Shengjie Wang, et al. Do you need proprioceptive states in visuomotor policies? *arXiv preprint arXiv:2509.18644*, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22508–22519, 2025a.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayi Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *ArXiv*, abs/2510.10274, 2025b.

APPENDIX

A FRAME-ALIGNMENT OFFSETS

Moz-1 defines its EEF poses in a right-handed world coordinate frame, with the x-, y-, and z-axes pointing leftward, backward, and upward, respectively. Its reference EEF orientation points vertically downward, with the camera located on the left side of the end effector. We use this convention as the target frame for aligning the other embodiments. Table 6 reports the axis and zero-orientation offsets, $R^{\text{axis}}_{s\to r}$ and $R^{\text{zero}}_{s\to r}$ in Eq. 2, represented as Euler-angle vectors.

Table 6: Frame-alignment offsets from each source embodiment to the Moz-1 convention. All offsets are represented as Euler-angle vectors in radians.

Source embodiment	$R^{\text{axis}}_{s\to r}$	$R^{\text{zero}}_{s\to r}$
Agilex Cobot Magic	$[0,\ 0,\ -\pi/2]$	$[\pi,\ 0,\ \pi/2]$
Agilex ALOHA	$[0,\ 0,\ -\pi/2]$	$[\pi,\ 0,\ \pi/2]$
Galaxea R1 Lite	$[0,\ 0,\ -\pi/2]$	$[0,\ -\pi/2,\ 0]$
Realman RMC	$[0,\ 0,\ -\pi/2]$	$[\pi,\ 0,\ \pi]$
Agibot G1	$[0,\ 0,\ -\pi/2]$	$[\pi,\ 0,\ -\pi/2]$
Galbot G1	$[0,\ 0,\ -\pi/2]$	$[0,\ \pi/2,\ 0]$
AlphaBot 2	$[0,\ 0,\ -\pi/2]$	Left Arm: $[-\pi/2,\ -\pi/2,\ 0]$ Right Arm: $[\pi/2,\ -\pi/2,\ 0]$
DROID (Franka)	$[0,\ 0,\ -\pi/2]$	$[0,\ \pi,\ 0]$
RoboMind (Bimanual Franka)	$[0,\ 0,\ -\pi/2]$	$[0,\ \pi,\ 0]$

B EVALUATION SKILLS AND OBJECTS

In this section, we list the skills and objects used in evaluations. Tables 9 and 7 report the skills and objects available only in the source embodiments and Moz-1, respectively. Table 8 lists the five skills selected for the controlled RoboTwin evaluation.

For object recognition, success requires the initial motion to the object and attempt to grasp it among five candidates, regardless of the subsequent grasp completion.

Table 7: Skills and objects available only in Moz-1 for real-world evaluation.

Skill	Embodiment	Object	Embodiment
Open a wooden drawer	Moz-1	Capybara toy	Moz-1
Place a stapler into a drawer	Moz-1	Barcode scanner	Moz-1
Pull a tissue from a tissue box	Moz-1	Green tea cup	Moz-1
Pick up a white bouquet vertically	Moz-1	Small tea canister	Moz-1
Put a napkin into the desk box	Moz-1	Bagged tea	Moz-1
Grasp a watering-can handle and water a plant	Moz-1		

Table 8: Five skills selected for the controlled RoboTwin evaluation.

Skill
Pick up a microphone with one arm and pass it to the other arm.
Open a laptop with one arm.
Rotate a QR-code sign to face the robot.
Pick up a scanner and an object with the two arms and scan the object.
Stack three bowls sequentially at the center of the table.

Table 9: Skills and objects only in the source embodiments for real-world transfer evaluation.

Skill	Source embodiment	Object	Source embodiment
Open a microwave door	Agibot G1	Egg-yolk pastry	Agilex Cobot Magic
Pour rice into a bowl	Agibot G1	Pencil sharpener	Agilex Cobot Magic
Wipe a blackboard	Agilex Cobot Magic	Knife holder	Agilex ALOHA
Unscrew a bottle cap	Agilex Cobot Magic	Monster Energy drink	Agilex ALOHA
Zip up a bag	Agilex ALOHA	Pipette	Agibot G1
Clink two cups together	Galbot G1	Angle bracket	Agibot G1
Stir water in a cup	AlphaBot 2	Docking station	AlphaBot 2
Pour water from a thermos into a cup	AlphaBot 2	Sandbag	AlphaBot 2
Fold a towel with both arms	Galaxea R1 Lite	Bath sponge	Galaxea R1 Lite
Pick up a box with both arms	Realman RMC	Drone controller	Galaxea R1 Lite
Open a cloth bag	Realman RMC	Mixue milk tea	Galbot G1
Pass a bowl between the two arms	Realman RMC	Utility knife	Realman RMC
Wipe a plate with a cloth	Bimanual Franka	Disposable toiletry kit	Bimanual Franka
Wipe the surface of a cup with a cloth	Bimanual Franka	Test tube	Bimanual Franka
Brush a glass cup	Bimanual Franka	Plastic measuring cup	Bimanual Franka
Fold a towel with one arm	DROID	Whisk	DROID
Cover a banana with a towel	DROID	Remote control	DROID
Throw a cleaning cloth into a large trash bin	DROID	Grapes	DROID
Open a laptop	DROID	Tape measure	DROID
Close a laptop	DROID	Screwdriver	DROID
Rotate a remote control by 90°	DROID		

C Model Implement Details

For the overall experiments, we train for 200k steps with a batch size of 256 on 8 A100 GPUs. For the Franka-Moz-1 experiments, we use the same batch size and GPU configuration and train for 100k steps. For multi-embodiment training, the sampling weight of each embodiment is proportional to its number of frames raised to the power of 0.43 (Black et al., 2024). Given the computational cost of multi-embodiment VLA training, we use one training run per setting and report performance over multiple rollouts. All models are trained to convergence under the same training protocol.

D Qualitative Analysis of Failure Cases

We further analyze failure cases to understand the behavioral differences associated with data alignment. Without sufficient alignment, failed transfer does not necessarily manifest as invalid robot motion: the policy often produces plausible reaching or manipulation behaviors, but fails to realize the task-specific object interactions or multi-stage structure of the source-only skill. With alignment, these task-relevant behaviors are more coherently expressed on the target embodiment, including object interactions, tool use, and coordinated multi-stage manipulation. These observations provide behavioral-level evidence that the gains from alignment reflect improved transfer of task-specific behaviors rather than merely improved low-level controllability or easy skill performance. Representative success and failure cases are provided in our supplementary videos, covering tasks with both pronounced and more modest behavioral differences between aligned and unaligned settings.

E Per-Task Evaluation Results

Tables 10-13 give the successful-rollout counts underlying the aggregate results. "*Target*", "*Naive*", "*HPT*", and "*Aligned*" denote "*Target Data Only*", "*Naive Data Mixing*", "*HPT-Style*", and "*Aligned*

Data Mixing", respectively. The per-task distribution suggests that source-task coverage alone is insufficient for reliable transfer. Relatively simple skills, such as pouring, wiping, and pick-and-place, can sometimes transfer through naive mixing, whereas more challenging skills involving folding, stirring, and bimanual manipulation often require better data compatibility despite being well represented in the source data. Alignment improves both types of skills across different source datasets, with particularly clear gains on tasks that fail under direct mixing. This indicates that the presence of a skill in source data does not imply its direct transferability, particularly for more challenging manipulation behaviors.

Table 10: Per-skill results in the controlled RoboTwin evaluation. All entries are successful rollouts out of 50. "No shared" retains aligned source data for the five evaluated skills only; "In-domain" retains target data for these skills.

RoboTwin skill	Target	Naive	Aligned	No shared	In-domain
Pick up a microphone with one arm and pass it to the other arm	0/50	0/50	28/50	19/50	47/50
Open a laptop with one arm	4/50	30/50	37/50	26/50	32/50
Rotate a QR-code sign to face the robot	0/50	19/50	33/50	27/50	28/50
Pick up a scanner and an object with the two arms and scan the object	0/50	0/50	19/50	11/50	24/50
Stack three bowls sequentially at the center of the table	5/50	16/50	33/50	21/50	30/50
Total	9/250	65/250	150/250	104/250	161/250

Table 11: Per-skill transfer to Moz-1 for the 21 source-only skills. All entries are successful rollouts out of 10.

Skill	Source embodiment	Target	Naive	HPT	Aligned
Open a microwave door	Agibot G1	0/10	0/10	2/10	9/10
Pour rice into a bowl	Agibot G1	8/10	7/10	8/10	9/10
Wipe a blackboard	Agilex Cobot Magic	0/10	9/10	7/10	10/10
Unscrew a bottle cap	Agilex Cobot Magic	0/10	5/10	5/10	9/10
Zip up a bag	Agilex ALOHA	0/10	6/10	8/10	9/10
Clink two cups together	Galbot G1	0/10	3/10	2/10	10/10
Stir water in a cup	AlphaBot 2	0/10	0/10	2/10	9/10
Pour water from a thermos into a cup	AlphaBot 2	0/10	10/10	1/10	9/10
Fold a towel with both arms	Galaxea R1 Lite	0/10	0/10	6/10	10/10
Pick up a box with both arms	Realman RMC	0/10	1/10	1/10	8/10
Open a cloth bag	Realman RMC	0/10	0/10	3/10	10/10
Pass a bowl between the two arms	Realman RMC	0/10	1/10	2/10	10/10
Wipe a plate with a cloth	Bimanual Franka	0/10	8/10	7/10	9/10
Wipe the surface of a cup with a cloth	Bimanual Franka	0/10	1/10	4/10	9/10
Brush a glass cup	Bimanual Franka	0/10	4/10	5/10	9/10
Fold a towel with one arm	DROID	0/10	3/10	3/10	8/10
Cover a banana with a towel	DROID	1/10	3/10	9/10	8/10
Throw a cleaning cloth into a large trash bin	DROID	0/10	7/10	2/10	9/10
Open a laptop	DROID	0/10	0/10	0/10	8/10
Close a laptop	DROID	0/10	2/10	0/10	8/10
Rotate a remote control by 90°	DROID	6/10	3/10	4/10	10/10
Total	21 skills	15/210	73/210	81/210	190/210

Table 12: Per-object recognition on Moz-1 for the 20 source-only objects. All entries are successful rollouts out of 10.

Object	Source embodiment	Target	Naive	HPT	Aligned
Egg-yolk pastry	Agilex Cobot Magic	5/10	1/10	8/10	10/10
Pencil sharpener	Agilex Cobot Magic	6/10	2/10	4/10	6/10
Knife holder	Agilex ALOHA	4/10	8/10	8/10	10/10
Monster Energy drink	Agilex ALOHA	10/10	9/10	10/10	10/10
Pipette	Agibot G1	1/10	5/10	0/10	4/10
Angle bracket	Agibot G1	6/10	4/10	9/10	10/10
Docking station	AlphaBot 2	6/10	8/10	8/10	8/10
Sandbag	AlphaBot 2	10/10	10/10	10/10	10/10
Bath sponge	Galaxea R1 Lite	6/10	10/10	10/10	10/10
Drone controller	Galaxea R1 Lite	9/10	10/10	10/10	10/10
Mixue milk tea	Galbot G1	10/10	10/10	6/10	10/10
Utility knife	Realman RMC	3/10	8/10	10/10	10/10
Disposable toiletry kit	Bimanual Franka	9/10	4/10	3/10	9/10
Test tube	Bimanual Franka	3/10	5/10	8/10	10/10
Plastic measuring cup	Bimanual Franka	9/10	9/10	9/10	10/10
Whisk	DROID	5/10	7/10	2/10	8/10
Remote control	DROID	2/10	4/10	3/10	8/10
Grapes	DROID	5/10	1/10	10/10	9/10
Tape measure	DROID	0/10	6/10	10/10	6/10
Screwdriver	DROID	2/10	2/10	0/10	9/10
Total	20 objects	111/200	123/200	138/200	177/200

Table 13: Per-item performance on skills and objects present in Moz-1 training data. All entries are successful rollouts out of 10. The two subtotals are over 60 and 50 rollouts, respectively.

Moz-1 item	Target	Naive	HPT	Aligned
Skills				
Open a wooden drawer	10/10	9/10	8/10	5/10
Place a stapler into a drawer	7/10	6/10	5/10	10/10
Pull a tissue from a tissue box	10/10	10/10	8/10	9/10
Pick up a white bouquet vertically	10/10	8/10	6/10	8/10
Grasp a watering-can handle and water a plant	10/10	10/10	7/10	9/10
Put a napkin into the desk box	10/10	10/10	10/10	10/10
Skill total	57/60	53/60	44/60	51/60
Objects				
Capybara toy	10/10	10/10	10/10	10/10
Barcode scanner	4/10	8/10	3/10	10/10
Green tea cup	10/10	7/10	4/10	10/10
Small tea canister	9/10	10/10	9/10	10/10
Bagged tea	10/10	10/10	10/10	10/10
Object total	43/50	45/50	36/50	50/50