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

Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos

Qixiu Li[1,2*†], Yu Deng[2*] Yaobo Liang[2*] Lin Luo[2*] Lei Zhou[2†] Chengtang Yao[2]
Lingqi Zeng[2†] Zhiyuan Feng[1,2†] Huizhi Liang[1,2†] Sicheng Xu[2] Yizhong Zhang[2] Xi Chen[2]
Hao Chen[2] Lily Sun[2] Dong Chen[2] Jiaolong Yang[2‡] Baining Guo[2]

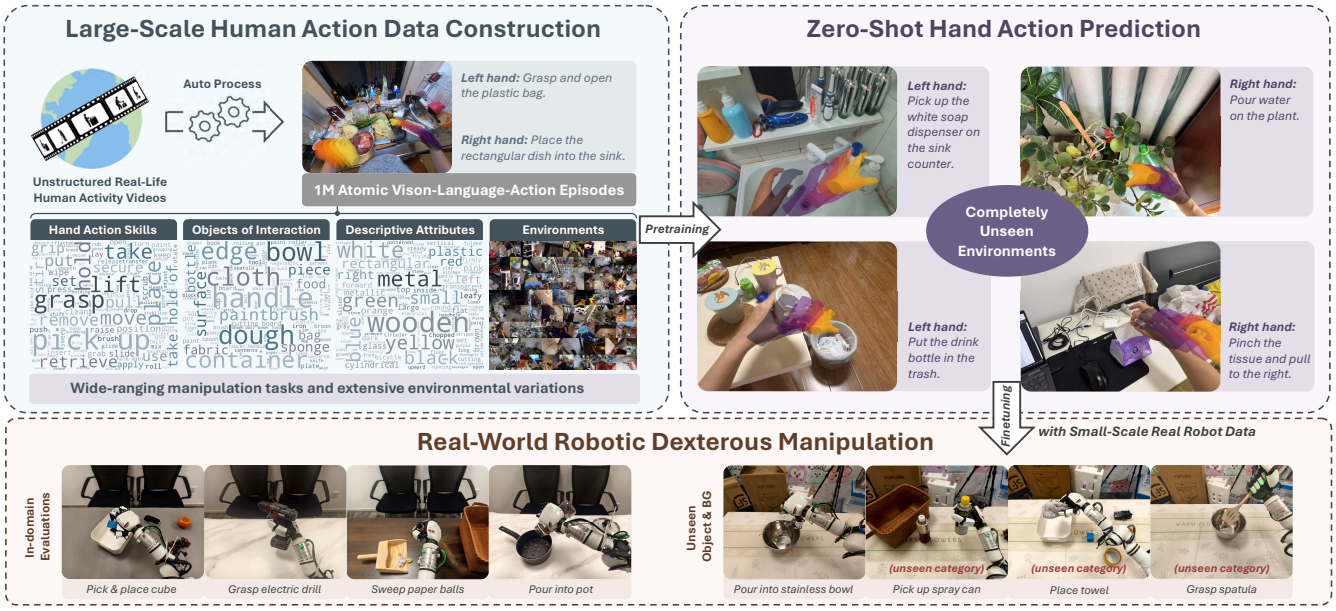

Fig. 1: We present a pretraining approach for robotic Vision-Language-Action (VLA) models by transforming unstructured real-life videos of human activity into structured V-L-A formats aligned with existing robot data. The pretrained model demonstrates strong zero-shot hand action prediction in unseen environments and can be effectively finetuned with dexterous robot hand data for real-world tasks, showing robust generalization to new objects and environments.

***Abstract*— This paper presents an approach for pretraining robotic manipulation Vision-Language-Action (VLA) models using a large corpus of unscripted real-life video recordings of human hand activities. Treating human hand as dexterous robot end-effector, we show that "in-the-wild" egocentric human videos without any annotations can be transformed into data formats fully aligned with existing robotic V-L-A training data in terms of task granularity and labels. This is achieved by the development of a fully-automated holistic human activity analysis approach for arbitrary human hand videos. This approach can generate atomic-level hand activity segments and their language descriptions, each accompanied with framewise 3D hand motion and camera motion. We process a large volume of egocentric videos and create a hand-VLA training dataset containing 1M episodes and 26M frames. This training data covers a wide range of objects and concepts, dexterous manipulation tasks, and environment variations in real life, vastly exceeding the coverage of existing robot data. We also design a dexterous hand VLA model architecture and pretrain the model on this dataset. The model exhibits strong zero-shot capabilities on completely unseen real-world observations. Additionally, fine-tuning it on a small amount of real robot action data significantly improves task success rates and generalization to novel objects in real robotic experiments. We believe this work lays a solid foundation for scaling up VLA pretraining towards generalizable embodied intelligence. The project website, which includes additional visualizations, models, datasets, and code, is available at: https://microsoft.github.io/VITRA/.**

[1]Tsinghua University. [2]Microsoft Research Asia. *Equal contribution. †Intern work done at Microsoft Research Asia. ‡Corresponding author: jiaoyan@microsoft.com

I. INTRODUCTION

Existing Vision-Language-Action data for robotic manipulation are typically collected in laboratory settings through human teleoperations [1]–[5]. Although such robot action data is invaluable, its high acquisition cost significantly limits both the scale of the collected data and its diversity in skills, object categories, and scene variations. Consequently, current V-L-A datasets lag far behind the Internet-scale language and VL data in terms of quantity and diversity, and they fall short of representing the complexity required for real-world robotic tasks. The V-L-A data for dexterous robot hands is even more scarce; to our knowledge there are no large-scale dexterous hand action datasets available for pretraining.

Meanwhile, There is a vast amount of real-life human videos on the web, containing rich examples of everyday

human actions and physical interactions with diverse environments. These videos are typically ***unstructured***: they come unscripted and unsegmented, vary in length and task granularity, contain noisy and irrelevant actions, and lack language instruction and 3D action labels. Although there have been numerous interests in utilizing human video for robot learning [6]–[17], no existing approaches leverage large-scale, unstructured videos without any human annotation for VLA model pretraining. This leads to a critical question: *can we transform these unstructured videos into data formats fully aligned with existing robotic V-L-A training data?*

This work is the first to address this question, and we provide an affirmative answer. For unstructured human videos, we consider human hand as robot end-effector and seek to achieve two types of alignment with real robot V-L-A data. *1) Task alignment*: we need meaningful segmentation and filtering of atomic-level human action sequences (short-horizon tasks), adhering to the recipe of existing robot data. This problem is closely related to temporal action segmentation from videos, which remains an open problem and there are no existing methods that meet our needs. *2) Label alignment*: we need to recover metric-space 3D hand motion accurately to the extent possible[1] to provide dense action labels. This is difficult as we often work with single, uncalibrated, and likely moving cameras. We also need precise language instruction labels to describe the actions.

To this end, we introduce a holistic human activity analytic framework that converts any human hand activity video of arbitrary length into multiple V-L-A trajectories of dexterous manipulation. It is a fully-automatic approach requiring no human intervention. In this framework, we first develop a monocular 3D camera and hand pose tracking approach leveraging recent advancements in 3D vision community, particularly deep visual SLAM, depth estimation, and hand reconstruction. The outputs include the camera FOV, the framewise camera pose, and the framewise hand pose (based on the 6D pose and full joint angles). For temporal atomic action segmentation, we propose a simple yet surprisingly effective algorithm based on the hand movement speed in the 3D space, obtained from the recovered 3D motion labels. Finally, for each segmented video clip, we visualize hand trajectories on sampled video frames and prompt VLM to determine whether the action constitutes meaningful manipulation and, if so, describe it in natural language.

One significant advantage of real-life video data is the inherent action diversity and scene variation it offers. As a starting point, we process a large volume of raw videos from existing egocentric human video datasets. The resultant hand V-L-A dataset contains about 1 million episodes and 26 million frames, which captures a broad spectrum of objects, concepts, skills, and environmental variations, vastly exceeding the coverage of existing robot data. We also develop a dexterous hand VLA model architecture with a Causal Action Transformer and pretrain the model on this dataset. The model exhibits strong zero-shot capabilities on observations of completely new scenes. We also conduct real-world robot experiments and show that fine-tuning the model on a small amount of real robot hand data significantly improves task success rates and generalization to novel objects and backgrounds.

Our work stands distinct from prior research that utilizes human video for training robotic manipulation models. The approaches that leverage egocentric human video for learning vision and language representations, affordances, point trajectories, *etc.* [7]–[9], [12]–[14] did not explore action pretraining for VLA models. Recent works that use latent actions from human videos [15]–[17] for pretraining do not provide explicit 3D action labels as we do. Our experiments demonstrate the superiority of our pretraining approach. Most recently, a few works that are concurrent to ours studied training VLA models with explicit 3D hand motions similar to ours [18]–[20], but their data is largely limited to scripted laboratory captures; a detailed discussion is provided in the next section.

Our approach offers a more tractable way for pretraining data scaling compared to existing techniques. Although this work uses videos from existing egocentric video datasets, there are no technical barriers preventing further data scaling. By not imposing constraints on the subjects' activities or environments and requiring only a single webcam, every life recorder can effectively become a robot teacher. We envision a future where robots can effectively learn from abundant, low-cost human video demonstrations to acquire diverse skills, complemented by targeted fine-tuning using a modest amount of real robot data or reinforcement learning. ***Our training dataset and pretrained VLA models are open-sourced to the community to facilitate further research.***

II. RELATED WORKS

a) Robotic VLA Model Pretraining: Robotic VLA models [1], [21]–[30] that can perform diverse language-instructed tasks typically need pretraining on large data. Incorporating VL-pretrained modules or backbones has been a common practice for VLA models, and here we focus on a brief overview of the pretraining with regard to *the action modality* or its proxy. Most recent VLA models with action pretraining [22], [23], [25], [26], [28], [29] have leveraged the Open X-Embodiment (OXE) dataset [2], which contains over 1M real robot trajectories collected on over twenty robots. This large-scale dataset provides diverse skills and environment variations well suited for pretraining. Some of these works [25], [27]–[29] also incorporate more open-source or in-house robot action data in addition to OXE. The work of [31] synthesized a large volume of V-L-A data in simulators for pretraining, but it handles the grasping task only. A line of works [3], [15]–[17], [32] studied learning latent action from human and/or robot videos in an unsupervised manner and pretraining models using the extracted latents as the proxy for action. Some other works

[1]While it's ideal to have 3D motion labels as accurate as possible, we believe some noise and imperfection are acceptable for pretraining, where the goal is to grasp common knowledge, learn motion patterns for diverse skills, and experience a wide spectrum of object and scene variations.

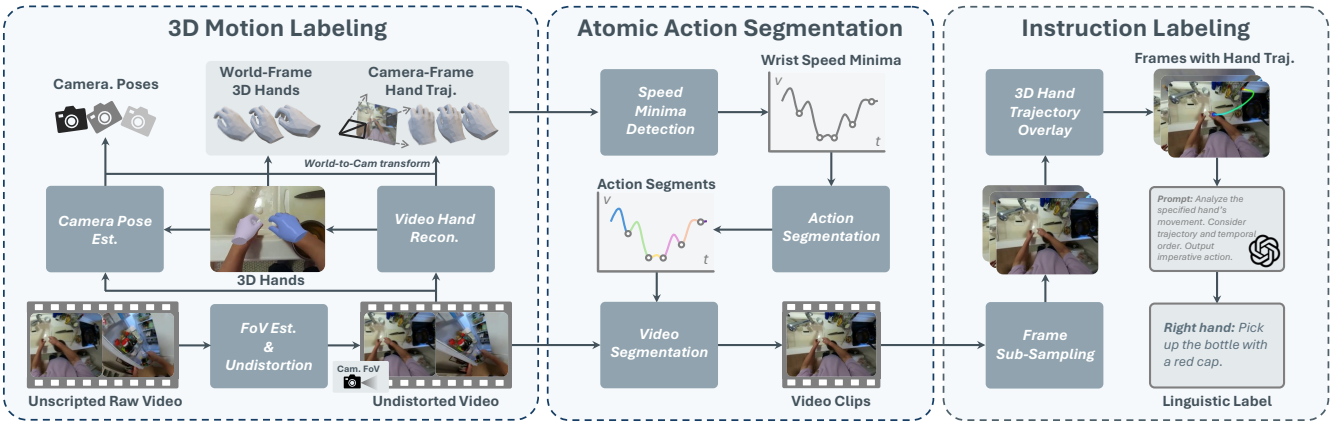

Fig. 2: Overview of our holistic human activity analysis framework, which transforms unscripted real-life human videos into VLA episodes of human hands aligned with typical robotic data.

propose to use the future frames in videos as the prediction target for pretraining [33], [34].

There are some recent attempts concurrent to ours which use 3D hand action labels of egocentric human videos for VLA pretraining [18]–[20]. They primarily use hand-object interaction videos captured in controlled environments with privileged information. For example, the videos are well segmented to language-instructed action clips since the tasks are pre-scripted, and 3D hand motions are typically obtained with advanced devices (*e.g.*, RGBD sensors, VR/AR headsets). We focus on a different goal, *i.e.*, harnessing unscripted real-life human videos for large-scale pretraining, which encompass a significantly broader range of tasks, objects, and real-world environments. This greatly enhances the zero-shot action perdition performance. Furthermore, their casual capture nature facilitates much greater scalability.

b) Dexterous Hand Manipulation: Dexterous manipulation with multi-fingered robot hands has been a vibrant area of research for decades. Earlier learning-based models with visual inputs were typically trained with reinforcement learning in simulators [35]–[37]. However, training dexterous RL policies requires sophisticated reward design and their applicability in real-world scenarios is often limited. Using human teleoperated demonstrations for imitation learning was also widely used to improve task performance [38], [39]. Methods that utilize human hand motion as demonstration data [6], [10], [40]–[44] were also actively studied. These previous works typically address a single or small set of tasks for a trained model. Recently, language instruction has been incorporated into dexterous hand manipulation models to handle more tasks with diverse objects [12], [45], [46].

c) Robot Learning from Human Videos: Exploiting human videos to train robotic models has been actively studied in recent years. Several studies [7]–[9], [47] leverage egocentric human videos for learning vision and language representations. Some methods use explicit human actions extracted from mocap videos [6], [10], [42], [43], [48], [49] or web videos [11], [50] to guide robot policy training with imitation learning frameworks. Instead of using explicit motions, other approaches learn affordances [13], [51]–[53], point trajectories [14], [54], [55], or hand-object masks [56] from human videos. Recently, a group of methods have emerged which learn latent actions from human videos in an unsupervised manner and pretrain action model with latent action labels [15]–[17], [57], [58]. Some recent attempts use extracted 3D hand action labels from egocentric human videos for VLA pretraining [18]–[20]. As mentioned earlier, these primarily involve videos captured in controlled environments with privileged information. In a different vein, some approaches utilize human videos to train video generation models for human-to-robot video transfer [59], [60], visual task planning [12], [61], or world models [16], [62], [63].

d) Temporal Action Segmentation for Videos: Temporal action segmentation, also known as temporal action detection or localization, is a technique for detecting action windows and classifying them from a long human video. Earlier approaches [64]–[68] have focused on predefined action classes. Recently, video-input VLMs [69], [70] with broad action understanding capabilities are proposed but they still face challenges in action localization accuracy. They do not meet our requirements in our preliminary tests.

III. Transforming Human Hand Video to VLA Data

Existing robotic manipulation V-L-A data [1]–[5] typically comprise simple, short-horizon tasks (*e.g., "pick up the sponge on table"*, *"wipe the stove with cloth"*), which can be composed to long-horizon tasks by a high-level planner. Each data episode comprises a language instruction, a video frame sequence, and frame-aligned 3D action chunks of the end-effector in the robot or camera coordinate system. Our approach analyzes an unscripted human video and generates V-L-A data in such format, treating the two human hands as the end-effector. The whole framework comprises three stages and an overview is shown in Fig. 2.

A. 3D Motion Labeling

The first stage of our approach extracts 3D motions from videos, including the motions of two hands and the camera. To achieve this, we first apply a simple algorithm to determine whether the camera is static or moving based on

background optical flow. Then we estimate camera intrinsics of the videos by applying DroidCalib [71] for moving cameras and MoGe-2 [72] and DeepCalib [73] for static cameras. The videos are then undistorted to conform to the pinhole camera model. Given the intrinsics and undistorted video, we proceed with video hand reconstruction and camera pose tracking. For the former, we employ HaWoR [74] to reconstruct per-frame camera-space 3D hands. Each reconstructed hand contains wrist 6D pose and joint angles represented with the MANO [75] parametric model. To track camera pose for moving cameras, we apply a modified version of MegaSAM [76] where we replace the depth estimation model providing depth priors to be MoGe-2 [72]. Then we can obtain a sequence of world-space 3D hands by combining the camera-space 3D hands and camera poses.

The world-space 3D hand sequence can be easily transformed into any video frame's camera space, effectively simulating a static camera as in most robot data. Moreover, it facilitates both the subsequent atomic action segmentation and instruction labeling, as will be described later. To enhance efficiency, we chop long videos into overlapping 20-second clips in this stage and recompose their results.

B. Atomic Action Segmentation

This stage aims to segment out simple, atomic-level hand action sequences from a long video. Our core idea for achieving this is to leverage the recovered 3D hand motions. During action transitions, human hands typically exhibit speed changes, with minima often indicating switches of action. This observation has inspired us to design the following algorithm which is simple yet surprisingly effective: we detect speed minima of the 3D hand wrists in the world space and use them as cutting points. We smooth the hand trajectory and select points that are local speed minima within a fixed window centered on each point. Segmentation is applied for the left and right hands independently with the other hand's motion ignored. This way, each segment captures the complete atomic action of at least one hand.

It is worth noting that this method is highly efficient and requires no additional model inference or pre-annotated text labels, making it particularly effective for the scalable segmentation of hand activity videos.

C. Instruction Labeling

Given the video segments and 3D hand action sequences, we create visualizations and utilize GPT-4.1 [77] for action captioning. From each segment, we evenly sample 8 frames and highlight hand trajectories on each frame by projecting the world-space trajectory of the hand palm from the current frame to the end of the clip (see Fig. 2 for an example). These frames are then fed into GPT, which is prompted to describe the specified hand's action in imperative form, taking into account both the content of the frames and the overlaid trajectories. We also instruct GPT to label clips lacking semantically meaningful action as "N/A".

Hand V-L-A Dataset Construction. We construct a large-scale human hand V-L-A datast by processing egocentric human videos from Ego4D [78], Epic-Kitchen [79], EgoExo4D [80], and Something-Something-V2 (SSV2) [81]. Note that *the human annotations for actions provided by these datasets are NOT used in this work; instead, we process the raw videos through our framework*. These annotations often do not match the desired task granularity or they lack precise start and end times for actions. Later we'll show in our experiments that training using these annotations results in obvious performance degradation compared to our approach. Our constructed dataset contains over 1M episodes with 26M frames (77% from Ego4D, 12% from Epic-Kitchen, 6% from EgoExo4D, and 5% from SSV2). It features diverse hand actions, objects, attributes, and environments, encompassing real-life activities such as cooking, cleaning, construction, repairing, crafting, and painting (Fig. 1).

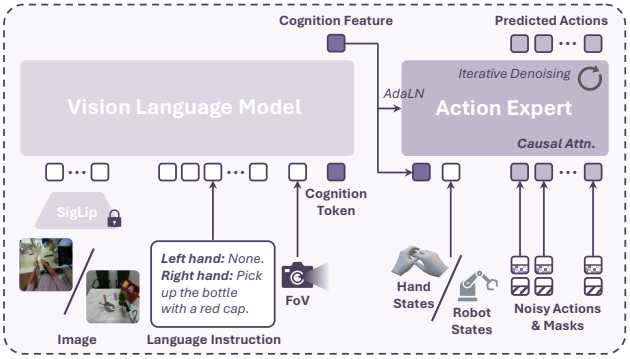

Fig. 3: Our VLA model architecture.

IV. Dexterous Hand VLA Model

We construct a VLA model $\boldsymbol{\pi}$ for dexterous manipulation:

$$\boldsymbol{\pi} : (\boldsymbol{l}, \boldsymbol{o}_t, \boldsymbol{s}_t) \rightarrow (\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, ..., \boldsymbol{a}_{t+N}), \tag{1}$$

which predicts a sequence of future end-effector actions $\boldsymbol{a}$ based on the current visual observation $\boldsymbol{o}_t$, the robot proprioceptive state $\boldsymbol{s}_t$, and a language instruction $\boldsymbol{l}$.

A. VLA Model Design

1) Model Architecture: An overview of our model architecture is presented in Fig. 3. Our model consists of a VLM backbone and a diffusion action expert. We use PaliGemma-2 [82] as the VLM, which combines a SigLIP [83] vision encoder with linear projection for alignment and a Gemma-2 [84] language model. We use the 3B-parameter model with an input image resolution of 224^2 as the default setting. We further incorporate camera FoV information as an extra token to the model. Following [26], we append a learnable "cognition" token as extra input to the VLM, whose output feature $\boldsymbol{f}^c$ serves as the condition for the action expert.

For the action expert, we apply a Diffusion Transformer (DiT) [85] and the DiT-Base model is used. The input is a concatenation of the cognition feature $\boldsymbol{f}^c$, the hand state $\boldsymbol{s}_t$, and a noisy action chunk $(\boldsymbol{a}_t^i, \boldsymbol{a}_{t+1}^i, \ldots, \boldsymbol{a}_{t+N}^i)$, where i denotes the denoising step. The hand state includes the wrist translation and rotation in camera space as well as hand joint angles. We additionally inject the cognition feature into the DiT using AdaLN [85] for enhanced conditioning. The

action expert predicts the added noise for iterative denoising, trained by optimizing an MSE loss $\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),i}||\hat{\boldsymbol{\epsilon}}^i - \boldsymbol{\epsilon}||_2$ where $\hat{\boldsymbol{\epsilon}}^i$ and $\boldsymbol{\epsilon}$ denote the predicted and ground-truth noise, respectively. The action expert and the VLM are trained end-to-end except for the vision encoder which remains frozen.

2) Hand Action Space: Our model predicts hand actions in the camera coordinate frame of the current observation $\boldsymbol{o}_t$. At time step t, the hand action $\boldsymbol{a}_t$ is defined as:

$$\boldsymbol{a}_t = [\Delta\boldsymbol{t}^l, \Delta\boldsymbol{r}^l, \boldsymbol{\theta}_h^l, \Delta\boldsymbol{t}^r, \Delta\boldsymbol{r}^r, \boldsymbol{\theta}_h^r] \in \mathbb{R}^{102}, \tag{2}$$

where $\Delta\boldsymbol{t} \in \mathbb{R}^3$ and $\Delta\boldsymbol{r} \in \mathbb{R}^3$ are the relative wrist translation and rotation between the consecutive frames, and $\boldsymbol{\theta}_h \in \mathbb{R}^{15\times3}$ represents the angles of 15 joints in the local frame of the MANO hand model (see Fig. 5). Superscripts l and r indicate the left and right hand, respectively.

3) Unified Single- and Dual-Hand Action Prediction: Our VLA pretraining data is at the level of single-hand atomic actions, with some episodes containing overlapping dual-hand actions. We introduce the following designs to handle different cases in a unified manner. Specifically, the VLM always receives language instructions in the format of `Left hand: <left-hand action>. Right hand: <right-hand action>`. For a video frame $\boldsymbol{o}_t$, left- and right-hand action descriptions are set to either `None` or the instructions of the corresponding atomic action chuck $\boldsymbol{o}_t$ falls within, with extra action masks (0 or 1) concatenated with noisy actions as input to the action expert (see Fig. 3).

4) Causal Action Denoising: Human hands move fast in real life activities, and many of the action clips in our pretraining dataset are as short as 1 second (∼30 frames). Consequently, many prediction chunks of the VLA model go beyond the episode end for a reasonable chunk length (*e.g.,* $N = 16$ in our setting). Naively padding with zero actions at the end can be problematic, as many atomic actions occur mid-task and should not conclude with no motion (*e.g.*, consider a wiping task where a hand moves back and forth). To address this issue, we employ *causal attention* for action denoising, ensuring that the token of each action step only attends to preceding actions. This prevents zero-padded positions from affecting earlier predictions in the bidirectional attention setting.

B. Pretraining with Human Hand VLA Data

We apply *trajectory-aware augmentation* to the training images and actions to enhance generalization. Specifically, input images are randomly cropped and perspective-warped with varying FoV, aspect ratio, and crop center, while keeping the principal point at image center. The action sequences are transformed to match the augmented camera parameters. During random cropping, we ensure that the projected hand trajectory from the current frame to episode end remains within the cropped image. Using this strategy, the objects of interaction are also mostly well-contained. We also apply random image flipping and make corresponding adjustments to hand actions and language instructions.

TABLE I: Evaluation and ablation study of hand action prediction for the pretrained model. Note that [19] is a concurrent work to ours. See text for details.

Method	Grasp	General action
	Avg./med. $d_{\text{hand-obj}}$ (cm) ↓	User Score ↑
Initial position	20.0 / 20.0	–
Being-H0 (8B) [19]	19.1 / 18.4	0.15
Ablations		
Lab data (EgoDex)	17.6 / 18.3	–
Human annotation	14.1 / 14.1	0.96
No augmentation	11.6 / 10.7	1.43
Bidirectional attention	9.3 / 7.2	1.69
Ours	**8.8 / 6.2**	**1.91**

Fig. 4: Examples of environments used in evaluation.

C. Fine-tuning for Robotic Dexterous Manipulation

After pretraining, the model can be fine-tuned on robot data for deployment. We consider the human hand action space as a superset of that of the robot hand and align the robot's action space with the human hand's as defined in Eq. (2). Specifically, robot end-effector 6D poses in camera coordinates are used to compute $\Delta\boldsymbol{t}$ and $\Delta\boldsymbol{r}$. For joint angles, a simple mapping strategy is applied: each joint of robot hand is mapped to its closest human joint in topology, and the corresponding dimension in human action $\boldsymbol{\theta}_h$ is used for fine-tuning. Unmapped dimensions in $\boldsymbol{\theta}_h$ are zero-padded and marked invalid in the action mask.

A Remark. Action space mapping between human hand and dexterous robot hand have been actively studied in the past [40]–[42]. We do not perform direct pose transfer (as done in teleoperation) and our fine-tuning can help mitigate the action space differences. Other strategies can also be employed and we leave it as our future work.

V. EXPERIMENTS

Computational Cost. The 3D motion labeling is conducted using 200 NVIDIA V100 GPUs and requires approximately one week. For model training, the pretraining stage is performed on 8 NVIDIA H100 GPUs with a batch size of 512 and requires approximately two days. The fine-tuning stage is conducted on 8 NVIDIA H100 GPUs with a batch size of 256 and takes around 8 hours. For real-robot deployment, the policy runs on a single NVIDIA RTX 4090 GPU.

A. Human Hand Action Prediction

1) Benchmark: We construct a benchmark to evaluate the pretrained VLA model for human hand action prediction in *unseen environments*. Two types of tasks are designed: *a) Grasping:* We instruct the model to grasp objects in the scene. We capture RGB-D images from 47 unseen environments and annotate 396 objects with captions and segmented 3D point clouds. Synthetic human hands are rendered onto the images at distances suitable for object grasping with a single action chunk (see Fig. 4). We compute

TABLE II: Success rates on real-world robot dexterous manipulation tasks (in %) .

Method	Seen				Average	Unseen Object & Background			Unseen Category	Average
	Pick & place (40 trials)	Functional grasp (24 trials)	Pour (8 trials)	Sweep (8 trials)		Pick & place (16 trials)	Functional grasp (16 trials)	Pour (8 trials)	Pick & place (24 trials)	
VPP [12]	57.5	29.2	12.5	0.0	24.8	12.5	0.0	0.0	8.3	5.2
π_0 [25]	37.5	25.0	**75.0**	50.0	46.9	0.0	6.2	25.0	33.3	16.1
No VLA pretrain	32.5	33.3	12.5	50.0	32.1	31.2	0.0	0.0	12.5	10.9
Latent action pretrain	42.5	41.7	37.5	**62.5**	46.0	0.0	0.0	0.0	0.0	0.0
OXE pretrain	40.0	37.5	62.5	25.0	41.3	12.5	6.3	0.0	12.5	7.8
Ours	**80.0**	**66.7**	**75.0**	**62.5**	**71.0**	**68.8**	**68.8**	**50.0**	**70.8**	**64.6**

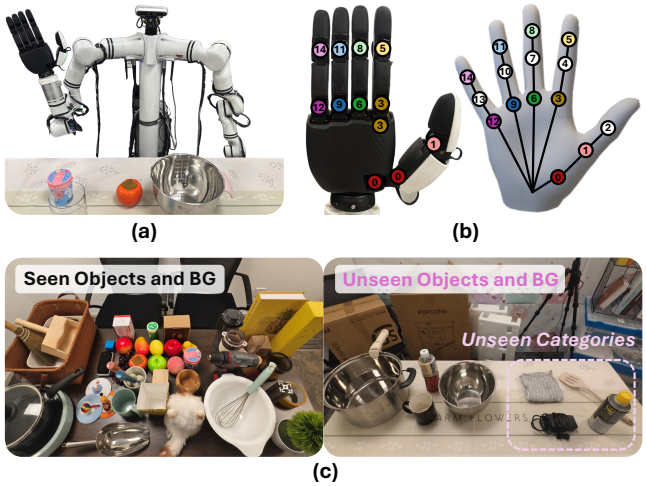

Fig. 5: (a) Robot setup. (b) Mapping between XHand and MANO, where joints sharing the same index indicate correspondence; white color denotes joints without counterparts. (c) Fine-tuning objects/environment and unseen objects/background for evaluation.

the minimum distance between predicted finger trajectories and target object points to evaluate movement plausibility. *b) General action:* We also evaluate the hand movements before and after contact using 117 unseen real-life environments captured with mobile phones (Fig. 1). For each scene, we prompt the model with annotated instructions and render predicted hand actions onto video frames. We conduct a user study with 23 participants and ask them to rank the top-3 actions for 30 randomly selected scenes. These actions will be assigned 3, 2, and 1 scores while 0 will be assigned to others. We report average scores across participants.

2) Baseline Methods: We compare our model with several baselines to validate the efficacy of our data and pretraining designs, including *a) Lab data*, which replaces our VLA data with the EgoDex [86] videos containing over 300K hand VLA episodes captured in lab environments; *b) Human annotation*, which uses human annotations in the original human video datasets for constructing VLA episodes (as mentioned previously, these annotations often do not match the desired task granularity or there's no precise start and end times); *c) No augmentation*, which discard the trajectory-aware data augmentation; *d) Bidirectional attention*, which uses bidirectional attention for the action expert; *e) Being-H0* [19], a recent hand VLA model pretrained on a large collection of scripted, laboratory human video data.

3) Results: Figure 1 shows our model's action predictions in these unseen environments, demonstrating its strong generalization to diverse scenarios. Table I reports the quantitative results. For *grasping*, we include the initial hand–object distance as a reference. For *general action*, the *Lab data* baseline is not included because the model only predicts keypoints using labels provided by EgoDex. Compared to models trained on EgoDex data, ours exhibit much stronger generalizability. Training with the original human annotations also underperforms, as its temporal or granularity misalignment between text and actions weakens instruction following. Replacing causal attention with bidirectional attention or removing augmentation reduces performance, highlighting the importance of these technical designs.

B. Real-World Robot Dexterous Manipulation

1) Robot Setup: We use a Realman robot with 12-DoF XHand dexterous hands and a RealSense head camera, as in Fig. 5 (a). The joint mapping between XHand and human hand for fine-tuning is shown in Fig. 5 (b).

2) Task Designs: We collected 1.2K teleoperated trajectories for four tasks: *i) General pick & place* – moving an object into a box with 3–4 distractors; *ii) Functional grasping* – grasping an object at a functional location (*e.g.*, handle); *iii) Pouring* – picking a bottle, pouring into another, and placing it back; *iv) Sweeping* – picking a broom, sweeping trash into a dustpan, and returning it. For evaluation, we perform the above four tasks in both *seen* and *unseen* settings (see Fig. 5 (c) for seen and unseen objects and backgrounds): *a) Seen:* Objects and backgrounds were observed during fine-tuning, with randomized positions and distractors. *b) Unseen:* Novel objects and backgrounds for evaluation, with two additional settings: *Unseen Objects*, where the objects are new but other objects of the same categories were seen in fine-tuning; and *Unseen Categories*, where the objects belong to categories not encountered before.

3) Results and Comparisons with Prior Art: Some representative execution results of our method are presented in Fig. 1. For quantitative comparison, we compare our method with several approaches: *a) No VLA pretrain*, directly fine-tuning from our model (initialized with Paligemma-2 VLM) without human data pretraining; *b) Latent action pretrain*, replacing 3D action labels with latent actions from LAPA [15] trained on our data; *c) OXE pretrain*, using Open X-Embodiment [2] data instead of our human VLA data; *d) VPP* [12], a recent dexterous hand manipulation model leveraging diffusion-based video generation [87]; and *e)* π_0 [25], a VLA model pretrained on large-scale robot data.

Table II compares the performance of different methods. Our method outperforms all other baselines, including latent

action pretraining on the same dataset. Compared to latent action, our approach provides more explicit action supervision and smaller pretraining-finetuning gap. In our tests, the VPP model lags behind LLM-based VLA models in instruction following and unseen object recognition. Compared to π_0 and OXE pretraining, our method achieves substantially stronger few-shot and unseen task performance. The OXE dataset contains data from gripper-based robots and offers far less diversity in objects, tasks, and environments compared to our human hand V-L-A dataset. Our model demonstrates robust generalization to unseen objects and environmental changes and even for objects from unseen categories, highlighting the effectiveness of leveraging human activity data for generalizable VLA learning.

VI. CONCLUSIONS AND LIMITATIONS

This paper introduces a novel approach for pretraining robotic manipulation VLA models using unstructured real-life human activity videos. We develop a fully-automatic pipeline to convert in-the-wild egocentric human videos into atomic-level VLA data aligned with existing robotic demonstrations. We also design a dexterous hand VLA model with tailored training strategies to effectively leverage human data for pretraining. Experiments show that our pretrained model exhibits strong zero-shot performance in unseen real-world environments and high task success after finetuning on limited robot data, demonstrating a highly promising and scalable approach for generalizable VLA pretraining.

Despite these encouraging results, our approach has several limitations. First, the accuracy of the proposed data transformation pipeline depends on the annotation tools, which inevitably introduce noise. In particular, the hand pose estimation in the camera coordinate frame relies on HaWoR [74], whose performance inherently determines our reconstruction precision. For the accuracy of atomic action segmentation and instruction labeling, we randomly sample 500 episodes and perform manual verification. Among them, 88% episodes have reasonable segmentation and descriptions. We also observe several common failure modes in the data transformation pipeline. (1) The reconstructed hand pose tends to exhibit larger errors with partially occluded hands. (2) The pipeline may mistakenly reconstruct hands belonging to different individuals, when multiple people's hands appear in the video. (3) Hands wearing gloves of unusual colors may fail to be detected. (4) GPT-4.1 [77] occasionally misidentifies finger types, such as misidentifying the middle finger as the index finger. Addressing these limitations remains an important direction for future work.

REFERENCES

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023.

[2] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *ICRA*, 2024.

[3] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, X. He, X. Huang, *et al.*, "AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," in *IROS*, 2025.

[4] A. Khazatsky, K. Pertsch, S. Nair, *et al.*, "DROID: A large-scale in-the-wild robot manipulation dataset," *arXiv:2403.12945*, 2024.

[5] H.-S. Fang, H. Fang, Z. Tang, J. Liu, *et al.*, "RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot," in *ICRA*, 2024.

[6] Y. Qin, Y.-H. Wu, S. Liu, *et al.*, "DexMV: Imitation learning for dexterous manipulation from human videos," in *ECCV*, 2022.

[7] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv:2203.06173*, 2022.

[8] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "VIP: Towards universal visual reward and representation via value-implicit pre-training," in *ICLR*, 2023.

[9] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, *et al.*, "R3M: A universal visual representation for robot manipulation," in *CoRL*, 2023.

[10] 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 *CoRL*, 2023.

[11] K. Shaw, S. Bahl, and D. Pathak, "VideoDex: Learning dexterity from internet videos," in *CoRL*, 2023.

[12] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," in *ICML*, 2025.

[13] S. Bahl, R. Mendonca, L. Chen, *et al.*, "Affordances from human videos as a versatile representation for robotics," in *CVPR*, 2023.

[14] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang, *et al.*, "Magma: A foundation model for multimodal AI agents," in *CVPR*, 2025.

[15] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, *et al.*, "Latent action pretraining from videos," in *ICLR*, 2025.

[16] X. Chen, J. Guo, T. He, C. Zhang, P. Zhang, D. C. Yang, L. Zhao, and J. Bian, "IGOR: Image-goal representations are the atomic control units for foundation models in embodied ai," *arXiv:2411.00785*, 2024.

[17] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, *et al.*, "GR00T N1: An open foundation model for generalist humanoid robots," *arXiv:2503.14734*, 2025.

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

[19] 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:2507.15597*, 2025.

[20] H. Bi, L. Wu, T. Lin, H. Tan, *et al.*, "H-RDT: Human manipulation enhanced bimanual robotic manipulation," *arXiv:2507.23523*, 2025.

[21] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv:2212.06817*, 2022.

[22] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv:2406.09246*, 2024.

[23] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv:2405.12213*, 2024.

[24] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, *et al.*, "Vision-language foundation models as effective robot imitators," in *ICLR*, 2022.

[25] 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:2410.24164*.

[26] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, *et al.*, "CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv:2411.19650*, 2024.

[27] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: A diffusion foundation model for bimanual manipulation," *ICLR*, 2025.

[28] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv:2504.16054*, 2025.

[29] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, *et al.*, "SpatialVLA: Exploring spatial representations for visual-language-action model," *arXiv:2501.15830*, 2025.

[30] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "DexVLA: Vision-language model with plug-in diffusion expert for general robot control," in *CoRL*, 2025.

[31] S. Deng, M. Yan, S. Wei, H. Ma, Y. Yang, J. Chen, Z. Zhang, T. Yang, X. Zhang, *et al.*, "GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv:2505.03233*, 2025.

[32] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Learning to act anywhere with task-centric latent actions," *arXiv:2502.14420*, 2025.

[33] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv:2312.13139*, 2023.

[34] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, *et al.*, "GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv:2410.06158*, 2024.

[35] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, *et al.*, "Solving rubik's cube with a robot hand," *arXiv:1910.07113*, 2019.

[36] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *IJRR*, 2020.

[37] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," *NeurIPS*, 2022.

[38] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *RSS*, 2018.

[39] D. Jain, A. Li, S. Singhal, A. Rajeswaran, *et al.*, "Learning deep visuomotor policies for dexterous hand manipulation," in *ICRA*, 2019.

[40] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *ICRA*, 2020.

[41] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and K. Liu, "DexCap: Scalable and portable mocap data collection system for dexterous manipulation," in *RSS Workshop*, 2024.

[42] 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:2503.13441*, 2025.

[43] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "EgoMimic: Scaling imitation learning via egocentric video," in *ICRA*, 2025.

[44] S. Park, S. Lee, M. Choi, J. Lee, J. Kim, J. Kim, and H. Joo, "Learning to transfer human hand skills for robot manipulations," *arXiv:2501.04169*, 2025.

[45] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, *et al.*, "DexGraspVLA: A vision-language-action framework towards general dexterous grasping," *arXiv:2502.20900*, 2025.

[46] V. de Bakker, J. Hejna, T. G. W. Lum, O. Celik, A. Taranovic, D. Blessing, G. Neumann, J. Bohg, and D. Sadigh, "Scaffolding dexterous manipulation with vision-language models," *arXiv:2506.19212*, 2025.

[47] J. Yang, B. Liu, J. Fu, B. Pan, G. Wu, *et al.*, "Spatiotemporal predictive pre-training for robotic motor control," *arXiv:2403.05304*, 2024.

[48] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, *et al.*, "DexTransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations," *arXiv:2209.14284*, 2022.

[49] Z. Chen, S. Chen, A. Etienne, I. Laptev, and C. Schmid, "ViViDex: Learning vision-based dexterous manipulation from human videos," in *ICRA*, 2025.

[50] A. Patel, A. Wang, I. Radosavovic, and J. Malik, "Learning to imitate object interactions from internet videos," *arXiv:2211.13225*, 2022.

[51] P. Mandikal and K. Grauman, "DexVIP: Learning dexterous grasping with human hand pose priors from video," in *CoRL*, 2022.

[52] A. Kannan, K. Shaw, S. Bahl, P. Mannam, and D. Pathak, "DEFT: Dexterous fine-tuning for hand policies," in *CoRL*, 2023.

[53] H. Chen, B. Sxun, 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 *CVPR*, 2025.

[54] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *ECCV*, 2024.

[55] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, *et al.*, "Any-point trajectory modeling for policy learning," *arXiv:2401.00025*, 2023.

[56] H. G. Singh, A. Loquercio, C. Sferrazza, J. Wu, H. Qi, *et al.*, "Hand-object interaction pretraining from videos," in *ICRA*, 2025.

[57] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne, *et al.*, "FLARE: Robot learning with implicit world modeling," *arXiv:2505.15659*, 2025.

[58] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "UniVLA: Learning to act anywhere with task-centric latent actions," *arXiv:2505.06111*, 2025.

[59] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *IROS*, 2021.

[60] S. Xie, H. Cao, Z. Weng, Z. Xing, H. Chen, S. Shen, J. Leng, Z. Wu, and Y.-G. Jiang, "Human2Robot: Learning robot actions from paired human-robot videos," *arXiv:2502.16587*, 2025.

[61] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, *et al.*, "Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation," *arXiv:2409.16283*, 2024.

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

[63] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, *et al.*, "DreamGen: Unlocking generalization in robot learning through video world models," *arXiv:2505.12705*, 2025.

[64] L. Wang, Y. Qiao, X. Tang, *et al.*, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognition Challenge*, 2014.

[65] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016.

[66] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *CVPR*, 2019.

[67] C.-L. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing moments of actions with transformers," in *ECCV*, 2022.

[68] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah, and C. Xu, "Diffusion action segmentation," in *ICCV*, 2023.

[69] J. Chen, Z. Lv, S. Wu, K. Q. Lin, C. Song, D. Gao, J.-W. Liu, Z. Gao, D. Mao, and M. Z. Shou, "VideoLLM-Online: Online video large language model for streaming video," in *CVPR*, 2024.

[70] G. Chen, Y.-D. Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qiao, T. Lu, *et al.*, "Videollm: Modeling video sequence with large language models," *arXiv:2305.13292*, 2023.

[71] A. Hagemann, M. Knorr, and C. Stiller, "Deep geometry-aware camera self-calibration from video," in *ICCV*, 2023.

[72] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang, "MoGe-2: Accurate monocular geometry with metric scale and sharp details," *arXiv:2507.02546*, 2025.

[73] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin, "DeepCalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras," in *CVMP*, 2018.

[74] J. Zhang, J. Deng, C. Ma, and R. A. Potamias, "HaWoR: World-space hand motion reconstruction from egocentric videos," in *CVPR*, 2025.

[75] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," *TOG*, 2017.

[76] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely, "MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos," in *CVPR*, 2025.

[77] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[78] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.

[79] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, *et al.*, "The EPIC-KITCHENS dataset: Collection, challenges and baselines," *TPAMI*, 2020.

[80] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, *et al.*, "Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives," in *CVPR*, 2024.

[81] 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 *ICCV*, 2017.

[82] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, *et al.*, "Paligemma 2: A family of versatile vlms for transfer," *arXiv:2412.03555*, 2024.

[83] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *ICCV*, 2023.

[84] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv:2408.00118*, 2024.

[85] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023.

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

[87] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv:2311.15127*, 2023.