2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Developing Vision-Language-Action Model from Egocentric Videos

Tomoya Yoshida[1,4], Shuhei Kurita[2,3,4], Taichi Nishimura[5], Shinsuke Mori[1]

Abstract— **Egocentric videos capture how humans manipulate objects and tools, providing diverse motion cues for learning object manipulation. Unlike the costly, expert-driven manual teleoperation commonly used in training Vision-Language-Action models (VLAs), egocentric videos offer a scalable alternative. However, prior studies that leverage such videos for training robot policies typically rely on auxiliary annotations, such as detailed hand-pose recordings. Consequently, it remains unclear whether VLAs can be trained directly from raw egocentric videos. In this work, we address this challenge by leveraging EgoScaler, a framework that extracts 6DoF object manipulation trajectories from egocentric videos without requiring auxiliary recordings. We apply EgoScaler to four large-scale egocentric video datasets and automatically refine noisy or incomplete trajectories, thereby constructing a new large-scale dataset for VLA pre-training. Our experiments with a state-of-the-art π_0 architecture in both simulated and real-robot environments yield three key findings: (i) pre-training on our dataset improves task success rates by over 20% compared to training from scratch, (ii) the performance is competitive with that achieved using real-robot datasets, and (iii) combining our dataset with real-robot data yields further improvements. These results demonstrate that egocentric videos constitute a promising and scalable resource for advancing VLA research.**

I. Introduction

Vision-Language-Action models (VLAs) aim to learn general-purpose robot behaviors that follow natural language instructions across environments [3], [4], [5], [6], [7], [8], [9], [10], [11]. Such models are pre-trained with large-scale, multi-embodiment datasets [5], [8], [11] and then fine-tuned on embodiment-specific datasets. However, most pre-training datasets for VLAs heavily rely on human teleoperation, where a number of experts directly manipulate robots to collect instances for imitation learning. This is inherently costly and labor-intensive, leaving a data scarcity problem.

One promising direction to address this problem is to leverage first-person perspective recordings of humans performing everyday tasks, enabled by the recent proliferation of AR/VR devices and smart glasses [12], [13], [14]. Particularly, such egocentric videos provide diverse human-object interactions at a close range and inherently provide motion cues for learning object manipulation. Several studies have begun to explore the use of egocentric videos in robot learning [15], [16], [17], [18]. For example, EgoMimic [16] and EgoVLA [17] leverage enriched egocentric recordings including hand poses to learn robot policies. These studies demonstrate that utilizing egocentric videos is more time- and scale-efficient than teleoperation-based data collection. However, these approaches often depend on dense auxiliary supervision, whose acquisition requires specialized hardware (e.g., multi-camera systems or depth sensors) as well as extensive manual annotation. In a recent study, LAPA [19] attempted to learn latent action representations from egocentric videos. While this approach is scalable due to the absence of auxiliary labels, such latent representations often struggle to capture fine-grained motions. For example, they perform well on simple actions such as pushing but only moderately on more complex skills like pick-and-place.

Considering the limited scalability of rich egocentric recordings and the lack of fine-grained motion cues in egocentric videos, existing methods provide valuable insights but may fall short of sufficiently detailed and diverse action examples for robotic foundation models (see Fig. 1). It is also notable that robot policies trained on diverse real-world egocentric recordings can fall short when evaluated within controlled environments, particularly simulators, due to simplified visual systems [20], [21]. Therefore, although egocentric recordings are valuable resources of human motion cues, they remain underexplored in the existing literature.

To address this issue, we focus on extracting explicit action trajectories, which provide supervision that represents how to move and rotate objects. We leverage EgoScaler [1], a framework designed to extract object manipulation trajectories from egocentric videos. Each pose in a trajectory represents the centroid and rotation of the manipulated object, approximated as the end-effector states of a robot, excluding the gripper. We apply this framework to four large egocentric video datasets, including Ego4D [22], Ego-Exo4D [23], HD-EPIC [24], and Nymeria [25]. The extracted trajectories are then curated by automatically removing noisy or incomplete instances. After this careful filtering process, we construct a new large-scale dataset for VLA pre-training.

We conduct our experiments based on a state-of-the-art π_0 [8] architecture. For comparison, we include three real-robot datasets—BC-Z [26], BridgeData V2 [27], and Fractal [3], which match our dataset in scale and diversity. We evaluate performance in both simulated (SIMPLER [20]) and real-robot environments. Our key findings are threefold:

1) We successfully train π_0 from egocentric videos without auxiliary labels, achieving significant improvements over both training from scratch and LAPA.
2) Our dataset achieves performance on par with leading real-robot datasets, while slightly outperforming BC-Z and BridgeData V2.
3) Combining our dataset with BridgeData V2 yields fur-

[1]Kyoto University, Kyoto 606-8501, Japan
[2]National Institute of Informatics, Tokyo 101-8430, Japan
[3]Institute of Science Tokyo, Tokyo, Japan, Tokyo 152-8550, Japan
[4]NII LLMC, Tokyo 100-0003, Japan
[5]Sony Interactive Entertainment, Tokyo 108-0075, Japan
Contact: yoshida.tomoya.25h@st.kyoto-u.ac.jp

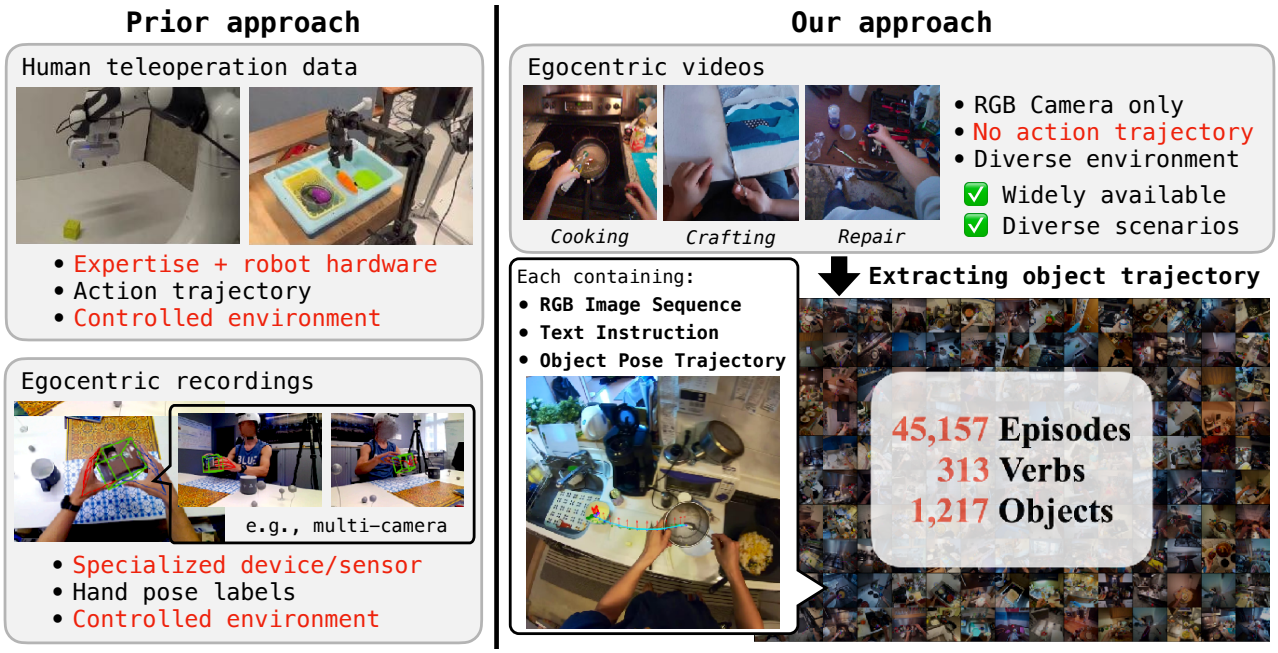


Fig. 1: **Comparison of conventional VLA pre-training sources and ours.** We leverage egocentric videos *without auxiliary labels* for VLA pre-training. Using EgoScaler [1] to extract 6DoF object manipulation trajectories, we construct a large-scale dataset. The multi-camera example is adapted from [2].

ther gains, surpassing the performance of pre-training on either dataset alone.

II. Related Work

A. Dataset for Robotic Foundation Model

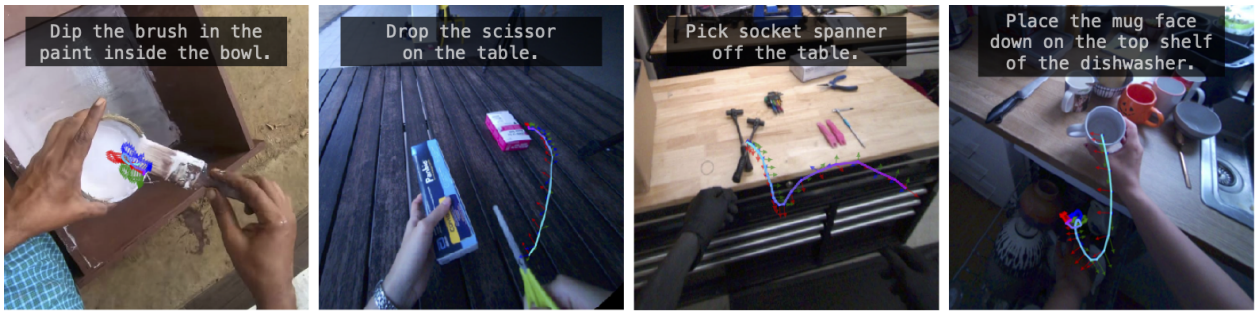
Robotic foundation models have been revolutionizing the robot learning literature, as they enable the execution of multi-purpose and environment-agnostic actions across diverse robotic hardware. Such robotic foundation models, however, rely on large-scale datasets for robotic actions that typically cover multiple robotic hardware and diverse tasks [5], [8], [11], [10]. As constructing such collections requires substantial effort from the robotics community, several datasets have been created through worldwide collaborations [28], [5]. Open X-Embodiment (OXE) dataset [5] aggregates over 60 individual datasets from various institutions into a unified collection that spans diverse manipulation tasks across multiple embodiments. In OXE, they demonstrate that scaling datasets is crucial for improving model performance. Similarly, π_0 model [8] depends on the internal dataset of 'π,' which consists of over 10,000 hours of dexterous manipulation data and makes it possible to execute environment-agnostic and long-horizon tasks. While these datasets have substantially advanced robot learning, the reliance on extensive human effort to construct large-scale robot datasets has become a major bottleneck to progress toward a general-purpose robotic foundation model. To address this inherent limitation of manual dataset collection, we leverage egocentric videos to automatically construct a pre-training dataset for robotic foundation models.

B. Egocentric Video Dataset

Egocentric vision captures fine-grained hand–object interactions, providing motion cues for action understanding [2], [29]. Reflecting its importance, a number of datasets have been introduced over the past decade [22], [23], [30], [31], [32], [33], [24], [34]. For example, Ego4D [22] comprises over 3,000 hours of human daily activity videos spanning hundreds of scenarios, including cooking and crafting. Although egocentric videos remain relatively limited in scale compared to internet videos, the domain is expected to grow with the advancement of AR/VR agents and smart glasses [35], [34], [36]. This anticipated growth will further expand egocentric video collections, underscoring their potential as a scalable resource for robot learning.

C. Robot Learning from Human Demonstration Recordings

Recordings of human everyday tasks in the real world have emerged as a promising source for robot learning, reducing reliance on costly teleoperation-based data collection [15], [19], [37]. In VRB [38] and HRP [39], they extract visual affordances during human-object interaction for grasping tasks from egocentric videos to predict potential interaction regions, providing implicit guidance for robotic dexterous manipulation tasks. Beyond grasping capabilities, a variety of approaches have been explored for more complex manipulation tasks [16], [40], [41], [18], [42], [43], [17], [44], [45]. MimicPlay [40] introduces an approach to learn high-level action plans by imitating hand trajectories from multi-view recordings of human demonstrations. By incorporating these high-level plans as latent representations into robot policies, this work achieves improved performance



Ego4D; *Crafting* Ego-Exo4D; *Experiments* Ego-Exo4D; *Repair* HD-EPIC; *Cooking*

Fig. 2: **Samples of extracted trajectories via EgoScaler [1].** The trajectory is color-coded from cyan (start) to purple (end) to indicate temporal progression. Red, green, and blue arrows represent the X, Y, and Z axes of the object's coordinate frame at each time step.

with minimal robot data. More recently, EgoVLA [17] pre-trains on egocentric recordings including hand poses, significantly enhancing VLA performance compared to training from scratch. Although the utilization of egovision in robot learning seems successful, these existing approaches rely on recordings captured with specialized hardware, such as multi-camera systems, depth sensors, or proprietary devices like Aria Glasses [13], which limits the scalability of the resulting datasets. In contrast, our work focuses on constructing a pre-training dataset from egocentric videos without auxiliary labels, thereby unlocking the scalability of large-scale egocentric video resources for robot learning.

III. Method

Considering the cost and scalability issues of collecting teleoperation data, we leverage egocentric videos for VLA pre-training. However, egocentric videos do not directly provide action trajectories. To address this, we use EgoScaler [1], a framework that extracts object manipulation trajectories from egocentric videos. Applying this framework to large and diverse egocentric video datasets, we construct a large-scale pre-training dataset. As shown in Fig. 1, each episode in this dataset comprises an image sequence, a text instruction, and a 6DoF object pose trajectory. We pre-train a VLA on this dataset and then post-train it on small embodiment-specific datasets.

A. Problem Definition

Following previous studies [46], [8], [47], we train a robot policy that outputs sequences of future robot actions (action chunks). Formally, at each timestep t, given a language instruction ℓ, RGB observations v_t, and proprioceptive state $\boldsymbol{\tau}_t$, we model a policy $\pi(\mathbf{a}_{1:H} \mid v_t, \boldsymbol{\tau}_t, \ell)$ that defines a distribution over the next H actions $\{\mathbf{a}_t, \ldots, \mathbf{a}_{t+H-1}\}$. In pre-training, v_t is a single image, and both $\boldsymbol{\tau}_t$ and $\mathbf{a}_t$ are approximated in an end-effector space without gripper states, derived from egocentric videos. In post-training and evaluation, v_t may comprise images from one or several cameras, depending on the hardware design, and both $\boldsymbol{\tau}_t$ and $\mathbf{a}_t$ are defined in the robot's native control space (e.g., joint space or end-effector space). Despite these gaps, recent studies have demonstrated that VLAs are capable of dealing with such modality differences [5], [8]. At inference time, an action chunk $a_{1:H} \sim \pi(a_{1:H} \mid v_t, \tau_t, \ell)$ is sampled from the trained policy and then executed sequentially.

B. Pre-training Dataset Construction

Re-visiting EgoScaler Framework. EgoScaler [1] extracts a 6DoF object manipulation trajectory from an egocentric video. This framework consists of four stages. First, given a video clip, the start and end timestamps of the action as well as the manipulated object within the scene are identified using GPT-4o [48]. Second, the position sequence of the manipulated object is extracted using an open-vocabulary segmentation model [49], [50] and a dense 3D point tracker [51]. Third, this sequence is projected into the camera coordinate system of the action start frame via point cloud registration [52], eliminating the camera-wearer's movement. Fourth, a rotation sequence is obtained by computing the transformation between consecutive object point clouds using singular value decomposition. Combining these steps yields a sequence of 6DoF poses $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_T\}$, where $\boldsymbol{\tau}_t = (x, y, z, \mathrm{roll}, \mathrm{pitch}, \mathrm{yaw})$. Here, (x, y, z) captures the translational components of the object's centroid position, and $(\mathrm{roll}, \mathrm{pitch}, \mathrm{yaw})$ represents the rotational components of the object. These trajectories represent the object's pose over time, enabling us to approximate the end-effector states during manipulation without gripper states. Fig. 2 illustrates action trajectories across diverse environments.

Egocentric Video Resources. EgoScaler can be applied to various types of egocentric videos. The original paper of EgoScaler targets only Exo-Ego4D [23], but this work expands it to four large egocentric video datasets with diverse activities and scenarios: Ego4D [22], Ego-Exo4D [23], HD-EPIC [24], and Nymeria [25]. These datasets focus on hand-object interactions and provide diverse activities and scenarios. Unlike the other datasets, Ego4D does not provide camera intrinsics required by EgoScaler to reconstruct 3D trajectories. We therefore estimate them in advance using COLMAP [53], [54]. By applying EgoScaler to these datasets, we initially obtained 124,559 episodes.

Data Curation Methods. We found that EgoScaler sometimes produces inaccurate trajectories, mainly due to object detection and point cloud registration errors. To remove

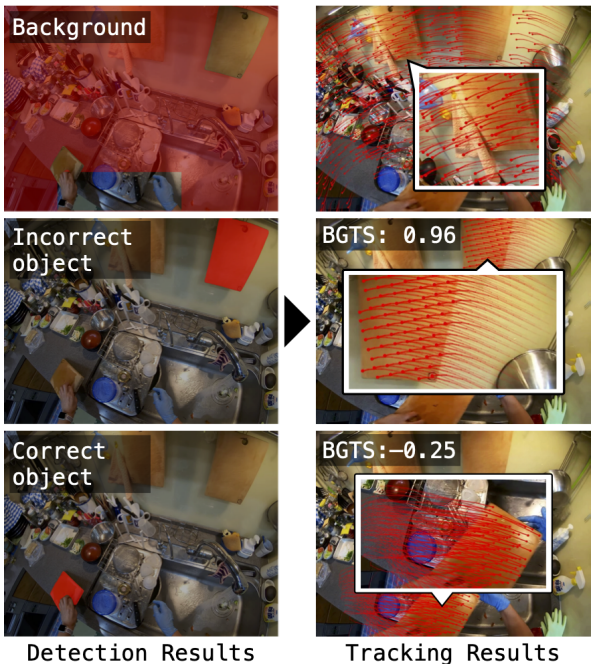


Fig. 3: **Samples of computing background track similarity (BGTS).** Tracked sequences are depicted in red. Low BGTS indicates the object moves due to hand interaction.

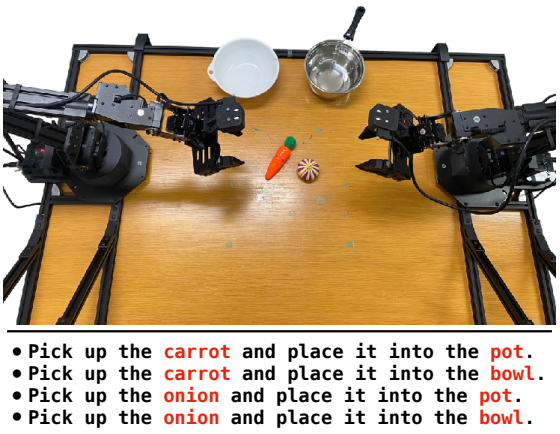


Fig. 4: **Overview of real-robot evaluation setting.**

TABLE I: **Statistics of previous robot datasets and ours.**

Dataset	#Episodes	#Verbs	#Objects
RoboTurk [55]	1,796	2	2
BC-Z [26]	39,350	9	17
BridgeData V2 [27]	53,192	270	749
Fractal [3]	87,212	6	13
DROID [28]	92,233	194	907
Ours	45,157	313	1,217

them automatically, we apply two rule-based filters: a travel distance threshold for registration errors and a background track similarity threshold for detection errors.

For the travel distance threshold (δ_{DT}), we define the travel distance D of a trajectory as the cumulative displacement of its positional component, $D = \sum_{t=1}^{T-1} \|\mathbf{p}_{t+1} - \mathbf{p}_t\|_2$, where $\mathbf{p}_t = (x_t, y_t, z_t)$ denotes the translational element of the object trajectory $\boldsymbol{\tau}_t$. Trajectories corrupted by registration errors often contain abrupt mismatches across consecutive frames, leading to abnormally large D. We therefore discard trajectories with $D > \delta_{\text{TD}}$.

For the background track similarity threshold (δ_{BGTS}), let the object track be $\{\mathbf{o}_t\}_{t=1}^{T}$ and the background track be $\{\mathbf{q}_t\}_{t=1}^{T}$, where $\mathbf{o}_t, \mathbf{q}_t \in \mathbb{R}^2$ are image-plane positions obtained by point tracker [51] within EgoScaler framework. We observed that tracks from detection errors often resemble those of the background, as shown in Fig. 3. This is because detection errors typically occur on non-interacted, static objects. To detect such cases, we compute the background track similarity (BGTS) as the average cosine similarity between the object and background displacements:

$$\text{BGTS} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\mathbf{u}_t \cdot \mathbf{b}_t}{\|\mathbf{u}_t\| \, \|\mathbf{b}_t\|}, \tag{1}$$

where $\mathbf{u}_t = \mathbf{o}_{t+1} - \mathbf{o}_t$ and $\mathbf{b}_t = \mathbf{q}_{t+1} - \mathbf{q}_t$ are velocity vectors from the object and background tracks. We discard episodes with BGTS $> \delta_{\text{BGTS}}$, where we empirically set $\delta_{\text{BGTS}} = 0.7$ based on simulator experiments.

Moreover, due to depth inconsistencies between consecutive frames, the translational components of the extracted trajectories often contain jitter noise. To suppress this, we apply a smoothing filter by averaging each translation vector over a five-frame window centered at the current frame. At sequence boundaries (i.e., $t = 1, 2, T-1, T$), the window size is reduced accordingly, increasing the influence of the central frame. Applying these curation methods, we finally obtain 45,157 episodes. The statistics of our dataset, along with teleoperation-based robot datasets, are shown in Table I.

C. Policy Training

In this work, we use a state-of-the-art VLA π_0 [8]. Built on the pre-trained VLM PaliGemma [56], π_0 employs a flow-matching–based [57], [58] action head that incorporates robot state and generates continuous, high-frequency actions.

Action Representation. During pre-training on our dataset, an action is represented as the displacement of the 6DoF object pose trajectory $\boldsymbol{\tau}$:

$$\mathbf{a}_t = \left[\Delta x_t,\ \Delta y_t,\ \Delta z_t,\ \Delta \text{rot6D}_t\right]. \tag{2}$$

Here, $\{x_t, y_t, z_t\}$ denotes the positional coordinates, and $\text{rot6D}_t \in \mathbb{R}^6$ is a flattened vector of the first two columns of the rotation matrix R_t. Because gripper states cannot be obtained from Section III-B, each action is represented by a 9-dimensional vector. For proprioceptive states, we use the original trajectories $\boldsymbol{\tau}$ for each timestep, converting rotational elements into rot6D representation. To mitigate distribution differences between human and robot data, all actions and proprioceptive states are normalized during training [16].

Dataset Merging for Pre-training. Pre-training VLAs often involves combining multiple robot-embodiment datasets to enable large-scale and diverse training [5], [8]. Following these, when merging our dataset with existing robot datasets, we pad and normalize action and proprioceptive vectors as

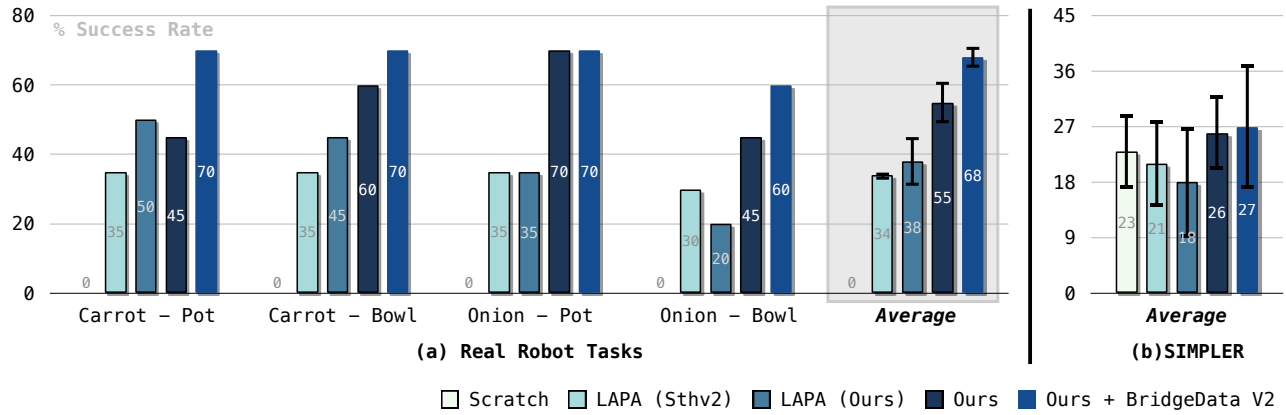


Fig. 5: **Performance of baselines and our dataset across manipulation tasks.** "`ObjectA – ObjectB`" denotes the task "`pick ObjectA and place it into ObjectB`." Parentheses indicate the used dataset for latent action pre-training. Average success rates (%) with standard error are reported.

needed to match dimensionalities across datasets, and then perform joint training on the concatenated data.

Training Objective. The model is trained to predict a sequence of future actions conditioned on language, vision, and proprioception inputs. We minimize the mean squared error between the predicted action $\hat{\mathbf{a}}_t$ and the ground truth action $\mathbf{a}_t$ across a chunk of H future steps:

$$\mathcal{L}_{\text{action}} = \frac{1}{H}\sum_{t=1}^{H} \|\hat{\mathbf{a}}_t - \mathbf{a}_t\|_2^2 . \tag{3}$$

This loss is used for both pre-training and post-training.

IV. Experiments

We evaluate the effectiveness of our dataset as a pre-training source in simulated and real-robot environments. For all experiments, we employ the π_0 [8] architecture. Unlike the conventional settings of fine-tuning a publicly available π_0 checkpoint, we **pre-train** a π_0 model within our dataset.

A. Experimental Setup

Manipulation Task Details. For simulated environments, we use SIMPLER [20] BridgeData V2 environment, which contains four pick-and-place tasks. Inspired by [19], we collected a small amount of successful episodes for post-training purposes. For this purpose, we used a pre-trained π_0 [8] on π [8] dataset and post-trained it on BridgeData V2 [27] dataset. Using the π_0, 25 successful episodes were collected for each task. For evaluation, performance is measured with 200 rollouts for each task.

For real-robot environments, as shown in Fig. 4, we use ALOHA [46] and design four language-aware pick-and-place tasks with four objects. In each rollout, all four objects are present simultaneously, requiring the model to interpret the given instruction correctly. We manually collect 200 episodes per task, totaling 800 episodes. For evaluation, performance is measured with 10 rollouts per task using fixed initial camera and object setups, with minor natural lighting variations. The success rate is calculated using a two-stage scoring scheme: 0.5 points for grasping the correct object, and 0.5 points for placing it in the correct location.

Implementation Details. For dataset reconstruction, we performed parallel processing using multiple single A100 40GB GPUs. We trained π_0 on 8×H200 GPUs using AdamW [59] optimizer with bfloat16 precision under a constant learning rate of 5×10^{-5}. We freeze the pre-trained VLM parameters during both pre-training and post-training. This design, inspired by SmolVLA [9], makes training GPU-friendly and time-efficient. Pre-training was conducted for 20,000 steps with a batch size of 1,024. In evaluation, we use 40,000 steps with a batch size of 128 for the real-robot setting. For the simulator setting, we select the best checkpoint on a validation set, evaluated every 10,000 steps between 10,000 and 50,000 steps.

B. Results

Comparison with Scratch and Ours. The scratch baseline is a model trained only on the post-training dataset. As shown in Fig. 5, our method outperforms the scratch baseline in real-robot tasks, while achieving smaller but consistent gains in simulation. Under an identical architecture and post-training setting, the scratch model attains 0% success in real-robot, indicating a failure to ground instructions and generate meaningful trajectories. The smaller performance gap in simulation is likely due to a visual domain gap, which we discuss in Section IV-C. Moreover, merging our dataset with BridgeData V2 [27] yields additional improvements compared to our dataset alone.

Comparison with LAPA and Ours. LAPA [19] is an implicit pre-training approach that does not rely on auxiliary labels, enabling the use of egocentric videos as VLA pre-training. It learns a discrete set of latent action tokens using a VQ-VAE [60] over temporally separated image pairs, and pre-trains a VLA to predict these tokens from image–text inputs. For fair comparison, we apply the approach to the same π_0 model instead of the original architecture. Experiments are conducted on both our dataset and Something-Something V2 (SthV2) [61], following the original setting.

TABLE II: **Comparison with robot datasets.** Successes out of 10 rollouts are reported for each task, with the final column showing the total.

Dataset	Carrot -Pot	Carrot -Bowl	Onion -Pot	Onion -Bowl	*Total*
BridgeData V2 [27]	4/10	3/10	6/10	4/10	17/40
BC-Z [26]	5/10	5/10	4/10	5/10	19/40
Fractal [3]	7/10	4/10	7/10	4/10	22/40
Ours	4/10	6/10	7/10	4/10	21/40
Ours + [27]	7/10	7/10	7/10	6/10	**27/40**

Fig. 5 presents the performance comparison between LAPA and ours. Our approach consistently outperforms LAPA in both real-robot and simulated environments. Although LAPA outperforms the scratch baseline in real-robot tasks, its performance degrades in simulation, likely due to the simplified visual systems in simulated environments. This suggests that leveraging rich, diverse egocentric video without action labels can harm performance in simulation. Similarly, the greater visual diversity in our dataset is more effective than SthV2 in the real-robot setting, but it leads to performance drops in simulation. These results indicate that explicit action trajectories provide more robust and informative supervision across environments.

Comparison with Robot Datasets and Ours. Table II summarizes the performance of π_0 pre-trained separately on three robot datasets (Fractal [3], BridgeData V2 [27], and BC-Z [26]) and on our dataset in real-robot tasks. This comparison reveals two key findings. First, merging ours with BridgeData V2 outperforms using any robot dataset alone. This result suggests that our dataset can *effectively complement robot data* and serve as a useful component within larger multi-embodiment collections such as OXE [5]. Second, pre-training on our dataset alone yields higher performance than BC-Z and BridgeData V2, but lower than Fractal, owing to its much larger scale. Together, these results highlight the importance of large-scale pre-training for VLAs and demonstrate that our dataset is effective both on its own and when combined with existing robot datasets.

C. Ablation Study

Dataset Scalability. As our framework automatically extracts object manipulation trajectories, we examine how dataset scale affects performance. We utilize data at ratios of 1.0, 0.5, and 0.1 of the full dataset (corresponding to 45K, 20K, and 5K episodes, respectively). Fig. 6 illustrates the results across different dataset sizes. In the real-robot setting, scaling our dataset significantly improves task performance. In contrast, in simulation, the full dataset slightly underperforms the 0.1-ratio subset by 1% on average. The limited improvement from scaling in simulation likely stems from a visual domain gap. While the simulator provides reduced noise and controlled variability, the rich and diverse cues in egocentric videos may hinder performance.

Hyperparameter of Background Trajectory Similarity. Setting an appropriate curation threshold is crucial to balancing the scale and quality of our dataset. We conduct an ablation study, varying the background trajectory similarity threshold $\delta_{\text{BGTS}} = \{0.5, 0.7, 1.0\}$. As shown in Table III, a lower threshold ($\delta_{\text{BGTS}} = 0.5$) removes noisy trajectories but significantly reduces the dataset scale. In contrast, a higher threshold ($\delta_{\text{BGTS}} = 1.0$) retains more episodes but leads to degraded performance due to increased noise. We find that $\delta_{\text{BGTS}} = 0.7$ achieves the best trade-off between dataset size and task performance in simulation experiments. We therefore adopt this value for constructing our dataset.

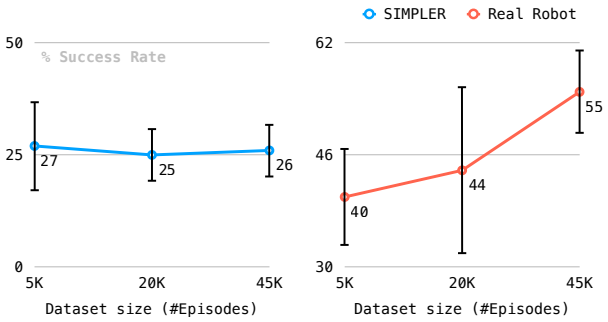


Fig. 6: **Dataset size and performance.** Average success rates (%) with standard error reported.

TABLE III: **Performance comparison across different background track similarity thresholds (BGTS).**

δ_{BGTS}	#Episodes	SIMPLER [20]	Real Robot
0.5	28,719	$25.8_{\pm 5.4}$	$53.8_{\pm 5.2}$
0.7	45,157	$\mathbf{26.0}_{\pm 5.9}$	$\mathbf{55.0}_{\pm 6.1}$
1.0	86,427	$22.5_{\pm 4.8}$	$38.8_{\pm 4.3}$

V. Conclusion

We demonstrated that egocentric videos without auxiliary labels are an effective resource for VLA pre-training. Using EgoScaler, we construct a large-scale dataset by extracting explicit object manipulation trajectories from egocentric videos. Pre-training on this dataset achieves performance competitive with real-robot data and significantly outperforms training from scratch and prior approaches. Moreover, merging our dataset with a single robot dataset proved effective, and extending this to multiple robot datasets remains an interesting direction for future work.

Limitation. Although our dataset is effective for VLA pre-training, it has three main limitations inherited from the EgoScaler pipeline. First, it is limited to manipulation of a single rigid object, leaving multi-object interactions and deformable objects for future work. Second, inaccuracies in some extracted trajectories reduce the usable dataset size, yet this limitation is not fundamental and is expected to improve with future advances in 4D reconstruction and tracking. Third, the dataset does not include gripper state information, which may limit applicability to tasks requiring explicit actuation modeling.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K17983, JP22KK0184 and JST CRONOS JPMJCS24K6.

REFERENCES

[1] T. Yoshida, S. Kurita, T. Nishimura, and S. Mori, "Generating 6DoF Object Manipulation Trajectories from Action Description in Egocentric Vision," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[2] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2O: Two Hands Manipulating Objects for First Person Interaction Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-1: Robotics Transformer for Real-World Control at Scale," in *arXiv preprint arXiv:2212.06817*, 2022.

[4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of The 7th Conference on Robot Learning*, 2023.

[5] Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[6] D. Ghosh, H. R. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, Q. Vuong, T. Xiao, P. R. Sanketi, D. Sadigh, C. Finn, and S. Levine, "Octo: An Open-Source Generalist Robot Policy," in *Robotics: Science and Systems*, 2024.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Proceedings of The 8th Conference on Robot Learning*, 2025.

[8] 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.

[9] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[10] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[11] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[12] Meta, "Quest 3," 2023. [Online]. Available: https://www.meta.com/quest/quest-3/

[13] 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.

[14] Apple, "Apple vision pro," 2024. [Online]. Available: https://www.apple.com/apple-vision-pro/

[15] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A Universal Visual Representation for Robot Manipulation," in *Proceedings of The 6th Conference on Robot Learning*, 2023.

[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.

[17] 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 preprint arXiv:2507.12440*, 2025.

[18] 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.

[19] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, L. Liden, K. Lee, J. Gao, L. Zettlemoyer, D. Fox, and M. Seo, "Latent Action Pretraining from Videos," in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, "Evaluating Real-World Robot Manipulation Policies in Simulation," in *Proceedings of The 8th Conference on Robot Learning*, 2025.

[21] S. Höfer, K. Bekris, A. Handa, J. C. Gamboa, M. Mozifian, F. Golemo, C. Atkeson, D. Fox, K. Goldberg, J. Leonard, C. Karen Liu, J. Peters, S. Song, P. Welinder, and M. White, "Sim2Real in Robotics and Automation: Applications and Challenges," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 398–400, 2021.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 995–19 012.

[23] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, J. Chen, F. Cheng, F.-J. Chu, S. Crane, A. Dasgupta, J. Dong, M. Escobar, C. Forigua, A. Gebreselasie, S. Haresh, J. Huang, M. M. Islam, S. Jain, R. Khirodkar, D. Kukreja, K. J. Liang, J.-W. Liu, S. Majumder, Y. Mao, M. Martin, E. Mavroudi, T. Nagarajan, F. Ragusa, S. K. Ramakrishnan, L. Seminara, A. Somayazulu, Y. Song, S. Su, Z. Xue, E. Zhang, J. Zhang, A. Castillo, C. Chen, X. Fu, R. Furuta, C. Gonzalez, P. Gupta, J. Hu, Y. Huang, Y. Huang, W. Khoo, A. Kumar, R. Kuo, S. Lakhavani, M. Liu, M. Luo, Z. Luo, B. Meredith, A. Miller, O. Oguntola, X. Pan, P. Peng, S. Pramanick, M. Ramazanova, F. Ryan, W. Shan, K. Somasundaram, C. Song, A. Southerland, M. Tateno, H. Wang, Y. Wang, T. Yagi, M. Yan, X. Yang, Z. Yu, S. C. Zha, C. Zhao, Z. Zhao, Z. Zhu, J. Zhuo, P. Arbelaez, G. Bertasius, D. Damen, J. Engel, G. M. Farinella, A. Furnari, B. Ghanem, J. Hoffman, C. Jawahar, R. Newcombe, H. S. Park, J. M. Rehg, Y. Sato, M. Savva, J. Shi, M. Z. Shou, and M. Wray, "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] T. Perrett, A. Darkhalil, S. Sinha, O. Emara, S. Pollard, K. K. Parida, K. Liu, P. Gatti, S. Bansal, K. Flanagan, J. Chalk, Z. Zhu, R. Guerrier, F. Abdelazim, B. Zhu, D. Moltisanti, M. Wray, H. Doughty, and D. Damen, "HD-EPIC: A Highly-Detailed Egocentric Video Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[25] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, *et al.*, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *European Conference on Computer Vision (ECCV)*, 2024.

[26] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning," in *Proceedings of the 5th Conference on Robot Learning*, 2022.

[27] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, A. Lee, K. Fang, C. Finn, and S. Levine, "BridgeData V2: A Dataset for Robot Learning at Scale," in *Proceedings of The 7th Conference on Robot Learning*, 2023.

[28] S. Jayanthi, L. Chen, N. Balabanska, V. Duong, E. Scarlatescu, E. Ameperosa, Z. H. Zaidi, D. Martin, T. K. D. Matto, M. Ono, and M. Gombolay, "DROID: Learning from Offline Heterogeneous Demonstrations via Reward-Policy Distillation," in *Proceedings of The 7th Conference on Robot Learning*, 2023, pp. 1547–1571.

[29] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "HOI4D: A 4D Egocentric Dataset for Category-

Level Human-Object Interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] 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 *European Conference on Computer Vision (ECCV)*, 2018.

[31] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100," *International Journal of Computer Vision (IJCV)*, vol. 130, p. 33–55, 2022.

[32] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] Y. Li, M. Liu, and J. M. Rehg, "In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[34] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, B. Ouyang, Z. Lin, M. Cominelli, Z. Cai, B. Li, Y. Zhang, P. Zhang, F. Hong, J. Widmer, F. Gringoli, L. Yang, and Z. Liu, "EgoLife: Towards Egocentric Life Assistant," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[35] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An Outlook into the Future of Egocentric Vision," *International Journal of Computer Vision*, vol. 132, no. 11, pp. 4880–4936, 2024.

[36] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, N. Joshi, and M. Pollefeys, "HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[37] H. Kim, J. Kang, H. Kang, M. Cho, S. J. Kim, and Y. Lee, "UniSkill: Imitating Human Videos via Cross-Embodiment Skill Representations," *arXiv preprint arXiv:2505.08787*, 2025.

[38] 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 (CVPR)*, 2023, pp. 13 778–13 790.

[39] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta, "HRP: Human affordances for Robotic Pre-training," in *Proceedings of Robotics: Science and Systems*, 2024.

[40] 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," in *Proceedings of The 7th Conference on Robot Learning*, 2023.

[41] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *arXiv preprint arXiv:2503.00779*, 2025.

[42] J. Shi, Z. Zhao, T. Wang, I. Pedroza, A. Luo, J. Wang, J. Ma, and D. Jayaraman, "ZeroMimic: Distilling Robotic Manipulation Skills from Web Videos," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[43] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. Leutenegger, "VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Zero-Shot Robotic Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[44] V. Liu, A. Adeniji, H. Zhan, S. Haldar, R. Bhirangi, P. Abbeel, and L. Pinto, "Egozero: Robot Learning from Smart Glasses," *arXiv preprint arXiv:2505.20290*, 2025.

[45] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from In-the-wild Human Videos using Data-Editing," *arXiv preprint arXiv:2508.09976*, 2025.

[46] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proceedings of Robotics: Science and Systems*, 2023.

[47] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient Action Tokenization for Vision-Language-Action Models," *arXiv preprint arXiv:2501.09747*, 2025.

[48] 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.

[49] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection," in *European Conference on Computer Vision (ECCV)*, 2024.

[50] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," *arXiv:2304.02643*, 2023.

[51] Y. Xiao, Q. Wang, S. Zhang, N. Xue, S. Peng, Y. Shen, and X. Zhou, "SpatialTracker: Tracking Any 2D Pixels in 3D Space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[52] J. Park, Q.-Y. Zhou, and V. Koltun, "Colored Point Cloud Registration Revisited," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[53] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[55] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, S. Savarese, and L. Fei-Fei, "RoboTurk: A Crowdsourcing Platform for Robotic Skill Learning through Imitation," in *Proceedings of The 2nd Conference on Robot Learning*, 2018.

[56] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[57] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[58] Q. Liu, "Rectified flow: A marginal preserving approach to optimal transport," *arXiv preprint arXiv:2209.14577*, 2022.

[59] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[60] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, 2017.

[61] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017.