Humanoid Policy $\sim$ Human Policy

Ri-Zhao Qiu[*1], Shiqi Yang[*1], Xuxin Cheng[*†1], Chaitanya Chawla[2], Jialong Li[1], Tairan He[2], Ge Yan[3],

Lars Paulsen[1], Ge Yang[4], Sha Yi[1], Guanya Shi[2], Xiaolong Wang[1]

[*]equal contribution   [†]project lead

[1]UC San Diego, [2]CMU, [3]University of Washington, [4]MIT

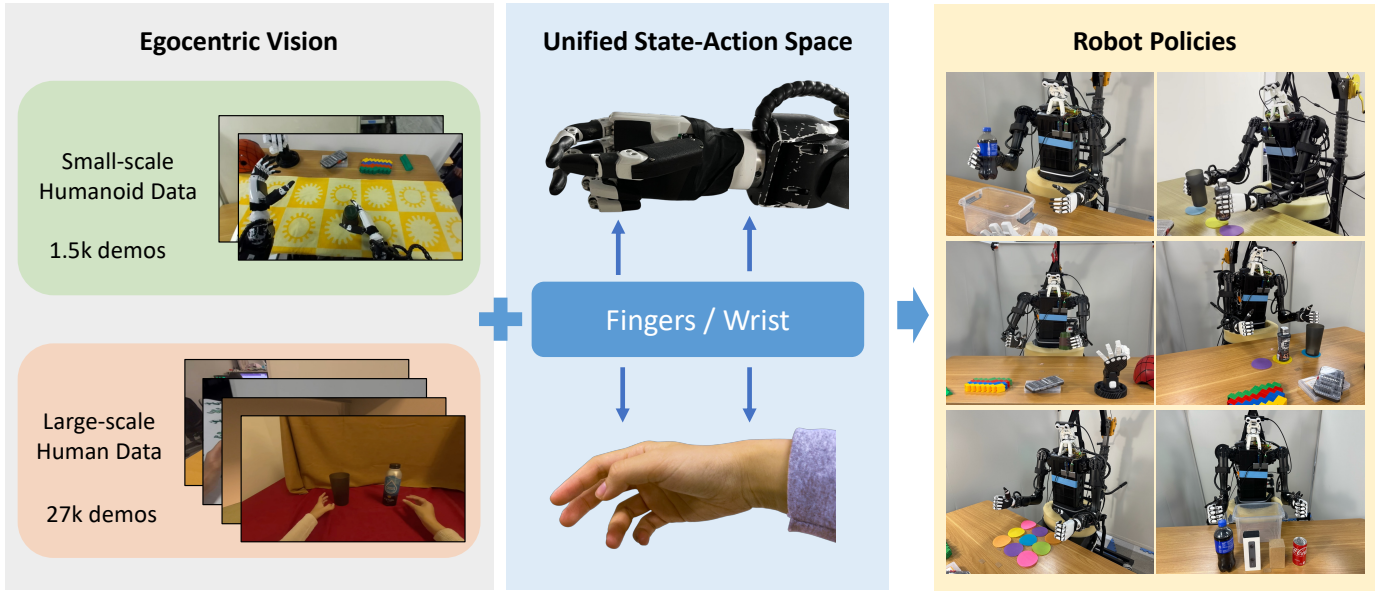https://human-as-robot.github.io

Fig. 1: We propose to use **task-oriented** egocentric human data to learn humanoid manipulation policies. 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. To facilitate such cross-embodiment learning, We collect a large-scale dataset, **P**hysical **H**uman-**H**umanoid **D**ata (**PH**^{2}**D**), with automatic hand-finger 3D poses from consumer-grade VR devices.

Abstract—**Humanoid** (*adj.*): **having human form or characteristics.** – *Merriam-Webster, 2025.*

Training manipulation policies for humanoid robots with diverse data enhance 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 improves both generalization and robustness of HAT with significantly better data collection efficiency.

I. INTRODUCTION

Learning from real robot demonstrations has led to great progress in robotic manipulation recently [37, 43, 6, 15]. One key advancement to enable such progress was hardware / software co-designs to scale up data collection using teleoperation [70, 20, 11, 62, 9, 25] and directly controlling the robot end effector [14, 5, 70, 20, 23, 11]. Instead of gathering data on a single robot, collective efforts have been made to merge diverse robot data and train foundational policies across embodiments [14, 44, 43, 37, 6, 15], 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 [50] and NLP [45].

If we examine humanoid robot teleoperation more closely, it involves robots mimicking human actions using geometric transforms or retargeting to control robot joints and endeffectors. From this perspective, **we propose to model robots in a human-centric representation**, and the robot action is

Dataset	Human		Robot	
	# Frames	# Demos	# Frames	# Demos
DexCap [55]	~378k	787	NA	NA
EgoMimic [29]	~432k[†]	2,150	**1.29M**[†]	1,000
PH²D (Ours)	**~3.02M**	**26,824**	~668k	**1,552**

TABLE I: **Comparisons of task-oriented egocentric human datasets.** Besides having the most demonstrations, PH²D is collected on various manipulation tasks, a diverse set of objects and receptacles, accompanied by accurate 3D hand-finger poses and language annotations. The number of frames is estimated assuming 30 Hz. [†]: estimated based on time reported for data collection; whereas DexCap [55] and PH²D report frames after initial processing.

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.

In this paper, we introduce to 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 [40, 3]. To realize this, we first collect an egocentric task-oriented dataset of **P**hysical **H**umanoid-**H**uman **D**ata, dubbed PH²D. 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 [22, 12], PH²D 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 convert human actions to robot actions. Such a conversion is a differentiable process that allows end-to-end training on different embodiments.

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. In addition, it enables generalization to objects completely unseen in robot 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**, PH²D, which is a large egocentric, task-oriented human-humanoid dataset with accurate hand and wrist poses for modeling human behavior (see Tab. I).
- **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.

II. 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 dexterity [71, 43, 56, 37, 10, 49, 9, 38] compared to RL- and planning-based approaches [48, 28, 36, 35, 68] due to direct alignment with the real world. The scale of data for imitation learning has grown substantially with recent advancements in data collection [1, 9, 11, 62], 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.

1) 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 [50] and natural language processing [45] 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 [7, 31, 32, 42, 52, 63]. Due to various discrepancies (*e.g.,* supervision and viewpoints) between egocentric robot views and internet videos, most existing work [40, 3] use modular approaches with intermediate representations as surrogates for training. The most representative ones are affordances [40, 3] for object interaction, object keypoints predictions [4, 57, 33, 13, 58], or other types of object representations [47, 41, 39].

2) 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 [29], aligning motion trajectories or flow [34, 51], object-centric actions [74, 26], or hand tracking with specialized hardware [55].

However, many existing papers focus on imitating humans in a single task, overlooking the potential of human data to be directly applied for larger scale cross-embodiment learning.

Cross-Embodiment. Cross-embodiment pre-training has been shown to improve adaptability and generalization over different embodiments [27, 8, 61, 60, 16, 18, 21, 53, 59, 64, 66, 67, 69]. When utilizing human videos, introducing intermediate representations can be prone to composite errors. Recent works investigate end-to-end approaches [43, 56, 37, 6] 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 [6]. Some other work [2, 33] 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 [55, 29, 54] also attempts to use egocentric human demonstrations for end-to-end cross-embodiment policy learning. DexCap [55] 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 [6, 37] have shown the scalability of 2D visual inputs. Most related to our work, EgoMimic [29] also proposes to collect data using wearable device [17] with 2D visual inputs. However, EgoMimic requires strict visual sensor alignment and heuristic designs such as visual masking. Such reliance on visual models during training and testing leads to composite failure similar to the modular approach. In addition, PH^{2}D 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.

III. 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 [6, 37, 15, 43, 30, 44]. This paper proposes egocentric human manipulation demonstrations as a scalable source of cross-embodiment training data. Sec. III-A 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. III-B describes various techniques to handle domain gaps to align human data and robot data for learning humanoid manipulation policy.

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

Though there has been existing work that collects egocentric human videos [29, 12, 22, 55], they either (1) provide demonstrations mostly for non-task-oriented skills and do not provide world-frame 3D head and hand poses estimations for imitation learning supervision [22, 12] or (2) require specialized hardware or robot setups [55, 29].

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.

a) Adapting Low-cost Commerical Devices: With development in pose estimation [73] 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
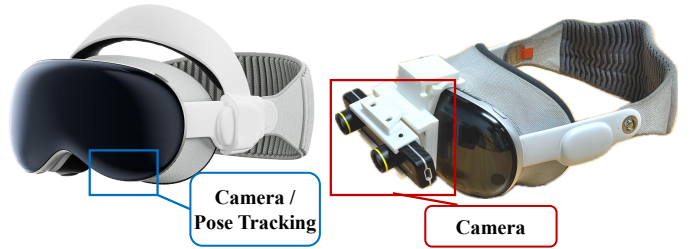

Fig. 2: **Adapting Consumer-grade Devices for Data Collection.** To avoid relying on specialized hardware for data collection and make our method more accessible, we design our data collection process using consumer-grade VR devices.

teleoperate robot in real-time [9, 23]. We design software and hardware to support convenient data collection across different devices:

- **Apple Vision Pro + Built-in Camera.** We developed a Vision OS App that accesses the bottom left camera for visual observation and uses the Apple ARKit to access 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.

Different cameras are intended to provide more visual diversity for data collection.

b) 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 closet 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. In addition, we place objects close to humans so that they are within the workspace of human arms without whole-body movement, which resembles the workspace of commercial humanoid robots without whole-body movement.

B. 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
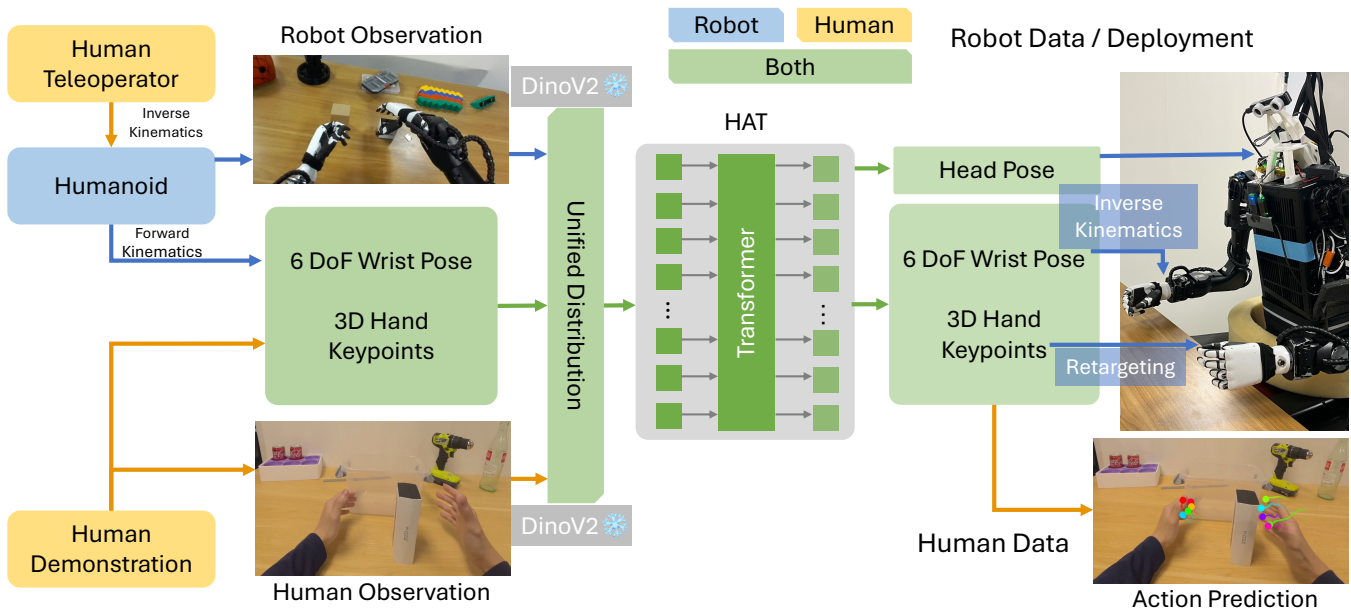
Fig. 3: **Overview of HAT.** Human Action Transformer (HAT) learns a robot policy by modeling humans. During training, we sample a state-action pair from either human data or robot data. The images are encoded by a frozen DinoV2 encoder [46]. The HAT model makes predictions in a human-centric action space using wrist 6 DoF poses and finger tips, which is retargeted to robot poses during real-robot deployment.

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^2D 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 $M \gg N$ because collecting egocentric human videos can be done more efficiently than both tele-operation.

The goal is to design a policy $\pi : \mathbf{S} \to \mathbf{A}$ 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). This paper uses HAT as the policy, which is based on Action Chunk Transformer [70]. We modified the original implementation to replace ResNet-18 [24] with frozen DinoV2 [46] backbones and added visual adaptor to intermediate layers. 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 robot is also equipped with an actuated 2DoF neck that can rotate, which resembles the autonomous whole-body 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 action space for prediction of HAT is a 54-dimensional vector. Rotations of the head, left wrist, and right wrist are represented as 6D rotations [72]; the translation of left and right wrists are represented as x/y/z vectors. In this work, since we deploy our policy on a robot with 5-fingered dexterous hands (shown in Fig. 3), there exists a bijective mapping between 10 finger tips

of robot dexterous hands and common human hands, which are represented as 3D x/y/z keypoints for regressing. Note that injective mapping is also possible (*e.g.,* mapping distance between the thumb finger and other fingers to parallel gripper distance, but we leave it for future study).

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. In addition, 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 policies such as adding visual artifacts [29] or generative methods [65] to train human-robot policies - basic image augmentations such as color jittering and Gaussian blurring are effective regularization of the visual inputs in this case.

Action Domain Gap. To mitigate the drastic speed difference between humans and humanoids, 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 $\alpha_{slow} = 4$ in all tasks.

Due to the difference in state space, we randomly dropout proprioception readings during training to avoid undesired reliance on low-dimensional state inputs.

Training. The final policy is denoted as $\pi : f_\theta(\cdot) \to \mathbf{A}$ for both human and robot policy. The trainable parameters of both the proprioceptive encoder θ and the transformer trunks are optimized jointly during training. The final loss is given
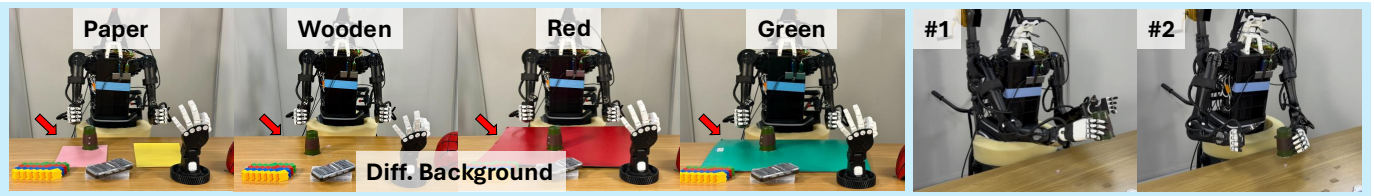
(a) The robot performs the **cup passing** task across four different backgrounds. The left side shows the four background variations, while the right side illustrates the two passing directions: (*#1* - Right hand passes the cup to the left hand, *#2* - Left hand passes the cup to the right hand).
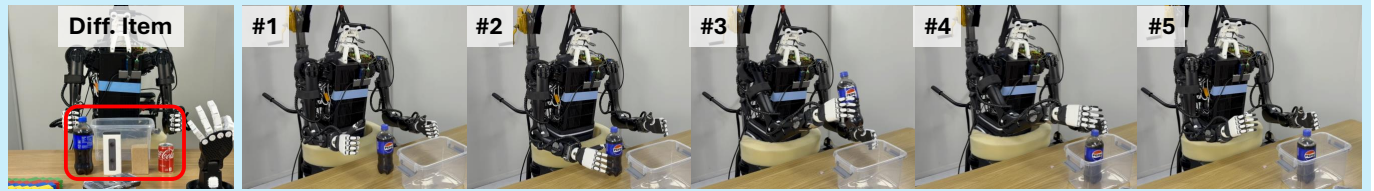
(b) The robot performs the **horizontal grasping** task with four different items: bottle, box_1, box_2, and can, as shown on the left. The right side illustrates the process: (*#1-#3* - The robot grasps the bottle, *#4-#5* - The robot places it into the plastic bin).
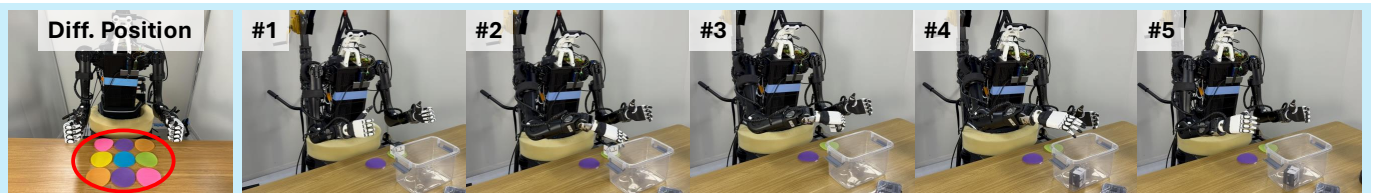
(c) The robot performs the **vertical grasping** task. As shown on the left, the Dynamixel box is placed in nine different positions for grasping. The right side illustrates the process: (*#1-#3* - The robot grasps the box, *#4-#5* - The robot places the box into the plastic bin).
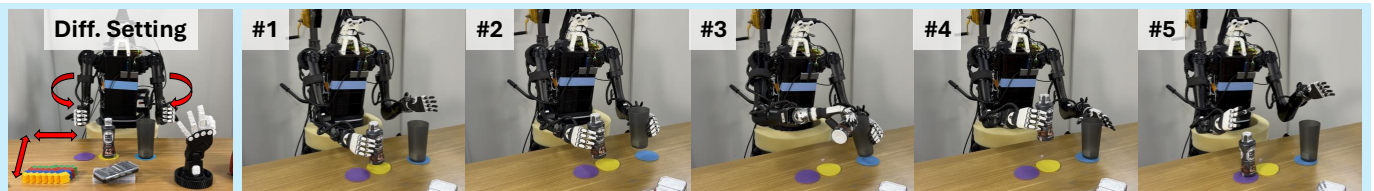
(d) The robot performs the **pouring** task. The left side shows different settings achieved by varying the robot's rotation and the table's position. The right side illustrates the pouring process: (*#1* - Right hand grasps the bottle, *#2* - Left hand grasps the cup, *#3* - Pouring the drink, *#4* - Left hand places the cup down, *#5* - Right hand places the bottle down).

Fig. 4: Illustrations of tasks used in quantitative evaluations. From top to bottom: cup passing, horizontal grasping, vertical grasping, and pouring.

by,

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

where $\lambda = 2$ is an (insensitive) hyperparameter used to balance loss to emphasize the importance of end effector positions over learning unnecessarily precise finger tip keypoints.

IV. EXPERIMENTS

This section investigates how cross-embodiment learning with egocentric task-oriented human demonstrations improves the robustness and generalizability of acquired policies, which highlights the potential of PH²D as a scalable data source. We also provide qualitative examples of manipulation tasks and ablate design choices.

a) Hardware Platforms.: We run our experiments on Unitree H1 bimanual humanoid robots equipped with a pair of 6-DOF Inspire dexterous hands. We focus on manipulation and do not actuate the lower body of the robot. Instead, we introduce a 2 DoF actuated neck similar to the design in [9], allowing the robot to visually focus on the objects it manipulates, similar to how humans do. This configuration results in a total of 28 degrees of freedom. The robot setup is illustrated in Fig. 3 and Fig. 4. Note that we rely solely on head cameras, without using wrist cameras, to study the manipulation capabilities based purely on egocentric vision.

b) Implementation Details.: We implement policy architecture by adopting the implementation from OpenTV [9]. In particular, we replace the trainable ImageNet-pretrained ResNet-18 backbone with frozen DinoV2 [46] ViT-S for its improved robustness against lighting and texture changes. Unless specifically noted, we use different means and standard deviations to normalize the state-action space. We implement

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**

TABLE II: **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. We also ablate the design choice of using different normalizations (D. Norm) for different embodiments.

two variants:

- ACT: baseline implementation using the architecture described above, trained using only robot data. Robot states are represented as joint positions.
- 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.

c) 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 Human-Distribution (H.D.) setting, we evaluate on-scene setups included in the human demonstrations but not robot demonstrations. In the Out-Of-Distribution (O.O.D.) setting, we test generalizability and robustness by introducing novel setups that were not present in any training demonstrations. Fig. 4 visualizes different manipulation tasks and how we define out-of-distribution settings for each task.

With the setup above, we aim to answer important research questions as follows:

- Does co-training with human data improve I.D. performance?
- Does co-training with human data improve O.O.D. generalization?
- How efficient is human demonstration collection compared to teleoperation [9]?
- How much does each design choice contribute?

A. Overall Evaluation

Human data has minor effects on I.D. testing. From Tab. II, we can see that I.D. performance with or without co-training with human data gives similar results. In the I.D. setting, we attempt to replicate the scene setups as training demonstrations, including both background, object types, and object placements. Thus, policies trained with only a small amount of robot data performed well in this setting. This finding is consistent with recent work [9, 11] that frozen visual

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 III: **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.

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 IV: **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.

foundation models [50, 46] improve robustness against certain perturbations such as lighting and similar textures.

Human data improves the O.O.D. settings. 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 setting to evaluate such generalization. From Tab. II, 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, we evaluate different types of generalization in separate tasks. We provide in-depth analyses of each type of generalization below.

B. Cross-embodiment 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. 4a.

In terms of training data distribution, the teleoperation data for this task was collected exclusively on the paper background
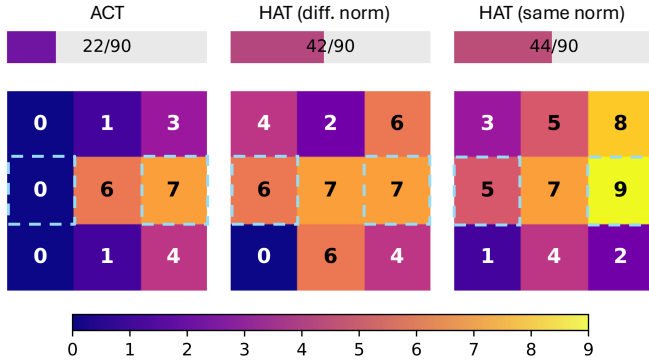
Fig. 5: **Object Placement Generalization.** Performance comparisons of models trained with and without human data on vertical grasping (picking). Each cell in the 3×3 grid represents a 10cm × 10cm region where the robot attempts to pick up a box, with numbers indicating successful attempts out of 10. The real-robot data is collected in two cells inside the dashed lines. Notably, our teleoperation data is intentionally imbalanced.

Method	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 V: **Ablation of how human whole-body movement in training demonstrations affects policy rollour.** We collect 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 natural whole-body movement like humans, it negatively influences the manipulation success rate.

shown in Fig. 4a, 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. III. , HAT consistently outperforms across all four backgrounds, demonstrating robustness to background variations. In addition, the overall 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. 4b. 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. 4b. 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. IV, 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 execution errors in grasping 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.

C. Ablation Study

Normalization of different embodiments. Tab. II suggests minor differences between using different normalization co-efficients for the states and actions vectors humans and humanoids. We take a closer look in Fig. 5, where we investigate the impact of different normalization strategies in the vertical grasping (picking) task. Noticeably, the same normalization

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

TABLE VI: **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. 5. Baselines use joint positions as state input or do not interpolate human motions.

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 VII: **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.

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. **Autonomous Whole-body Movement.** In Tab. V, 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, but for the currently available platforms, we instruct operators to minimize body movement as much as possible in our dataset. **State-Action Design.** In Tab. VI, 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 implicitly learns to differentiate between different embodiments. Rather than developing a shared representation across embodiments, it appears to rely on a form of conditional switching—activating the corresponding embodiment-specific policy when observing a familiar state. While this approach results in high success rates for in-distribution evaluation, it provides limited benefits for cross-embodiment generalization, as the policy struggles to transfer knowledge effectively between different embodiments. **Efficiency of Data Collection.** In Tab. VII, 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 [19, 25], and puppeting [62, 70].

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.

D. Few-Shot Transfer across Heterogenous Embodiments

To further evaluate our method's cross-embodiment adaptability, we conducted few-shot generalization experiments on a distinct humanoid platform (Unitree H1_2, termed Humanoid B), contrasting it with our primary platform (modified Unitree H1, Humanoid A). Notably, Humanoid B's demonstration data were collected in an entirely separate environment with only a single object, introducing both embodiment and environmental shifts.

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. 6 co-training policies (ii) and (iii) substantially outperformed the Humanoid B-only baseline. Notably, co-training improved success rates not only
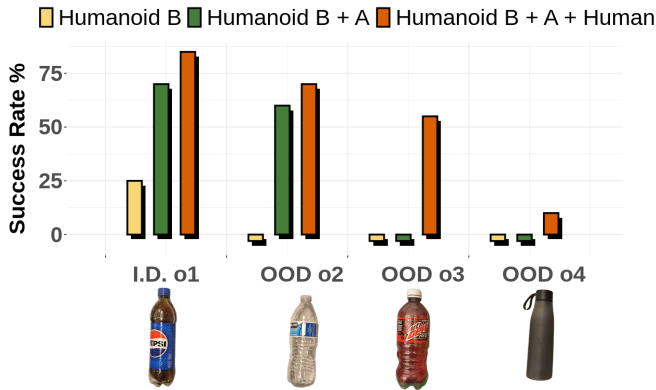
Fig. 6: **Co-training Efficacy and Generalization across Objects.** Performance of Humanoid B on co-training with Humanoid A and Human on vertical grasping across 4 objects. ID o1 is an object common to all embodiments. OOD o2 and OOD o3 are from Humanoid A demonstrations and Human demonstrations respectively. OOD o4 is out-of distribution for all embodiments.
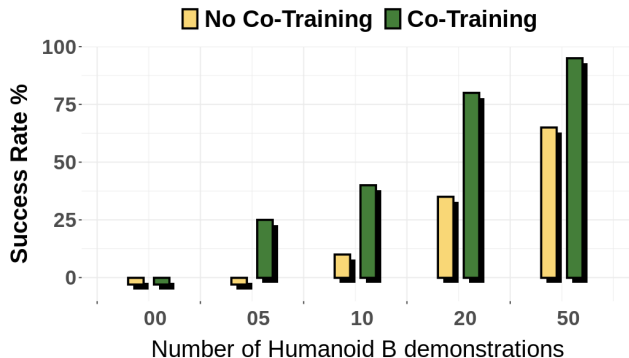

Fig. 7: **Few-Shot Adaptation.** Co-training consistently outperforms isolated training as Humanoid B demonstrations increase, achieving robust success rates even in low-data regimes.

on in-distribution objects (seen by Humanoid B) but also on novel objects never encountered by Humanoid B during training, 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. While holding Humanoid A (100 demos) and human (50 demos) datasets fixed, we incrementally increased Humanoid B's demonstrations (Fig. 7). Co-training (Humanoid B + A + Human) consistently outperformed isolated training on Humanoid B across all data regimes. Crucially, even with as few as 5–10 demonstrations from Humanoid B, co-training achieved more than twice the successful rollouts on novel objects, whereas zero-shot transfer (isolated training) remained near chance (12%). This demonstrates that our method enables data-efficient adaptation: by leveraging shared representations from diverse embodiments,

minimal target-domain demonstrations suffice to anchor robust policies.

These results highlight two key advantages of our approach: (1) the ability to unify heterogeneous demonstration sources (robots, humans) into a generalizable policy framework, and (2) the capacity to rapidly adapt to new embodiments with drastically reduced data requirements. Such capabilities are critical for scaling robot learning to real-world settings, where per-platform data collection is often prohibitively expensive.

V. 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.

VI. LIMITATION

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. In addition, current evaluations are done on robots equipped with dexterous hands, which are more aligned with humans. One future direction is to validate whether PH^{2}D also facilitates cross-embodiment learning of other forms of robots, such as those with parallel grippers.

REFERENCES

[1] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel 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.

[2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *RSS*, 2022.

[3] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

[4] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. In *ECCV*, 2024.

[5] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash 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.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

[8] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.

[9] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning (CoRL)*, 2024.

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

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

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[13] Neha Das, Sarah Bechtle, Todor Davchev, Dinesh Jayaraman, Akshara Rai, and Franziska Meier. Model-based inverse reinforcement learning from visual demonstrations. In *Conference on Robot Learning*, pages 1930–1942. PMLR, 2021.

[14] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

[15] Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*, 2024.

[16] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[17] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

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

[19] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024.

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

[21] Ali Ghadirzadeh, Xi Chen, Petra Poklukar, Chelsea Finn, Mårten Björkman, and Danica 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.

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[23] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[25] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[26] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.

[27] Wenlong Huang, Igor Mordatch, and Deepak 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.

[28] Mazeyu Ji, Ri-Zhao Qiu, Xueyan Zou, and Xiaolong Wang. Graspsplats: Efficient manipulation with 3d feature splatting. *arXiv preprint arXiv:2409.02084*, 2024.

[29] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024. URL https://arxiv.org/abs/2410.24221.

[30] 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.

[31] Jangwon Lee and Michael 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] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 61(12):1323–1334, 2013.

[33] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024.

[34] Li-Heng Lin, Yuchen Cui, Amber Xie, Tianyu Hua, and Dorsa Sadigh. Flowretrieval: Flow-guided data retrieval for few-shot imitation learning. *arXiv preprint arXiv:2408.16944*, 2024.

[35] Minghuan Liu, Zixuan Chen, Xuxin Cheng, Yandong Ji, Ri-Zhao Qiu, Ruihan Yang, and Xiaolong Wang. Visual whole-body control for legged loco-manipulation. In *CoRL*, 2024.

[36] Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. *arXiv preprint arXiv:2411.04999*, 2024.

[37] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[38] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. In *ICRA*, 2025.

[39] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[40] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *RSS*, 2023.

[41] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[42] Anh Nguyen, Dimitrios Kanoulas, Luca Muratore, Darwin G Caldwell, and Nikos 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.

[43] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.

[44] 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.

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

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[47] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.

[48] Ri-Zhao Qiu, Yafei Hu, Yuchen Song, Ge Yang, Yang Fu, Jianglong Ye, Jiteng Mu, Ruihan Yang, Nikolay Atanasov, Sebastian Scherer, et al. Learning generalizable feature fields for mobile manipulation. *arXiv preprint arXiv:2403.07563*, 2024.

[49] Ri-Zhao Qiu, Yuchen Song, Xuanbin Peng, Sai Aneesh Suryadevara, Ge Yang, Minghuan Liu, Mazeyu Ji, Chengzhe Jia, Ruihan Yang, Xueyan Zou, et al. Wildlma: Long horizon loco-manipulation in the wild. *arXiv preprint arXiv:2411.15131*, 2024.

[50] 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 *ICML*. PMLR, 2021.

[51] Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.

[52] Jonas Rothfuss, Fabio Ferreira, Eren Erdal Aksoy, You Zhou, and Tamim Asfour. Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, 3(4):4007–4014, 2018.

[53] Tanmay Shankar, Yixin Lin, Aravind Rajeswaran, Vikash Kumar, Stuart Anderson, and Jean Oh. Translating robot skills: Learning unsupervised skill correspondences across robots. In *International Conference on Machine Learning*, pages 19626–19644. PMLR, 2022.

[54] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anand-

kumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[55] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[56] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024.

[57] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[58] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh 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.

[59] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.

[60] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[61] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024.

[62] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024.

[63] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis 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.

[64] Zhao-Heng Yin, Lingfeng Sun, Hengbo Ma, Masayoshi Tomizuka, and Wu-Jun Li. Cross domain robot imitation with invariant representation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 455–461. IEEE, 2022.

[65] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

[66] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pages 537–546. PMLR, 2022.

[67] Grace Zhang, Linghan Zhong, Youngwoon Lee, and Joseph J Lim. Policy transfer across visual and dynamics domain gaps via iterative grounding. *arXiv preprint arXiv:2107.00339*, 2021.

[68] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In *RSS*, 2024.

[69] Qiang Zhang, Tete Xiao, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Learning cross-domain correspondence for control with dynamics cycle-consistency. *arXiv preprint arXiv:2012.09811*, 2020.

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

[71] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.

[72] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

[73] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023.

[74] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.