



METIS: Multi-Source Egocentric Training for Integrated Dexterous Vision-Language-Action Model

Yankai Fu^{1,2*}, Ning Chen^{1,2*}, Junkai Zhao^{2*†}, Shaozhe Shan¹,
Guocai Yao², Pengwei Wang², Zhongyuan Wang², Shanghang Zhang^{1,2,✉}

¹State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University; ²Beijing Academy of Artificial Intelligence

*Equal contribution, †Project leader, ✉Corresponding author

Project Webpage: <https://aureleopku.github.io/METIS>

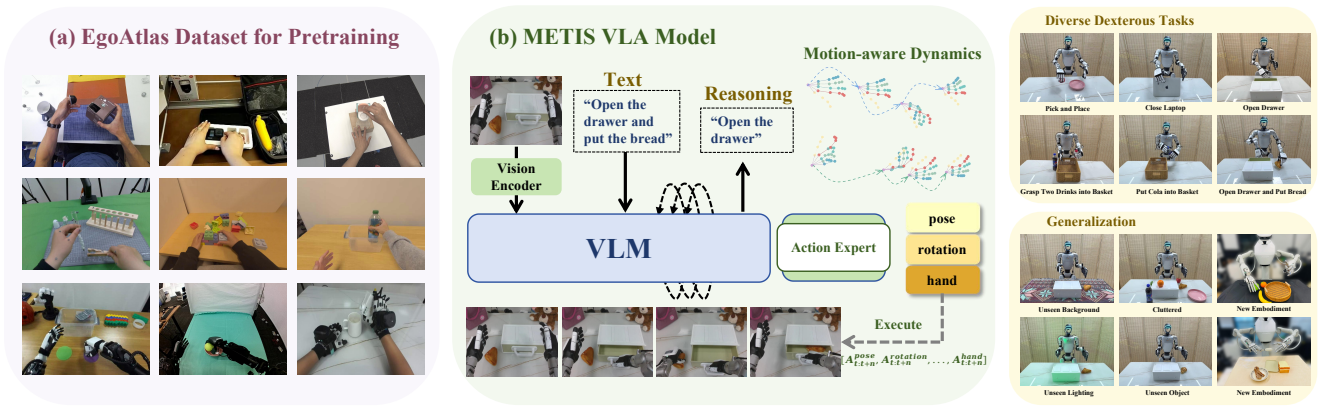


Figure 1. METIS is trained on a multi-source egocentric manipulation dataset EgoAtlas. It leverages motion-aware dynamics to extract manipulation-relevant dexterous features, and integrates reasoning and acting within a unified framework. METIS achieves strong performance across diverse dexterous manipulation tasks and exhibits remarkable generalization capability.

Abstract

Building a generalist robot that can perceive, reason, and act across diverse tasks remains an open challenge, especially for dexterous manipulation. A major bottleneck lies in the scarcity of large-scale, action-annotated data for dexterous skills, as teleoperation is difficult and costly. Human data, with its vast scale and diverse manipulation behaviors, provides rich priors for learning robotic actions. While prior works have explored leveraging human demonstrations, they are often constrained by limited scenarios and a large visual gap between human and robots. To eliminate these limitations, we propose METIS, a vision-language-action (VLA) model for dexterous manipulation pretrained on multi-source egocentric datasets. We first construct EgoAtlas, which integrates large-scale human and robotic data from multiple sources, all unified under a consistent action space. We further extract motion-aware dynamics, a compact and

discretized motion representation, which provides efficient and expressive supervision for VLA training. Built upon them, METIS integrates reasoning and acting into a unified framework, enabling effective deployment to downstream dexterous manipulation tasks. Our method demonstrates exceptional dexterous manipulation capabilities, achieving highest average success rate in six real-world tasks. Experimental results also highlight the superior generalization and robustness to out-of-distribution scenarios. These findings emphasize METIS as a promising step toward a generalist model for dexterous manipulation.

1. Introduction

Recent advances in vision-language-action (VLA) models have achieved remarkable progress toward general-purpose embodied intelligence[14, 18, 43, 67]. However, such mod-

els heavily rely on scarce and expensive real-world robot data, often collected via human teleoperation. Despite extensive efforts to build large-scale robotic datasets[4, 16, 34, 52], their scale and diversity are still two orders of magnitude lower than those used for LLM training, severely limiting the scalability and generalization of current VLAs. This issue is even more pronounced in dexterous manipulation, where acquiring high-quality demonstrations is extremely costly and complex. As a result, most existing VLA research focuses on simple gripper-based tasks[5, 13, 25, 29], leaving dexterous manipulation largely unexplored.

In contrast to the scarcity of robotic datasets, human data are vastly abundant and rich in semantic information, providing valuable behavioral priors. Previous works have explored learning task-relevant representations from human videos, such as affordance[1, 20], latent action[5, 58], and key-point flow[60, 63]. Recent progress such as EgoVLA[57], Being-H0[31], and H-RDT[2] leverage large-scale human datasets to pretrain VLA, providing a new paradigm for scaling robotic learning. However, human videos are often confined to specific household or workspace scenes (*e.g.*, tabletop or kitchen), exhibiting uneven scene coverage and strong contextual biases. Moreover, there exists a substantial gap between human data and robot data, in both visual appearance and action space. Another line of research explores co-training strategies that jointly utilize self-collected human and robotic data. For example, HAT[38] learns a human policy and demonstrates that incorporating human data can significantly enhance the generalization and robustness of robotic policies, while MotionTrans [61] aligns human motions to robot-specific embodiments to achieve zero-shot skill transfer. These approaches collect manipulation-related human data but do not fully exploit the vast amount of human data available on the internet.

In this work, we investigate these problems by introducing **METIS, a vision-language-action (VLA) model** for dexterous manipulation pretrained on multi-source egocentric datasets. We first construct **EgoAtlas**, a multi-source egocentric dataset that covers large-scale internet human data, robot data, and our enhanced human data collected through a wearable system. EgoAtlas spans four major categories and eight sources, all aligned under a unified action space. We further propose motion-aware dynamics, a compact and discretized representation designed for dexterous manipulation. It captures both visual and motion dynamics, providing efficient and expressive supervision for training VLA models. Built upon them, METIS is pretrained on EgoAtlas, unifying reasoning and acting within a single framework. This design enables efficient fine-tuning and deployment on downstream dexterous manipulation tasks.

To comprehensively evaluate METIS, we conduct extensive real-world experiments on dexterous tasks. The comparative results show that our model achieves superior perfor-

mance and efficiency, attaining the highest average success rate among all baseline VLAs. Besides, METIS also exhibits strong generalization to out-of-distribution scenarios, including unseen background, unseen object, unseen lighting condition, cluttered environment, and can be transferred to higher-DoF embodiments. Finally, ablation studies verify the contributions of both the multi-source egocentric dataset and the motion-aware dynamics to the overall system, highlighting the potential of learning robotic motion priors from human data. In summary, our contributions are as follows:

- We construct a multi-source egocentric manipulation dataset EgoAtlas, which integrates diverse human and robotic data sources under a unified action space.
- We propose to extract motion-aware dynamics, a compact and discretized representation of dexterous hand motion.
- We present METIS, a VLA model for dexterous manipulation, pretrained on large-scale multi-source egocentric data. It integrates reasoning and acting within a unified framework.
- We demonstrate the effectiveness and generalization of our method through a range of real-world experiments.

2. Related Work

2.1. Dexterous Manipulation

Dexterous manipulation is a key challenge in robotics, focusing on achieving human-like fine-grained manipulation skills. Traditional approaches are based on optimization and control algorithms[21, 26, 49, 54], typically assuming access to known dynamics and object models, and focusing on trajectory planning under physical constraints. While achieving strong performance in specific scenarios, these methods often struggle to generalize across diverse real-world settings. Recent learning-based approaches, including reinforcement learning (RL) and imitation learning (IL), have made remarkable progress across various tasks such as grasping[46, 53], in-hand manipulation[35, 36], and tool use[40, 59]. RL methods learn precise and dexterous skills through large-scale training in simulation, but often suffer from a significant sim-to-real gap. In contrast, IL methods leverage expert demonstrations to achieve robust performance in real-world dexterous manipulation tasks[9, 10, 17, 27], but typically rely on expensive human teleoperation data. In this work, we address this data bottleneck by pretraining our model on large-scale multi-source egocentric data that combines human and robotic demonstration to learn motion priors efficiently.

2.2. Learning Dexterity from Human Data

Human data offer valuable priors for dexterous manipulation, featuring abundant samples, fine-grained hand motion, and rich semantic cues. There has been works on learning task-relevant representations from human videos, such as

affordance[1, 20], latent action[5, 58], and keypoint flow[60, 63]. Ego-Only[48] utilizes the MAE framework[11] to extract actionable information from egocentric videos, while LAPA[58] and UniVLA[5] apply VQ-VAE[45] to extract latent action from large-scale unlabeled human data. However, learning directly from human videos introduces redundant, manipulation-irrelevant content and underemphasizes the critical role of hand motion. To mitigate this gap, HAT[38] and MotionTrans[61] adopt a co-training strategy on both human and robotic data with explicit motion information, improving the robustness of the policy and enabling zero-shot skill transfer. In this work, we learn action priors from human data through joint modeling of visual and motion dynamics to support VLA learning.

2.3. Vision-Language-Action Model

Vision-Language-Action (VLA) models have achieved unprecedented progress in recent years[14, 18, 43, 64], driven by advances in Vision-Language Models (VLMs)[41, 42, 55] and the availability of large-scale robotic datasets[4, 16, 34, 52]. By processing multimodal inputs—such as visual observations and language instructions—these models enable robots to autonomously perform a wide range of tasks. Representative VLA models, such as RT-2[67] and OpenVLA [18] generate action sequences autoregressively. In contrast, π_0 [3] and DexVLA[51] employ a diffusion-based policy, fitting continuous action distributions via iterative denoising. However, these approaches primarily focus on gripper-based manipulation, overlooking the rich motion and interaction dynamics inherent in dexterous tasks. Recently, several studies have extended VLA frameworks to dexterous manipulation. For example, GR00T N1[32] learns latent representations from hybrid data to train a humanoid manipulation policy, while EgoVLA[57] and Being-H0[31] leverage large-scale human demonstrations for VLA pretraining, enabling the acquisition of motion priors. Despite their impressive results, the scene homogeneity of human videos leads to bias and a large visual gap from robotic observations. To overcome this issue, we leverage multi-source egocentric data, introduce an enhanced human dataset, and develop an integrated VLA model for unified reasoning and acting.

3. EgoAtlas Dataset

In this section, we introduce **EgoAtlas**, a large-scale, multi-source egocentric dataset designed to bridge human and robotic dexterous manipulation. EgoAtlas integrates data from multiple modalities, which are aligned within a unified action space for consistent VLA training.

3.1. Wearable System for Enhanced Human Data

Traditional human hand motion datasets rely on multi-camera or VR-based tracking systems, which suffer from viewpoint dependency, occlusion, and restricted capture

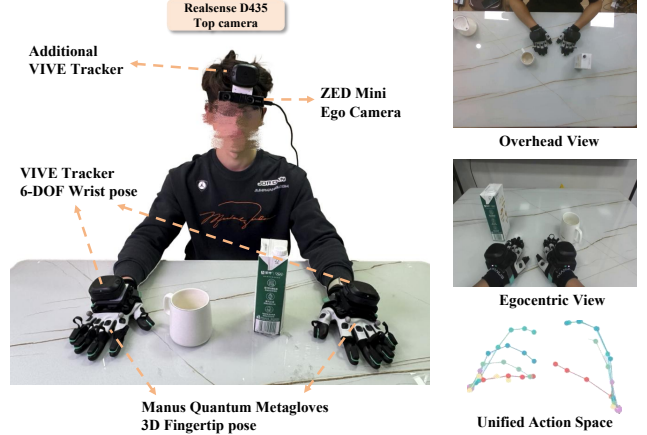


Figure 2. Wearable Hand Motion Collection System.

space. To overcome these issues, we develop a wearable glove–tracker system that enables portable, high-fidelity human motion capture. This system allows data to be collected anytime and anywhere.

Wearable Hand Motion Collection System. We use Manus Quantum Metagloves to record precise 3D positions of hand keypoints, with 25 keypoints per hand. A VIVE Tracker mounted on each glove provides the 6-DoF wrist pose, enabling global hand localization in space. To capture egocentric visual observations, a head-mounted camera is used to record the first-person view during manipulation, providing visual information closely aligned with the operator’s perspective, while another VIVE Tracker is attached to the headset to perform extrinsic calibration. This calibration aligns the motion-capture and ego-camera coordinate frames, enabling the entire system to operate in the wild and ensuring scene diversity. We also provide a top-down third-person camera view to facilitate future research on learning from human videos. The entire system operates at 20HZ, balancing motion fidelity with reliable multi-sensor synchronization.

Subtask-Level Annotation. We performed subtask-level annotation to enrich the semantic structure of the dataset. Each trajectory is paired with a language instruction that describes the overall episode, along with a fine-grained segmentation into multiple subtasks. This annotation design enables the VLA to focus on hand motion and supports hierarchical reasoning over long-horizon manipulation tasks.

3.2. Data Sources and Statistics

EgoAtlas integrates data from four major sources, covering both human and robotic domains with diverse sensing modalities: (1) **Vision-based motion capture datasets**[7, 22, 30, 50, 56], which use multi-camera optical systems to capture precise 3D hand annotations and object interactions. However, they are typically confined to small tabletop environments with limited diversity. (2) **VR-based human datasets**[12], which utilize on-device SLAM and calibrated

Table 1. **Statistics of EgoAtlas**. We use in-the-wild to denote data collected in diverse, unconstrained real-world scenes.

Method	Trajs	Frames	Pose	Subtask	Human	Robot	In-the-wild
ARCTIC	296	214.5K	✓	✗	100%	0%	✗
H2O	109	65.3K	✓	✓	100%	0%	✗
HoloAssist	100	777.3K	✓	✓	100%	0%	✗
Oakink	134	146K	✓	✓	100%	0%	✗
EgoDex	314.8K	77.9M	✓	✗	100%	0%	✓
PH2D	1.8K	416.5K	✓	✗	66.1%	33.9%	✓
ActionNet	15.7K	7.4M	✓	✗	0%	100%	✓
Ours	10K	2.8M	✓	✓	100%	0%	✓

cameras to capture hand and wrist pose. VR-based setups exhibit fewer scene constraints and allow more flexible data collection across different environments. **(3) Teleoperated robot data**[8, 39], which involve human operators remotely controlling dexterous robotic hands to execute manipulation tasks. **(4) Self-collected enhanced motion dataset**, as described in Sec. 3.1, we collected 10K high-fidelity human hand motion trajectories. This enhanced dataset brings three main benefits: (1) it is robust to occlusions and visual ambiguities; (2) it introduces visual diversity via wearable gloves; (3) it provides rich semantic annotations.

The composition of EgoAtlas dataset is summarized in Tab. 1, which includes 8 sources with detailed annotations of hand motions. In total, EgoAtlas contains 343K trajectories and 89.72M image–action pairs. Weighted sampling is applied to maintain a balanced source distribution, with detailed weights listed in the supplementary material.

3.3. Data Processing

Different embodiments exhibit distinct action spaces. To enable a generalizable VLA model across heterogeneous embodiments, we construct a unified action space that bridges the gap between human and robot motion representations. For the wrist pose (18 dim), we unify all representations into the ego-camera coordinate frame, which consists of a 3D position and a 6D rotation vector. For the hand (30 dim), we calibrate the motion to the wrist coordinate frame, using the 3D positions of each fingertip.

Dexterous hand’s joint angles can be mapped to fingertip positions through forward kinematics (FK). During inference, the process is inverted—fingertip targets predicted by the policy are converted back to joint angles via inverse kinematics (IK). Notably, the wrist coordinate frames of humans and robots are aligned through calibration to ensure cross-embodiment consistency.

4. Method

In this section, we present our proposed METIS, a VLA model which learns from multi-source egocentric manipulation dataset, and can be efficiently deployed on real robots. We first introduce the necessary preliminaries in Sec. 4.1. To facilitate effective learning of dexterous manipulation,

we develop motion-aware dynamics for VLA training in Sec. 4.2. We further detail the model architecture and the cross-modal reasoning of METIS in Sec. 4.3, which enable unified multimodal perception and action generation.

4.1. Problem Formulation

Our goal is to learn a VLA model for dexterous manipulation from multi-source egocentric data. Formally, given a collection of human and robot trajectories $D = D_{robot} \cup D_{human}$ from the EgoAtlas dataset, each trajectory can be represented as a sequence of observation–action pairs $\tau = \{(o_t, a_t)\}_{t=1}^T$, where observation $o_t = \{I_t, S_t\}$ consists of the egocentric image input and the proprioceptive state. The training objective is to learn a policy $\pi_\phi(a_t|o_t, l)$ that predicts actions conditioned on egocentric observations o_t and language instruction l .

To bridge the embodiment gap between human and robot, we construct a unified proprioception–action space which includes an 18D wrist pose P_t^w (3D position and 6D rotation vector per hand) and a 30D finger pose P_t^f capturing the 3D position of individual fingertips. Here we describe the design of the three components in detail.

- **Image Observation I_t** : We use egocentric image from a first-person camera, which provides a viewpoint that focus on fine-grained interaction details.
- **Wrist Pose P_t^w** : We represent the wrist pose as the 3D position and the 6D rotation vector following Zhou et al. [66]. They are relative to the camera coordinate frame.
- **Finger Pose P_t^f** : We represent the finger pose using 3D fingertip positions, defined in the wrist coordinate frame.

4.2. Motion-Aware Dynamics Construction

Recent VLA models typically employ a tokenizer to discretize continuous actions, enabling autoregressive sequence modeling for policy learning. However, as the number of action chunks and the system’s degrees of freedom increase, the length of these discrete action sequences grows, which significantly slows down autoregressive generation. Additionally, this tokenization approach often struggles to capture fine-grained motion details, which is particularly critical in dexterous manipulation, where subtle finger movements and contact interactions are essential.

To address these challenges, we propose a compact and discretized representation that effectively constructs motion-aware dynamics for dexterous manipulation. This representation provides an efficient and expressive supervision signal for VLA pretraining. The proposed dynamics model consists of two key components, detailed as follows.

Visual Dynamics Discretization. The causal relationship between motion and visual change is crucial for learning a generalizable VLA model. Therefore, it is important to model visual dynamics while incorporating motion information,

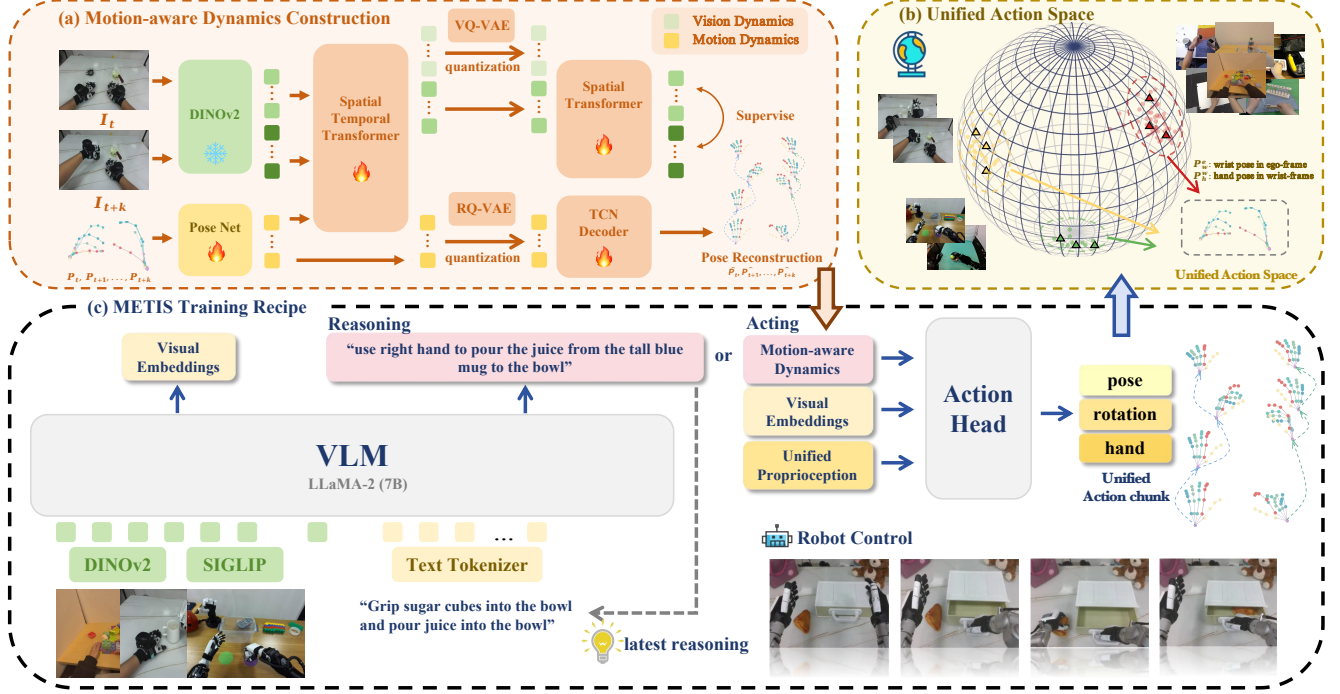


Figure 3. **Overview Framework** (a) We construct an expressive yet compact representation to capture the dynamics involved in dexterous manipulation. (b) METIS is pretrained on multi-source EgoAtlas dataset, where human and robot actions are align under a unified action space. (c) METIS integrates reasoning and acting within a framework, enabling effective deployment to downstream dexterous tasks.

especially in egocentric dexterous manipulation scenarios where subtle hand–object interactions drive task progression.

Specifically, we employ an Inverse Dynamics Model based encoder $\mathcal{I}(D_{vis}|I_t, I_{t+k}, P_{t,t+1,\dots,t+k})$ and a Forward Dynamics Model based decoder $\mathcal{F}(I_{t+k}|I_t, D_{vis})$. The encoder captures motion-relevant visual dynamics by integrating visual observations with continuous motion information, while the decoder is trained to predict future observation given the constructed visual dynamics. Our encoder includes both spatial and temporal transformer while the decoder only contains spatial transformer.

The visual dynamics D_{vis} are quantized with the VQ-VAE objective [45]. It maps continuous latent features into a finite set of discrete codebook embeddings, thereby enabling efficient and stable auto-regressive training for VLA. Following Bu et al. [5], we do not reconstruct image pixels, as raw pixels contain substantial redundancy and details irrelevant to the manipulation task. Instead, we use pretrained DINOv2[33] to extract high-level semantic representations, that better reflect task-relevant visual dynamics.

Motion Dynamics Quantization. Learning motion priors is a challenging yet crucial problem for VLA models, therefore we introduce a separate set of discretized codebook embeddings D_{mot} to focus on motion capturing. We discretize hand motion trajectories into compact dynamics tokens that effectively capture proprioceptive information and reflect the fine-grained features of dexterous manipulation. We

use PoseNet as the encoder for 3D hand motion, combining multi-scale temporal convolutions with trajectory self-attention to capture spatio-temporal dynamics. The continuous motion features are then quantized using RQ-VAE[24], which not only effectively prevents codebook collapse but also captures hierarchical motion patterns, from coarse to fine levels. A temporal convolutional network (TCN)[23] is used as the decoder to reconstruct motion trajectories during training, as illustrated in Fig. 3(a).

4.3. METIS Model

As shown in Fig. 3(c), our proposed METIS model is built upon a VLM, where the parameters are initialized from the Prismatic-7B[15]. It incorporates a hybrid vision encoder that integrates SigLIP[62] and DINOv2[33], capturing both global semantics and fine-grained spatial details. The resulting visual features $f^{SigLIP} \in \mathbb{R}^{N_v \times 1024}$ and $f^{DINO} \in \mathbb{R}^{N_v \times 1152}$, where N_v represents the visual token dimension, are concatenated along the channel dimension. A projection layer is applied to align visual embeddings with the language modality. The 7B LLaMA-2[44] large language model is adopted as the LLM backbone, which employs a decoder-only Transformer architecture with 32 sequential blocks for auto-regressive language modeling.

Previous work such as OpenVLA[18] and RT-2[67] map the infrequently used words in LLaMA-2 to action bins uniformly distributed within $[-1, 1]$. While effective for

low-dimensional control, they struggle to scale to high-dimensional, continuous actions, often resulting in inefficient and unstable behavior. METIS first extends the LLaMA tokenizer vocabulary with $|C_1 + C_2|$ special tokens, where $|C_1|$ and $|C_2|$ correspond to the codebook sizes of the visual dynamics tokens and motion dynamics tokens, respectively. Using the dynamics model introduced in the previous section, we discretize each egocentric manipulation sequence into motion-aware action tokens, denoted as D_{vis} and D_{mot} . Each dynamics feature is assigned to its nearest codebook entry, and the resulting discrete index is mapped to a unique special token. This design fully leverages the original VLM architecture, enabling autoregressive supervision for training. By preserving the language priors of the language model while injecting motion information, METIS learns to model the fine-grained dynamics essential for dexterous manipulation. The auto-regressive objective of METIS π is to minimize the sum of next-dynamics negative log-probabilities:

$$\mathcal{L}_{ar} = \mathbb{E}_{o_t, l, a_{d, < i}} \left[- \sum_{i=1}^N \log \pi_{\phi}(\hat{a}_{d, i} | o_t, l, a_{d, < i}) \right] \quad (1)$$

where N represents the total length of dynamic tokens, which is 44 in our settings (4 for visual dynamics and 40 for motion dynamics). More details can be found in the supplementary material.

Action Decoder. The Action Decoder translates motion-aware dynamic tokens into executable low-level actions. It takes as input the dynamics token, visual embeddings, and the current proprioception. visual and dynamic features are aggregated through multi-head attention pooling, while proprioceptive inputs are projected into the hidden space via a two-layer MLP. The fused representation is passed through a linear projection head to predict a sequence of actions over one second (30 future steps at 30 Hz). The final loss is $\mathcal{L} = \mathcal{L}_{ar} + \lambda \mathcal{L}_{action}$.

Chain-of-Thought Reasoning for Action. Inspired by chain-of-thought prompting in large language models, we enable the METIS to decompose high-level manipulation instructions into shorter subtasks. The subtask is accompanied by fine-grained hand-level descriptions that provide explicit guidance for action prediction, such as "use right hand to pour the juice from the tall blue mug to the bowl". To support this reasoning process, we manually annotate our self-collected dataset with hand-level subtask labels describing intermediate manipulation actions such as grasping, lifting, pouring, and rotating. Following Lin et al. [28], METIS integrates reasoning and acting within a unified framework that autonomously determines at each timestep t whether to reason or act. We introduce two special tokens: beginning of reasoning([BOA]), beginning of dynamics([BOD]). METIS enters reasoning mode only when a subtask transition occurs, which substantially reduces inference latency. When [BOD] is predicted, the VLM directly outputs motion-

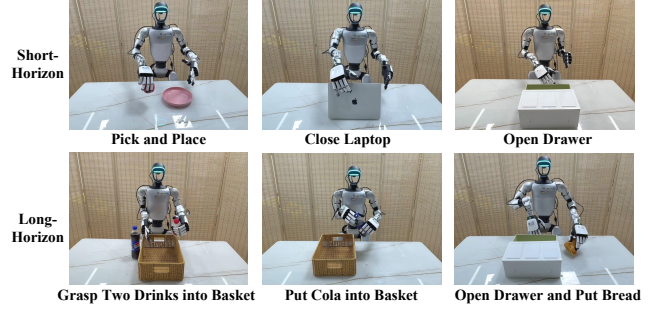


Figure 4. **Visualization of dexterous manipulation tasks**, including three short-horizon three long-horizon tasks.

aware dynamics, which are decoded by the action decoder into low-level actions for execution. This adaptive switching effectively enhances the mutual understanding between reasoning and control, while reducing latency during the inference phase.

5. Experiments

We conduct comprehensive experiments to answer the following questions:

- How does METIS perform in real-world dexterous manipulation experiments (Sec. 5.2)?
- How promising is METIS in terms of sample efficiency and generalizability (Sec. 5.3 and Sec. 5.4)?
- How do multi-source egocentric data and motion-aware dynamics contribute to overall performance (Sec. 5.5)?

5.1. Experiment Setup

Hardware Platform. For hardware platform, we use a Unitree G1 humanoid robot equipped with a pair of Inspire 6-DoF dexterous hands for fine-grained manipulation. An Intel RealSense D435 camera is mounted on the robot’s head to capture egocentric RGB observations.

Self-collected Robot Data. We collect robot demonstrations through human teleoperation with a glove-tracker system[47]. The tracker attached to the operator’s wrist captures precise wrist poses, which are converted into the robot arm’s joint configurations via inverse kinematics(IK). Simultaneously, a motion-capture glove records the fingertip trajectories, which are mapped to the dexterous hand joint space using an IK-based retargeting algorithm[37].

Tasks. We evaluated METIS on six dexterous manipulation tasks including three short-horizon and three long-horizon tasks, as shown in Fig. 4: (1) Pick and Place, (2) Close Laptop, (3) Open Drawer, (4) Grasp Two Drinks into Basket, (5) Put Cola into Basket, (6) Open Drawer and Put Bread. Each task is collected with 100 high-quality demonstrations and is evaluated with 20 trials by default. We provide more details about these tasks in the supplementary material.

Baselines. We compare METIS with four representative baselines: (1) ACT[65], an action chunking transformer that

Table 2. **Main results of six real-world tasks.** Each experiment is evaluated with 20 trials. SR denotes Success Rate, and PSR denotes Progress Success Rate. Generally, METIS achieves highest average success rate among all tasks.

Method	Pick and Place	Close Laptop	Open Drawer	Grasp Two Drinks into Basket		Put Cola into Basket		Open Drawer and Put Bread	
	SR	SR	SR	SR	PSR	SR	PSR	SR	PSR
ACT	35.0%	65.0%	95.0%	25.0%	40.0%	50.0%	53.3%	5.0%	5.0%
OpenVLA-OFT	50.0%	80.0%	10.0%	40.0%	57.5%	55.0%	56.7%	0.0%	1.0%
$\pi_{0.5}$	60.0%	85.0%	70.0%	65.0%	72.5%	75.0%	76.7%	60.0%	65.0%
GR00T N1.5	70.0%	80.0%	80.0%	65.0%	70.0%	70.0%	73.3%	70.0%	72.5%
METIS (Ours)	85.0%	95.0%	90.0%	75.0%	85.0%	70.0%	76.7%	75.0%	82.5%

learns low-level visuomotor policy. (2) OpenVLA-OFT[19], an enhanced open-source VLA model via optimized fine-tuning; (3) $\pi_{0.5}$ [14], a VLA flow model for general robot control; (4) Gr00t N1.5[32], an improved VLA model for generalist humanoid robots.

Evaluation Metrics. We use two metrics to evaluate model performance: Success Rate (SR), indicating the entire task is successfully completed, and Progress Rate (PSR), capturing the average completion ratio of sub-tasks relative to the overall task in long-horizon settings.

5.2. Performance in Real-World Experiments

We conduct experiments on real-world dexterous manipulation tasks, as shown in Tab. 2. Our model, achieves the highest average success rate and consistently outperforms existing state-of-the-art VLA models on most tasks. The specialist model ACT performs well on short-horizon manipulation task but struggles on long-horizon tasks, highlighting the importance of integrated reasoning capabilities. $\pi_{0.5}$ underperforms on tasks requiring precise dexterous manipulation, as it has not been pretrained on large-scale dexterous datasets. GR00T N1.5 achieves competitive results through large-scale pretraining on real humanoid robot and human data with implicit latent representations. Nevertheless, the absence of explicit reasoning and feedback correction limits its performance on long-horizon tasks. METIS, benefiting from its motion-aware dynamics and explicit reasoning mechanism, achieves outstanding performance across both short and long-horizon tasks. Additionally, METIS attains highest PSR across all long-horizon tasks, demonstrating it can reason and act coherently across long-horizon sequences, maintaining stable control and minimizing error accumulation during task execution.

Instruction Following. We further evaluate the instruction following capability of METIS. In this setup, three fruits of different colors (apples, orange, and lemon) are placed on the table. As shown in Fig. 5, when given different language instructions (*e.g.*, "place the red apple on the plate"), METIS can identify the target fruit and execute the corresponding grasping action.

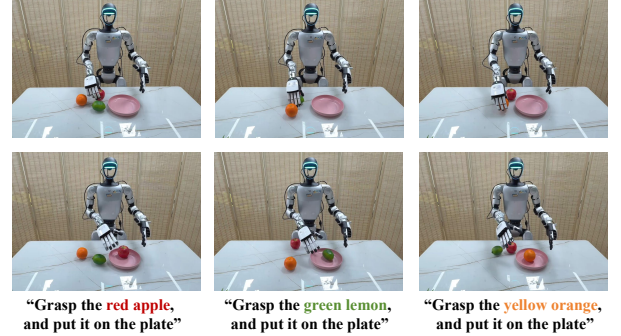


Figure 5. **Instruction following results.** Each task is collected with 100 demonstrations, jointly trained, and evaluated using different language instructions.

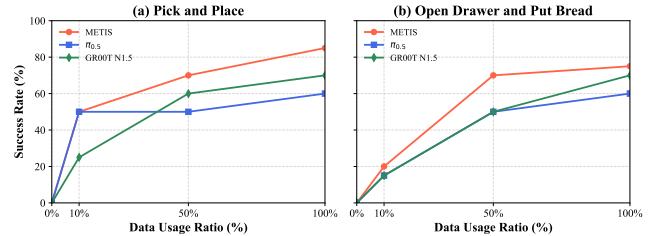
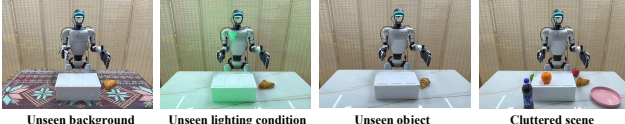


Figure 6. **Experimental results of efficiency.** We train the VLA model with an increasing number of demonstrations.

5.3. Efficiency

We evaluate the sample efficiency of METIS by training the model with varying amounts of data on downstream dexterous manipulation tasks, as shown in Fig. 6. METIS exhibits excellent efficiency, achieving superior performance even with limited training data. Notably, when fine-tuned with only 10% of the data, METIS still achieves a 50% success rate on the Pick and Place task. This results strongly indicate that pretraining on diverse multi-source egocentric data with unified action space provides valuable prior knowledge, including spatial reasoning, visual-hand coordination, and motion dynamics. Such priors form consistent visuomotor representations and facilitate rapid transfer to new downstream tasks.

Table 3. **Generalization results of OOD scenarios.** We evaluate the model in four unseen settings: unseen background, unseen lighting, unseen object, and cluttered scene.



Method	unseen background	unseen lighting	unseen object	cluttered
$\pi_{0.5}$	50.0%	70.0%	65.0%	55.0%
GR00T N1.5	65.0%	65.0%	65.0%	60.0%
METIS (Ours)	70.0%	65.0%	70.0%	70.0%

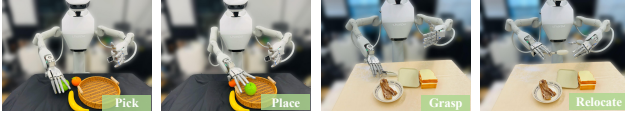


Figure 7. **Cross-Embodiment Generalization.** METIS demonstrates transferability to 22-DoF dexterous hands, achieving stable performance on Grasp Apple into Basket and Tool Use tasks.

5.4. Generalization

Besides the remarkable effectiveness and efficiency, METIS also showcases excellent generalization capabilities. We evaluate our model in four *out-of-distribution*(OOD) scenarios, including (1)**Unseen background**, we cover the tabletop with a colorful patterned tablecloth to introduce significant visual distractions and background textures unseen during training. (2)**Unseen lighting condition**, we illuminate the scene with color-changing and flickering lights to simulate dynamic and diverse lighting situations. (3)**Unseen object**, we replace the original scene bread with a visually distinct croissant to assess object-level generalization. (4)**Cluttered environment**, we randomly place distractor objects such as a plate and an apple near the drawer, increasing visual and spatial complexity. Taking the Open Drawer and Put Bread task as an example, we compare and report the success rates of METIS and GR00T N1.5 across all four out-of-distribution scenarios, as shown in Tab. 3. The results demonstrate that METIS effectively adapts to various distributional shifts, maintaining stable visuomotor grounding and task execution even under significant visual or physical variations.

Corss-Embodiment Generalization. METIS also generalizes effectively to higher-DoF dexterous hands, demonstrating strong adaptability across embodiments. We evaluate METIS on the Sharpa Beta embodiment equipped with a pair of 22-DoF SharpaWave Dexterous hands, as shown in Fig. 7. The model achieves 85.0% success rate on the Grasp Apple into Basket task and 70.0% on the Tool Use task, respectively. As METIS predicts fingertip trajectories rather than direct joint angles, the policy is naturally transferable and remains unaffected by variations in hand kinematics.

Table 4. **Ablation on multi-source egocentric pretraining.**

Method	Pick and Place	Open Drawer and Put Bread
METIS-NoPretrain	60.0%	35.0%
METIS-HumanPretrain	70.0%	60.0%
METIS-FullPretrain	85.0%	75.0%

Table 5. **Ablation on motion-aware dynamics.** Comparison between METIS w/ and w/o motion-aware dynamics module.

Method	Pick and Place	Open Drawer and Put Bread
METIS w/o motion-aware dynamics	30.0%	0.0%
METIS w/ motion-aware dynamics	85.0%	75.0%

5.5. Ablations

We have demonstrated the effectiveness of METIS in terms of overall performance, sample efficiency, and generalization. However, the contribution of its core components remains unexplored. In this section, we perform ablation studies focusing on two aspects: (1) the effect of multi-source egocentric pretraining, (2) the impact of the motion-aware dynamics.

Multi-source Egocentric Pretraining. To investigate the role of multi-source egocentric pretraining, we finetune and evaluate three variants of the model on downstream dexterous manipulation tasks: (a) METIS without any pretraining, (b) METIS pretrained only on the open-source human data, (c) METIS pretrained on EgoAtlas dataset. As presented in Tab. 4, pretraining on multi-source egocentric datasets improves downstream performance, indicating that training on diverse visual and action distributions enables the model to learn more robust visuomotor priors and generalize effectively across a wide range of tasks. Notably, although the no-pretraining VLA achieves a certain level of success rate on some tasks, it exhibits significant loss fluctuations during the post-training stage and unstable joint jitters during real-world deployment.

Motion-aware Dynamics. We further conduct ablation experiments on the motion-aware dynamics. During the post-training stage, we remove the autoregressive supervision of dynamics and only supervise the continuous actions. As shown in Tab. 5, METIS shows a substantial performance drop, particularly on long-horizon manipulation tasks. This result clearly demonstrates that motion-aware dynamics capture an expressive and compact motion representation, which plays a crucial role in learning temporal consistency and guiding fine-grained action prediction during dexterous manipulation.

6. Conclusions and Limitations

In this paper, we present METIS, a vision-language-action model for dexterous manipulation pretrained on multi-source egocentric dataset, integrating reasoning and acting within a unified framework. To support large-scale pretraining, we construct EgoAtlas, a comprehensive multi-source ego-

centric dataset that aligns human and robotic data under a consistent action space. By extracting motion-aware dynamics from manipulation trajectories, METIS acquires compact and expressive representations of hand motion, enabling precise and coordinated action generation. Experimental results demonstrate that METIS achieves strong performance across diverse dexterous manipulation tasks and exhibits robust generalization to out-of-distribution scenarios.

Limitations. Despite the exceptional performance demonstrated by METIS, several limitations remain. First, our model relies solely on egocentric observations, which may restrict its ability to perceive complete object geometry and fine interaction details. This limitation could be mitigated by incorporating additional wrist-mounted or external cameras. Second, the pretraining process currently excludes large-scale third-person data available online. Extending pretraining to broader multi-view manipulation datasets represents a promising direction for future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62476011).

References

- [1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 2, 3
- [2] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint arXiv:2507.23523*, 2025. 2
- [3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 3
- [4] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. 2, 3
- [5] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025. 2, 3, 5
- [6] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. <https://github.com/huggingface/lerobot>, 2024. 16
- [7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 3, 16
- [8] Yao Mu Fourier ActionNet Team. Actionnet: A dataset for dexterous bimanual manipulation. 2025. 4, 16
- [9] Yankai Fu, Qiuxuan Feng, Ning Chen, Zichen Zhou, Mengzhen Liu, Mingdong Wu, Tianxing Chen, Shanyu Rong, Jiaming Liu, Hao Dong, et al. Cordvip: Correspondence-based visuomotor policy for dexterous manipulation in real-world. *arXiv preprint arXiv:2502.08449*, 2025. 2
- [10] Koffivi Fidèle Gbagbe, Miguel Altamirano Cabrera, Ali Alabbas, Oussama Alyunes, Artem Lykov, and Dzmitry Tsetserukou. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2864–2869. IEEE, 2024. 2
- [11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3
- [12] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 3, 16
- [13] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 2
- [14] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π_0 : 5: a vision-language-action model with open-world generalization, 2025. URL <https://arxiv.org/abs/2504.16054>, 1(2):3. 1, 3, 7, 16
- [15] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024. 5
- [16] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2, 3
- [17] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation. *IEEE Transactions on Robotics*, 40:2287–2305, 2024. 2
- [18] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan

- Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1, 3, 5
- [19] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 7, 16
- [20] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024. 2, 3
- [21] Vikash Kumar, Yuval Tassa, Tom Erez, and Emanuel Todorov. Real-time behaviour synthesis for dynamic hand-manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6808–6815. IEEE, 2014. 2
- [22] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021. 3, 16
- [23] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 5
- [24] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022. 5
- [25] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. 2
- [26] Yunshuang Li, Yiyang Ling, Gaurav S Sukhatme, and Daniel Seita. Learning geometry-aware nonprehensile pushing and pulling with dexterous hands. *arXiv preprint arXiv:2509.18455*, 2025. 2
- [27] Davide Liconti, Yasunori Tshimitsu, and Robert Katzschmann. Leveraging pretrained latent representations for few-shot imitation learning on a dexterous robotic hand. *arXiv preprint arXiv:2404.16483*, 2024. 2
- [28] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025. 6
- [29] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025. 2
- [30] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li 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*, pages 21013–21022, 2022. 3
- [31] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pre-training from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 2, 3
- [32] NVIDIA, Nikita Cherniadev Johan Bjorck and Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llon-top, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzheng Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, 2025. 3, 7, 16
- [33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [34] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 2, 3
- [35] Johannes Pitz, Lennart Röstel, Leon Sievers, and Berthold Bäuml. Dextrous tactile in-hand manipulation using a modular reinforcement learning architecture. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1852–1858. IEEE, 2023. 2
- [36] Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation. In *Conference on Robot Learning*, pages 1722–1732. PMLR, 2023. 2
- [37] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems*, 2023. 6
- [38] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy ~ human policy. *arXiv preprint arXiv:2503.13441*, 2025. 2, 3, 16
- [39] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy ~ human policy. *arXiv preprint arXiv:2503.13441*, 2025. 4
- [40] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Sri-rama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. *arXiv preprint arXiv:2411.13677*, 2024. 2

- [41] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 3
- [42] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 3
- [43] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 1, 3
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5
- [45] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3, 5
- [46] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3891–3902, 2023. 2
- [47] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024. 6
- [48] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023. 3
- [49] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*, 2022. 2
- [50] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. 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*, pages 20270–20281, 2023. 3, 16
- [51] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025. 3
- [52] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024. 2, 3
- [53] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 2
- [54] Fan Yang, Thomas Power, Sergio Aguilera Marinovic, Soshi Iba, Rana Soltani Zarrin, and Dmitry Berenson. Multi-finger manipulation via trajectory optimization with differentiable rolling and geometric constraints. *IEEE Robotics and Automation Letters*, 2025. 2
- [55] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14203–14214, 2025. 3
- [56] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 3, 16
- [57] Ruihan Yang, Qinxu Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 2, 3
- [58] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejeun Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. 2, 3
- [59] Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, et al. Dexteritygen: Foundation controller for unprecedented dexterity. *arXiv preprint arXiv:2502.04307*, 2025. 2
- [60] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024. 2, 3
- [61] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025. 2, 3
- [62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 5
- [63] Anran Zhang, Hanzhi Chen, Yannick Burkhardt, Yao Zhong, Johannes Betz, Helen Oleynikova, and Stefan Leutenegger. Actron3d: Learning actionable neural functions from videos for transferable robotic manipulation. *arXiv preprint arXiv:2510.12971*, 2025. 2, 3
- [64] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025. 3
- [65] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-

cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. [6](#), [16](#)

- [66] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. [4](#)
- [67] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. [1](#), [3](#), [5](#)

Appendix

A. Data Process

Data Calibration. We perform a comprehensive calibration procedure using three VIVE trackers within our motion-capture setup. Two trackers are mounted on the wrists to provide the 6-DoF global wrist poses in the world coordinate frame. Another tracker is rigidly attached to the head-mounted egocentric camera via a custom-designed mount. Given this fixed transformation, we obtain the extrinsic parameters of the egocentric camera with respect to the global coordinate frame. By combining these measurements, we transform the wrist poses from the world coordinate system into the egocentric camera coordinate frame, achieving consistent spatial correspondence between hand motion and egocentric observations. For the finger poses, the Manus Quantum gloves directly provide 3D fingertip positions relative to the wrist frame.

Coordinate System Definition. The egocentric camera coordinate system follows the conventional computer vision convention: the origin is located at the optical center of the camera, the z-axis points forward along the optical axis (outward from the camera), the x-axis points to the right of the camera, and the y-axis points downward in the image plane.

When computing the fingertip positions of the dexterous hand, we employ forward kinematics (FK) with the root joint of the hand as the origin. However, the base of the robotic hand is typically not perfectly aligned with the wrist pose. To resolve this, we apply a fixed transformation between the wrist and the hand base to ensure geometric consistency. For clarity, we define the wrist coordinate systems as follows:

- **Left Hand:** the **x-axis** points toward the fingertips, the **y-axis** from the palm to the back of the hand, and the **z-axis** from the little finger toward the thumb.
- **Right Hand:** the **x-axis** points along the forearm direction, the **y-axis** from the palm to the back of the hand, and the **z-axis** from the little finger toward the thumb.

We provide a visualization of the hand motion in our dataset (see Fig. 8).

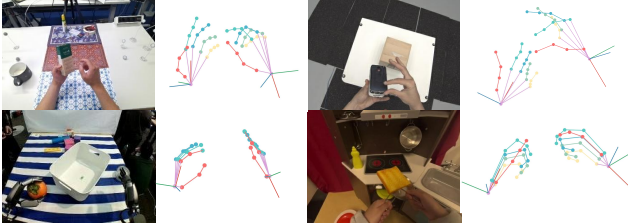


Figure 8. **Visualization of hand motion.** We use 3D keypoints to represent hand motion. To ensure consistency across embodiments, the wrist frames of human and robot hands are strictly aligned.

Data Collection Interface. To ensure data quality, we built a data collection interface that visualizes the video stream, wrist poses and fingertip keypoints in real time. This platform allows the operator to monitor the tracking status of all sensors. When the operator identifies tracking failures or degraded quality, the recording session can be manually terminated anytime. The entire system operates at 20HZ, balancing motion fidelity with reliable multi-sensor synchronization.

B. Motion-aware dynamics tokens

To empower the Vision-Language Model (VLM) with enhanced motion comprehension during action execution, we introduce a **Motion-aware Dynamics Tokens** framework as a supervisory mechanism. This framework is structured to decompose dynamic information into two complementary components: **Vision Dynamics** and **Motion Dynamics**.

The **Vision Dynamics** component is designed to learn a compact, discrete representation of the visual change between consecutive frames. This process begins with an encoder $\text{Enc}_V(\cdot)$ based on a self-attention mechanism. The encoder’s input is a concatenation of three elements: the visual features I_t, I_{t+k} from the image sequence, the corresponding motion features $P_{t,t+1,\dots,t+k}$, and a set of randomly initialized, learnable parameters D_{vis} :

$$\hat{D}_{vis} = \text{Enc}_V(I_t, I_{t+k}, P_{t,t+1,\dots,t+k}, D_{vis})$$

These parameters are then discretized using a VQ-VAE objective, resulting in a final representation composed of $V = 4$ tokens selected from a candidate codebook of size $|C_v| = 16$

$$D'_{vis} = \mathbf{VQ}(\hat{D}_{vis})$$

This architecture allows the visual information to be deeply infused with motion characteristics. Subsequently, a decoder $\text{Dec}_V(\cdot)$ takes the first frame I_t along with the refined parameters D'_{vis} and is tasked with reconstructing the final frame I_{t+k} :

$$I_{t+k} = \text{Dec}_V(I_t, D'_{vis})$$

This auto-encoding objective forces D'_{vis} to encapsulate the essential visual transformation between the two frames. Conversely, the **Motion Dynamics** component aims to capture more complex and granular motion patterns directly from raw action data. We employ a two-layer residual quantization (RQ) architecture for this purpose. To preserve the most authentic motion characteristics, the original action information $M_{t,t+1,\dots,t+k}$ is first processed through a Pose Network $\text{PoseNet}(\cdot)$ and then fed directly into the RQ network $\text{RQ}(\cdot)$:

$$P_{t,t+1,\dots,t+k} = \text{PoseNet}(M_{t,t+1,\dots,t+k})$$

$$D_{mot} = \mathbf{RQ}(P_{t,t+1,\dots,t+k})$$

This architecture selects a total of $R = 40$ tokens from a larger, shared codebook of size $|C_m| = 512$, enabling a hierarchical and fine-grained representation of motion. The motion dynamics tokens D_{mot} are then decoded using a Temporal Convolutional Network (TCN) to reconstruct the original motion sequence, serving as the supervision signal to ensure the quantized representation preserves essential temporal dynamics.

$$M_{t,t+1,\dots,t+k} = \text{TCN}(D_{mot})$$

A key design principle is uniformity: all codebooks maintain a unified feature dimension $d = 128$. This ensures consistent representation for downstream processing while simultaneously enabling each component to effectively capture the hierarchical spatiotemporal patterns crucial for advanced action-aware modeling.

C. Task details

Pick and Place. In this task, the robot performs a complete pick-and-place operation involving a common household object—an apple. The robot first identifies the position of the apple on the tabletop through egocentric visual observation and moves its hand toward the target. It then adjusts its wrist orientation and finger configuration to achieve a stable grasp. After successfully lifting the apple, the robot transports it to the target area and gently places it into a plate. This task assesses the model’s capability to perceive and manipulate objects accurately, as well as its ability to coordinate visual perception and fine-grained hand control in a sequential manipulation process. Success is achieved if the apple is successfully placed into the plate.

Close laptop. In this task, the robot is required to close a partially open laptop. The robot places four fingers along the back edge of the laptop cover and applies a downward motion by rotating the wrist, allowing the lid to close smoothly. This task evaluates the model’s ability to control contact-rich interactions and perform coordinated multi-finger and wrist motions for precise object manipulation. Success is achieved if the laptop is fully closed.

Open Drawer. In this task, the robot is required to open a partially closed drawer. The robot first bends its four fingers and positions the hand so that the fingertips can reach into the handle of the drawer. It then moves the wrist backward while maintaining a stable grip on the handle, gradually pulling the drawer open. This task requires precise hand movements guided by visual feedback and involves contact-rich manipulation that demands fine coordination between finger articulation and wrist motion. Success is achieved if the drawer is pulled out to a fully open position.

Grasp two drinks into basket. In this bimanual manipulation task, the robot is required to sequentially grasp and

place two bottled drinks into a basket. The robot first raises its left hand, reaches toward the bottle, and performs a stable grasp before lifting it and placing it into the basket. It then raises the right hand to repeat the same motion for the second bottle, ensuring symmetric and coordinated control between both hands. After both bottles are successfully placed, the robot closes both hands into fists and pushes the basket forward to complete the task. This task evaluates the model’s ability to perform coordinated bimanual manipulation and to plan sequential actions that involve precise grasping and contact-rich object placement. Success is achieved if both bottles are correctly placed into the basket and the basket is pushed forward.

Put Cola into Basket. This task requires precise bimanual cooperation to manipulate and place a bottle of cola into a basket. The robot first moves its left hand toward the cola bottle and grasps it from the lower side, lifting it to chest height. The right hand then approaches slowly and grasps the upper part of the bottle, ensuring a stable handover between both hands. After securing the bottle, the robot releases the left hand while the right hand moves toward the basket and places the bottle inside. This task evaluates the model’s ability to perform coordinated bimanual-hand manipulation, including object handover, grasp stability, and visually guided placement. Success is achieved if the cola bottle is stably placed inside the basket.

Open drawer and put bread. This long-horizon task requires the robot to perform a sequence of coordinated actions involving both hands. The robot first bends the fingers of its right hand and moves the wrist forward so that the fingertips can reach into the drawer handle. It then pulls the drawer open smoothly through a controlled backward motion. Next, the robot moves its left hand to grasp a piece of bread randomly placed on the tabletop, lifts it, and positions it above the open drawer. The bread is then released into the drawer, after which the robot uses its right hand to close the drawer. This task challenges the model’s ability to perform long-horizon visuomotor control, requiring precise spatial reasoning, bimanual coordination, and contact-rich manipulation across multiple sub-actions. Success is achieved if the bread is successfully placed inside the drawer and the drawer is fully closed.

Grasp Apple into Basket. In this task, the robot is required to accurately identify the position of an apple among three objects (orange, apple, and banana) present simultaneously in its field of view, then precisely grasp the apple using four fingers. After successfully securing the apple, the robot transports it above a basket and gently places it inside. This task evaluates the model’s capability for object perception and manipulation in cluttered environments, as well as its ability to execute targeted grasping and placement operations amid visual distractions. Success is achieved when the apple is securely deposited into the basket.



Figure 9. Visualization of task progress across embodiments.

Tool Use. In this task, the robot is required to reposition a bread clip into an optimal grasping posture for functional use. The robot first precisely pinches the bread clip with its right hand, then transfers the tool’s end to its left hand for stabilization. Finally, the right hand regrips the bread clip at an operational angle suitable for precise manipulation. This task evaluates the model’s capability for bimanual coordination and its ability to dynamically adjust grasping strategies during tool transfer. Success is achieved when the tool is securely held in a functionally optimal posture through seamless handover between both hands.

More visualization of dexterous manipulation tasks can be found in Fig. 9.

D. Policy Implementation Details

D.1. Pretraining details

We use the EgoAtlas dataset to pretrain METIS, which integrates eight heterogeneous data sources covering both human and robot egocentric manipulation trajectories. To ensuring training efficiency and maintain a balanced distribution across data domains, we apply sampling to construct the

final training mixture. The detailed dataset composition and mixture weights are provided in Tab. 6.

Table 6. The dataset name and sampling weights used in EgoAtlas pretraining dataset.

Training Dataset Mixture	
H2O[22]	0.8%
OAKINK[56]	1.9%
PH2D[38]	5.4%
ARCTIC[7]	2.8%
EgoDex[12]	40.3%
Holoassist[50]	10.6%
ActionNet[8]	13.4%
Our Enhanced Data	25.4%

During pretraining, we jointly optimize all model parameters, including the vision encoder, the LLM backbone, and the action decoder. We use a global batch size of 768 (32 per device) and use distributed training under a fully sharded data-parallel (FSDP) setup. Optimization is performed with AdamW using a learning rate of $2e^{-5}$, no weight decay, and gradient clipping at a max-norm of 1.0. We train on a cluster of 24 NVIDIA H100 GPUs. Empirically, we find that 60k training steps are sufficient to achieve strong performance, which requires approximately 72 hours.

D.2. Posttraining details

For post-training, we use an 8-GPU setup with a per-device batch size of 4. We apply LoRA updates with a rank of 32 to both the LLM backbone and the vision encoder, while performing full-parameter fine-tuning on the action decoder. Only the LoRA-adapted weights and the action decoder parameters are included in the trainable set. We optimize the model using AdamW with a learning rate of $3.5e^{-4}$ and a weight decay of $1e^{-3}$. The learning rate schedule follows a StepLR policy, where the learning rate is decayed by a factor of 0.1 after 80% of the total training steps. To improve training efficiency, we adopt mixed-precision training: both the LLM backbone and the vision encoder are trained in bfloat16 precision, whereas the action decoder is kept in float32 for numerical stability. This configuration provides a favorable trade-off between speed and accuracy during post-training.

For the robot observation–action interface, we set the proprioceptive history to 1 and use an action chunk length of 32. Real-world data are collected at 30 Hz. The recorded images are first center-cropped to a resolution of 480×640 , and subsequently resized to 224×224 before being fed into the visual encoder.

E. Baseline Settings

For OpenVLA-OFT[19], all experiments are conducted on 4 NVIDIA H100 GPUs. We use the full joint configuration

of the dual-arm and dual-hand system as the policy action space (2×7 DoF for arm joints and 2×6 DoF for hand joints). We apply LoRA-based fine-tuning to the VLA and adopt an L1 regression–style action head for continuous action prediction. The action chunk length is set to 30. We use the following hyperparameters for training OpenVLA-OFT.

Table 7. Hyperparameters for OpenVLA-OFT.

Name	Value
Steps	40000
Batch Size	4
Learning Rate	$5e^{-4}$
Action Chunks	30
LoRA Rank	32
Action Head	<i>l1_regression</i>

For $\pi_{0.5}$ [14], post-training uses 1 NVIDIA H100 GPU across all tasks. We use wrist pose for arms and joints for dexterous hands (2×7 DoF for wrist poses and 2×6 DoF for hand joints). Action chunk size is set to 10. Hyperparameter configuration is as follows.

Table 8. Hyperparameters for $\pi_{0.5}$.

Name	Value
Steps	30000
Batch Size	32
Learning Rate	$5e^{-5}$
Action Chunks	10
Ema decay	0.999
Warm-up steps	10000

For Gr00t-N1.5[32], All experiments are conducted on 4 NVIDIA H100 GPUs. We use joint angles as the policy action space, which contains two 7-DoF arm joints and two 6-DoF hand joints. Fine-tuning is set to freeze the language model backbone and vision tower, only fine-tuning the projector and diffusion model. The action chunk length is set to 16, and some hyperparameters are set as follows.

Table 9. Hyperparameters for Gr00t-N1.5.

Name	Value
Steps	10000
Batch Size	32
Learning Rate	$1e^{-4}$
Weight Decay	$1e^{-5}$
Warmup Ratio	0.05
Action Chunks	16

For ACT[65], we choose to follow the lerobot[6] framework. All training is conducted on a single NVIDIA H100 GPU. At this point, we unify the action space with METIS, using

the 9D pose of the wrist and the 3D XYZ pose of the fingers as the action space. We use ResNet18 as the vision backbone for training, with the remaining transformer layers trained from scratch. The action chunk length is set to 100, and the hyperparameters are as follows.

Table 10. Hyperparameters for ACT.

Name	Value
Steps	600000
Batch Size	8
Learning Rate	$1e^{-5}$
Weight Decay	$1e^{-4}$
Warmup Ratio	0.05
Action Chunks	100
Vision Backbone	ResNet-18

F. Cross-Embodiment Settings

In the cross-embodiment experiments, we employ the Sharpa Beta embodiment equipped with a pair of 22-DoF SharpaWave Dexterous Hands. During the data collection phase, we capture first-person view images at 30 Hz along with joint state-action pairs at 60 Hz. The data processing pipeline begins with temporal alignment between the 30 Hz images and 60 Hz joint angles, synchronizing both modalities to 30 Hz. For the joint angle information, we perform forward kinematics (FK) calculations and coordinate frame transformations: the fingertip poses (3-dimensional) are transformed into the wrist coordinate frame, while the wrist poses (9-dimensional) are converted to the camera coordinate frame, ensuring a unified action space. During inference, the model’s output fingertip and wrist poses undergo inverse kinematics (IK) computations to convert them back to joint angles for robot control.