香 港 大 學
THE UNIVERSITY OF HONG KONG

June 1, 2026
@CVPRinParis Event

Whole-body Intelligence with Human-centric Data at Scale

Hongyang Li
Assistant Professor
The University of Hong Kong

Tianyu Li
Senior Research Assistant
The University of Hong Kong

Outline

1. Background
- Time for loco-manipulation, aka whole-body control
- Why and why now?
- VLA and Data effort

2. Algorithm part: WholeBodyVLA (ICLR 2026)

3. Data part: Human-centric data collection (RSS 2026)

4. Challenges and Future Work

Impressive Skills Showcase

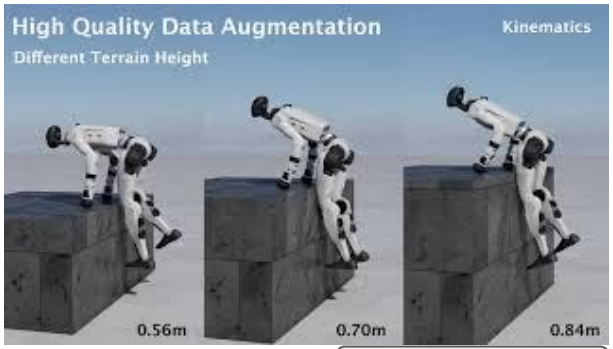
Excellent Performance, Limited Task Generalization

Carnegie Mellon University *ASAP, RSS'25* — Agility

清華大學 Tsinghua University *HuB, CoRL'25* — Balance

SKILD AI *LocoFormer, CoRL'25* — Adaptability

Berkeley UNIVERSITY OF CALIFORNIA *BeyondMimic, arXiv2508* — Agility



amazon *OmniRetarget, ICRA'26* — Interaction

HoST, RSS'25 — Recovery

Motivation

Autonomous Humanoid System in Early 2025

GR00T N1

Figure

Upper-body tasks only

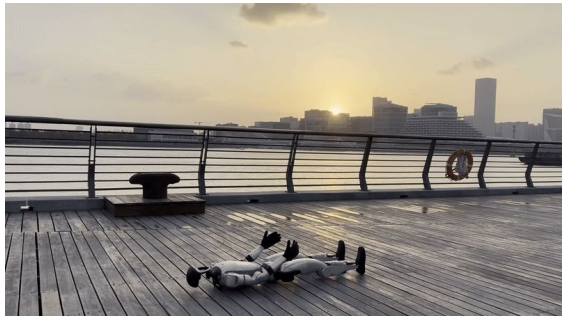
Motivation

Autonomous Humanoid System in Early 2025

GR00T N1

Figure

Upper-body tasks only

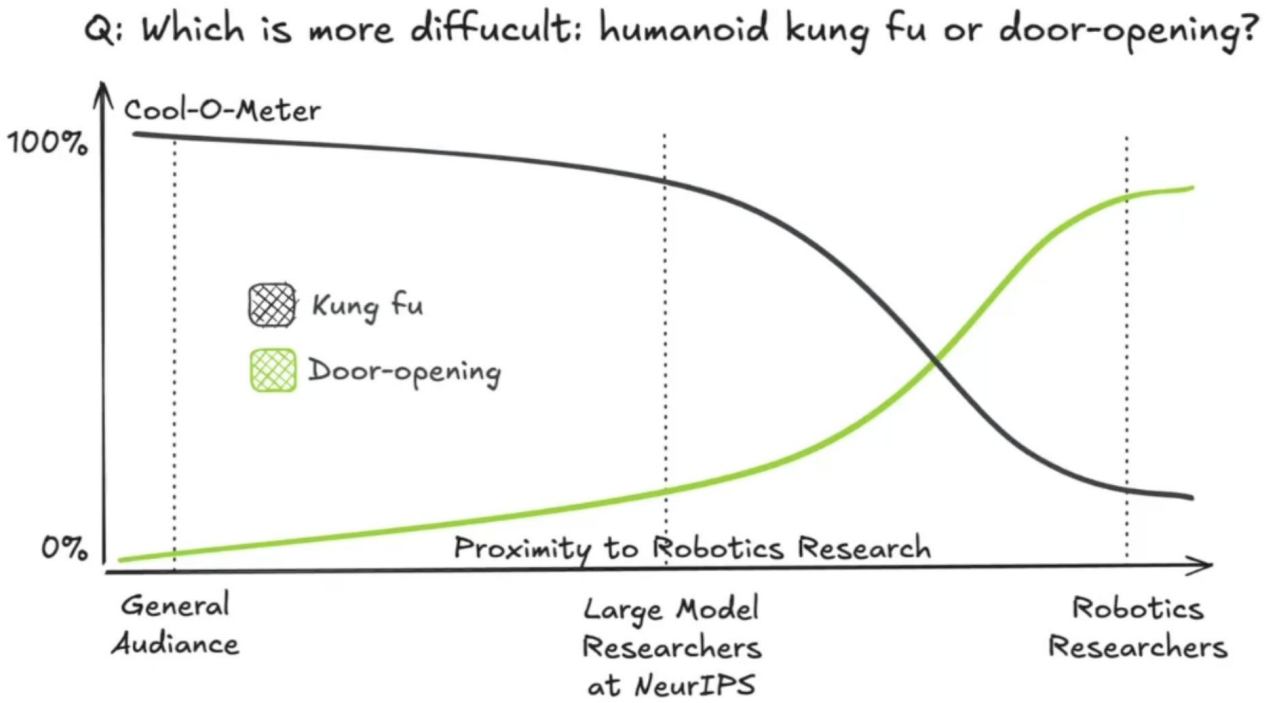


HoST

BeyondMimic

Pure locomotion

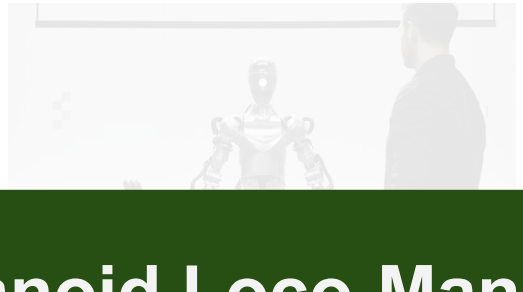
Motivation



from DoorMan, Haoru Xue, Nvidia,
https://www.youtube.com/watch?v=0lHC6Et10V4

Motivation: Simultaneous Manipulation + Locomotion

Autonomous Humanoid System in Early 2025

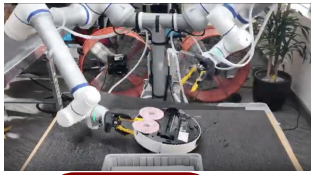


Upper-body tasks only

Humanoid Loco-Manipulation

Pure locomotion

Paradigm Shift: from Upper Body Manipulation to Whole-body Intelligence



Upper body Manipulation

Current Physical World Models VLA / World Model

Digital World

Physical World

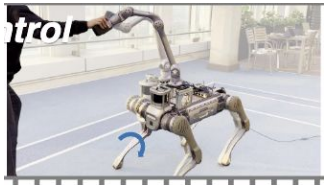
?

2023-2025

2026 - 2030

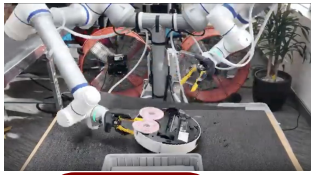
Lower body: locomotion

2010～ - 2024



- Learned to dance, punch
- Spring Festival Gala performance

Paradigm Shift: from Upper Body Manipulation to Whole-body Intelligence



Upper body Manipulation

Current Physical World Models VLA / World Model

Digital World

Physical World

?

2023-2025

2026 - 2030

Fully body Coordination

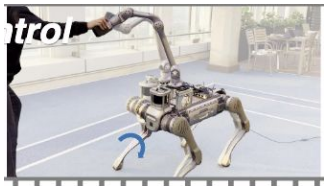
Large Humanoid Model (LHM)

End-to-End Full-Body Manipulation & Motion Control

- Realize Humanoid Robot Scaling Law
- Environment, Object, and Human Interaction

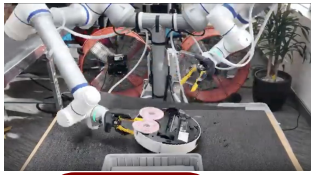
2010～ - 2024

Lower body: locomotion



- Learned to dance, punch
- Spring Festival Gala performance

Paradigm Shift: from Upper Body Manipulation to Whole-body Intelligence



Upper body Manipulation

Digital World

2023-2025

2010～ - 2024

Lower body: locomotion

Physical World

2026 - 2030

Large Humanoid Model (LHM)

End-to-End Full-Body Manipulation & Motion Control

Fully body Coordination

Data? Algorithm? Compute?

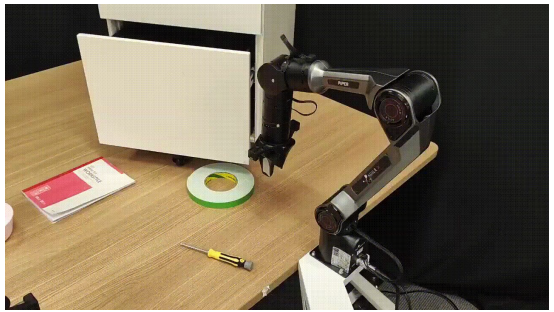
Algorithm challenge: VLA the only solution?

VLA is cool, but….

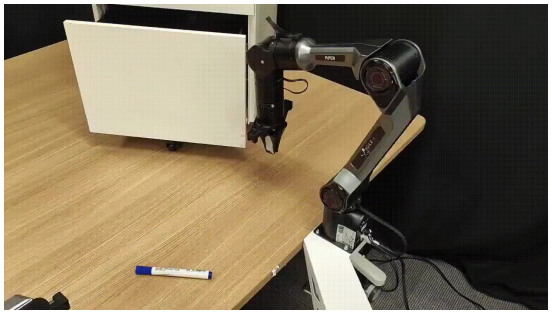
Our “toy” demos:

Lightning Variation

Visual Distractors



Novel Objects

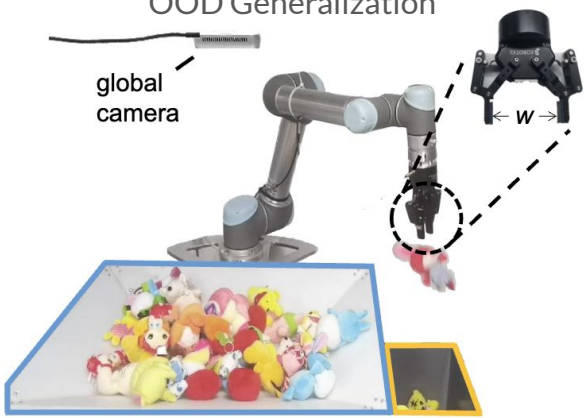


Algorithm challenge: VLA the only solution?

VLA is cool, but….

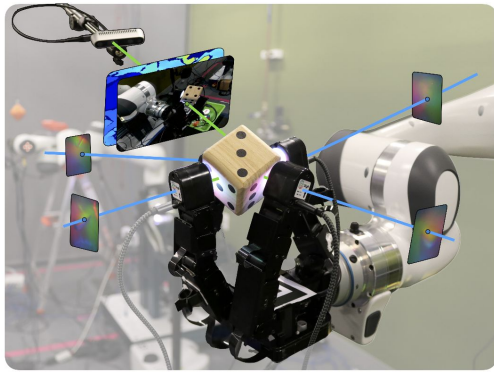
- ✕ Struggles to Generalize
- ✕ Not Really Dextrous
- ✕ Edge-side Responsive Control

OOD Generalization



AnyGrasp

In-hand Manipulation



NeuralFeels

Highly Dynamic Tasks

Aerobatic-Planner

Fang, Hao-Shu, et al. "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains." IEEE *TRO (2023)*

Suresh, Sudharshan, et al. "NeuralFeels with neural fields: Visuotactile perception for in-hand manipulation." *Science Robotics* (2024)

Wang, Mingyang, et al. "Unlocking aerobatic potential of quadcopters: Autonomous freestyle flight generation and execution." *Science Robotics*（2025）

Data Perspective: Where is RoboGPT?

Assuming 238 words/minute, 1.33 token/word

Open-X-embodiment (OXE)
4k Hours

GPT-2
475k Hours

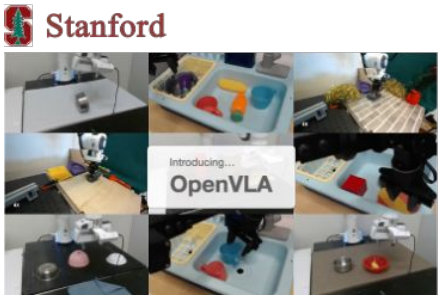
Pi0 Data
10k Hours

LLaMA-3

790M Hours
Or 0.79B Hours

Task/Generalization: Where is RoboGPT?

- **Short-horizon**
- **Simple task** (e.g. pick and place)
- **Controlled Lab Environment**



Vuong Q, Levine S, Walke H R, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. ICRA 2024

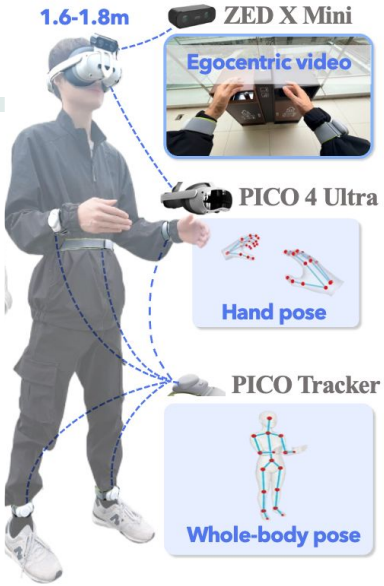
Data Challenge

Task / Embodiment Specific Data

X-Embodiment Robot Data

Open X-Embodiment Dataset

Action-less Web Videos



Human-centric data learning

Outline

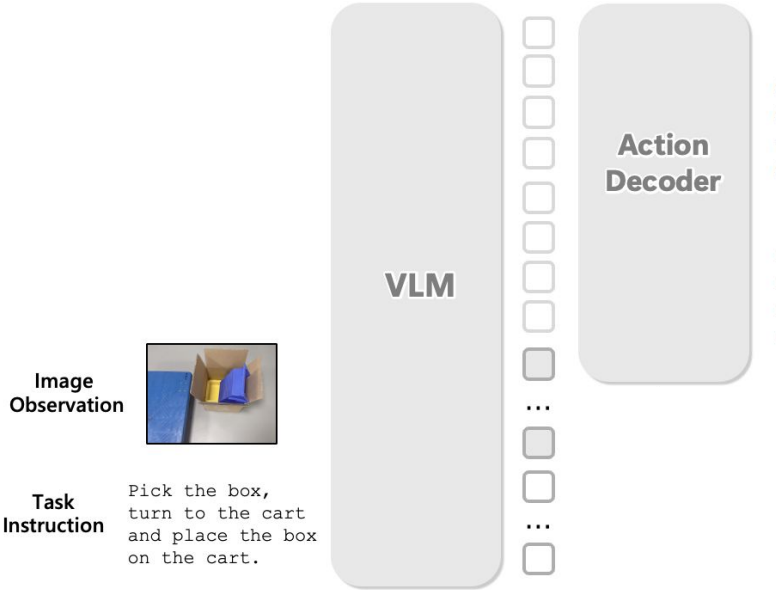
1. Background
   - Time for loco-manipulation, aka whole-body control
   - Why and why now?
   - VLA and Data effort

2. Algorithm part: WholebodyVLA (ICLR 2026)

3. Data part: Human-centric data collection (RSS 2026)

4. Challenges and Future Work

Basic: Vision-Language-Action (VLA) Framework



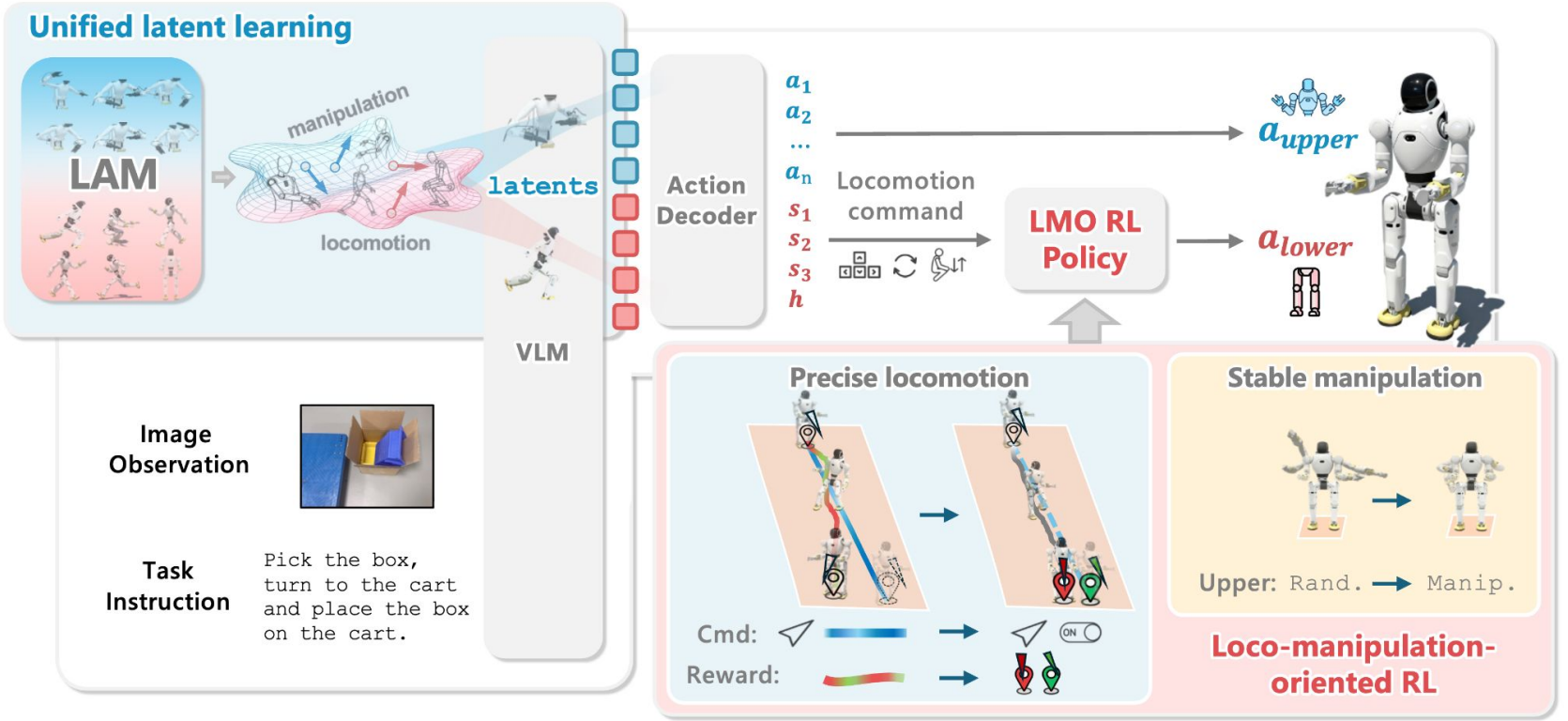
π_0

GEN-0

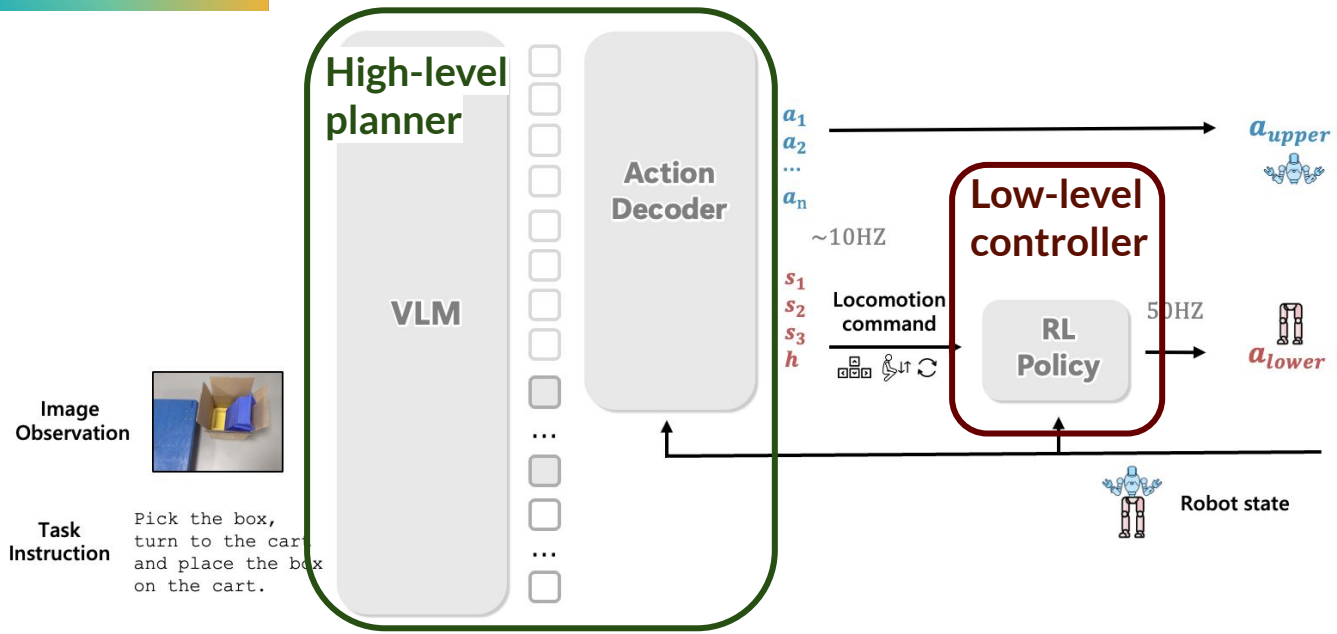
WholeBodyVLA

https://opendrivelab.com/WholeBodyVLA

ICLR 2026



Humanoid Pipeline using VLA: The Setup

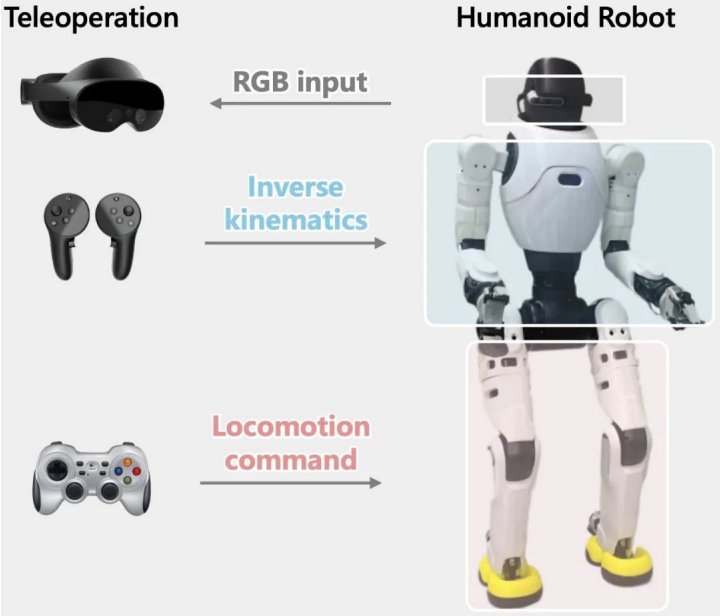


High-level planner (VLM + action decoder): **perception, planning** →
- Upper-body action
- Locomotion command

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

Challenge - Data

Real robot data is too costly to scale-up



TWIST2

To teleoperate the humanoid robot, we need at least 2 human teleoperators with a humanoid robot.

Challenge - Data (cont'd)

Scarcity of loco-manipulation data

Agibot World

Tabletop manipulation

Large-scale data

Room-to-Room

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Room-Across-Room

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Navigation on wheeled or quadruped platform

Challenge - Data (cont'd)

Scarcity of loco-manipulation data

Agibot World

Tabletop manipulation

Large-scale data

Room-to-Room

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Room-Across-Room

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Navigation on wheeled or quadruped platform

Manipulation and navigation are **Separate tasks.**

Challenge - Data (cont'd)

Scarcity of loco-manipulation data

Agibot World

Tabletop manipulation

Large-scale data

Room-to-Room

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Room-Across-Room

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Navigation on wheeled or quadruped platform

Manipulation and navigation are **Separate tasks.**

Learn from human/action-free videos!

How to obtain loco-manip. knowledge?

Learn from human/action-free videos

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

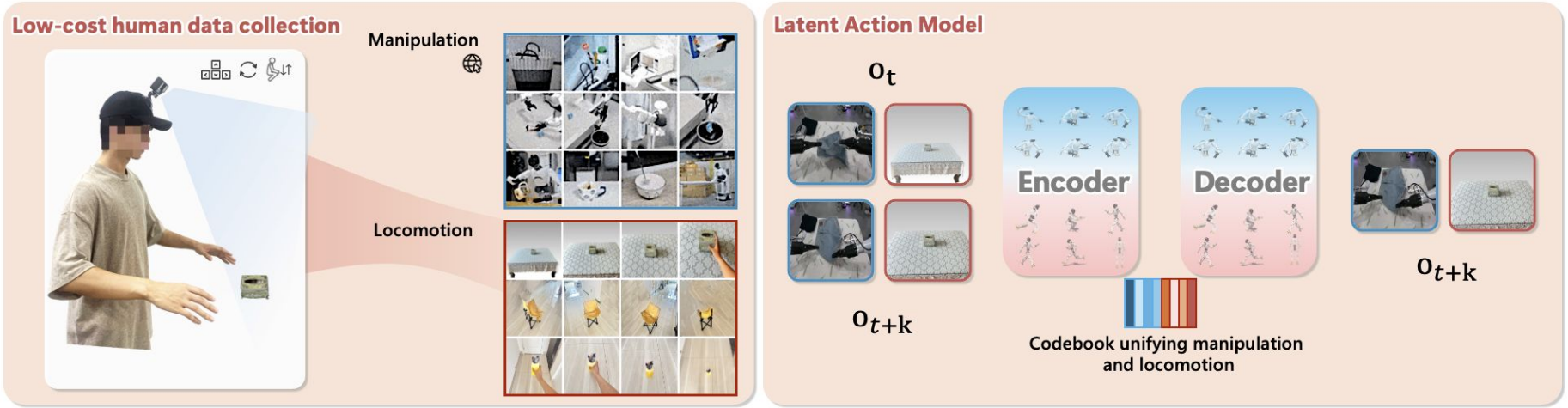
How to obtain loco-manip. knowledge? (cont'd)

Learn from human/action-free videos

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

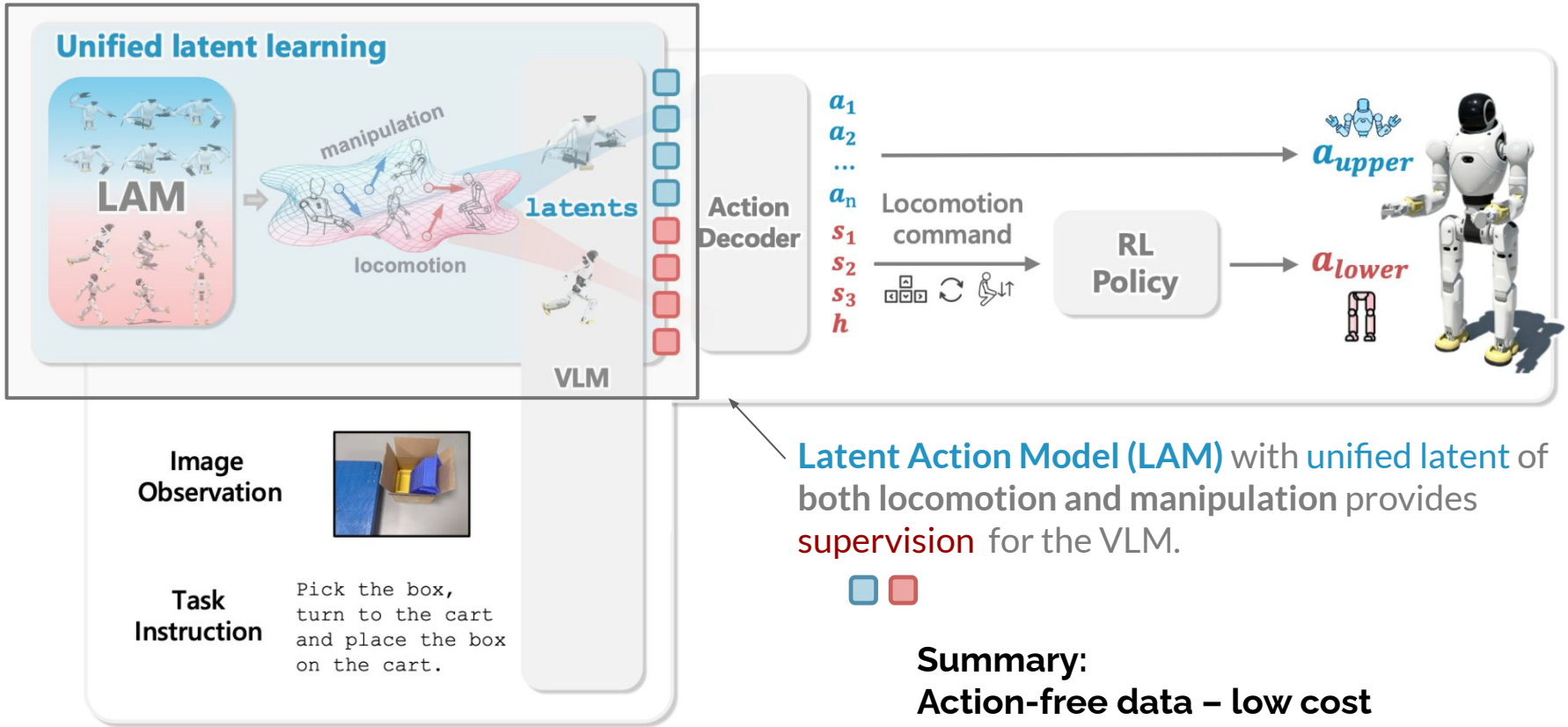
Manipulation: Large-scale tabletop manipulation dataset (*e.g.*, AgibotWorld)

Method | Unified Latent Learning



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

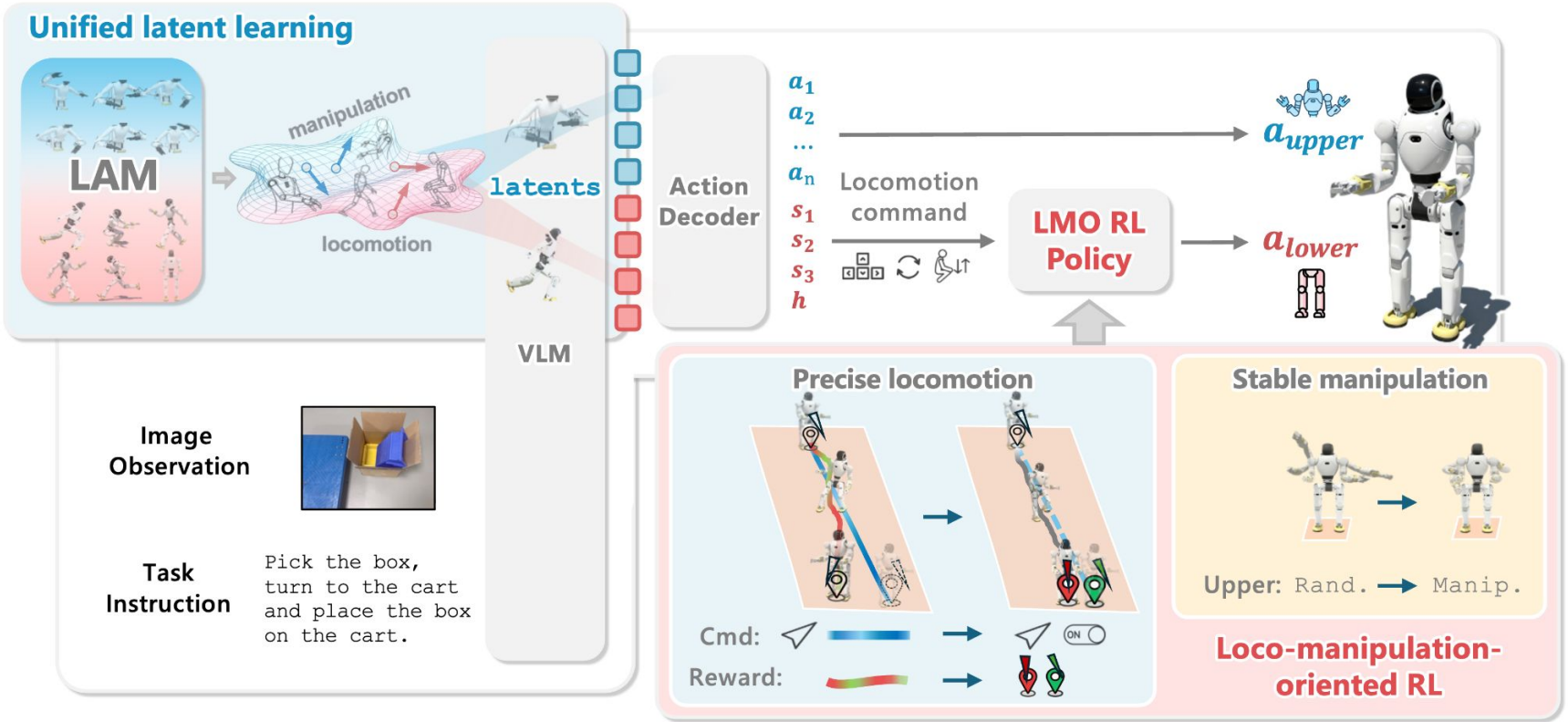
Method | Unified Latent Learning(cont'd)



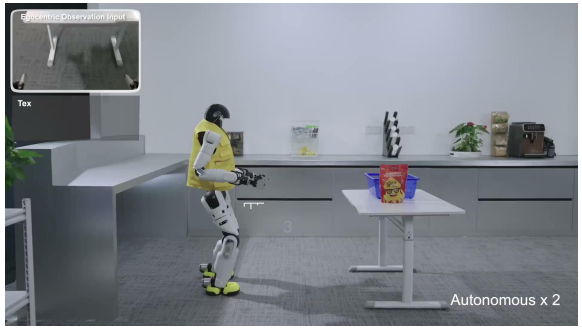
Latent Action Model (LAM) with unified latent of **both locomotion and manipulation** provides supervision for the VLM.

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

The Full Picture



Demo | WholeBodyVLA

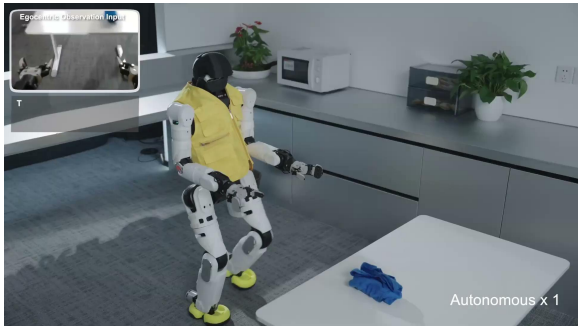


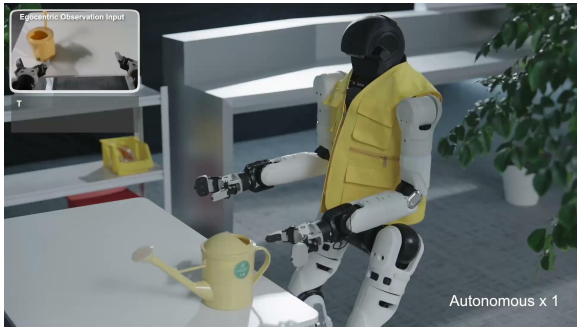
Start-pose Gen.

Object Gen.

Long-horizon

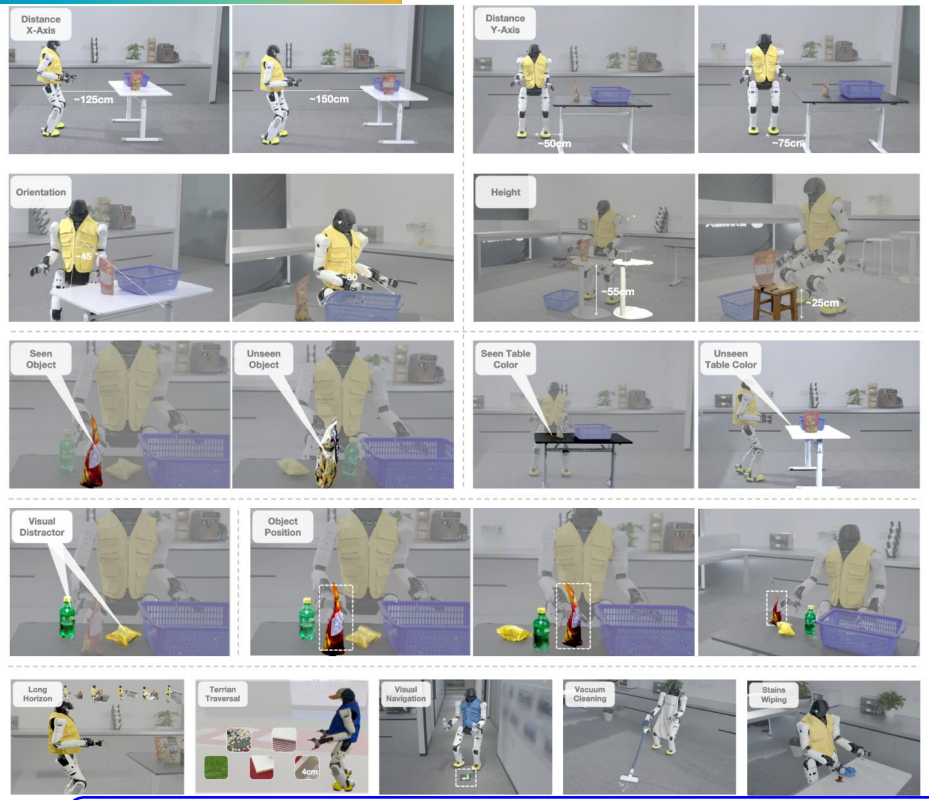
Go-around

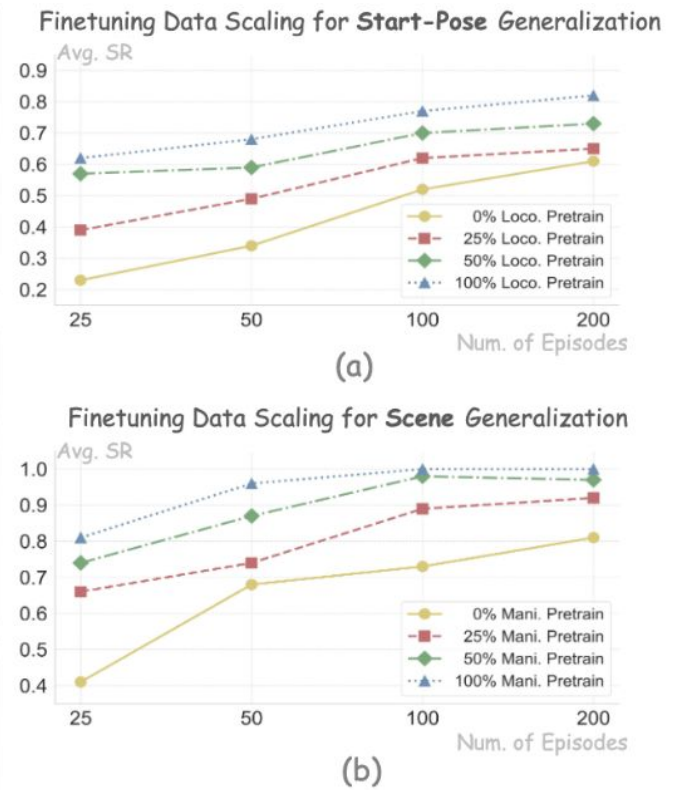




Extend to everyday task

Experiment | Generalization and Scaling





Unified latent learning reduces the reliance on real robot data.

Take-aways for WholebodyVLA

- By learning the **unified** latents and manipulation-aware locomotion policy, we *enable* loco-manipulation with a single VLA framework.
- By learning the unified latents of locomotion and manipulation, we *reduce* the reliance on large-scale loco-manipulation data.

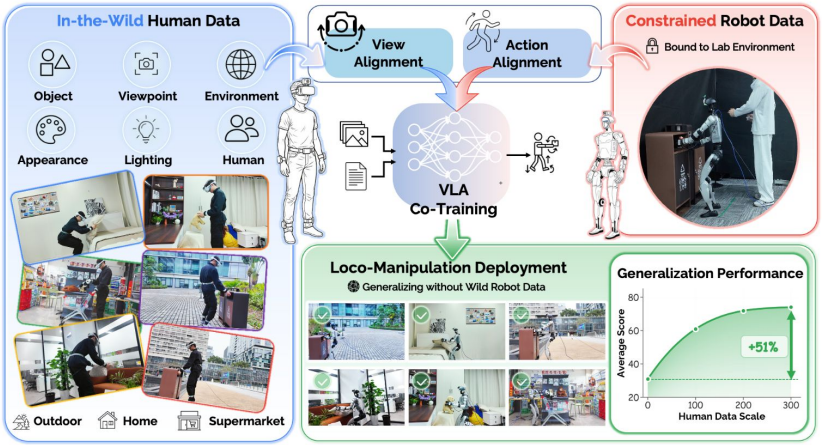
Outline

1. Background
   - Time for loco-manipulation, aka whole-body control
   - Why and why now?
   - VLA and Data effort
2. Algorithm part: WholebodyVLA (ICLR 2026)
3. **Data part: Human-centric data collection (RSS 2026)**
4. Challenges and Future Work

EgoHumanoid

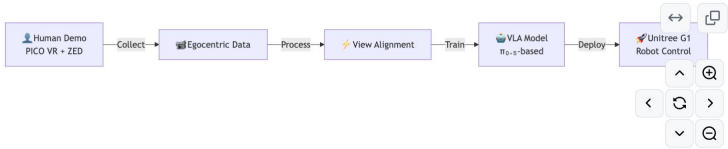
https://github.com/OpenDriveLab/EgoHumanoid

Framework Overview



System Architecture

📖 Overview



EgoHumanoid consists of **four main components**:

1 Data Collection	2 Data Processing	3 Model Training	4 Deployment
Collect synchronized multi-modal data from both humanoid robots (Unitree G1) and human demonstrators (PICO VR + ZED Mini)	Process, align, and retarget human demonstrations to robot action space	Fine-tune vision-language-action models $\pi_{0.5}$ on the processed datasets	Deploy trained policies on real humanoid robots with real-time inference

Motivation | Data Bottleneck

Robot Teleoperation

❌ High cost & operational complexity

❌ Confined to laboratory environments

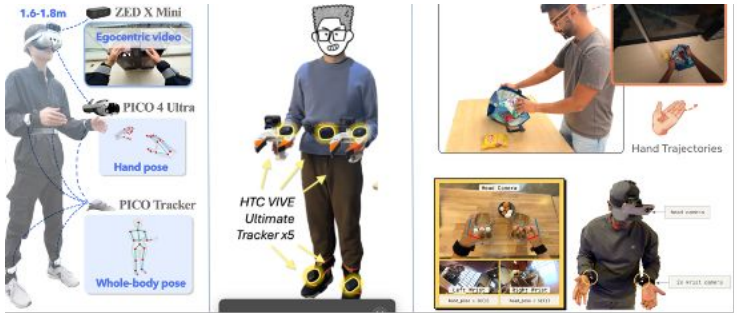
✅ No training-deployment gap

Human Data

✅ Low-cost portable sensors

✅ Scalable across diverse environments

❌ Human-robot embodiment gap



Motivation | EgoDataset

General Purpose

- Epic-Kitchens: egocentric cooking activities and actions (55hrs)
- **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.(166hrs)

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/pdf/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/pdf/2602.06949)

Motivation | EgoDataset

General Purpose

- Epic-Kitchens: egocentric cooking activities and actions (55hrs)
- **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

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)

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

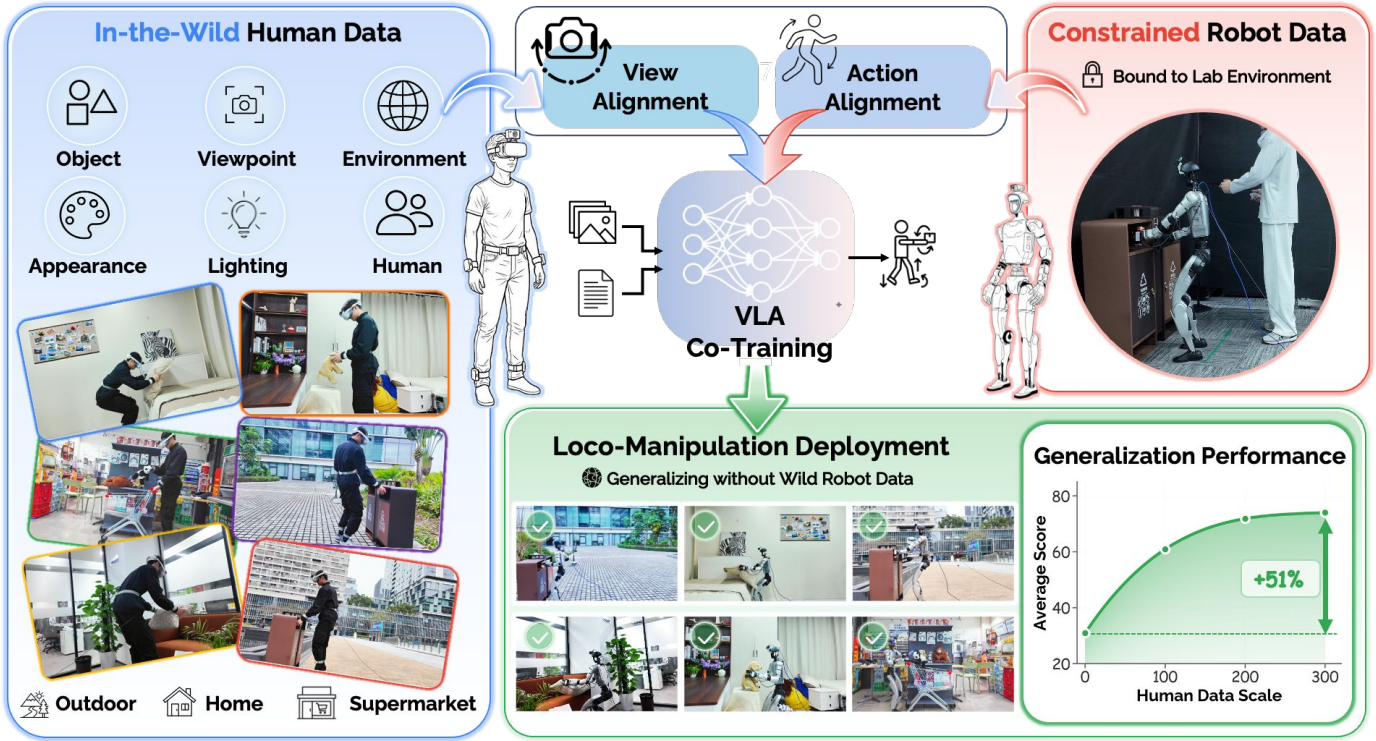
09)

				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

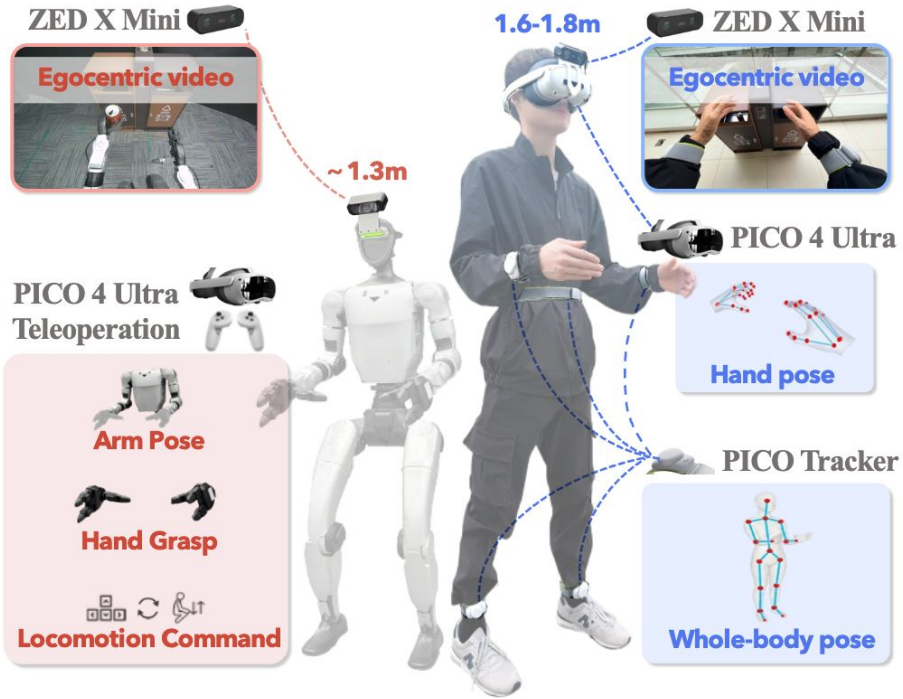
DreamDojo(https://arxiv.org/pdf/2602.06949)

Our Solution | Co-training

Human data with whole-body pose & Robot data



Data Collection



Robot Data Collection

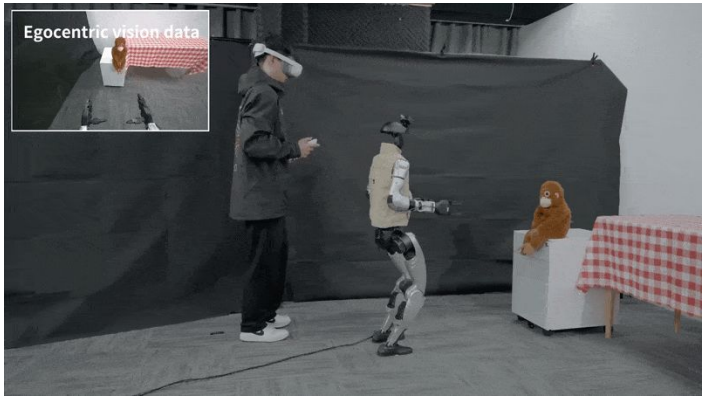
- ZED Camera
- UniTree G1
- Pico4 Ultra
- Pico Controller * 2

Human Data Collection

- ZED Camera
- Pico4 Ultra
- Pico Motion Tracker * 5

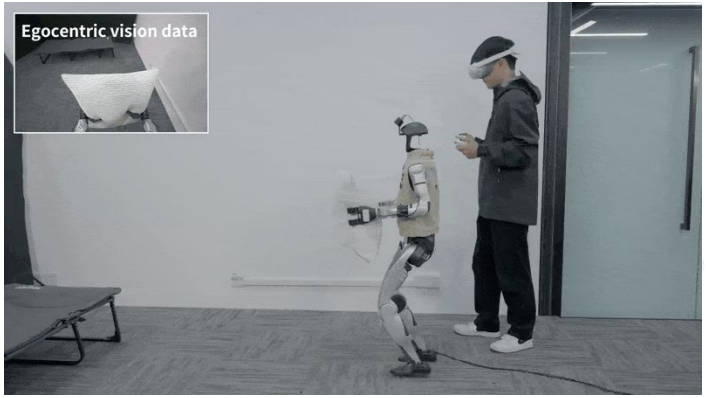
Data Collection | Laboratary Robot Data

Cart Stowing



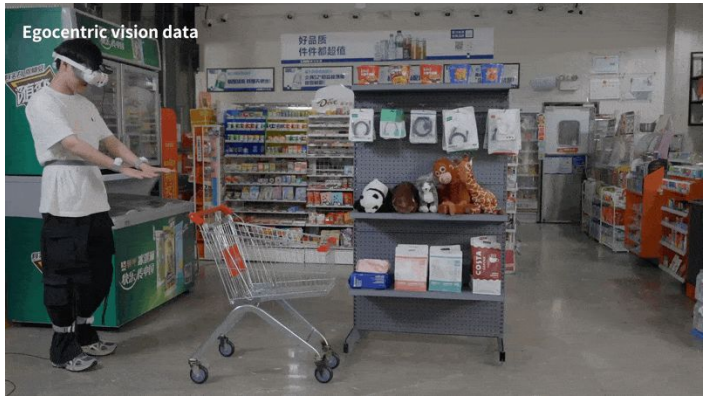
Toy Transfer

Trash Disposal



Pillow Placement

Data Collection | Diverse Human Data



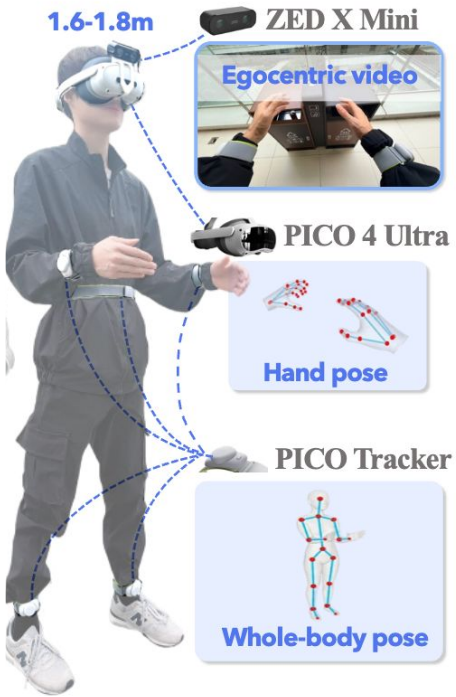
Cart Stowing

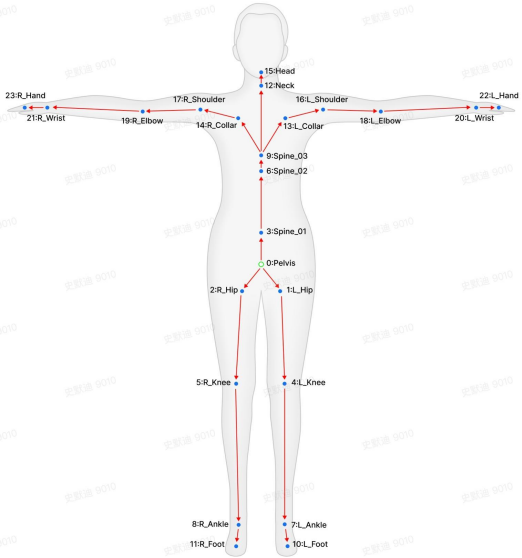
Toy Transfer

Trash Disposal

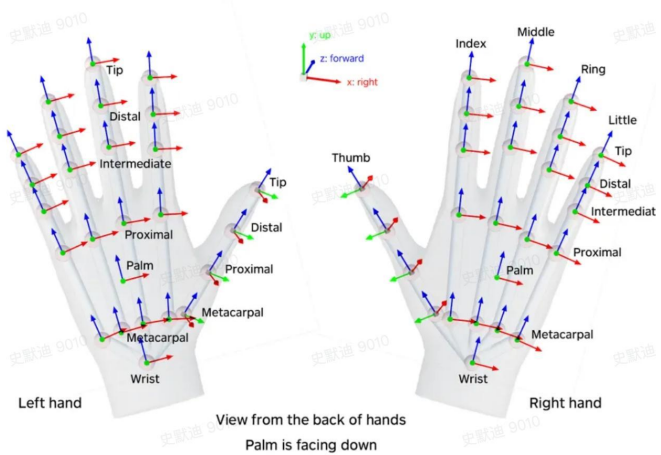
Pillow Placement

Data Collection



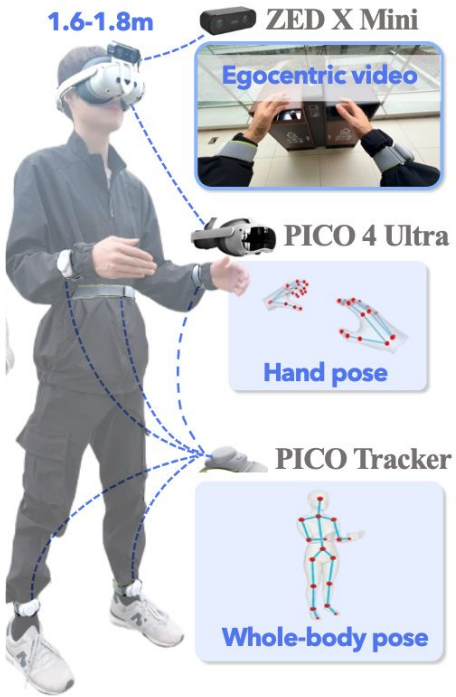


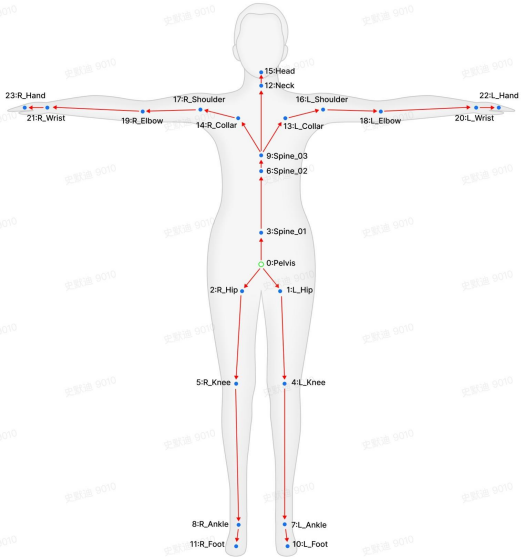
24 dof whole body pose



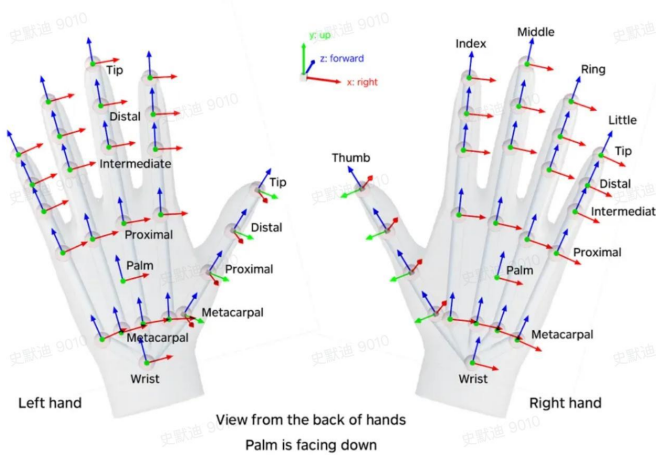
2*26 dof hand pose

Data Collection





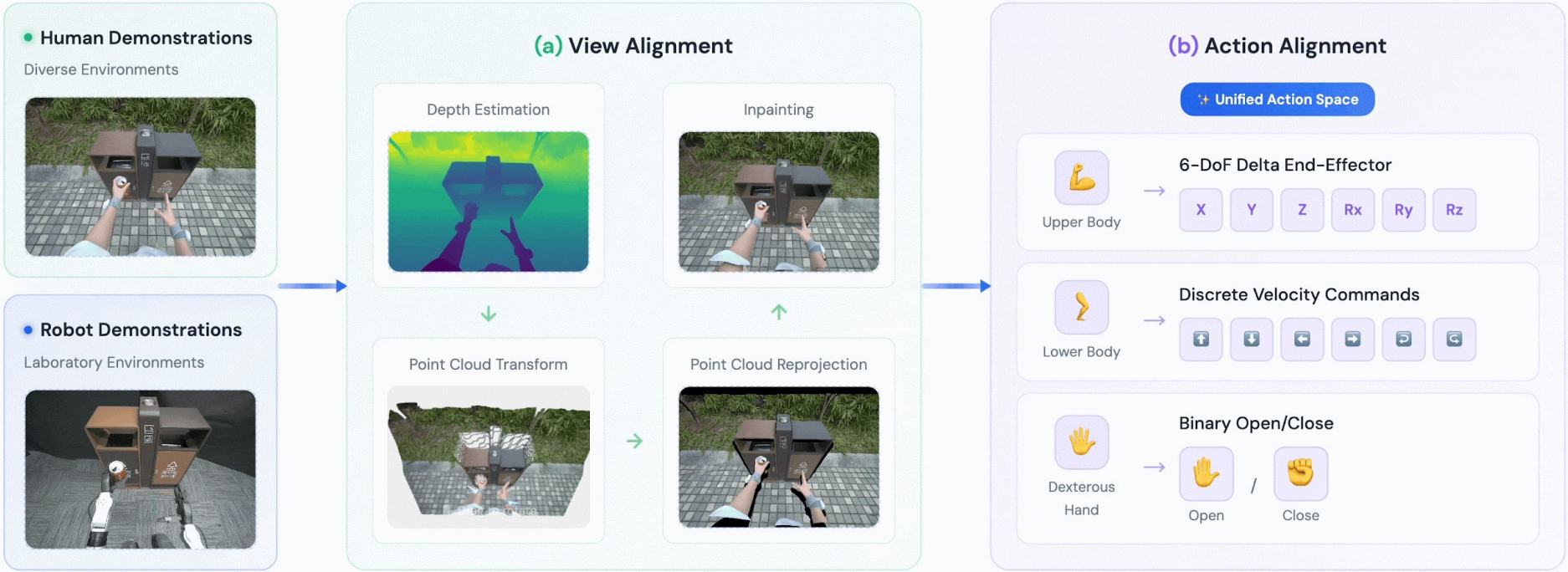
24 dof whole body pose



2*26 dof hand pose

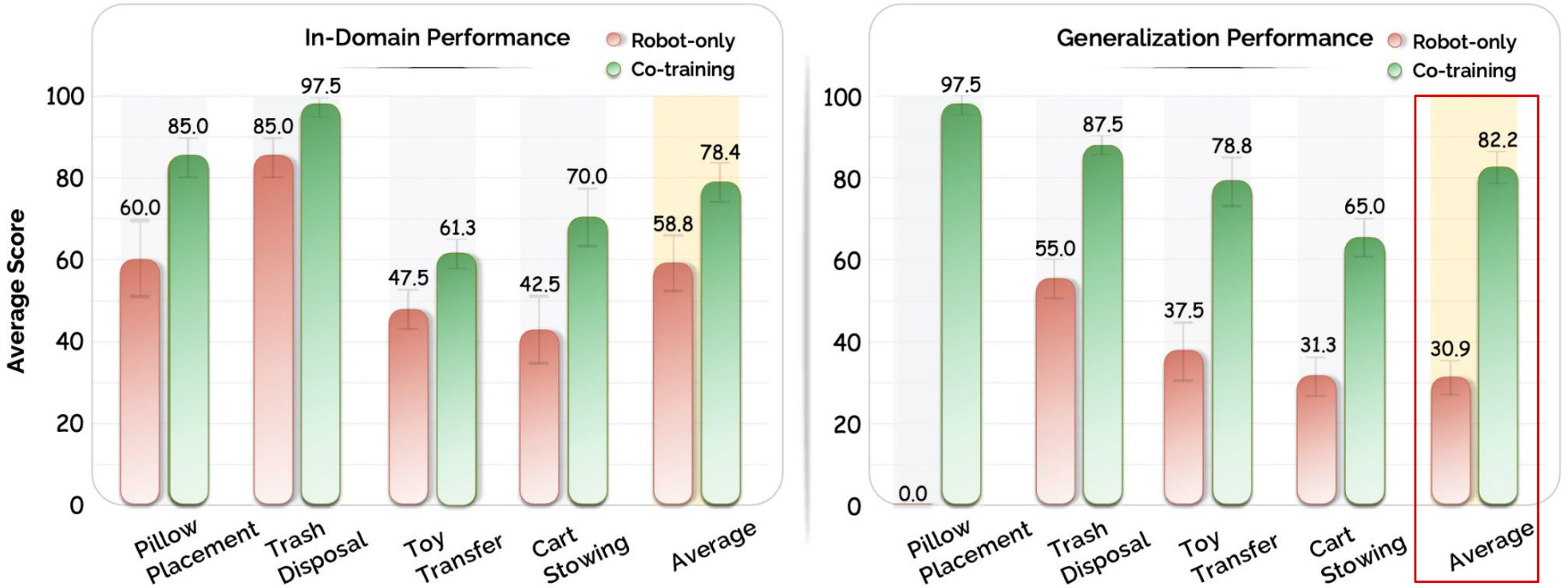
Human & Robot Domain Gap: Visual Viewpoint & Action representation

Pipeline Overview



Visual Gap **Action Gap**

Experiment

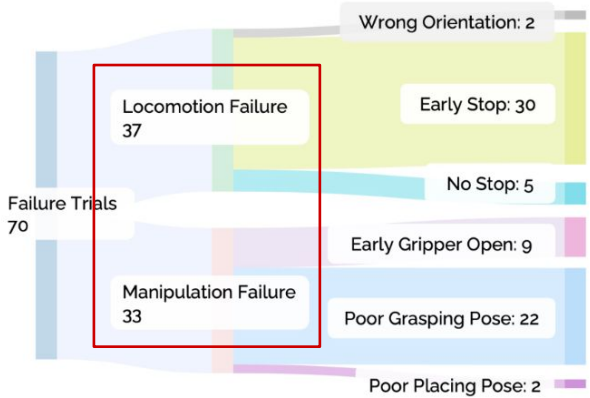


Co-training boosts robot performance, especially in unseen settings (+51.3%).

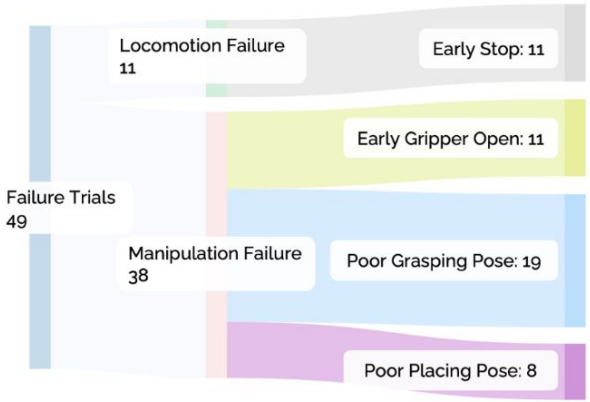
Experiment

● Navigation ● Manipulation

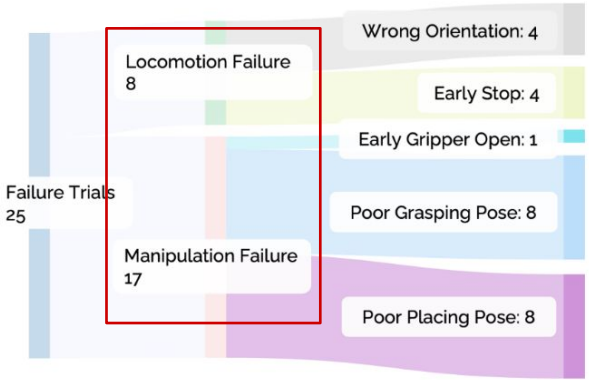
Training Data	Pillow Placement		Trash Disposal		Toy Transfer				Cart Stowing			
	s1	s2	s1	s2	s1	s2	s3	s4	s1	s2	s3	s4
Robot-only	0	0	65	45	100	50	0	0	100	15	5	5
Human-only	100	95	100	80	100	100	45	35	100	5	0	0
Co-training	**100**	**95**	**100**	**75**	**100**	**100**	**60**	**55**	**100**	**60**	**50**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



(a) Robot-only



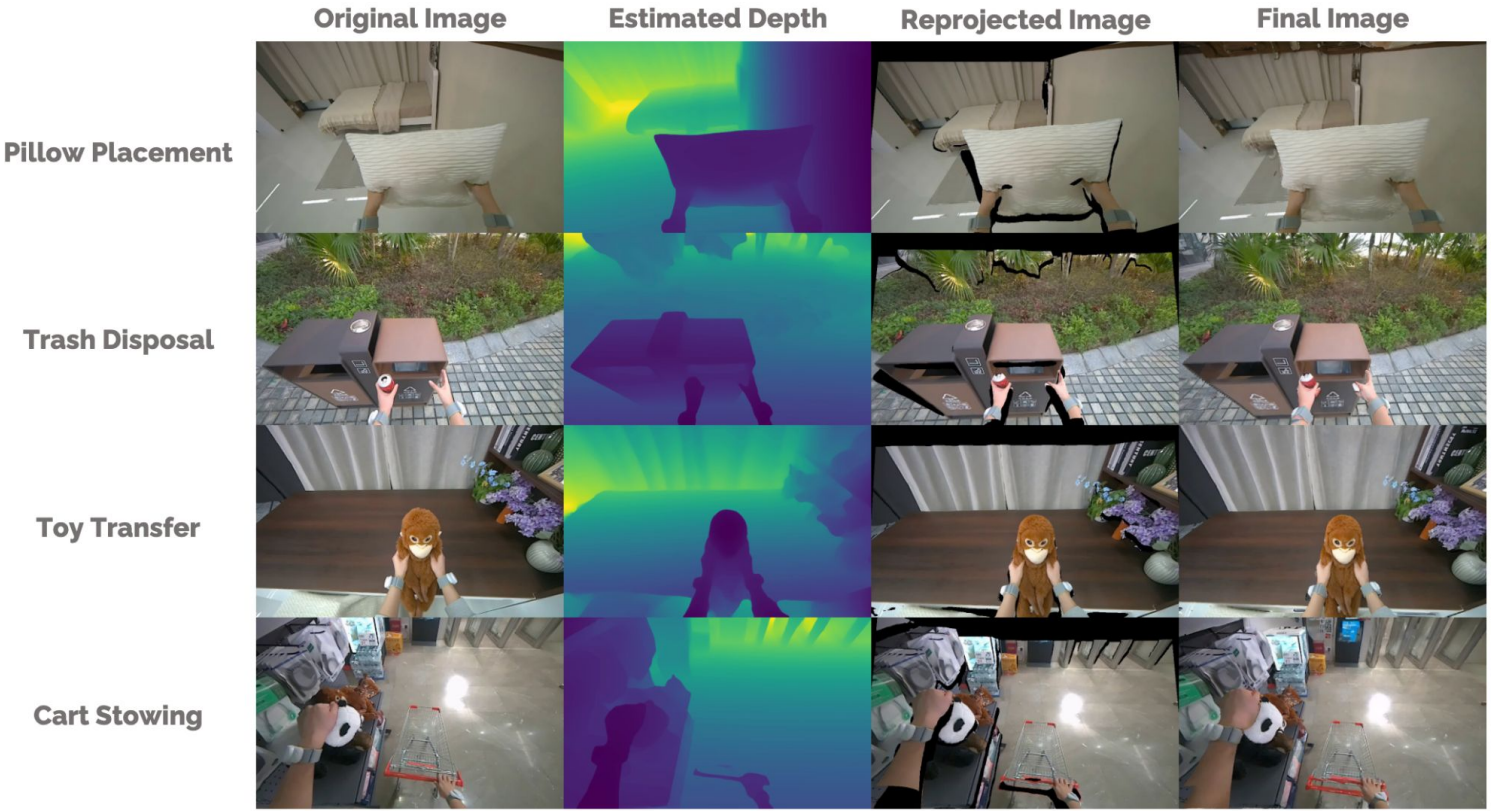
(b) Human-only



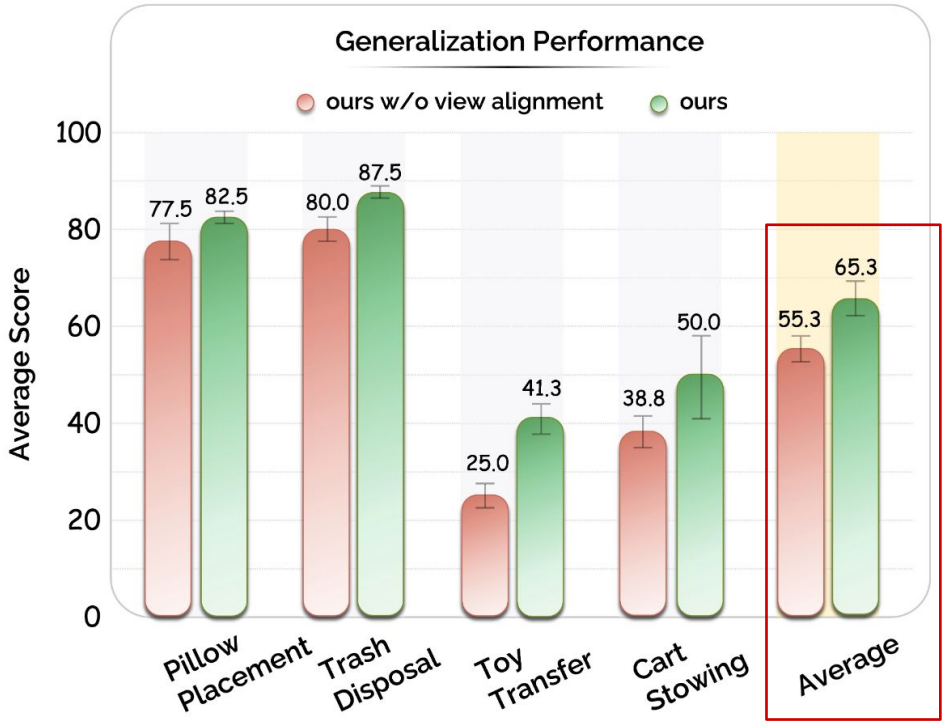
(c) Co-training

Navigation transfers better from human data

Experiment | View Alignment Visualization



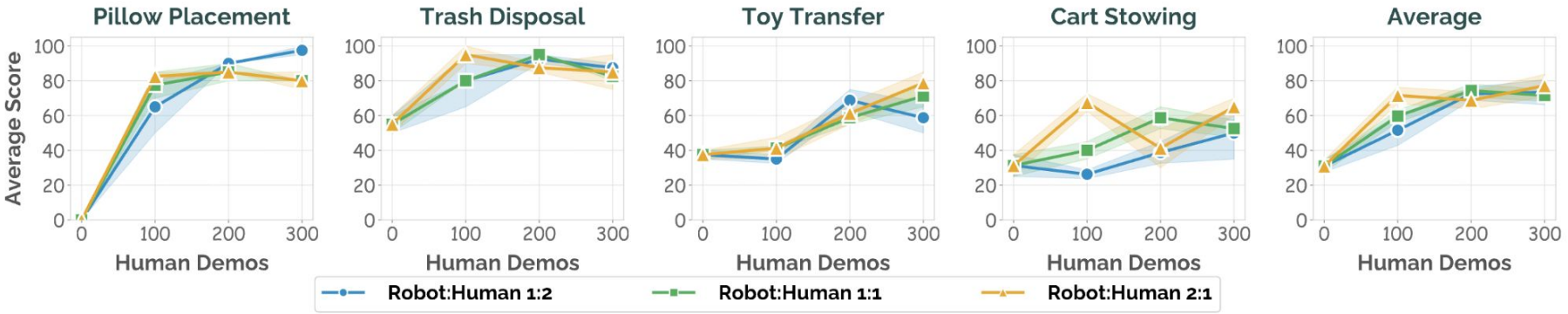
Experiment



View alignment improves performance by bridging camera height gaps

Zero-Shot Deployment

Experiment



Performance scales with human data;
optimal sampling ratio depends on task precision requirements.

Outline

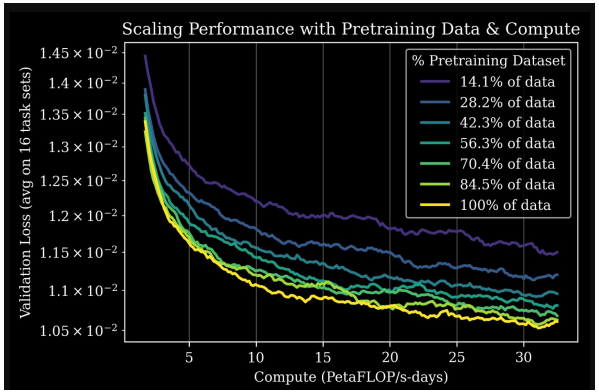
1. Background
   - Time for loco-manipulation, aka whole-body control
   - Why and why now?
   - VLA and Data effort
2. Algorithm part: WholebodyVLA (ICLR 2026)
3. Data part: Human-centric data collection (RSS 2026)
4. **Challenges and Future Work**

Roadmap & Future Work

Scale up egocentric human data

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**



GEN-0, Scaling Law, Generalist

Object, Environment Robustness -> Generalization, Emergent Capability

Roadmap & Future Work

Towards generalizable whole body intelligence:

- Scalable human-centric data collection system
- Human-level embodiment
- Scale up pre-training
- Align action space to humanoid embodiment


香 港 大 學
THE UNIVERSITY OF HONG KONG
END
Q&A
hongyang@hku.hk
tianyu@opendrivelab.com