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

Unlocking In-the-Wild Loco-Manipulation with Robot-Free Egocentric Demonstration

Modi Shi[2,3*] Shijia Peng[1*†] Jin Chen[2*] Haoran Jiang[2] Tianyu Li[2]
Di Huang[3] Ping Luo[1†] Hongyang Li[1‡] Li Chen[1‡]
[1] The University of Hong Kong [2] Shanghai Innovation Institute [3] Beihang University
https://opendrivelab.com/EgoHumanoid

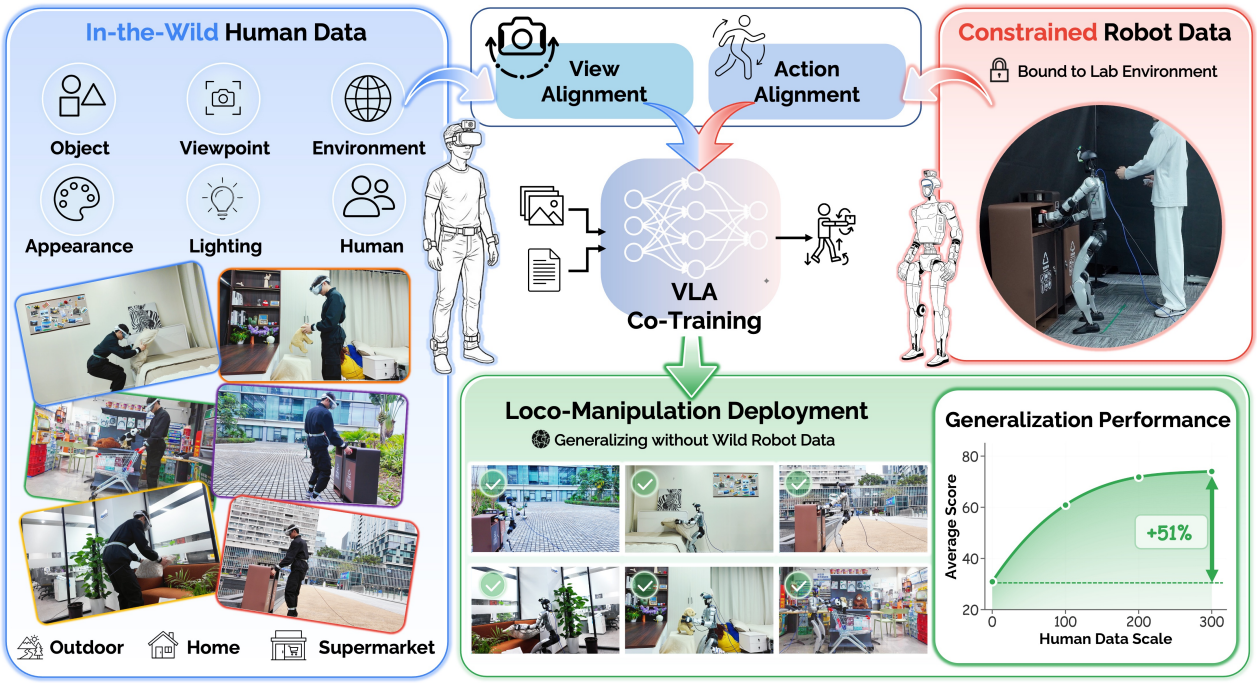


Fig. 1: Introducing **EGOHUMANOID**, the first investigation on human-to-humanoid transfer for whole-body loco-manipulation. Robot teleoperation data collection is constrained to laboratory environment due to hardware and safety limitations, while in-the-wild human demonstrations offer scalable diversity in objects, scenes, lighting, and viewpoints. Our alignment pipeline bridges the embodiment gap through view and action alignment, enabling vision-language-action (VLA) co-training on both data sources. Real-world loco-manipulation deployment validates that egocentric human demonstrations invigorate generalization *without* scene-specific robot data, outperforming robot-only baselines by 51% with consistent scaling behavior.

***Abstract*—Human demonstrations offer rich environmental diversity and scale naturally, making them an appealing alternative to robot teleoperation. While this paradigm has advanced robot-arm manipulation, its potential for the more challenging, data-hungry problem of humanoid loco-manipulation remains largely unexplored. We present EGOHUMANOID, the first framework to co-train a vision-language-action policy using abundant egocentric human demonstrations together with a limited amount of robot data, enabling humanoids to perform loco-manipulation across diverse real-world environments. To bridge the embodiment gap between humans and robots, including discrepancies in physical morphology and viewpoint, we introduce a systematic alignment pipeline spanning from hardware design to data processing. A portable system for scalable human data collection is developed, and we establish practical collection protocols to improve transferability. At the core of our human-to-humanoid alignment pipeline lies two key components. The view alignment reduces visual domain discrepancies caused by camera height and perspective variation. The action alignment maps human motions into a unified, kinematically feasible action space for humanoid control. Extensive real-world experiments demonstrate that incorporating robot-free egocentric data significantly outperforms robot-only baselines by 51%, particularly in unseen environments. Our analysis further reveals which behaviors transfer effectively and the potential for scaling human data.**

*Equal Contribution. ‡Project lead. †Work done while at Kinetix AI.

I. INTRODUCTION

Humanoid robots hold immense promise for operating across diverse human-centric environments, from household assistance to outdoor service scenarios [23]. These applications fundamentally demand *loco-manipulation*, which involves the

close orchestration of whole-body locomotion and dexterous manipulation [38, 54, 77]. Unlike fixed-base manipulators [31, 5, 70], robots must navigate through varied spaces, adjust their body posture for reachability, and manipulate objects, all while maintaining dynamic balance in unstructured environments.

Despite rapid progress in model architectures and control algorithms, learning humanoid loco-manipulation remains bottlenecked by the *scarcity* of diverse, large-scale demonstration data. Existing approaches [18, 3] primarily rely on robot teleoperation, which provides embodiment-consistent supervision but suffers from high cost, operational complexity, and hardware instability [38, 25, 73]. Moreover, it confines data collection to the laboratory environment, as transporting humanoid platforms and teleoperation equipment such as motion capture suits to diverse real-world scenarios, *e.g.*, homes, parks, stores, outdoor spaces, is often impractical [81, 74].

We instead provide a compelling alternative based on the simple observation: humans naturally perform loco-manipulation tasks every day, exactly across the environments where robots are expected to operate. Recent advances in wearable sensing [15, 22, 50] make it possible to capture egocentric human demonstrations using lightweight, portable devices, without requiring any robot hardware. This paradigm offers scalable access to behaviorally rich and environmentally diverse data, and has induced progress in robot manipulation and navigation [32, 29, 55, 83, 49, 30, 10].

However, applying egocentric human demonstrations to humanoid control is far from straightforward due to the fundamental *embodiment gap*. Humanoid robots differ substantially from humans in morphology and kinematics, including limb proportions and joint limits [78, 65]. Egocentric visual observations also differ – humans observe their own hands and body, whereas robots perceive metallic manipulators from distinct viewpoints [12, 81]. Moreover, motion dynamics diverge, as human walking patterns, body sway, and balance strategies do not directly transfer to robots with different mass distributions and actuation constraints [49, 83]. Collectively, these discrepancies are pronounced in loco-manipulation, where whole-body movement amplifies viewpoint change and introduces substantial motion variation in the egocentric observations.

To this end, we present **EGOHUMANOID**, a systematic framework for human–humanoid co-training as described in Fig. 1. Our key insight is that while low-level actions are embodiment-specific, high-level behavioral structures (*e.g.*, navigation routes, object approach strategies, and task decomposition) *could* transfer reliably when observations and motions are properly aligned. Moreover, human demonstrations in diverse real-world settings expose policies to variations that robot-only data rarely covers, enabling stronger in-the-wild generalization. Accordingly, we build a data collection system that captures both robot-free human demonstrations and teleoperated humanoid robot data. The portable human setup integrates a VR headset, body trackers for pose estimation, and an egocentric camera, enabling scalable collection in varied scenarios without robot hardware. Complementary robot data, collected via VR-based teleoperation, provides embodiment-accurate supervision for manipulation-intensive behaviors.

The embodiment alignment pipeline is proposed to convert human demonstrations into robot-compatible training signals. Our approach focuses on two core components. The view alignment reduces visual discrepancies by transforming human egocentric observations to approximate robot viewpoints using depth-based reprojection and inpainting. The action alignment employs a unified action space shared by humans and robots, using delta end-effector poses for upper body control and discrete commands for locomotion. Together, these mechanisms guarantee the VLA policy co-training on both data sources.

EGOHUMANOID is validated on a Unitree G1 humanoid robot across four indoor and outdoor loco-manipulation tasks. Experiments demonstrate that incorporating human data significantly improves policy performance by 20% on average and enables generalization to scenes not covered by the robot data, yielding a 51% performance gain. Through extensive ablation studies, we scrutinize the transfer mechanism for sub-tasks, validate the scaling effectiveness, and identify critical design choices in our alignment pipeline.

In summary, the contributions are: **(a) The first endorsement of human-to-humanoid transfer for whole-body loco-manipulation tasks.** We show that egocentric human data effectively enhances humanoid policy through co-training, establishing the feasibility of cross-embodiment transfer with an integrated data collection and training framework. **(b) A principled embodiment alignment pipeline.** We propose practical strategies to bridge human-robot embodiment gaps, combining view alignment to mitigate visual disparities and action alignment to enforce kinematic feasibility. **(c) Comprehensive real-world evaluation and analysis.** We characterize which behaviors transfer effectively, how to scale human data, and exhibit that human data substantially boosts generalization beyond the robot-only training environment.

We hope this study could encourage broader exploration of egocentric human data as a scalable pathway toward generalizable humanoid control, and lay the groundwork for transferring human behaviors to complex loco-manipulation skills.

II. RELATED WORK

A. Egocentric Human Data for Robots

Egocentric human data has emerged as a promising source for robot learning, offering scalable collection in diverse environments without robot hardware. Various systems enable robot-free data capture, including handheld grippers with mounted cameras [12, 80], wearable glasses [15], and VR/AR headsets such as Apple Vision Pro, Meta Quest, and PICO. Such data has been leveraged in multiple ways for embodied AI. One line of work trains vision-language models on large-scale egocentric datasets [21, 22, 67, 39, 36, 76] to enhance scene understanding and spatial reasoning, with applications in embodied question answering [42] and egocentric decision-making [62]. For robot policy learning, prior work primarily adopts a pretraining-then-finetuning paradigm, using human data to learn visual representations [43, 41, 75], motion priors [44, 82, 66, 71], or latent action embeddings [68, 6, 27],

discarding action information during pretraining. Co-training, in contrast, leverages actions from both human and robot demonstrations as supervision through aligned observation and action spaces, enabling more effective knowledge transfer [49, 30, 83, 72]. However, existing co-training approaches focus on fixed-base manipulation. Our work presents the first validation of human-robot co-training for humanoid loco-manipulation, addressing the unique challenges of transferring whole-body human demonstrations to mobile humanoid platforms.

B. Cross-Embodiment Data Alignment

In robot-to-robot transfer, Mirage [8] reduces visual mismatch by rewriting robots in images through cross-painting and inpainting, while Chen *et al.* [9] use generative models to augment robot appearances and viewpoints. For action alignment, common approaches either adopt transferable task-space or end-effector control interfaces [8, 60, 53], or project heterogeneous state–action spaces into a shared latent representation [59, 37, 13]. Human-to-robot transfer is more challenging due to large gaps in morphology, viewpoint, and motion dynamics, but it is attractive given the scale of human data. The UMI series [12, 80, 63, 61] makes human demonstrations robot-compatible at collection time using hand-held hardware that provides accurate global positions. However, this hand-held hardware ties the collected data to a specific gripper morphology. Another line of work relies on motion retargeting, mapping human motions to robot configurations [2, 65, 19], which is usually employed for training low-level humanoid controllers [24, 11, 57]. This paradigm is further used by several concurrent manipulation or navigation-centered works, through estimating hand poses from egocentric human videos for joint training with robot data [29, 83, 30, 72]. They focus on manipulation solely or decouple these two subtasks for mobile robots to prevent base shift. Contrarily, our work targets humanoid loco-manipulation with tightly-coupled whole-body coordination. We address the larger morphological gap through dedicated view transformation, including reprojection and generation, and action retargeting to fully exploit egocentric human demonstrations.

C. Humanoid Loco-Manipulation

Humanoid loco-manipulation couples locomotion and manipulation, as well as task planning, where both reliable execution and scalable supervision remain challenging. Note that "whole-body" here is used at the task level rather than full joint-level actuation. Inspired by humanoid whole-body control [24, 1, 57, 48], emerging works build loco-manipulation skills with reinforcement learning, typically relying on specifically modeled objects in simulation, carefully designed rewards, and learning curricula to handle the strong coupling between base motion and arm–hand interaction [25, 77, 69, 78]. However, these methods face scaling bottlenecks such as per-contact physics tuning, intractable affordance modeling, and unmodeled hardware non-idealities. In parallel, recent works explore scaling supervision through motion- or language-conditioned generation, aiming to lower the burden of expensive robot teleoperation needed for new tasks [17, 14, 64, 56]. VLA policies have been developed from manipulation methods as well, by extending models with lower-body robot action commands [27, 81, 46] or motion targets [40, 4, 16]. Despite this progress, most pipelines still scale primarily through robot-centered data or constrained setups, enduring the lack of environmental diversity and laborious data collection process, which prevents studying generalization in human-centric scenes. In sharp contrast, we treat robot-free egocentric human demonstrations as a scalable source of diverse whole-body behaviors to complement robot data, and we provide a starting point to co-train a humanoid VLA with two sources of data corpus through explicit view and action alignment.

III. EgoHumanoid Framework

We present the EgoHumanoid framework for training humanoid loco-manipulation policies by leveraging both egocentric human demonstrations and teleoperated robot data. The overview is illustrated in Fig. 1. We first formalize the problem setup (Sec. III-A), then describe our data collection system (Sec. III-B), embodiment alignment pipeline (Sec. III-C), and policy training procedure (Sec. III-D).

A. Problem Setup

Our goal is to train a VLA model capable of completing loco-manipulation tasks across novel real-world environments. The policy is trained on a combined dataset $\mathcal{D} = \mathcal{D}_{\text{robot}} \cup \mathcal{D}_{\text{human}}$, where $\mathcal{D}_{\text{robot}}$ consists of teleoperated robot demonstrations collected in constrained laboratory settings, and $\mathcal{D}_{\text{human}}$ comprises egocentric human demonstrations of the same tasks captured across diverse scenarios, *e.g.*, homes, stores, and outdoor environments. Each data episode in $\mathcal{D}$ contains egocentric videos and synchronized whole-body actions.

After training, the policy is deployed under two settings: (1) *in-domain* evaluation in laboratory environments similar to those in $\mathcal{D}_{\text{robot}}$, and (2) *generalization* evaluation in scenes covered by $\mathcal{D}_{\text{human}}$ but absent from $\mathcal{D}_{\text{robot}}$. This setup directly tests whether human data enables generalization beyond the limited scenes where robot teleoperation is feasible.

We use a mid-sized (1.3m) Unitree G1 humanoid robot as our hardware platform, on account of its robust performance in low-level controller and durability. Meanwhile, the robot bares huge embodiment gap with normal teleoperators considering its size and degree of freedom (29 DoF with 3-finger Dex3 dexterous hands). Unlike prior work on tabletop bimanual manipulation, where the robot base remains stationary, we focus on tasks that require whole-body coordination—the robot must locomote to target locations while simultaneously or sequentially performing manipulation. This fundamental requirement for lower-body mobility distinguishes humanoid loco-manipulation from fixed-base settings and motivates our investigation of human-to-humanoid transfer.

B. Data Collection System

Crucially, robots and humans differ in practicality and embodiment. Robot data collection typically requires expensive

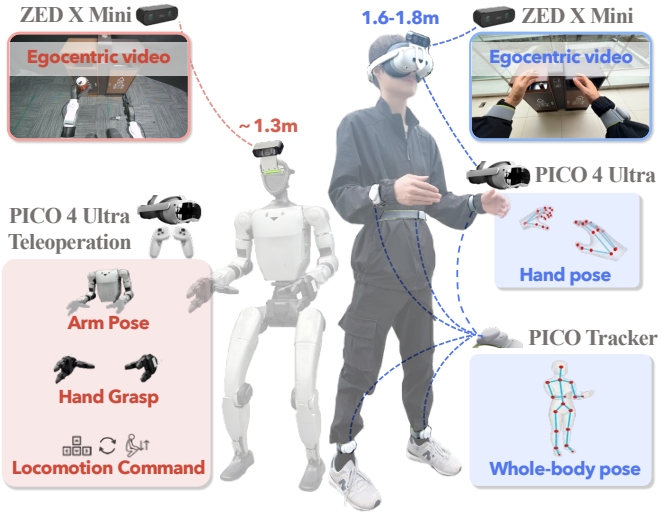


Fig. 2: **Hardware setup for data collection.** Humans and the G1 humanoid robot are equipped with an integrated VR-based system for portable usage and agile development. The same camera captures egocentric recordings. The VR headset and trackers provide coarse human poses, while the VR controller is employed to teleoperate the robot.

hardware, careful setup, and is mostly bound to laboratories, while human data can be collected cheaply and flexibly using wearable devices in diverse settings. To mitigate the embodiment and visual gap in the data collection level and support human-robot co-training, we develop a unified and portable hardware setup with a VR-based system for swift adaptation between the two collection modes, shown in Fig. 2. We intentionally drop wrist cameras as egocentric-only is a more universal setup, and the benefits of wrist views are not deterministic due to the enormous gaps, as shown in previous works [30] (See more discussion in Sec. V). The efficiency of human demonstration contrast to robot teleoperation is pronounced ($\sim 2\times$) and detailed in Table III.

Human Data Collection. We use a portable PICO VR setup that enables low-cost and unconstrained recording of human demonstrations in both indoor and outdoor settings. The subject wears the headset with five PICO Motion Trackers, while a head-mounted ZED X Mini camera captures synchronized egocentric RGB images. Using the PICO SDK [79], we record full-body human motion in real time, including 24 body keypoints and detailed hand poses with 26 keypoints per hand.

Robot Data Collection. Robot demonstrations are collected through VR-based teleoperation. The operator wears a PICO VR headset and uses handheld controllers to issue navigation commands (forward/backward, lateral movement, turning, standing, squatting) and wrist pose commands derived from the controller-to-headset relative pose. These commands are converted to joint-level actions through inverse kinematics and executed on a Unitree G1 humanoid, while the controller trigger controls grasping of a Dex3 dexterous hand. A low-level locomotion policy [45] ensures stable whole-body execution. We record navigation commands, wrist poses of the end-effector, hand grasp states, and synchronized egocentric RGB images from a head-mounted ZED X Mini camera.

C. Human-to-Humanoid Alignment

To transform human demonstrations captured by the VR setup in Sec. III-B into robot-compatible training data, we develop an alignment pipeline consisting of two main modules, view alignment and action alignment for addressing visual domain gaps from differing camera viewpoints and action representation gaps from morphological discrepancies, respectively. The procedure is depicted in Fig. 3 and described below. More implementation details are provided in Appendix C.

View Alignment. The substantial height discrepancy between humans and humanoid robots results in pronounced differences in egocentric visual observations. To bridge this visual gap, we transform human egocentric images to approximate robot camera viewpoints through a three-stage process. First, we use MoGe [58] to infer an affine-invariant per-pixel 3D point map via its reprojection-based focal/shift recovery, and derive a scale-invariant depth map. We then transform the recovered 3D points into the target robot camera frame and project them onto the target image plane. During training, random perturbations are applied to the target pose to encourage robustness to viewpoint variations. Note that the reprojection can produce missing regions due to invalid 3D predictions indicated by MoGe's validity mask and view-dependent disocclusions introduced by the pose change, leaving target pixels with no source correspondence. Therefore, the final step involves employing latent diffusion-based inpainting [51] to hallucinate the missing regions, conditioning on the observed context and the missing region mask. Throughout this, complete RGB images are generated for better mimicking robot egocentric inputs.

Action Alignment. We design a unified action space that accommodates both human and robot demonstrations while respecting their kinematic differences.

Upper Body. The actions are parameterized as 6-DoF delta end-effector poses, consistent with concurrent cross-embodiment works [83, 30, 72, 71]. Using deltas avoids dependence on a globally aligned base frame, which could be ill-defined between human and robot recordings, enabling direct comparability across embodiments. We also avoid joint-level retargeting [2] as it may introduce artifacts and perturb interaction-relevant hand–object geometry [65]. Explicitly, we express human wrist poses in a pelvis-centric frame, then smooth translations with a Savitzky-Golay filter [52]. For rotations, we filter in the $\mathrm{SO}(3)$ tangent space using log/exp maps to prevent quaternion interpolation ambiguities. Finally, the transformed data is downsampled from 100 Hz to 20 Hz to match robot control and output the delta pose between consecutive frames as the action.

Lower body. To collect robot data, the operator issues navigation commands using a discrete set of constant-velocity primitives (*i.e.*, forward/backward, left/right, turn left/right, stand/squat) through the VR setup, following common practice in humanoid loco-manipulation [3, 27]. To align human

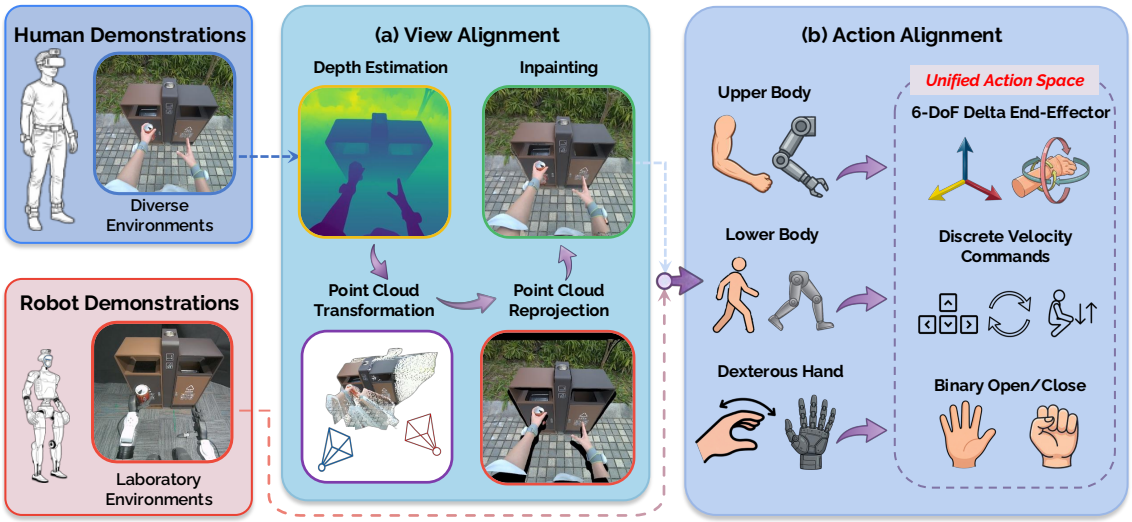


Fig. 3: **Pipeline of human-to-humanoid alignment. (a) View Alignment**: Egocentric images are transformed to approximate robot viewpoints by reprojecting estimated depth points and generative inpainting to fill in blank holes. **(b) Action Alignment**: We employ relative end-effector poses to unify the upper body action space, and discrete commands for lower-body locomotion.

demonstrations with this robot action space, we convert the human pelvis trajectory into the same discrete commands. Specifically, we apply Savitzky–Golay smoothing to suppress jitter, then estimate the instantaneous heading via centered differences with a continuity constraint to prevent direction flips. Displacements in the world frame are projected on the local frame to obtain forward and lateral velocities. Yaw rate is computed from inter-frame heading changes. We downsample these continuous commands to 20 Hz by averaging within each control window, and quantize them into discrete bins of horizontal movements. Finally, a discrete stand/squat primitive is derived by thresholding inter-frame changes in pelvis height, matching the robot teleoperation interface.

Gripper. We represent the gripper action as a binary variable $a_t = \{0, 1\}$, where 1 corresponds to the gripper being closed and 0 corresponds to the gripper being open. During robot data collection, a teleoperator controls the gripper via a handheld controller, and the resulting binary command sequence $\{a_t\}$ is recorded. Human demonstrations are captured using the VR system that tracks 26 hand joints per hand. To robustly infer the hand grasping state, we first apply low-pass filtering followed by Savitzky-Golay smoothing to the raw joint trajectories. We then compute the finger-level curvature κ_f, defined as the curvature evaluated at the midpoint of a quadratic polynomial fitted to each finger's joint polyline. The hand-level grasping state is obtained by averaging κ_f across all fingers and thresholding the resulting scalar $\bar{\kappa}$ to produce a binary open/close label. This curvature-based representation mitigates noise and enables reliable supervision extraction from human demonstrations. Similarly, all human-derived signals are downsampled to 20 Hz to align temporally with the robot data.

D. Loco-Manipulation Policy Co-Training

With identical visual observation formats and action dimensions owing to the alignment procedure, we co-train a single policy on both human and robot demonstrations. Explicitly, we adapt $\pi_{0.5}$ [26], a state-of-the-art VLA model, for humanoid loco-manipulation via finetuning. The policy receives egocentric RGB observations and language instructions as input, and outputs actions in our unified action space. We intentionally omit proprioceptive states, as morphological differences between humans and humanoids yield incompatible proprioceptive distributions that would confound co-training.

Multi-source Data Sampling. In this case, the amount of human demonstrations may be substantially larger and more diverse than robot data, creating a highly imbalanced dataset. Prior work has shown that balanced sampling is essential for imbalanced multi-source training in robot manipulation [20, 28]. Meanwhile, loco-manipulation tasks require both high-level navigation or locomotion and intricate manipulation, where human and robot data may feature distinct advantages in these aspects. Thereafter, we examine the human-to-robot sampling ratio within each mini-batch, preserving sufficient exposure to robot demonstrations (See Sec. IV-D).

IV. Evaluations

We conduct real-world experiments to evaluate the effectiveness of leveraging egocentric human data for humanoid loco-manipulation. Our evaluation addresses three key questions:

- **Q1: Generalization.** Does egocentric human data enable humanoid policies to generalize to diverse human-centric environments beyond laboratory settings?
- **Q2: Transfer Analysis.** What behaviors transfer effectively from human demonstrations to humanoid policies?
- **Q3: Data Scaling.** Does policy performance scale with increasing amounts of human demonstration data?

Fig. 4: **Humanoid loco-manipulation tasks for evaluation.** We design four tasks which span varying levels of difficulty for large-space movement and dexterous manipulation. Robots are teleoperated in laboratories as the source domain (**Top**). Human-centric scenes occur in human demonstrations only **(Middle)**, and set as testbeds for generalization evaluation (**Bottom**).

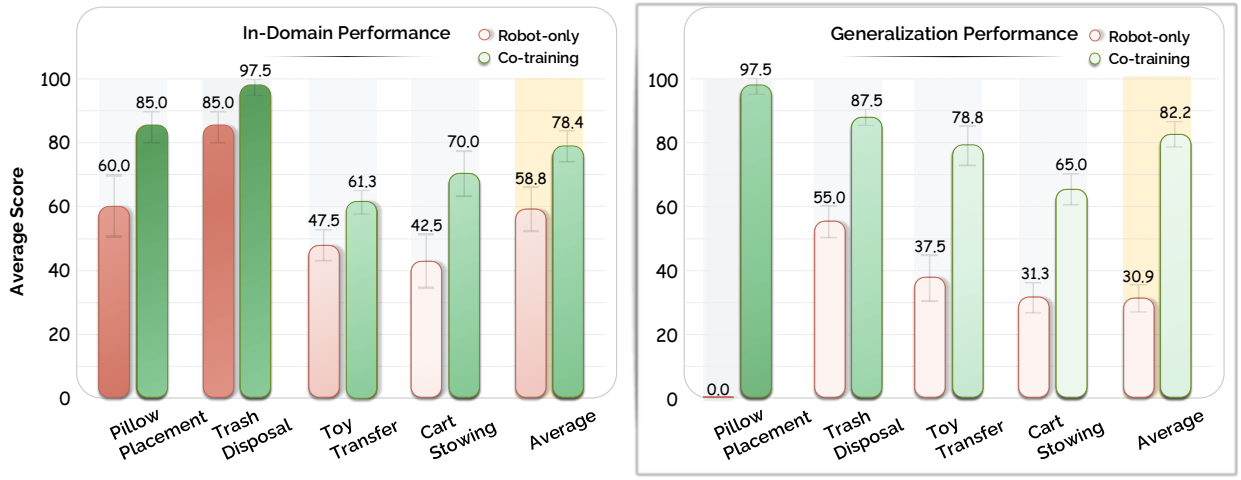


Fig. 5: **Performance of human-robot data co-training with EGOHUMANOID.** Our pipeline achieves unanimous improvements over robot-only baselines across in-domain and generalized environments. The boost is amplified in robots' unseen settings.

A. Real-world Experimental Setup

We design four loco-manipulation tasks requiring tight coordination between locomotion and manipulation, illustrated in Fig. 4. All tasks involve non-trivial locomotion (1–5m), distinguishing them from fixed-base bimanual manipulation. Navigation accuracy subsequently impacts manipulation success: suboptimal stopping positions render subsequent manipulation infeasible. The variance in stopping positions across trials further requires policies to generalize across viewpoints and object configurations.

The four tasks are: (1) `Pillow Placement`. The robot carries a pillow to a bed and places it at a target location. This task tests stable locomotion with a bulky object and placement on a deformable surface. (2) `Trash Disposal`. The robot carries garbage (crumpled paper or a can) to a waste container with lids, inserts it horizontally into the hole rather than dropping from above, requiring precise localization and end-effector control. (3) `Toy Transfer`. The robot approaches a toy on a surface, grasps it, reorients, walks to a distant table, and places the toy down. This task evaluates sequential coordination across approach, grasp, in-hand locomotion, and placement. (4) `Cart Stowing`. The robot pushes a cart to a product display, grasps a toy, places it in the cart, and pushes the cart away. This task involves sustained contact during locomotion and multi-phase manipulation.

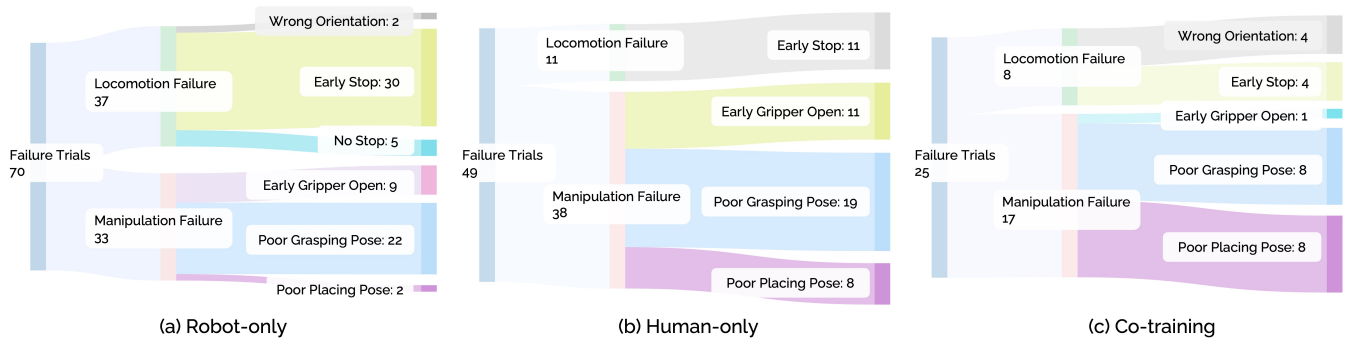
We conduct 20 trials per setting with position perturbations, and normalized scores are adopted to evaluate the performance of each trial. We delineate complementary experimental details in Appendix C2, and list our scoring criteria in Appendix D.

B. Will human data improve humanoid loco-manipulation?

We evaluate whether co-training with robot-free egocentric demonstrations enables humanoid policies to generalize

TABLE I: **Granular performance across tasks and subtasks.** Each task is roughly categorized into sub-steps of locomotion (including high-level navigation) or manipulation, represented in blue and red, respectively, and denoted as s1, s2, *etc*.

Training Data	Pillow Placement		Trash Disposal		Toy Transfer				Cart Stowing			
	s1	s2	s1	s2	s1	s2	s3	s4	s1	s2	s3	s4
Robot-only	0	0	65	45	100	50	0	0	100	15	5	5
Human-only	100	95	100	80	100	100	45	35	100	5	0	0
Co-training	**100**	**95**	**100**	**75**	**100**	**100**	**60**	**55**	**100**	**60**	**50**	**50**



(a) Robot-only (b) Human-only (c) Co-training

Fig. 6: **Failure mode analysis.** For each setting—(a) Robot-only, (b) Human-only, and (c) Co-training—we log all failed episodes. The Sankey diagrams break failures down into locomotion vs. manipulation errors and finer causes.

beyond laboratory environments. We compare *Robot-only*, trained exclusively on robot data, with *Co-training*, trained on both sources using our alignment pipeline. Policies are evaluated under two settings: *In-Domain* (laboratory environments) and *Generalization* (diverse scenes covered only by human data). For these experiments, robot-only models use 100 episodes per task, while co-training absorbs 300 episodes of human demonstrations additionally.

As shown in Fig. 5, co-training consistently improves performance across both settings. For in-domain evaluation, co-training achieves 78% average score versus 59% for robot-only. The gap widens remarkably in generalization settings: co-training reaches 82% compared to 31% for robot-only. Notably, generalization performance exceeds in-domain results, indicating that human data effectively bridges the domain gap to novel environments. These results demonstrate that humanoid robots can operate in diverse real-world settings without in-the-wild robot data collection.

C. Which subskill benefits most from the data transfer?

Delving deeper into the results in Fig. 5, we examine which skills are primarily learned from human demos. We train a *Human-only* model and analyze performance separately across locomotion and manipulation stages. Table I shows consecutive success rates of sub-steps. Though dependence exists among steps leading to a not strictly aligned 'apple-to-apple' comparison, two observations emerge.

First, navigation transfers effectively from human data alone. Human-only achieves 100% on navigation-dominated subtasks (Pillow Placement s1, Trash Disposal s1, Toy Transfer s1, Cart Stowing s1). Even on Toy Transfer s3, requiring two consecutive turns, Human-only attains 45%, only 15% below Co-training. In contrast, Robot-only yields near-zero success on these stages, confirming that human demonstrations provide transferable navigation knowledge.

Second, manipulation also transfers but effectiveness diminishes with increasing precision requirements. Human-only outperforms Robot-only on coarse manipulation subtasks (Pillow Placement s2, Trash Disposal s2, Toy Transfer s2/s4), where approximate contact suffices. However, for precision-critical phases such as Cart Stowing s2, Human-only achieves only 5% versus Robot-only's 15%. Notably, Co-training reaches 60% on this subtask, indicating that human data provides useful priors that become effective when combined with robot demos.

The failure analysis in Fig. 6 corroborates these findings: Human-only exhibits three times more manipulation failures than locomotion failures, whereas Robot-only failures are balanced across both phases.

D. Does Performance Scale with Human Data?

We investigate how human demonstration quantity and mini-batch sampling ratio jointly affect policy performance. Fig. 7 presents results across four tasks as human demonstrations scale from 0 to 300 under varying sampling ratios. The optimal ratio depends on task characteristics: tasks requiring only coarse grasping (*e.g.*, Pillow Delivery) benefit from higher human data ratios (Robot:Human 1:2), while fine manipulation tasks (*e.g.*, Toy Transfer, Cart Shopping) favor higher robot ratios (Robot:Human 2:1). This aligns with our subtask analysis (Sec. IV-C), confirming that the embodiment gap disproportionately impacts precision-critical manipulation.

Across all sampling ratios, results improves consistently as human demonstrations accumulate, verifying that our method scales effectively without overfitting to dominant human data.

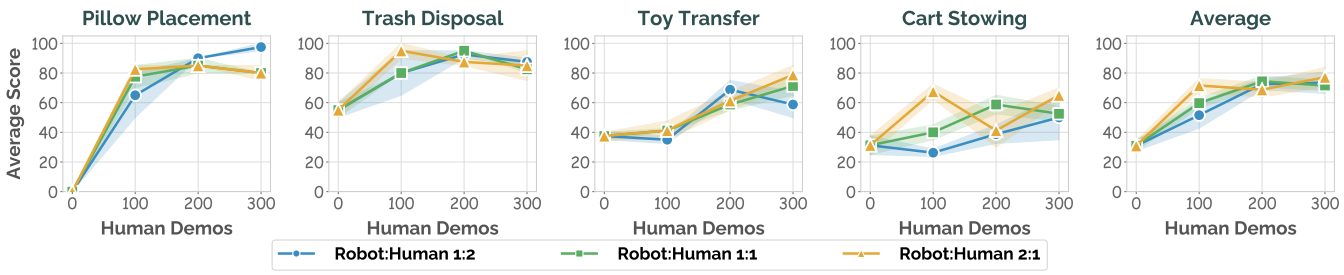


Fig. 7: **Scaling human demonstrations with various training-time data sampling strategies.** More human demonstrations, from robot-only (0 Human Demos) to 3× robot data (300 Human Demos), principally yield improved loco-manipulation performance in generalized scenarios. The optimal sampling ratio depends on task characteristics: manipulation-heavy tasks favor higher robot data ratios, while navigation-dominant tasks benefit from more human data.

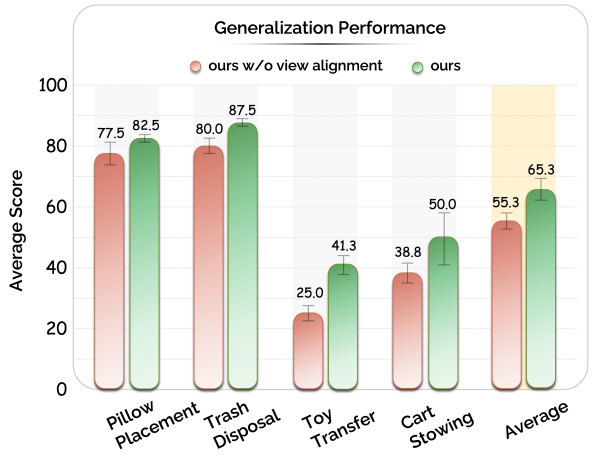


Fig. 8: **View alignment ablation.** View alignment is essential for effective human-robot co-training, especially for tasks involving variable object heights such as Toy Transfer.

E. View Alignment Ablation

We ablate the effect of view alignment in Fig. 8. View alignment yields consistent improvements across all tasks, with the largest gains on Toy Transfer and Cart Stowing. Both tasks involve objects placed at varying heights, creating deployment viewpoints that differ from both data sources: human demonstrations are captured from a higher camera position, while robot data is collected with objects at different heights. Without view alignment, the policy encounters out-of-distribution viewpoints from either source. Our view transformation bridges this gap by reprojecting human observations to the robot's camera height with added pose perturbations, exposing the policy to broader viewpoint variations during training. Visualizations are provided in Appendix E.

F. Scene Diversity in Human Data

We further investigate how scene diversity affects co-training. Fixing the data quantity at 100 robot and 300 human demonstrations, we vary only the number of distinct human scenes and evaluate zero-shot in a novel scene unseen during training (Table II). The score rises monotonically from 57.5% (robot-only) to 82.5% with three scenes, indicating that scene diversity is a key driver of generalization even at fixed data quantity.

G. Data Collection Efficiency

As shown in Table III, human demonstrations are roughly 2× faster than robot teleoperation across all four tasks, reducing the per-episode collection time from 62.1 s to 39.7 s on average.

V. Discussions

Our work demonstrates inspiring pioneering results of learning humanoid loco-manipulation from human data, boosting robot-only baselines by a large margin as shown in Fig. 5. Yet, the overall success rate is not satisfactory, and typical failure cases are characterized in Fig. 6. Due to the instability of the humanoid and the huge embodiment gap, we find a bunch of critical factors that affect the performance significantly, even when equipped with the proposed pipeline. We present the devil in the details below, and also incorporate practical guidelines for data collection in Appendix B.

Action Representation. Two candidate action spaces exist for human-robot alignment: absolute end-effector pose in camera frame and delta end-effector pose. We adopt the latter because human and humanoid morphologies yield systematically different hand-to-camera distances—human arms are typically longer relative to torso height, resulting in different absolute position distributions even for functionally equivalent poses. Delta end-effector naturally shares an identical action space.

However, delta end-effector poses present a fundamental limitation: the correspondence between human hand orientation and robot gripper orientation is ambiguous. While we define explicit pose correspondences, these mappings are difficult to infer from visual observations alone. Consequently, without proprioceptive input, precise rotation transfer remains challenging—the policy cannot reliably disambiguate intended end-effector orientations from egocentric images, limiting fine-grained rotational control in manipulation tasks.

Data Scale Requirements. Humanoid loco-manipulation presents unique data efficiency challenges compared to fixed-base manipulation. These tasks require the robot to first

TABLE II: **Effect of human demonstration scene diversity on zero-shot generalization.** We evaluate the Trash Disposal task in a novel scene unseen during training, progressively increasing the number of distinct human demonstration scenes while keeping robot data fixed. Sub-steps s1 (locomotion) and s2 (manipulation) are reported alongside the average task score.

Training Data	# Human Scenes	s1	s2	Average Score
Robot-only	0	70	45	57.5
Co-training	1	90	60	75.0
Co-training	2	**100**	50	75.0
Co-training	3	**100**	**65**	**82.5**

TABLE III: **Comparison of data collection efficiency.** The average time (s) for collection of each data episode shows the preeminent efficiency ($\sim 2\times$) of gathering human demonstrations over teleoperating robots.

Method	Pillow Placement	Trash Disposal	Toy Transfer	Cart Stowing	Average
Robot teleop	35.7	43.1	66.2	103.5	62.1
Human demo	16.2	18.3	34.5	89.9	**39.7**

navigate to the target location before performing manipulation—yet the robot rarely stops at precisely the same position, resulting in highly diverse viewpoints during the manipulation phase. This viewpoint variability raises substantially increasing data requirements. In our experiments, achieving comparable success rates on loco-manipulation tasks requires approximately 2–3× more demonstrations than fixed-base manipulation tasks of similar difficulty, highlighting the compounded challenge of learning both robust navigation and precise manipulation from egocentric observations.

VI. Conclusion

We introduce EgoHumanoid, the first framework to realize human-to-humanoid transfer for loco-manipulation. By aligning egocentric human demonstrations with robot data through view transformation and unified action space, our approach enables effective VLA co-training. Experiments demonstrate substantial generalization improvements, enabling deployment in diverse environments without in-the-wild robot data. We hope this work encourages broader exploration of egocentric human data toward generalizable humanoid control.

Limitations and Future Work. As an early endeavor, our method still bears limitations with respect to orientation ambiguity in delta end-effector representations, scaling laws with human data, and expressive whole-body controls. We will take scaling up this paradigm with advanced egocentric hardware as key future direction. More discussions are in Appendix A.

Acknowledgments

This study is supported by National Natural Science Foundation of China (62206172). This work is in part supported by the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust. We gratefully acknowledge our robot operators for data collection and evaluations. We also extend our thanks to Jiazhi Yang, Tianyu Li, Yixuan Pan, Junli Ren, Penglin Fu, and Kaiyang Wu for their insightful feedback and fruitful discussions.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *CoRL*, 2025. 3

[2] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting Matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025. 3, 4

[3] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. HOMIE: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. In *RSS*, 2025. 2, 4

[4] Boston Dynamics. Large behavior models help atlas find new footing. https://bostondynamics.com/blog/large-behavior-models-atlas-find-new-footing/, 2025. 3

[5] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IROS*, 2025. 2

[6] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. In *RSS*, 2025. 2

[7] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-N-On: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025. 13

[8] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. In *RSS*, 2024. 3

[9] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken

Goldberg. RoVi-Aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *CoRL*, 2024. 3

[10] Li Chen, Chonghao Sima, Kashyap Chitta, Antonio Loquercio, Ping Luo, Yi Ma, and Hongyang Li. Intelligent robot manipulation requires self-directed learning. *Authorea Preprints*, 2025. 2

[11] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. GMT: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025. 3

[12] 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 *RSS*, 2024. 2, 3

[13] Apan Dastider, Hao Fang, and Mingjie Lin. Cross-embodiment robotic manipulation synthesis via guided demonstrations through cyclevae and human behavior transformer. *arXiv preprint arXiv:2503.08622*, 2025. 3

[14] Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-VLA: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025. 3

[15] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2

[16] Figure. Introducing Helix 02: Full-body autonomy. https://www.figure.ai/news/helix-02, 2026. 3

[17] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. DemoHLM: From one demonstration to generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2510.11258*, 2025. 3

[18] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *CoRL*, 2024. 2

[19] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *CoRL*, 2025. 3

[20] Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *RSS*, 2024. 5

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[22] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 2

[23] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *IEEE/ASME TMECH*, 2025. 1

[24] 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 *CoRL*, 2025. 3, 13

[25] Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, et al. VIRAL: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025. 2, 3

[26] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *CoRL*, 2025. 5, 14, 17

[27] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, et al. WholeBodyVLA: Towards unified latent vla for whole-body loco-manipulation control. In *ICLR*, 2026. 2, 3, 4

[28] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. MT-Opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021. 5

[29] 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 *ICRA*, 2025. 2, 3

[30] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025. 2, 3, 4, 13

[31] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *CoRL*, 2024. 2

[32] Ashish Kumar, Saurabh Gupta, and Jitendra Malik. Learning navigation subroutines from egocentric videos. In *CoRL*, 2019. 2

[33] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025. 13

[34] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025. 13

[35] Guangrun Li, Yaoxu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2r: A human-

to-robot data augmentation for robot pre-training from videos. *arXiv preprint arXiv:2505.11920*, 2025. 13

[36] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2

[37] Junjia Liu, Zhuo Li, Minghao Yu, Zhipeng Dong, Sylvain Calinon, Darwin Caldwell, and Fei Chen. Human-humanoid robots cross-embodiment behavior-skill transfer using decomposed adversarial learning from demonstration. *RAM*, 2025. 3

[38] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-TeleVision: Predictive motion priors for humanoid whole-body control. In *ICRA*, 2025. 2

[39] Hao Luo, Zihao Yue, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Deheng Ye, and Zongqing Lu. OpenMMEgo: Enhancing egocentric understanding for lmms with open weights and data. In *NeurIPS*, 2025. 2

[40] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025. 3, 13

[41] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023. 2

[42] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. OpenEQA: Embodied question answering in the era of foundation models. In *CVPR*, 2024. 2

[43] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022. 2

[44] Yaru Niu, Yunzhe Zhang, Mingyang Yu, Changyi Lin, Chenhao Li, Yikai Wang, Yuxiang Yang, Wenhao Yu, Tingnan Zhang, Zhenzhen Li, Jonathan Francis, Bingqing Chen, Jie Tan, and Ding Zhao. Human2LocoMan: Learning versatile quadrupedal manipulation with human pretraining. In *RSS*, 2025. 2

[45] NVIDIA. GR00T-WholeBodyControl. https://github.com/NVlabs/GR00T-WholeBodyControl, 2025. 4, 13, 15, 17

[46] NVIDIA. GR00T N1.6: An improved open foundation model for generalist humanoid robots. https://research.nvidia.com/labs/gear/gr00t-n1_6/, 2025. 3

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 17

[48] Yixuan Pan, Ruoyi Qiao, Li Chen, Kashyap Chitta, Liang Pan, Haoguang Mai, Qingwen Bu, Cunyuan Zheng, Hao Zhao, Ping Luo, and Hongyang Li. Agility Meets Stability: Versatile humanoid control with heterogeneous data. In *ICRA*, 2026. 3, 13

[49] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy ˜ human policy. In *CoRL*, 2025. 2, 3

[50] TARS Robotics, Yuhang Zheng, Jichao Peng, Weize Li, Yupeng Zheng, Xiang Li, Yujie Jin, Julong Wei, Guanhua Zhang, Ruiling Zheng, et al. World In Your Hands: A large-scale and open-source ecosystem for learning human-centric manipulation in the wild. *arXiv preprint arXiv:2512.24310*, 2025. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4, 14, 17

[52] Ronald W. Schafer. What is a savitzky-golay filter? [lecture notes]. *IEEE SPM*, 2011. 4

[53] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025. 3

[54] Wandong Sun, Luying Feng, Baoshi Cao, Yang Liu, Yaochu Jin, and Zongwu Xie. ULC: A unified and fine-grained controller for humanoid loco-manipulation. *arXiv preprint arXiv:2507.06905*, 2025. 2

[55] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. DexWild: Dexterous human interactions for in-the-wild robot policies. In *RSS*, 2025. 2

[56] Ilyass Taouil, Haizhou Zhao, Angela Dai, and Majid Khadiv. Physically consistent humanoid loco-manipulation using latent diffusion models. In *HUMANOIDS*, 2025. 3

[57] Takara E Truong, Qiayuan Liao, Xiaoyu Huang, Guy Tevet, C Karen Liu, and Koushil Sreenath. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. 3

[58] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025. 4, 14, 17

[59] Tianyu Wang, Dwait Bhatt, Xiaolong Wang, and Nikolay Atanasov. Cross-embodiment robot manipulation skill transfer using latent space alignment. *arXiv preprint arXiv:2406.01968*, 2024. 3

[60] Zhenyu Wei, Zhixuan Xu, Jingxiang Guo, Yiwen Hou, Chongkai Gao, Zhehao Cai, Jiayu Luo, and Lin Shao. $\mathcal{D}(\mathcal{R}, \mathcal{O})$ Grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In *ICRA*, 2025. 3

[61] Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Yufei Jiang, Ran Huang, Guoying Gu, and Hongyang Li. FreeTacMan: Robot-free visuo-tactile data collection system for contact-rich manipulation. *arXiv preprint arXiv:2506.01941*, 2025. 3

[62] Boshen Xu, Yuting Mei, Xinbi Liu, Sipeng Zheng, and Qin Jin. EgoDTM: Towards 3d-aware egocentric video-language pretraining. In *NeurIPS*, 2025. 2

[63] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. In *CoRL*, 2025. 3

[64] Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, et al. LeVERB: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025. 3

[65] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025. 2, 3, 4

[66] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 2

[67] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *ICLR*, 2025. 2

[68] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *ICLR*, 2025. 2

[69] Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. VisualMimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025. 3

[70] Checheng Yu, Chonghao Sima, Gangcheng Jiang, Hai Zhang, Haoguang Mai, Hongyang Li, Huijie Wang, Jin Chen, Kaiyang Wu, Li Chen, Lirui Zhao, Modi Shi, Ping Luo, Qingwen Bu, Shijia Peng, Tianyu Li, and Yibo Yuan. χ_0: Resource-aware robust manipulation via taming distributional inconsistencies. *arXiv preprint arXiv:2602.09021*, 2026. 2

[71] Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. EgoMI: Learning active vision and whole-body manipulation from egocentric human demonstrations. *arXiv preprint arXiv:2511.00153*, 2025. 2, 4

[72] Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. MotionTrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025. 3, 4, 13

[73] Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. TWIST: Teleoperated whole-body imitation system. In *CoRL*, 2025. 2, 13

[74] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. TWIST2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025. 2

[75] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. In *RSS*, 2024. 2

[76] Hai Zhang, Siqi Liang, Li Chen, Yuxian Li, Yukang Xu, Yichao Zhong, Fu Zhang, and Hongyang Li. Sparse video generation propels real-world beyond-the-view vision-language navigation. *arXiv preprint arXiv:2602.05827*, 2026. 2

[77] Yuanhang Zhang, Yifu Yuan, Prajwal Gurunath, Tairan He, Shayegan Omidshafiei, Ali-akbar Agha-mohammadi, Marcell Vazquez-Chanlatte, Liam Pedersen, and Guanya Shi. FALCON: Learning force-adaptive humanoid loco-manipulation. *arXiv preprint arXiv:2505.06776*, 2025. 2, 3

[78] Siheng Zhao, Yanjie Ze, Yue Wang, C Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025. 2, 3

[79] Zhigen Zhao, Liuchuan Yu, Ke Jing, and Ning Yang. XRoboToolkit: A cross-platform framework for robot teleoperation. *arXiv preprint arXiv:2508.00097*, 2025. 4, 17

[80] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan CHEN, Pingrui Zhang, et al. FastUMI: A scalable and hardware-independent universal manipulation interface with dataset. In *CoRL*, 2025. 2, 3

[81] Rui Zhong, Yizhe Sun, Junjie Wen, Jinming Li, Chuang Cheng, Wei Dai, Zhiwen Zeng, Huimin Lu, Yichen Zhu, and Yi Xu. HumanoidExo: Scalable whole-body humanoid manipulation via wearable exoskeleton. *arXiv preprint arXiv:2510.03022*, 2025. 2, 3

[82] Jiaming Zhou, Teli Ma, Kun-Yu Lin, Zifan Wang, Ronghe Qiu, and Junwei Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In *CVPR*, 2025. 2

[83] 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. 2, 3, 4, 13