July 13, 2026
@RSS ExWBC Workshop

ARCHON

# Scaling Whole-Body Humanoid Skills with Human Demonstration

**Tianyu Li**
Co-Founder & CEO
Archon Robotics, Shanghai, China

PhD Researcher
OpenDriveLab, The University of Hong Kong

# Outline

1. **Background: Scaling Loco-Manipulation with Human Data**
2. WholebodyVLA (ICLR 2026): Pre-training from Egocentric Human Videos
3. EgoHumanoid (RSS 2026): Human-Centric Whole-Body Data Collection
4. The Road to Whole-Body Intelligence

# Impressive Skills Showcase for Humanoid

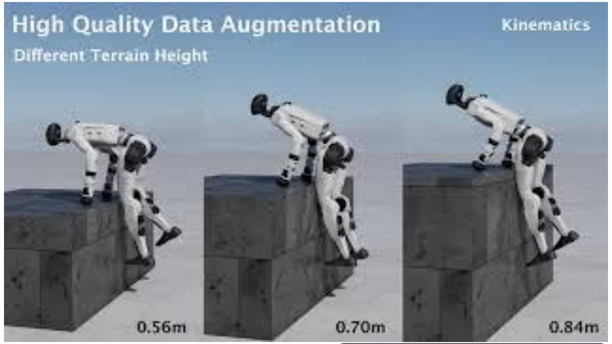
## Excellent Performance, Limited Task Generalization

ASAP, RSS'25 — Agility


HuB, CoRL'25 — Balance

SKILD AI LocoFormer, CoRL'25 — Adaptability

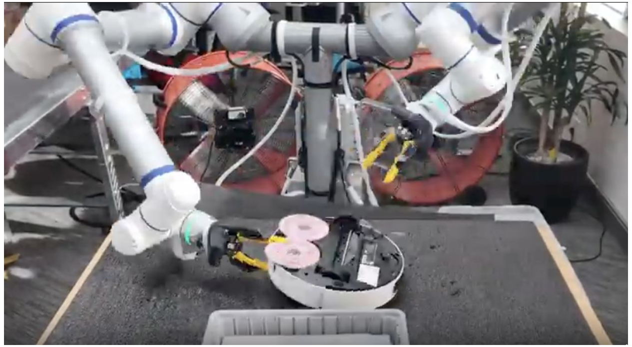

BeyondMimic, arXiv2508 — Agility




amazon OmniRetarget, ICRA'26 — Interaction

HoST, RSS'25 — Recovery

# Loco-manipulation systems need more generalizability



Generalist, manipulation

Figure, loco-manipulation

- A fixed-base manipulator's task distribution is bounded by its reach.
- A humanoid robot can walk to any scene and manipulate any object.
- **Loco-manipulation demands broader generalization because mobility exposes the robot to a far wider distribution of environments, objects, and tasks.**

# Ego-centric Dataset

## General Purpose

- Epic-Kitchens: egocentric cooking activities and actions (55 hrs)
- **Ego4D: large-scale egocentric video (3.7k hrs)**🌟
- Ego-Exo4D: paired egocentric + exocentric (2k hrs)
- HoloAssist: egocentric assistive tasks (AR/egocentric), coarse hand pose annot. (166 hrs)

| Dataset | # Traj. | # Tasks | # Frames | Lang. Annot. | Cam. Ext. | Dexterous Annot. | Collection Method |
|---|---|---|---|---|---|---|---|
| **RoboTurk** [21] | 2k | 3 | 12M | ✗ | ✗ | ✗ | teleoperation |
| **RoboNet** [9] | 162k | n/a | 15M | ✗ | ✗ | ✗ | scripted |
| **BridgeData V2** [38] | 60k | 13 | 2M | ✓ | ✗ | ✗ | teleop+scripted |
| **DROID** [16] | 76k | 86 | 19M | ✓ | ✓ | ✗ | teleoperation |
| **EgoMimic** [15] | 2k | 3 | 0.4M | ✗ | ✓ | ✗ | egocentric video |
| **EPIC-KITCHENS** [8] | 40k | 125 | 12M | ✓ | ✗ | ✗ | egocentric video |
| **HOI4D** [19] | 4k | 54 | 2M | ✗ | ✗ | ✓ | egocentric video |
| **Ego4D (HOI)** [11] | 89k | n/a | 21M | ✓ | ✗ | ✗ | egocentric video |
| **EgoDex (ours)** | **338k** | **194** | **90M** | ✓ | ✓ | ✓ | egocentric video |

EgoDex (https://arxiv.org/abs/2505.11709)

## Manipulation-oriented (3D Hand Pose)

- HOT3D: egocentric tasks with hand/object annotations (Meta) (833 mins)
- HOI4D: 4D hand-object interaction dataset (includes egocentric views) (4k seq)
- **EgoDex: largest w/ dexterous annotation. (829 hrs)**🌟
- TACO: tool-use/action-centric dataset (2.5k seq)

| Dataset | Type | Prior Works | # Hour | # Trajectory | # Skill | # Scene |
|---|---|---|---|---|---|---|
| RT-1 | Robot | IRASim, UniSim | 900 | 130k | 8 | 2 |
| BridgeData V2 | Robot | IRASim, UniSim, WorldGym, WMPE | 130 | 60.1k | 13 | 24 |
| Language-Table | Robot | IRASim, UniSim, HMA | 2.7k | 442k | - | - |
| RoboNet | Robot | iVideoGPT, IRASim | - | 162k | - | 10 |
| DROID | Robot | Ctrl-World, UWM | 350 | 76k | 86 | 564 |
| AgiBot-World | Robot | EnerVerse-AC | 2.9k | 1,000k | 87 | 106 |
| Nymeria | Human | PEVA, EgoTwin, EgoControl | 300 | 1.2k | - | 50 |
| In-lab | Human | - | 55 | 13.9k | 35 | 1 |
| EgoDex | Human | - | 829 | 30k | 194 | 5 |
| DreamDojo-HV | Human | - | 43,827 | 1,135k | 6,015† | 1,135k |
| **Our Mixture** | Human | - | 44,711 | 1,179k | ≥6,015† | ≥1,135k |

DreamDojo (https://arxiv.org/abs/2602.06949)

# Ego-centric Dataset

### General Purpose

- Epic-Kitchens: egocentric cooking activities and actions (55 hrs)
- **Ego4D: large-scale egocentric video (3.7k hrs)**🌟
- Ego-Exo4D: paired egocentric + exocentric (2k hrs)
- HoloAssist: egocentric assistive tasks (AR/egocentric), coarse hand pose an

| Dataset | # Traj. | # Tasks | # Frames | Lang. Annot. | Cam. Ext. | Dexterous Annot. | Collection Method |
|---|---|---|---|---|---|---|---|
| **RoboTurk** [21] | 2k | 3 | 12M | ✗ | ✗ | ✗ | teleoperation |
| **RoboNet** [9] | 162k | n/a | 15M | ✗ | ✗ | ✗ | scripted |
| **BridgeData V2** [38] | 60k | 13 | 2M | ✓ | ✗ | ✗ | teleop+scripted |
| **DROID** [16] | 76k | 86 | 19M | ✓ | ✓ | ✗ | teleoperation |
| **EgoMimic** [15] | 2k | 3 | 0.4M | ✗ | ✓ | ✗ | egocentric video |
| **EPIC-KITCHENS** [8] | 40k | 125 | 12M | ✓ | ✗ | ✗ | egocentric video |
| **HOI4D** [19] | 4k | 54 | 2M | ✗ | ✗ | ✓ | egocentric video |
| **Ego4D (HOI)** [11] | 89k | n/a | 21M | ✓ | ✗ | ✗ | egocentric video |
| | | | | | | ✓ | egocentric video |

709)

**No whole-body Pose !**

### Manipulation-oriente

- HOT3D: egocentric t annotations (Meta) (833 mins)
- HOI4D: 4D hand-object interaction dataset (includes egocentric views) (4k seq)
- **EgoDex: largest w/ dexterous annotation. (829 hrs)**🌟
- TACO: tool-use/action-centric dataset (2.5k seq)

| | | | | ectory | # Skill | # Scene |
|---|---|---|---|---|---|---|
| RT-1 | Robot | IRASim, UniSim | 900 | 130k | 8 | 2 |
| BridgeData V2 | Robot | IRASim, UniSim, WorldGym, WMPE | 130 | 60.1k | 13 | 24 |
| Language-Table | Robot | IRASim, UniSim, HMA | 2.7k | 442k | - | - |
| RoboNet | Robot | iVideoGPT, IRASim | - | 162k | - | 10 |
| DROID | Robot | Ctrl-World, UWM | 350 | 76k | 86 | 564 |
| AgiBot-World | Robot | EnerVerse-AC | 2.9k | 1,000k | 87 | 106 |
| Nymeria | Human | PEVA, EgoTwin, EgoControl | 300 | 1.2k | - | 50 |
| In-lab | Human | - | 55 | 13.9k | 35 | 1 |
| EgoDex | Human | - | 829 | 30k | 194 | 5 |
| DreamDojo-HV | Human | - | 43,827 | 1,135k | 6,015† | 1,135k |
| **Our Mixture** | Human | - | 44,711 | 1,179k | ≥6,015† | ≥1,135k |

DreamDojo (https://arxiv.org/abs/2602.06949)

# Richer supervision preserves more of the strategy

LEVEL 1

## Ego video only

**RGB + task context**

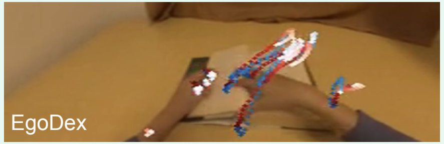
LEARN

- Task semantics
- Visual intent
- Action structure

**Action is inferred indirectly**

LEVEL 2

## Ego + EE pose




**RGB + hand / wrist trajectory**

LEARN

- Manipulation trajectory
- Hand–eye coordination
- Task-space action

**Lower-body motion is compressed**

LEVEL 3

## EgoBody

**RGB + whole-body pose**

LEARN

- Weight shift & stepping
- Leaning, kneeling, bracing
- Whole-body coordination

**Highest capture burden—until now**

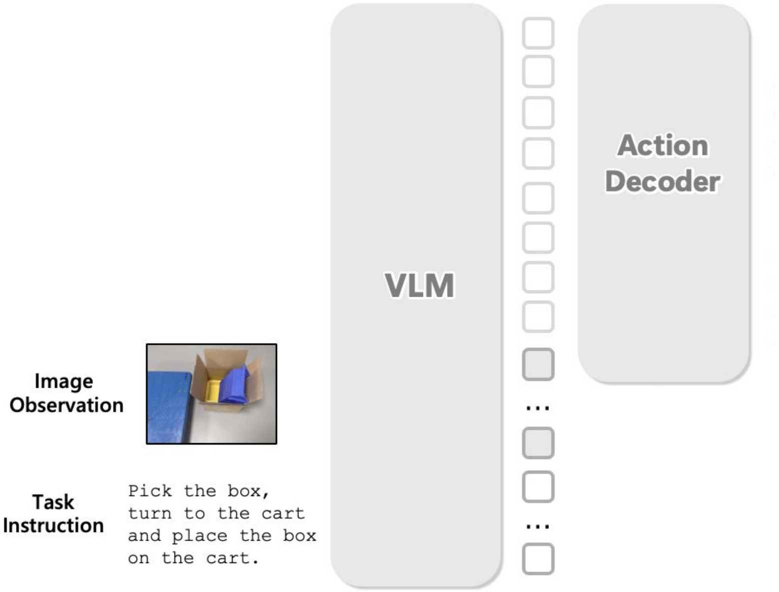
**More human motion preserved →**

**Data scale determines coverage. Body supervision determines what can be learned.**

# Outline

1. Background: Scaling Humanoid Loco-Manipulation
2. **WholebodyVLA (ICLR 2026):** Pre-training from Egocentric Human Videos
3. EgoHumanoid (RSS 2026): Human-Centric Whole-Body Data Collection
4. The Road to Whole-Body Intelligence

# Basic: Vision-Language-Action (VLA) Framework

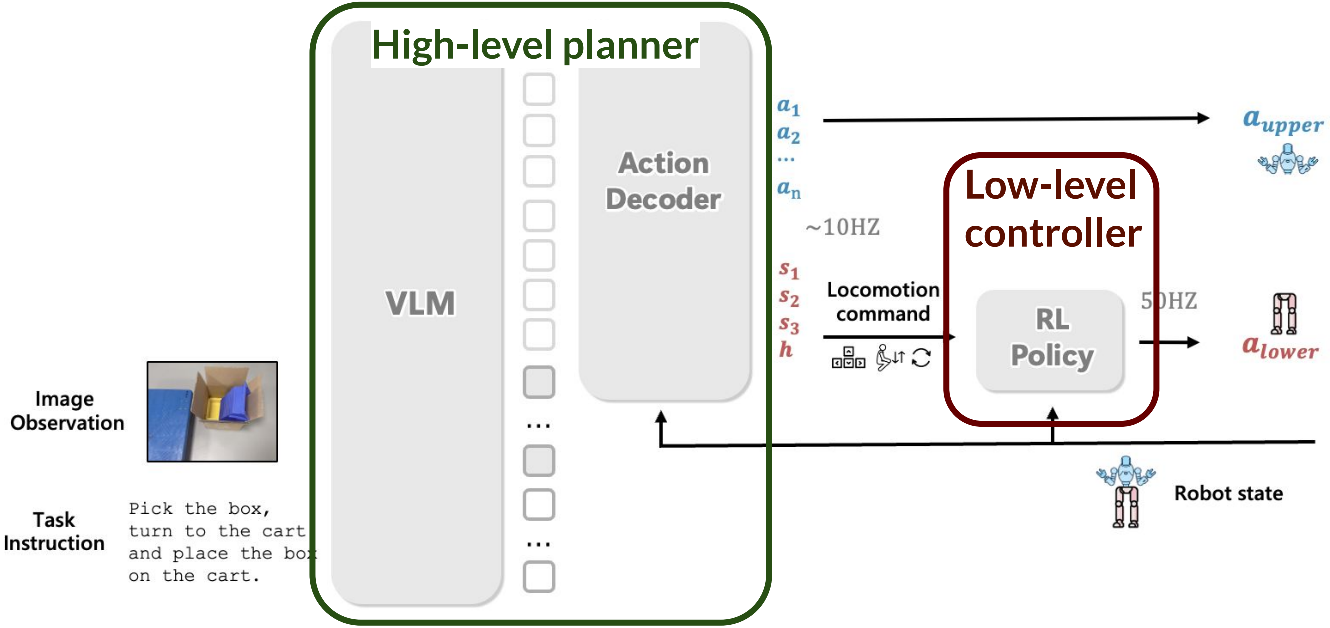




$\pi_0$


**GEN-0**

# Humanoid Pipeline using VLA: The Setup




**High-level planner** (VLM + action decoder): **perception, planning** ⟶
- Upper-body action
- Locomotion command

**Low-level controller** (RL policy): **locomotion under command** ⟶
- Lower-body action

# Challenge - Data

## Scarcity of loco-manipulation data

### Agibot World


Tabletop manipulation

Large-scale data

### Room-to-Room


### Room-Across-Room


Navigation on wheeled or quadruped platform

# Challenge - Data

## Scarcity of loco-manipulation data

### Agibot World


Tabletop manipulation

**Large-scale data**

### Room-to-Room


### Room-Across-Room


Navigation on wheeled or quadruped platform

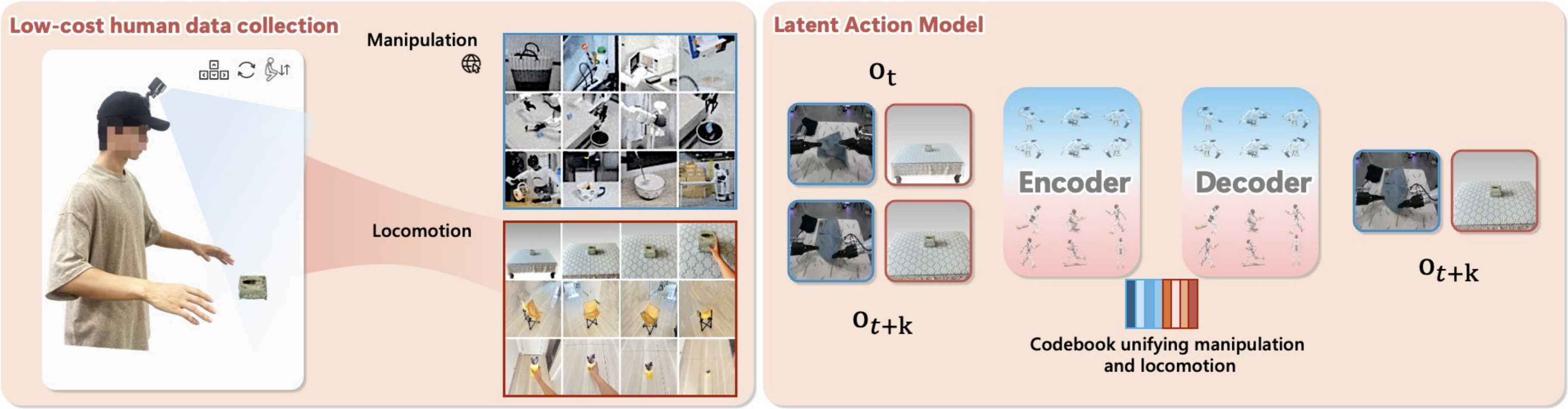
**Manipulation and navigation are **separate** tasks.**

# How to obtain loco-manip. knowledge?

## Learn from human/action-free videos

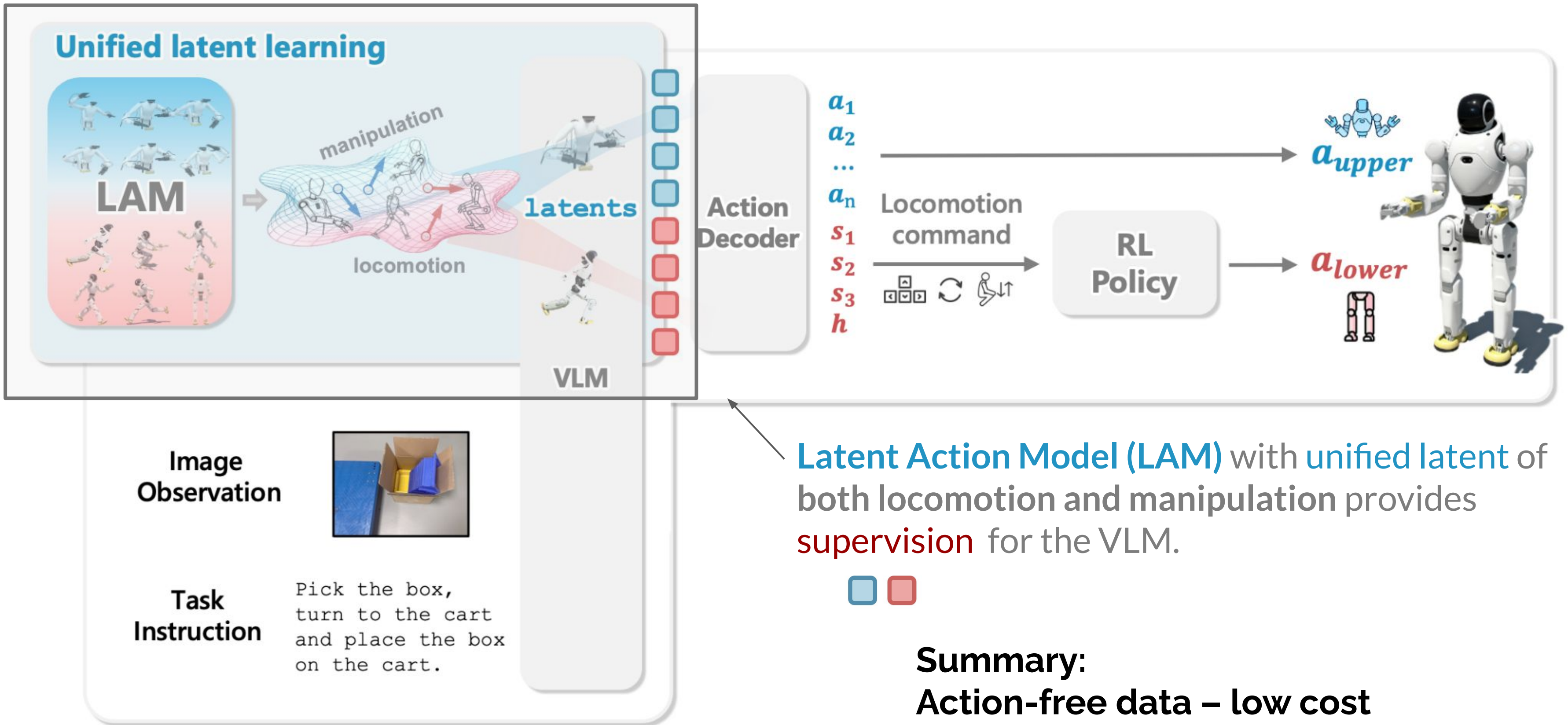

Locomotion:
Low-cost egocentric video - locomotion with manipulation task involved

# Method | Unified Latent Learning




Latent Action Model (LAM) is pretrained, following LAPA (https://arxiv.org/abs/2410.11758).

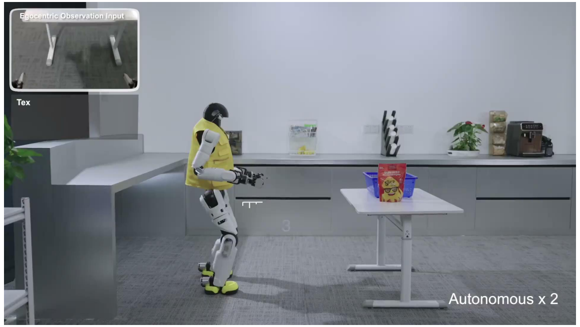
# Method | Unified Latent Learning




Latent Action Model (LAM) with unified latent of **both locomotion and manipulation** provides supervision for the VLM.

**Summary:**
**Action-free data – low cost**
**Fewer data due to LAM structure**

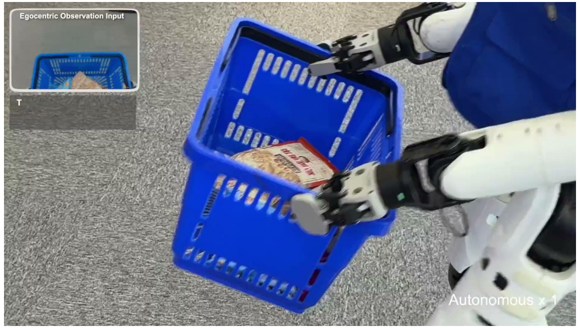
# Demo | WholeBodyVLA



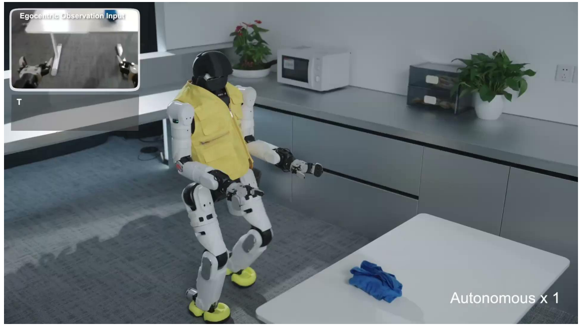

Start-pose Gen.


Object Gen.


Long-horizon




Go-around





Extend to everyday task

# WholeBodyVLA

## WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control

**Haoran Jiang**[1,2,4*] **Jin Chen**[1,2,4*] **Qingwen Bu**[2] **Li Chen**[2] **Modi Shi**[4,2]
**Yanjie Zhang**[3] **Delong Li**[3] **Chuanzhe Suo**[3] **Chuang Wang**[3] **Zhihui Peng**[3†] **Hongyang Li**[2†]

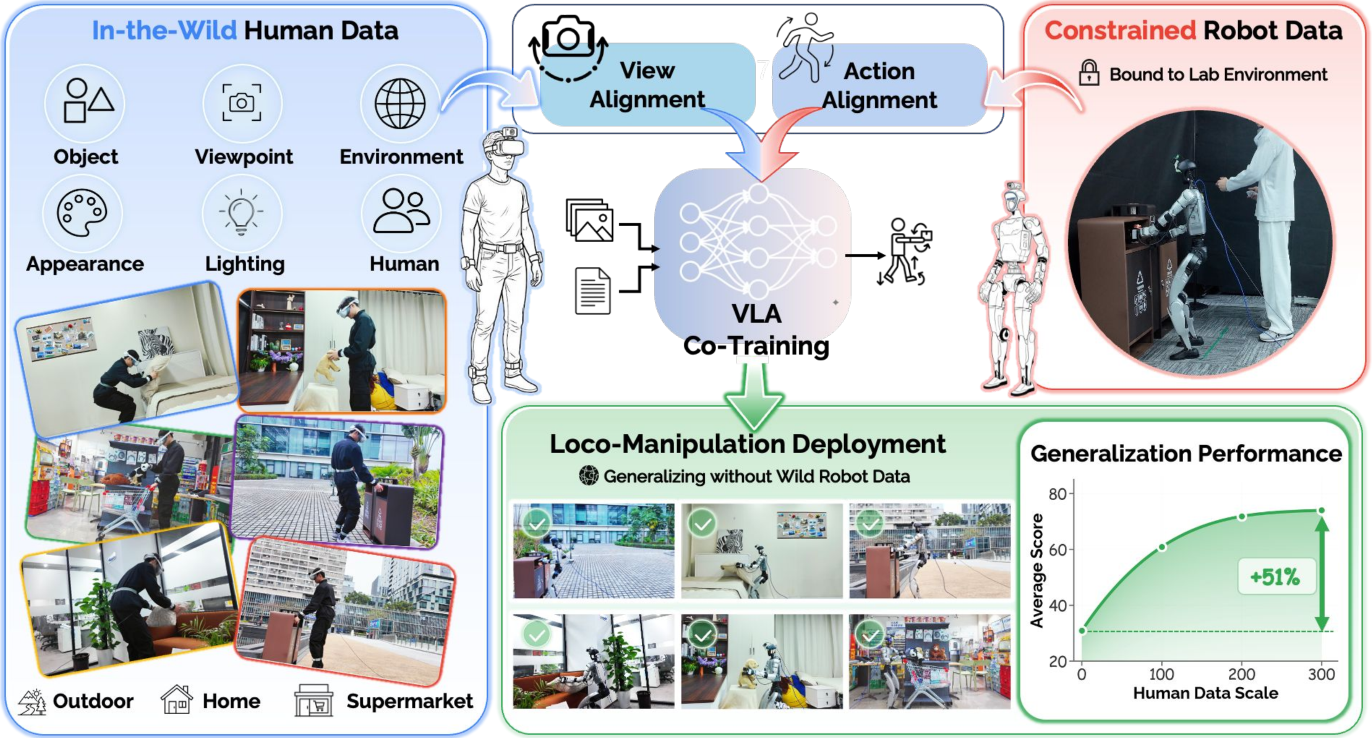
[1]Fudan University [2]OpenDriveLab & MMLab at The University of Hong Kong [3]AgiBot [4]SII
[*]Equal contribution [†]Project co-lead

# Outline

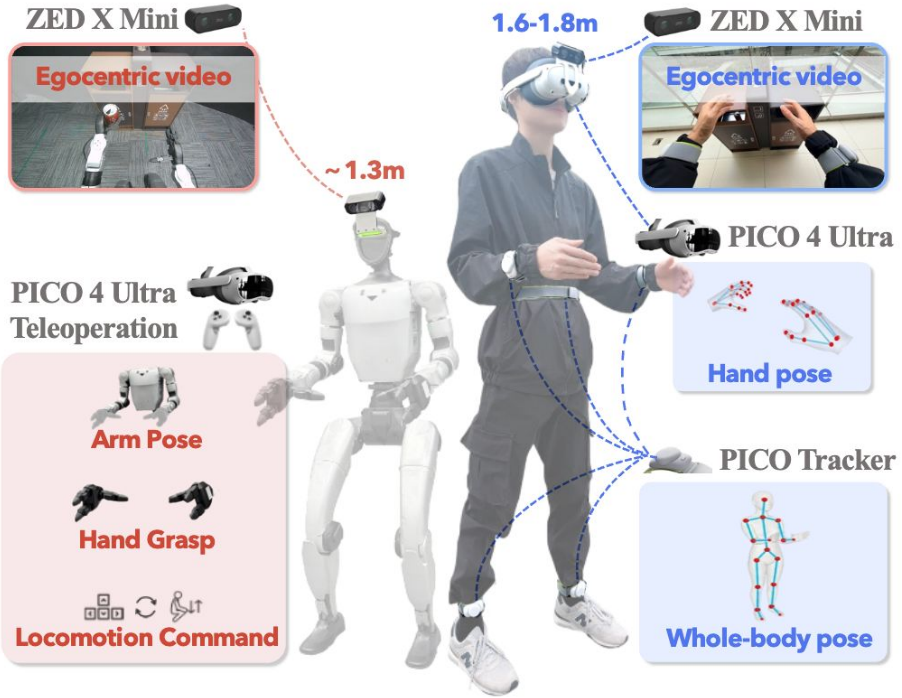
1. Background: Scaling Humanoid Loco-Manipulation
2. WholebodyVLA (ICLR 2026)**:** Pre-training from Egocentric Human Videos
3. **EgoHumanoid (RSS 2026)**: Human-Centric Whole-Body Data Collection
4. The Road to Whole-Body Intelligence

# EgoHumanoid | Human Robot Co-training

## Human data with whole-body pose & Robot data



# Data Collection




## Robot Data Collection

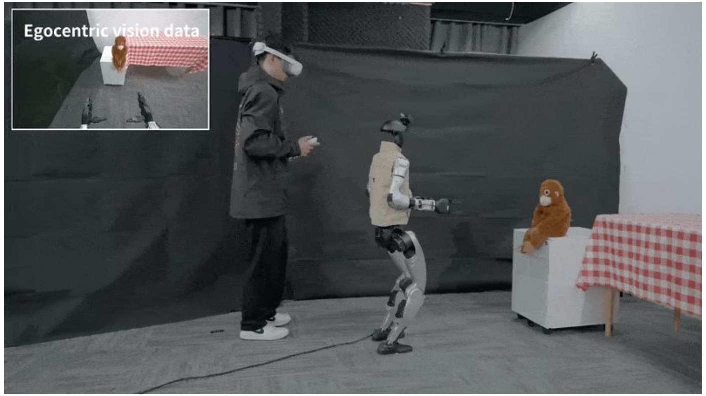
- ZED Camera
- UniTree G1
- Pico 4 Ultra
- Pico Controller * 2

## Human Data Collection

- ZED Camera
- Pico 4 Ultra
- Pico Motion Tracker * 5

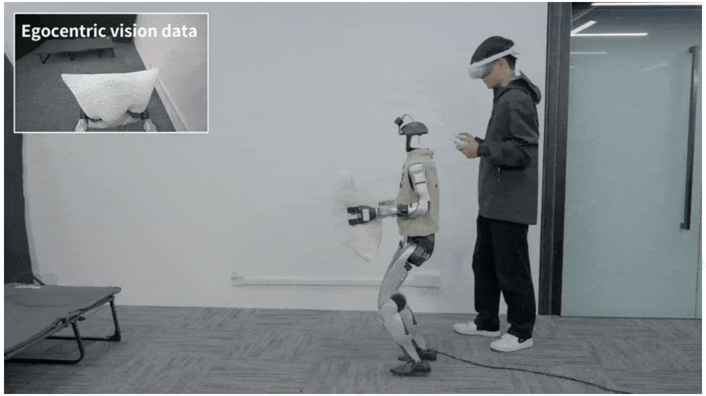
# Data Collection | Laboratory Robot Data


Cart Stowing



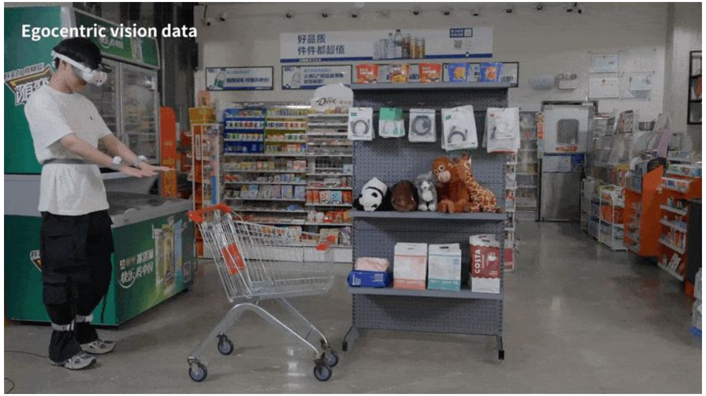

Toy Transfer


Trash Disposal




Pillow Placement

# Data Collection | Diverse Human Data



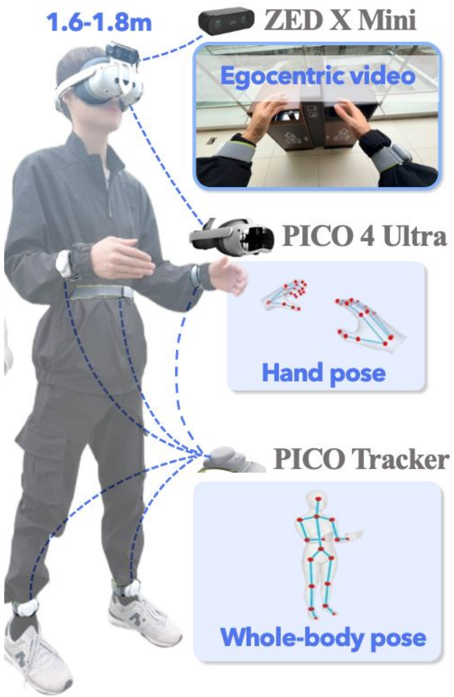

Cart Stowing

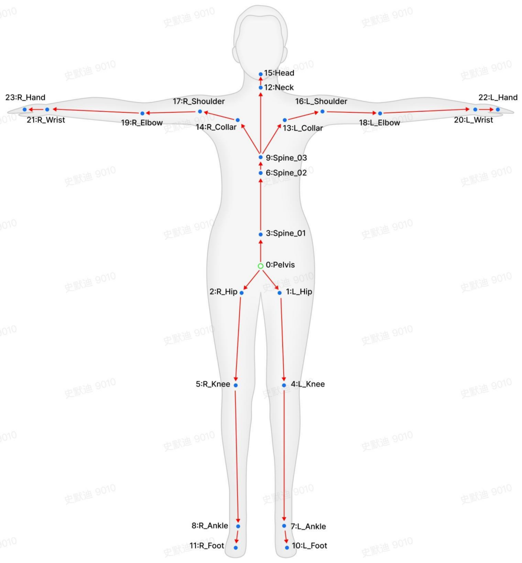

Toy Transfer


Trash Disposal

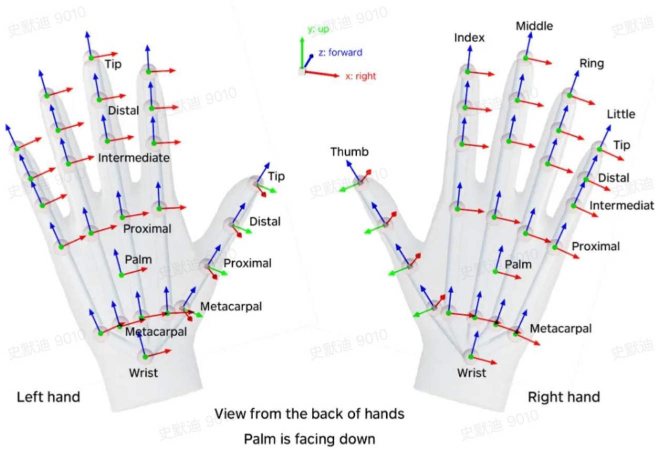

Pillow Placement

# Data Collection



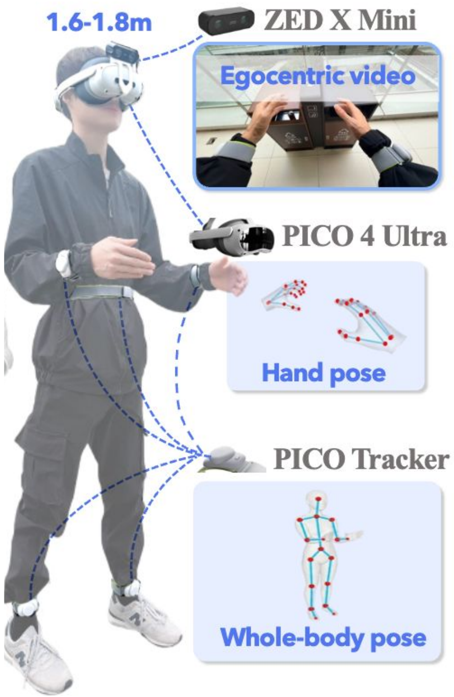



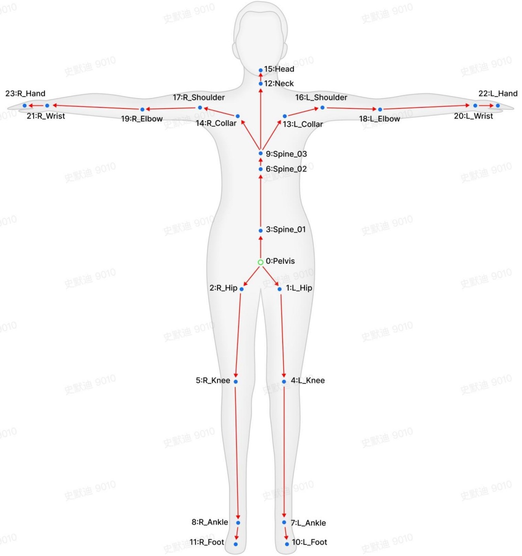

24 dof whole body pose



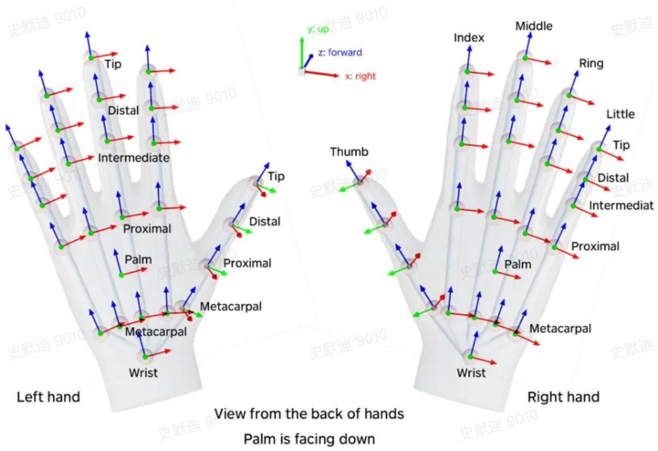

2*26 dof hand pose

# Data Collection



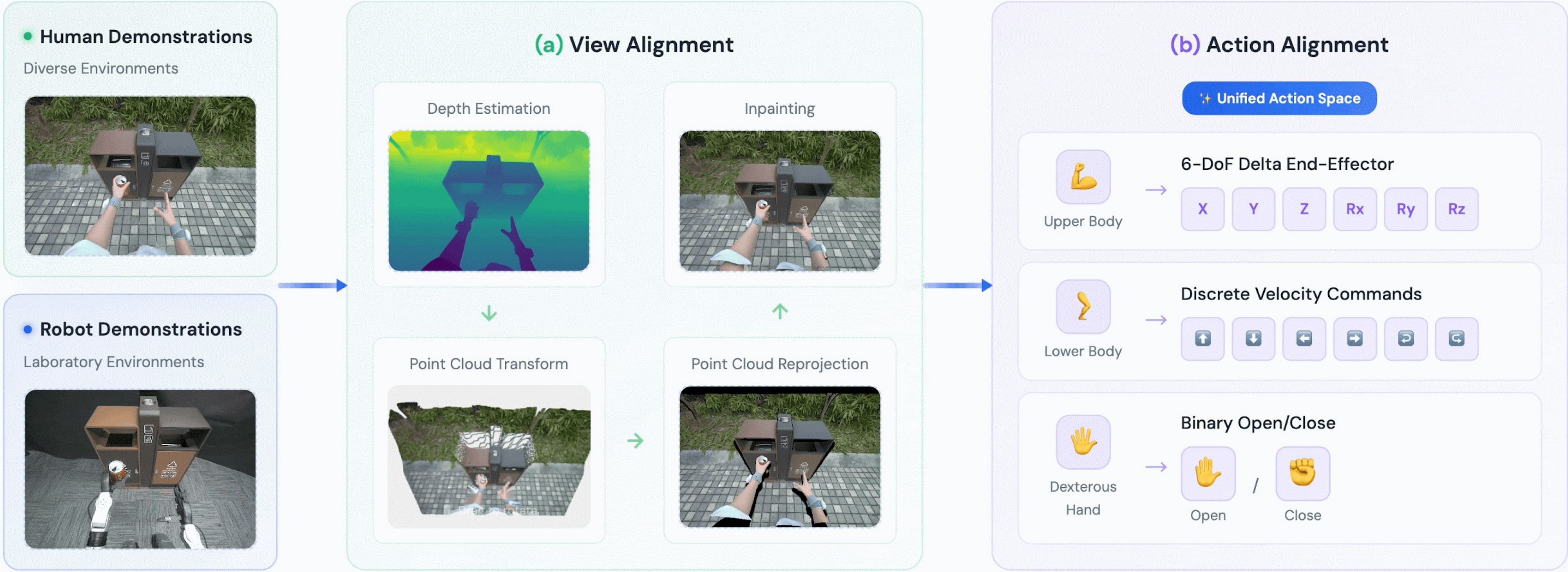

24 dof whole body pose




2*26 dof hand pose

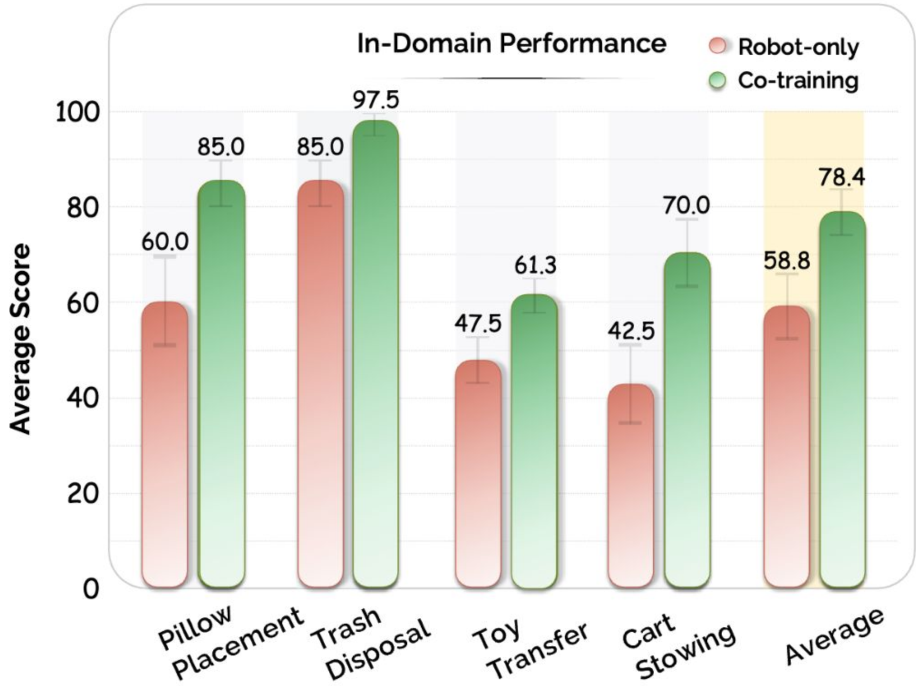
**Human & Robot Domain Gap: Visual Viewpoint & Action representation**

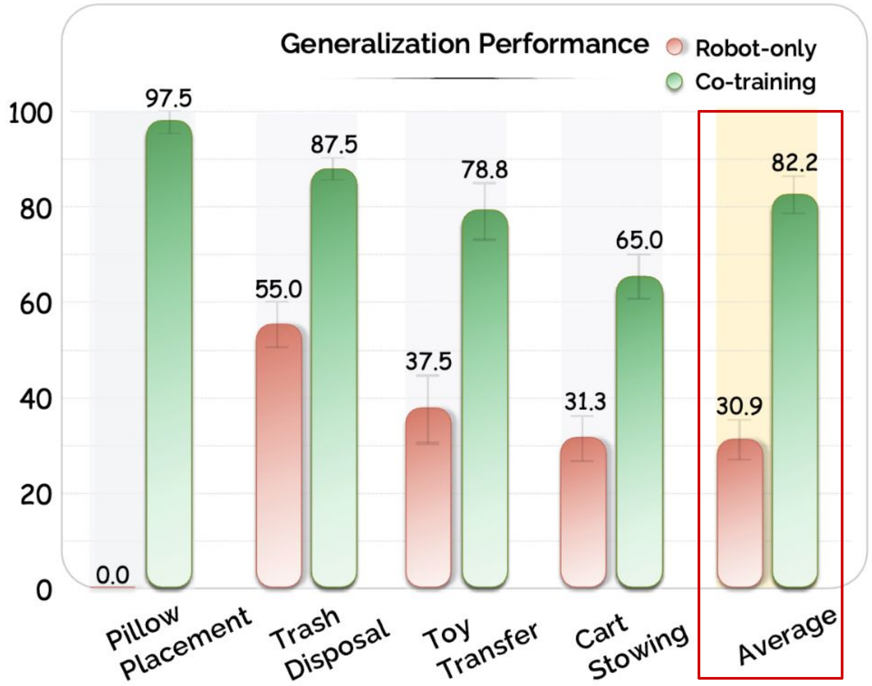
# Pipeline Overview




**Visual Gap** **Action Gap**

# Experiment



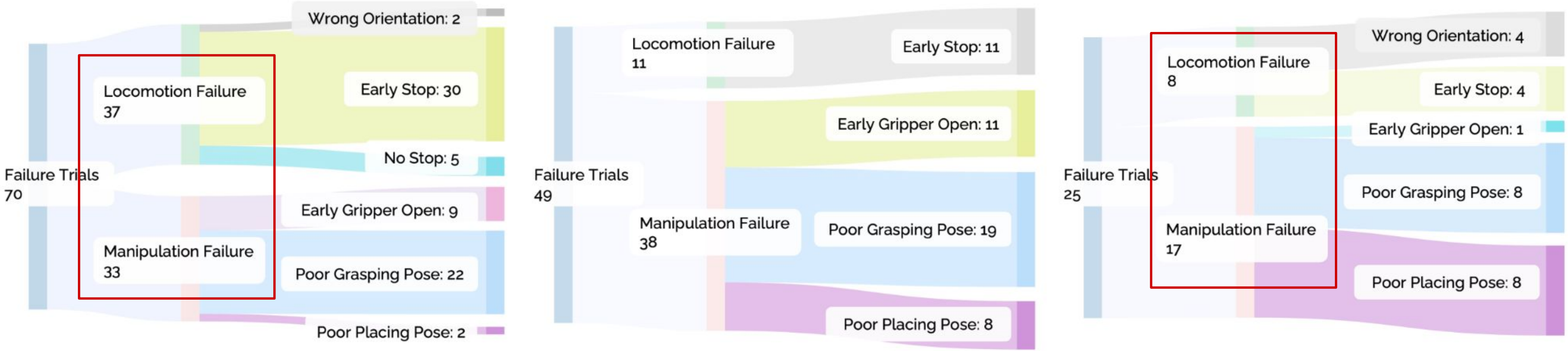




**Co-training boosts robot performance, especially in unseen settings (+51.3%).**

# Experiment

● Navigation ● Manipulation

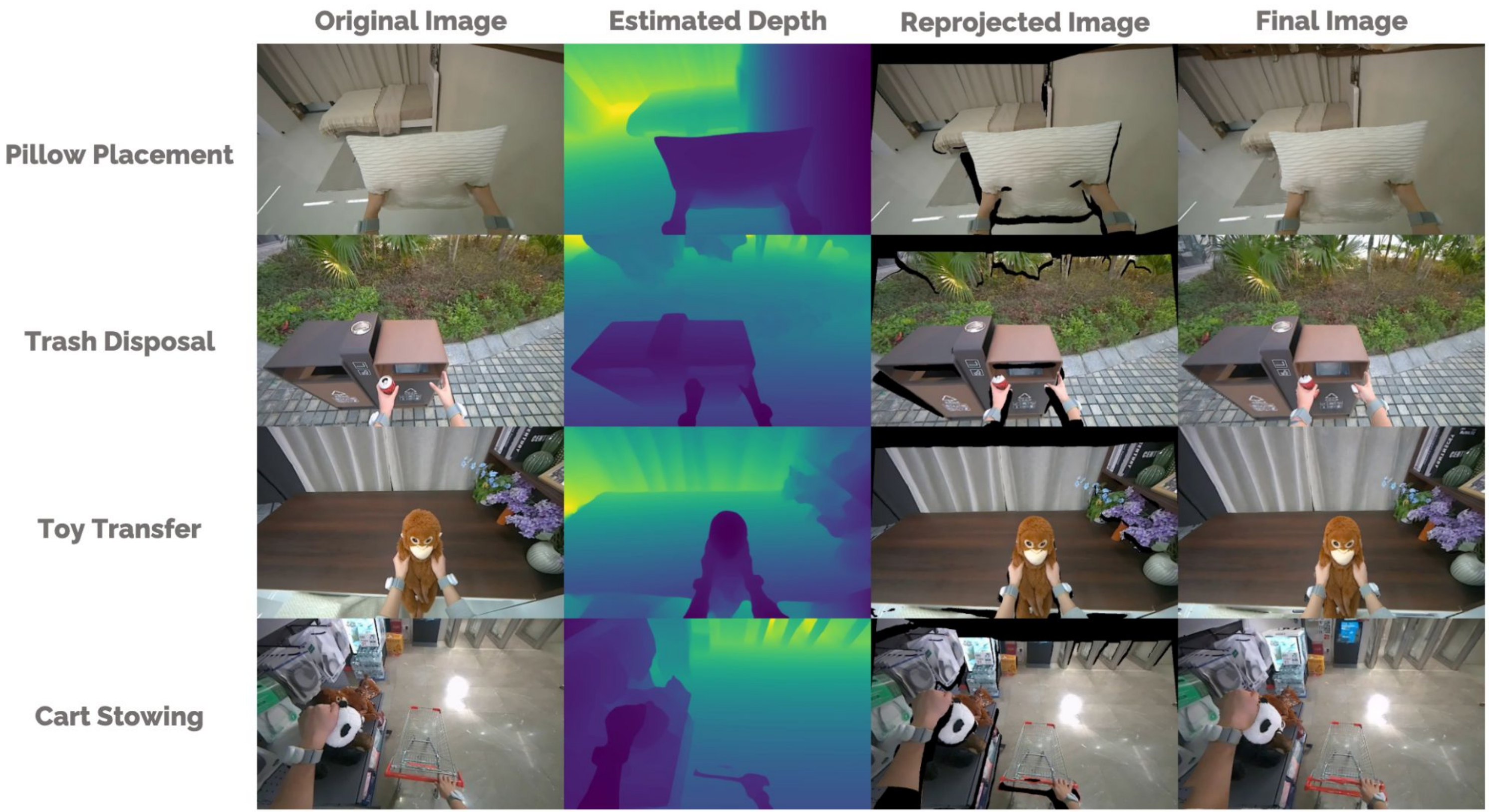
| Training Data | Pillow Placement | | Trash Disposal | | Toy Transfer | | | | Cart Stowing | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | s1 | s2 | s1 | s2 | s1 | s2 | s3 | s4 | s1 | s2 | s3 | s4 |
| Robot-only | 0 | 0 | 65 | 45 | 100 | 50 | 0 | 0 | 100 | 15 | 5 | 5 |
| Human-only | 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** |



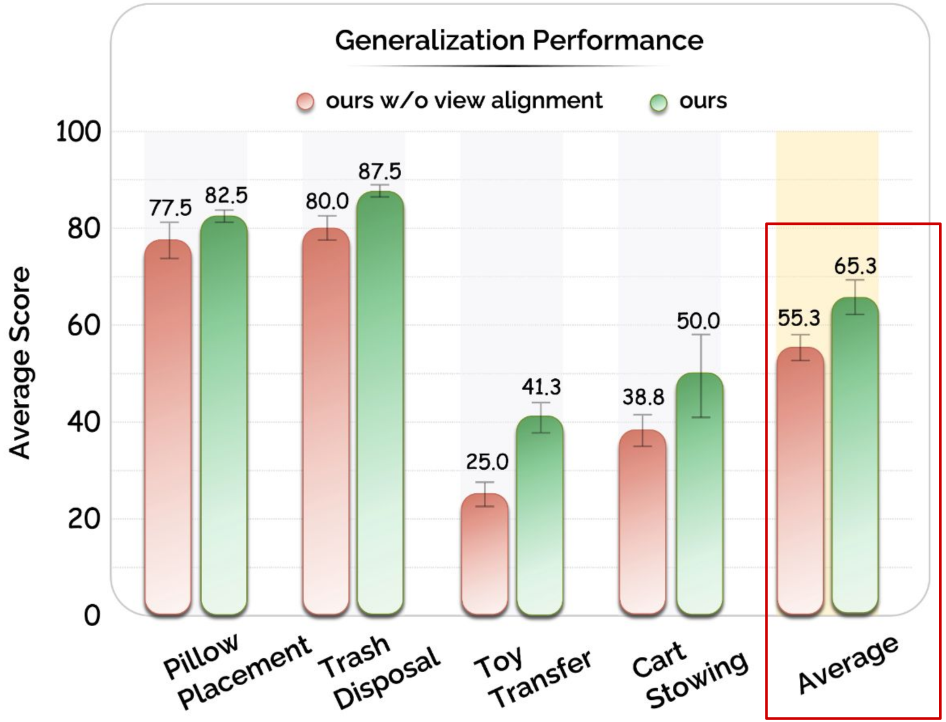

(a) Robot-only (b) Human-only (c) Co-training

**Navigation transfers better from human data**

# Experiment | View Alignment Visualization



# Experiment




**View alignment improves performance by bridging camera height gaps**

# Zero-Shot Deployment

# EgoHumanoid

## 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

# Outline

1. Background: Scaling Humanoid Loco-Manipulation
2. WholebodyVLA (ICLR 2026): Pre-training from Egocentric Human Videos
3. EgoHumanoid (RSS 2026): Human-Centric Whole-Body Data Collection
4. **The Road to Whole-Body Intelligence**

# Loco-manipulation is not locomotion + manipulation

**Shift weight while pulling dish rack**

**Use foot as an additional contact**

A decomposition into **upper-body manipulation** and **lower-body commands** inevitably discards useful whole-body coordination.

**@Figure, Helix-02**

# Our path climbs the human-data hierarchy

STEP 1 · EGO VIDEO

## WholeBodyVLA

### Learn intent at scale

Action-free human videos
+ robot post-training

POLICY OUTPUT

Upper-body action
+ discrete locomotion

STEP 2 · WHOLE-BODY DATA

## EgoHumanoid

### Capture coordination

Robot-free EgoBody data
+ human–robot co-training

POLICY OUTPUT

Upper-body action
+ discrete locomotion

NEXT · WHOLE-BODY INTELLIGENCE

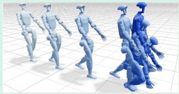
## Humanoid Foundation Model

### Preserve the full strategy

EgoBody pre-training
+ robot grounding

POLICY OUTPUT

**Whole-body motion trajectory**



More human motion preserved →

richer human supervision

**From learning human intent to generating executable whole-body motion.**

# The missing piece is whole-body intelligence.

Not a hand-off between an upper-body policy and a lower-body command — one interface, the whole body.

STAGE 1 · DATA

### EgoBody data pretrain

Egocentric + whole-body human pose demonstration + co-train with robot data

STAGE 2 · Planner

### Humanoid Foundation Model

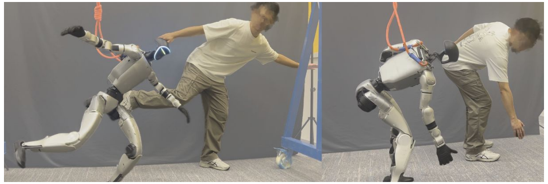
Outputs **whole-body robot motion trajectory**

STAGE 3 · Controller

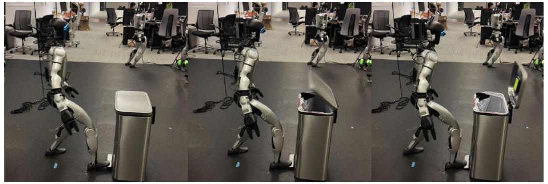
### Universal tracker

RL whole-body tracking controller — **credit: TWIST2, SONIC, AMS...**

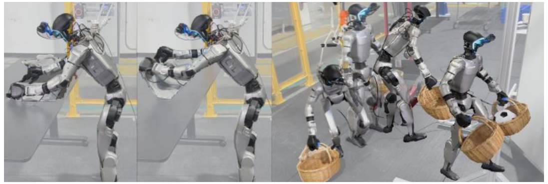
## Universal trackers:



Agility Meets Stability



SONIC



TWIST2


ARCHON
Whole-body Manipulation of Heavy Objects
Narrow Space Traversal
Environment-Leveraged Loco-Manipulation
Long-range Reachability

# Thanks

Q&A

tianyu@archon.tech
https://sephyli.github.io/