MotionTrans: Human VR Data Enable Motion-Level Learning for Robotic Manipulation Policies

Anonymous Submission

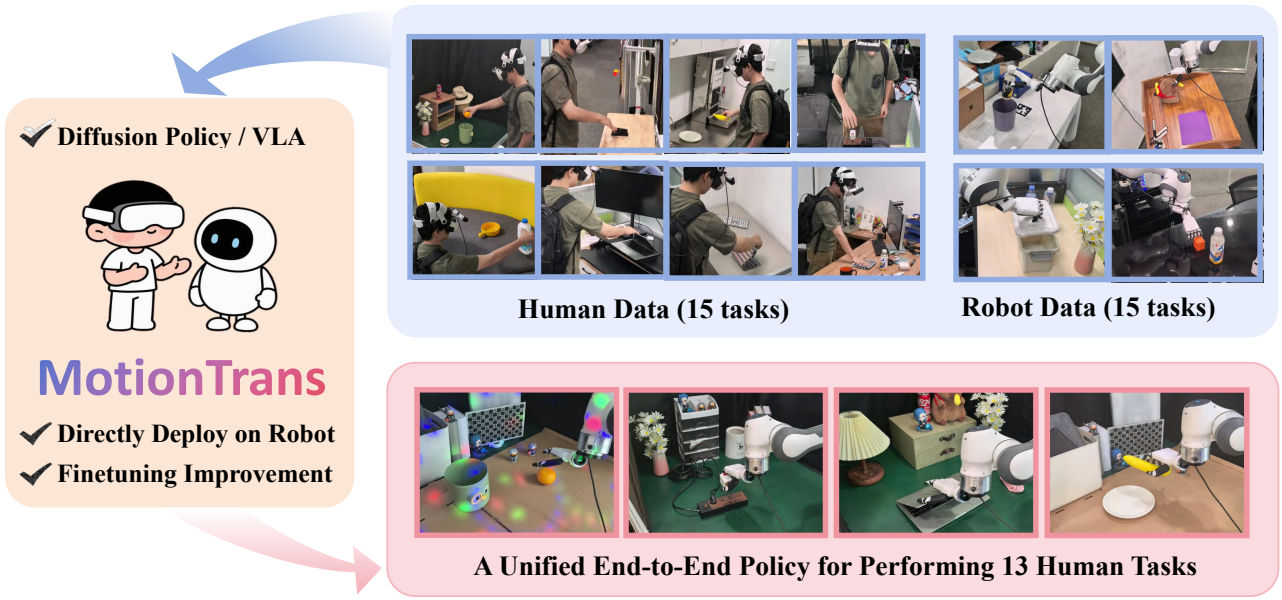

Fig. 1: We propose *MotionTrans*, a framework for **zero-shot** motion transfer from VR human data. By cotraining on 30 human / robot tasks, we enable end-to-end manipulation policies to directly perform tasks in human data on real robots.

***Abstract*— Scaling real robot data is a key bottleneck in imitation learning, leading to the use of auxiliary data for policy training. While previous works show that using human data can bring benefits, such as improving robustness and training efficiency, it remains unclear whether it can realize its greatest advantage: *enabling robot policies to directly learn new motions for task completion*. In this paper, we systematically explore this potential through multi-task human-robot cotraining. We introduce *MotionTrans*, a framework that includes a data collection system, a human data transformation pipeline, and a weighted cotraining strategy. By cotraining 30 human-robot tasks simultaneously, we direcly transfer motions of 13 tasks from human data to deployable end-to-end robot policies. Notably, 9 tasks achieve non-trivial success rates in zero-shot manner. All data, code, and model weights will be open-sourced.**

I. Introduction

Learning robotic manipulation policies from teleoperated demonstrations has advanced rapidly in recent years [1]–[3]. However, collecting large-scale robot datasets is costly and labor-intensive [4], creating a major bottleneck for scaling manipulation capabilities. Human data [5], [6] is abundant, easy to collect, and rich in diverse manipulation behaviors [6], making it a particularly promising source. Prior work leverages human demonstrations to extract task-aware intermediates—e.g., affordances [7] and keypoint flows [8]—to support motion transfer, but introducing intermediate representations hinders seamless integration with mainstream end-to-end policies. More recently, advances in wearable sensing have enabled direct use of human motion data (e.g., VR-tracked hand poses) for robot policy pretraining or cotraining [5], [9]–[12], yielding gains in visual grounding [11], robustness [10], and data efficiency [12]. Yet, it remains unclear whether such data can deliver its greatest advantage: *enabling robot policies to directly acquire new task motions*.

We address this question with *MotionTrans*, a framework designed to **directly learn 10+ robot-executable motions from human data within a unified, end-to-end policy**. We achieve this via multi-task human–robot cotraining. Concretely, we build a VR-based teleoperation system and data pipeline to construct the *MotionTrans Dataset*, comprising 3,213 demonstrations across 15 human tasks and 15 robot tasks in over 10 scenes. We introduce a transformation procedure that maps human demonstrations into the robot observation–action space, making them compatible with mainstream end-to-end policies such as Diffusion Policy [2] and the Vision–Language–Action model (π_0-VLA) [3]. Finally, we adopt a weighted cotraining strategy that jointly optimizes over human and robot tasks.

We evaluate **zero-shot transfer** performance on all human tasks, directly deploying policies to the robot without collecting any robot data for those tasks. Results show that both Diffusion Policy [2] and π_0-VLA [3] achieve non-trivial success on 9 tasks. Even when unsuccessful, the policies often exhibit meaningful task-directed motions, such as reaching target objects. These findings demonstrate the feasibility

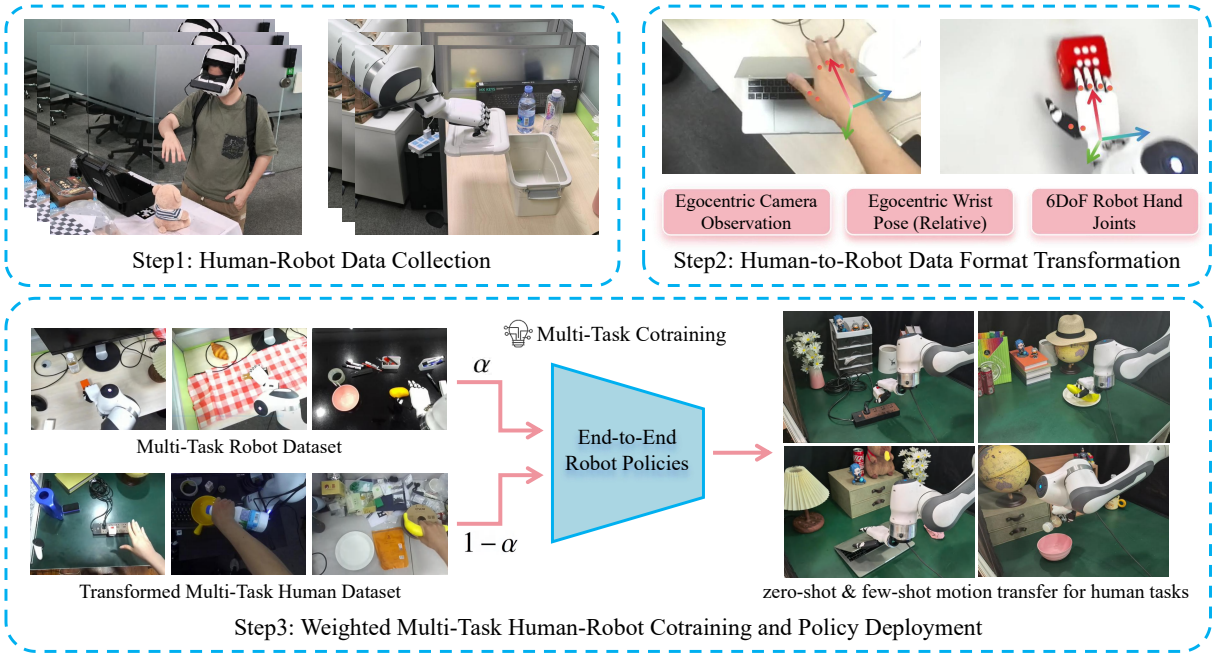

Fig. 2: Overview of *MotionTrans*: a human–robot data collection system, a pipeline that transforms human data into a robot-compatible format, and a weighted human–robot multi-task cotraining strategy. The trained policies can be directly deployed to execute tasks appearing only in the human dataset on real robots.

of motion-level learning from human data and establish a practical framework and principles for explicit human-to-robot transfer. We will also open-source the full pipeline to enable reproducibility by the community.

II. MotionTrans

A. Problem Definition

Our goal is to enable explicit human-to-robot motion transfer. We train a policy P_{policy} on $D = D_{\text{robot}} \cup D_{\text{human}}$, where $D_{\text{robot}} = \{D^i_{\text{robot}} \mid i = 1, \ldots, N_{\text{robot}}\}$ and $D_{\text{human}} = \{D^i_{\text{human}} \mid i = 1, \ldots, N_{\text{human}}\}$. Each D^i denotes a task-specific subset, and the human and robot task sets are **non-overlapping**. After training, we deploy P_{policy} on robot and evaluate it on **tasks from D_{human}** to assess the effectiveness of motion transfer. This constitutes the **zero-shot** setting, as these evaluated human tasks have no robot demonstrations during training. We define the input and output of our policies within the robot observation-action space $S = (I_t, P_t, A_t)$. At each timestamp t, the policy receives an egocentric RGB image $I_t \in \mathbb{R}^{H \times W \times 3}$ and proprioceptive states $P_t \in \mathbb{R}^{T_P \times D}$. The policy outputs an action chunk prediction $A_t \in \mathbb{R}^{T_A \times D}$ [2].

B. Human-Robot Data Collection System

For human-robot cotraining, we need to collect both robot and human data [9]. Here we describe the data collection system, illustrated in the top-left of Figure 2. We extend ARCap [13] to build our human data collection system (Figure 4(a)), incorporating a portable VR headset for recording hand keypoint positions K_t, wrist poses W_t and camera pose, and an RGB camera for the image stream I_t. The view of collectors are provided in Figure 4(b). For robot data (Figure 4(c)), we develop our teleoperation system based on VR-driven method Open-Television [1].

C. Transforming Human Data to the Robot Format

Raw human demonstrations collected with VR differ in format from robot demonstrations, preventing direct cotraining with robot policies [10], [11]. We therefore first transform human data into the **robot observation–action space** [1]. After transformation, human data could acts as "supplementary robot data", thus can be used to train any mainstream end-to-end **robot** policy. The observation-action space of the robot includes three components: image observation, wrist pose and hand joints. For image, we use **egocentric** view for both human and robot data. For wrist pose, we use the **egocentric** camera coordinate system for both human and robot data. This allows for the measurement of wrist poses in a unified coordinate system, ensuring that the spatial definitions of human and robot data are consistent. For hand joints, we employ the dex-retargeting [14], an optimization-based inverse kinematics solver, to map human hand keypoints from VR to robot hand joint state.

D. Weighted Multi-Task Human–Robot cotraining

By unifying the observation and action spaces, we enable joint training of human and robot data under a shared end-to-end robot policy. This section introduce the multi-task policy architectures we use and how we train these policies. We evaluate two representative policy architectures: **Diffusion Policy (DP)** [2] and **Vision–Language–Action model (π_0-VLA)** [3]. Following [15], we adopt a size-aware weighted objective over $D = D_{\text{robot}} \cup D_{\text{human}}$: $\mathcal{L}_D = \alpha \mathcal{L}_{D_{\text{robot}}} + (1-\alpha)\mathcal{L}_{D_{\text{human}}}$, where $\mathcal{L}$ denotes the loss function of imitation learning [2], [3]. We set $\alpha = \frac{|D_{\text{human}}|}{|D_{\text{human}}|+|D_{\text{robot}}|}$, where $|D_{\text{robot}}|$ and $|D_{\text{human}}|$ denote dataset sizes.

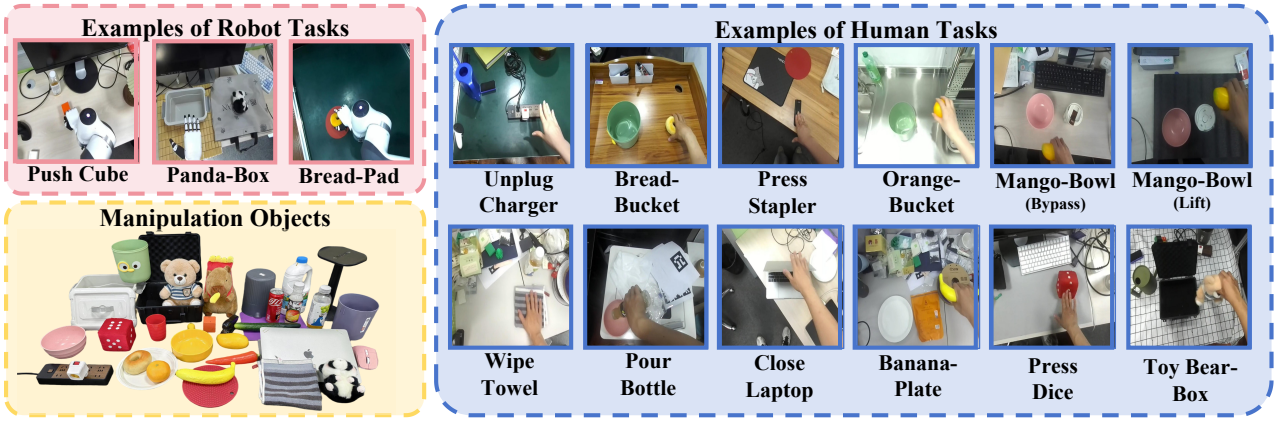

Fig. 3: The *MotionTrans Dataset* comprises 3,213 demonstrations spanning 15 human tasks and 15 robot tasks across more than 10 scenes. For statistical analysis, tasks are grouped by motion-similar skill categories. For human task "Open Box+Panda-Box", it contains both open and pick-place skills.

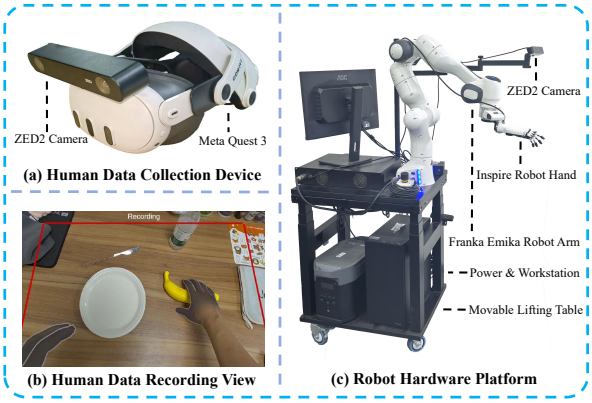

Fig. 4: Illustration of our hardware system, which includes a human VR-based data collection device and a single-arm robot platform. A screenshot of the VR device during human data collection is also provided.

III. Experiments

A. Experiment Setup

Hardware Platform. For the robot hardware (Figure 4(c)), we use a Franka Emika robot arm in combination with a 6DoF Inspired Dexterous (Right) Hand [1]. A ZED2 camera is fixed to the table in an egocentric view to provide an image observation stream. For human data collection (Figure 4(a)), we use the Meta Quest 3 as our VR headset. To ensure consistency in image observations, we also employ a ZED2 camera to record RGB images.

***MotionTrans* Multi-Task Dataset.** Here we introduce the *MotionTrans Dataset*, which is used to train our policies. The dataset contains 3,213 demonstrations across more than 10 scenes, covering 15 human tasks and 15 robot tasks. A brief summary of the dataset is shown in Figure 3. The number of demonstrations for each human / robot task ranges from 40 to 150. To enrich language instructions for VLA training, we leverage GPT-4o [16] to paraphrase and expand task descriptions in the dataset. For tasks, the human and robot task sets are non-overlapping. For motions, similar tasks across human and robot data (e.g., pick-and-place) share similar motion patterns but still exhibit notable differences. In addition, some motions appear only in the human dataset but not in the robot dataset, such as unplugging, closing, lifting, etc.

Evaluation Tasks and Metrics. Since our goal is to understanding the effectiveness of human-to-robot motion transfer, we focus on evaluating robot policies on the human tasks. The list of all evaluated tasks can be found in Figure 5. We use the *Success Rate (SR)* to evaluate the policy performance in accomplishing specific tasks. We also define a *Motion Progress Score (Score)* to quantify the quality of policy motion for task completion.

B. Experiment Results

The goal of the zero-shot experiment is to verify the effectiveness of direct human-to-robot motion transfer. We train policies using our *MotionTrans Dataset*. Subsequently, we directly deploy policies to real robot hardware and evaluate the performance of tasks in human data. We refer to this as zero-shot setting because the policies learn motions from humans without any robot data collected for these human tasks. We seek to answer the following questions:

- (Q1.1) Can policies directly learn to accomplish tasks in human data by human-robot cotraining?
- (Q1.2) For tasks that cannot be accomplished, can the policies learn meaningful motion for task completion?
- (Q1.3) Is cotraining with robot data the key factor for achieving explicit motion transfer?
- (Q1.4) What is the difference in motion transfer effectiveness between different policy architectures?

(Q1.1) *MotionTrans* enables policies to achieve non-trivial success rate across 9 tasks in the human dataset. The results of the zero-shot experiment are shown in Figure 5. We can see that 9 tasks achieve a non-trivial success rate. Among these tasks, pick-and-place tasks account for the vast majority. This can be attributed to (1) the simplicity of pick-and-place motion and (2) the large number of such tasks in our dataset. Notably, for the cases where even if both the pick objects and place targets are not seen in robot tasks (e.g.,

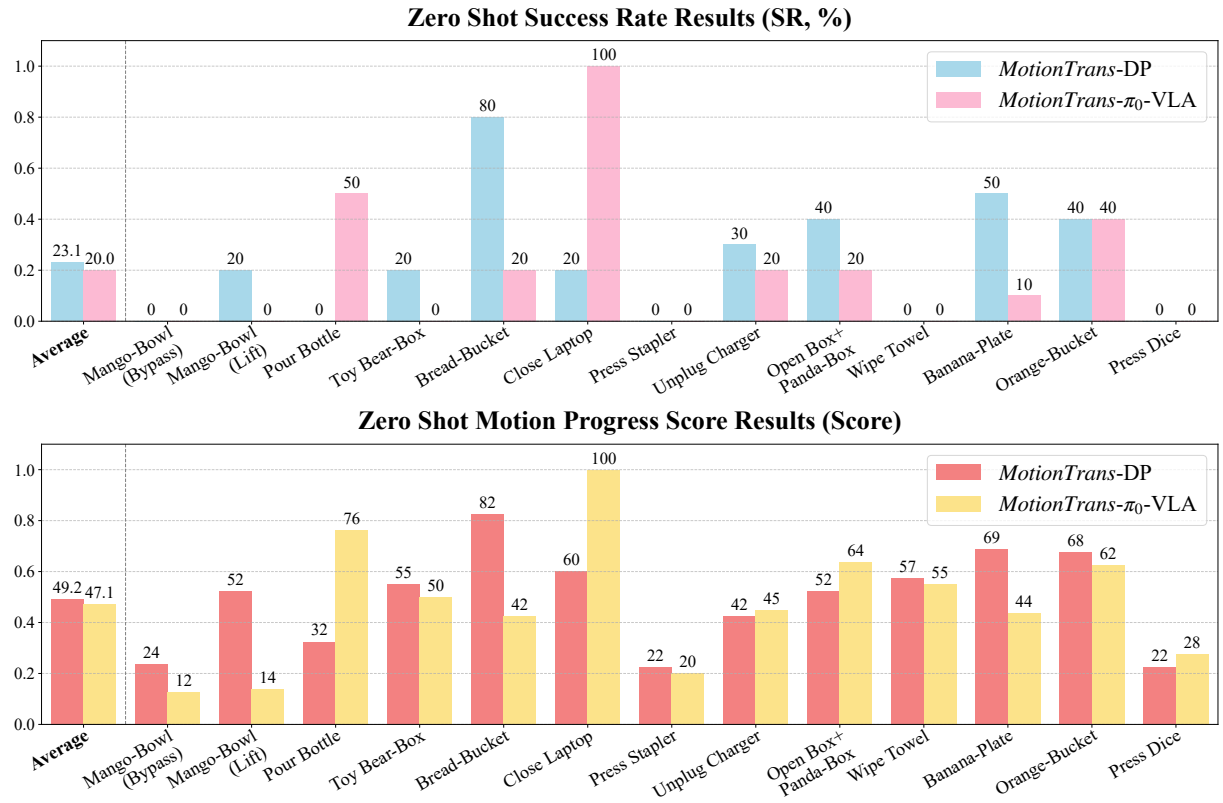

Fig. 5: Results of *MotionTrans* in the zero-shot experiment setting. The results show that both Diffusion Policy (DP) [2] and π_0-VLA [3] achieve successful human-to-robot motion transfer. Even without any robot data for these human tasks, 9 tasks attain a non-zero success rate. For the remaining tasks, *MotionTrans* still generates meaningful motion for task accomplishment, as indicated by a non-trivial Motion Progress Score.

the "Orange-Bucket" task), this type of task-level transfer is still possible. Other accomplished tasks includes motions like pouring, unplugging, lifting, opening and closing (pressing).

(Q1.2) For unsuccessful tasks, *MotionTrans* enables policies to learn meaningful motions toward task completion. Figure 5 shows that both DP and π_0-VLA achieve positive Motion Progress Scores across all tasks, with an overall average of about 0.5. This indicates that the policies are able to complete certain sub-processes for all evaluation tasks. For instance, in the "Wipe Towel" task, both DP and π_0-VLA learn the motion of "push towel forward" to some extent. Moreover, we observe that human data enables the policy to identify spatial locations for almost all human tasks, which is represented as reaching the target manipulated objects (may only appear in human data) to some extent. An example of this is the "Press Stapler" task: although the stapler is not seen in the robot data, the policy still performs approaching behavior.

(Q1.3) Cotraining with robot data is the key factor for successful motion transfer. We find that when robot data is not included for cotraining, the success rate across all tasks is 0% for zero-shot setting. Generally, the policy trained solely on human data exhibits random motion when deployed on the robot. This demonstrates that cotraining with robot data is essential for explicit human-to-robot motion transfer, which could bridge the gap between humans and robots, allowing human motions to adapt to robot embodiment.

(Q1.4) DP and π_0-VLA each have their own advantages (manipulation precision and task adherence). No single model excels across all tasks. On average, the performance of the two models is nearly identical. However, we observe that different models demonstrate their strengths on different tasks. Generally, DP performs better than π_0-VLA in precise manipulation stage, such as grasping, and exhibits stronger spatial location capabilities. In contrast, π_0-VLA shows stronger instruction following for motion generation, such as 'Pour Bottle" task. We hypothesize that the model focuses more on visual perception (DP) tends to achieve greater manipulation precision, whereas the model that emphasizes task semantics and instruction following (π_0-VLA) can adhere to task requirements more stringently.

IV. Conclusion

In this paper, we propose *MotionTrans*, a framework that achieves motion-level learning from human data for end-to-end robot policies. The experiments show that our method achieves explicit human-to-robot motion transfer in a zero-shot setting and significantly improves finetuning performance in a few-shot setting. We hope that the new motion-centric insights that we propose could enhance the utilization of human data in robot policy learning in more effective ways.

References

[1] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[2] 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*, p. 02783649241273668, 2023.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "pi_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

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

[5] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, *et al.*, "Humanoid policy˜ human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[6] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," *arXiv preprint arXiv:2505.11709*, 2025.

[7] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 778–13 790.

[8] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General flow as foundation affordance for scalable robot learning," *arXiv preprint arXiv:2401.11439*, 2024.

[9] 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.

[10] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, H. Yin, S. Liu, *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv:2507.12440*, 2025.

[11] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: Vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[12] H. Bi, L. Wu, T. Lin, H. Tan, Z. Su, H. Su, and J. Zhu, "H-rdt: Human manipulation enhanced bimanual robotic manipulation," *arXiv preprint arXiv:2507.23523*, 2025.

[13] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu, "Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," *arXiv preprint arXiv:2410.08464*, 2024.

[14] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *arXiv preprint arXiv:2307.04577*, 2023.

[15] A. Wei, A. Agarwal, B. Chen, R. Bosworth, N. Pfaff, and R. Tedrake, "Empirical analysis of sim-and-real cotraining of diffusion policies for planar pushing from pixels," *arXiv preprint arXiv:2503.22634*, 2025.

[16] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.