WHAT MAKES A GOOD PRE-TRAINING? DISSECTING PRE-TRAINING SCIENCE FOR CROSS-EMBODIMENT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of vision-language-action (VLA) models has fueled the consensus that aggregating cross-embodiment data is key to generalist policies. However, given the significant structural heterogeneity across robot embodiments, the efficacy of simple data scaling remains questionable. **To address this, we systematically investigate the efficacy of cross-embodiment pre-training.** Through real-world manipulation experiments and controlled comparisons on RoboTwin 2.0, we identify kinematic similarity as an important consideration for pre-training data selection. In simulation, pre-training on ARX, which is kinematically closer to Piper, improves downstream performance, whereas FR3 pre-training yields negative transfer. Furthermore, we demonstrate that scaling data diversity, dataset size, and training duration all improve performance on downstream tasks. **Our findings suggest that future generalist robot learning paradigms must shift from random data collection to strategic data curation.**

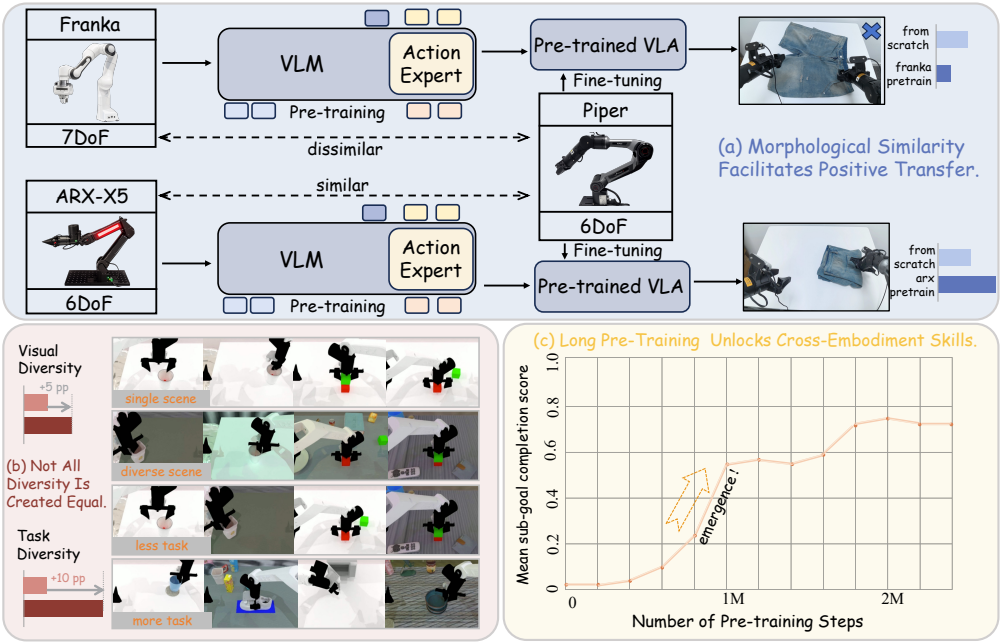


Figure 1: **An overview of the determinants for cross-embodiment transfer.** We analyze the impact of morphological similarity, data diversity, and training duration. Our results reveal that (a) pre-training on morphologically similar robots improves downstream performance, (b) task breadth is more critical than visual variation, and (c) long-horizon capabilities emerge after extended pre-training. Gains in (b) are averaged over clean and randomized evaluation conditions; pp denotes percentage points, indicating absolute differences in downstream task success rates.

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

The success of foundation models has demonstrated the transformative impact of scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Brown et al., 2020). Inspired by this, recent works have focused on using cross-embodiment datasets to pre-train robot foundation models. Recent works such as π_0 (Black et al., 2024) exemplify this trend. These initiatives rest on a core assumption: given sufficient data scale, models can bridge the gaps between robot morphologies and develop motion control capabilities. Consequently, pooling millions of trajectories from diverse embodiments—such as Franka Research 3 (FR3), UR5e, and ARX-X5—has become the prevailing approach for developing generalist policies, aiming to replicate the emergent abilities observed in LLM.

However, transferring such abilities to robotics faces a fundamental challenge: distinct embodiments. Unlike the unified semantic vocabularies found in language models, robots exhibit inherently distinct kinematics, joint configurations, and action space topologies. For example, the Franka Research 3 features 7-DoF with kinematic redundancy (null-space), whereas the ARX-X5 operates with a constrained 6-DoF configuration. This discrepancy implies that optimal policies for the same task can lie on fundamentally different manifolds. This physical "Tower of Babel" raises a critical question: How does cross-embodiment pretrain data affect the downstream tasks performance? While recent research suggests that naive aggregation may yield detrimental results (Shi et al., 2025), the community still lacks an understanding of exactly **how different pre-training strategies impact downstream tasks performance**.

To address this challenge, we present a systematic empirical study on the factors of cross-embodiment transfer. We adopt a rigorous "pre-training $\rightarrow$ fine-tuning" evaluation paradigm to isolate cross-embodiment transferability. Our evaluation spans a broad spectrum of tasks, ranging from pick-and-place operations to complex, long-horizon manipulation skills such as folding pants. Through this setup, we systematically analyze three pivotal dimensions: morphological similarity, the nature of data diversity (visual vs. semantic), and the interplay between dataset scale and training duration in shaping downstream manipulation performance.

We derive a series of counter-intuitive yet actionable insights, challenging the naive data scaling assumption:

- **Pre-training on morphologically similar robots improves downstream performance.** Our results identify kinematic similarity as an important consideration when selecting source embodiments for pre-training. Across six Piper tasks, ARX pre-training raises mean success from 0.28 to 0.38, whereas FR3 pre-training reduces it to 0.04. This performance contrast is consistent with ARX's smaller kinematic distance to Piper. Under a fixed pre-training budget, a 50:50 ARX+FR3 mixture reaches 0.30, showing that incorporating an additional source embodiment does not necessarily improve transfer.
- **Task diversity is more critical than visual diversity**: We argue that not all forms of diversity are created equal. Our results show that expanding tasks yields significantly higher generalization gains (+10 percentage points on average) compared to merely increasing visual domain randomization (e.g., lighting and textures, +5 percentage points on average). This suggests that learning diverse motor primitives is more critical than visual robustness for foundation models.
- **Larger scale data and longer training drive superior performance**: We observe clear benefits from larger pre-training datasets and longer training, especially for long-horizon tasks. In particular, compositional abilities (e.g., folding pants) exhibit an emergence-like pattern: while atomic skills are acquired early, successful composition appears only after extended optimization.

In conclusion, our study highlights the limitations of naive data scaling in cross-embodiment robot learning. We conducted experiments in both real-world and simulation environments, collecting 500 hours of real-robot data and 35 hours of simulation data to examine how pre-training data composition affects downstream transfer. We demonstrate that building generalist foundation models requires shifting from mere data accumulation to strategic curation emphasizing morphological compatibility and targeted task diversity. These findings lay the groundwork for refining scaling laws in robotics, ultimately guiding the creation of more efficient and versatile policies.

2 RELATED WORK

Large-scale robot dataset. The robotics community is actively building large-scale, diverse trajectory datasets (O'Neill et al., 2023; Khazatsky et al., 2024; Bu et al., 2025; Mu et al., 2025; Chen et al., 2025; Wu et al., 2024; Hou et al., 2025) to develop robot foundation models. Open X-Embodiment (OXE) (O'Neill et al., 2023) aggregates over one million trajectories across 22 embodiments, establishing a unified data interface (RLDS) that effectively bridges data silos between laboratories. Furthermore, DROID (Khazatsky et al., 2024) standardizes a Franka-based data-collection platform deployed across 13 institutions, yielding 76k demonstrations over 564 scenes and 86 tasks to improve in-the-wild generalization. AgiBot World (Bu et al., 2025) targets data at scale with a human-in-the-loop collection pipeline, reporting data from 100 robots and 100+ real-world scenarios. Additionally, RoboTwin (Mu et al., 2025; Chen et al., 2025) proposes a generative AI-based paradigm, constructing diverse dual-arm simulation environments to address real-world data scarcity and improve sim-to-real generalization.

Robot foundation model with cross embodiment pre-training. Large-scale pre-training has become a key paradigm for enabling generalist capabilities in robotics (Black et al., 2024; Brohan et al., 2023; Kareer et al., 2025). Modern vision-language-action (VLA) models (Black et al., 2024; Brohan et al., 2022; 2023; Reed et al., 2022; Intelligence et al., 2025; Bjorck et al., 2025; Zhou et al., 2025) typically leverage pre-trained vision-language backbones (Bai et al., 2023; Wang et al., 2024; Beyer et al., 2024) as an initialization and undergo further pre-training on large-scale cross-embodiment robot trajectories. The underlying hypothesis of these works is often that scaling data volume implicitly bridges the gap between high-level semantics and low-level control across diverse robots.

Recent studies demonstrate that data composition matters beyond scale. Abe et al. (2026) analyze morphology-related gradient conflicts in cross-embodiment offline reinforcement learning for locomotion and mitigate them through embodiment-based grouping. Shi et al. (2025) examine task and embodiment diversity in manipulation, showing that single-embodiment pre-training can support cross-embodiment transfer and that scene diversity benefits generalization under distribution shifts.

Our study examines source-robot choice and data diversity as two considerations for pre-training data selection in VLA manipulation. By relating transfer outcomes to kinematic similarity and distinguishing visual from behavioral diversity, we derive practical guidance for curating cross-embodiment pre-training data.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given the availability of large-scale expert demonstration datasets $\mathcal{D}_{\text{pre-train}}$ from a set of diverse source embodiments $\mathcal{E}_{\text{source}}$ (e.g., ARX-X5 and FR3), we aim to leverage this data for knowledge transfer to a novel target embodiment e_{target}. This strategy seeks to mitigate the need for collecting extensive and costly new data for e_{target} from scratch. This raises a central research question: What strategies of pre-training $\mathcal{D}_{\text{pre-train}}$ are most effective for enabling cross-embodiment knowledge transfer?

We evaluate the efficacy of this transfer using a two-stage paradigm:

1. **Pre-training:** The policy π_θ is first trained on $\mathcal{D}_{\text{pre-train}}$, where the target robot is strictly unseen (i.e., $e_{\text{target}} \notin \mathcal{E}_{\text{source}}$). This yields the pre-trained weights θ_{pre}.

2. **Fine-tuning:** The pre-trained model $\pi_{\theta_{\text{pre}}}$ is subsequently fine-tuned on a limited dataset $\mathcal{D}_{\text{target}}$ collected from e_{target}, resulting in the final policy $\pi_{\theta_{\text{fine}}}$.

Our experiments systematically compare how different $\mathcal{D}_{\text{pre-train}}$ configurations impact the final performance of $\pi_{\theta_{\text{fine}}}$ on the target embodiment e_{target}.

3.2 Pre-training Data Composition

Real-World Teleoperation Dataset. We compiled and aggregated a large-scale, real-world manipulation dataset totaling nearly 500 hours of demonstrations. This dataset is a composite of publicly available and internally collected data (see fig. 2 for a detailed breakdown). The embodiment composition is dominated by the ARX robot, but it also includes data from other morphologically diverse arms, such as the FR3 and UR5e. In terms of task distribution, the data covers a variety of manipulation skills, including common actions like picking and placing, as well as complex, dexterous tasks like Folding Pants. Folding tasks account for approximately 70% of the pre-training data, primarily involving clothes folding, with pants-folding demonstrations making up a smaller portion.

Crucially, this pre-training dataset does not contain any data from the target embodiment, Piper, whose demonstrations are introduced only during downstream fine-tuning. This exclusion is essential to ensure that our evaluation measures pure cross-embodiment transfer capabilities.

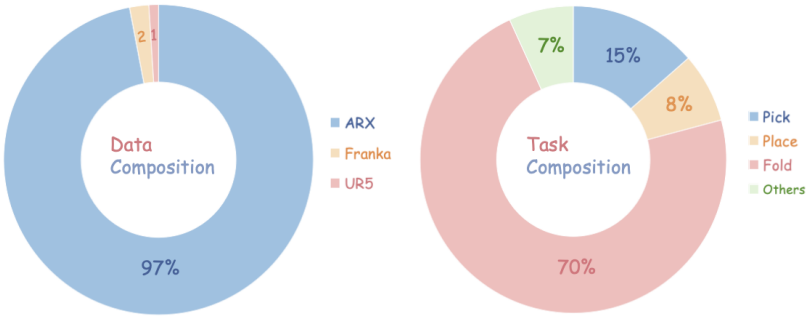


Figure 2: **Overview of Teleoperation Dataset.** (a) Dataset composition by embodiment type. (b) Dataset composition by task type. Percentages are rounded to the nearest whole number.

Simulation Dataset. We utilized the official pre-collected dataset for RoboTwin 2.0 as our Simulation pre-training dataset. This dataset encompasses 50 dual-arm manipulation tasks within the simulated environment. For each task, it provides 50 clean (non-randomized) trajectories and 500 domain-randomized trajectories. To facilitate controlled comparisons of morphological similarity, we extracted the demonstration data specifically for the ARX and FR3 embodiments, amounting to approximately 3.5 million steps for each. These datasets were subsequently used to train three independent sets of pre-trained weights: ARX-only, FR3-only, and a mixed ARX+FR3 configuration.

3.3 Model Architecture

We adopt the architecture of π_0, a state-of-the-art vision-language-action (VLA) model designed for general-purpose robot control with strong performance on benchmarks such as RoboTwin 2.0. Since our core objective is to evaluate the impact of the robotic pre-training data $\mathcal{D}_{\text{pre-train}}$ on downstream task performance rather than the model architecture, we initialize all models from the pre-trained weights of PaliGemma (Beyer et al., 2024). This ensures that any observed performance differences (e.g., between "Direct Fine-tuning" and "Pre-trained" configurations) can be attributed to the robotic data pre-training phase rather than architectural choices. The proprioceptive input consists of the joint positions of both arms and the gripper states. The action vectors have 14 dimensions for ARX and Piper and 16 dimensions for FR3. Each vector concatenates the left-arm joint positions, the left gripper command, the right-arm joint positions, and the right gripper command, in that order. Actions are normalized using embodiment-specific statistics and padded with trailing zeros to a shared 32-dimensional output space. Padded dimensions are excluded from the training loss. All models share a single action head that outputs actions in this unified space. During fine-tuning on Piper, we retain the pre-trained parameters, including those of the action head, and update all model parameters, jointly adapting the vision-language backbone and action expert to Piper.



Figure 3: **The experiment setup and real-world task.** Left: The piper robot equipped with three cameras. Right: Sequential frames illustrating a successful rollout for the folding clothes task.

4 EXPERIMENTS

In this section, we dissect the determinants of transfer, moving beyond naive scaling to analyze the role of data composition and training dynamics. Specifically, our investigation addresses the following core questions:

- To what extent does cross-embodiment pre-training boost performance on challenging, real-world dexterous manipulation tasks?
- How critical is morphological similarity between the pre-training and fine-tuning embodiments for successful cross-embodiment transfer?
- Under a fixed pre-training budget, how do task diversity versus visual diversity affect transfer performance?
- Holding data composition fixed, how do dataset size and pre-training steps influence transfer performance?

4.1 REAL-WORLD EXPERIMENT

4.1.1 EXPERIMENTAL SETUP

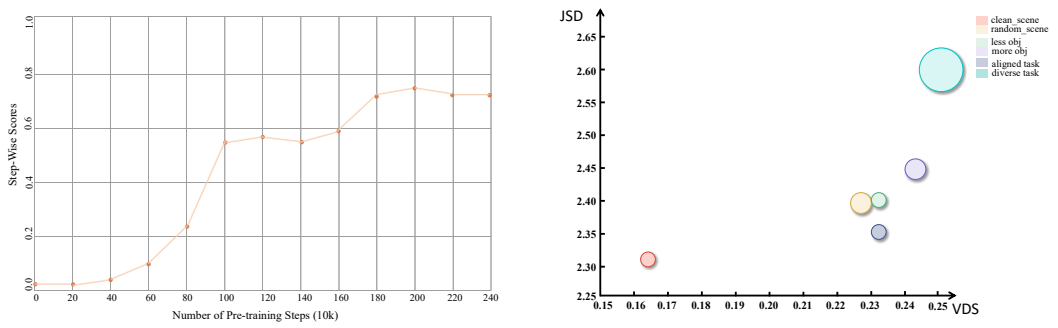
Real-world task. To rigorously evaluate cross-embodiment transfer, we employ "Folding Pants," a challenging long-horizon dexterity task. In this task, a pair of pants is placed on the workspace, and the model is required to execute a complex bimanual sequence: grasping the pants, folding them with both arms, and placing the folded pants on the right side of the table. For all model comparisons in this real-world evaluation, we use a dedicated dataset of 300 expert trajectories for fine-tuning. Both policies are fine-tuned for 40,000 steps using the same learning rate.

Real robot setup. Experiments are conducted on a Piper robot with two 6-DoF arms, each with a gripper. We selected Piper due to its **morphological similarity** to the ARX robot (the primary source embodiment in our pre-training data). Camera configuration details are provided in Appendix.

Evaluation criterion. We evaluate each fine-tuned policy on 10 real-robot trials. Each trial receives the average of three binary sub-goal completion indicators: grasping, folding, and placing. We report the mean sub-goal completion score across the 10 trials. A trial score of 1.0 indicates that all three sub-goals were completed. Detailed scoring criteria are provided in Appendix.

Experimental results.

The experimental results provide compelling evidence for the efficacy of cross-embodiment pre-training. Upon fine-tuning, our pre-trained policy demonstrated superior performance, achieving a mean sub-goal completion score of 0.73 across 10 trials. In stark contrast, the direct fine-tuning baseline struggled to acquire effective behaviors, yielding a mean sub-goal completion score of 0.03 under the same evaluation protocol. The baseline frequently failed during the initial grasping phase, thereby precluding the successful execution of subsequent folding and placing sequences.



(a) Pre-training duration vs. mean sub-goal completion score on the Folding Pants task.

(b) Visual vs. behavioral diversity.

Figure 4: **Analysis of pre-training dynamics.** (a) Mean sub-goal completion score across 10 trials on the real-world "Folding Pants" task as a function of pre-training duration. (b) Six dataset configurations in VDS–JSD space; bubble size indicates success rate.

Table 1: **Impact of Pre-training Data Composition on Downstream Performance.** We report the success rate across six distinct downstream tasks.

Pre-training Strategy	Dump Bin Bigbin	Grab Roller	Handover Mic	Lift Pot	Place Shoe	Stack blocks Two	Avg.
Direct Fine-tuning	0.20	0.60	0.00	0.25	0.65	0.00	0.28
ARX-only	0.65	0.65	0.20	0.15	0.55	0.10	0.38
FR3-only	0.00	0.20	0.00	0.00	0.05	0.00	0.04
ARX+FR3	0.40	0.60	0.05	0.05	0.70	0.00	0.30

This substantial performance gap indicates that, despite the exclusion of the target Piper robot from the pre-training dataset, large-scale cross-embodiment data imparts essential motor primitives and physical priors. These transferred capabilities proved critical for enabling the model to master complex, long-horizon manipulation tasks.

4.1.2 Effect of Pre-training Duration

To investigate the effect of pre-training duration on transfer efficacy, we analyzed the performance on the "Folding Pants" task using models pre-trained for different numbers of steps. As shown in fig. 4a, the mean sub-goal completion scores increase with pre-training duration, indicating that the model gradually learns atomic skills.

With shorter pre-training, policies execute individual components of the task but often fail to complete the full sequence. With extended pre-training, policies begin to execute grasping, folding, and placing in sequence, completing the full task. This qualitative progression accompanies the increase in mean sub-goal completion scores.

4.2 Simulation Experiment

4.2.1 The Role of Morphological Similarity in Pre-training

Experimental setup. To investigate this impact, we designed a rigorous experiment in the RoboTwin 2.0 environment. RoboTwin 2.0 is designed specifically for bimanual manipulation and is capable of simulating diverse manipulator embodiments. This experiment compares the pre-training effectiveness of morphologically similar, dissimilar, and mixed data configurations against a direct fine-tuning baseline. We selected the Piper as our target embodiment and evaluated performance across six distinct downstream tasks with varying levels of difficulty. In RoboTwin, ARX, FR3, and Piper use the same number, positions, and types of cameras. Each robot uses its default control configuration provided by RoboTwin, without manual modifications. All models undergo multi-task fine-tuning on a combined dataset consisting of 50 clean expert demonstrations for each of the six tasks. We compare four distinct model configurations:

Table 2: **Kinematic distances to Piper.** All comparisons use the same 242 matched end-effector poses, selected from 512 candidates sampled using Piper's kinematics. Distances characterize single-arm kinematics with fixed tool offsets; collisions, gripper articulation, and dual-arm mounting geometry are excluded.

Source embodiment	Kinematic distance to Piper ↓
ARX	**2.773**
FR3	3.956

We compare four configurations: **(i) Direct Fine-tuning**, the baseline initialized from PaliGemma without robotic pre-training; **(ii) ARX-only**, pre-trained exclusively on ARX data; **(iii) FR3-only**, pre-trained exclusively on FR3 data; and **(iv) ARX+FR3**, pre-trained on a mixture of both.

All three pre-training configurations (ARX-only, FR3-only, ARX+FR3) use 80,000 optimization steps and the same batch size. The ARX+FR3 configuration samples ARX and FR3 data with equal probability (50% each). Under this fixed training budget, the mixture therefore receives approximately half as many ARX samples as ARX-only. Subsequently, all pre-trained checkpoints, as well as the direct fine-tuning baseline, are fine-tuned on the combined dataset of the six downstream tasks for a uniform 10,000 steps. All fine-tuned policies are then evaluated on these tasks for 20 rollouts under the clean condition. We report the Final Success Rate, defined as the average success rate over these 20 trials. Detailed hyperparameter settings for pre-training and fine-tuning, including learning rates, batch sizes, and training steps, are provided in Appendix B.

Quantifying kinematic similarity. To quantify one aspect of embodiment similarity, we compare local end-effector motion capabilities at matched end-effector poses using manipulability matrices (Jaquier et al., 2018). We co-locate the robot base origins while retaining their configured orientations. From each robot's URDF model, we compute the geometric Jacobian J and construct $M = \widetilde{J}\widetilde{J}^\top + \epsilon I$, where $\widetilde{J} = \mathrm{diag}(L^{-1}, L^{-1}, L^{-1}, 1, 1, 1)J$, $L = 0.5\,\mathrm{m}$, and $\epsilon = 10^{-6}$. We define the kinematic distance as

$$D(s,t) = \frac{1}{|\mathcal{P}|} \sum_{x \in \mathcal{P}} \left\| \log\Big(M_s(x)^{-1/2} M_t(x) M_s(x)^{-1/2} \Big) \right\|_F , \tag{1}$$

where $\mathcal{P}$ contains matched poses with valid inverse-kinematics solutions found for all compared robots. A smaller distance indicates more similar local motion capabilities under an equal joint-velocity norm budget. Under this protocol, ARX exhibits a smaller distance to Piper than either FR3 or UR5 (table 2). This ordering is preserved across the tested characteristic lengths and regularization strengths. Appendix D details the sampling and sensitivity analysis.

Results and analysis. Results are presented in table 1. Here, morphological similarity refers to the kinematic resemblance between embodiments, including degrees of freedom, joint configurations, and link proportions. ARX and Piper share a similar 6-DoF structure, whereas FR3 is a 7-DoF arm with kinematic redundancy and a fundamentally different joint topology.

The ARX-only model achieved the highest mean success rate (0.38), substantially outperforming the direct fine-tuning baseline (0.28), with gains most pronounced on challenging tasks (e.g., handover mic, dump bin bigbin). FR3-only pre-training reduces mean success from 0.28 to 0.04, demonstrating negative transfer under the shared joint-space action representation. ARX, which has a smaller kinematic distance to Piper, instead improves mean success to 0.38. The ARX+FR3 mixture achieves a mean success rate of 0.30, above the direct fine-tuning baseline (0.28) but below ARX-only (0.38). Thus, replacing half of the ARX sampling probability with FR3 data reduces transfer performance under a fixed training budget. This comparison measures the combined effect of changing the source-data composition and reducing ARX exposure; it does not isolate interference from FR3 data from the effect of fewer ARX samples.

4.2.2 Beyond Morphology: Unpacking the Impact of Data Diversity

Having established the transfer benefit of ARX pre-training, we next examine how scene, object, and task diversity affect downstream performance while keeping the source and target embodiments

Table 3: **Impact of pre-training data diversity.** Success rates on Piper across three diversity axes. Task diversity yields the most significant gains.

	Scene		Object		Task	
	Clean	Diverse	Less	More	Aligned	Diverse
Clean Eval	0.38	0.44	0.34	0.37	0.34	**0.48**
Rand. Eval	0.34	0.38	0.33	0.37	0.32	**0.38**

fixed. To answer this, we investigate the impact of scene, object, and task diversity on 10 bimanual manipulation tasks from RoboTwin 2.0.

Experimental setup. We designed a controlled study using ARX as the source embodiment and Piper as the target, maintaining a consistent data budget of approximately 5,000 trajectories across all settings. To systematically isolate distinct diversity dimensions, we designed three comparisons:

(i) Scene diversity: standard static background vs. domain-randomized scenes (varying lighting and textures), using additional data collected for this experiment, with the same 10 tasks and 500 trajectories per task in each condition; **(ii) Object diversity:** we recollected data in RoboTwin using limited and expanded sets of object instances within the same object categories. The expanded set contains twice as many distinct object instances as the limited set, with scene and task composition held constant; **(iii) Task diversity:** aligned tasks (10 target tasks only) vs. diverse tasks (47 tasks, including 37 unseen during fine-tuning), with both groups using the same total number of trajectories and the same scene-diversity setting throughout pre-training. The 37 additional tasks primarily combine existing manipulation skills with different objects or task goals.

Experimental results. Table 3 quantifies the impact of three diversity dimensions under a fixed pre-training data budget. The results indicate a clear hierarchy of influence. Visual diversity (scene and object) yielded only marginal gains, improving success rates by approximately 0.02 to 0.06. In contrast, scaling task diversity resulted in a substantial performance boost, raising the success rate in clean scenes from 0.34 to 0.48.

Regarding scene diversity, we observe a notable finding: although the model pre-trained on domain-randomized data had never encountered clean backgrounds during pre-training, it outperformed the model pre-trained exclusively on clean data when evaluated in clean scenes (0.38 vs. 0.44). This demonstrates that for cross-embodiment pre-training, leveraging randomized scenes is significantly more effective than relying on a single, static environment.

Furthermore, the task diversity experiments reveal a counter-intuitive trend. In the aligned task setting, pre-training tasks aligned perfectly with the downstream fine-tuning tasks, ensuring high data density per task. Conversely, the diverse task setting included numerous tasks unseen during fine-tuning. This reduces the demonstrations allocated to the target tasks while broadening the combinations of manipulation skills, objects, and task goals represented in pre-training. Despite this, the diverse task model achieved significantly better fine-tuning performance. This indicates that exposing the policy to a broader distribution of skills facilitates cross-embodiment transfer more effectively than simply increasing data repetition on specific tasks.

Experimental analysis: Decoupling visual and behavioral diversity. To understand why task diversity yields superior gains, we analyze pre-training datasets using two metrics: **Visual Diversity Score (VDS)**, defined as the average pairwise cosine distance between trajectory-level representations obtained by averaging frame embeddings from CLIP ViT-B/32 (Radford et al., 2021) within each trajectory, and **Joint Space Diversity (JSD)**, defined as the average pairwise dynamic time warping distance between trajectories of the 12 arm-joint positions, standardized using a common set of means and standard deviations across all six dataset configurations, with gripper channels excluded.

As illustrated in fig. 4b, VDS and JSD distinguish the visual and behavioral changes introduced by different data-diversity interventions. Specifically, compared to the clean scene, the diverse scene group exhibits a substantial surge in VDS yet shows stagnant JSD. The scene-diversity intervention produces smaller downstream gains than task diversification, indicating that a larger increase in visual diversity alone does not guarantee a larger transfer benefit. Conversely, the transition from

Table 4: **Experimental results for pre-training data quantity.** We report the average success rate for each task under clean and randomized settings, based on different pre-training fractions (25%, 50%, 75%, 100% of pre-training data).

Task	ARX-25%		ARX-50%		ARX-75%		ARX-100%	
	Clean	Randomized	Clean	Randomized	Clean	Randomized	Clean	Randomized
Dump Bin Bigbin	0.40	0.30	0.40	0.35	0.50	0.50	0.65	0.60
Grab Roller	0.60	0.65	0.60	0.60	0.80	0.75	0.75	0.60
Handover Mic	0.70	0.40	0.60	0.25	0.60	0.45	0.75	0.45
Lift Pot	0.15	0.25	0.45	0.20	0.25	0.25	0.85	0.4
Open Microwave	0.30	0.30	0.10	0.10	0.05	0.25	0.1	0.05
Place Empty Cup	0.35	0.20	0.40	0.25	0.30	0.15	0.55	0.15
Place Shoe	0.35	0.20	0.50	0.60	0.65	0.20	0.25	0.35
Press Stapler	0.45	0.60	0.40	0.55	0.30	0.50	0.30	0.70
Shake Bottle	0.80	0.55	0.75	0.55	0.75	0.45	0.85	0.40
Stack Blocks Two	0	0	0	0	0	0	0	0
Avg.	0.41	0.35	0.42	0.35	0.42	0.35	**0.51**	**0.37**

aligned task to diverse task yields a significant leap in JSD despite minimal fluctuation in VDS. The higher JSD of the diverse-task dataset accompanies stronger downstream performance, highlighting behavioral coverage as an important consideration for pre-training data curation. Task diversity therefore offers a practical way to broaden the joint-motion patterns represented in a fixed data budget.

4.2.3 Impact of Pre-training Data Quantity on Transfer Performance

Experimental setup. To test whether scaling morphologically similar data improves transfer, we subsample the ARX-only dataset into four scales (25%, 50%, 75%, 100%) and pre-train models for 80,000 steps each. We evaluate on 10 downstream tasks from RoboTwin 2.0, fine-tuning each model with 50 clean expert demonstrations for 10,000 steps and reporting success rates over 20 rollouts.

Results and analysis. table 4 details the performance scaling across 10 downstream tasks as we increase the ARX pre-training data volume from 25% to 100%. Quantitative results reveal a clear positive correlation: scaling the data yield a improvement in average success rates, rising from 0.41 to 0.51 in the clean setting and from 0.35 to 0.37 in the randomized setting. These findings validate that the "More is Better" principle holds true, provided that the pre-training data is morphologically aligned with the downstream embodiment.

5 Limitations

Our study suggests several directions for future work. The current experiments focus on a controlled setting with two source embodiments, ARX, FR3, and one target platform, Piper. This design enables a focused analysis of data and embodiment factors, while broader coverage across additional robot families would further test the generality of the observed trends. We also use a single VLA backbone, π_0, to avoid conflating data effects with architectural differences; evaluating other VLA and policy architectures is an important next step.

6 Conclusion

In this work, we challenge the "naive data scaling" assumption by dissecting the determinants of cross-embodiment transfer. Our findings highlight the importance of considering kinematic similarity when selecting source embodiments: ARX pre-training improves transfer to Piper, whereas FR3 pre-training yields negative transfer under the shared joint-space action representation. Furthermore, we reveal that task breadth outperforms visual diversity in facilitating cross-embodiment transfer. Finally, we confirm that under morphological alignment, scaling data size boosts performance, while complex skills emerge after extended training. Consequently, we advocate for a paradigm shift from blind scaling to "embodiment-aware data curation", prioritizing morphological similarity and task diversity to build efficient generalist policies.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools to assist with writing and language editing, including refining phrasing, correcting grammar, improving clarity, and organizing the text. All AI-assisted revisions included in the manuscript were reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript, including its scientific claims, experimental results, and references.

REFERENCES

Haruki Abe, Takayuki Osa, Yusuke Mukuta, and Tatsuya Harada. Cross-embodiment offline reinforcement learning for heterogeneous robot datasets. *arXiv preprint arXiv:2602.18025*, 2026.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 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*, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Noémie Jaquier, Leonel Rozo, Darwin G. Caldwell, and Sylvain Calinon. Geometry-aware manipulability learning, tracking and transfer. *arXiv preprint arXiv:1811.11050*, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

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.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27649–27660, 2025.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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.

Zhongyi Zhou, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Vision-language-action model with open-world embodied reasoning from pretrained knowledge. *arXiv preprint arXiv:2505.21906*, 2025.

A Experimental Setup Details

A.1 Real-World Setup

Camera configuration. Our visual setup consists of a top-down RealSense D405 camera for a global workspace view, alongside two RealSense D435i cameras mounted on the left and right wrists for close-up manipulation observation.

Task description. We evaluate cross-embodiment transfer using a long-horizon "Folding Pants" task. Unlike simple pick-and-place tasks, this activity requires precise bimanual coordination and sequential planning.

Scoring criteria. We evaluate each fine-tuned policy on 10 real-robot trials under identical settings. Instead of a sparse binary success signal, we adopt a fine-grained metric by decomposing the long-horizon task into three sequential sub-goals, each evaluated on a binary basis. For trial i, the score is $s_i = (g_i + f_i + p_i)/3$, where g_i, f_i, and p_i indicate completion of grasping, folding, and placing, respectively. We report $\bar{s} = \frac{1}{10}\sum_{i=1}^{10} s_i$. The reported value of 0.73 is this mean sub-goal completion score, not the proportion of trials completing all three sub-goals. The three stages are:

- **Grasping:** The robot successfully grasps the two ends of the pants and lifts the garment completely off the table surface, ensuring a stable grip without slippage.
- **Folding:** First, the robot grasps the top and bottom corners of one side and folds them inward. Second, the right arm grasps the waistband area and executes a final fold towards the left. Success requires both folding actions to be completed.
- **Placing:** The robot successfully places the folded pants onto the right side of the table.

A.2 Simulation Setup



Figure 5: **RoboTwin 2.0 Evaluation Tasks.** The set of 10 bimanual manipulation tasks used for the simulation benchmark.

Diversity experiment details. All models were pre-trained for a uniform 80,000 steps and fine-tuned on the Piper robot (using a mix of single and diverse scene data, 100 trajectories per task) for 10,000 steps. We evaluated each fine-tuned policy over 40 trials per task under each of the clean and randomized conditions, totaling 80 trials per task across the 10 target tasks. The three comparison settings are:

- **Scene diversity:** We compared a model pre-trained on data with a standard static background against one pre-trained on domain-randomized scenes (varying lighting and textures). The datasets for this comparison were collected in a subsequent data collection stage, separately from the official dataset described in Section 3.2. Both datasets consisted of the same 10 tasks, with 500 trajectories per task under each scene condition.
- **Object diversity:** We recollected data in RoboTwin using two sets of object instances from the same object categories. The diverse set contained twice as many distinct instances as the limited set. Both datasets covered the same 10 tasks and included both clean and randomized scenes, with scene and task composition held constant.

- **Task diversity:** We compared an aligned task dataset (containing only the 10 target tasks) against a diverse task dataset (containing 47 tasks, including 37 unseen during fine-tuning). The total data volume was strictly controlled at 5,000 trajectories and scene diversity was held constant for both groups, to decouple task diversity from data quantity and visual variations. The 37 additional tasks primarily combine existing manipulation skills with different objects or task goals.

B Training Hyperparameters

All models were initialized with PaliGemma pre-trained weights. We used a unified set of hyperparameters within each training stage to ensure fair comparisons. table 5 summarize the settings for real-world pre-training, simulation pre-training, and fine-tuning, respectively.

Table 5: **Training Hyperparameters.** Settings for real-world pre-training, simulation pre-training, and fine-tuning.

Hyperparameter	Real-world Pre-train	Sim Pre-train	Fine-tuning
Optimizer	AdamW	AdamW	AdamW
Learning Rate	2e-5	5e-5	2e-5
LR Schedule	constant	constant	cosine
Batch Size	96	96	96
Total Training Steps	2,400,000	80,000	10,000 / 40,000
Action Output Space	Joint Positions	Joint Positions	Joint Positions
Action Chunking Size	50	50	50

Fine-tuning uses 40,000 steps for the real-world experiments and 10,000 steps for all simulation experiments.

C Per-Task Experimental Results

We present detailed per-task success rates for the diversity experiments in Section 4.2.2. table 6 compares Scene Diversity settings, table 7 examines Object Diversity, and table 8 details Task Diversity.

D Kinematic Similarity Analysis

Models and coordinate conventions. We use the RoboTwin URDF models and embodiment configurations for ARX, FR3, UR5, and Piper; ARX uses the ARX-X5 model. Base translations are removed and configured base rotations retained. We construct the joint anchor frame and apply the configured end-effector rotation corrections and gripper offset. The comparison includes the arm joints and a fixed tool offset. All arms use the same Euclidean joint-velocity norm budget, without weighting by model-specific velocity limits.

Pose sampling and inverse kinematics. We generate scrambled Sobol samples (seed 20260922) over the central 90% of each Piper joint's configured range. We retain the first 512 end-effector poses within the base-relative region $|x| \leq 0.35\,\mathrm{m}$, $0.10 \leq y \leq 0.65\,\mathrm{m}$, and $0.02 \leq z \leq 0.60\,\mathrm{m}$. This is a joint-space sample conditioned on a workspace region, rather than a sample from task demonstrations. For each robot and pose, bounded nonlinear least squares uses 16 initial configurations, including the configured home pose, with at most 160 function evaluations per start. Random initial angles lie in the intersection of the URDF limits and $[-\pi, \pi]$; optimization uses the full URDF limits. Solutions must have position error below $10^{-4}\,\mathrm{m}$ and rotation error below $10^{-3}\,\mathrm{rad}$. Among accepted solutions, we select the one closest to the configured home pose in wrapped joint-angle root-mean-square distance.

Valid solutions are found for 508, 324, 305, and 509 of the 512 candidates for ARX, FR3, UR5, and Piper, respectively. All comparisons use the intersection of 242 poses. Failure to find a solution

Table 6: **Impact of Scene Diversity.** Success rates of policies pre-trained on Clean vs. Diverse scenes, evaluated under both clean and randomized settings.

	Eval on Clean Scene		Eval on Diverse Scene	
Task	**Pre-train: Clean**	**Pre-train: Random**	**Pre-train: Clean**	**Pre-train: Random**
Dump Bin Bigbin	0.60	0.60	0.30	0.80
Handover Mic	0.30	0.20	0.10	0.10
Place Empty Cup	0.10	0.60	0.00	0.20
Stack Blocks Two	0.10	0.20	0.20	0.20
Place Shoe	0.70	0.80	0.70	0.60
Open Microwave	0.00	0.00	0.00	0.00
Grab Roller	0.70	0.70	0.70	0.50
Press Stapler	0.50	0.30	0.50	0.70
Lift Pot	0.20	0.00	0.20	0.00
Shake Bottle	0.60	1.00	0.70	0.70
Average	**0.38**	**0.44**	**0.34**	**0.38**

Table 7: **Impact of Object Diversity.** Success rates comparison between policies pre-trained with limited (Less Obj) vs. diverse (More Obj) objects.

	Eval on Clean Scene		Eval on Random Scene	
Task	**Pre-train: Less Obj**	**Pre-train: More Obj**	**Pre-train: Less Obj**	**Pre-train: More Obj**
Dump Bin Bigbin	0.55	0.70	0.30	0.60
Handover Mic	0.15	0.25	0.05	0.10
Place Empty Cup	0.15	0.15	0.20	0.10
Move Can Pot	0.00	0.00	0.00	0.00
Place Shoe	0.45	0.35	0.45	0.40
Open Microwave	0.35	0.45	0.10	0.50
Grab Roller	0.50	0.65	0.60	0.75
Press Stapler	0.30	0.40	0.40	0.45
Lift Pot	0.25	0.05	0.40	0.00
Shake Bottle	0.65	0.65	0.75	0.75
Average	**0.335**	**0.365**	**0.325**	**0.365**

Table 8: **Impact of Task Diversity.** Success rates comparison when scaling the breadth of pre-training tasks (Less Task vs. More Task).

	Eval on Clean Scene		Eval on Random Scene	
Task	**Pre-train: Less Task**	**Pre-train: More Task**	**Pre-train: Less Task**	**Pre-train: More Task**
Dump Bin Bigbin	0.45	0.85	0.20	0.50
Handover Mic	0.00	0.05	0.05	0.05
Place Empty Cup	0.15	0.45	0.20	0.20
Stack Blocks Two	0.05	0.00	0.00	0.05
Place Shoe	0.55	0.50	0.50	0.50
Open Microwave	0.00	0.00	0.10	0.00
Grab Roller	0.60	0.65	0.60	0.60
Press Stapler	0.70	0.55	0.40	0.40
Lift Pot	0.20	0.80	0.40	0.70
Shake Bottle	0.65	0.95	0.75	0.75
Average	**0.335**	**0.48**	**0.32**	**0.375**

is a numerical outcome rather than a proof of unreachability. The analytic Jacobians and inverse-kinematics residual derivatives agree with central finite differences to within 10^{-9} in maximum absolute error.

Uncertainty and sensitivity. Bootstrapping the 242 common poses with 3,000 resamples gives 95% intervals of $[2.596, 2.959]$, $[3.795, 4.126]$, and $[3.813, 4.243]$ for ARX, FR3, and UR5, respectively. These intervals describe variation across sampled poses. Table 9 reports sensitivity to the characteristic length, regularization, and Jacobian components. Averaging over all pairs of distinct

IK solutions found at each pose, followed by averaging across poses, gives distances of 3.365, 3.977, and 4.074, respectively. ARX has the smallest mean distance in each of these comparisons.

Table 9: Sensitivity of mean kinematic distance to Piper. Translation-only and rotation-only distances use the corresponding three rows of the Jacobian.

Setting	ARX	FR3	UR5
$L = 0.5\,\mathrm{m},\ \epsilon = 10^{-6}$	2.773	3.956	4.023
$L = 0.25\,\mathrm{m},\ \epsilon = 10^{-6}$	2.776	3.958	4.026
$L = 1.0\,\mathrm{m},\ \epsilon = 10^{-6}$	2.766	3.949	4.016
$L = 0.5\,\mathrm{m},\ \epsilon = 10^{-8}$	2.784	3.967	4.034
$L = 0.5\,\mathrm{m},\ \epsilon = 10^{-4}$	2.706	3.891	3.950
Translation only	0.497	2.184	1.282
Rotation only	0.876	0.941	1.283