

λ_0 : Scaling Open-World Egocentric Human Data for Humanoid Loco-Manipulation



Alibaba Token Hub  and Amap  Alibaba Group

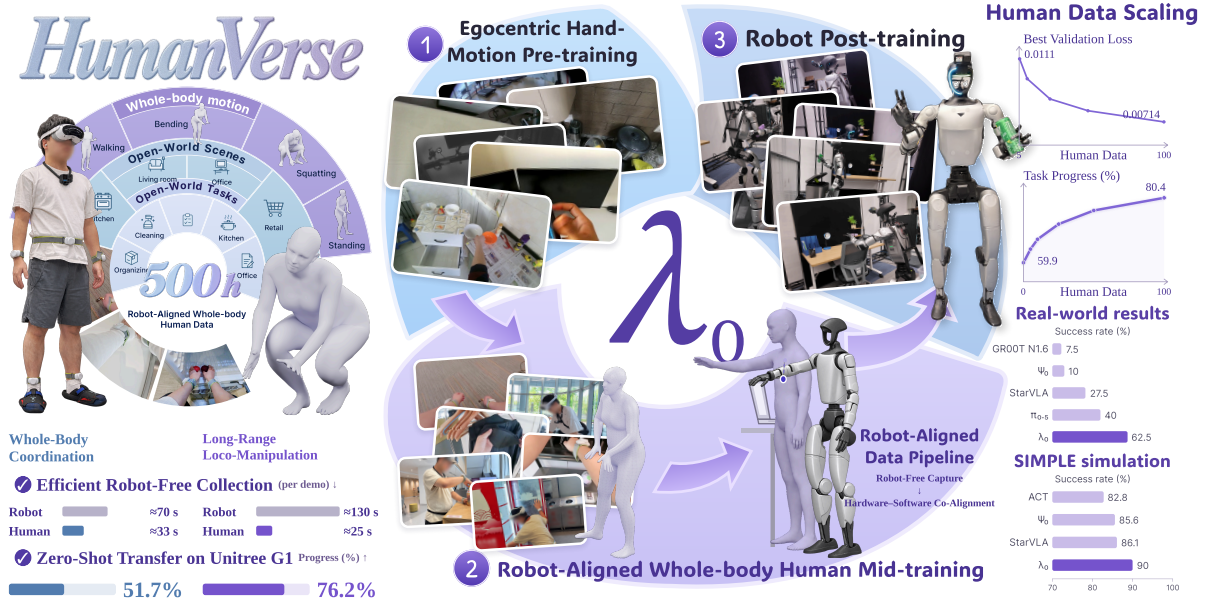


Figure 1: Introducing λ_0 with HUMANVERSE-500. Robot-free egocentric capture yields 500 hours of whole-body human loco-manipulation data. Three-stage training transfers this experience to λ_0 for humanoid loco-manipulation in simulation and the real world.

Abstract

Humanoid loco-manipulation requires coordinated locomotion, posture, and manipulation, yet collecting robot demonstrations of these behaviors is costly. Human demonstrations offer a scalable alternative, but existing egocentric human supervision often omits locomotion and posture. We introduce HUMANVERSE, a 500-hour dataset of open-world human loco-manipulation pairing egocentric video with synchronized body and hand motion. The data pipeline combines efficient robot-free capture with hardware–software co-alignment of visual observations and actions. Policies trained on the aligned human demonstrations transfer to Unitree G1 across whole-body manipulation and long-range loco-manipulation without target-task robot demonstrations. Building on this pipeline, we develop λ_0 , a humanoid foundation model for whole-body loco-manipulation, with robot-aligned human demonstrations providing supervision for mid-training. Increasing whole-body human data consistently improves downstream task progress. λ_0 achieves 90.0% success on SIMPLE and 62.5% across real-world loco-manipulation tasks. Our findings establish whole-body human demonstrations as a scalable source of supervision for humanoid loco-manipulation.

Code: <https://github.com/LambdaHumanoid/openlambda>

Homepage: <https://lambdahumanoid.github.io>

Contents

1	Introduction	3
2	Related Work	4
2.1	Generalist Robot Policies	4
2.2	Learning from Human Videos	4
2.3	Humanoid Whole-Body Control	4
3	HumanVerse-500: Scaling Egocentric Whole-Body Data Collection in the Open World	5
3.1	Robot-Free Capture and Collection Efficiency	5
3.2	Dataset Scope and Coverage	6
3.3	Human Motion Reconstruction and Annotation	6
3.4	Robot-Aligned Data Pipeline.	7
4	λ_0: Learning Humanoid Control from Human Experience	9
4.1	Policy Architecture and Action Interfaces.	9
4.2	Three-Stage Training	9
4.3	Training Objective and Deployment	11
5	Experiments on Humanoid: Alignment, Scale, and Transfer	11
5.1	Effectiveness and Efficiency of Robot-Aligned Human Data.	11
5.2	Q1: Can Human Supervision Advance Humanoid Loco-Manipulation?	11
5.3	Q2: How Do the Scale and Training Stages of Human Supervision Shape Humanoid Learning?	13
5.4	Q3: Can Human Data Improve Control and Generalization?	14
6	Conclusion.	16
7	Contributions.	16
	References.	16
	Appendix	22
A	Dataset and Related-Work Comparisons	22
A.1	Egocentric Dataset Comparison	22
A.2	Related-Method Comparison	22
B	Action Interfaces and Training Details	23
B.1	State–Action Definitions.	23
B.2	Model Architecture	23
B.3	Training Configuration.	24
C	Evaluation Protocols	24
C.1	Baseline Implementations.	24
C.2	Simulation Evaluation	24
C.3	Real-World Evaluation and Scoring	25
D	Human-Data Scaling and Ablation.	26
D.1	Data Quality and Scaling Protocol	26
D.2	Training-Stage Ablations	26
D.3	Generalization	27

1 Introduction

Learning humanoid loco-manipulation across tasks and environments requires demonstrations that capture coordinated locomotion, posture, and manipulation (Jiang et al., 2026a; Shi et al., 2026; Wei et al., 2026b; Hu et al., 2026; Figure AI, 2026a;b). Whole-body teleoperation provides visual observations paired with robot actions, but requires operators to guide a physical humanoid through each demonstration (Fu et al., 2025a; He et al., 2025; Luo et al., 2026b; GEAR Team et al., 2025; Li et al., 2026b). This dependence on robot availability and operation makes broader task and environment coverage costly to obtain (Shi et al., 2026; Punamiya et al., 2026).

Egocentric video records people’s everyday interactions across tasks and environments 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., 2026b; Wang et al., 2026a). Pretraining on these action-labeled videos can improve downstream manipulation (Zheng et al., 2026b; Punamiya et al., 2026). For humanoid loco-manipulation, however, wrist and hand trajectories alone omit the locomotion and postural changes that support whole-body interaction (Shi et al., 2026; Zhao et al., 2026a). Capturing synchronized body and hand motion preserves the demonstrated coordination, but using it for robot learning also requires accounting for differences in visual observations, body proportions, kinematics, and execution.

We study whether increasing robot-aligned whole-body human data improves humanoid control when robot supervision is held fixed. We develop a data pipeline that captures human demonstrations without a robot and aligns their observations and actions to the downstream humanoid (Sec. 3.4). Matching the capture camera and viewpoint, together with visual processing, brings human observations closer to the robot’s visual input. For actions, we preserve the demonstrated interaction targets while accounting for body proportions, robot kinematics, and controller dynamics to construct robot-compatible state-action supervision. We evaluate policies trained with this supervision on Unitree G1 in coordinated whole-body manipulation and long-range loco-manipulation (Sec. 5.1). Using task-specific human demonstrations without target-task robot demonstrations, the evaluations achieve mean progress of 51.7% and 76.2%, respectively.

Using this capture setup, we build HUMANVERSE, a 500-hour corpus of human loco-manipulation spanning open-world scenes and tasks. We record egocentric video and body motion with lightweight wearable sensors, synchronize the streams, and reconstruct hand motion from video. Collection requires no robot at the recording site, allowing people to demonstrate tasks directly in everyday environments.

We develop λ_0 , a whole-body humanoid vision-language-action policy that learns from these demonstrations (Fig. 1). We train the policy in three stages: egocentric hand-motion pretraining on diverse datasets, robot-aligned whole-body human mid-training on HUMANVERSE demonstrations, and robot post-training. We vary the amount of whole-body human data while holding robot demonstrations and the robot post-training budget fixed.

Increasing whole-body human data reduces validation loss and raises real-world task progress monotonically from 59.9% without whole-body mid-training to 80.4% with the full corpus. Ablations identify whole-body mid-training as the larger contributor to real-world success among the two human-data stages. The full λ_0 policy achieves 90.0% success on SIMPLE and 62.5% across four real-world loco-manipulation tasks, improving over the strongest evaluated baselines by 3.9% and 22.5%, respectively. Human-data training also improves progress on robot-unseen object instances observed in human demonstrations. Our contributions are:

- We develop a robot-aligned data pipeline and validate both efficient robot-free collection and zero-shot transfer on Unitree G1 across two representative task families: whole-body manipulation and long-range loco-manipulation. Transfer uses no target-task robot demonstrations.
- We introduce HUMANVERSE, a 500-hour dataset pairing egocentric video with synchronized body and hand motion across open-world scenes and tasks.
- We develop λ_0 through three-stage training and show that scaling whole-body human supervision improves downstream humanoid control under fixed robot supervision, supported by scaling, ablation, and generalization evaluations.

2 Related Work

2.1 Generalist Robot Policies

Large-scale robot datasets combine 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 scalable 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). UMI on Legs extends robot-free hand-held demonstrations to quadruped manipulation by coupling a task-level manipulation policy with a simulation-trained whole-body controller (Ha 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. Nymeria pairs synchronized multimodal egocentric recordings with full-body motion across 300 hours of daily activities in the wild (Ma et al., 2024), providing a foundation for studying context-dependent human motion. 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., 2026a). OpenEgo standardizes hand-pose annotations and temporally localized action descriptions across egocentric datasets (Jawaid & Xiang, 2025). 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., 2026b). HAT unifies human and humanoid state-action spaces for co-training on aligned PH2D demonstrations (Qiu et al., 2025), while DexWild combines wearable human interaction data with robot demonstrations for dexterous manipulation across environments (Tao et al., 2025). EgoZero learns closed-loop manipulation from smart-glasses demonstrations without robot data using morphology-agnostic representations (Liu et al., 2025), while HumanEgo uses hand-object interaction representations to learn robot-data-free policies from short human recordings (Wang et al., 2026b). Beyond stationary manipulation, EMMA co-trains human full-body mobile manipulation demonstrations with static robot data, avoiding mobile robot teleoperation and reporting improved performance as human data increases (Zhu et al., 2025). Building on these results, we study how explicit body-hand supervision supports coordinated bipedal humanoid interaction under a fixed robot-demonstration budget.

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). Motion retargeting is a key interface: GMR studies how retargeting artifacts affect downstream tracking (Araujo et al., 2026), and Human2Humanoid combines morphology-invariant trajectory alignment with physics-aware feasibility constraints (Huang et al., 2026). 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., 2026a; Wei et al., 2026b; Shi et al., 2026; Hu et al., 2026; Bjorck et al., 2025; Luo et al., 2026a). ZeroWBC generates and tracks whole-body motion from an initial egocentric image and language instruction without robot teleoperation demonstrations (Yang et al., 2026a); HALOMI couples egocentric demonstration capture with view alignment and controller-aware reference adaptation for active-perception loco-manipulation (Zhao et al., 2026a). Complementary simulation-mediated approaches include VideoMimic, which reconstructs human motion and scenes for contextual control (Allshire et al., 2025), and SUGAR, which physically refines human-video interaction priors before distilling autonomous policies (Wu et al., 2026). GRAIL instead synthesizes interactions from 3D assets and video priors to train visual policies through simulation (Xie et al., 2026). LEGS combines procedurally generated demonstrations with photorealistic Gaussian-splatting scene backgrounds to fine-tune humanoid loco-manipulation VLAs without teleoperation (Kim et al., 2026). Predictive models and integrated whole-body systems offer complementary ways to connect perception, motion, and

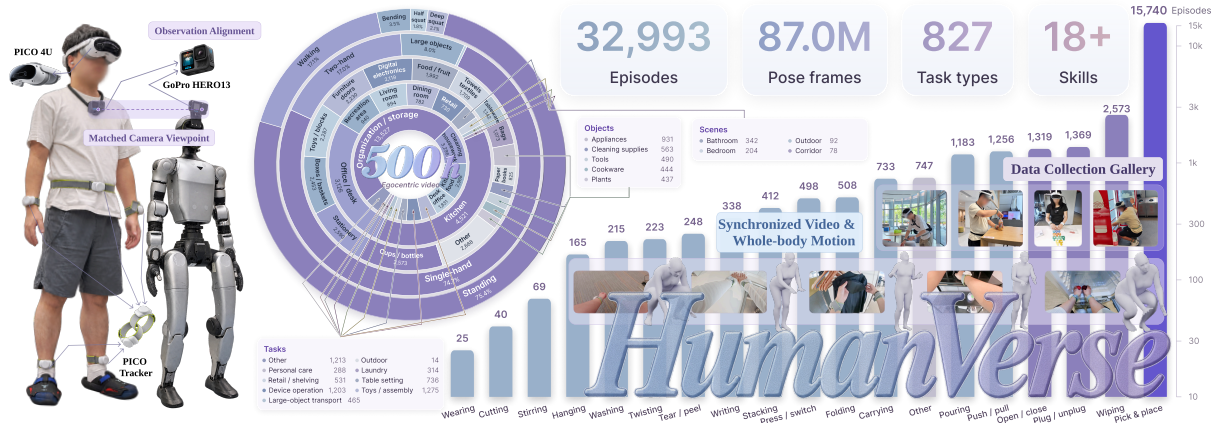


Figure 2: **HUMANVERSE**. The wearable capture setup, task and motion distributions, and paired egocentric video and whole-body motions illustrate the dataset’s scale and diversity.

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). Our focus is on collecting synchronized egocentric body–hand experience in open-world environments, aligning it for robot execution, and measuring the downstream benefit of scaling this supervision under a fixed robot-demonstration budget.

Supplementary Table 5 compares the human-data interfaces, training strategies, and scaling axes of related approaches.

3 HumanVerse-500: Scaling Egocentric Whole-Body Data Collection in the Open World

HUMANVERSE contains 500 hours of egocentric human loco-manipulation, comprising 32,993 episodes and 87 million pose frames (Fig. 2). The dataset covers 827 task types across diverse objects and scenes, pairing first-person video with synchronized body and hand motion. We organize the corpus around scalable acquisition, traceable motion supervision, and conversion to the downstream robot embodiment. We refer to the resulting training representation as *robot-aligned whole-body human data*; the 500-hour figure describes the recorded human corpus before conversion and validity filtering.

3.1 Robot-Free Capture and Collection Efficiency

We choose the GoPro HERO13 for convenient human data collection and equip our 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 reconstruct hand motion from the same video. Figure 3 summarizes the workflow from open-world scene and task selection to robot-free capture, data curation, and robot alignment.

Collection protocol. Collectors receive activity-level prompts rather than prescribed joint trajectories, allowing variation in walking paths, initial posture, object instances, placements, grasp strategies, and intermediate corrections. Each episode preserves the approach, repositioning, and transitions needed to complete an interaction, rather than only the final hand motion. Because capture requires no on-site robot, collectors carry the wearable interface into everyday locations and record open-world activity directly where it occurs. They are directed toward loco-manipulation activities—carrying, tidying, cleaning, and operating objects and fixtures—so that episodes stay manipulation-dense while preserving the natural walking and repositioning between interactions. Original streams are retained with collector and session identifiers, activity annotations, calibration, and device-health records; on-site checks flag interrupted recordings, missing modalities, and failed calibration.

Data collection efficiency. Section 5.1 compares collection effort and downstream performance on whole-body manipulation tasks, complemented by EgoAlign’s collection-time results for long-range object relocation.

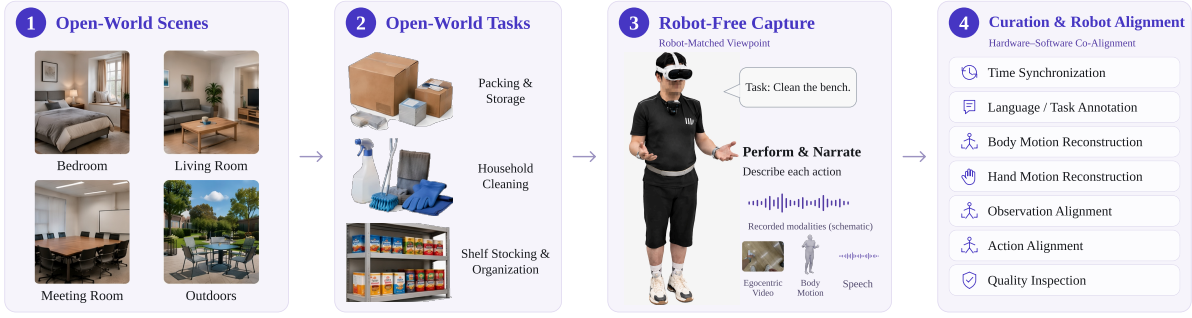


Figure 3: **Robot-free capture, curation, and robot alignment.** We select open-world scenes and tasks and record egocentric video, body motion, and speech with a robot-matched camera viewpoint. Curation includes synchronization, language/task annotation, body and hand motion reconstruction, observation alignment, action alignment, and quality inspection. Capture hardware and processing jointly support hardware–software co-alignment to the downstream robot. Modality thumbnails are schematic.

3.2 Dataset Scope and Coverage

The three stages use distinct data sources. Stage I draws on the seven heterogeneous egocentric datasets listed in the main paper. These public sources are not counted as part of HUMANVERSE. HUMANVERSE 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 duration and its scaling subsets. This separation distinguishes human-data scaling in Stage II from robot supervision in Stage III.

HUMANVERSE captures everyday activities across diverse open-world settings—including homes, retail and supermarket spaces, industrial sites, offices, and outdoor areas—covering 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 Supplementary Table 4 compares the dataset’s capture setup and motion coverage with existing egocentric datasets.

Qualitative task diversity. Figure 4 presents 60 egocentric frames illustrating everyday tasks, scenes, and manipulated objects. Each tile depicts one moment from a capture session. These examples illustrate qualitative diversity and do not estimate dataset-wide task frequencies.



Figure 4: **Representative examples from HUMANVERSE.** Sixty first-person frames are arranged in a 12×5 gallery to illustrate diversity in tasks, scenes, objects, and hand–object interactions.

3.3 Human Motion Reconstruction and Annotation

Temporal synchronization and language annotation. The GoPro and PICO run on independent clocks. We assign each episode a time offset under the convention $t_{\text{pico}} = t_{\text{gopro}} + \Delta$ and refine a capture-time prior by cross-correlating GoPro gyroscope measurements with tracker-derived body angular velocity. After applying this offset, video and body pose are resampled onto a common timeline over their overlapping interval, with synchronization quality retained as an explicit diagnostic. Spoken instructions

on the GoPro audio track are transcribed into episode-level language/task labels. The records retain stable frame indices for reconstruction and alignment.

Body and hand motion. We reconstruct a synchronized human motion sequence from wearable body tracking and egocentric video. The PICO system (PICO, n.d.) supplies timestamped 24-joint body kinematics, while Dyn-HaMR (Yu et al., 2025) estimates 21 joints per hand using camera motion from ViPE (Huang et al., 2025). Coordinate transforms register these trajectories to the synchronized video frames. These body and hand trajectories provide the reference for robot alignment in Sec. 3.4.

Coordinate representations. We retain the camera trajectory and available camera-to-world and body-frame transforms. Hand trajectories are stored both in current-camera coordinates for reprojection and in an episode-fixed frame anchored at the first valid camera pose to preserve displacement over time. Finger geometry is represented in wrist-local, palm-aligned, scale-normalized coordinates. These records support the query-time camera coordinates and relative wrist targets defined in Sec. 4.1.

Annotation records. Each derived motion sequence retains source-frame indices, coordinate transforms, conversion versions, per-track visibility, and estimator confidence. Raw recordings remain the canonical source. Separate body and hand validity masks preserve usable body motion when hands are missing or occluded, without filling missing poses by temporal interpolation. The body stream consists of wearable-device kinematics; the hand trajectories are reconstructed from RGB. Appendix D.1 describes their quality checks and data splits.

Figure 5 shows 16 examples pairing egocentric frames with their reconstructed motions.

3.4 Robot-Aligned Data Pipeline

Our robot-aligned data pipeline combines robot-free capture with hardware–software co-alignment to the downstream embodiment (Fig. 3). Capture hardware and processing are designed together: matched cameras and capture height reduce viewpoint differences, observation alignment addresses visual appearance, and action alignment constructs compatible robot states and executable motion targets. The two alignment dimensions are detailed below.

Action Alignment. We convert reconstructed human motion into G1 state–action supervision while retaining the demonstrated end-effector targets (Fig. 6). Human-reference alignment first adapts SMPL poses to robot body proportions. Robot-reference alignment then corrects residual end-effector errors using G1’s kinematic model. The resulting reference is encoded into SONIC motion tokens (Luo et al., 2026b), which dynamics-aware optimization refines using execution feedback. Targets from the original human sequence (Sec. 3.3) are registered in local coordinates and kept fixed across stages.

Human-reference alignment. Starting from the reconstructed SMPL sequence, we uniformly scale a copy of the human body to G1 height while retaining its joint angles and relative body proportions. The



Figure 5: **Synchronized human motion and egocentric observations.** Top: egocentric RGB frames. Bottom: corresponding SMPL body meshes with MANO hand motion and illustrative camera frusta. Camera frusta schematically illustrate viewing direction.

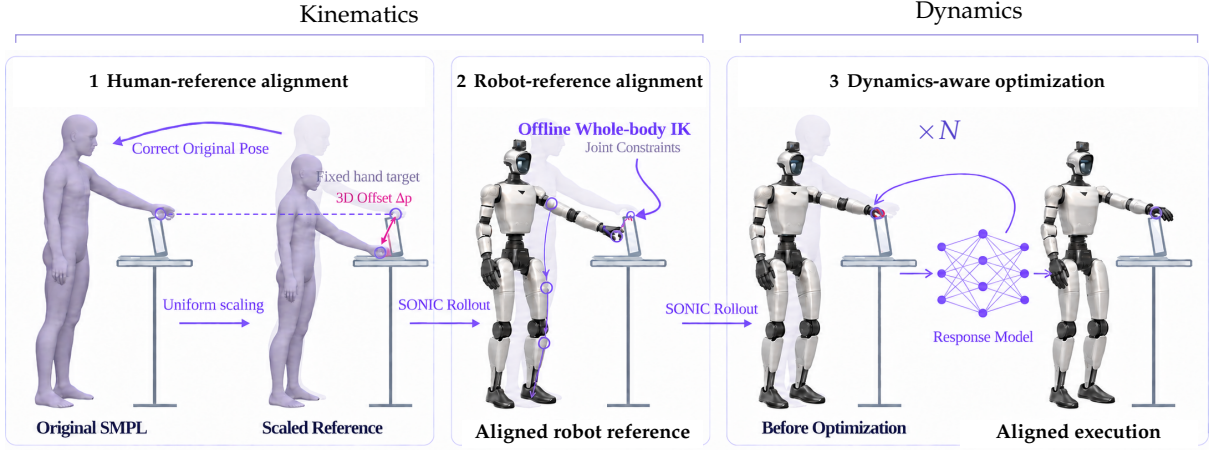


Figure 6: **Human-to-humanoid action alignment.** (1) Human-reference alignment adapts SMPL poses; (2) robot-reference alignment corrects robot joint trajectories. Both use kinematic corrections, followed by (3) dynamics-aware optimization of motion tokens using execution feedback. Human end-effector targets are retained throughout; poses and overlays are schematic.

demonstrated hand targets remain fixed in the interaction frame, so scaling can lower the hand below its intended contact position. The hand offset exposed by the scaled body guides bounded corrections to the original, unscaled SMPL pose through selected collar, shoulder, and elbow joints. Regularization limits the correction magnitude and its temporal variation, while root travel, lower-body motion, spine trajectories, and wrist-local rotations retain their recorded values. We correct the SMPL pose sequence, encode it, and execute the motion in simulation:

$$\tilde{\mathbf{q}} = \mathbf{q} + \Delta\mathbf{q}, \quad \mathbf{b}^0 = \mathcal{S}(\mathcal{E}_h(\tilde{\mathbf{q}})). \quad (1)$$

Here, $\mathbf{q}$ and $\tilde{\mathbf{q}}$ denote the original and corrected SMPL pose sequences, and $\Delta\mathbf{q}$ contains the corrections to the selected upper-body joints. SONIC’s human-motion encoder $\mathcal{E}_h$ converts the corrected human motion into motion tokens. The simulation operator $\mathcal{S}$ executes these tokens in MuJoCo through the frozen SONIC decoder in closed loop, returning the actual robot joint trajectory $\mathbf{b}^0$. This trajectory initializes robot-reference alignment with G1’s actual geometry.

Robot-reference alignment. We refine the simulated joint trajectory using constrained whole-body inverse kinematics (IK) with G1’s actual kinematic model. The optimization reduces task-relevant end-effector errors while respecting joint and correction bounds, limiting deviations from $\mathbf{b}^0$, and preserving temporal smoothness:

$$\mathbf{b}^{\text{ref}} = \text{IK}(\mathbf{b}^0, \mathbf{p}), \quad \mathbf{z}^0 = \mathcal{E}_r(\mathbf{b}^{\text{ref}}). \quad (2)$$

Here, $\mathbf{b}^{\text{ref}}$ is the corrected kinematic reference trajectory, and $\mathbf{p}$ denotes the fixed human end-effector targets registered in compatible local coordinates. The targets specify task-dependent end-effector positions and orientations where available. SONIC’s robot-motion encoder $\mathcal{E}_r$ converts $\mathbf{b}^{\text{ref}}$ into motion tokens $\mathbf{z}^0$, which initialize dynamics-aware optimization to address errors during controller execution.

Dynamics-aware optimization. We replay $\mathbf{z}^0$ and optimize motion tokens to reduce the discrepancy between realized end effectors and the fixed human targets. At iteration k , a rollout provides endpoint residuals $\mathbf{r}^k$ and controller observation history $\mathbf{s}^k$. With this history held fixed, the frozen SONIC decoder maps candidate tokens to changes in joint commands. A local kinematic Jacobian and a learned response model predict how those changes affect executed endpoints; token optimization seeks a predicted response that compensates for $\mathbf{r}^k$, with regularization limiting the update. Gradients pass through the fixed decoder and response model, while each candidate is evaluated in a fresh simulation rollout. Candidate selection retains earlier solutions, including the initialization, and the selected sequence undergoes independent replay to check endpoint tracking and stability.

Aligned state-action supervision. The accepted replay pairs the selected body-motion tokens with the robot states realized under the same controller history. Finger motion from Sec. 3.3 is retargeted to the bilateral BrainCo Revo2 hands (BrainCo, n.d.). The tokens are encoded from the temporal motion sequence, preserving its evolution throughout the demonstration. We associate the body tokens, hand commands, and realized states with the source video frames and retain their validity masks, yielding

the human-derived Stage-II supervision (Sec. 4.2). Supervision is contributed only by episodes that pass acceptance and independent replay after dynamics-aware optimization; failed conversions and invalid targets are masked out rather than imputed, so HUMANVERSE’s recorded duration upper-bounds rather than equals the usable supervision.

Observation Alignment. We reduce the visual gap between human demonstrations and robot deployment through robot-arm rendering and training-time image augmentation (Chen et al., 2026). The shared camera setup and matched camera height reduce the initial viewpoint difference. We use SAM2 to segment human forearms and wrist-mounted PICO trackers, remove them with DiffuEraser, and render G1 arms from the aligned robot joint positions. Translation, rotation, and uniform scale are adjusted to match the human wrists and forearms before the rendered arms are composited onto the inpainted backgrounds, retaining the surrounding scene and demonstrated interaction. During training, random resized crops, small rotations, and photometric jitter account for residual viewpoint and appearance differences caused by human and robot motion. The processed frames remain paired with the aligned state-action supervision. At deployment, the policy takes robot camera images directly.

Data quality and splits. We partition HUMANVERSE by recording session before extracting temporal windows and draw nested scaling subsets only from the frozen general-purpose training partition. Task-specific human demonstrations for downstream post-training remain separate from this corpus, and human-data selection does not consult the robot evaluation manifest. Appendix D.1 details the quality checks, validity masking, and split protocol.

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 egocentric hand-motion pretraining, robot-aligned whole-body human mid-training, and robot post-training (Fig. 7). The robot-compatible human supervision constructed in Sec. 3.4 connects the dataset to Stage II: it supplies whole-body motion targets in the downstream embodiment’s format before native robot adaptation in Stage III.

4.1 Policy Architecture and Action Interfaces

Architecture and inputs. λ_0 combines a Qwen3.5-2B (Qwen Team, 2026) vision-language backbone with a generative action expert. In Stages II and III, the policy takes an egocentric image, a language instruction, and the current state. For G1, the 44-dimensional state contains 29 body-joint positions, six hand commands per side, and three projected-gravity components. Following $\pi_{0.5}$ (Black et al., 2025), we discretize the state and append it to the instruction. The backbone supplies layer-wise features to the action expert.

Action outputs and domain interfaces. The action expert predicts a 50-step action chunk; for G1, each step contains a 64-dimensional SONIC motion latent and two 6-dimensional Revo2 hand commands, totaling 76 dimensions. Stage I instead predicts future discrete bilateral hand-state tokens using only the backbone. The action expert is introduced in Stage II and shared with Stage III. Human-derived mid-training data and native G1 post-training data use the same state-action format (Appendix B.1) but distinct domain identifiers and input/output projections. Following X-VLA (Zheng et al., 2025), domain-specific projections and learned soft prompts support cross-domain transfer.

4.2 Three-Stage Training

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

Stage I predicts future hand-state tokens in the vision-language backbone. Stages II and III use the masked velocity-field loss $\mathcal{L}^q$ defined in Sec. 4.3, with different data sources and active action interfaces. Let $\phi^{(1)}$ denote the Stage-I backbone parameters and $\theta^{(k)}$ the action-policy parameters learned in Stages II and III. Stage II initializes its backbone from the Stage-I EMA checkpoint and introduces the action expert. Stage III initializes from the Stage-II EMA checkpoint.

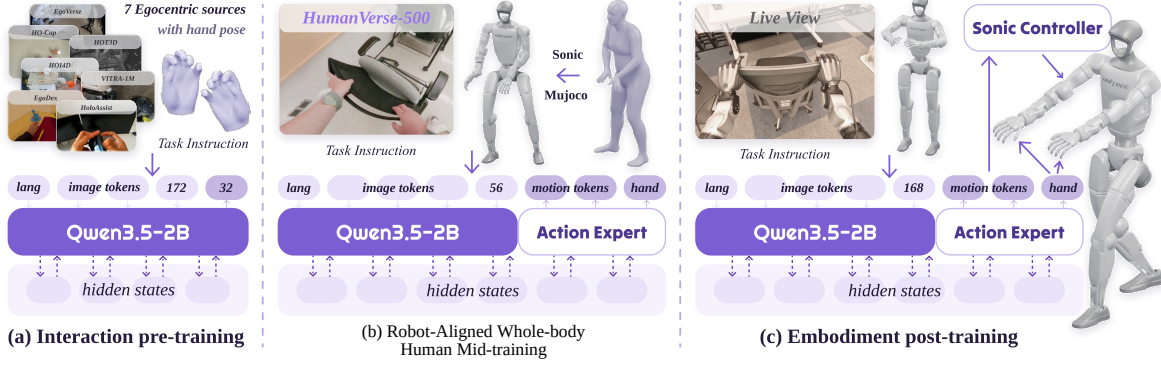


Figure 7: **Three-stage training of λ_0 .** The policy learns interaction from egocentric data, whole-body coordination from HUMANVERSE, and downstream control from real-robot data.

Stage I: egocentric hand-motion pretraining. We initialize the vision-language backbone from a vision-language checkpoint and train it to predict a future sequence of discrete bilateral hand-state tokens from egocentric observations, the current hand state, and language (Fig. 7(a)). No action expert is trained in this stage. 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:

$$p_1(d) = \frac{\sqrt{N_d}}{\sum_{d' \in \mathcal{S}_E} \sqrt{N_{d'}}}, \quad \phi^{(1)} = \arg \min_{\phi} \mathbb{E}_{d \sim p_1, \tau \sim \mathcal{D}_d} [\mathcal{L}_{\text{token}}^E(\tau; \phi)]. \quad (3)$$

Here $\mathcal{L}_{\text{token}}^E$ is the future hand-state token prediction loss. Each continuous bilateral hand state has 138 dimensions before discretization: 3D wrist position, 6D wrist rotation, and 20×3 local hand-keypoint coordinates per hand.

Stage II: robot-aligned whole-body human mid-training. During mid-training, HUMANVERSE demonstrations are converted into the G1 state-action format (Fig. 7(b)). The robot-aligned data pipeline combines robot-free capture with hardware-software co-alignment to the downstream whole-body embodiment. Observation alignment combines matched camera configuration and viewpoint with visual processing; action alignment converts human motion into robot-compatible states and action targets through reference correction and execution refinement. Stage II learns from this aligned interface before robot post-training. The pipeline separates human-data acquisition from the target embodiment; here its concrete realization uses the G1 observation and action spaces. This uses the body-motion and hand-command conversion in Sec. 3.4; the supervision remains human-derived rather than native robot experience. Each 76-dimensional action comprises a 64-dimensional SONIC motion token and two 6-dimensional Revo2 hand commands. Human-derived and native robot data use distinct domain identifiers and separate input and output projections for scale differences. Denoting the conversion by Φ_R and the human-domain loss by $\mathcal{L}^H$, we initialize the backbone from Stage I, introduce the action expert and human-data projections, and optimize

$$\theta^{(2)} = \arg \min_{\theta: \phi \leftarrow \text{EMA}(\phi^{(1)})} \mathbb{E}_{\tau \sim \Phi_R(\mathcal{D}_W)} [\mathcal{L}^H(\tau; \theta)]. \quad (4)$$

We update the backbone, action expert, and human-data projections during this stage.

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. 7(c)). The post-training objective is

$$\theta^{(3)} = \arg \min_{\theta \leftarrow \text{EMA}(\theta^{(2)})} \mathbb{E}_{\tau \sim \mathcal{D}_R} [\mathcal{L}^R(\tau; \theta)]. \quad (5)$$

We jointly optimize the model, embodiment conditioning, and active interfaces.

4.3 Training Objective and Deployment

Masked velocity-field objective for Stages II and III. Stages II and III use flow matching following $\pi_{0.5}$ (Black et al., 2025). Our loss $\mathcal{L}^q$ averages squared prediction error over valid target entries, excluding missing annotations and any provided action prefix. Inference uses 10 Euler steps from Gaussian noise to actions, keeping any provided action prefix fixed.

Deployment. At deployment, λ_0 uses G1’s egocentric image, language instruction, and proprioceptive state with the robot-domain identifier and projections (Fig. 7(c)). The predicted 64-dimensional motion token drives the SONIC whole-body controller, and the two 6-dimensional hand commands drive the Revo2 hands. Deployment requires no human motion streams.

5 Experiments on Humanoid: Alignment, Scale, and Transfer

We first validate aligned human supervision across whole-body manipulation and long-range loco-manipulation (Sec. 5.1), then evaluate λ_0 ’s three-stage recipe under the task and scoring protocols below. We organize the evaluation around three questions: whether human supervision advances humanoid capabilities, how its scale and training stages shape learning, and whether these learning gains translate into better control and generalization. We ask: **Q1.** *Can human supervision advance humanoid loco-manipulation?* (Sec. 5.2); **Q2.** *How do the scale and training stages of human supervision shape humanoid learning?* (Sec. 5.3); **Q3.** *Can human data improve control and generalization?* (Sec. 5.4).

5.1 Effectiveness and Efficiency of Robot-Aligned Human Data

We assess robot-aligned human supervision along two dimensions: the behaviors it enables without target-task robot demonstrations, and the collection effort needed to learn them. EgoHumanoid-V2 (Chen et al., 2026) and EgoAlign (Jiang et al., 2026b) provide complementary evidence across whole-body manipulation and long-range loco-manipulation.

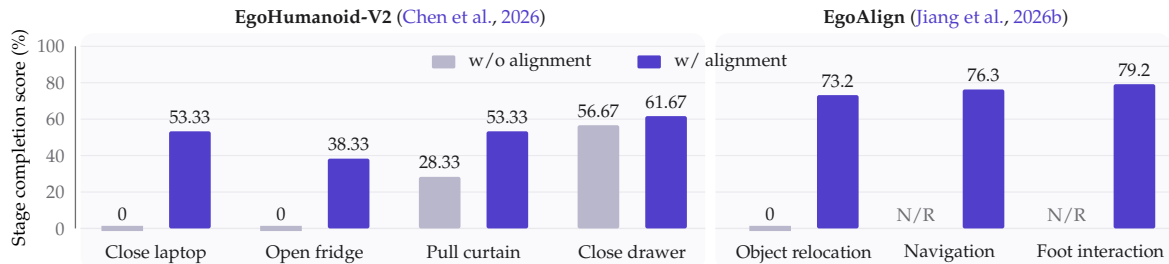


Figure 8: **Task progress from aligned human demonstrations.** Scores include partial completion; long-range results average seen and unseen target locations. N/R indicates missing before-alignment scores.

Effectiveness across task families. Figure 8 covers four whole-body tasks and three long-range tasks, with mean progress of 51.7% and 76.2%, respectively. Paired whole-body comparisons show the benefit of alignment, while long-range evaluations cover relocation, navigation, and foot interaction at seen and unseen target locations. These results motivate the subsequent training and scaling studies.

Performance relative to collection effort. Figure 9 shows competitive robot performance with substantially lower human-data collection costs. In EgoHumanoid-V2 (Chen et al., 2026), averaging Close Laptop and Pull Curtain, human collection reduces mean labor per demonstration by 76.4%; policies trained on 200 human demonstrations reach a mean task score of 53.3%, compared with 60.0% from 50 robot demonstrations. For long-range object relocation, EgoAlign (Jiang et al., 2026b) reduces labor per demonstration by 90.4% (25 versus 260 person-seconds, with one person for human capture and two for teleoperation), including resets.

5.2 Q1: Can Human Supervision Advance Humanoid Loco-Manipulation?

We evaluate whether the human-supervised training recipe improves humanoid loco-manipulation across simulation and real-world tasks by comparing λ_0 with established policy baselines.

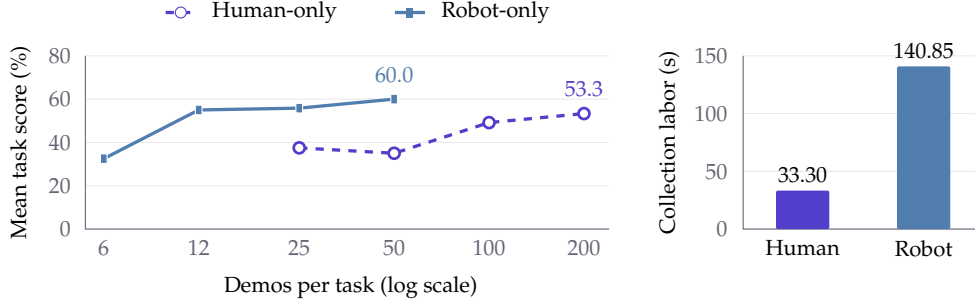


Figure 9: **Task performance and collection labor with human or robot supervision.** Equal-weight means over Close Laptop and Pull Curtain. Left: task scores versus demonstrations per task, evaluated over 20 trials per policy. Right: person-seconds per demonstration, including capture, amortized setup, calibration, troubleshooting, and resets, with one person for human capture and two for teleoperation.

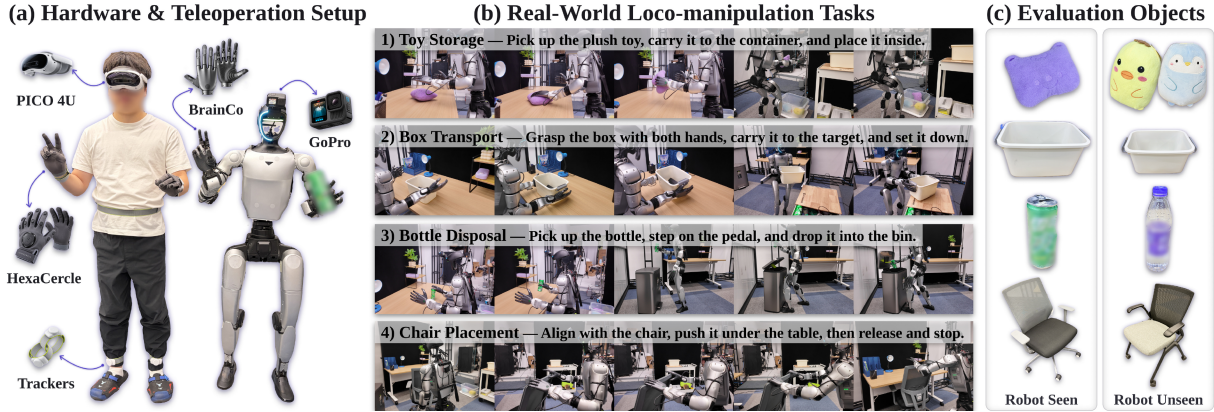


Figure 10: **Real-world data collection and evaluation.** (a) Egocentric teleoperation collects whole-body demonstrations. (b) Four loco-manipulation tasks evaluate task success. (c) Robot-seen and robot-unseen objects are used for standard task and generalization evaluation, respectively.

Hardware. A PICO 4 Ultra with five trackers captures body motion, which SONIC converts into Unitree G1 control signals (Fig. 10(a)). We use HexaCercle M11 gloves to capture hand poses robustly under occlusion and map the captured finger motion to control the BrainCo Revo2 hands. A head-mounted GoPro provides egocentric RGB, synchronized with whole-body demonstrations on G1’s computer.

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. 10(b), with ten frozen initial configurations per task and policy. Fixed initial conditions support consistent policy comparisons. We report task success and progress to capture full and partial completion. Progress sums fixed milestone scores, normalized to 100 per task and averaged equally across tasks. Standard trials use robot-seen objects in Fig. 10(c). Generalization uses objects seen 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.

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%, respectively. It achieves the highest success count on all four tasks, with the largest progress gain on bottle disposal (51.0%), and keeps progress above 75% on every task. Chair placement separates the two metrics: λ_0 completes 6/10 trials to StarVLA’s 3/10 yet trails on progress (77.0% versus 82.5%).

Simulation evaluation. Table 2 reports 90.0% overall success for λ_0 , exceeding StarVLA and Ψ_0 by 3.9% and 4.4%, respectively. It leads StarVLA at every level (90.0%, 93.3%, and 86.7% at Levels 0–2), though

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

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

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

Model	XMovePick	BendPick	Handover	Mobile P&P	Grasp	XMoveBendPick	Average	Overall
H-RDT (Bi et al., 2026)	0/0/20	0/0/10	0/10/0	0/0/0	0/0/0	0/0/0	0.0/1.7/5.0	2.2
InternVLA-M1 (Chen et al., 2025)	0/0/0	50/50/0	0/0/0	0/0/0	0/0/0	30/50/70	13.3/16.7/11.7	13.9
EgoVLA (Yang et al., 2026b)	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

the margin narrows to 1.7% at Level 2. It reaches 100% on BendPick at all levels and on Handover at Levels 1 and 2; Mobile P&P is the weakest case, trailing StarVLA at Levels 1 and 2 (70% versus 80%).

5.3 Q2: How Do the Scale and Training Stages of Human Supervision Shape Humanoid Learning?

We examine how the amount and staging of human supervision affect learning. Controlled Stage-II scaling experiments measure validation loss and training dynamics, while stage ablations assess the contribution of each human-data stage to downstream performance.

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.

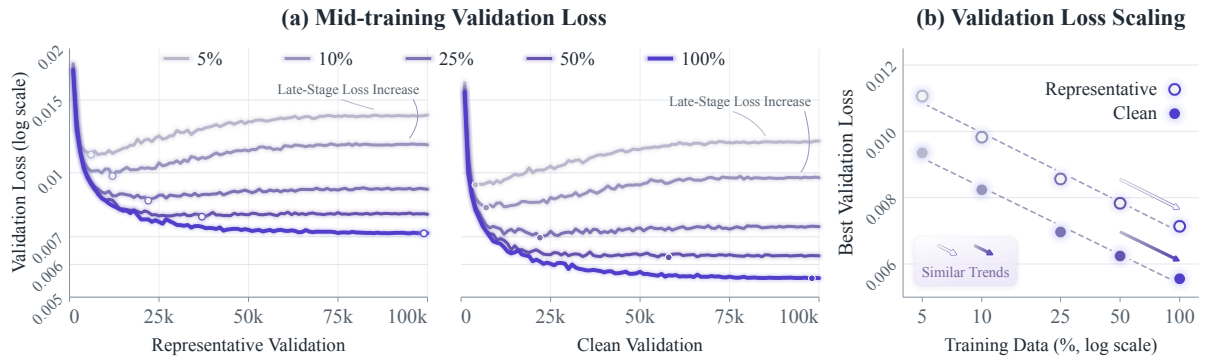


Figure 11: **Human-data scaling of validation loss.** (a) Training curves on representative and clean holdouts at five data fractions. (b) Minimum validation loss versus data scale with descriptive log-linear fits over the measured range. Minima are selected independently on each holdout.

Training dynamics. Fig. 11(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 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. 11(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 $\hat{L}_h(D) = a_h + b_h \ln D$ 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 far more than removing Stage I (40.0% versus 12.5%): whole-body mid-training contributes more than the interaction priors 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: Can Human Data Improve Control and Generalization?

We test whether human-data training translates into better downstream control and generalization. We measure task progress as Stage-II data increases, then evaluate human-guided transfer to instances absent from robot demonstrations.

Downstream task-progress scaling. Fig. 12 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% from no Stage II to the full corpus.

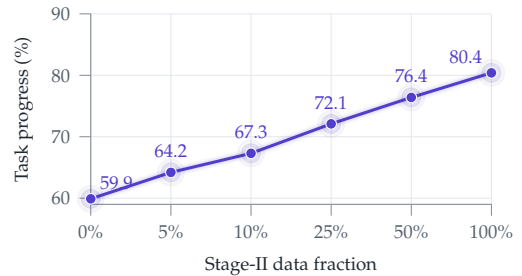


Figure 12: Average real-world task progress versus Stage-II human-data fraction.

Generalization with human demonstrations. To assess generalization, we evaluate whether human demonstrations improve performance on object instances absent from robot demonstrations across the four real-world tasks. For each method, we train a policy using robot demonstrations of object instances A and human demonstrations of A and B. Thus, B is observed in human demonstrations but absent from robot demonstrations. *Seen* (A, A, B) and *Human-guided* (A, A, B) evaluate this same policy on A and B, respectively. *Robot-only* (A, B) evaluates a robot-only variant on B under matched initializations and scoring; unlike Sec. 5.1, it uses robot demonstrations of other objects within the same task.

λ_0 achieves the highest mean *Human-guided* progress at 55.8%, compared with 46.5% for the strongest baseline, $\pi_{0.5}$ (Fig. 13). Compared with its *Robot-only* variant, λ_0 improves progress on all four tasks, increasing the average from 48.4% to 55.8%, a gain of 7.4%. These results support human-data training as a means of improving performance on robot-unseen, human-observed object instances, although performance remains below the *Seen* setting. The comparison measures 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, nor does it vary human-data scale.

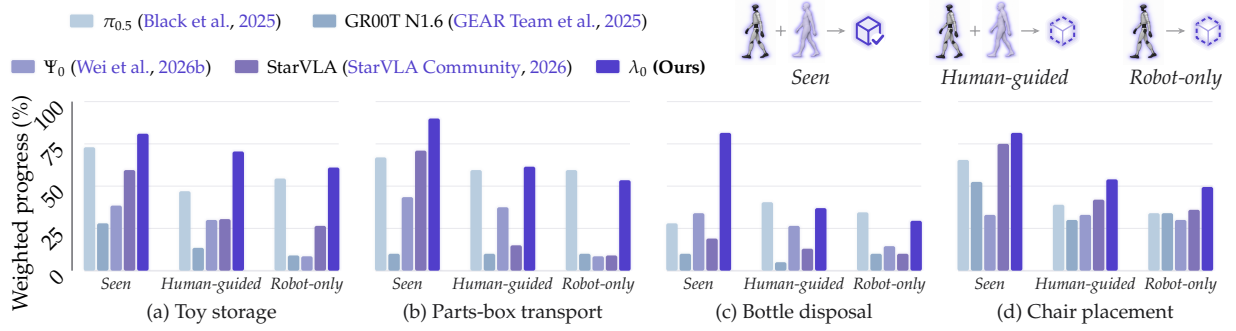


Figure 13: **Generalization with human demonstrations.** Seen and Human-guided evaluate the same policy, trained on robot demonstrations of A and human demonstrations of A and B, on A and B respectively. Robot-only evaluates a policy trained without human data, using robot demonstrations of A, on B. B is absent from robot demonstrations in both training recipes; it is human-observed only in the joint-data recipe.

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, a 500-hour dataset of diverse open-world human activities with synchronized egocentric video, body motion, and hand motion. We document collection throughput and data-quality controls, while our human-data validation achieves 51.7% and 76.2% progress on whole-body manipulation and long-range loco-manipulation, respectively, without target-task robot demonstrations. For λ_0 , robot-compatible supervision derived from HUMANVERSE supports robot-aligned whole-body human mid-training before native robot adaptation. Experiments on SIMPLE and four real-world tasks demonstrated improved performance on loco-manipulation tasks, including evaluations with robot-unseen, human-observed object instances. 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.

7 Contributions

Core Contributors.

Contributors.

Advisors.

References

- Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025. URL <https://arxiv.org/abs/2505.03729>.
- Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026.
- 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. Revo2: 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/iroso60139.2025.11247088.

- Jin Chen, Yiming Jiang, Chongyang Xu, Modi Shi, Shijia Peng, Li Chen, Tianyu Li, Mu Xu, Yilun Chen, Steven Hoi, and Hongyang Li. EgoHumanoid-V2: Human-to-humanoid transfer of coordinated whole-body skills for loco-manipulation. *arXiv preprint arXiv:2609.37181*, 2026. URL <https://arxiv.org/abs/2609.37181>.
- 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.
- Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. UMI on Legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024. URL <https://arxiv.org/abs/2407.10353>.
- 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/>.
- Tianchen Huang, Feiyang Yuan, Junchi Gu, Shurui Fang, Xiaohu Zhang, Yu Wang, Wei Gao, and Shiwu Zhang. Human2Humanoid: Physics-aware cross-morphology motion retargeting for humanoid robots. *arXiv preprint arXiv:2606.03476*, 2026. URL <https://arxiv.org/abs/2606.03476>.
- 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.
- Ahad Jawaaid and Yu Xiang. OpenEgo: A large-scale multimodal egocentric dataset for dexterous manipulation. *arXiv preprint arXiv:2509.05513*, 2025. URL <https://arxiv.org/abs/2509.05513>.
- 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*, 2026a.
- Yiming Jiang, Jin Chen, Chongyang Xu, Yilun Chen, Aimin Hao, and Yisheng He. EgoAlign: Bridging the human-humanoid gap for long-range loco-manipulation. *arXiv preprint arXiv:2609.38046*, 2026b. URL <https://arxiv.org/abs/2609.38046>.
- 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.
- Hojune Kim, Timothy Chen, Jiankai Sun, Lars W. Osterberg, Qianzhong Chen, Ke Wang, and Mac Schwager. LEGS: Fine-tuning teleop-free VLAs for humanoid loco-manipulation in an embodied gaussian splatting world. *arXiv preprint arXiv:2606.01458*, 2026. URL <https://arxiv.org/abs/2606.01458>.
- 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.
- Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025. URL <https://arxiv.org/abs/2505.20290>.
- 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):eae4592, 2026b. doi: 10.1126/scirobotics.aed4592.
- Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. URL <https://arxiv.org/abs/2406.09905>.
- 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.
- Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy $\sim$ human policy. *arXiv preprint arXiv:2503.13441*, 2025. URL <https://arxiv.org/abs/2503.13441>.

- 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.
- Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. DexWild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025. URL <https://arxiv.org/abs/2505.07813>.
- 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*, 2026a.
- Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. HumanEgo: Zero-shot robot learning from minutes of human egocentric videos. *arXiv preprint arXiv:2605.24934*, 2026b. URL <https://arxiv.org/abs/2605.24934>.
- 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.
- Tianshu Wu, Xiangqi Kong, Yue Chen, Qize Yu, Hang Ye, Jia Li, Yizhou Wang, and Hao Dong. SUGAR: A scalable human-video-driven generalizable humanoid loco-manipulation learning framework. *arXiv preprint arXiv:2605.20373*, 2026. URL <https://arxiv.org/abs/2605.20373>.

- Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026. URL <https://arxiv.org/abs/2606.05160>.
- Haoran Yang, Jiacheng Bao, Yucheng Xin, Haoming Song, Yuyang Tian, Bin Zhao, Dong Wang, and Xuelong Li. ZeroWBC: Learning natural whole-body humanoid interaction from human egocentric data. *arXiv preprint arXiv:2603.09170*, 2026a. URL <https://arxiv.org/abs/2603.09170>.
- 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*, 2026b. 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.
- Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025. URL <https://arxiv.org/abs/2510.10274>.
- Ruijie Zheng et al. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. In *Proceedings of the Conference on Robot Learning*, 2026b. To appear.
- Lawrence Y. Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. EMMA: Scaling mobile manipulation via egocentric human data. *arXiv preprint arXiv:2509.04443*, 2025. URL <https://arxiv.org/abs/2509.04443>.
- 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. Dataset construction is presented in Sec. 3, policy architecture and action interfaces in Sec. 4.1, and evaluation settings and results in Sec. 5. The supplementary material below provides data-quality and split-validation details, additional training formalism and implementation details, baseline implementations, and validation-scaling analysis details.

A Dataset and Related-Work Comparisons

Dataset scope, motion reconstruction, and robot alignment are described in Sec. 3. The following comparisons summarize the data interfaces and experimental questions of related resources and methods.

A.1 Egocentric Dataset Comparison

Table 4 compares capture setups, motion coverage, and documented humanoid interfaces.

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 (Ours)	500	Tracked full body and reconstructed hands	Compatible	SONIC body tokens and G1 training views	Portable PICO + GoPro

^aNymeriaPlus upgrades the 300+ h Nymeria corpus. Its retained duration after sequence filtering is not separately specified here. ^bThe 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.

A.2 Related-Method Comparison

Table 5 compares the human supervision, training strategies, and controlled scaling axes of related approaches.

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., 2026a)	300 h action-free egocentric video	latent locomotion pre-training, robot manipulation latents, and teleop grounding	0–100% video $\times$ 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?
$\omega-0$ (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

The interfaces below specify the hand-state tokens used in Stage I and the G1-compatible supervision used in Stages II and III. Architecture and training details complement Sec. 4.

B.1 State–Action Definitions

Egocentric hand-state token interface. For hand $\chi \in \{L, R\}$, let $\mathbf{p}_t^\chi \in \mathbb{R}^3$ and $\mathbf{R}_t^\chi \in \text{SO}(3)$ 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

$$\mathbf{h}_t = [\mathbf{p}_t^L, \rho_6(\mathbf{R}_t^L), \mathbf{g}_t^L, \mathbf{p}_t^R, \rho_6(\mathbf{R}_t^R), \mathbf{g}_t^R] \in \mathbb{R}^{138}. \quad (6)$$

We follow the state discretization of $\pi_{0.5}$ (Black et al., 2025): each normalized scalar is assigned to one of 256 uniform bins over $[-1, 1]$, and the bin indices are serialized as text and passed through the vision–language tokenizer. We apply this rule to the current hand state in the input and to each hand state in the future sequence used as the Stage-I prediction target. Stage I predicts future hand-state tokens without an action expert or continuous action head.

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 $\mathbf{g}_t^{\text{proj}} \in \mathbb{R}^3$, while hand retargeting provides six normalized Revo2 commands per side. Its interface is

$$\begin{aligned} \mathbf{s}_t^R &= [\mathbf{q}_t^{1:29}, \mathbf{u}_t^L, \mathbf{u}_t^R, \mathbf{g}_t^{\text{proj}}] \in \mathbb{R}^{44}, \\ \mathbf{a}_{t+h}^R &= [\mathbf{z}_{t+h}, \mathbf{u}_{t+h}^L, \mathbf{u}_{t+h}^R] \in \mathbb{R}^{76}. \end{aligned} \quad (7)$$

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.2 Model Architecture

Backbone and action expert. The Qwen3.5-2B backbone supplies projected features to corresponding layers of the shared action expert. Domain-specific input projections encode noisy actions and sinusoidal noise embeddings, while output projections predict action-space velocities. Following X-VLA (Zheng et al., 2025), we use embodiment-conditioned soft prompts; our configuration uses 32 prompt tokens and a learned embodiment bias.

Action prediction. Stages II and III use flow matching following $\pi_{0.5}$ (Black et al., 2025). The loss averages squared velocity error over valid scalar targets, excluding missing annotations, padding, and action prefixes, without action-block reweighting. For real-time chunking (RTC), half of the training samples have no prefix; the remainder use up to eight valid prefix steps, kept noise-free and excluded from the loss. Each observation uses four independent noise realizations during mid-training and post-training. Stage I predicts discrete hand-state tokens within the backbone.

B.3 Training Configuration

Table 6 summarizes the default 2B training schedule. The EMA is updated after each optimizer step with decay 0.999; the Stage-I backbone EMA initializes Stage II, and the Stage-II policy EMA initializes Stage III.

Table 6: Default training schedule for the 2B model. Stages II and III predict 50-step action chunks.

Stage	Optimizer updates	Updated modules
I	200k	Vision-language backbone
II	100k	Backbone, action expert, human-data projections
III	50k	Backbone, action expert, robot interfaces and conditioning

Stage I 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. Data sources and square-root sampling are defined in Sec. 4.2. 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). Stage-II supervision is constructed offline using the frozen SONIC controller in a G1 MuJoCo environment and bilateral hand retargeting. Stage III uses native robot demonstrations, with the same state-action format but distinct domain identifiers and projections.

C Evaluation Protocols

C.1 Baseline Implementations

We adapt the four baselines to the G1 interface in Eq. 7, using absolute actions sampled at 50 Hz and a 50-frame future buffer. Dataset-specific normalization uses training statistics. Each baseline retains its upstream training implementation and checkpoint format.

GR00T N1.6. We use the official Eagle vision-language backbone and N1.6 flow action expert with its 128-dimensional action-expert capacity. The action projector is reinitialized when its dimensions change. Training uses a 50-step action horizon, learning rate 10^{-4} , and effective batch size 64. 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 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, an effective batch size of 32, and the native action-sampling code.

StarVLA. We use Qwen3-VL-4B with the QwenPI-v3 layerwise flow-matching action model, a 50-step action horizon, four noise realizations per observation, and ten inference steps. The released configuration uses 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 a checkpoint is supplied.

C.2 Simulation Evaluation

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. 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 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. 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 the same task-wise summary. 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.

C.3 Real-World Evaluation and Scoring

Evaluation setup. Each policy is evaluated on ten fixed initial configurations per task, with shared objects, target regions, instructions, timeouts, and success criteria. Configuration order is randomized per policy; trial videos and state logs retain identifiers, while manual adjudicators are blinded to model identity. We report success counts and mean progress over the ten trials, averaging equally across tasks without pooling training seeds or reporting seed-based confidence intervals. This protocol applies to the baseline comparisons and training-stage ablations in Tables 1 and 3.

Progress and autonomous success. Progress sums the points for completed stages in Table 7, totaling 100 per task; each stage is scored independently of preceding stages, with the same weights across methods and data scales. Contact alone does not count as a stable grasp, and lifting requires clearance from the supporting surface. For bottle disposal, approach requires the hand to enter the grasp region, arrival requires a pose suitable for pedal operation and disposal, and pedal contact scores only if the lid opens enough to admit the bottle. The five-point stability stage requires three seconds of unsupported standing without unintended collision. Full-task success is recorded separately and requires the table’s terminal condition to hold for three seconds. Human contact with the robot, object, or environment after execution begins disqualifies autonomous success; assisted continuations are reported separately, and safety and intervention are logged for every task.

Table 7: 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 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

D Human-Data Scaling and Ablation

D.1 Data Quality and Scaling Protocol

Data quality. A frozen manifest tracks each episode from capture through reconstruction and robot conversion. It records the participant, session, environment, activity, duration, frame count, body-tracking coverage, per-hand visibility, coordinate metadata, conversion version, and train or validation assignment.

Structural checks require monotonic timestamps, overlapping video and tracking intervals, declared coordinate conventions, consistent stream lengths, and finite unmasked values. Geometric checks reproject reconstructed hands into RGB; a synchronized viewer supports manual inspection of images, camera-frame and episode-frame hand trajectories, and body motion. Reprojection error, estimator confidence, tracking coverage, and conversion failures remain explicit diagnostic fields. The predefined data contract determines inclusion: invalid target dimensions are masked, while conversion failures are recorded. These checks assess data integrity; synchronized visualizations provide qualitative reconstruction diagnostics. Dataset summaries report duration, episode count, participant and environment counts, activity distribution, bilateral-hand visibility, tracking coverage, and conversion failure rate.

Data splits and validation. We partition HUMANVERSE by recording session before extracting temporal windows, so all windows from an episode inherit the same split. Task-specific human demonstrations collected for downstream post-training remain separate from the general-purpose mid-training corpus. Session, object-instance, and evaluation-configuration identifiers track experimental separation; human-data selection does not consult the robot evaluation manifest.

Clean contains held-out sessions whose frame-weighted validity ratios for both SONIC whole-body and bilateral hand annotations reach their respective upper-quartile thresholds in the original split. Here, *clean* refers to annotation validity rather than appearance or task difficulty. *Representative* is stratified by data source and quartile bins of the two validity ratios. The 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 data fractions.

Scaling protocol. The human-data scaling axis is expressed as a fraction of the frozen HUMANVERSE training partition, with validation sessions excluded before subset construction. The 0%, 5%, 10%, 25%, 50%, and 100% conditions define the evaluation grid for Fig. 12. Nonzero subsets are nested, sampled at the episode level, and stratified by activity and recording session before chunk extraction. The Stage-I checkpoint, robot post-training set, Stage-III update budget, validation sets, 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.

Fig. 11 reports the 2B model at the five nonzero fractions, each trained for 100k updates with seed 43. Total validation loss is evaluated every 1k updates with validation seed 20260813.

Following EgoScale (Zheng et al., 2026b), we analyze validation dynamics and minimum loss separately. Each holdout’s minimum is selected independently over the full 0–100k-update window, so the selected steps may differ. We fit $\hat{L}(D) = a + b \ln D$ by unweighted ordinary least squares to the five measured minima, where $D \in \{5, 10, 25, 50, 100\}$ is the training-data percentage. Each fraction has one run, without across-seed error bars. Fits describe only the observed range; their R^2 values measure in-sample fit, without establishing cross-seed reproducibility, equal slopes, or extrapolation accuracy. Similar absolute improvements need not imply similar relative improvements. Closed-loop robot performance is evaluated separately from these offline prediction metrics.

D.2 Training-Stage Ablations

The study varies the inclusion of Stage-I pretraining 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; the other two ablations omit Stage I or Stage II individually. Architecture and action representations are unchanged. Table 3 reports these variants and the full recipe on the four real-world tasks, with ten trials per task and progress averaged equally across tasks. SIMPLE configuration-selection limitations are described in Appendix C.2.

D.3 Generalization

The protocol covers all four real-world tasks. Let $\mathcal{D}_A^R$ denote robot demonstrations of 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 *Robot-only* evaluates this robot-only policy on B. The Robot-only condition therefore measures object-instance generalization within a demonstrated task using a policy trained without human data, distinct from the target-task-robot-free setting in Sec. 5.1.

The Human-guided and Robot-only 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. 13 reports mean weighted progress over 10 trials per task, method, and condition (600 trials). The Seen values come from the generalization evaluation, whereas Table 1 uses the standard-task evaluation records. These results are reported separately. 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.