Humanoid Policy ~ Human Policy

Ri-Zhao Qiu[*,1] **Shiqi Yang**[*,1] **Xuxin Cheng**[*,1] **Chaitanya Chawla**[*,2] **Jialong Li**[1]
Tairan He[2] **Ge Yan**[4] **David Yoon**[3] **Ryan Hoque**[3] **Lars Paulsen**[1]
Ge Yang[5] **Jian Zhang**[3] **Sha Yi**[1] **Guanya Shi**[2] **Xiaolong Wang**[1]
[1] UC San Diego, [2] CMU, [3] Apple, [4] University of Washington, [5] MIT
https://human-as-robot.github.io/

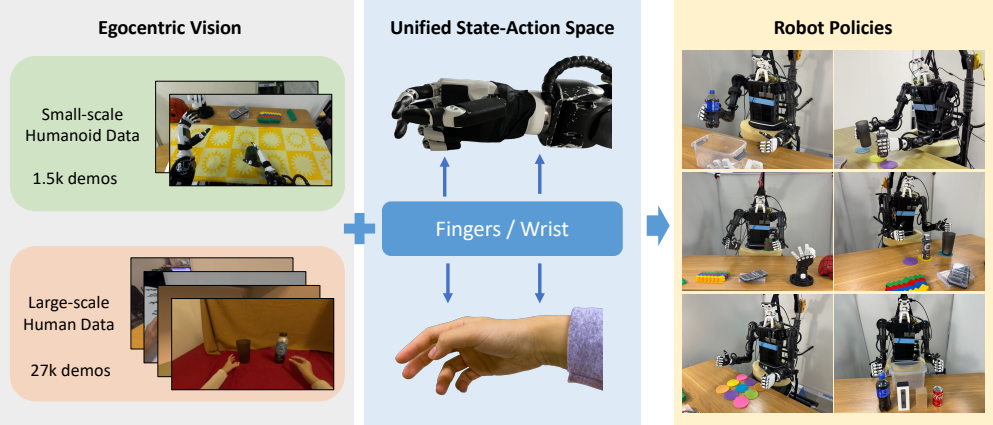


Figure 1: This paper advocates high-quality human data as a data source for cross-embodiment learning - **task-oriented** egocentric human data. We collect a large-scale dataset, **P**hysical **H**uman-**H**umanoid **D**ata (**PH^{2}D**), with hand-finger 3D poses from consumer-grade VR devices on well-defined manipulation tasks directly aligned with robots. Without relying on modular perception, we train a Human Action Transformer (**HAT**) manipulation policy by directly modeling humans as a different humanoid embodiment in an end-to-end manner.

Abstract: Training manipulation policies for humanoid robots with diverse data enhances their robustness and generalization across tasks and platforms. However, learning solely from robot demonstrations is labor-intensive, requiring expensive tele-operated data collection, which is difficult to scale. This paper investigates a more scalable data source, egocentric human demonstrations, to serve as cross-embodiment training data for robot learning. We mitigate the embodiment gap between humanoids and humans from both the data and modeling perspectives. We collect an egocentric **task-oriented** dataset (**PH^{2}D**) that is directly aligned with humanoid manipulation demonstrations. We then train a human-humanoid behavior policy, which we term Human Action Transformer (**HAT**). The state-action space of HAT is unified for both humans and humanoid robots and can be differentiably retargeted to robot actions. Co-trained with smaller-scale robot data, HAT directly models humanoid robots and humans as different embodiments without additional supervision. We show that human data improve both generalization and robustness of HAT with significantly better data collection efficiency.

Keywords: Robot Manipulation, Cross-Embodiment, Humanoid

1 Introduction

Learning from real robot demonstrations has led to great progress in robotic manipulation recently [1, 2, 3, 4]. One key advancement to enable such progress was hardware / software co-designs to scale up data collection using teleoperation [5, 6, 7, 8, 9, 10] and directly controlling

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

the robot end effector [11, 12, 5, 6, 13, 7]. Instead of gathering data on a single robot, collective efforts have been made to merge diverse robot data and train foundational policies across embodiments [11, 14, 2, 1, 3, 4], which have shown to improve cross-embodiment and cross-task generalizability.

However, collecting structured real-robot data is expensive and time-consuming. We are still far away from building a robust and generalizable model as what has been achieved in Computer Vision [17] and NLP [18]. If we examine humanoid robot teleoperation more closely, it involves robots mimicking human actions using geometric transforms or retargeting to control robot joints and end-effectors. From this perspective, **we propose to model robots in a human-centric representation**, and the robot action is just a transformation away from the human action. If we can accurately capture the end-effector and head poses of humans, egocentric human demonstrations will be a more scalable source of training data, as we can collect them efficiently, in any place, and without a robot.

Dataset	Human		Robot	
	# Frames	# Demos	# Frames	# Demos
DexCap [15]	∼378k	787	NA	NA
EgoMimic [16]	∼432k†	2,150	**1.29M**†	1,000
PH^{2}D (Ours)	**∼3.02M**	**26,824**	∼668k	**1,552**

Table 1: **Comparisons of task-oriented egocentric human datasets.** Besides having the most demonstrations, PH^{2}D is collected on various manipulation tasks, diverse objects and scenes, with accurate 3D hand-finger poses and language annotations. †: estimated based on reported data collection time with 30 Hz; whereas DexCap [15] and PH^{2}D report processed frames for training.

In this paper, we perform cross-human and humanoid embodiment training for robotic manipulation. Our key insight is to model bimanual humanoid behaviors by *directly imitating human behaviors without using learning surrogates* such as affordances [19, 20]. To realize this, we first collect an egocentric task-oriented dataset of **P**hysical **H**umanoid-**H**uman **D**ata, dubbed PH^{2}D. We adapt consumer-grade VR devices to collect egocentric videos with automatic but accurate hand pose and end effector (*i.e.,* hand) annotations. Compared to existing human daily behavior datasets [21, 22], PH^{2}D is task-oriented so that it can be directly used for co-training. The same VR hardwares are then used to perform teleoperation to collect smaller-scale humanoid data for better alignment. We then train a Human-humanoid Action Transformer (HAT), which predicts future hand-finger trajectories in a unified human-centric state-action representation space. To obtain robot actions, we simply apply inverse kinematics and hand retargeting to differentiably convert human actions to robot actions for deployment.

We conduct real-robot evaluations on different manipulation tasks with extensive ablation studies to investigate how to best align human and humanoid demonstrations. In particular, we found that co-training with diverse human data improves robustness against spatial variance and background perturbation, generalizing in settings unseen in robot data but seen in human data. We believe that these findings highlight the potential of using human data for large-scale cross-embodiment learning.

In summary, our contributions are:

- **A dataset**, PH^{2}D, which is a large egocentric, task-oriented human-humanoid dataset with accurate hand and wrist poses for modeling human behavior (see Tab. 1).
- **A cross human-humanoid manipulation policy**, HAT, that introduces a unified state-action space and other alignment techniques for humanoid manipulation.
- **Improved policy robustness and generalization** validated by extensive experiments and ablation studies to show the benefits of co-training with human data.

2 Related Work

Imitation Learning for Robot Manipulation. Recently, learning robot policy with data gathered directly from the multiple and target robot embodiment has shown impressive robustness and dex-

terity [23, 2, 24, 1, 25, 26, 9, 27, 28]. The scale of data for imitation learning has grown substantially with recent advancements in data collection [29, 9, 7, 8], where human operators can efficiently collect large amounts of high-quality, task-oriented data. Despite these advances, achieving open-world generalization still remains a significant challenge due to lack of internet-scale training data.

Learning from Human Videos. Learning policies from human videos is a long-standing topic in both computer vision and robotics due to the vast existence of human data. Existing works can be approximately divided into two categories: aligning observations or actions.

Learn from Human - Aligning Observations. While teleoperating the actual robot platform allows learning policy with great dexterity, there is still a long way to go to achieve higher levels of generalization across diverse tasks, environments, and platforms. Unlike fields such as computer vision [17] and natural language processing [18] benefiting from internet-scale data, robot data collection in the real world is far more constrained. Various approaches have attempted to use internet-scale human videos to train robot policies [30, 31, 32, 33, 34, 35]. Due to various discrepancies (*e.g.,* supervision and viewpoints) between egocentric robot views and internet videos, most existing work [19, 20] use modular approaches with intermediate representations as surrogates for training. The most representative ones are affordances [19, 20] for object interaction, object keypoints predictions [36, 37, 38, 39, 40], or other types of object representations [41, 42, 43].

Learn from Human - Aligning Actions. Beyond observation alignment, transferring human demonstrations to robotic platforms introduces additional challenges due to differences in embodiment, actuation, and control dynamics. Specific alignment of human and robot actions is required to overcome these disparities. Approaches have employed masking in egocentric views [16], aligning motion trajectories or flow [44, 45], object-centric actions [46, 47], or hand tracking with specialized hardware [15]. Most closely related to our work, HumanPlus [48] designs a remapping method from 3D human pose estimation to tele-operate humanoid robots. Compared to HumanPlus, the insight of our method is to waive the requirement for robot hardware in collecting human data and collect diverse human data directly for co-training. In contrast to HumanPlus, we intentionally avoid performing retargeting on human demonstrations and designed the policy to directly use human hand poses as states/actions. On the other hand, the 'human shadowing' retargeting in HumanPlus is a teleoperation method that still requires robots, leading to lower collection efficiency than ours.

Cross-Embodiment. Cross-embodiment pre-training has been shown to improve adaptability and generalization over different embodiments [49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]. When utilizing human videos, introducing intermediate representations can be prone to composite errors. Recent works investigate end-to-end approaches [2, 24, 1, 3] using cross-embodied robot data to reduce such compounding perceptive errors. Noticeably, these works have found that such end-to-end learning leads to desired behaviors such as retrying [3]. Some other work [62, 38] enforces viewpoint constraints between training human demonstrations and test-time robot deployment to allow learning on human data but it trades off the scalability of the data collection process.

Concurrent Work. Some concurrent work [15, 16, 63] also attempts to use egocentric human demonstrations for end-to-end cross-embodiment policy learning. DexCap [15] uses gloves to track 3D hand poses with a chest-mounted RGBD camera to capture egocentric human videos. However, DexCap relies on 3D inputs, whereas some recent works [3, 1] have shown the scalability of 2D visual inputs. Most related to our work, EgoMimic [16] also proposes to collect data using wearable device [64] with 2D visual inputs. However, EgoMimic requires strict visual sensor alignments; whereas we show that scaling up diverse observations with different cameras makes the policy more robust. In addition, PH^2D is also greater in dataset scale and object diversity. We also show our policy can be deployed on real robots without strict requirements of visual sensors and heuristics, which paves the way for scalable data collection.

3 Method

To collect more data to train generalizable robot policies, recent research has explored cross-embodiment learning, enabling policies to generalize across diverse physical forms [3, 1, 4, 2, 65, 14]. This paper proposes egocentric human manipulation demonstrations as a scalable source of cross-embodiment training data. Sec. 3.1 describes our approach to adapt consumer-grade VR devices to scale up human data collection conveniently for a dataset of task-oriented egocentric human demonstrations. Sec. 3.2 describes various techniques to handle domain gaps to align human data and robot data for learning humanoid manipulation policy.

3.1 PH^{2}D: Task-oriented Physical Humanoid-Human Data

Though there has been existing work that collects egocentric human videos [16, 22, 21, 15], they either (1) provide demonstrations mostly for non-task-oriented skills (*e.g., dancing*) and do not provide world-frame 3D head and hand poses estimations for imitation learning supervision [21, 22] or (2) require specialized hardware or robot setups [15, 16].

To address these issues, we propose PH^{2}D. PH^{2}D address these two issues by (1) collecting task-oriented human demonstrations that are directly related to robot execution, (2) adapting well-engineered SDKs of VR devices (illustrated in Fig. 2) to provide supervision, and (3) diversifying tasks, camera sensors, and reducing whole-body movement to reduce domain gaps in both vision and behaviors.

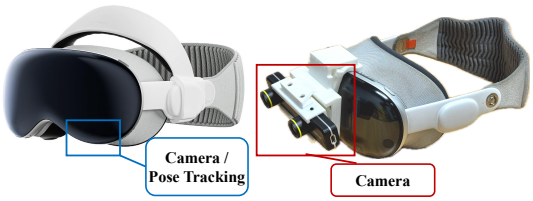


Figure 2: **Consumer-grade Devices for Data Collection.** To avoid relying on specialized hardware for data collection to make our method scalable, we design our data collection process using consumer-grade VR devices.

Adapting Low-cost Commerical Devices With development in pose estimation [66] and system engineering, modern mobile devices are capable of providing accurate on-device world frame 3D head pose tracking and 3D hand keypoint tracking [9], which has proved to be stable enough to teleoperate robot in real-time [9, 13]. We design software and hardware to support convenient data collection across different devices. Different cameras provide better visual diversity.

- **Apple Vision Pro + Built-in Camera.** We developed a Vision OS App that uses the built-in camera for visual observation and uses the Apple ARKit for 3D head and hand poses.
- **Meta Quest 3 / Apple Vision Pro + ZED Camera.** We developed a web-based application based on OpenTelevision [9] to gather 3D head and hand poses. We also designed a 3D-printed holder to mount ZED Mini Stereo cameras on these devices. This configuration is both low-cost ($<$700\$) and introduces more diversity with stereo cameras.

Data Collection Pipeline We collect task-oriented egocentric human demonstrations by asking human operators to perform tasks overlapping with robot execution (*e.g.,* grasping and pouring) when wearing the VR devices. For every demonstration, we provide language instructions (*e.g., grasp a can of coke zero with right hand*), and synchronize proprioception inputs and visual inputs by closest timestamps.

Action Domain Gap. Human actions and tele-operated robot actions exhibit two distinct characteristics: (1) human manipulation usually involves involuntary whole-body movement, and (2) humans are more dexterous than robots and have significantly faster task completion time than robots. We mitigate the first gap by requesting the human data collectors to sit in an upright position. For the second speed gap, we interpolate translation and rotations of human data during training (effectively 'slowing down' actions). The slow-down factors α_{slow} are obtained by normalizing the average task completion time of humans and humanoids, which is empirically distributed around 4. For consistency, we use $\alpha_{\text{slow}} = 4$ in all tasks.

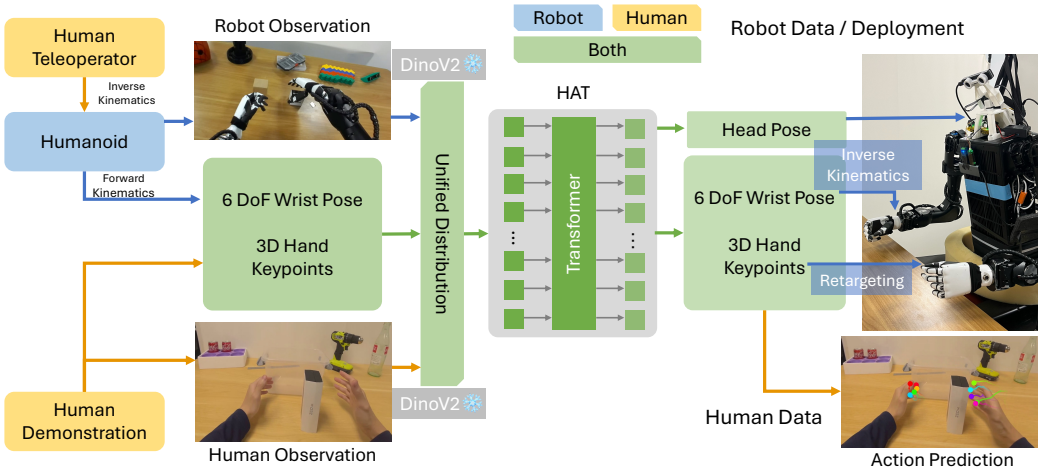


Figure 3: **Overview of HAT.** Human Action Transformer (HAT) learns a robot policy by modeling humans. During training, we sample a state-action pair from either human data or robot data. The images are encoded by a frozen DinoV2 encoder [67]. The HAT model makes predictions in a human-centric action space using wrist 6 DoF poses and finger tips, which is retargeted to robot poses during real-robot deployment.

3.2 HAT: Human Action Transformer

HAT learns cross-embodied robot policy by modeling humans. We demonstrate that treating bimanual humanoid robots and humans as different robot embodiments via retargeting improves both generalizability and robustness of HAT.

More concretely, let $\mathcal{D}_{robot} = \{(\mathbf{S}_i, \mathbf{A}_i)\}_{i=1}^{N}$ be the set of data collected from real bimanual humanoid robots using teleoperation [9], where $\mathbf{S}_i$ is the states including proprioceptive and visual observations of i-th demonstration and $\mathbf{A}_i$ be the actions. The collected PH^{2}D dataset, $\mathcal{D}_{human} = \{(\tilde{\mathbf{S}}_i, \tilde{\mathbf{A}}_i)\}_{i=1}^{M}$ is used to augment the training process. Note that it is reasonable to assume $M \gg N$ due to the significantly better human data collection efficiency.

The goal is to design a policy $\pi : \mathbf{S} \rightarrow \mathbf{A}$ that predicts future robot actions $\mathbf{a}_t$ given current robot observation $\mathbf{s}_t$ at time t, where the future actions $\mathbf{a}_{t+1}$ is usually a chunk of actions for multi-step execution (with slight abuse of notation). We model π using HAT, which is a transformer-based architecture predicting action chunks [5]. The overview of the model is illustrated in Fig. 3. We discuss key design choices of HAT with experimental ablations.

Unified State-Action Space. Both bimanual robots and humans have two end effectors. In our case, our robots are also equipped with an actuated 2DoF neck that can rotate, which resembles the autonomous head movement when humans perform manipulation. Therefore, we design a unified state-action space (*i.e.,* $(\mathbf{S}, \mathbf{A}) \equiv (\tilde{\mathbf{S}}, \tilde{\mathbf{A}})$) for both bimanual robots and humans. More concretely, the proprioceptive observation is a 54-dimensional vector (6D rotations [68] of the head, left wrist, and right wrist; x/y/z of left and right wrists and 10 finger tips). In this work, since we deploy our policy on robots with 5-fingered dexterous hands (shown in Fig. 4), there exists a bijective mapping between the finger tips of robot hands and human hands. Note that injective mapping is also possible (*e.g.,* mapping distance between the thumb finger and other fingers to parallel gripper distance).

Visual Domain Gap. Two types of domain gaps exist for co-training on human/humanoid data: camera sensors and end effector appearance. Since our human data collection process includes cameras different from robot deployment, this leads to camera domain gaps such as tones. Also, the appearances of human and humanoid end effectors are different. However, with sufficiently large and diverse data, we find it not a strict necessity to apply heuristic processing such as visual artifacts [16] or generative methods [69] to train human-robot policies - basic image augmentations such as color jittering and Gaussian blurring are effective regularization.

Meth.	H. Data	D. Norm	Passing		Horizontal Grasp		Vertical Grasp		Pouring		Ovr. Succ.	
			I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.	I.D.	O.O.D.
ACT	✗	NA	19/20	36/60	8/10	7/30	7/20	15/70	**8/10**	1/10	42/60	59/170
HAT	✓	✗	17/20	51/60	**9/10**	11/30	**14/20**	**30/70**	5/10	5/10	45/60	97/170
HAT	✓	✓	**20/20**	**52/60**	8/10	**12/30**	13/20	29/70	**8/10**	**8/10**	**49/60**	**101/170**
Type of Generalization			Background		Texture		Obj. Placement					

Table 2: **Success rate of autonomous skill execution**. Co-training with human data (H. Data) significantly improves the Out-Of-Distribution (O.O.D.) performance with nearly 100% relative improvement on all tasks on Humanoid A. We also ablate the design choice of using different normalizations (D. Norm) for different embodiments. We designate each task setting to investigate a single type of generalization. Detailed analysis of each type of generalization is presented in Sec. C.

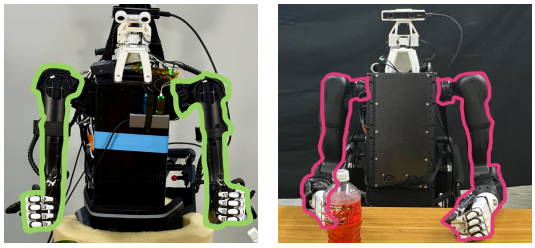
Training. The final policy is denoted as $\pi : f_\theta(\cdot) \rightarrow \mathbf{A}$ for both human and robot policy, where f_θ is a transformer-based neural network parametrized by θ. The final loss is given by,

$$\mathcal{L} = \ell_1(\pi(s_i), a_i) + \lambda \cdot \ell_1(\pi(s_i)_{\text{EEF}}, a_{i,\text{EEF}}) \,, \tag{1}$$

where EEF are the indices of the translation vectors of the left and right wrists, and $\lambda = 2$ is an (insensitive) hyperparameter used to balance loss to emphasize the importance of end effector positions over learning unnecessarily precise finger tip keypoints.

4 Experiments

Hardware Platforms. We run our experiments on two humanoid robots (Humanoid A and Humanoid B shown in Fig. 4) equipped with 6-DOF Inspire dexterous hands. Humanoid A is a Unitree H1 robot and Humanoid B is a Unitree H1_2 robot with different arm configurations. Similar to humans, both robots (1) are equipped with actuated necks [9] to get make use of egocentric views and (2) do not have wrist cameras. Unless otherwise noted, most humanoid data collection is done with Humanoid A. We use Humanoid B mainly for testing cross-humanoid generalization.



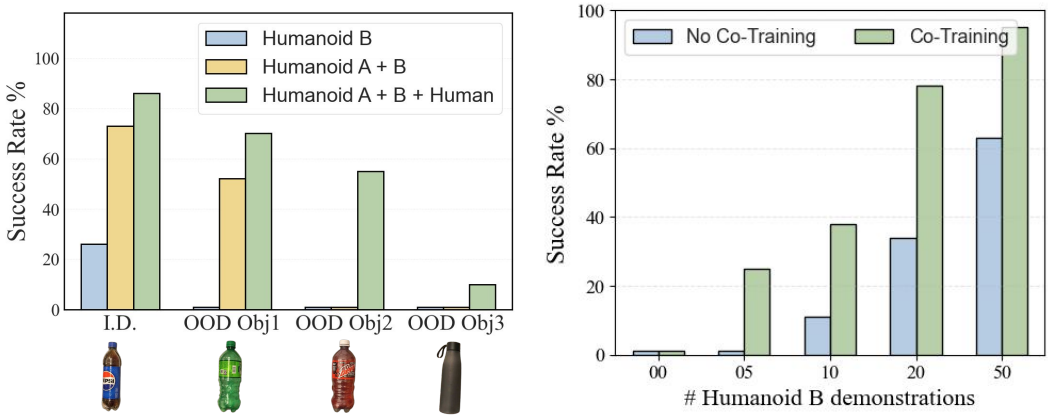
(a) Humanoid A (b) Humanoid B

Figure 4: **Hardware Illustration.** Most robot data attributes to Humanoid A, a Unitree H1 robot. Humanoid B, a Unitree H1-2 robot with **different** arm motor configurations, is used to evaluate few-shot cross-humanoid transfer. Detailed comparisons in Sec. D

Implementation Details. We implement policy architecture by adopting an transformer-based architecture predicting future action chunks [5]. We use a frozen DinoV2 ViT-S [67] as the visual backbone. We implement two variants: (1) **ACT**: baseline implementation using the Action Chunk Transformer [5], trained using only robot data. Robot states are represented as joint positions. (2) **HAT**: same architecture as ACT, but the state encoder operates in the unified state-action space. Unless otherwise stated, HAT is co-trained on robot and human data. A checkpoint is trained for each task with approximately 250-400 robot demonstrations.

Experimental Protocol. We collect robot and human demonstrations in different object sets. Since human demonstrations are easier to collect, the settings in human demonstrations are generally more diverse, which include background, object types, object positions, and the relative position of the human to the table.

We experimented with four different dexterous manipulation tasks and investigated in-distribution and out-of-distribution setups. The *in-distribution (I.D.)* setting tests the learned skills with backgrounds and object arrangements approximately similar to the training demonstrations presented in the real-robot data. In the Out-Of-Distribution (O.O.D.) setting, we test generalizability and robustness by introducing novel setups that were presented in human data but not in robot data. Fig. 7 visualizes different manipulation tasks and how we define out-of-distribution settings for each task.



(a) Performance of Humanoid B co-trained with PH^2D on horizontal grasping. o1 is seen by Humanoid B. o2 and o3 seen in human data. o4 is unseen in all data.

(b) Co-training consistently outperforms isolated training as Humanoid B demonstrations increase, achieving good success rates even in low-data regimes.

Figure 5: **Few-Shot Adaptation.** Co-training consistently outperforms isolated training as Humanoid B demonstrations increase, achieving robust success rates even in low-data regimes.

4.1 Main Evaluation

Human data has minor effects on I.D. testing. From Tab. 2, we can see that I.D. performance with or without co-training with human data gives similar results. In the I.D. setting, we closely match the scene setups as training demonstrations, including both background, object types, and object placements. Thus, policies trained with only a small amount of Humanoid A data performed well in this setting. This finding is consistent with recent work [9, 7] that frozen visual foundation models [17, 67] improve robustness against certain external perturbations such as lighting.

Human data improves the O.O.D. settings with many generalizations. One common challenge in imitation learning is overfitting to only in-distribution task settings. Hence, it is crucial for a robot policy to generalize beyond the scene setups seen in a limited set of single-embodiment data. To demonstrate how co-training with human data reduces such overfitting, we introduce O.O.D. task settings to evaluate such generalization. From Tab. 2, we can see that co-training drastically improves O.O.D. settings, achieving nearly 100% relative improvement in settings unseen by the robot data. In particular, we find that human data improves three types of generalization: **background, object placement, and appearance**. To isolate the effect of each variable, each task focuses on a specific type of generalization as listed in Tab. 2, with in-depth analyses in Sec. C.

4.2 Few-Shot Transfer across Heterogenous Embodiments

We conducted few-shot generalization experiments on a distinct humanoid platform (Humanoid B), contrasting it with our primary platform, Humanoid A. Notably, Humanoid B's demonstration data were collected in an entirely separate environment, introducing both embodiment and environmental shifts. We highlight two key advantages of our approach: (1) the ability to unify heterogeneous human-centric data sources (humanoids and humans) into a generalizable policy framework, and (2) the capacity to rapidly adapt to new embodiments with drastically reduced data requirements.

Experiment 1: Cross-embodiment co-training efficacy Using only 20 demonstrations from Humanoid B, we trained 3 policies - respectively on data from (i) Humanoid B only, (ii) Humanoid B + Humanoid A (cross-embodiment), and (iii) Humanoid B + Humanoid A + Human (cross-embodiment and human priors). As shown in Fig. 5a, co-trained policies (ii) and (iii) substantially outperformed the Humanoid B-only baselines on all task settings, underscoring the method's ability to transfer latent task structure across embodiments.

Experiment 2: Scaling Demonstrations for Few-Shot Adaptation We further quantified the relationship between required for few-shot generalization. We hold Humanoid A and human datasets fixed

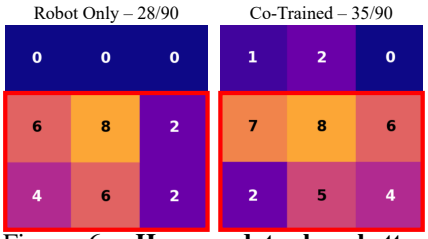


Figure 6: **Human data has better sampling efficiency**. Per-grid vertical grasping successes out of 10 trials with models trained with robot-only data and mixed data. Red boxes indicate where training data is collected.

Task	State Space	Action Speed	Success
Vertical Grasping	✓	✗	1/10
	✗	✓	0/10
	✓	✓	**4/10**

Table 3: **Importance of unifying policy inputs and outputs.** We report the number of successes of vertical grasping objects in the upper-left block as illustrated in Fig. 8. Baselines use joint positions as state input or do not interpolate human motions.

for the horizontal grasping task and ablate number of demonstrations required for Humanoid B in Fig. 5. Co-training (Humanoid B + A + Human) consistently outperformed isolated training on Humanoid B across all settings, especially in the few-data regime.

4.3 Ablation Study

Sampling Efficiency of Human and Humanoid Data. Conceptually, collecting human data is less expensive, not just because it can be done faster, but also because it can be done in in-the-wild scenes; reduces setup cost before every data collection; and avoids the hardware cost to equip every operator with robots.

We perform additional experiments to show that even in the lab setting, human data can have better sampling efficiency in unit time. In particular, we provide a small-scale experiment on the vertical grasping task. Allocating 20 minutes for two settings, we collected (1) 60 Humanoid A demonstrations, (2) 30 Humanoid A demonstrations, and 120 human demonstrations. To avoid conflating diversity and data size, the object placements in all demonstrations are evenly distributed at the bottom 6 cells. The results are given in Fig. 6. The policy trained with mixed robot and human data performs significantly better, which validates the sampling efficiency of human data over robot data. Each cell represents a 10cm × 10cm region where the robot attempts to pick up a box.

State-Action Design. In Tab. 3, we ablate the design choices of the proprioception state space and the speed of output actions. In particular, using the same set of robot and human data, we implement two baselines: 1) a unified state-action space, but does not interpolate (*i.e.,* slow down) the human actions; and 2) a baseline that interpolates human actions but uses separate state representation for humanoid (joint positions) and humans (EEF representation). The policies exhibit different failure patterns during the rollout of these two baselines. Without interpolating human actions, the speed of the predicted actions fluctuates between fast (resembling humans) and slow (resembling teleoperation), which leads to instability. Without a unified state space, the policy is given a 'shortcut' to distinguish between embodiments, which leads to on-par in-distribution performance and significantly worse OOD performance.

More Ablation Study. Due to space limit, please refer to the appendix and the supplementary for more qualitative visualization and quantitative ablation studies.

5 Conclusions

This paper proposes PH^{2}D, an effort to construct a large-scale human task-oriented behavior dataset, along with the training pipeline HAT, which leverages PH^{2}D and robot data to show how humans can be treated as a data source for cross-embodiment learning. We show that it is possible to directly train an imitation learning model with mixed human-humanoid data without any training surrogates when the human data are aligned with the robot data. The learned policy shows improved generalization and robustness compared to the counterpart trained using only real-robot data.

6 Limitations

Although we also collect language instructions in PH^{2}D, due to our focus on investigating the embodiment gap between humans and humanoids, one limitation of the current version of the paper uses a relatively simple architecture for learning policy. In the near future, we plan to expand the policy learning process to train a large language-conditioned cross-embodiment policy to investigate generalization to novel language using human demonstrations. The collection of human data relies on off-the-shelf VR hardwares and their hand tracking SDKs. Since these SDKs were trained mostly for VR applications, hand keypoint tracking can fail for certain motions with heavy occlusion. In addition, though the proposed method conceptually extends to more robot morphologies, current evaluations are done on robots equipped with dexterous hands.

7 Acknowledgment

This work was supported, in part, by NSF CAREER Award IIS-2240014, NSF CCF-2112665 (TILOS), and gifts from Amazon, Meta and Apple.

References

[1] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

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

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

[4] S. Dasari, O. Mees, S. Zhao, M. K. Srirama, and S. Levine. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*, 2024.

[5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[6] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[7] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[8] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024.

[9] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning (CoRL)*, 2024.

[10] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[11] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

[12] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[13] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[14] A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[15] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[16] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024. URL https://arxiv.org/abs/2410.24221.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[18] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023.

[19] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. In *RSS*, 2023.

[20] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

[21] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[22] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[23] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.

[24] L. Wang, X. Chen, J. Zhao, and K. He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024.

[25] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[26] R.-Z. Qiu, Y. Song, X. Peng, S. A. Suryadevara, G. Yang, M. Liu, M. Ji, C. Jia, R. Yang, X. Zou, et al. Wildlma: Long horizon loco-manipulation in the wild. *arXiv preprint arXiv:2411.15131*, 2024.

[27] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. In *ICRA*, 2025.

[28] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.

[29] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *2023 ieee international conference on robotics and automation (icra)*, pages 5954–5961. IEEE, 2023.

[30] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[31] J. Lee and M. S. Ryoo. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–2, 2017.

[32] K. Lee, Y. Su, T.-K. Kim, and Y. Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 61(12):1323–1334, 2013.

[33] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis. Translating videos to commands for robotic manipulation with deep recurrent neural networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3782–3788. IEEE, 2018.

[34] J. Rothfuss, F. Ferreira, E. E. Aksoy, Y. Zhou, and T. Asfour. Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, 3(4):4007–4014, 2018.

[35] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[36] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. In *ECCV*, 2024.

[37] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[38] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024.

[39] N. Das, S. Bechtle, T. Davchev, D. Jayaraman, A. Rai, and F. Meier. Model-based inverse reinforcement learning from visual demonstrations. In *Conference on Robot Learning*, pages 1930–1942. PMLR, 2021.

[40] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[41] S. Pirk, M. Khansari, Y. Bai, C. Lynch, and P. Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.

[42] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[43] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[44] L.-H. Lin, Y. Cui, A. Xie, T. Hua, and D. Sadigh. Flowretrieval: Flow-guided data retrieval for few-shot imitation learning. *arXiv preprint arXiv:2408.16944*, 2024.

[45] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.

[46] Y. Zhu, A. Lim, P. Stone, and Y. Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.

[47] C.-C. Hsu, B. Wen, J. Xu, Y. Narang, X. Wang, Y. Zhu, J. Biswas, and S. Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.

[48] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024.

[49] W. Huang, I. Mordatch, and D. Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *International Conference on Machine Learning*, pages 4455–4464. PMLR, 2020.

[50] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.

[51] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[52] J. Yang, D. Sadigh, and C. Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[53] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[54] T. Franzmeyer, P. Torr, and J. F. Henriques. Learn what matters: cross-domain imitation learning with task-relevant embeddings. *Advances in Neural Information Processing Systems*, 35: 26283–26294, 2022.

[55] A. Ghadirzadeh, X. Chen, P. Poklukar, C. Finn, M. Björkman, and D. Kragic. Bayesian meta-learning for few-shot policy adaptation across robotic platforms. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1274–1280. IEEE, 2021.

[56] T. Shankar, Y. Lin, A. Rajeswaran, V. Kumar, S. Anderson, and J. Oh. Translating robot skills: Learning unsupervised skill correspondences across robots. In *International Conference on Machine Learning*, pages 19626–19644. PMLR, 2022.

[57] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.

[58] Z.-H. Yin, L. Sun, H. Ma, M. Tomizuka, and W.-J. Li. Cross domain robot imitation with invariant representation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 455–461. IEEE, 2022.

[59] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pages 537–546. PMLR, 2022.

[60] G. Zhang, L. Zhong, Y. Lee, and J. J. Lim. Policy transfer across visual and dynamics domain gaps via iterative grounding. *arXiv preprint arXiv:2107.00339*, 2021.

[61] Q. Zhang, T. Xiao, A. A. Efros, L. Pinto, and X. Wang. Learning cross-domain correspondence for control with dynamics cycle-consistency. *arXiv preprint arXiv:2012.09811*, 2020.

[62] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. In *RSS*, 2022.

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

[64] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

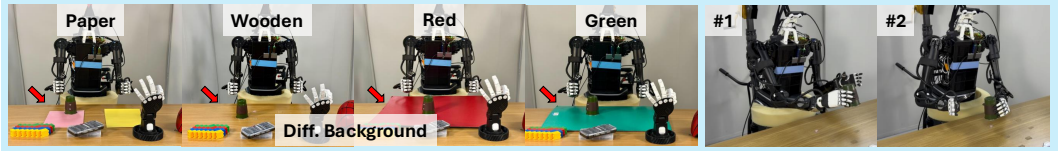
[65] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[66] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023.

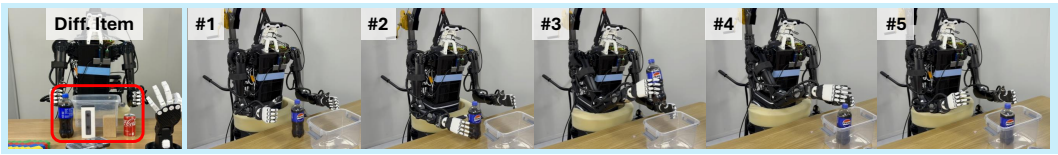
[67] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[68] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

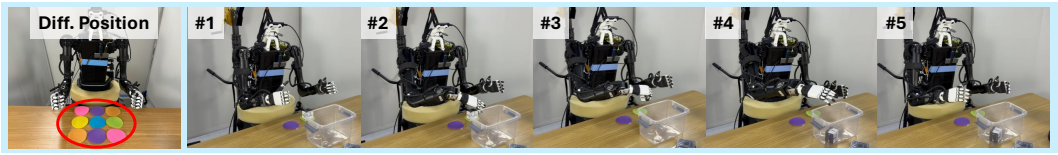
[69] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.



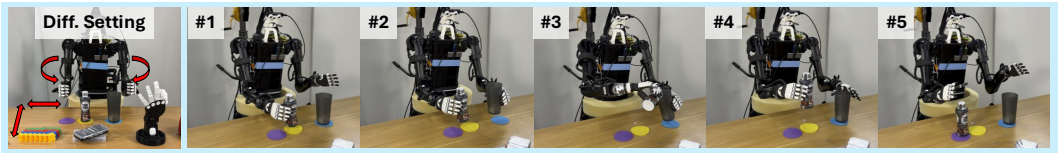
(a) The robot performs the **cup passing** task across four different backgrounds. The left side shows the four background variations, while the right side illustrates the two passing directions: (*#1* - Right hand passes the cup to the left hand, *#2* - Left hand passes the cup to the right hand).



(b) The robot performs the **horizontal grasping** task with four different items: bottle, box_1, box_2, and can, as shown on the left. The right side illustrates the process: (*#1-#3* - The robot grasps the bottle, *#4-#5* - The robot places it into the plastic bin).



(c) The robot performs the **vertical grasping** task. As shown on the left, the Dynamixel box is placed in nine different positions for grasping. The right side illustrates the process: (*#1-#3* - The robot grasps the box, *#4-#5* - The robot places the box into the plastic bin).



(d) The robot performs the **pouring** task. The left side shows different settings achieved by varying the robot's rotation and the table's position. The right side illustrates the pouring process: (*#1* - Right hand grasps the bottle, *#2* - Left hand grasps the cup, *#3* - Pouring the drink, *#4* - Left hand places the cup down, *#5* - Right hand places the bottle down).

Figure 7: Illustrations of tasks used in quantitative evaluations. From top to bottom: cup passing, horizontal grasping, vertical grasping, and pouring.

Method	Bottle	Box_1	Box_2	Can	Ovr. Succ.
	I.D.	H.D.	H.D.	H.D.	
Without whole-body	8/10	**6/10**	0/10	**7/10**	**21/40**
With whole-body	9/10	3/10	**3/10**	3/10	18/40

Table 4: **Ablation of how human whole-body movement in training demonstrations affects policy rollout**. We collect the same number of demonstrations on the same set of objects for the *grasping* task with or without whole-body movement. Since the robot does not have a natural whole-body movement like humans, it negatively influences the manipulation success rate.

A More Ablation Study - Data Collection

Autonomous Whole-body Movement. In Tab. 4, we justify the necessity to minimize body movement in human data collection. Humans tend to move their upper body unconsciously during manipulation (including shoulder and waist movement). However, existing humanoid robots have yet to reach such a level of dexterity. Thus, having these difficult-to-replicate actions in the human demonstrations leads to degraded performance. We hypothesize that such a necessity would be greatly reduced with the development of both whole-body locomotion methods and mechanical designs,

Method	Grasping (secs)	Pouring (secs)
Human Demo	3.79±0.27	4.81±0.35
Human Demo with VR	4.09±0.30	4.90±0.26
Humanoid Demo (VR Teleop)	19.72±1.65	37.31±6.25

Table 5: **Amortized mean and standard deviation of the time required to collect a single demonstration**, including scene resets. The first row shows the time for regular human to complete corresponding tasks in real world. The second row represents our human data when wearing VR for data collection, demonstrating that egocentric human demonstrations provide a more scalable data source compared to robot teleoperation.

but for the currently available platforms, we instruct operators to minimize body movement as much as possible in our dataset.

Efficiency of Data Collection. In Tab. 5, we compare task completion times across different setups, including standard human manipulation, human demonstrations performed while wearing a VR device, and robot teleoperation. This analysis highlights how task-oriented human demonstrations can be a scalable data source for cross-embodiment learning. Notably, wearing a VR device does not significantly impact human manipulation speed, as the completion time remains nearly the same as in standard human demonstrations.

Among different data collection schemes, we find that most overhead arises during the retargeting process from human actions to robot actions. This is primarily due to latency and the constrained workspace of 7-DoF robotic arms, which are inherent challenges in existing data collection methods such as VR teleoperation [9], motion tracking [48, 10], and puppeting [8, 5].

Beyond data collection speed, human demonstrations offer several additional advantages over teleoperation. They provide a safer alternative, reducing risks associated with real-robot execution. They are also more labor-efficient, as they do not require additional personnel for supervision. Furthermore, human demonstrations allow for greater flexibility in settings, enabling a diverse range of environments without requiring robot-specific adaptations. Additionally, human demonstrations achieve a higher demonstration success rate, and the required hardware (such as motion capture or VR devices) is more accessible and cost-effective compared to full robotic setups. These factors collectively make human data a more scalable solution for large-scale data collection.

B Normalization of different embodiments.

Tab. 2 suggests minor differences between using different normalization coefficients for the states and actions vectors of humans and humanoids. We take a closer look in Fig. 8, where we investigate the impact of different normalization strategies in the vertical grasping (picking) task. Noticeably, the same normalization approach achieved the highest overall success rate, but the success distribution is biased towards the upper-right region of the grid.

We hypothesize that this is because humans have a larger workspace than humanoid robots. Thus, human data encompasses humanoid proprioception as a subset, which results in a relatively smaller distribution for the robot state-action space.

C In-Depth Analysis of Different Types of Generalization

Human data improves background generalization. We chose to use the *cup passing* task to test background generalization. We prepared four different tablecloths as backgrounds, as shown in Fig. 7a. In terms of training data distribution, the teleoperation data for this task was collected exclusively on the paper background shown in Fig. 7a, whereas the human data includes more than five different backgrounds. This diverse human dataset significantly enhances the generalization ability of the co-trained HAT policy. As shown in Tab. 7. , HAT consistently outperforms across all four backgrounds, demonstrating robustness to background variations. In addition, the overall

Method	Bottle	Box_1	Box_2	Can	Ovr. Succ.
	I.D.	H.D.	O.O.D.	O.O.D.	
ACT	8/10	5/10	**1/10**	1/10	16/40
HAT	8/10	**7/10**	**1/10**	**4/10**	**21/40**

Table 6: **Object Appearance Generalization:** In the horizontal grasping task, we evaluated the grasping performance by attempting to grasp each object 10 times and recorded the success rate.

success rate increases by nearly 50% compared to training without human data, highlighting the advantage of utilizing diverse human demonstrations.

Human data improves appearance generalization. To test how co-training improves robustness to perturbations in object textures, we evaluate the *horizontal grasping* policy on novel objects, as shown in Fig. 7b. Specifically, we compare the policy's performance on the bottle, box_1, box_2, and can, as shown left to right in the first image in Fig. 7b. These objects differ significantly in both color and shape from the bottle used in the teleoperation data distribution.

Since grasping is a relatively simple task, our adjusted policy demonstrates strong learning capabilities even with only 50 teleoperation data samples. The policy can successfully grasp most bottles despite the limited training set. To better highlight the impact of human data, we selected more challenging objects for evaluation. As shown in Tab. 6, human data significantly enhances the policy's ability to grasp these more difficult objects.

Notably, box_1 appears in the human data, while box_2 does not. Despite this, we observe that co-training with human data still improves overall performance, even on box_2, though its success rate does not increase. This suggests that, beyond direct experience with specific objects, the human data helps the policy learn broader visual priors that enable more proactive and stable grasping behaviors. For box_2, while the success rate remains low—partially due to its low height and color similarity to the table—the co-trained HAT policy demonstrates fewer out-of-distribution (OOD) failures and more actively searches for graspable regions. The failures on box_2 are primarily due to unstable grasping and the small box slipping from the hand, rather than the inability to perceive or locate the object.

Furthermore, adding more human data not only improves performance on objects seen in human training demonstrations (e.g., box_1) but also enhances generalization to completely novel objects (e.g., box_2 and can). We hypothesize that, as the number of objects grows, HAT starts to learn inter-category visual priors that guide it to grasp objects more effectively, even when they were not explicitly present in the training set.

Human data improves object placement generalization. Finally, we introduce variations in object placements that are not present in the real-robot training demonstrations and specifically investigate this in the *vertical grasping (picking)* task. In this task, we intentionally constrain the robot data collection to object placements within a subset of cells, while human vertical grasping data covers a much more diverse range of settings.

To systematically analyze the impact of human data, we evaluate model performance on a structured 3×3 grid, where each cell represents a 10cm × 10cm region for grasping attempts. The numbers in each cell indicate the number of successful picks out of 10 trials. Real-robot training data is collected from only two specific cells, highlighted with dashed lines.

A key detail in our teleoperation data distribution is that 50 picking attempts are collected from the right-hand side grid and only 10 from the left-hand side grid. This imbalance explains why policies trained purely on teleoperation data struggle to grasp objects in the left-side grid. We observe that models trained solely on robot data fail to generalize to unseen cells, whereas cross-embodiment learning with human data significantly improves generalization, doubling the overall success rate.

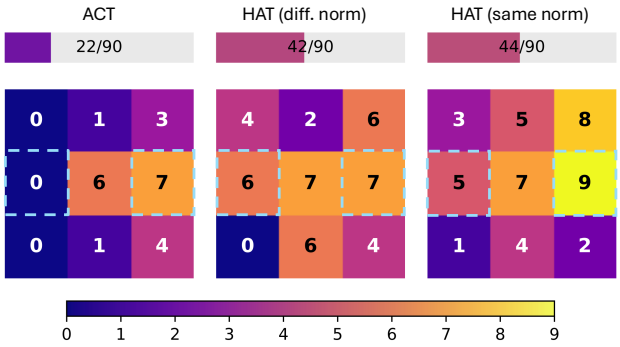


Figure 8: **Object Placement Generalization.** Performance comparisons of models trained with and without human data on vertical grasping (picking). Each cell in the 3×3 grid represents a 10cm × 10cm region where the robot attempts to pick up a box, with numbers indicating successful attempts out of 10. The real-robot data is collected in two cells inside the dashed lines. Notably, our teleoperation data is intentionally imbalanced.

Method	Paper	Wooden	Red	Green	Ovr. Succ.
	I.D.	H.D.	O.O.D.	O.O.D.	
ACT	19/20	14/20	12/20	10/20	55/80
HAT	**20/20**	**16/20**	**18/20**	**18/20**	**72/80**

Table 7: **Background Generalization:** In the cup passing task, we evaluate the passing performance by recording the number of failures or retries needed to complete 20 cup-passing trials.

D In-Depth Comparison between Humanoid A and Humanoid B configurations

This section presents a detailed comparison of the two humanoid platforms, referred to as Humanoid A and Humanoid B, with a focus on joint structure and implications for manipulation capabilities. We restrict our analysis to the arm configurations, as other parts of the body were not exclusively explored in this work.

While morphologically similar, these two humanoids have drastically different arm configurations that create hurdles in direct policy transfer. Besides differences in motor technical specs such as torque and types of encoder (Humanoid B has absolute motor position encoders), they also have different mechanical limits. The range of motion (ROM) for the first four proximal joints—shoulder_pitch, shoulder_roll, shoulder_yaw, and elbow—differs across the two platforms. Humanoid B exhibits a consistently wider ROM, which allows a wider set of reachable configurations and increases the manipulability of the arm in constrained environments. Table 8 summarizes the ROM values for these shared joints.

A more significant architectural divergence is observed at the wrist. Humanoid A includes a single distal joint—wrist_roll—providing limited wrist articulation. This restricts end-effector dexterity and constrains in-hand manipulation strategies to a single rotational degree of freedom. In contrast, Humanoid B is equipped with a complete wrist mechanism composed of three independently actuated joints: wrist_pitch, wrist_roll, and wrist_yaw. These additional degrees of freedom allow for full orientation control of the end-effector, enabling tasks that require precise alignment, rotation, and fine adjustment of object poses.

Joint	Humanoid A	Humanoid B
shoulder_pitch	-164° to $+164^\circ$	-180° to $+90^\circ$
shoulder_roll	-19° to $+178^\circ$	-21° to $+194^\circ$
shoulder_yaw	-74° to $+255^\circ$	-152° to $+172^\circ$
elbow	-71° to 150°	-54° to 182°
wrist_roll	-175° to 175°	-172° to 157°

Table 8: Joint Range of Motion Comparison between Humanoid A and B (in degrees)