Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Ψ_0: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation

Songlin Wei[1*], Hongyi Jing[1*], Boqian Li[1*], Zhenyu Zhao[1*], Jiageng Mao[1], Zhenhao Ni[1], Sicheng He[1], Jie Liu[1], Xiawei Liu[1], Kaidi Kang[1], Sheng Zang[1], Weiduo Yuan[1], Marco Pavone[2], Di Huang[3], Yue Wang[1†]
[1]USC Physical Superintelligence (PSI) Lab [2]NVIDIA [3]WorldEngine
* Equal Contribution † Corresponding Author
https://psi-lab.ai/Psi0

Fig. 1: **Humanoid Loco-Manipulation.** Ψ_0 performs diverse loco-manipulation tasks in a pantry, including taking a cup from the coffee machine, pushing a cart, wiping the table, grasping a bottle and placing it in the sink, and pushing the fridge door.

***Abstract*—We introduce Ψ_0 (*Psi-Zero*), an open foundation model to address challenging humanoid loco-manipulation tasks. While existing approaches often attempt to address this fundamental problem by co-training on large and diverse human and humanoid data, we argue that this strategy is suboptimal due to the fundamental kinematic and motion disparities between humans and humanoid robots. Therefore, data efficiency and model performance remain unsatisfactory despite the considerable data volume. To address this challenge, Ψ_0 decouples the learning process to maximize the utility of heterogeneous data sources. Specifically, we propose a staged training paradigm with different learning objectives: First, we autoregressively pre-train a VLM backbone on large-scale egocentric human videos to acquire generalizable visual-action representations. Then, we post-train a flow-based action expert on high-quality humanoid robot data to learn precise robot joint control. Our research further identifies a critical yet often overlooked data recipe: in contrast to approaches that scale with noisy Internet clips or heterogeneous cross-embodiment robot datasets, we demonstrate that pre-training on high-quality egocentric human manipulation data followed by post-training on domain-specific real-world humanoid trajectories yields superior performance. Extensive real-world experiments demonstrate that Ψ_0 achieves the best performance using only about 800 hours of human video data and 30 hours of real-world robot data, outperforming baselines pre-trained on more than 10$\times$ as much data by over 40% in overall success rate across multiple tasks. We will open-source the entire ecosystem to the community, including a data processing and training pipeline, a humanoid foundation model, and a real-time action inference engine.**

I. Introduction

Humanoid robots, endowed with human-like morphology and dexterity, have achieved remarkable progress in whole-body motion control [12, 28, 22, 14]. However, their manipulation capabilities, which could eventually unlock enormous potential for society, have received less attention and faced greater challenges. Recent advances in large language models (LLMs) have illuminated a promising path towards intelligence: by scaling both data and model capacity, general intelligence can emerge. Inspired by this paradigm, the robotics community has begun exploring scaling laws that are suitable for agents with physical bodies. Recently, works such as RT 1-2 [8, 47], OpenVLA [24], Gemini Robotics [37], GR00T [4], and Physical Intelligence's $\pi_0, \pi_{0.5}$ [5, 21] have advocated training large action models using massive amounts of real robot data. These approaches provide early evidence that the reasoning and planning abilities of large models can significantly improve generalization in robotic manipulation. However, these methods often rely on large-scale teleoperation data, which is prohibitively costly and challenging to acquire for humanoid loco-manipulation.

Fortunately, human egocentric videos provide a scalable alternative as they capture abundant natural motion patterns and rich behavior-level information without the expense of robot teleoperation. However, directly transferring knowledge from human videos to humanoid control is non-trivial due to the substantial embodiment gap between humans and robots. Early efforts [10, 39, 3] attempt to learn from human videos by adopting a unified human-centric state-action representation. Nevertheless, learning from such heterogeneous data remains challenging due to intrinsic discrepancies between humans and humanoids, including differences in action frequency, motion dynamics, and degrees of freedom. Although these approaches employ domain adaptation [10] or co-training strategies that mix human and robot data [39], a single monolithic policy that models two fundamentally different action distributions is inherently suboptimal. As a result, the learned policies still struggle to control humanoids to perform complex, long-horizon tasks. Therefore, this paper studies a fundamental question: *how can we effectively distill motion priors and world knowledge from human egocentric videos to enable robust whole-body control for humanoid robots?*

To that end, we propose a novel multi-stage training paradigm with different learning goals for each stage: we first pre-train a VLM to predict next-step actions using the human-robot unified action space. The objective of this stage is to enable the model to learn task-level motion priors across diverse activities, while also learning visual representations aligned with downstream robotic tasks. We then train a separate flow-based action expert using real humanoid robot data to predict action sequences directly in the joint space. This post-training stage includes both task-agnostic training on cross-task humanoid data and task-specific fine-tuning on in-domain teleoperated demonstrations. We implement our action expert as a multi-modal diffusion transformer (MM-DiT) [15], which is more capable than a naive DiT. Conditioned on the visual-language features from the VLM, the action expert efficiently and concurrently outputs joint-space action chunks. This stage enables the action expert to capture embodiment-specific dynamics. As a result, only a small amount of additional real-robot data is required for task-specific fine-tuning, after which the model can rapidly acquire long-horizon, dexterous loco-manipulation skills.

To enable effective training and deployment of our humanoid VLA, we make several key contributions. First, we optimize a manipulation-oriented teleoperation pipeline that improves lower-body stability during whole-body manipulation. Second, to ensure smooth execution in the real world at inference time, we introduce training-time real-time action chunking, which mitigates motion jitter caused by model inference latency. Finally, we deploy our model on a real humanoid robot and benchmark it against state-of-the-art methods on several complex, long-horizon tasks. Our experiments suggest that, using only 800 hours of human egocentric video and 30 hours of real-robot data, our model achieves significantly better performance than existing methods trained with more than 10 $\times$ as much data on long-horizon loco-manipulation tasks. These results reveal that **effective scaling requires scaling the right data in the right way.** We will release the full training pipeline, pre-trained model weights, deployment code to facilitate future research.

II. Related Works

A. Whole-Body Dexterous Manipulation

Humanoid whole-body control has witnessed significant progress in recent works [41, 12, 26, 27, 16, 1, 44, 35]. Humanoid robots are now able to mimic diverse human motions like running, dancing, and even flipping. Despite significant progress in locomotion, researchers have struggled to achieve comparable success in humanoid dexterous loco-manipulation. LangWBC [36] and LeVERB [38] introduce language-conditioned whole-body control policies, allowing humanoid robots to robustly execute high-level and language-specified behaviors. However, these methods primarily focus on locomotion and navigation, with limited emphasis on dexterous manipulation scenarios. In parallel, AMO [25] and TWIST2 [42] enable humanoid whole-body control through VR-based teleoperation, providing an effective framework for collecting loco-manipulation data. However, they emphasize more on low-level control, rather than learning a precise policy for long-horizon dexterous loco-manipulation.

Dexterous manipulation [18], on the other hand, poses a long-standing challenge due to the high degree-of-freedom control and frequent self-occlusion between palms and fingers, which make vision-based dexterous manipulation extremely challenging. Being-H0 [29] proposes to learn from human video by curating a large amount of hand-object interaction videos and fine-tuning a pre-trained VLM using multiple task data like motion-infilling and translation. However, this method is limited to single-arm tabletop manipulation. To address the mentioned challenges, we propose to build a

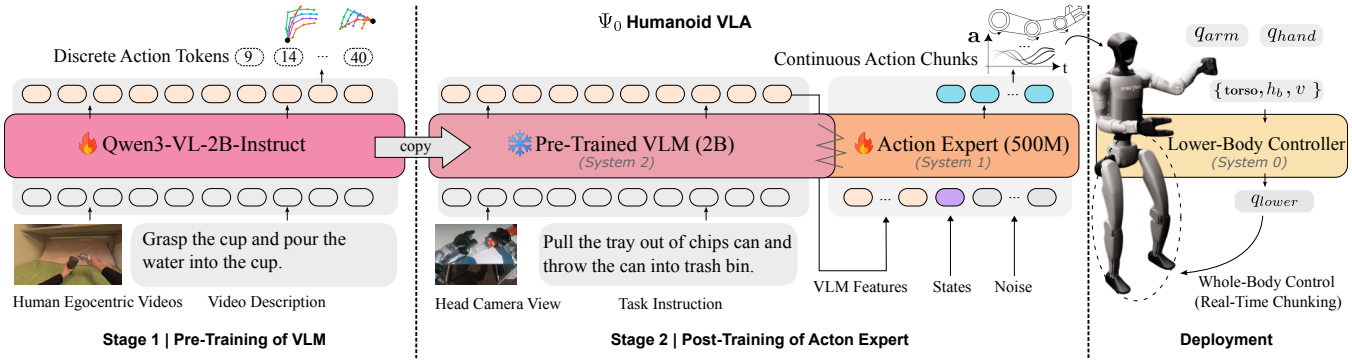


Fig. 2: **Model Training and Deployment**: First, we pre-train the VLM on the EgoDex [20] dataset to autoregressively predict the next-action tokens in the task space. Then, we post-train the flow-based action expert using robotic data to predict action chunks in the joint space. Finally, we implement a real-time chunking mechanism that leverages the lower-body controller to achieve smooth whole-body control.

unified VLA model for humanoid whole-body dexterous manipulation.

B. Humanoid VLAs

Inspired by the remarkable success of foundation models, VLAs [47, 24, 5, 43, 17, 37] have emerged as a promising direction toward bringing artificial intelligence into the physical world. π series [5, 21] demonstrate exceptional generalization and robustness across challenging manipulation scenarios, including dual-arm and mobile manipulation. GR00T [4] further open-sources the first foundation model for humanoid robots, trained on a large-scale mixture of real-world and synthetic data generated from videos. However, in contrast to them, we find that training on higher-quality data is more critical than simply scaling to large volumes of heterogeneous cross-embodiment data. In this work, we explore a new paradigm for training humanoid VLAs that leverages large-scale human egocentric video data, complemented by a smaller amount of real robot interaction data.

C. Learning From Egocentric Videos

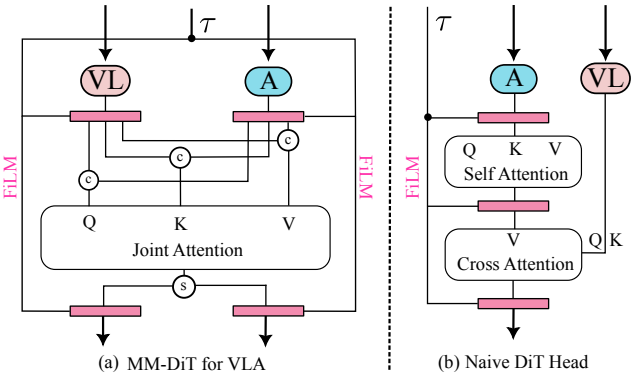
Data scarcity remains a fundamental constraint in training VLAs, as teleoperation data collection is less efficient and more expensive to scale. In contrast, human video data contains rich prior knowledge of human–object interactions [32, 23, 40], providing a scalable alternative. Recent approaches, such as EgoVLA [39] and In-n-On [10], co-train their models on human video and robot data to predict unified human wrist and hand actions, followed by inverse kinematics (IK) during inference to map these predictions to robot actions. Similarly, H-RDT [3] trains a large diffusion transformer (DiT) to predict arm and hand actions in the end-effector space. However, co-training the model end-to-end on a mixture of humanoid and non-humanoid robot data is suboptimal, as the model must simultaneously learn two fundamentally different action distributions. Instead, we identify a critical yet overlooked training recipe: after pre-training with next-action prediction to learn task semantics and visual representations, we post-train the action expert to directly model actions in the joint space, thereby avoiding the inefficiencies of co-training.

III. The Ψ_0 Foundation Model

In this section, we introduce Ψ_0 (*Psi-Zero*), a VLA model for humanoid dexterous loco-manipulation. Given a natural language task instruction ℓ and the current observation $\mathbf{o}_t$, our model predicts the whole-body action chunk $\mathbf{a}_{t:t+H}$. The observation $\mathbf{o}_t$ contains the current head camera image $\mathbf{I}_t$ and the whole-body proprioceptive state $\mathbf{q}_t$, including upper joint state, torso roll, pitch, yaw, and the base height. The action $\mathbf{a} \in \mathbb{R}^{36}$ is defined as $\{\mathbf{q}_{hand}, \mathbf{q}_{arm}, \mathbf{torso}_{rpy}, h_b, v_x, v_y, v_{yaw}, p_{yaw}\}$, where $\mathbf{q}_{hand} \in \mathbb{R}^{14}$ and $\mathbf{q}_{arm} \in \mathbb{R}^{14}$ are the two hand and arm joints respectively, $\mathbf{torso}_{rpy} \in \mathbb{R}^3$ is the torso roll, pitch, yaw. $h_b \in \mathbb{R}$ is the base height of the humanoid and $v_x, v_y \in \mathbb{R}$ are the horizontal linear velocities, and $v_{yaw} \in \mathbb{R}$ denotes angular velocity around the upward direction. $p_{yaw} \in \mathbb{R}$ is the target yaw rotation. We employ an RL-based control policy [25] to control the lower body and torso joints throughout data collection and policy evaluation.

A. Model Architecture

Ψ_0 is a foundation model that adopts a *triple-system* architecture, following prior work [21, 4]. As shown in Fig. 2, the high-level policy consists of two end-to-end–trained components: a vision–language backbone (*system-2*) and a multi-modal diffusion transformer (MM-DiT) action expert (*system-1*). We use the state-of-the-art vision–language foundation model Qwen3-VL-2B-Instruct [2] as *system-2*. The action expert is implemented as a flow-based MM-DiT inspired by Stable Diffusion 3 [15], containing approximately 500M parameters. Compared to a naive DiT-based action head, this design enables more efficient fusion of action and vision–language features. Conditioned on hidden features from the VLM backbone, the action expert predicts future whole-body action chunks $\mathbf{a}_{t:t+H}$. The 8-DoF lower-body actions $\{\mathbf{torso}_{rpy}, h_b, v_x, v_y, v_{yaw}, p_{yaw}\}$ are passed to *system-0*, a RL–based tracking policy. We adopt the off-the-shelf controller AMO [25], which maps these inputs to 15-DoF lower-body joint angles $\mathbf{q}_{lower} \in \mathbb{R}^{15}$, including 3 DoF waist and 12 DoF leg joint. Together with the 28-DoF upper-body joints $(\mathbf{q}_{arm}, \mathbf{q}_{hand})$, the system outputs 43-DoF actions for whole-body control.



(a) MM-DiT for VLA (b) Naive DiT Head

Fig. 3: **MM-DiT for VLA:** Comparison of MM-DiT architecture with naive DiT. τ is the flow timestep and **VL** and **A** denotes hidden states of the vision-language and action respectively.

B. Training Recipe

We present an efficient training recipe for learning humanoid loco-manipulation skills from both human videos and real robot data. The overall training procedure consists of three stages: first, pre-training the VLM backbone on the large-scale high-quality and diverse human egocentric videos; second, post-training the flow-based action expert on cross-task real humanoid data; and third, fine-tuning the action expert using a small amount of in-domain task data, which enables rapid adaptation to new tasks.

1) Pre-Training on Egocentric Human Video: Training a humanoid foundation model faces a significant data scarcity bottleneck. Human egocentric videos, which are much cheaper to scale than real-world robotics data, offer a promising alternative. Therefore, we leverage EgoDex [20], which contains approximately 829 hours of human egocentric video capturing human hands performing diverse dexterous manipulation tasks. To further mitigate the visual gap between human videos and robotic observations, we incorporate Humanoid Everyday [46], which contains 31 hours of humanoid data covering 260 diverse tasks, ranging from human–object interactions to manipulations of deformable and articulated objects. We use a shared action representation for both human hands and robot end-effectors. Specifically, the 48-DoF action in task space is defined as $\mathbf{a} \triangleq \{\mathbf{a}_l, \mathbf{a}_r\}$ and each $\mathbf{a}_l$ or $\mathbf{a}_r \in \mathbb{R}^{24}$ is $\{\mathbf{T}_{wrist}, \mathbf{P}_{thumb}, \mathbf{P}_{index}, \mathbf{P}_{middle}, \mathbf{P}_{ring}, \mathbf{P}_{pinky}\}$. The $\mathbf{T} \in \mathbb{R}^9$ is the 9-DoF wrist pose vector consisting of 3D position and 6D rotation. Each $\mathbf{P} \in \mathbb{R}^3$ is a 3D fingertip position. Such unified action representation enables joint training of human and robot data and achieves stable training.

However, naively training the model to autoregressively predict multiple high-dimensional action chunks is very computationally expensive and drastically slows down pre-training. Our key insight is that the goal of pre-training the VLM backbone is to learn the task semantics of the language instruction and the visual representation for downstream real-robot manipulations. Predicting a single next-step action suffices for such a goal. Therefore, we train the VLM to predict only a single-step action $\mathbf{a}_t$ instead of $\mathbf{a}_{t:t+H}$, which requires much less computation. We use FAST [33] to tokenize continuous actions into discrete tokens. We train the FAST tokenizer on 500,000 randomly sampled actions from EgoDex [20]. The final trained tokenizer achieves an average L1 reconstruction loss of 0.005, and compresses each action sequence from 48 tokens to a variable token length $N \approx 20$. Then, the VLM is trained autoregressively to predict next-action tokens, *i.e.,* to maximize

$$p_\theta(\mathbf{a}) = \prod_{t=1}^{N} p_\theta(\mathbf{a}_t | \mathbf{a}_{<t}, \ell, \mathbf{o}_t). \tag{1}$$

2) Post-Training on Cross-Task Real Humanoid Data: After the VLM backbone is trained, we freeze its parameters and train the action expert from scratch. Conditioning on the hidden feature extracted from the VLM backbone $\mathbf{z}_t = f_\theta^{vlm}(\mathbf{o}_t, \ell)$, and a uniformly sampled flow timestep $\tau \in [0, 1]$, the flow-matching training objective is

$$\mathcal{L}_{fm} = \mathbb{E}\left[\| v_\rho^{flow}(\mathbf{z}_t, \mathbf{a}_t^\tau, \tau) - (\boldsymbol{\epsilon} - \mathbf{a}_t) \|\right] \tag{2}$$

where $\boldsymbol{\epsilon}$ is Gaussian noise and $\mathbf{a}_t^\tau = \tau \mathbf{a}_t + (1-\tau)\boldsymbol{\epsilon}$ is the noised action. We adapt the MM-DiT architecture [15] to implement the action expert network v_ρ^{flow}, as illustrated in Fig. 3. Specifically, the model uses the time-conditioning feature τ to modulate the action (A) feature and the vision–language (VL) features separately. During each transformer block, the action tokens and VL tokens perform joint global attention, which facilitates more effective fusion of visual information compared to the naive DiT.

3) Fine-Tuning on In-domain Teleoperation Data: With the pre-trained VLM and the post-trained action expert, our model can be fine-tuned further end-to-end using a small amount of in-domain data and rapidly learn long-horizon, dexterous loco-manipulation tasks. We evaluate the model on eight real-world tasks (as illustrated in Fig. 6), each posing distinct challenges: some require precise arm coordination, while others demand long-distance navigation. Most tasks exceed 2,000 steps at 30Hz, rendering them truly long-horizon. Each task contains three to five sub-tasks, and each sub-task corresponds to a skill such as grasping or pushing.

C. Real-Time Action Chunking

Humanoid robots require smooth and reactive control, particularly when executing long-horizon, dexterous manipulation tasks. However, existing VLAs typically contain billions of parameters, which inevitably introduce a "stop-and-think" behavior due to inference latency. Our Ψ_0 model similarly comprises over 2.5 billion parameters, with a single forward pass taking approximately 160 ms. To enable smooth policy rollout despite this latency, we adopt training-time real-time chunking (RTC) following [7]. With RTC, each action prediction is conditioned on the previously committed action chunk and outputs a consistent chunk of future actions, as illustrated in Fig. 4. To faithfully simulate inference delay during training, we randomly remove diffusion noise from the first $d = \texttt{uniform}(0, d_{\max})$ tokens and mask them out in the loss computation in Eq. 2. Here, $d_{\max} \in [0, H - s)$ denotes the maximum inference delay in timesteps, while H and s correspond to the action chunk prediction horizon and the execution horizon, respectively.

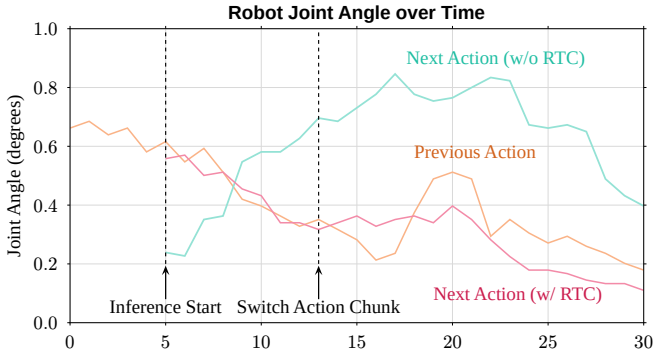


Fig. 4: **Real-Time Chunking:** Given that the previous action is being executed (yellow line), the next action can diverge significantly (cyan line) without RTC, which leads to control jitter. With RTC (red line), the divergence between two consecutive actions is strongly suppressed, resulting in smoother and more stable behavior.

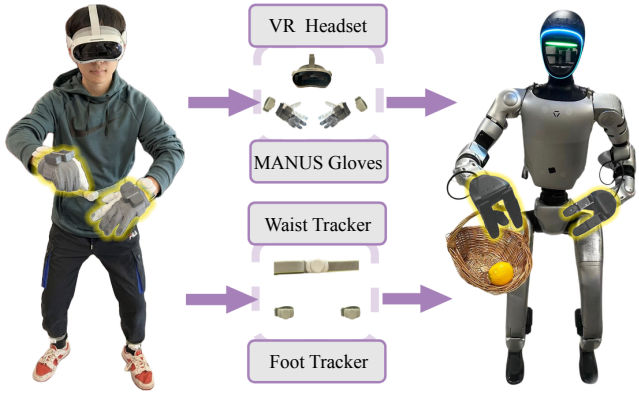


Fig. 5: **Real-Robot Teleoperation Setup:** We use MANUS gloves for dexterous hand retargeting; a VR headset and wrist trackers capture upper-body poses for inverse kinematics, while waist and foot trackers provide high-level locomotion commands.

D. Tailoring Teleoperation for Loco-Manipulation

Efficiently learning a long-horizon loco-manipulation task critically depends on the quality of in-domain data for fine-tuning. However, existing teleoperation systems are primarily designed for locomotion and lack the stability and adaptability required for dexterous manipulation. Designing an effective teleoperation system for humanoid loco-manipulation requires balancing whole-body expressiveness, locomotion stability, and operational simplicity. Existing end-to-end whole-body teleoperation pipelines [42, 30] that directly map full-body human motion to humanoid control through reinforcement learning often suffer from limited robustness due to noisy tracking signals and unstable whole-body motion patterns. Moreover, these systems rely on handheld controllers and reduce dexterous hand control to low-dimensional gripper-like commands, limiting manipulation expressiveness. On the other hand, systems that decouple manipulation from locomotion through explicit base commands [25] improve lower-body stability, but typically require additional controllers or multiple operators and thus reduce practicality.

To address these limitations, we propose a tailored teleoperation framework that explicitly separates upper-body pose tracking, dexterous manipulation, and locomotion commands, while enabling single-operator whole-body control. As shown in Fig. 5, the teleoperator's upper-body pose is captured using a PICO headset [34] and wrist trackers, and a multi-target inverse kinematics solver is implemented to compute the humanoid's arm and torso configurations. Fine-grained finger motions are acquired using MANUS gloves [31], allowing direct control over all degrees of freedom of the dexterous hands. Locomotion commands, including translational velocity and turning orientation, are directly inferred from waist and foot trackers and provided as high-level commands to a RL policy [25] responsible for stable lower-body control.

By using a small set of wearable trackers and separating locomotion from in-place whole-body actions, our framework enables single-operator humanoid teleoperation with improved locomotion stability across diverse task scenarios. Furthermore, the combination of wrist trackers and MANUS gloves alleviates common occlusion and out-of-view issues in vision-based VR tracking, enabling accurate and reliable upper-body and hand tracking. Together, these design choices support robust and practical humanoid whole-body teleoperation for complex loco-manipulation tasks.

IV. Experiments

A. Implementation

1) Hardware Platform: Throughout all real-world experiments, we employ the Unitree G1 humanoid platform, which provides 29 degrees of freedom for whole-body control. In addition, each arm is equipped with a 7-DoF Dex3-1 dexterous hand. Visual observations are obtained using the default head-mounted Intel RealSense D435i camera.

2) Data Preparation: The EgoDex dataset contains approximately 900M frames and provides per-frame global transformation matrices for the upper humanoid body, including 7 spine joints, 2 arms, and 21 joints for each hand. To improve pre-training efficiency, all actions are transformed into the current head-camera coordinate frame, and the frame rate is upsampled by a factor of 3. Due to the presence of extreme outliers in EgoDex, action values are normalized using the 1st and 99th quantiles. We omit state inputs during the pre-training stage. We use the Humanoid Everyday dataset [46] for task-agnostic post-training, which contains approximately 3 million frames of real-world teleoperated data. Actions are represented as 36-DoF joint-space vectors $a = \{\mathbf{q}_{hand}, \mathbf{q}_{arm}, \mathbf{torso}_{rpy}, h_b, v_x, v_y, v_{yaw}, p_{yaw}\}$. Since Humanoid Everyday only provides upper-body motion, we similarly pad missing lower-body action components. States consist of 28-DoF joint positions of both hands and arms from the current frame and are fed into the model without normalization.

3) Training Details: Training begins by fitting a FAST tokenizer using 500,000 randomly sampled actions from EgoDex. The resulting L1 reconstruction loss on held-out action data is approximately 0.005, improving upon the 0.01 using the original open-source FAST tokenizer. The FAST tokenizer

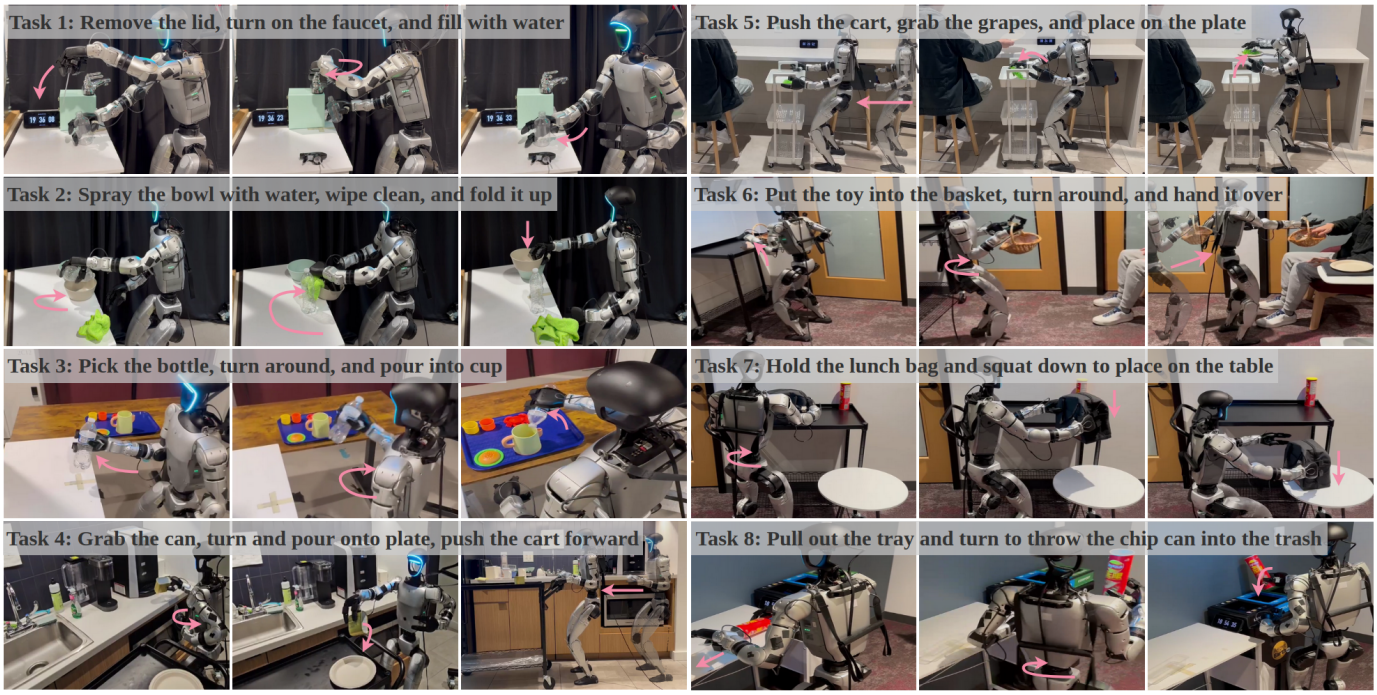


Fig. 6: **Real-World Task Setup:** We evaluate Ψ_0 on eight diverse long-horizon dexterous loco-manipulation tasks involving manipulation, whole-body motion, and locomotion. The task instruction is overlayed on the task images and each sub-task is denoted with marker for better visualization. Our policy rollout videos are included in the Supplementary Materials.

compresses each action sequence into 20 tokens which accelerates subsequent training. Then, we fine-tune Qwen3-VL-2B-Instruct [2] during the pre-training stage using 64 A100 GPUs for 10 days. Training is formulated as next-action prediction only, and we avoid action chunking to reduce computational overhead. The learning rate is fixed at 0.0001 and the global batch size is 1024. Next, we post-train the action expert, containing approximately 500M parameters, on the Humanoid Everyday dataset. During this stage, the VLM backbone is frozen, the learning rate is fixed at 0.0001, and the global batch size is set to 2048. This stage takes approximately 30 hours on a single node with 32 A100 GPUs. Finally, we fine-tune only the action expert for each downstream task for 40,000 steps, using a cosine learning rate scheduler with an initial learning rate of 0.0001.

B. Real-World Humanoid Experiments

1) Task Description: As shown in Fig. 6, we evaluate Ψ_0 on eight real-world long-horizon manipulation tasks spanning diverse daily scenarios. The tasks range from simple interactions, such as pick-and-place, pushing, and wiping, to more challenging dexterous manipulations requiring precise finger-object coordination, including turning a faucet and pulling out a chip tray. Beyond upper-body manipulation, the tasks also involve whole-body motions, such as torso rotation and squatting, as well as lower-body locomotion and turning. Overall, this evaluation benchmarks model performance on complex long-horizon dexterous loco-manipulation tasks across multiple real-world environments.

2) Evaluation Protocols: We collect 80 teleoperated trajectories for each task. All baseline models are fine-tuned on the same dataset, using identical image observations as well as the same action and state representations. Each long-horizon task consists of three to five sub-tasks involving dexterous manipulation, dual-arm coordination, and locomotion. As a result, policies may fail at early sub-tasks, which can lead to complete rollout failure. To fully assess the capabilities of each baseline, the evaluator is allowed to intervene and assist the policy in progressing past failed sub-tasks so that execution can continue. We therefore report success rates for individual sub-tasks in addition to the overall task success rate. For each task, we perform 10 rollout trials per model. A rollout is considered successful only if all sub-tasks are completed. All baselines, including Ψ_0, are deployed using the same client code to control the robot.

3) Baselines: We conduct comprehensive real-world benchmarking against most recent open-source baselines. We invest huge effort to reproduce the best possible results for each.

a) $\pi 0.5$: demonstrates strong generalization on mobile robot platforms with dual arms and grippers. However, the released model and checkpoint are limited to 30-dimensional action spaces. To adapt the model to humanoid tasks, we expand the action dimension to 36 and set the action chunk size to 16. The checkpoint weights of the corresponding linear layers are padded accordingly to accommodate the expanded action space. To account for the embodiment gap between the original training data and humanoids, we increase the learning rate from 1e-5 to 1e-4 and the global batch size from 32 to 128

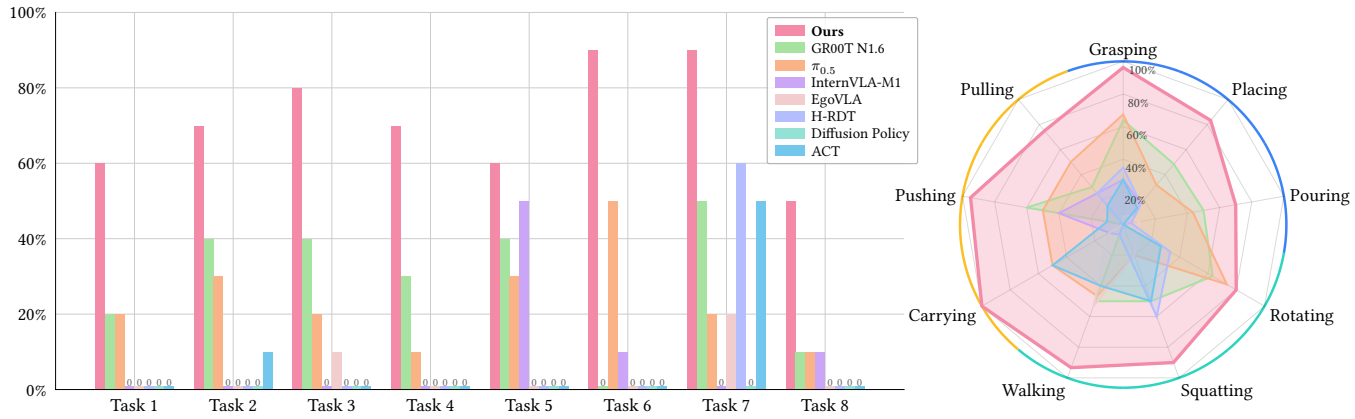


Fig. 7: **Real-World Benchmark:** Evaluation results of policies across our eight tasks, showing task-wise success rates (%) (left) and aggregated skill-level success rates (%) (right). The task descriptions are shown in Fig. 6. Detailed results for each task including all sub-task progress are included in the **Supplementary Materials**.

for better performance, ensuring fair comparison. We fine-tune the *Pi05_DROID* checkpoint, which we convert to a PyTorch implementation.

b) GR00T N1.6: shows strong performance in grasping and loco-manipulation, with robust spatial generalization. We use all the default hyperparameters for fine-tuning in the release code. We initialize the model from the GR00T N1.6 3B pre-trained checkpoint and fine-tune it on our teleoperated data for 20,000 steps with a global batch size of 24 on three NVIDIA A100 GPUs. We use cosine scheduling for the learning rate at 1e-4. As the RTC inference code for GR00T N1.6 is not publicly available in the official repository, we adopt a standard sequential inference scheme, in which the observation corresponding to the most recently executed action is used to condition the prediction of subsequent actions.

c) InternVLA-M1 [11]: is a unified framework for spatial grounding and robot control, which demonstrates strong spatial reasoning capabilities. However, it is only pre-trained on spatial reasoning and robotic arm data which limits its performance on humanoid tasks. We start with the checkpoint pre-trained on the RT-1 Bridge dataset, freeze the VLM backbone and fine-tune the action head for 30,000 steps with a batch size of 64 on a single NVIDIA A100 GPU. In our experiments, InternVLA-M1 exhibits action jitter across consecutive action chunks, resulting in unstable executions.

d) H-RDT: is a single large DiT action expert with 2B parameters. We train the model for 10,000 training steps with a batch size of 32 on a single NVIDIA A100 GPU. The resulting policy excels at tasks that do not require precise movements. However, it struggles with manipulation tasks that require high-precision across many joints.

e) EgoVLA: is a vision–language–action model pre-trained on egocentric human manipulation videos using EgoDex and additional data sources. Since the original codebase predicts only end-effector wrist and hand poses, we adapt the action decoder to output robot joint-space commands required by downstream tasks. We fine-tune the pre-trained EgoVLA on our teleoperated downstream tasks following the training configuration reported in the original paper, training for 115 epochs with an effective batch size of $16\times8\times4$. In our experiments, EgoVLA shows limited performance on lower-body commands, likely because its pre-training primarily captures upper-body and hand manipulation skills and does not provide strong priors for coordinated lower-body motion.

f) Diffusion Policy (DP) [13]: For visual feature extraction, we employ a pre-trained ResNet-18 [19] as the visual encoder. We set the learning rate to 1×10^{-4} and the global batch size to 32. Training is conducted for 40,000 steps using two A100 GPUs, with each task trained for approximately 15 hours. We observe that DP fails on most tasks, even though it can reasonably fit the training data. We conjecture that the UNet-based DP model has insufficient visual capacity. During inference, we perform 100 iterative denoising steps to progressively transform random noise into actionable trajectories.

g) Action Chunking with Transformers (ACT) [45]: To adapt to the humanoid locomotion and manipulation tasks, we reconfigure the action head to output 36-dimensional actions and tune the chunk size to 100, and initialize the transformer block with a configuration of 4 encoder layers and 1 decoder layer, aligning with the publicly released ACT framework [9]. Other training hyper-parameters like learning rate, batch size and training steps are kept the same as DP.

4) Comparisons to Baselines: As illustrated in Fig. 7, our model outperforms all baselines by a large margin. Our model exhibits the most stable performance across all eight long-horizon dexterous loco-manipulation tasks. Notably, it achieves an average overall success rate that is at least 40% higher than that of the second-best baseline, GR00T-N1.6 [4], which is the most recently released humanoid foundation model. These results highlight the effectiveness of our training paradigm, despite using a relatively small amount of robotic data in both the pre-training and post-training stages. We attribute this success to the unique training recipe. A key insight is that pre-training the VLM on large-scale human video enables it to learn domain-aligned visual representations for downstream manipulation tasks, while avoiding the

Pre-Training		Post-Training	Real-Time	MM-DiT	Naive DiT	Right-Arm	Left-Arm	Dual-Arm	Overall
EgoDex	HE	(On HE)	Chunking	Action Head	Action Head	Pick-n-Place	Pick-n-Place	Carry	Success Rate
✗	✗	✗	✗	✗	✓	1/10	1/10	1/10	0/10
✗	✗	✗	✗	✓	✗	9/10	2/10	3/10	2/10
✓	✗	✗	✗	✓	✗	8/10	6/10	6/10	6/10
✓	✓	✗	✗	✓	✗	8/10	8/10	9/10	8/10
✓	✓	✓	✗	✓	✗	9/10	9/10	10/10	9/10
✓	✓	✓	✓	✓	✗	9/10	9/10	9/10	9/10

TABLE I: **Ablation Studies**. We study the effects of pre-training, post-training, and real-time chunking on a dual-arm long-horizon task which consists of three steps: right-arm pick and place, left-arm pick-and-place and dual-arm lift.

hazardous and difficult co-training of two fundamentally different distributions. With language and visual representations extracted from the pre-trained VLM, we further post-train only the action expert in the joint space using high-quality real-robot data, enabling it to learn a strong prior for embodied control. More detailed results, including per-subtask progress and policy rollout videos, are provided in the **Supplemental Material**.

C. Ablation Studies

Due to limited compute and time, we perform our ablation study using a single real-world task: *pick toys into a box and lift it*. This task consists of three sub-tasks: (1) picking up a toy dumpling with the right arm and placing it into the box; (2) picking up a toy hippopotamus with the left arm and placing it into the box; and (3) carrying the box with both arms. This task consists of multiple execution stages and requires the policy to handle single-arm pick-and-place and dual-arm coordination.

a) The Role of Pre-Training and Post-Training: First, we study how the original Qwen3-VL VLM pre-trained on text-generation tasks performs in our settings. As shown in Table I, freezing the pre-trained Qwen3-VL backbone and fine-tuning only the action head yields the poorest performance, achieving an overall success rate of only 0.2. This result highlights the importance of pre-training the VLM backbone on human data to learn how to generate action tokens. After pre-training on EgoDex for task-space next-action prediction, the model achieves a substantial performance improvement. Notably, even though the VLM backbone is trained to predict a different action representation than that used by the downstream action head, supervising it with next-step 48-DoF actions still enables the model to learn meaningful visual representations for robotic tasks. These findings suggest an effective pathway for learning from large-scale human video data while avoiding the inference latency associated with fully autoregressive VLM action generation. With post-training of the action expert on high-quality robot data, overall performance is further improved.

b) MM-DiT versus Naive DiT: We also ablate the effectiveness of the proposed MM-DiT action head by comparing it with a naive DiT for action prediction. The results show that MM-DiT consistently outperforms the DiT variant. This improvement can be attributed to MM-DiT's dual-modulation design and its joint attention mechanism, which integrates VL features from the VLM backbone with Action (A) branch representations. Our analysis suggests that naive DiT, originally designed for text-conditioned image generation, provides weaker conditioning when applied to VL-guided action prediction. Additional ablation studies on the action expert are provided in the **Supplementary Material**.

c) Real-Time Chunking Behaviors: VLAs typically suffer from slow inference due to their large model size. When receiving a new query to generate actions, inference can take more than 200 ms, during which the humanoid robot must pause while waiting for the actions to become available, introducing jitter and unstable behavior in whole-body control tasks. One solution is test-time real-time chunking [6]. It employs inference-time gradient guidance to the flow-based action generation to steer the future actions to be consistent with past ones, therefore achieving smooth execution of the joint commands. However, we found that our model can not be guided at test time stably; as a result, we implemented training-time real-time chunking [7]. We observed that real-time chunking mitigates physical collisions during policy execution and increases policy rollout throughput without performance degradation.

D. Generalization Experiments

(1) Cross embodiment. Our training data spans at least two embodiments. We further include a cross-embodiment experiment on a new platform, Dexmate Vega. We finetuned the pick-and-place task on Vega, as shown in Fig. 8, with a 100% success rate in 10 trials. Even though the Dexmate Vega has a different morphology and different camera configuration from the Unitree G1, our model facilitates quick adaptation to a new embodiment.

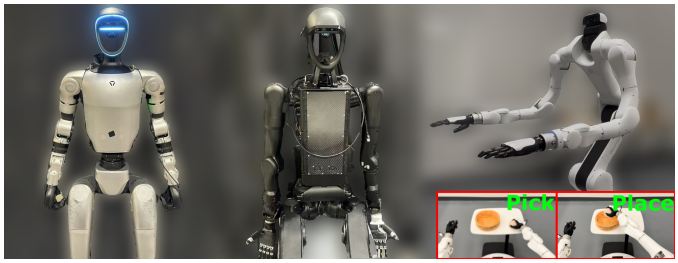


Fig. 8: Robot platforms used in our data or experiment. From left to right: Unitree G1, Unitree H1, Dexmate Vega.

(2) Zero-shot and few-shot experiments. We include zero-shot experiments in Table II and Fig. 9. This demonstrates that our model generalizes to novel objects, backgrounds, and poses. This echoes the few-shot experiments included in the supplementary video (pantry tasks at the beginning), where some tasks are trained using as few as 10 trajectories, e.g., pushing the refrigerator door and moving the chair. We have found that collecting more diverse training data in terms of target object pose, object classes, and background would further improve generalizability.

Zero-shot Generalization Settings	**Success Rate**
Background Generalization (Wooden + Black)	2/10
Object Generalization (Toy + Box)	4/10
Spatial Generalization (Position + Orientation)	2/10

TABLE II: Zero-shot generalization, long-horizon tasks.

Fig. 9: Zero-shot generalization setups.

(3) Ablation on long-horizon real tasks. To further evaluate the pre-training recipe, we include more ablation studies on more difficult long-horizon tasks in Fig. 10. Herein, "Scratch" indicates the VLM backbone is initialized from the pre-trained Qwen, and the action expert is initialized from scratch. "EgoDex-only" indicates the VLM is pre-trained on EgoDex data only, while the "EgoDex+HE" indicates the action expert is further fine-tuned with HE data. As can be observed from the figure, all the variants are under-performed with the full Ψ_0 baseline. Once again proves that both pre-training on human egocentric data and post-training on real robot data contribute to the overall performance of Ψ_0.

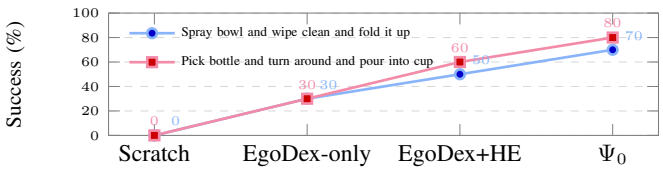


Fig. 10: Ablations on additional long-horizon tasks.

V. Conclusion

We introduce Ψ_0 , an open foundation model accompanied by a complete open-source suite for teleoperation, learning infrastructure, and deployment. Through extensive experiments, our results suggest that scaling humanoid learning requires scaling the right data in the right way. In contrast to blindly increasing the volume of teleoperation data at substantial cost, we leverage affordable, high-quality egocentric videos to learn human motion priors and human-object interaction knowledge. Our work further introduces several novel and empirically validated techniques that significantly improve the effectiveness of humanoid VLAs, including efficient whole-body dexterous teleoperation, MM-DiT-based action experts, and real-time control at deployment. Together, our training recipe and model architecture achieve state-of-the-art performance on challenging, complex, long-horizon tasks, while relying on substantially less real-world robotic data. We hope this work can serve as a foundation for humanoid learning, accelerating the development of humanoids capable of assisting with everyday tasks.

Limitation. Due to compute and time constraints, we are unable to further scale training to larger collections of human videos and real-world robotic data, which we leave for future work. Another limitation stems from the hardware platform, whose payload capacity constrains the execution of potentially more capable manipulation behaviors.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of The Conference on Robot Learning*, Proceedings of Machine Learning Research, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] 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.

[4] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[5] 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.

[6] Kevin Black, Manuel Y Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. *arXiv preprint arXiv:2506.07339*, 2025.

[7] Kevin Black, Allen Z Ren, Michael Equi, and Sergey Levine. Training-time action conditioning for efficient real-time chunking. *arXiv preprint arXiv:2512.05964*, 2025.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[9] 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.

[10] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

[11] Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, Yang Tian, Bin Wang, Bolun Wang, Fangjing Wang, Hanqing Wang, Tai Wang, Ziqin Wang, Xueyuan Wei, Chao Wu, Shuai Yang, Jinhui Ye, Junqiu Yu, Jia Zeng, Jingjing Zhang, Jinyu Zhang, Shi Zhang, Feng Zheng, Bowen Zhou, and Yangkun Zhu. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy, 2025. URL https://arxiv.org/abs/2510.13778.

[12] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

[13] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2024. URL https://arxiv.org/abs/2303.04137.

[14] Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[16] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

[17] Haoran Geng, Songlin Wei, Congyue Deng, Bokui Shen, He Wang, and Leonidas Guibas. Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions. *arXiv preprint arXiv:2312.01307*, 2, 2023.

[18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[20] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709.

[21] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[22] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025.

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

[24] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[25] Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Rizhao Qiu, and Xiaolong Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. *Robotics: Science and Systems 2025*, 2025.

[26] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon

tasks, 2025.

[27] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[28] Minghuan Liu, Zixuan Chen, Xuxin Cheng, Yandong Ji, Ri-Zhao Qiu, Ruihan Yang, and Xiaolong Wang. Visual whole-body control for legged loco-manipulation. *arXiv preprint arXiv:2403.16967*, 2024.

[29] 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 pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

[30] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[31] MANUS Technology Group. MANUS – High-Precision Data Gloves for Robotics, VR & Mocap. https://www.manus-meta.com/, 2024.

[32] Jiageng Mao, Siheng Zhao, Siqi Song, Chuye Hong, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, et al. Universal humanoid robot pose learning from internet human videos. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2025.

[33] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[34] PICO Immersive Pte. Ltd. PICO 4 Ultra: An All-New Mixed Reality Experience. https://www.picoxr.com/global/products/pico4-ultra, 2023.

[35] Haozhi Qi, Yen-Jen Wang, Toru Lin, Yi Brent, Yi Ma, Koushil Sreenath, and Jitendra Malik. Coordinated humanoid manipulation with choice policies. *arXiv:2512.25072*, 2025.

[36] Yiyang Shao, Xiaoyu Huang, Bike Zhang, Qiayuan Liao, Yuman Gao, Yufeng Chi, Zhongyu Li, Sophia Shao, and Koushil Sreenath. Langwbc: Language-directed humanoid whole-body control via end-to-end learning. *arXiv preprint arXiv:2504.21738*, 2025.

[37] Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025.

[38] Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

[39] Ruihan Yang, Qinxi 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.

[40] Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. Egomi: Learning active vision and whole-body manipulation from egocentric human demonstrations, 2025. URL https://arxiv.org/abs/2511.00153.

[41] Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025.

[42] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.

[43] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

[44] Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning, 2025. URL https://arxiv.org/abs/2510.05070.

[45] 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.

[46] Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharor, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.

[47] 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.