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

Anonymous authors
Paper under double-blind review



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.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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

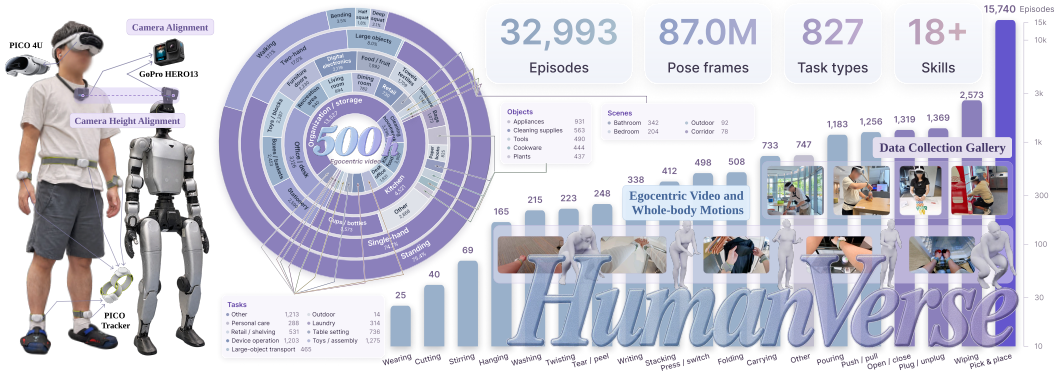

Figure 2: **HumanVerse-500.** The wearable capture setup, task and motion distributions, and paired egocentric video and whole-body motions illustrate the dataset's scale and diversity.

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, while Appendix Table 4 compares the dataset's capture setup and motion coverage with existing egocentric datasets.

3.2 WHOLE-BODY MOTION RECONSTRUCTION

We construct whole-body motion records 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. Separate 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. Appendix A details coordinate conventions, quality checks, and session-level data splits.

3.3 HUMAN-TO-ROBOT RETARGETING

We further convert the captured human motion into training data for the downstream 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 3 shows the corresponding retargeting, with visualization details in Appendix A.5.



Figure 3: 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. 4).

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

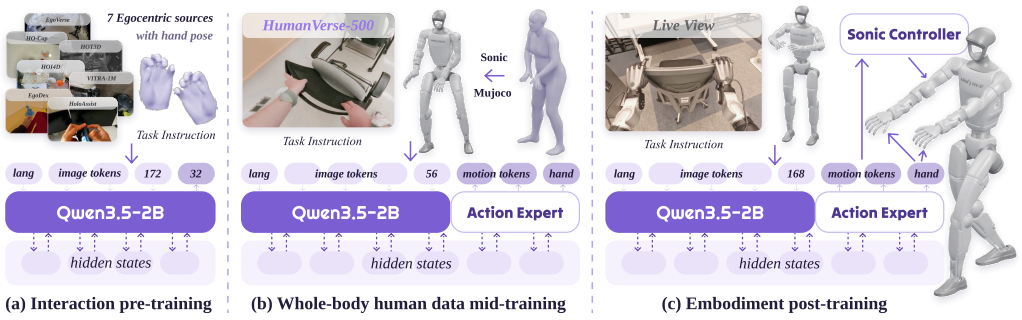

Figure 4: **Three-stage training of λ_0.** The policy learns interaction from egocentric data, whole-body coordination from HUMANVERSE-500, and downstream control from real-robot data.

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 pre-training 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. 4(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 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. 4(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. 4(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 $\widehat{\mathbf{v}}$ 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:

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

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. 4(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

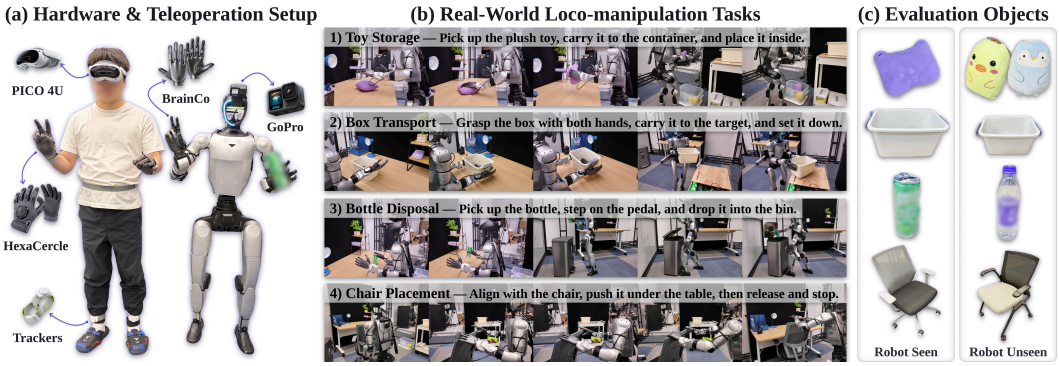

Figure 5: **Real-world data collection and evaluation.** (a) Egocentric teleoperation collects whole-body demonstrations. (b) Four loco-manipulation tasks evaluate task success. (c) Robot-seen and robot-unseen objects are used for standard task and object-generalization evaluation, respectively.

Hardware. As shown in Fig. 5(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. 5(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. 5(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**

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%),

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**

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.

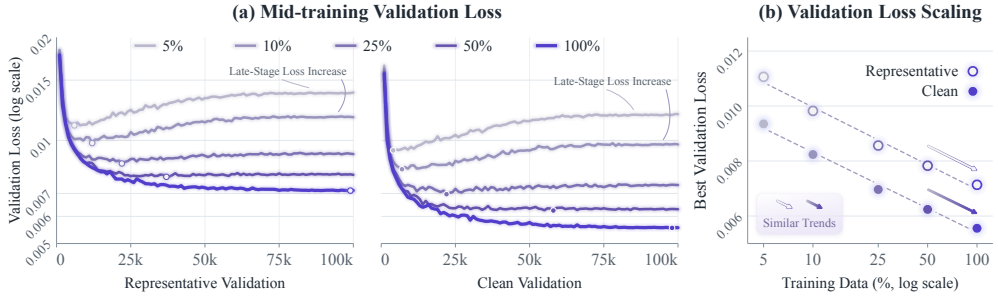

Figure 6: **Human-data scaling of validation loss.** (a) Training curves on representative and clean holdouts at five data fractions. (b) Minimum validation loss versus data scale with descriptive log-linear fits over the measured range. Minima are selected independently on each holdout.

Training dynamics. Fig. 6(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. 6(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. 7 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.

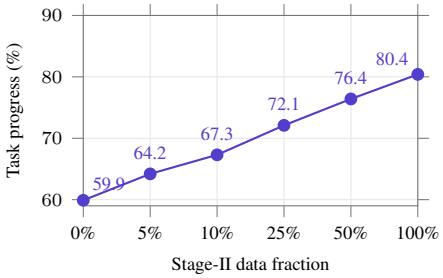

Figure 7: Average real-world task progress versus Stage-II human-data fraction.

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.

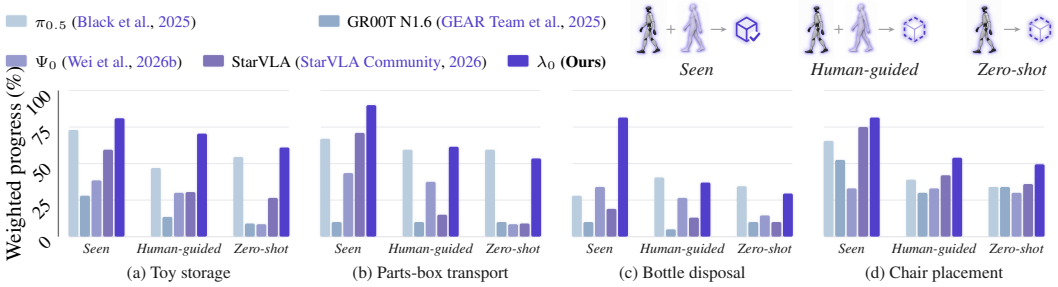

Figure 8: **Object-transfer progress.** Seen and Human-guided evaluate a joint policy trained on robot-seen object A and human-only data for object B. Zero-shot tests a robot-only policy on B.

As shown in Fig. 8, 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.

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.

AI USE STATEMENT

Generative AI tools were used to assist with experimental code development and debugging, data analysis, manuscript drafting and language editing, and figure preparation. All reported experimental results were obtained through actual model training and evaluation, not generated by AI tools. The authors reviewed and verified all AI-assisted content and take full responsibility for the methods, results, and conclusions.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe data collection, policy training, and experimental evaluation. Appendices A and B document data provenance, quality checks, action interfaces, and training and baseline implementation details. Appendix C specifies the data-scaling controls, while Appendices D and E describe the simulation and real-world evaluation protocols and scoring criteria. Appendices G and H detail the training-stage ablations and object-generalization protocol.

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.

APPENDIX

We provide a project website and an overview video to accompany the paper. The supplementary material below presents additional details on data collection, model training, and evaluation, together with qualitative visualizations and demonstration sequences.

A Dataset Scope, Manifest, and Provenance **17**

A.1 Qualitative Examples of Task Diversity 17

A.2 Synchronized Human and Camera Visualization 17

A.3 Annotation Provenance and Coordinate Records 19

A.4 Quality Assurance and Split Construction 19

A.5 Human–Robot Motion Visualization 19

B Action Interfaces and Training Details **20**

B.1 Trajectory Definitions and Transfer Paradigm 20

B.2 Velocity-Field Objective and Valid-Target Normalization 20

B.3 Stage-Specific Training Objectives 21

B.4 Physical Grounding and Deployment 22

B.5 State–Action Definitions 22

B.6 Layer-Wise Architecture and Conditioning 22

B.7 Baseline Implementation Details 23

C Scaling Controls **24**

C.1 Validation Experiments and Plotting Protocol 24

D Simulation Evaluation Protocol **25**

E Real-World Evaluation Protocol **25**

E.1 Weighted Task Progress 25

E.2 Stage Criteria and Autonomous Success 26

F Trial Adjudication and Statistical Analysis **27**

G Data Ablation Protocol **27**

H Object-Exposure Generalization Protocol **27**

I Failure Taxonomy **28**

J Real-World Demonstration Sequences **29**

A Dataset Scope, Manifest, and Provenance

The three training stages use disjointly defined data resources. Stage I draws on the seven heterogeneous egocentric datasets listed in the main paper. These public sources are not counted as part of HumanVerse-500. HumanVerse-500 denotes only the 500-hour synchronized egocentric whole-body corpus used in Stage II. Robot demonstrations are used only in Stage III and are excluded from both the HumanVerse-500 duration and its scaling subsets. This separation distinguishes human-data scaling in Stage II from robot supervision in Stage III.

The frozen HumanVerse-500 manifest records each episode's participant, recording session, environment, activity label, duration, video frame count, body-tracking coverage, per-hand visibility, coordinate metadata, conversion version, and train or validation assignment. Temporal windows inherit the split of their source episode and are never partitioned independently.

An episode is structurally valid only when timestamps are monotonic, video and tracking intervals overlap, coordinate conventions are declared, and every unmasked value is finite. Missing or occluded hands mask only the affected supervision blocks. Reprojection error, estimator confidence, body-tracking coverage, and conversion failures remain explicit diagnostic fields rather than undocumented sampling filters.

Data collection efficiency. Based on our collection experience, a four-hour human-data collection session typically yields 1.5–2 hours of usable recordings, corresponding to 37.5–50% of the session duration. A robot teleoperation session of the same duration typically yields about 50 demonstrations totaling 30–50 minutes, or 12.5–20.8% of the session duration. Human recording therefore provides approximately 1.8–4 times as much usable recording time per collection hour. These figures are approximate operational estimates rather than a controlled cost study. They compare collection time, excluding hardware and downstream processing costs, and do not imply equal supervision value per recorded minute.

A.1 Qualitative Examples of Task Diversity

Figure 9 presents selected egocentric frames illustrating everyday tasks and manipulated objects. The labels identify the source tasks. Each frame depicts one moment within the corresponding task. These examples illustrate qualitative diversity and do not estimate dataset-wide task frequencies.

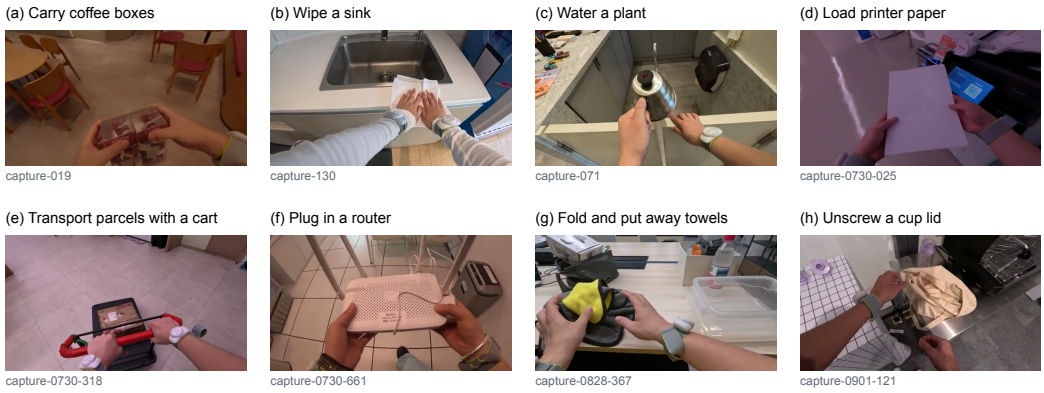

Figure 9: **Representative examples from HumanVerse-500.** Selected frames illustrate carrying coffee boxes, wiping a sink, watering a plant, loading printer paper, transporting parcels, connecting a router, folding towels, and opening a cup. Source clip identifiers are retained beneath each panel.

A.2 Synchronized Human and Camera Visualization

Figure 10 pairs three human-motion visualizations from the project-page viewer with their synchronized egocentric frames. The examples show sampled moments from carrying coffee boxes, wiping a sink, and watering a plant.

Table 4: Egocentric human-data resources and documented training interfaces. Body and hand coverage refers to supervision, not what appears in RGB. NR: not reported in the cited source. This does not imply incompatibility.

Resource	Hours	Body / hands	SMPL	Humanoid interface	Capture setup
EgoDex (Hoque et al., 2026)	829	Upper body and articulated hands	NR	NR in dataset release	Apple Vision Pro
HOT3D (Banerjee et al., 2025)	13.9	Hands and objects without full-body coverage	NR	NR	Aria / Quest 3 with external tracking
NymeriaPlus (DeTone et al., 2026)	NR[a]	Full body, finger motion NR	Provided	NR	Aria, wristbands, XSens
EgoHumanoid (Shi et al., 2026)	NR	Full body and articulated hands	NR	G1: wrist targets + locomotion commands	PICO + trackers, ZED X Mini
EgoScale (Zheng et al., 2026b)	20,854	Wrist and hand action supervision	NR	G1 transfer via wrist / hand interface	Multiple human-data sources
Index (Figure, 2026)	NR	Pose schema NR	NR	Figure / Helix, G1 NR	Distributed creator recordings
EgoSuite-Open100K (Lightwheel AI, 2026)	10k[b]	Hands, full body in body subsets	NR	NR (LeRobot is a storage format)	Head camera, wrist views in Pro
HUMANVERSE-500 (Ours)	500	Tracked full body and reconstructed hands	Compatible	SONIC body tokens and G1 training views	Portable PICO + GoPro

[a]NymeriaPlus upgrades the 300+ h Nymeria corpus. Its retained duration after sequence filtering is not separately specified here. [b]The August 26, 2026 announcement reports 10k h initially available out of 100k h planned, including 12k h of planned full-body subsets. These planned hours are not counted as released full-body data. SMPL compatibility and a demonstrated robot interface are distinct. Our body stream is device-derived kinematics, not optical motion-capture ground truth. Hand geometry is reconstructed offline. Capture hardware is listed without an unverified cost ranking.

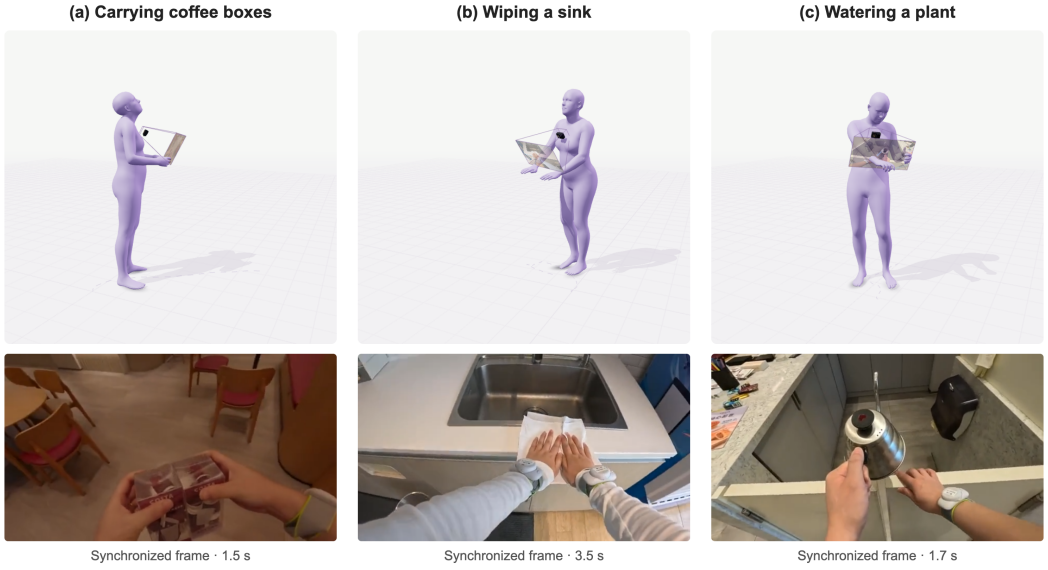

Figure 10: **Synchronized human motion and egocentric observations.** Top: SMPL body meshes with MANO hand motion and illustrative camera frusta. Bottom: RGB frames at the same timestamps. Camera frusta indicate the viewing direction schematically and do not represent calibrated camera parameters.

A.3 Annotation Provenance and Coordinate Records

The wearable body stream is synchronized device kinematics, not optical motion-capture ground truth. Dyn-HaMR reconstructs 21 joints per hand from RGB using ViPE camera motion. Records retain the camera trajectory used for estimation, per-track visibility, and diagnostic confidence. Raw streams are the canonical dataset record. Derived views are versioned with source-frame indices and coordinate transforms. Unavailable or occluded observations remain masked rather than being replaced by temporally imputed poses.

Camera-to-world and body-frame metadata are preserved whenever available. An episode-fixed hand-trajectory view is anchored at the first valid camera pose to retain global displacement, while current-camera coordinates support geometric verification. These stored views are distinct from the query-time camera frame and relative wrist targets used by the training interface in Sec. 4.1. Finger geometry is wrist-local, palm-aligned, and scale-normalized. Whole-body tokens are computed from the tracked temporal sequence using a frozen motion encoder, not from an episode-level pose summary. Body and hand targets retain separate validity masks because their sensors and estimators have different failure modes.

A.4 Quality Assurance and Split Construction

Structural checks verify stream lengths, timestamp monotonicity, finite values, coordinate conventions, and source identity. Geometric checks reproject reconstructed hands into RGB and record reprojection and visibility diagnostics. A synchronized viewer exposes RGB, current-camera 3D, episode-frame 3D, and body trajectories for manual audit. Diagnostics affect inclusion only through the pre-specified data contract: invalid target dimensions are masked and conversion failures are reported explicitly.

HumanVerse-500 is partitioned into training and validation sets by recording session before temporal window extraction. Separate session, object-instance, and evaluation-configuration identifiers support experimental separation. The frozen robot evaluation manifest is not consulted during human-data selection. Dataset reporting includes duration, episode count, participant and environment counts, activity distribution, bilateral-hand visibility, body-tracking coverage, and conversion failure rate. Nested scaling subsets are drawn only from the frozen general-purpose training partition, as detailed in Appendix C.

A.5 Human–Robot Motion Visualization

Figure 3 shows human and corresponding G1 poses at 0, 3, 7, and 8 s with increasing opacity. For visualization, G1 poses follow the human root path, horizontal displacement is amplified $4\times$, and body heights are normalized. The displayed paths indicate pose correspondence, not measured robot trajectory tracking.

Table 5: Problem-level comparison with the closest humanoid VLA and world-action studies. "General" denotes human data not collected for the downstream evaluation tasks. "Task-targeted" denotes demonstrations whose task or object coverage is selected for the downstream study. This table compares experimental questions rather than reported scores.

Method	Human signal	Role in training	Controlled scale	Primary question
WholeBodyVLA (Jiang et al., 2026)	300 h action-free egocentric video	latent locomotion pre-training, robot manipulation latents, and teleop grounding	0–100% video × teleop trajectories	Can visual latent actions reduce loco-manipulation teleoperation?
EgoHumanoid (Shi et al., 2026)	same-task RGB and wearable motion	direct human–robot co-training after view/action alignment	0–300 human demonstrations and sampling ratio	Can task-matched human data improve scene generalization?
Ψ_0 (Wei et al., 2026b)	829 h wrist/fingertip-labeled EgoDex video	task-space VLM pre-training before robot joint-action learning	full EgoDex versus 10%	Which staged data recipe yields an effective humanoid foundation model?
OpenHLM (Hu et al., 2026)	task-targeted stationary HuMI motion	heterogeneous co-training after whole-body VLA adaptation	whole-body teleop count and data-source choice	Which interfaces and data sources yield a whole-body-native VLA?
ω-0 (Li et al., 2026b)	public human video and whole-body motion (ARCTIC, Xperience-10M, Motion-X)	action-token VLM pre-training and SONIC replay for predictive action-latent learning	human-data volume scaling not reported; data-source and model ablations	Can latent visual foresight improve concurrent whole-body loco-manipulation?
λ_0	500 h synchronized RGB, full-body motion, and bilateral hands	human-data mid-training followed by robot post-training	0–100% Stage-II human data with a fixed robot budget	What scales with explicit whole-body human experience under fixed grounding?

B Action Interfaces and Training Details

λ_0 uses a Qwen3.5-2B vision–language backbone, layer-wise visual–language features, a generative action expert, and explicit embodiment conditioning. The Stage-I hand, G1-compatible whole-body, and SIMPLE interfaces retain their corresponding state–action contracts. The robot-compatible Stage-II view and native G1 Stage-III data use the same robot state–action format, with distinct training-domain identifiers. All interfaces predict 50-step action chunks, with masks excluding unavailable blocks and padded dimensions from the loss.

Stage-I pre-training uses a global batch size of 1024, a base learning rate of 10^{-5}, and a 0.1 learning-rate multiplier on the vision–language backbone. An exponential moving average with decay 0.999 is updated after each optimizer step and initializes the next stage. Checkpoints contain the online parameters, EMA parameters, optimizer state, global step, data-state metadata, and random-number states required for deterministic resumption.

Stage-II supervision is constructed offline. The wearable body stream is processed by the frozen SONIC controller in a G1 MuJoCo environment, while hand articulation is retargeted to the bilateral hand interface. The robot-compatible record stores 29-DoF body state, bilateral hand state, projected gravity, the body-motion token, and bilateral hand commands. Training reads only finalized LeRobot v2.1 records. Hand estimation, simulation, and retargeting are never executed online.

Stage III uses native robot demonstrations to adapt the policy to downstream execution. The robot-domain projections, shared backbone, and action expert are optimized using the valid-target masks. Checkpoints retain the robot dataset version, training state, and optimizer state for resumption.

As in Sec. 4.2, Stage III minimizes

$$\mathcal{L}_{\text{post}} = \mathbb{E}_{\tau\sim\mathcal{D}_R}[\mathcal{L}^R(\tau;\boldsymbol{\theta})], \tag{6}$$

where $\mathcal{L}^R$ uses the valid-target masked MSE in the main text. The robot demonstrations and Stage-III update budget remain fixed for every HumanVerse-500 scaling condition.

B.1 Trajectory Definitions and Transfer Paradigm

A training domain d provides trajectories

$$\tau_i^d = \left\{ \left(\mathbf{x}_{i,t}^d, \mathbf{s}_{i,t}^d, \ell_i^d, \mathbf{a}_{i,t+1:t+H}^d, \mathbf{m}_{i,t+1:t+H}^d\right) \right\}_{t=0}^{T_i-1}, \qquad H = 50, \tag{7}$$

where $\mathbf{x}_{i,t}^d$ is the egocentric RGB observation, $\mathbf{s}_{i,t}^d$ is the current state, ℓ_i^d is the language instruction, $\mathbf{a}_{i,t+1:t+H}^d$ is a future action chunk, and $\mathbf{m}_{i,t+1:t+H}^d$ marks its valid temporal and feature dimensions. Each domain is associated with an embodiment identifier $e(d)$ and an action space $q(d)$. Neither identifier implies numerical correspondence between domains.

We distinguish three data collections. $\mathcal{D}_E$ is the union of seven public egocentric hand–object datasets, $\mathcal{D}_W$ is the whole-body human-motion dataset HumanVerse-500, and $\mathcal{D}_R$ contains native robot demonstrations. Each trajectory in $\mathcal{D}_W$ is converted to the G1-compatible format by Φ_R. This conversion uses frozen retargeting and simulation modules. The resulting data remain human-derived. Our transfer paradigm is therefore

$$\boldsymbol{\theta}^{(0)} \xrightarrow{\mathcal{D}_E} \boldsymbol{\theta}^{(1)} \xrightarrow{\Phi_R(\mathcal{D}_W)} \boldsymbol{\theta}^{(2)} \xrightarrow{\mathcal{D}_R} \boldsymbol{\theta}^{(3)}. \tag{8}$$

Stage I learns interaction geometry, Stage II couples manipulation with whole-body motion, and Stage III grounds the transferred representation in a task and a physical embodiment.

B.2 Velocity-Field Objective and Valid-Target Normalization

A common learning objective. All λ_0 checkpoints and the four adapted baselines use velocity-field prediction. The target is $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{a}$. The action $\mathbf{a}$ is the data target used to construct the noisy input, not the quantity predicted directly. For each sample, we draw $u \sim \mathcal{U}[0,1)$ and set $\sigma = 0.001 + 0.999u^{2/3}$, corresponding to a rescaled $\mathrm{Beta}(1.5,1)$ distribution. Given $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, we construct

$$\widetilde{\mathbf{a}}_{h,\sigma} = (1-\sigma_h)\mathbf{a}_h + \sigma_h\boldsymbol{\epsilon}_h, \qquad \mathbf{v}_h = \boldsymbol{\epsilon}_h - \mathbf{a}_h, \qquad \widehat{\mathbf{v}} = f_{\boldsymbol{\theta}}(\mathbf{x},\ell,\mathbf{s},e,q,\widetilde{\mathbf{a}}_\sigma,\boldsymbol{\sigma}). \tag{9}$$

Let r denote the sampled action-prefix length. For RTC training, half of the samples use no action prefix ($r = 0$). For the remainder, we sample up to eight valid prefix steps, set their noise level to zero, and exclude them from supervision. These zero-noise prefix actions condition prediction of the remaining suffix, and prefix positions are excluded from the velocity loss.

Let M contain every valid, non-prefix scalar target (i, h, j) in a batch. The implemented objective is

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\sigma,\boldsymbol{\epsilon},r}\left[\frac{1}{|M|}\sum_{(i,h,j)\in M}\left(\widehat{v}_{i,h,j} - v_{i,h,j}\right)^2\right]. \tag{10}$$

The validity mask removes unavailable dimensions, missing-hand annotations, and cross-interface padding. Action-space blocks are used to report diagnostic losses, but the optimizer averages all valid scalar targets without block reweighting. The mid-training and post-training recipes draw four independent noise realizations per observation before applying the same masked velocity-field objective.

B.3 Stage-Specific Training Objectives

All stages use the conditional velocity-field objective above, but differ in the data distribution, prediction interface, and role of supervision. Stage transitions initialize from the preceding exponential-moving-average (EMA) checkpoint, preserving learned representations while activating the interfaces needed by the next stage.

Stage I: interaction-prior pre-training. This stage learns visual and language-conditioned hand–object interaction structure before introducing whole-body targets. We initialize the shared backbone from a vision–language checkpoint and optimize it jointly with the action expert and egocentric hand interface on 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). VITRA-1M additionally indexes activity from Ego4D, EPIC-KITCHENS, Ego-Exo4D, and Something-Something V2 (Grauman et al., 2022; Damen et al., 2022; Grauman et al., 2024; Goyal et al., 2017). If N_d is the number of usable frames in source d, square-root sampling gives

$$p_1(d) = \frac{\sqrt{N_d}}{\sum_{d'\in\mathcal{S}_E}\sqrt{N_{d'}}}, \qquad \boldsymbol{\theta}^{(1)} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{d\sim p_1,\,\tau\sim\mathcal{D}_d,\,t}\left[\mathcal{L}^E_{\text{denoise}}(\tau, t)\right]. \tag{11}$$

The prediction target is future relative wrist motion and local finger geometry, not humanoid control. Square-root sampling retains coverage from small sources while preventing the largest corpus from setting the entire training distribution. Dataset identity remains available to the conditioning module, and all supervision retains its native frame index, provenance, and validity mask.

Stage II: whole-body motion-prior mid-training. Starting from the Stage-I EMA checkpoint, Stage II uses human demonstrations converted to the G1-compatible state–action format, as defined in Sec. 4.2. Human-derived data use distinct domain identifiers and separate input and output projections from native robot data. The shared backbone, action expert, and human-data projections are updated while the Stage-I hand head remains frozen. The resulting EMA checkpoint initializes Stage III.

Stage III: embodiment-grounded post-training. We initialize from the Stage-II exponential-moving-average checkpoint and adapt to target embodiment e. This stage addresses the remaining gap between motion supervision and execution in the target robot's physical environment. Native robot trajectories provide executable state transitions through the robot interface. The robot post-training objective is

$$\mathcal{L}_3(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\mathcal{D}_R}[\mathcal{L}^R_{\text{denoise}}(\tau;\boldsymbol{\theta})], \qquad \boldsymbol{\theta}^{(3)} = \arg\min_{\boldsymbol{\theta}\leftarrow\text{EMA}(\boldsymbol{\theta}^{(2)})} \mathcal{L}_3(\boldsymbol{\theta}). \tag{12}$$

The shared backbone, action expert, embodiment conditioning, and active interfaces are optimized jointly. We hold the Stage-III update budget and $\mathcal{D}_R$ fixed in the human-data scaling comparison. The output is a task- and embodiment-adapted policy, rather than direct deployment of a human-only model. In the default 2B recipe, Stages I–III use 200k, 100k, and 50k optimizer updates, respectively, with the same 50-step prediction horizon.

B.4 Physical Grounding and Deployment

Fig. 5 shows the target humanoid and teleoperation setup used to collect $\mathcal{D}_R$ for Stage III. Unlike the simulated view in Stage II, these demonstrations pair observations and robot states with actions recorded during physical task execution. They ground the representations learned in the preceding stages in physical robot execution.

At deployment, the policy is conditioned on the target robot's observation, instruction, state, and embodiment identifier, and uses only its robot output interface. For G1, the predicted body-motion tokens are interpreted by the SONIC control interface and paired with robot hand commands. Reconstructed human joints are not sent directly to actuators. Human streams are training resources, not required deployment inputs. Other embodiments retain their native interfaces. Controller timing and task-evaluation constraints are separate from the shared action-prediction objective.

B.5 State–Action Definitions

Egocentric hand interface. For hand $\chi \in \{L, R\}$, let $\mathbf{p}_t^\chi \in \mathbb{R}^3$ and $\mathbf{R}_t^\chi \in \mathrm{SO}(3)$ denote the wrist pose, let $\rho_6(\mathbf{R}) \in \mathbb{R}^6$ be its continuous 6D representation, and let $\mathbf{g}_t^\chi \in \mathbb{R}^{60}$ contain the 20 non-wrist joints in a wrist-local, palm-aligned, scale-normalized frame. The current hand state is

$$\mathbf{h}_t = \left[\mathbf{p}_t^L, \rho_6(\mathbf{R}_t^L), \mathbf{g}_t^L, \mathbf{p}_t^R, \rho_6(\mathbf{R}_t^R), \mathbf{g}_t^R\right] \in \mathbb{R}^{138}. \tag{13}$$

All future wrists are expressed in the camera frame at the query time t. Stage-I actions retain relative wrist motion and absolute local finger shape:

$$\mathbf{a}_{t+h}^{E,\chi} = \left[\mathbf{p}_{t+h}^\chi - \mathbf{p}_t^\chi, \rho_6\big((\mathbf{R}_t^\chi)^\top \mathbf{R}_{t+h}^\chi\big), \mathbf{g}_{t+h}^\chi\right], \qquad \mathbf{a}_{t+h}^E = \left[\mathbf{a}_{t+h}^{E,L}, \mathbf{a}_{t+h}^{E,R}\right] \in \mathbb{R}^{138}. \tag{14}$$

This construction removes arbitrary camera-world offsets while retaining the geometry needed to distinguish reaching, grasping, and in-hand interaction.

G1-compatible whole-body interface. Let $\mathbf{z}_t \in \mathbb{R}^{64}$ be the motion token produced by the frozen SONIC encoder. For converted human data, a closed-loop free-base G1 rollout provides 29 body-joint positions and projected gravity $\mathbf{g}_t^{\mathrm{proj}} \in \mathbb{R}^3$, while hand retargeting provides six normalized Revo2 commands per side. Its interface is

$$\begin{aligned} \mathbf{s}_t^R &= \left[\mathbf{q}_t^{1:29}, \mathbf{u}_t^L, \mathbf{u}_t^R, \mathbf{g}_t^{\mathrm{proj}}\right] \in \mathbb{R}^{44}, \\ \mathbf{a}_{t+h}^R &= \left[\mathbf{z}_{t+h}, \mathbf{u}_{t+h}^L, \mathbf{u}_{t+h}^R\right] \in \mathbb{R}^{76}. \end{aligned} \tag{15}$$

Native G1 post-training data use the same state–action format, with distinct domain identifiers and separate input and output projections. Other embodiments, including SIMPLE, retain their native state and actuator blocks. Variable-width states and actions are padded only at collation time. Masks remove padding, episode boundaries, failed conversions, and invisible hands from supervision.

B.6 Layer-Wise Architecture and Conditioning

Figure 11 summarizes the shared policy and its inference path; the equations below detail the input encoding and layer-wise conditioning.

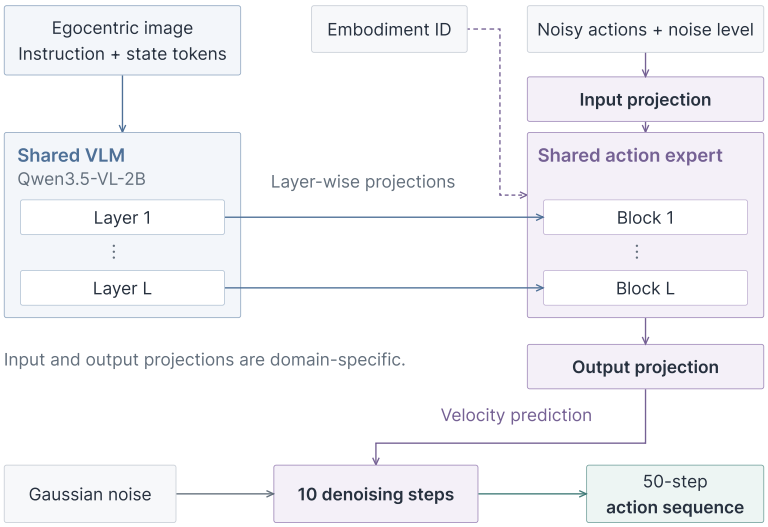

Figure 11: **Network architecture of λ_0.** The shared vision–language backbone encodes the egocentric image and state-aware instruction, providing projected features to the corresponding action-expert layers. Domain-specific input and output projections connect the shared expert to each action space, with embodiment conditioning. At inference, velocity predictions guide ten denoising steps from Gaussian noise to a 50-step action sequence.

λ_0 combines a shared Qwen3.5-2B vision–language backbone with a layer-wise generative action expert. Normalized proprioception is quantized into 256 bins and appended to the instruction. The backbone then produces hidden states $\mathbf{C}^{1:L} = V_\phi(\mathbf{x}, [\ell; Q(\mathbf{s}^q)])$. For action space q, a domain-specific encoder maps the noisy action and a sinusoidal noise embedding to action tokens,

$$\mathbf{u}_h^{q,0} = E_q([\tilde{\mathbf{a}}_h^q, \gamma(\sigma_h)]) \,. \tag{16}$$

An embodiment-specific 32-token soft prompt and a learned embodiment bias condition the shared expert. At VLM layer l, the corresponding projected context is injected into the lth expert block,

$$\left(\mathbf{P}_e^l, \mathbf{U}^{q,l}\right) = F_\psi^l\left(\mathrm{Proj}_l(\mathbf{C}^l), \mathbf{P}_e^{l-1}, \mathbf{U}^{q,l-1}\right), \qquad \hat{\mathbf{v}}^q = D_q(\mathbf{U}^{q,L}). \tag{17}$$

Thus V_ϕ and $F_\psi^{1:L}$ transfer across stages, while (E_q, D_q) preserve the semantics and dimensionality of each state–action space. A new embodiment appends conditioning rows. If it introduces a new action contract, its interface is appended without overwriting previous ones.

B.7 Baseline Implementation Details

We implement the four baselines through a shared data adapter, which preserves each upstream trainer and checkpoint format while enforcing one G1 real-data contract. The contract has a 44-dimensional state (29 body-position values, 6 left-hand values, 6 right-hand values, and 3 projected-gravity values) and a 76-dimensional absolute action (64 SONIC motion values plus 6 values for each Revo2 hand), sampled at 50 Hz with a 50-frame future buffer. Dataset-specific normalizers use the recorded training statistics. The physical evaluation uses the same four tasks and ten trials per task, without averaging over training seeds.

GR00T N1.6. We use the official Eagle vision–language backbone and N1.6 flow action expert with its 128-dimensional action-expert capacity. Relative-action conversion is disabled, the Lambda-0 action projector is reinitialized when its dimensions change, and the adapter uses a 50-step action horizon, learning rate 10^{-4}, and its two-GPU effective batch of 64 (micro-batch 1 with gradient accumulation 32). We initialize from the official pretrained GR00T N1.6 weights.

Ψ_0. We use Qwen3-VL-2B with the official Ψ_0 flow action expert and SonicTrainer, with a 44-dimensional state and a 76-dimensional action, a 30-frame real-SONIC action chunk, and RTC enabled with maximum delay 8. The VLM is frozen, the action expert is adapted with learning rate 10^{-4}, 1,000 warm-up steps, cosine decay, and the released vision–language and action-expert checkpoints.

$\pi_{0.5}$. We use the openpi JAX implementation with the official pretrained $\pi_{0.5}$ parameters, the 76-dimensional absolute action interface, and the adapter's per-dataset normalization statistics. The default two-GPU batch is 32 (16 examples per device). The native optimizer, checkpointing, and action-sampling code are unchanged.

StarVLA. We use Qwen3-VL-4B with the QwenPI-v3 layerwise flow-matching action model, state dimension 44, action dimension 76, horizon 50, four repeated diffusion noises, and ten inference steps. The released configuration uses ZeRO-2, learning rates 10^{-5} for the VLM and 10^{-4} for the action model, cosine decay, gradient clipping at 1, and a 50k-step budget. The action expert is randomly initialized unless an explicit checkpoint is supplied. All baselines retain their native optimizer, distributed-training, and checkpoint semantics. The adapter only standardizes data views, dimensions, normalization, and evaluation commands.

C Scaling Controls

The human-data scaling axis is expressed as a fraction of the frozen HumanVerse-500 training partition. Validation sessions are excluded before forming the subsets. The 0%, 5%, 10%, 25%, 50%, and 100% conditions define the evaluation grid for Fig. 7. Nonzero subsets are nested, sampled at the episode level from the frozen training partition, and stratified by activity and recording session before chunk extraction. The Stage-I checkpoint, robot post-training set, Stage-III update budget, validation set, and evaluation manifest remain fixed.

The primary compute-matched comparison uses the same number of Stage-II optimizer updates for every nonzero fraction. The 0% reference omits Stage II. Validation loss and closed-loop metrics are analyzed separately so a reduction in the Stage-II objective is not treated as equivalent to improved robot control.

C.1 Validation Experiments and Plotting Protocol

Fig. 6 reports validation results at 5%, 10%, 25%, 50%, and 100% training-data fractions. Each run uses the 2B variant, training seed 43, and 100,000 optimization steps. We report the total validation loss on the clean and representative validation sets, evaluated every 1,000 steps with validation seed 20260813. Data fractions are reported directly rather than converted to hours. These are offline prediction metrics, not closed-loop robot success rates.

The two validation sets probe different data-quality regimes. *Clean* contains held-out capture sessions selected for high frame-weighted validity of both SONIC whole-body action annotations and bilateral hand annotations: both session-level validity ratios must reach the respective upper-quartile thresholds in the original split. Thus, "clean" refers to annotation validity, not visual appearance or task difficulty. *Representative* is a separate holdout stratified by data source and quartile bins of the two validity ratios, covering a broader range of annotation quality. The frozen sets contain 1,008 and 513 episodes, respectively, each from 44 capture sessions. Their sessions are mutually disjoint and excluded from training. Both sets remain fixed across all five training-data fractions.

Following the diagnostic decomposition used by EgoScale (Zheng et al., 2026b), we distinguish validation dynamics from the best recorded validation error at each data scale. The best validation loss on both holdouts decreases across the five data scales. We show measured minima with descriptive log-linear fits. These are not evidence of a universal scaling law. Downstream transfer is evaluated separately.

The two plots in panel (a) show the 100 unsmoothed evaluations from step 1,000 through 100,000 on a shared logarithmic loss axis spanning 0.005–0.02 and a step axis starting at zero. Step 0 is omitted from these panels for readability but retained in the source data. Panel (b) selects each holdout's minimum independently over the full 0–100,000-step window. Minima need not occur at the same step. Solid circles denote clean and pale-gray-centered circles representative validation. Each data fraction has one run, so no across-seed error bars are shown.

For each holdout, we fit $\widehat{L}(D) = a + b \ln D$ by unweighted ordinary least squares to the five measured minima, where $D \in \{5, 10, 25, 50, 100\}$ is the training-data percentage. Clean gives $a = 0.0112283$, $b = -0.00126478$, and $R^2 = 0.9901$. Representative gives $a = 0.0129477$, $b = -0.00129865$, and $R^2 = 0.9868$. Pale-purple dashed lines show these fits only within the

observed range. Their slope magnitudes differ by approximately 2.7% relative to clean. A doubling of data predicts an absolute loss reduction of $-b \ln 2$: 0.000900 for representative and 0.000877 for clean. Thus, the two holdouts exhibit similar absolute, not necessarily relative, improvements over the observed data range. With one run per scale, this is a descriptive comparison, not a statistical test of equal slopes. The reported R^2 values measure in-sample fit, not cross-seed reproducibility or extrapolation accuracy.

D Simulation Evaluation Protocol

We evaluate the six core SIMPLE leaderboard tasks reported in Table 2. Each task has ten rollouts at each of Levels 0, 1, and 2. An entry $a/b/c$ in Table 2 denotes success rates in percent at these three levels. With ten trials, each success contributes ten percentage points. For successful-trial count $n_{k,l}$ on task k at level l, the main-table metrics are $S_l = 100 \sum_{k=1}^{6} n_{k,l}/60$ and $S_{\text{overall}} = (S_0 + S_1 + S_2)/3$. λ_0 uses single-task training. StarVLA and λ_0 entries are from our evaluations. Public baseline entries retain their source protocols.

The current λ_0 summary selects the best recorded configuration separately for each task, rather than a shared checkpoint and inference setting. All six SIMPLE tasks use the velocity-field action head. No legacy x_0 action head is used. The reported row retains the task-wise checkpoint and inference selection recorded for each task. The six tasks sum to 162 successes in 180 trials. Configuration selection was not established here as validation-only. These numbers should not be interpreted as a fixed-configuration comparison. The Full Recipe row in Table 3 uses this updated task-wise summary (90.0%, 93.3%, and 86.7% at Levels 0–2, with 90.0% overall). Matched checkpoint and inference selection across ablations remains to be verified.

Public baseline counts follow the official leaderboard accessed September 13, 2026, including its updated ACT results. Mobile P&P corresponds to locomotion pick between tables.

E Real-World Evaluation Protocol

For each reported policy and task, we run ten trials from ten frozen initial configurations. Results are aggregated over these ten trials. We do not average across training seeds. Configuration identities, object instances, target regions, task instructions, timeouts, and terminal predicates are fixed before evaluating any policy. The following stage-completion criteria and terminal predicates are shared by all methods.

E.1 Weighted Task Progress

For task t with K_t stages, the evaluator assigns $s_{t,k} \in \{0, 1\}$ to stage k: a completed stage receives its full preassigned point value $w_{t,k}$, and an uncompleted stage receives zero. The point value is the weight, not a separate subjective score. The weights satisfy $w_{t,k} > 0$ and $\sum_{k=1}^{K_t} w_{t,k} = 100$. For rollout r, we report

$$P_t^{(r)} = \sum_{k=1}^{K_t} w_{t,k} s_{t,k}^{(r)} \quad \in [0, 100]. \tag{18}$$

Thus, 75 points is reported as 75% progress. This is a weighted sum of binary stage outcomes, not an equal-weight stage count or a subjective continuous quality rating. A stage earns credit only when its own observable criterion is met. The sum does not additionally apply a completed-prefix gate. Repeated completion within one rollout does not earn additional points. All stages must be adjudicated before a record is considered complete. A fully adjudicated record is not necessarily a successful rollout.

For R_t rollouts per task and T evaluated tasks, task means and the macro-average are

$$\overline{P}_t = \frac{1}{R_t} \sum_{r=1}^{R_t} P_t^{(r)}, \qquad \overline{P} = \frac{1}{T} \sum_{t=1}^{T} \overline{P}_t. \tag{19}$$

Each task therefore has equal weight regardless of its number of stages. The same stage definitions and point allocations must be fixed before evaluation and used for every method and data-scale

condition. Table 6 lists every stage and its point value. The point values sum to 100 for each task. Bottle disposal has six task-execution stages and a separate five-point post-disposal stability stage.

E.2 Stage Criteria and Autonomous Success

Table 6 summarizes the task stages. For bottle disposal, approaching the bottle means bringing the hand into its effective grasp region. A stable grasp requires reliable bottle retention, and lifting requires the bottle to clear the table. Arrival at the bin requires a pose from which the robot can both operate the pedal and deliver the bottle. Pedal opening is a separate stage: touching the pedal is insufficient. The lid must open far enough to admit the bottle. The disposal stage is worth 20 points and requires release followed by the bottle entering the bin. Release alone is insufficient. The final five-point stage requires the robot to remain stable for three seconds after disposal, without human support or unintended collision. Standing still without disposing of the bottle does not satisfy this post-disposal criterion.

For toy storage, contact is distinct from a stable grasp, lifting requires table clearance, and transport requires the toy to reach the box opening. Placement and release are distinct: the toy must enter the box and remain there after the hand lets go. For bimanual box transport, both hands must establish contact, the box must remain stable relative to the grasp, and its base must clear the table before transport. Placement requires support by the target surface, followed by stable release. For chair placement, turning and alignment, bilateral contact, stable grasp and push initiation, sustained pushing, final insertion, and safe release are assessed separately.

Full-task success is a separate binary outcome, requiring the task-specific terminal predicate to hold for three seconds without human intervention. Safety and intervention are recorded separately for all tasks, even when a task includes an explicitly scored stability stage. The bottle task awards five points for post-disposal stability. The chair task assigns 20 points to final insertion and five points to release and safe termination. A progress score alone must not be used to infer autonomous success.

Table 6: Operational definitions for the real-world evaluation. Progress is the sum of point values for completed stages (shown in parentheses, totaling 100 per task). Terminal success additionally requires the task-specific predicate to remain satisfied for three seconds.

Task	Scored task stages (points)	Terminal success predicate
Bottle disposal	1. Approach the bottle (10) 2. Grasp stably (15) 3. Lift clear of the table (15) 4. Arrive at the bin (15) 5. Press the pedal to open the lid (20) 6. Release the bottle into the bin (20) 7. Remain stable after disposal (5)	Bottle fully inside the trash can and robot standing without contact assistance
Single-hand toy storage	1. Approach the toy (10) 2. Establish contact (15) 3. Grasp stably (20) 4. Lift clear of the table (15) 5. Transport above the box (15) 6. Place inside the box (20) 7. Release with the toy remaining inside (5)	Toy fully inside the box, released, and not displaced during hand withdrawal
Bimanual parts-box transport	1. Approach the box (10) 2. Establish bilateral contact (15) 3. Establish a stable grasp (20) 4. Lift clear of the table (15) 5. Transport above the target surface (20) 6. Place on the target surface (15) 7. Release with the box remaining stable (5)	Box inside the target region, both hands released, and no contained parts spilled
Chair placement	1. Approach the chair (10) 2. Turn and align with the chair (15) 3. Establish bilateral contact (10) 4. Grasp stably and initiate pushing (20) 5. Push continuously toward the table (20) 6. Complete insertion under the table (20) 7. Release and stop safely (5)	Chair footprint, position, and yaw satisfy pre-registered under-table tolerances, with the robot released and stable

Main-paper Table 1 reports the paired evaluation against external baselines, and Table 3 reports the primary matched ablations under the same protocol.

F Trial Adjudication and Statistical Analysis

Each evaluation configuration receives a persistent identifier shared by all methods. Configuration order is randomized independently for each policy. Trial videos and state logs retain configuration and model identifiers, while model identity is hidden from manual adjudicators. Human contact with the robot, object, or environment after policy execution begins is an intervention and makes the rollout ineligible for autonomous success. Assisted continuations may be retained for diagnosis but are reported separately.

We report the raw ten-trial success counts and task means. No seed-based confidence interval or seed-resampling bootstrap is applied. Macro-averages weight tasks equally, independent of task duration. The validation-loss scaling plot reports one run per data fraction, without across-seed error bars. Its regression fits are descriptive only over the measured range (Appendix C.1).

G Data Ablation Protocol

The ablation study varies the inclusion of Stage-I pre-training and Stage-II mid-training while retaining vision–language initialization and Stage III in every variant. The w/o human-data pretraining condition omits both Stages I and II. The other two ablations omit Stage I or Stage II individually. No action-representation or architecture ablation is included. Main-paper Table 3 reports the w/o human-data pretraining, w/o Stage I, w/o Stage II, and full-recipe conditions on the four real-world tasks, with ten trials per task. Their total success counts are 6/40, 20/40, 9/40, and 25/40, respectively. Progress is averaged equally across the four tasks.

H Object-Exposure Generalization Protocol

The protocol covers all four real-world tasks. Let $\mathcal{D}_A^R$ denote robot demonstrations of physical object instance A, and let $\mathcal{D}_A^H$ and $\mathcal{D}_B^H$ denote human demonstrations of A and B. No robot demonstrations of B are used. A joint-data policy is trained with $\mathcal{D}_A^R \cup \mathcal{D}_A^H \cup \mathcal{D}_B^H$. A robot-only comparison uses $\mathcal{D}_A^R$ without human-data training. *Seen* evaluates the joint-data policy on A, *Human-guided* evaluates that same policy on B, and *Zero-shot* evaluates the robot-only policy on B. Zero-shot therefore refers to object-instance transfer within a demonstrated task, not to task-level zero-shot learning or the absence of generic visual-language pretraining.

The Human-guided and Zero-shot conditions use the same B instances and evaluation configurations, holding the instruction, scene, target region, initial robot configuration, and success predicate fixed. Robot A data are shared across the two training recipes. Success and weighted progress follow the paired real-world protocol. This comparison tests the combined benefit of human-data training. It does not isolate the contribution of human B demonstrations from human A demonstrations or earlier human pretraining. Fig. 8 reports mean weighted progress over 10 trials per task, method, and condition (600 trials in total). The Seen values come from the object-generalization evaluation, whereas Table 1 uses the standard-task evaluation records. The two sets of measurements are reported separately. Each bar is computed from the recorded milestone scores. Its solid checked and dashed object icons indicate presence and absence in robot demonstrations only. B is human-observed for Human-guided but unobserved in the robot-only recipe.

I FAILURE TAXONOMY

Table 7: Operational failure taxonomy for real-world evaluation. Each failed rollout is assigned its earliest unrecovered primary failure, while recovery attempts are recorded separately.

Category	Operational rule
Perception / grounding	Wrong target, receptacle, or task interpretation before purposeful contact
Locomotion	Fails to enter the manipulation workspace, loses balance, or cannot maintain a usable base pose
Reach / contact	Selects the correct target but fails to establish the intended hand or body contact
Grasp	Contact is established but the object is not secured or immediately slips before transport
Transport	A secured object is dropped, spilled, or carried along an invalid path
Placement	The object reaches the target workspace but violates the terminal position, orientation, containment, release, or stability predicate
Recovery	A recoverable error is detected but the policy fails to regrasp, reposition, or resume within the timeout
System / safety	Emergency stop, unintended collision, controller fault, or other execution failure after the policy begins acting

J Real-World Demonstration Sequences

Figure 12 shows temporally ordered frames from the four project-page demonstration videos. Frames are sampled at 0%, 25%, 50%, 75%, and 98% of each clip's duration. Timestamps refer to the supplied video files. These clips illustrate task execution and are not used to compute the aggregate evaluation metrics.

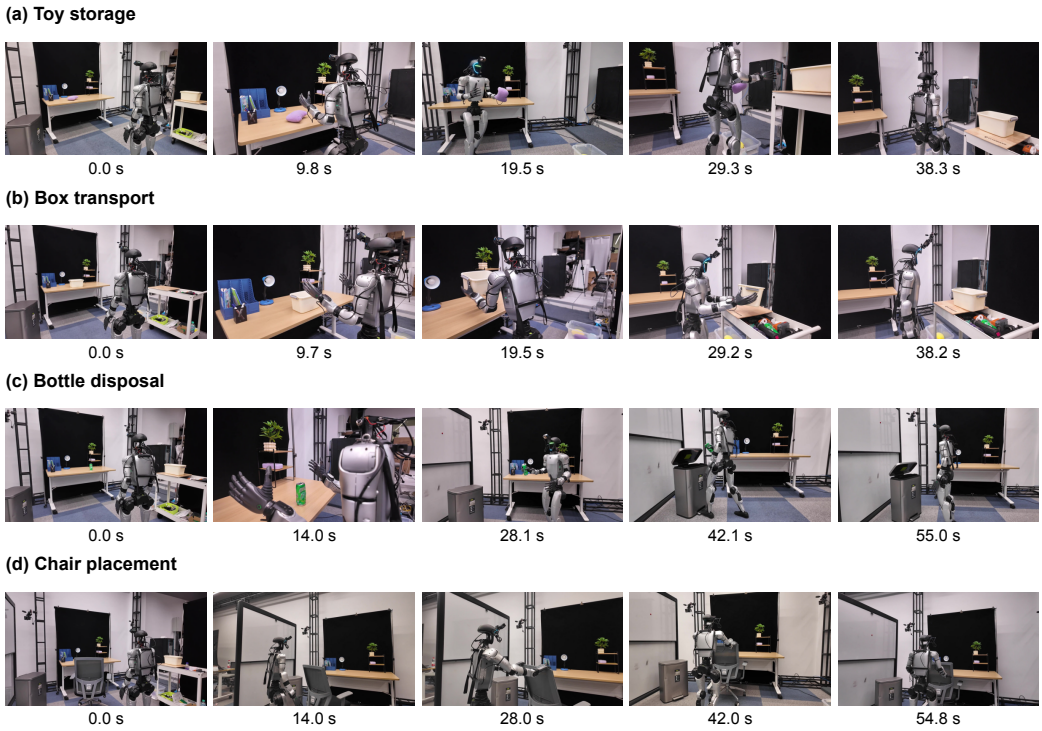

Figure 12: **Real-world demonstration sequences.** Rows show toy storage, box transport, bottle disposal, and chair placement. Time increases from left to right within each row. The frames retain the original camera view and display the elapsed clip time.