arXiv:2610.00438v1 [cs.RO] 30 Sep 2026

TOWARDS A GENERAL HUMANOID LOCO-MANIPULATION MODEL VIA EGOCENTRIC WHOLE-BODY HUMAN DATA PRETRAINING

Chongyang Xu[1,2]*, Zhao Wu[1], Jin Chen[1,3], Yiming Jiang[1,4], Jinhui Ye[5], Yuming Jiang[1], Shifeng Zhang[1], Ziliang Feng[2], Mu Xu[1], Yilun Chen[1†✉], Li Lu[2✉], Steven C.H. Hoi[1]

[1]Alibaba Group [2]Sichuan University [3]Shanghai Innovation Institute
[4]Beihang University [5]Hong Kong University of Science and Technology

[✉]Corresponding authors. [†]Project leader.

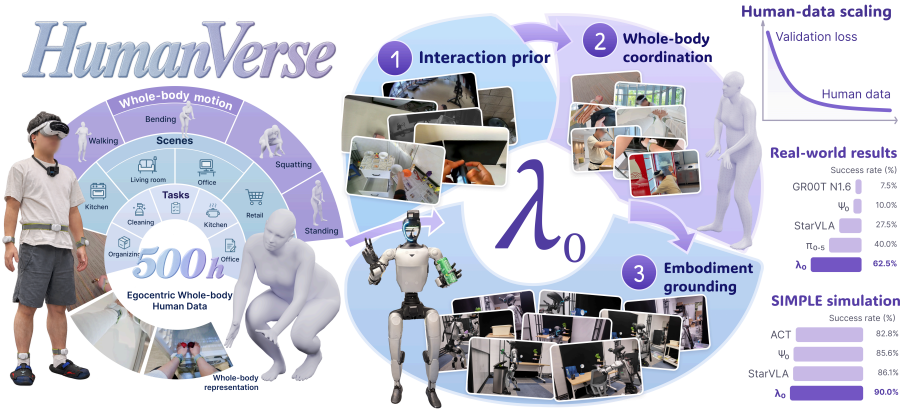


Figure 1: Introducing $\boldsymbol{\lambda_0}$ with **HUMANVERSE-500**. We introduce a robot-free egocentric human-data collection paradigm and use it to collect HUMANVERSE-500, a 500-hour dataset of whole-body human loco-manipulation. Then, we develop λ_0, a unified whole-body humanoid vision–language–action policy, through a three-stage recipe combining large-scale Internet manipulation data, HUMANVERSE-500, and embodiment-specific data. Finally, λ_0 achieves state-of-the-art performance and strong generalization across simulated and real-world loco-manipulation tasks.

ABSTRACT

Humanoid whole-body manipulation has advanced rapidly, enabling policies to coordinate locomotion, posture, bimanual interaction, and dexterous hand movements. Meanwhile, egocentric human videos provide diverse examples of everyday interactions across objects and scenes, offering scalable supervision without robot operation. However, existing supervision from these videos provides limited coverage of whole-body movement and coordination with hand–object interaction, while obtaining such supervision through humanoid teleoperation is also costly and difficult to scale. We therefore explore how human experience can support scalable learning of humanoid loco-manipulation. To support this study, we introduce HUMANVERSE-500, a 500-hour dataset of diverse human loco-manipulation behaviors in open-world environments, collected with a lightweight wearable system that synchronizes egocentric video with body and hand motion. Building on this dataset, we develop λ_0, a whole-body humanoid vision–language–action policy, through three-stage training that first learns interaction from diverse egocentric datasets, then coordinates body and hand motion using HUMANVERSE-500, and finally adapts the policy to downstream tasks and robot embodiments. Across these stages, λ_0 learns a shared representation space for human experience transfer, while domain-specific interfaces handle differences between human and robot states and actions. We evaluate λ_0 on SIMPLE and 4 real-world loco-manipulation tasks, achieving state-of-the-art performance, and further analyze its scaling behavior, generalization, and training-stage contributions to understand how human data support downstream whole-body humanoid control. We will release our code, models, and data to support further research.

*This work was done during Chongyang Xu's internship at Alibaba Token Hub.

1 INTRODUCTION

Humanoid whole-body manipulation has advanced rapidly, enabling robots to coordinate locomotion, posture, and dexterous manipulation within a unified policy (Jiang et al., 2026; Shi et al., 2026; Wei et al., 2026b; Hu et al., 2026; Figure AI, 2026a;b). Learning these coordinated behaviors requires demonstrations that capture both whole-body movement and interaction with objects. Such demonstrations can be collected through whole-body teleoperation, where human operators guide humanoids through coordinated movements and object interactions (Fu et al., 2025a; He et al., 2025; Luo et al., 2026b; GEAR Team et al., 2025; Li et al., 2026b). Collecting diverse demonstrations this way requires substantial robot time and human effort, making it difficult to scale across tasks and environments (Shi et al., 2026; Punamiya et al., 2026).

Egocentric human videos offer a scalable alternative, capturing diverse everyday interactions without requiring robot operation (Grauman et al., 2022; Kareer et al., 2025; Hoque et al., 2026; Punamiya et al., 2026; Figure, 2026). Recent methods recover wrist and hand trajectories from these videos, turning observed interactions into action supervision for robot learning (Li et al., 2026a; Yang et al., 2026; Wang et al., 2026). Pretraining on these action-labeled videos can improve downstream manipulation, supporting the use of human experience for scalable robot learning (Zheng et al., 2026b; Punamiya et al., 2026). However, wrist and hand trajectories alone do not fully capture the whole-body coordination required for humanoid loco-manipulation. A humanoid must approach an object, adjust its posture to reach it, and maintain its grasp while carrying it elsewhere (Shi et al., 2026; Zhao et al., 2026a). Learning these behaviors requires supervision that preserves the temporal coordination between body movement and hand–object interaction. We therefore ask how diverse human experience can provide scalable whole-body supervision for humanoid loco-manipulation.

To address this question, we introduce HUMANVERSE-500, a 500-hour dataset of human loco-manipulation collected across diverse activities in open-world environments. We use a lightweight wearable system to capture egocentric video and body motion, then reconstruct hand motion from video to provide synchronized whole-body supervision. Building on HUMANVERSE-500, we develop λ_0, a whole-body humanoid vision–language–action policy that transfers human experience to robot control (Fig. 1). We train the policy in three stages to learn interaction from diverse egocentric datasets, coordinate body and hand motion using HUMANVERSE-500, and adapt to downstream tasks and robot embodiments. A shared representation space enables transfer across stages, while domain-specific interfaces handle differences between human and robot states and actions.

We evaluate λ_0 on SIMPLE and four real-world loco-manipulation tasks, improving success rates over the strongest evaluated baselines by 3.9 and 22.5 percentage points, respectively. Across the evaluated data scales, more whole-body human data reduces validation loss and improves real-world task progress. Ablations further identify whole-body mid-training as the larger contributor to real-world success among the two human-data stages, while human demonstrations improve progress on robot-unseen objects. These results support scalable whole-body human data as a source of supervision for humanoid loco-manipulation. Our contributions are summarised as follows:

- We introduce HUMANVERSE-500, a 500-hour dataset of open-world human loco-manipulation with synchronized egocentric video, body motion, and hand motion.
- We develop λ_0, a whole-body humanoid vision–language–action policy that transfers human experience through three-stage training in a shared human–robot representation space.
- We reveal a consistent scaling trend in whole-body human pretraining and demonstrate its benefits for downstream performance and generalization in humanoid loco-manipulation.

2 RELATED WORK

2.1 GENERALIST ROBOT POLICIES

Large-scale robot datasets bring together demonstrations across tasks and embodiments, supporting the development of generalist policies (Dasari et al., 2020; Jang et al., 2022; Walke et al., 2023; Open X-Embodiment Collaboration, 2024; Brohan et al., 2023; Zitkovich et al., 2023; Khazatsky et al., 2024; Bu et al., 2025; Zhao et al., 2026b). These policies learn from diverse demonstrations within a shared model (Ghosh et al., 2024; Kim et al., 2025; Chi et al., 2025), motivating more scal-

able ways to collect training data. Advances in teleoperation and portable demonstration interfaces reduce collection effort (Dass et al., 2023; Qin et al., 2023; Zhao et al., 2023; Fu et al., 2025b; Chi et al., 2024). However, robot-based collection remains tied to hardware availability and repeated environment resets, while portable interfaces capture motion within their supported action spaces.

2.2 Learning from Human Videos

Human video datasets capture diverse activities, objects, and environments, providing rich experience for robot learning (Grauman et al., 2022; Damen et al., 2022; Goyal et al., 2017; Grauman et al., 2024; Wang et al., 2023b; DeTone et al., 2026; Lightwheel AI, 2026). Learning from it requires connecting observations to actions. Temporal and geometric annotations describe when interactions occur and how hands and objects move (Liu et al., 2022; Wang et al., 2025; Sener et al., 2022; Ragusa et al., 2021; Huang et al., 2024; Banerjee et al., 2025), while motion tracking and reconstruction provide human trajectories for action learning (Hoque et al., 2026; Punamiya et al., 2026; Li et al., 2026a; Zheng et al., 2026b; Wang et al., 2026). Complementary approaches learn visual representations from human videos (Nair et al., 2023; Sontakke et al., 2023; Kumar et al., 2020) or connect human demonstrations to robot control through imitation and alignment (Mandlekar et al., 2022; Shao et al., 2020; Bahl et al., 2022; Wang et al., 2023a; Bharadhwaj et al., 2023; Kareer et al., 2025; Lepert et al., 2025; Yang et al., 2026). Human videos offer scalable supervision for manipulation, but capturing whole-body coordination remains challenging for humanoid learning.

2.3 Humanoid Whole-Body Control

Humanoid whole-body control coordinates locomotion, posture, and manipulation to support complex physical interactions (Fu et al., 2025a; He et al., 2025; Luo et al., 2026b). Building on this control foundation, humanoid policies learn task-level behaviors from human experience through latent video actions, geometric pretraining, or joint training on human and robot demonstrations (Jiang et al., 2026; Wei et al., 2026b; Shi et al., 2026; Hu et al., 2026; Bjorck et al., 2025; Luo et al., 2026a). Predictive models and integrated whole-body systems offer complementary ways to connect perception, motion, and control (Li et al., 2026b; Symbiosis Robotics, 2026; Figure AI, 2026a; Parada, 2026; Zheng et al., 2026a; Ye et al., 2026; Gao et al., 2026). A remaining challenge is to turn diverse human experience into scalable supervision for humanoid whole-body control.

3 HumanVerse-500: Egocentric Whole-Body Human Data

As shown in Fig. 2, HumanVerse-500 contains 500 hours of egocentric human loco-manipulation, comprising 32,993 episodes and 87 million pose frames. The dataset covers 827 task types across diverse objects and scenes, pairing first-person video with synchronized body and hand motion. We choose the GoPro HERO13 for convenient human data collection and equip our

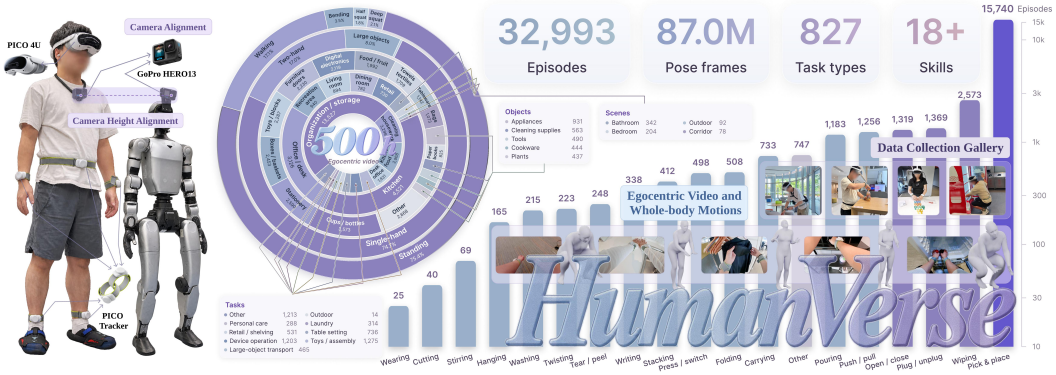


Figure 2: **HumanVerse-500.** The wearable capture setup, task and motion distributions, and paired egocentric video and whole-body motions illustrate the dataset's scale and diversity.

downstream robot, Unitree G1, with the same camera on its head. We align the human camera height with the robot camera using a neck mount, while a PICO 4 Ultra headset and body trackers capture body motion alongside the video. We then reconstruct hand motion from the same video.

3.1 DATA COVERAGE

HUMANVERSE-500 captures everyday activities across indoor and outdoor environments, including cleaning, cooking, personal care, laundry, table setting, and large-object transport. These activities involve objects of different sizes and require varied combinations of locomotion, posture changes, and hand–object interaction. For example, interacting with objects at different heights involves bending or reaching, while transporting objects couples walking with sustained grasping. Continuous recordings capture transitions between interactions and body–hand coordination throughout each task. Figure 2 summarizes the task, scene, object, and motion distributions. Figure 3 shows representative egocentric frames spanning household, office, and object-handling activities.



Figure 3: **Egocentric task diversity in HUMANVERSE-500.** Sixty egocentric frames span carrying, cleaning, plant care, office tasks, and object handling.

3.2 WHOLE-BODY MOTION RECONSTRUCTION

We construct whole-body motion by combining wearable-tracked body kinematics with hand motion reconstructed from video. The PICO system (PICO, n.d.) provides timestamped 24-joint body motion, while Dyn-HaMR (Yu et al., 2025) reconstructs 21 joints per hand using camera motion estimated by ViPE (Huang et al., 2025). Timestamps and coordinate transforms align body and hand trajectories with video frames, while an episode-fixed frame preserves hand displacement. Validity masks retain usable body supervision when hands are missing or occluded. Quality checks cover timestamp consistency, finite values, hand reprojections, and synchronized motion overlays. Figure 4 shows synchronized body motion, hand motion, and egocentric RGB frames.

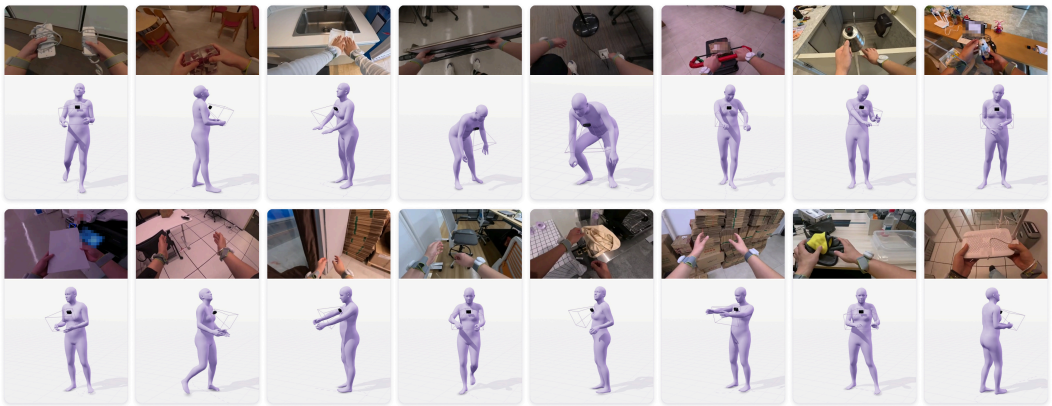


Figure 4: **Whole-body Human Data.** Egocentric RGB frames, body motion, and hand motion.

3.3 HUMAN-TO-ROBOT RETARGETING

We convert the captured human motion into training data for the Unitree G1 policy using SONIC (Luo et al., 2026b). Its frozen encoder produces body-motion tokens, and its controller tracks human motion in MuJoCo to obtain G1 states. We retarget the reconstructed hand motion to the bilateral BrainCo Revo 2 dexterous hands (BrainCo, n.d.) to obtain robot hand commands. The paired records share egocentric video and timestamps while retaining distinct body and hand representations. Figure 5 shows the corresponding retargeting.



Figure 5: Whole-body motion retargeting to G1 with dexterous hands.

4 λ_0: LEARNING HUMANOID CONTROL FROM HUMAN EXPERIENCE

We develop λ_0 with a shared vision–language–action architecture and domain-specific interfaces to transfer human experience to humanoid control. Training proceeds through interaction pre-training, whole-body mid-training, and embodiment-specific post-training (Fig. 6).

4.1 POLICY ARCHITECTURE AND ACTION INTERFACES

Given an egocentric image $\mathbf{x}$, a language instruction ℓ, and the current state $\mathbf{s}^q$, λ_0 predicts a 50-step action chunk in the active action space q. The backbone and action expert are shared across stages, while domain-specific action projections handle differences in target dimensions and semantics.

Shared transfer core. λ_0 consists of a Qwen3.5-2B (Qwen Team, 2026) vision–language backbone V_ϕ and a generative action expert F_ψ. The current state is quantized into 256 bins and appended to the language instruction. The backbone encodes this state-aware instruction together with the image $\mathbf{x}$, then supplies layer-wise features to the action expert. For each action space q, an input encoder E_q maps the noisy action and noise level into action tokens, while an output decoder D_q maps the shared features to a velocity prediction in that action space:

$$\widehat{\mathbf{v}}^q = D_q(F_\psi(V_\phi(\mathbf{x}, [\ell; Q(\mathbf{s}^q)]), E_q(\widetilde{\mathbf{a}}^q, \boldsymbol{\sigma}), e)) \,. \tag{1}$$

Here e is the embodiment identifier. The backbone V_ϕ and action expert F_ψ are shared across stages, while separate input encoders E_q and output decoders D_q handle the corresponding action spaces.

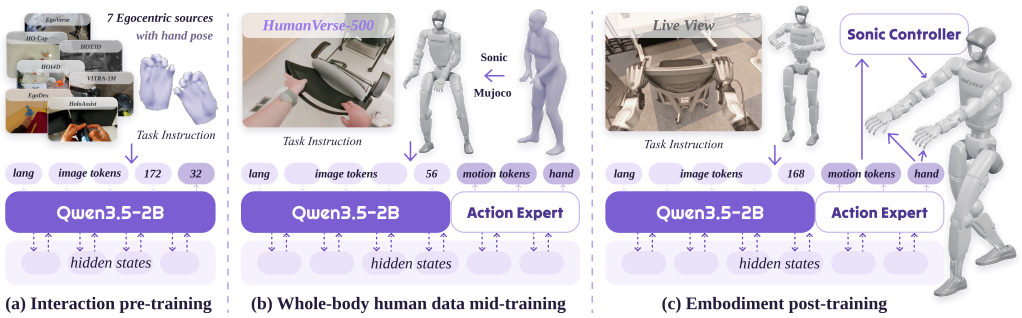
Action interfaces. The egocentric interface predicts hand motion conditioned on video, language, and bilateral hand states. The robot-compatible view of mid-training data uses the same state–action format as downstream G1 post-training data, while the two stages use distinct training domains.

4.2 THREE-STAGE TRAINING

Our training pipeline uses three data sources: public egocentric hand–object datasets ($\mathcal{D}_E$) for pretraining on egocentric videos and hand poses, HUMANVERSE-500 ($\mathcal{D}_W$) for mid-training on whole-body human motion, and robot demonstrations ($\mathcal{D}_R$) for robot post-training.

All stages use the masked velocity-field loss $\mathcal{L}^q$ defined in Sec. 4.3, with different data sources and active interfaces. Let $\boldsymbol{\theta}^{(k)}$ denote the policy parameters learned in Stage k. Stages II and III initialize from the preceding EMA checkpoint while changing the supervision and active interfaces.

Stage I: egocentric hand-motion pretraining. We initialize the backbone from a vision–language checkpoint and train it with the action expert and egocentric hand interface (Fig. 6(a)). Training uses EgoDex, EgoVerse, HO-Cap, HOI4D, HoloAssist, HOT3D, and VITRA-1M (Hoque et al., 2026; Punamiya et al., 2026; Wang et al., 2025; Liu et al., 2022; Wang et al., 2023b; Banerjee et al., 2025; Li et al., 2026a), with hand-pose supervision from either native ground-truth annotations or pseudo-labels generated by open-source methods. Let $\mathcal{S}_E$ denote these seven data sources. We sample datasets in proportion to the square root of their usable frame counts N_d to increase the



(a) Interaction pre-training (b) Whole-body human data mid-training (c) Embodiment post-training

Figure 6: **Three-stage training of λ_0.** The policy learns interaction from egocentric data, whole-body coordination from HUMANVERSE-500, and downstream control from real-robot data.

relative contribution of smaller sources:

$$p_1(d) = \frac{\sqrt{N_d}}{\sum_{d' \in \mathcal{S}_E} \sqrt{N_{d'}}}, \qquad \boldsymbol{\theta}^{(1)} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{d \sim p_1,\, \tau \sim \mathcal{D}_d} \left[\mathcal{L}^E(\tau; \boldsymbol{\theta}) \right]. \tag{2}$$

Stage I predicts a 138-dimensional bimanual action at each time step, with each hand represented by 3D relative wrist translation, 6D relative wrist rotation, and 20×3 local hand-keypoint coordinates.

Stage II: whole-body human-motion mid-training. During mid-training, HUMANVERSE-500 demonstrations are converted into the G1 state–action format (Fig. 6(b)). Each action comprises a 64-dimensional SONIC motion latent and two 6-dimensional Revo2 hand commands, totaling 76 dimensions. Human-derived and native robot data use distinct domain identifiers and separate input and output projections to accommodate scale differences. Denoting the conversion by Φ_R and the human-domain loss by $\mathcal{L}^H$, we optimize

$$\boldsymbol{\theta}^{(2)} = \arg \min_{\boldsymbol{\theta} \leftarrow \mathrm{EMA}(\boldsymbol{\theta}^{(1)})} \mathbb{E}_{\tau \sim \Phi_R(\mathcal{D}_W)} [\mathcal{L}^H(\tau; \boldsymbol{\theta})]. \tag{3}$$

We update the backbone, action expert, and human-data projections while keeping the Stage-I hand head frozen. This stage uses only human data. Robot demonstrations are introduced in Stage III.

Stage III: robot post-training. We post-train the policy on robot demonstrations to adapt the motion and interaction patterns learned in the preceding stages to downstream tasks and executable robot control (Fig. 6(c)). The post-training objective is

$$\boldsymbol{\theta}^{(3)} = \arg \min_{\boldsymbol{\theta} \leftarrow \mathrm{EMA}(\boldsymbol{\theta}^{(2)})} \mathbb{E}_{\tau \sim \mathcal{D}_R} [\mathcal{L}^R(\tau; \boldsymbol{\theta})]. \tag{4}$$

The shared model, embodiment conditioning, and active interfaces are optimized jointly.

4.3 TRAINING OBJECTIVE AND DEPLOYMENT

Masked velocity-field objective. We interpolate between clean actions and Gaussian noise and train the model to predict the velocity target $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{a}$. Let $\hat{\mathbf{v}}$ denote the predicted velocity and M^q the set of valid, non-prefix sample–time–feature indices for action space q, excluding missing annotations. We minimize the mean squared error over these entries:

$$\mathcal{L}^q = \mathbb{E}\left[\frac{1}{|M^q|} \sum_{(i,h,j) \in M^q} \left(\hat{v}^q_{i,h,j} - v^q_{i,h,j} \right)^2 \right]. \tag{5}$$

At inference, we initialize the action chunk from Gaussian noise and denoise it using 10 Euler steps from noise level one to zero. Any provided action prefix remains fixed throughout denoising.

Deployment. At deployment, λ_0 receives G1's egocentric observation, language instruction, and proprioceptive state, using the robot-domain identifier and its corresponding input and output projections (Fig. 6(c)). The predicted 64-dimensional motion latent is passed to the SONIC controller for whole-body execution, while the two 6-dimensional hand commands are sent to the Revo2 hands.

5 EXPERIMENTS

We evaluate whether the three-stage recipe turns human data into improved humanoid control and object transfer. Specifically, we ask: **Q1.** *How well does the model perform on humanoid loco-manipulation?* (Sec. 5.2); **Q2.** *How do human-data scale and the training recipe affect learning?* (Sec. 5.3); **Q3.** *Does human-data scaling improve downstream transfer?* (Sec. 5.4). Accordingly, we evaluate held-out learning dynamics, task performance, and unseen-object generalization.

5.1 EVALUATION SETUP AND METRICS

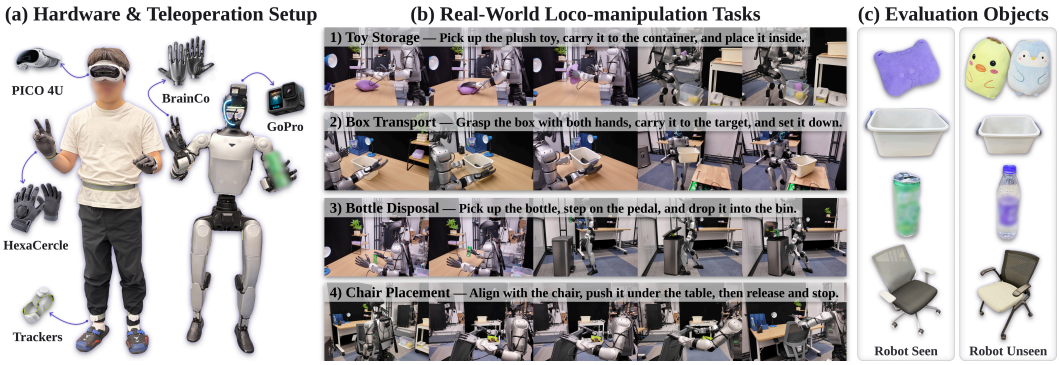


Figure 7: **Real-world data collection and evaluation.** (a) Egocentric teleoperation collects whole-body demonstrations. (b) Four loco-manipulation tasks evaluate task success. (c) Robot-seen and robot-unseen objects are used for standard task and object-generalization evaluation, respectively.

Hardware. As shown in Fig. 7(a), a PICO 4U with five trackers captures body motion, which SONIC converts into Unitree G1 control signals. We use HexaCercle M11 gloves to capture hand poses robustly under occlusion and map the captured finger motion to control the BrainCo Revo 2 hands. A GoPro mounted on the G1's head provides egocentric RGB, while the G1's computer records synchronized whole-body expert demonstrations.

Evaluation. We evaluate six core SIMPLE tasks (Wei et al., 2026a) at Levels 0–2 with ten rollouts per task–level pair, and four real-world tasks shown in Fig. 7(b), with ten frozen initial configurations per task and policy. These fixed configurations support consistent comparisons across policies by controlling variation in the initial conditions. We report task success rate and task progress, capturing both full task completion and partial execution of multi-stage behaviors. Progress is computed as the sum of fixed milestone scores, normalized to 100 per task and averaged equally across tasks. Standard trials use robot-seen objects in Fig. 7(c). Generalization uses robot-unseen objects appearing only in human demonstrations.

Baselines. Simulation baselines come from the public SIMPLE leaderboard (Wei et al., 2026a). Public entries retain their source protocols. We evaluate StarVLA and λ_0 ourselves. Real-world baselines include $\pi_{0.5}$ (Black et al., 2025), GR00T N1.6 (GEAR Team et al., 2025), Ψ_0 (Wei et al., 2026b), and StarVLA (StarVLA Community, 2026). All methods use identical initial configurations and scoring criteria for consistent comparisons.

5.2 Q1: HOW WELL DOES THE MODEL PERFORM ON HUMANOID LOCO-MANIPULATION?

Real-world evaluation. Table 1 reports the highest average success rate (62.5%) and weighted progress (80.4%) for λ_0, exceeding the strongest external baselines by 22.5 and 25.8 percentage points, respectively. It achieves the highest success count on all four tasks, indicating consistent gains across tasks, with the largest progress improvement on bottle disposal (51.0 points). Progress remains above 75% on every task, suggesting broad milestone coverage even where full completion remains challenging. On chair placement, it completes 6/10 trials versus StarVLA's 3/10 despite lower progress (77.0% versus 82.5%), illustrating that partial-task progress and full completion capture different aspects of performance and should be reported together for a fuller assessment.

Table 1: **Real-world task performance.** We report success rates over ten trials and weighted progress (%) for each task. Averages weight the four tasks equally.

Method	Bottle Disposal		Toy Storage		Box Transport		Chair Placement		Average	
$\pi_{0.5}$ (Black et al., 2025)	2/10	28.0	5/10	64.5	5/10	58.0	4/10	65.5	16/40	54.0
GR00T N1.6 (GEAR Team et al., 2025)	0/10	10.0	1/10	19.0	0/10	10.0	2/10	46.0	3/40	21.3
Ψ_0 (Wei et al., 2026b)	1/10	26.5	2/10	30.0	1/10	37.5	0/10	33.0	4/40	31.8
StarVLA (StarVLA Community, 2026)	0/10	10.0	5/10	55.0	3/10	71.0	3/10	**82.5**	11/40	54.6
λ_0 **(Ours)**	**5/10**	**79.0**	**7/10**	**81.0**	**7/10**	**84.5**	**6/10**	77.0	**25/40**	**80.4**

Table 2: **SIMPLE task performance.** Success rates (%) at Levels 0–2 with ten trials per task–level pair. Average is computed per level. Overall averages all 18 task–level pairs.

Model	XMovePick	BendPick	Handover	Mobile P&P	Grasp	XMoveBendPick	Average	Overall
H-RDT (Bi et al., 2026)	0/0/20	0/0/10	0/10/0	0/0/0	0/0/0	0/0/0	0.0/1.7/5.0	2.2
InternVLA-M1 (Chen et al., 2025)	0/0/0	50/50/0	0/0/0	0/0/0	0/0/0	30/50/70	13.3/16.7/11.7	13.9
EgoVLA (Yang et al., 2026)	0/10/20	70/50/80	0/40/30	0/0/0	100/100/70	30/50/40	33.3/41.7/40.0	38.3
Diffusion Policy (Chi et al., 2025)	30/30/20	100/80/60	30/20/40	40/0/0	80/90/80	0/0/0	46.7/36.7/33.3	38.9
GR00T N1.6 (GEAR Team et al., 2025)	100/100/70	70/70/60	10/30/30	0/0/0	90/90/70	40/40/10	51.7/55.0/40.0	48.9
$\pi_{0.5}$ (Black et al., 2025)	70/50/10	100/100/80	50/40/50	30/30/30	100/100/80	0/0/0	58.3/53.3/41.7	51.1
DreamZero (Ye et al., 2026)	100/100/100	90/90/80	70/80/90	50/30/30	90/100/70	0/0/10	66.7/66.7/63.3	65.6
ACT (Zhao et al., 2023)	100/100/50	100/90/90	70/70/100	50/50/50	100/100/80	90/100/100	85.0/85.0/78.3	82.8
Ψ_0 (Wei et al., 2026b)	100/100/60	100/100/100	70/70/100	70/50/60	100/100/80	100/90/90	**90.0**/85.0/81.7	85.6
StarVLA (StarVLA Community, 2026)	100/90/90	70/90/70	80/80/100	90/80/80	90/100/70	70/100/100	83.3/90.0/85.0	86.1
λ_0 **(Ours)**	100/100/80	100/100/100	90/100/100	90/70/70	100/90/80	60/100/90	**90.0/93.3/86.7**	**90.0**

Simulation evaluation. Table 2 reports 90.0% overall success for λ_0, exceeding StarVLA and Ψ_0 by 3.9 and 4.4 percentage points, respectively. Its success rates of 90.0%, 93.3%, and 86.7% at Levels 0–2 lead StarVLA at every level, showing that the advantage extends across evaluation levels. The margin over StarVLA narrows to 1.7 points at Level 2, indicating that the relative gain varies with the evaluation setting. It achieves 100% success on BendPick at all levels and Handover at Levels 1 and 2, while Mobile P&P remains below StarVLA at Levels 1 and 2 (70% versus 80%), indicating room for improvement in mobile manipulation. BendPick and Handover maintain strong performance across all three levels.

5.3 Q2: How Do Human-Data Scale and the Training Recipe Affect Learning?

Validation-loss scaling. To isolate human-data scale in Stage II, we train the same 2B model for 100k updates on nested 5%, 10%, 25%, 50%, and 100% subsets with a fixed seed and optimization schedule. Every 1k updates, we evaluate two fixed, session-disjoint holdouts: *clean* contains sessions with upper-quartile whole-body and bilateral-hand annotation validity, while *representative* is stratified by data source and annotation validity to reflect the full data distribution. Following EgoScale (Zheng et al., 2026b), we analyze training dynamics and minimum loss at each scale.

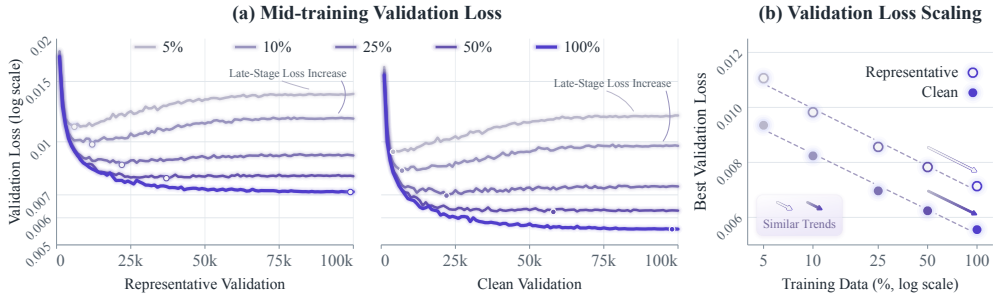


Figure 8: **Human-data scaling of validation loss.** (a) Training curves on representative and clean holdouts at five data fractions. (b) Minimum validation loss versus data scale with descriptive log-linear fits over the measured range. Minima are selected independently on each holdout.

Training dynamics. Fig. 8(a) shows later validation-loss minima at larger data scales. At 5% and 10%, representative loss reaches its minimum at 6k and 12k updates, then finishes 24.7% and 19.0% higher. On clean, the corresponding increases are 27.6% and 18.1%. At 50% and 100%, final

losses remain within 1.4% of the minima on representative and 0.9% on clean. At 50%, the minima occur at 37k/58k updates on representative/clean. At 100%, they occur at 99k/98k. The 25% run forms an intermediate regime: both holdouts reach their minima near 22k updates and finish only 6.4–6.5% higher. Larger datasets thus support longer training before validation loss rises, while smaller subsets benefit from earlier checkpoint selection. Comparing only final checkpoints would confound the effect of data scale with late-stage degradation.

Scaling trend. Fig. 8(b) compares minimum validation losses across data scales. From 5% to 100%, loss falls by 35.5% on representative and 40.6% on clean. Fitting $\widehat{L}_h(D) = a_h + b_h \ln D$ gives $R^2 = 0.9868/0.9901$ and per-doubling reductions of $0.000900/0.000877$ for representative/clean. Clean has lower loss throughout, while slope magnitudes differ by only 2.7%, indicating similar scaling responses despite their different annotation-validity distributions. These single-run fits describe the measured range without establishing cross-seed reproducibility or extrapolation.

Training-stage ablations under fixed robot supervision. Table 3 ablates the two human-data stages. SIMPLE success is 90.0% with the full recipe, 87.8% without Stage I, 82.3% without Stage II, and 79.4% without both. Real-world success/progress is 62.5%/80.4% for the full recipe, 50.0%/75.9% without Stage I, 22.5%/59.9% without Stage II, and 15.0%/54.0% without both. Removing Stage II reduces real-world success more than removing Stage I (40.0 versus 12.5 percentage points), highlighting the importance of whole-body mid-training beyond the interaction priors learned from egocentric hand data.

Table 3: **Training-stage ablations with fixed robot supervision.** Filled/open circles indicate included/removed data. Metrics report SIMPLE success and real-world averages across four tasks.

Variant	S1	S2	Robot	SIMPLE (%)				Real-world (%)	
				L0	L1	L2	Overall	Success	Progress
w/o human-data pretraining	○	○	●	83.3	83.3	71.7	79.4	15.0	54.0
w/o Stage II mid-training	●	○	●	83.6	86.7	76.7	82.3	22.5	59.9
w/o Stage I pretraining	○	●	●	88.0	91.7	83.7	87.8	50.0	75.9
λ_0 **(Full Recipe)**	●	●	●	**90.0**	**93.3**	**86.7**	**90.0**	**62.5**	**80.4**

5.4 Q3: Does Human-Data Scaling Improve Downstream Transfer?

Downstream task-progress scaling. Fig. 9 reports task progress of 59.9%, 64.2%, 67.3%, 72.1%, 76.4%, and 80.4% at 0%, 5%, 10%, 25%, 50%, and 100% Stage-II data, respectively. The endpoints match the no-Stage-II and full-recipe results in Table 3. We fix Stage-I initialization, model capacity, robot demonstrations, Stage-III update budget, and evaluation configurations. Nonzero fractions share Stage-II update budgets and checkpoint selection. Progress increases monotonically across these fractions, gaining 20.5 percentage points from no Stage II to the full corpus.

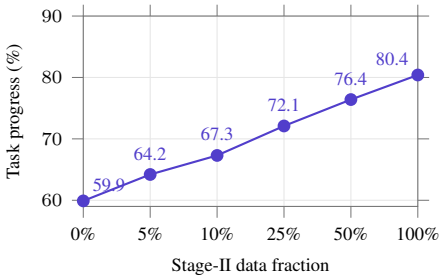


Figure 9: Average real-world task progress versus Stage-II human-data fraction.

Object generalization. We evaluate whether human-data training improves transfer to objects absent from robot demonstrations across the four real-world tasks. For each method, we train a policy using robot demonstrations of objects A and human demonstrations of A and B. *Seen* (🧍+🧍→📦) and *Human-guided* (🧍+🧍→📦) evaluate this same policy on A and B, respectively. *Zero-shot* (🧍→📦) evaluates a robot-only variant on B under matched initializations and scoring.

As shown in Fig. 10, λ_0 achieves the highest mean *Human-guided* progress at 55.8%, compared with 46.5% for the strongest baseline, $\pi_{0.5}$. Relative to its *Zero-shot* variant, λ_0 improves progress on all four tasks, increasing the average from 48.4% to 55.8%, a gain of 7.4 percentage points. These results support human-data training as a means of transferring manipulation skills to robot-unseen objects, although performance remains below the *Seen* setting.

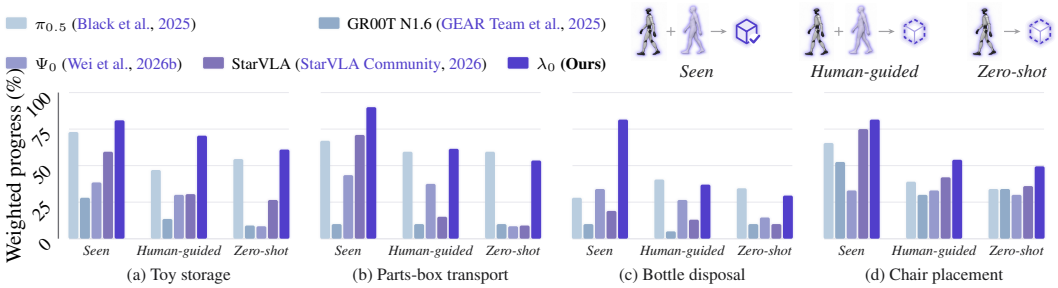


Figure 10: **Object-transfer progress.** Seen and Human-guided evaluate a joint policy trained on robot-seen object A and human-only data for object B. Zero-shot tests a robot-only policy on B.

6 Conclusion

We presented λ_0, a whole-body humanoid vision–language–action policy that transfers human experience through three-stage training. We introduced HumanVerse-500, a 500-hour dataset of diverse open-world human activities with synchronized egocentric video, body motion, and hand motion. Experiments on SIMPLE and four real-world tasks demonstrated improved loco-manipulation performance and transfer to robot-unseen objects. Controlled experiments revealed consistent benefits from scaling whole-body human data and highlighted the contribution of whole-body mid-training. These findings support scalable human activity as rich supervision for humanoid control.

Limitations and future work. Our evaluation focuses on G1, leaving transfer to full-sized humanoids untested. Despite camera alignment and motion retargeting, differences in body proportions, viewpoints, and feasible motions leave human–robot alignment imperfect. Future work will extend λ_0 to full-sized humanoids through improved motion retargeting and closer alignment between egocentric observations, whole-body trajectories, and physically executable robot actions.

References

Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In *Proceedings of Robotics: Science and Systems*, volume 18, 2022. doi: 10.15607/RSS.2022.XVIII.026.

Prithviraj Banerjee et al. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7071, 2025.

Homanga Bharadhwaj, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-RDT: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18135–18143, 2026. doi: 10.1609/aaai.v40i22.38875.

Johan Bjorck et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of the Conference on Robot Learning*, volume 305, pp. 17–40, 2025.

BrainCo. Revo 2: Overview. BrainCo Hand Documentation, n.d. URL `https://www.brainco-hz.com/docs/revolimb-hand-dev/en/revo2/parameters.html`. Accessed September 26, 2026.

Anthony Brohan et al. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, volume 19, 2023. doi: 10.15607/RSS.2023.XIX.025.

Qingwen Bu et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3549–3556, 2025. doi: 10.1109/iros60139.2025.11247088.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. InternVLA-M1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.045.

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. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668.

Dima Damen et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *Proceedings of the Conference on Robot Learning*, volume 100, pp. 885–897, 2020.

Shivin Dass, Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J. Lim, and Stefanos Nikolaidis. PATO: Policy assisted teleoperation for scalable robot data collection. In *Proceedings of Robotics: Science and Systems*, volume 19, 2023. doi: 10.15607/RSS.2023.XIX.013.

Daniel DeTone, Federica Bogo, Eric-Tuan Le, Duncan Frost, Julian Straub, Yawar Siddiqui, Yuting Ye, Jakob Engel, Richard Newcombe, and Lingni Ma. NymeriaPlus: Enriching nymeria dataset with additional annotations and data. *arXiv preprint arXiv:2603.18496*, 2026.

Figure. Introducing Index: Building the world's largest and most diverse physical dataset. Figure AI, August 2026. URL https://www.figure.ai/news/introducing-index. Accessed September 26, 2026.

Figure AI. Introducing Helix 02: Full-body autonomy. Figure AI, January 2026a. URL https://www.figure.ai/news/helix-02. Accessed September 26, 2026.

Figure AI. Helix 2.5: Zero-shot 30-home generalization. Figure AI, September 2026b. URL https://www.figure.ai/news/helix-2-5-zero-shot-30-home-generalization. Published September 17, 2026. Accessed September 26, 2026.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Proceedings of the Conference on Robot Learning*, volume 270, pp. 2828–2844, 2025a.

Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *Proceedings of the Conference on Robot Learning*, volume 270, pp. 4066–4083, 2025b.

Shenyuan Gao, William Liang, Kaiyuan Zheng, et al. DreamDojo: A generalist robot world model from large-scale human videos. In *Proceedings of the International Conference on Machine Learning*, 2026.

GEAR Team, Allison Azzolini, Johan Bjorck, et al. GR00T N1.6: An improved open foundation model for generalist humanoid robots. NVIDIA Research, 2025. URL https://research.nvidia.com/labs/gear/gr00t-n1_6/. Accessed September 26, 2026.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.090.

Raghav Goyal et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the International Conference on Computer Vision*, pp. 5842–5850, 2017.

Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kristen Grauman 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*, pp. 19383–19400, 2024.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Proceedings of the Conference on Robot Learning*, volume 270, pp. 1516–1540, 2025.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *Proceedings of the International Conference on Learning Representations*, 2026.

Yingdong Hu et al. OpenHLM: An empirical recipe for whole-body humanoid loco-manipulation. *arXiv preprint arXiv:2606.22174*, 2026.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. In *NVIDIA Research Whitepapers*, 2025. URL `https://research.nvidia.com/labs/toronto-ai/vipe/`.

Yifei Huang et al. EgoExoLearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22072–22086, 2024.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Proceedings of the Conference on Robot Learning*, volume 164, pp. 991–1002, 2022.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. In *Proceedings of the International Conference on Learning Representations*, 2026.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 13226–13233, 2025. doi: 10.1109/icra55743.2025.11127989.

Alexander Khazatsky et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.120.

Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning*, volume 270, pp. 2679–2713, 2025.

Ashish Kumar, Saurabh Gupta, and Jitendra Malik. Learning navigation subroutines from egocentric videos. In *Proceedings of the Conference on Robot Learning*, volume 100, pp. 617–626, 2020.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. In *Proceedings of the Conference on Robot Learning*, volume 305, pp. 4545–4565, 2025.

Qixiu Li et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026a.

Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, and Shanghang Zhang. ω-0: A latent predictive world action model for concurrent humanoid loco-manipulation. *arXiv preprint arXiv:2608.06375*, 2026b.

Lightwheel AI. EgoSuite-Open100K: 100,000 hours of egocentric human data for physical AI. Lightwheel AI, August 2026. URL https://huggingface.co/blog/LightwheelAI/egosuite-open100k. Accessed September 26, 2026.

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*, pp. 21013–21022, 2022.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-H0.5: Scaling human-centric robot learning for cross-embodiment generalization. In *Proceedings of the Conference on Robot Learning*, 2026a. To appear.

Zhengyi Luo et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026b. doi: 10.1126/scirobotics.aed4592.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Proceedings of the Conference on Robot Learning*, volume 164, pp. 1678–1690, 2022.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Proceedings of the Conference on Robot Learning*, volume 205, pp. 892–909, 2023.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 6892–6903, 2024. doi: 10.1109/icra57147.2024.10611477.

Carolina Parada. Gemini Robotics 2 brings whole body intelligence to robots. Google DeepMind, July 2026. URL https://deepmind.google/blog/gemini-robotics-2-brings-whole-body-intelligence-to-robots/. Accessed September 26, 2026.

PICO. Body tracking. PICO Developer Documentation, n.d. URL https://developer.picoxr.com/document/unity/body-tracking/. Accessed September 26, 2026.

Ryan Punamiya et al. EgoVerse: An egocentric human dataset for robot learning from around the world. In *Proceedings of Robotics: Science and Systems*, volume 22, 2026. doi: 10.15607/RSS.2026.XXII.092.

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 *Proceedings of Robotics: Science and Systems*, volume 19, 2023. doi: 10.15607/RSS.2023.XIX.015.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The MECCANO dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1569–1578, 2021.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2022.

Lin Shao, Toki Migimatsu, Qiang Zhang, Kaiyuan Yang, and Jeannette Bohg. Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems*, volume 16, 2020. doi: 10.15607/RSS.2020.XVI.082.

Modi Shi et al. Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. In *Proceedings of Robotics: Science and Systems*, volume 22, 2026. doi: 10.15607/RSS.2026.XXII.204.

Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. RoboCLIP: One demonstration is enough to learn robot policies. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55681–55693, 2023. doi: 10.52202/075280-2430.

StarVLA Community. StarVLA: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Symbiosis Robotics. Direct perception control model. Symbiosis Robotics, 2026. URL `https://symbiosis-robotics.com/research/dpc/en/`. Accessed September 26, 2026.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the Conference on Robot Learning*, volume 229, pp. 1723–1736, 2023.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-horizon imitation learning by watching human play. In *Proceedings of the Conference on Robot Learning*, volume 229, pp. 201–221, 2023a.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A capture system and dataset for 3D reconstruction and pose tracking of hand-object interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5683.

Xin Wang et al. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the International Conference on Computer Vision*, pp. 20270–20281, 2023b.

Ye Wang, Pei Lin, Xiong-Hui Chen, et al. Ego2Robot: Scalable robot data synthesis from egocentric human data. *arXiv preprint arXiv:2608.02580*, 2026.

Songlin Wei, Zhenhao Ni, Jie Liu, Zhenyu Zhao, Junjie Ye, Hongyi Jing, Junkai Xia, Xiawei Liu, Michael Leong, Liang Heng, Di Huang, and Yue Wang. SIMPLE: Simulation-based policy learning and evaluation for humanoid loco-manipulation. In *Proceedings of the Conference on Robot Learning*, 2026a. To appear.

Songlin Wei et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. In *Proceedings of Robotics: Science and Systems*, volume 22, 2026b. doi: 10.15607/RSS.2026.XXII.021.

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. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2026. To appear.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. In *Proceedings of the Conference on Robot Learning*, 2026. To appear.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HaMR: Recovering 4D interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2025. URL https://dyn-hamr.github.io/.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, volume 19, 2023. doi: 10.15607/RSS.2023.XIX.016.

Zehui Zhao, Yuxuan Zhao, Gaojing Zhang, Chenxi Liu, Maolin Zheng, and Wenzhao Lian. HALOMI: Learning humanoid loco-manipulation with active perception from human demonstrations. *arXiv preprint arXiv:2606.18772*, 2026a. URL https://arxiv.org/abs/2606.18772.

Zhenyu Zhao et al. Humanoid Everyday: A comprehensive robotic dataset for open-world humanoid manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026b.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. MotionWAM: Towards foundation world action models for real-time humanoid loco-manipulation. *arXiv preprint arXiv:2606.09215*, 2026a.

Ruijie Zheng et al. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. In *Proceedings of the Conference on Robot Learning*, 2026b. To appear.

Brianna Zitkovich et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the Conference on Robot Learning*, volume 229, pp. 2165–2183, 2023.