HUGE: SCALING HUMANOID GENERALIST VISION-LANGUAGE-ACTION MODELS WITH EGOCENTRIC EXPERIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending Vision-Language-Action models (VLAs) to humanoid robots requires coordinated whole-body control and diverse demonstrations that are costly to collect in real-world scenarios. Although egocentric human demonstrations provide a complementary source of task experience, differences in visual embodiment and missing robot proprioception complicate transfer to humanoid policies. To address these challenges, we introduce HUGE, a family of **Hu**manoid **Ge**neralist VLAs pre-trained on over 2,000 hours of egocentric human and robot data. Specifically, HUGE employs three existing representative VLA architectures to predict latent motion tokens for a pretrained whole-body controller, together with hand-control signals. Based on this, we introduce a three-stage training pipeline that combines egocentric pre-training, humanoid adaptation, and downstream fine-tuning. To facilitate transfer across embodiments, we construct pseudo-proprioceptive states through simulation playback during pre-training and modify arm appearance in egocentric observations through re-rendering during downstream fine-tuning. With this pipeline, HUGE achieves the state-of-the-art on simulation benchmark HumanoidArena. Furthermore, real-world experiments on the Unitree G1 indicate that augmenting robot demonstrations with visually adapted egocentric demonstrations significantly increases success on a pick-and-drop task.

1 INTRODUCTION

Recent advances in Vision-Language-Action models (VLAs) (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; Bjorck et al., 2025; Ye et al., 2026; Li et al., 2026a) fueled by Vision-Language Models (VLMs) (Beyer et al., 2024; Qwen Team, 2025) and World Models (WMs) (Wang et al., 2025; Assran et al., 2025) have enabled language-conditioned manipulation across a wide range of tabletop tasks. Building on this progress, extending VLAs to humanoid robots offers the potential to perform diverse tasks in human environments by combining manipulation with whole-body mobility. Despite this potential, adapting VLAs to humanoid robots remains challenging because task execution requires coordinated whole-body motion while maintaining balance and stability. To meet these control requirements, existing humanoid loco-manipulation approaches commonly employ hierarchical architectures that combine high-level policies with reinforcement learning (RL)-based controllers (Jiang et al., 2026; Wei et al., 2026; Bai et al., 2026). To this end, a central research question is how to train generalist VLAs that translate language instructions into coordinated whole-body behaviors across diverse tasks.

Realizing this goal requires addressing two key challenges. First, the high-dimensional action spaces and complex dynamics of humanoid robots make it difficult to directly transfer VLA policies developed for robotic manipulators to whole-body control. In particular, arm movements must be coordinated with the torso and lower body to accomplish manipulation objectives while maintaining balance and stability. Consequently, VLA-based humanoid control requires an action interface that connects task-level predictions to coordinated whole-body execution. Second, beyond these control challenges, learning generalist humanoid behaviors requires diverse whole-body demonstrations, yet collecting such data on real-world robots is costly and time-consuming. To alleviate this data bottleneck, egocentric human demonstrations offer a complementary source of task experience for large-scale training. However, differences in visual embodiment and the absence of robot

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

proprioceptive observations complicate the direct use of these demonstrations for humanoid policy learning. These challenges motivate an approach that integrates VLA policies with whole-body control and supports effective transfer from egocentric human experience to humanoid robots through a scalable training pipeline.

To address these challenges, we introduce HUGE, a family of **Hu**manoid **Ge**neralist VLAs that operate through a whole-body controller and learn from a large-scale mixture of egocentric human and robot data. Specifically, HUGE builds on existing representative VLA architectures and follows a three-stage training pipeline that sequentially incorporates egocentric human data, humanoid teleoperation data, and a mixture of downstream demonstrations. To support this pipeline, we assemble a training corpus comprising over 2,000 hours of egocentric and robot data. Furthermore, we prepare egocentric data for humanoid policy learning during both pre-training and downstream fine-tuning. In particular, during pre-training, we use simulation playback to construct pseudo-proprioceptive states for egocentric data that lack robot proprioceptive observations. Building on this pre-training stage, we re-render and replace arm regions in egocentric observations during downstream fine-tuning. To assess the effectiveness of this approach, we compare HUGE with baseline methods on the HumanoidArena simulation benchmark and further evaluate its performance on real-world humanoid tasks. Overall, HUGE achieves the state-of-the-art performance on HumanoidArena, demonstrating its capability for humanoid whole-body control, while comprehensive real-world experiments and solid analyses further demonstrate the effectiveness of the training recipe.

Our main contributions are summarized as follows:
• We introduce HUGE, a family of humanoid generalist VLAs, trained on over 2,000 hours of egocentric human and robot data through a three-stage training pipeline.
• We develop data adaptation methods that facilitate the transfer of egocentric human experience to humanoid policy learning.
• HUGE achieves the state-of-the-art performance against baselines on HumanoidArena, while comprehensive real-world experiments and analyses validate our design choices and training paradigm.

2 RELATED WORK

2.1 VISION-LANGUAGE-ACTION MODELS

Vision-Language-Action models (VLAs) (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024; Bjorck et al., 2025) built on Vision-Language Models (VLMs) (Beyer et al., 2024; Qwen Team, 2025) and World-Action Models (WAMs) (Ye et al., 2026; Li et al., 2026a; Yuan et al., 2026) based on video-prediction World Models (WMs) (Wang et al., 2025; Assran et al., 2025) have advanced rapidly in fixed-base single-arm and bimanual manipulation. Beyond language-conditioned task execution, these approaches combine robot demonstrations with pretrained multimodal knowledge or predictions of future observations to support policy learning across diverse manipulation tasks (Tian et al., 2025; Wang et al., 2026b; Liu et al., 2026). These advances have enabled robotic manipulators to autonomously fold and organize deformable objects such as clothes and towels, perform precise operations such as connector insertion and small-part assembly, and complete long-horizon tasks such as table setting and coffee preparation (Physical Intelligence et al., 2025a; Lin et al., 2026; Physical Intelligence et al., 2026; Song et al., 2026). For humanoid whole-body control, we adopt three representative VLA architectures, OFT (Kim et al., 2025), GR00T (Bjorck et al., 2025), and π (Black et al., 2024), without introducing additional modules or modifying the architectural designs. Using these architectures, we investigate humanoid whole-body control and training strategies centered on egocentric human demonstrations, complemented by robot data for embodiment adaptation and downstream task learning.

2.2 HUMANOID WHOLE-BODY CONTROL

Early reinforcement-learning approaches to humanoid control focused on locomotion, training policies in simulation to maintain balance, follow velocity commands, and adapt to terrain variations during real-world deployment (Radosavovic et al., 2024). Subsequent work expanded these capabilities to whole-body motion tracking, using retargeted human motions to train controllers that support coordinated teleoperation and autonomous skill learning from demonstrations (He et al., 2024; Ding et al., 2025; Luo et al., 2026b; Qi et al., 2026). Building on these control capabilities, one

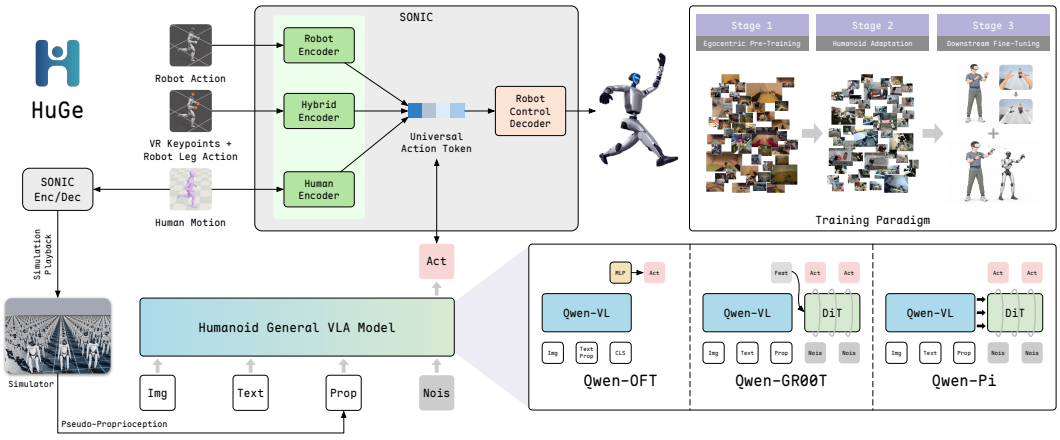


Figure 1: Schematic of the HUGE framework, architecture and training pipeline.

line of hierarchical VLA approaches predicts upper-body manipulation targets alongside lower-body motion commands, relying on reinforcement-learning controllers to execute locomotion and posture adjustments (Liu et al., 2025; Jiang et al., 2026; Wei et al., 2026; Bai et al., 2026). Alongside this direction, advances in whole-body motion models have supported VLA and world-action approaches that generate full-body motion references or unified motion latents, which whole-body controllers translate into coordinated joint commands (Zheng et al., 2026a; Li et al., 2026b). In contrast, we focus on training strategies centered on egocentric human demonstrations for humanoid whole-body control, examining three established VLA architectures without adding modules or modifying the architectural designs.

2.3 EGOCENTRIC LEARNING

Egocentric human demonstrations provide diverse observations and a complementary source of experience for robot learning. Early approaches use human videos to learn visual representations and motion priors that support downstream policy learning (Nair et al., 2022; Radosavovic et al., 2022; Ma et al., 2023; Wang et al., 2023a). Building on these efforts, more direct approaches learn action prediction from annotated human motions or inferred latent actions, incorporating human and robot data through pre-training, co-training, and task-specific adaptation (Luo et al., 2025; Li et al., 2025; Luo et al., 2026a; Zheng et al., 2026b). To address embodiment differences, another line of work converts human demonstrations into robot-compatible data through motion retargeting and visual adaptation, including human-arm removal, robot rendering, and embodiment masking (Lepert et al., 2025; Kareer et al., 2025; Zhang et al., 2026). Beyond fixed-base manipulation, related efforts extend human-data learning to humanoid manipulation and loco-manipulation through aligned action representations and staged training with robot demonstrations (Wei et al., 2026). Within this setting, we investigate how egocentric human demonstrations can support humanoid whole-body policy learning through controller-compatible action supervision, simulation-derived pseudo-proprioceptive states, and visual adaptation across training stages.

3 METHODOLOGY

In this section, we present the model architecture and training recipe of HUGE, as shown in Figure 1. We first describe how three representative VLA architectures are instantiated for humanoid control. Building on these model instantiations, we introduce a three-stage training pipeline that progressively incorporates egocentric human experience, humanoid demonstrations, and task-specific mixture data.

3.1 ARCHITECTURE

Given a visual observation o_t, a proprioceptive state s_t, and a natural-language task instruction l, HUGE predicts a 66-dimensional action vector a_t for humanoid control at time step t. The action

consists of a 64-dimensional latent motion token and two binary variables specifying whether the left and right hands are open or closed. The motion token lies in the universal token space of SONIC (Luo et al., 2026b), a pretrained whole-body controller. Conditioned on the proprioceptive state of robot, control decoder of SONIC converts the predicted motion token into joint-level commands for whole-body motion, while the two binary variables separately control hand opening and closing. This interface allows HUGE to predict task-conditioned whole-body actions while relying on the low-level controller for coordinated motion execution.

We instantiate HUGE with three existing representative VLA designs: OFT (Kim et al., 2025), GR00T (Bjorck et al., 2025), and π (Black et al., 2024). Across these variants, Qwen-based vision-language backbones encode visual observations, proprioceptive state and language instructions, while different action prediction modules map the resulting representations to humanoid actions (StarVLA Community, 2026). Specifically, the OFT-style variant uses a lightweight multilayer perceptron (MLP) to map the hidden states of predefined action tokens directly to continuous actions, predicting an action chunk in parallel. In contrast, the GR00T-style variant separates vision-language processing from action generation: the backbone provides contextual features, and a diffusion transformer (DiT) conditions on these features through cross-attention to generate action chunks via flow matching. Compared with this use of backbone output features, the π-style variant connects a DiT action expert to hidden states from multiple backbone layers through layer-wise cross-attention, progressively refining noisy action sequences into continuous action chunks. For humanoid control, all three variants use the action interface defined above, with each predicted time step containing a 64-dimensional latent motion token for SONIC and two hand-control signals.

3.2 TRAINING RECIPE

We train HUGE through a three-stage pipeline that progressively incorporates egocentric human experience, humanoid demonstrations, and task-specific mixture data. Specifically, egocentric pre-training utilizes large-scale egocentric demonstrations to learn language-conditioned action prediction. Subsequently, during humanoid adaptation, we continue training the policy on humanoid teleoperation demonstrations to adapt action prediction to robot observations and states. Finally, downstream fine-tuning adapts the policy to target tasks using task-specific demonstrations. Table 1 summarizes the data sources and dataset sizes for each training stage.

Table 1: Training data used across the three-stage training pipeline.

Training Stage	Source	Hours	Episodes	Frames
Stage 1	EgoDex (Hoque et al., 2026)	821.800	334,991	147,923,137
	EgoVerse (Punamiya et al., 2026)	47.080	1,704	5,127,761
	Xperience-10M (Ropedia, 2026)	478.330	10,651	34,589,556
	BEHAVIOR-1K (Li et al., 2022)	322.400	3,426	34,818,979
	TOTAL	*1,669.610*	*350,772*	*222,459,433*
Stage 2	Humanoid Everyday (Zhao et al., 2026)	16.470	4,064	1,779,287
	UniFoLM-WBT (Unitree Robotics, 2026)	22.960	3,656	2,479,519
	HIW-500 (BitRobot and Unitree and Hugging Face, 2026)	378.150	23,743	40,839,947
	TOTAL	*417.580*	*31,463*	*45,098,753*
Stage 3	HumanoidArena (Wang et al., 2026a)	3.030	700	545,454
	Pick-and-Drop-Cola Task (Ego)	2.456	400	442,074
	Pick-and-Drop-Cola Task (Tele-Op)	0.826	110	148,701

3.2.1 STAGE 1: EGOCENTRIC PRE-TRAINING

Egocentric pre-training aims to initialize HUGE for language-conditioned humanoid action prediction using large-scale egocentric experience. To this end, we combine human demonstrations from EgoDex (Hoque et al., 2026), EgoVerse (Punamiya et al., 2026), and Xperience-10M (Ropedia, 2026) with simulated robot demonstrations from BEHAVIOR-1K (Li et al., 2022), expressing heterogeneous motion supervision in the SONIC action space. Although the BEHAVIOR-1K demonstrations originate from robot simulation, we include these demonstrations into the first training stage, broadly

defined here as demonstrations collected without teleoperating real-world robots. Specifically, we use SONIC motion tokens for EgoDex and construct corresponding motion supervision for EgoVerse from the Aria bimanual subset. After checking tracking validity, calibration stability, segment duration, and G1 retargeting constraints, we retain 47.08 hours of EgoVerse segments. We then encode the retained head and wrist trajectories into SONIC motion tokens with a fixed lower-body standing reference. In comparison, Xperience-10M provides full-body motion references, which we convert into the reference format required by the SONIC encoder. For BEHAVIOR-1K, a SONIC planner generates G1 motion from upper-body pose, base-motion, and height references. Alongside these motion tokens, we derive binary hand-control labels from hand keypoints for EgoVerse and Xperience-10M and from gripper commands for BEHAVIOR-1K, while excluding unavailable EgoDex hand-action labels from supervision. To provide robot-state inputs, we replay the motion tokens on a simulated G1 using the SONIC controller in simulator; for BEHAVIOR-1K, we collect paired motion tokens and states during planner-driven execution. From these simulation records, we construct a 34-dimensional pseudo-proprioceptive state comprising 29 joint angles, two hand-state entries, and the three-dimensional gravity direction expressed in the body frame. Within this representation, the hand-state entries use action-label proxies when available and retain simulated hand states for EgoDex. Finally, we train each VLA variant to predict 50-step action chunks on a 50 Hz control timeline from the corresponding image, task instruction, and normalized pseudo-proprioceptive state, masking padded time steps and unavailable hand-action labels.

3.2.2 Stage 2: Humanoid Adaptation

Humanoid adaptation aims to transfer the policy learned from human and simulated experience to real humanoid operation. Although egocentric pre-training provides diverse motion supervision, the visual observations and pseudo-proprioceptive states can differ from those encountered by a real-world robot. To address this gap, we continue training each pretrained VLA variant on real G1 demonstrations, organizing heterogeneous data sources into a shared action and state interface. Specifically, for Humanoid Everyday (Zhao et al., 2026), we select the G1 subset and use SONIC actions and corresponding proprioceptive states. For UniFoLM-WBT (Unitree Robotics, 2026), we combine demonstrations collected with different hand types using SONIC motion tokens and standardized binary hand-control labels. For HIW-500 (BitRobot and Unitree and Hugging Face, 2026), we augment motion tokens with hand-control labels reconstructed from controller events, retaining each hand state until the next opening or closing event. Alongside these labels, we construct proprioceptive inputs from recorded body joint positions, observation-side hand events, and gravity directions derived from the recorded controller pivot orientation. Using this common representation, we jointly adapt the vision-language backbone and action prediction modules to real robot observations and motion supervision before downstream task fine-tuning.

3.2.3 Stage 3: Downstream Fine-Tuning

Downstream fine-tuning adapts the policy to target tasks and environments. Specifically, we continue training the policy separately for HumanoidArena (Wang et al., 2026a) and the real-world cola pick-and-drop task. For real-world adaptation, collecting task-specific robot demonstrations through teleoperation is costly and time-consuming. In particular, joint overheating during data collection can trigger robot shutdowns, interrupting demonstrations and raising safety concerns. To reduce reliance on robot teleoperation, we supplement robot demonstrations with egocentric human demonstrations collected using a PICO 4 Ultra and trackers attached to the wrists, waist, and ankles. To incorporate these demonstrations into the shared policy interface, we pair the recordings with SONIC action supervision and obtain pseudo-proprioceptive states through simulated G1 replay. However, the visible human hands and arms introduce an appearance mismatch with robot observations. To address this mismatch, we employ the visual editing pipeline of Phantom (Lepert et al., 2025) to segment and remove human hands and arms, inpaint the exposed regions, and composite rendered robotic hands into the scene. Within this pipeline, temporally smoothed hand keypoints guide robotic hand rendering, while task-specific grasp constraints and a thumb-visibility prior govern hand configuration and object occlusion in the cola grasping setting. Throughout this visual transformation, the synthesized videos retain the action targets, pseudo-proprioceptive states, and task instructions associated with the original recordings. Finally, we combine the synthesized human demonstrations with real robot demonstrations for supervised task adaptation, enabling human task experience to supplement robot data under a visual representation that more closely resembles robot observations.

4 EXPERIMENTATION

We first conduct simulation experiments to iteratively refine the training strategy and egocentric pre-training recipe for humanoid VLAs. Specifically, experiments on HumanoidArena examine training strategies, the effectiveness of pseudo-proprioceptive states during pre-training, and the scale of egocentric pre-training data. Building on these investigations, we then explore downstream fine-tuning and task adaptation on the real-world Unitree G1 robot, focusing on the composition of robot and egocentric demonstrations and the visual adaptation of egocentric observations. Additional implementation details are provided in Appendix A.

4.1 SIMULATION EVALUATION

We evaluate HUGE on seven HumanoidArena tasks (Wang et al., 2026a): three Human-Object Interaction (HOI) tasks and four Human-Scene Interaction (HSI) tasks. Using monocular egocentric RGB observations, language instructions, and robot proprioceptive states, we train a multitask policy for each variant in one run with 100 demonstrations per task. We then evaluate 60 episodes per task (three seed groups of 20), sharing task configurations across variants. Finally, we report task success rates, and unweighted task averages for HOI, HSI, and all seven tasks.

4.1.1 TRAINING STRATEGY

We first investigate which training strategy is most effective for learning humanoid VLAs. To this end, we compare frozen-backbone training (Frozen), low-rank adaptation (LoRA) (Hu et al., 2022), and full-parameter fine-tuning (Full) across the three HUGE variants on HumanoidArena. As shown in Table 2, LoRA achieves a higher performance than frozen-backbone training for all three variants, supporting the value of backbone adaptation in this setting. However, full-parameter fine-tuning yields further improvements, achieving average success rates of 55.00%, 58.33%, and 60.24% for HUGE-OFT, HUGE-GR00T, and HUGE-Pi, respectively, and exceeding LoRA by 19.29–38.81 percentage points. The consistent gains over LoRA suggest that the restricted update space may limit the adaptation of pretrained vision-language representations to humanoid action prediction. Accordingly, we adopt full-parameter fine-tuning for all subsequent experiments.

Table 2: Comparison of training strategies for HUGE trained without pre-training on HumanoidArena.

Model	Strategy	HOI				HSI					Overall AVG (↑)
		Football	DoubleDesk	P&PBox	AVG (↑)	OpenDoor	SitSofa	Boxing	VisNavi	AVG (↑)	
HUGE-OFT	Frozen	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	LoRA	0.00	0.00	65.00	21.67	0.00	8.33	38.33	1.67	12.08	16.19
	Full	36.67	0.00	76.67	37.78	88.33	53.33	80.00	50.00	67.92	**55.00**
HUGE-GR00T	Frozen	6.67	0.00	53.33	20.00	13.33	16.67	60.00	18.33	27.08	24.05
	LoRA	16.67	0.00	61.67	26.11	33.33	25.00	70.00	31.67	40.00	34.05
	Full	40.00	8.33	85.00	44.44	78.33	70.00	76.67	50.00	68.75	**58.33**
HUGE-Pi	Frozen	6.67	1.67	68.33	25.56	31.67	18.33	65.00	20.00	33.75	30.24
	LoRA	16.67	3.33	70.00	30.00	40.00	43.33	76.67	36.67	49.17	40.95
	Full	50.00	8.33	81.67	46.67	81.67	70.00	78.33	51.67	70.42	**60.24**

4.1.2 PSEUDO PROPRIOCEPTIVE STATES

Prior studies of robotic manipulation show that VLA policies can rely strongly on the current proprioceptive state when predicting actions (Lu et al., 2026). However, egocentric human demonstrations lack the corresponding robot-state information. To assess whether pseudo-proprioceptive states can compensate for this missing input during large-scale egocentric pre-training, we compare the three HUGE variants with and without pseudo-proprioceptive states during the egocentric pre-training stage. As shown in Table 3, incorporating these states improves average success rate by 5.00 and 2.86 percentage points for HUGE-GR00T and HUGE-Pi, respectively, suggesting that simulated body configurations provide useful context for learning humanoid action mappings from human demonstrations. More specifically, SitSofa success increases from 61.67% to 78.33% and from 68.33% to 83.33%, respectively, consistent with the potential value of body-state information for tasks involving substantial posture changes. In contrast, HUGE-OFT decreases from 56.43% to

54.05%. This degradation is associated with the textual state interface. Specifically, proprioceptive states are serialized as text, and input truncation leaves incomplete state descriptions that introduce ambiguity into action prediction.

Table 3: Effect of pseudo-proprioceptive states during egocentric pre-training on HumanoidArena.

Model	Ego Proprio	HOI				HSI					Overall AVG (↑)
		Football	DoubleDesk	P&PBox	AVG (↑)	OpenDoor	SitSofa	Boxing	VisNavi	AVG (↑)	
HuGE-OFT	×	33.33	0.00	76.67	36.67	86.67	65.00	73.33	60.00	71.25	56.43
	✓	48.33	0.00	53.33	33.89	75.00	55.00	73.33	73.33	69.17	54.05 (↓2.38)
HuGE-GR00T	×	63.33	3.33	76.67	47.78	93.33	61.67	71.67	61.67	72.09	61.67
	✓	71.67	15.00	81.67	56.11	95.00	78.33	68.33	56.67	74.58	66.67 (↑5.00)
HuGE-Pi	×	60.00	5.00	78.33	47.78	90.00	68.33	73.33	60.00	72.92	62.14
	✓	63.33	11.67	76.67	50.56	88.33	83.33	68.33	63.33	75.83	65.00 (↑2.86)

4.1.3 Scaling Egocentric Pre-Training

We next examine whether scaling egocentric pre-training improves downstream humanoid control. As shown in Figure 2, HuGE-GR00T and HuGE-Pi show progressively better downstream performance as egocentric pre-training data increases across the evaluated scales, reaching average performance of 66.67% and 65.00%, respectively, with the full corpus. These gains suggest that additional egocentric experience provides transferable information that benefits downstream control after humanoid adaptation and task-specific fine-tuning. However, HuGE-OFT declines at the larger scale, consistent with the previously discussed state truncation issue.

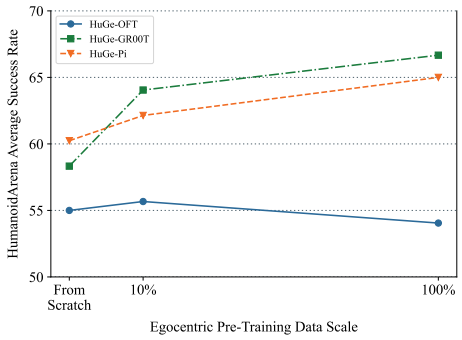


Figure 2: Effect of egocentric pre-training scale.

4.1.4 Performance Comparison

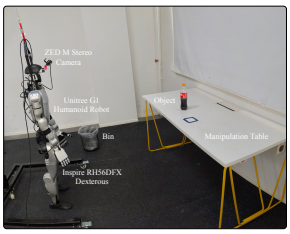
To put our results in perspective, we compare the three HuGE variants to a number of baselines on HumanoidArena. Specifically, Table 4 includes ACT (Zhao et al., 2023), DP (Chi et al., 2023), FM (Barreiros et al., 2026), and $\pi_{0.5}$ (Physical Intelligence et al., 2025b) results reported in the original HumanoidArena study (Wang et al., 2026a). Additionally, we evaluate GR00T (Bjorck et al., 2025) baselines trained from scratch or initialized from GR00T N1.7. Overall, HuGE-GR00T and HuGE-Pi achieve average success rates of 66.67% and 65.00%, respectively, exceeding the strongest baseline, $\pi_{0.5}$, by 12.86 and 11.19 percentage points. Moreover, both variants outperform all listed baselines in the HOI and HSI averages, indicating that the overall gains extend across both interaction categories. Meanwhile, HuGE-OFT achieves a lower overall average of 54.05% but leads on Boxing and VisNavi with 73.33% success on each task, showing that the strongest overall variant does not dominate every task. However, all three variants achieve success rates between 0% and 15.00% on DoubleDesk. This limited performance may be related to the prevalence of demonstrations containing repeated grasping attempts and the emphasis on whole-body humanoid data during pre-training. To address this limitation, future work could incorporate large-scale robotic arm manipulation data into pre-training to improve fine-grained grasping.

4.2 Real-World Evaluation

Since real human egocentric demonstrations cannot be collected directly in simulated task environments, we conduct real-world experiments to investigate downstream fine-tuning with egocentric data. We evaluate HuGE on a cola pick-and-drop task using a Unitree G1 equipped with Inspire dexterous hands, as shown in Figure 3(a). Specifically, the robot must approach a table, pick up a bottle, turn and walk toward a bin, and crouch to place the bottle inside. By combining locomotion, grasping, and object placement, this task is designed to assess coordinated whole-body control over

Table 4: Main results on HumanoidArena. Results marked with † are retrieved from the original HumanoidArena paper. All values are success rates (%).

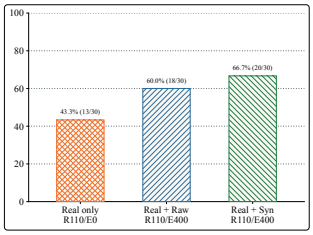
Model	Pre-Trained	HOI				HSI					Overall AVG (↑)
		Football	DoubleDesk	P&PBox	AVG (↑)	OpenDoor	SitSofa	Boxing	VisNavi	AVG (↑)	
ACT (Zhao et al., 2023)†	N/A	20.00	30.00	60.00	36.67	58.30	21.70	58.30	28.30	41.65	39.51
DP (Chi et al., 2023)†	N/A	53.30	35.00	0.00	29.43	83.30	83.30	21.70	50.00	59.58	46.66
FM (Barreiros et al., 2026)†	N/A	21.70	20.00	0.00	13.90	76.70	48.30	18.30	38.30	45.40	31.90
$\pi_{0.5}$ (Physical Intelligence et al., 2025b)†	N/A	10.00	30.00	75.00	38.33	85.00	70.00	71.70	35.00	65.43	53.81
GR00T (Bjorck et al., 2025)	N/A	0.00	0.00	75.00	25.00	11.67	13.33	68.33	36.67	32.50	29.29
GR00T (Bjorck et al., 2025)	GR00T N1.7	13.33	0.00	76.67	30.00	8.33	75.00	11.67	26.67	30.42	30.24
HUGE-OFT	✓	48.33	0.00	53.33	33.89	75.00	55.00	73.33	73.33	69.17	54.05
HUGE-GR00T	✓	71.67	15.00	81.67	56.11	95.00	78.33	68.33	56.67	74.58	**66.67**
HUGE-Pi	✓	63.33	11.67	76.67	50.56	88.33	83.33	68.33	63.33	75.83	65.00



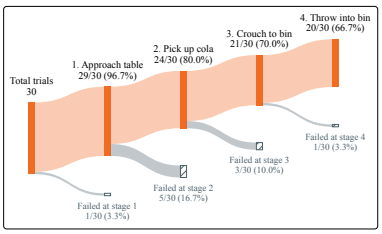
(a) Real-world experiment setup



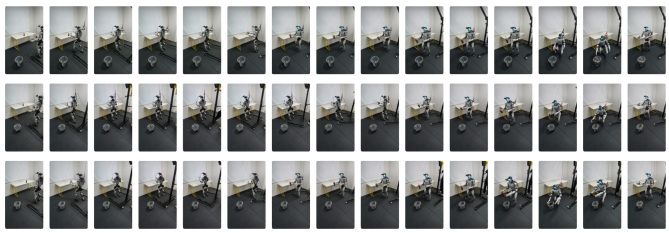
(b) Object placements for evaluating spatial generalization



(c) Success rates across fine-tuning data compositions



(d) Stage-wise performance over 30 trial



(e) Real-world execution sequences on the Unitree G1

Figure 3: Real-world evaluation of the cola pick-and-drop task on the Unitree G1. In (c), R and E denote real-world robot and egocentric demonstration counts, respectively, while Raw and Syn denote raw and visually synthesized egocentric demonstrations. In (d), flow widths and percentages are relative to all 30 trials, with failures recorded at the first failed stage.

a long-horizon sequence. Furthermore, we randomize the bottle position on the tabletop to assess generalization, as illustrated in Figure 3(b). For these real-world experiments, we use only the π-style architecture of HUGE. To assess the contribution of egocentric data, we compare fine-tuning on 110 robot demonstrations alone with augmentation using 400 raw or visually adapted egocentric demonstrations. For each configuration, we conduct 30 trials and report the success rate as the fraction of trials that complete the full task.

4.2.1 VISUAL ADAPTATION

During downstream fine-tuning, we augment egocentric human demonstrations with pseudo-proprioceptive states similar to the pre-training. Beyond these state inputs, we address a prominent visual discrepancy between human and robot observations caused by differences in hand and arm appearance. Specifically, we remove the visible human hands and arms, inpaint the exposed regions, and composite simulation-rendered robotic hands into the videos while preserving the original action, state, and language supervision. Figure 4 illustrates corresponding frames before and after visual adaptation, illustrating the replacement of human hands and arms with rendered robotic counterparts. To quantify the resulting visual changes, we compare the Fréchet distance (FD) from raw and visually adapted egocentric feature distributions to real robot feature distributions. Specifically, visual adaptation reduces FD from 2.0311 to 1.3499 (33.54%) for Wan2.2 (Wang et al., 2025), from 7.5697 to 5.5204 (27.07%) for VideoMAE V2 (Wang et al., 2023b), and from 305.5143 to 263.2697 (13.83%) for V-JEPA 2 (Assran et al., 2025). These reductions indicate that visually adapted ego-

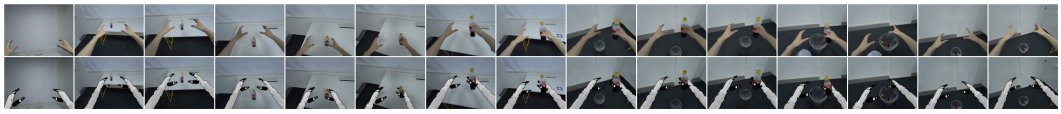


Figure 4: Egocentric observations before (top) and after (bottom) visual adaptation.

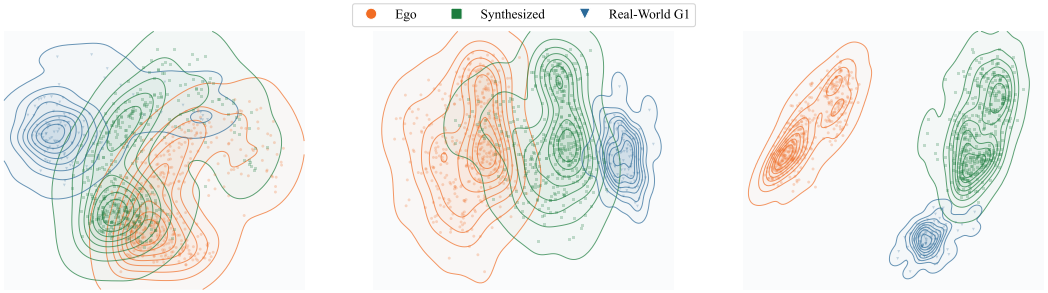


Figure 5: PCA projections of ℓ_2-normalized features from raw egocentric (Ego), visually adapted (Synthesized), and real-world G1 demonstrations, using Wan2.2 (Wang et al., 2025), VideoMAE V2 (Wang et al., 2023b), and V-JEPA 2 (Assran et al., 2025) from left to right. Contours show Gaussian kernel density estimates.

centric observations more closely resemble real robot observations across all three evaluated feature spaces. Consistently, Figure 5 shows synthesized features shifting toward the real robot distributions, providing complementary qualitative evidence of reduced visual discrepancy.

4.2.2 Downstream Fine-Tuning

Based on this, we evaluate whether egocentric demonstrations and visual adaptation improve downstream fine-tuning on the real-world cola pick-and-drop task. As shown in Figure 3(c), supplementing 110 robot demonstrations with 400 raw egocentric demonstrations increases success from 43.3% (13/30) to 60.0% (18/30), supporting the value of human task experience under a fixed robot demonstration budget. Furthermore, using visually adapted versions of the same egocentric demonstrations raises success to 66.7% (20/30). With action, state, and language supervision unchanged, these two additional successful trials suggest that visual adaptation improves the use of human demonstrations for downstream control, consistent with the reduced visual discrepancy observed in the preceding analysis. At the task level, Figure 3(d) reports 20 successful completions across 30 trials, with bottle pickup accounting for five of the ten failed trials and representing the largest source of failure. During deployment, we observed balance instability during bottle grasping, which we attribute in part to the weight of the dexterous hands. This instability contributes to unsuccessful grasp attempts and highlights the physical challenge of maintaining balance during manipulation. Finally, Figure 3(e) presents three execution sequences under different task conditions. Specifically, the first sequence includes a deliberate perturbation absent from the fine-tuning data: the bottle is knocked over during grasping. Meanwhile, the second and third sequences use different bottle positions on the tabletop. These demonstrations provide qualitative evidence of adaptability to object disturbances and spatial generalization across the evaluated bottle placements.

5 Conclusion

We presented HuGe, a family of humanoid VLAs trained on egocentric human and robot demonstrations through a three-stage pipeline. Building on established VLA architectures and a pretrained whole-body controller, HuGe incorporates pseudo-proprioceptive states and visual adaptation to support human-to-humanoid transfer. Experiments on HumanoidArena and a real-world G1 pick-and-drop task support the value of egocentric human experience for whole-body policy learning.

AI USE STATEMENT

The authors wrote and iteratively revised the manuscript, using generative AI tools, primarily GPT 6 Astra, to assist with wording and language polishing throughout the writing process. Additionally, Claude Opus 4.1, Claude Opus 4.5, Claude Opus 4.7, Claude Opus 4.8, Claude Fable 5, GPT 5.4, GPT 5.5, GPT 5.6 Sol, and GPT 6 Astra assisted with implementing specific functionality according to project requirements via vibe coding defined by the authors. For schematic illustration, AI assistance was limited to generating the Ego and Tele-op data-collection illustrations in Stage 3 of Figure 1. Beyond these assistance tasks, generative AI was not used to formulate research questions or hypotheses, design the methodology or experiments, or interpret the results. The authors take full responsibility for the final manuscript and associated research artifacts, including AI-assisted text, code, and illustrations.

ETHICS STATEMENT

All human demonstrations newly collected for this work were provided by the authors. For physical safety during real-world robot experiments, a suspension safety frame was used to provide fall protection. Additionally, we will provide data processing scripts for the existing datasets without directly redistributing the processed data.

REPRODUCIBILITY STATEMENT

The main paper and appendices document the experimental setup to support reimplementation and verification of the reported results. Specifically, Section 3 describes the model architectures, action representations, data sources and processing procedures, and three-stage training pipeline. Complementing these descriptions, Appendix A provides training hyperparameters, hardware configurations, image preprocessing, and downstream data sampling, while Appendix B specifies the proprioceptive state representation and normalization. In addition, Section 4 describes the HumanoidArena and real-world evaluation settings, episode and trial counts, and success-rate computation, with the corresponding task instructions provided in Appendix C.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *CoRR*, 2025.

Shuanghao Bai, Meng Li, Xinyuan Lv, Jiawei Wang, Xinhua Wang, Fei Liao, Chengkai Hou, Langzhe Gu, Wanqi Zhou, Kun Wu, Ziluo Ding, Zhiyuan Xu, Lei Sun, Shanghang Zhang, Zhengping Che, Jian Tang, and Badong Chen. HEX: Humanoid-Aligned Experts for Cross-Embodiment Whole-Body Manipulation. *CoRR*, 2026.

Jose A. Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro M. Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Masayuki Masuda, Jean Mercat, Patrick Tree Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Özgün Önol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael A. Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas

Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A careful examination of large behavior models for multitask dexterous manipulation. *Sci. Robotics*, 2026.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bosnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *CoRR*, 2024.

BitRobot and Unitree and Hugging Face. HIW-500: Humanoids In-the-Wild Dataset for Robot Learning. Hugging Face, 2026. Dataset.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith LLontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *CoRR*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. *CoRR*, 2024.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proc. of RSS*, 2023.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Huaicheng Zhou, Wenshuo Feng, Jiacheng Liu, Siteng Huang, and Donglin Wang. Humanoid-VLA: Towards Universal Humanoid Control with Visual Integration. *CoRR*, 2025.

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 *Proc. of CoRL*, pp. 1516–1540, 2024.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli sivapurapu, and Jian Zhang. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. In *Proc. of ICLR*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of ICLR*, 2022.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, wang chuang, zhihui peng, and Hongyang Li. WholeBodyVLA: Towards Unified Latent VLA for Whole-body Loco-manipulation Control. In *Proc. of ICLR*, 2026.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling Imitation Learning via Egocentric Video. In *Proc. of ICRA*, pp. 13226–13233, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, pp. 2679–2713, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. *CoRR*, 2025.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training Robots Without Robots Using Only Human Videos. In *9th Annual Conference on Robot Learning*, 2025.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *Proc. of CoRL*, 2022.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. *CoRR*, 2026a.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos. *CoRR*, 2025.

Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, and Shanghang Zhang. ω-0: A Latent Predictive World Action Model for Concurrent Humanoid Loco-Manipulation. *CoRR*, 2026b.

Minghui Lin, Pengxiang Ding, Shu Wang, Zifeng Zhuang, Yang Liu, Xinyang Tong, Wenxuan Song, Shangke Lyu, Siteng Huang, and Donglin Wang. HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models. In *Proc. of CVPR*, pp. 20732–20742, 2026.

Jiacheng Liu, Pengxiang Ding, Qihang Zhou, Yuxuan Wu, Da Huang, Zimian Peng, Wei Xiao, Weinan Zhang, Lixin Yang, Cewu Lu, and Donglin Wang. TrajBooster: Boosting Humanoid Whole-Body Manipulation via Trajectory-Centric Learning. *CoRR*, 2025.

Yang Liu, Pengxiang Ding, Tengyue Jiang, Xudong Wang, Wenxuan Song, Minghui Lin, Han Zhao, Hongyin Zhang, Zifeng Zhuang, Wei Zhao, Siteng Huang, Jinkui Shi, and Donglin Wang. MMaDA-VLA: Large Diffusion Vision-Language-Action Model with Unified Multi-Modal Instruction and Generation. *CoRR*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proc. of ICLR*, 2019.

Jingxian Lu, Wenke Xia, Yuxuan Wu, Zhiwu Lu, and Di Hu. When would Vision-Proprioception Policies Fail in Robotic Manipulation? In *Proc. of ICLR*, 2026.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos. *CoRR*, 2025.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-H0.5: Scaling Human-Centric Robot Learning for Cross-Embodiment Generalization. *CoRR*, 2026a.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Jim Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Sci. Robotics*, 2026b.

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 *Proc. of ICLR*, 2023.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *Proc. of CoRL*, pp. 892–909, 2022.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a VLA That Learns From Experience. *CoRR*, 2025a.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *CoRR*, 2025b.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. *CoRR*, 2026.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Chenyu Yang, Xi Wang, James Fort, Richard A. Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert K. Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An Egocentric Human Dataset for Robot Learning from Around the World. *CoRR*, 2026.

Zekun Qi, Xuchuan Chen, Dairu Liu, Chenghuai Lin, Yunrui Lian, Sikai Liang, Zhikai Zhang, Yu Guan, Jilong Wang, Wenyao Zhang, Xinqiang Yu, He Wang, and Li Yi. Humanoid-GPT: Scaling Data and Structure for Zero-Shot Motion Tracking. *CoRR*, 2026.

Qwen Team. Qwen3-VL Technical Report. *CoRR*, 2025.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-World Robot Learning with Masked Visual Pre-training. In *Proc. of CoRL*, pp. 416–426, 2022.

Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Sci. Robotics*, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, pp. 20, 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proc. of KDD*, pp. 3505–3506, 2020.

Ropedia. Xperience-10M: A Large-Scale Egocentric Multimodal Dataset with Structured 3D/4D Annotations. Hugging Face, 2026. Dataset.

Wenxuan Song, Han Zhao, Fuhao Li, Ziyang Zhou, Xi Wang, Jing Lyu, Pengxiang Ding, Yan Wang, Donglin Wang, and Haoang Li. CapVector: Learning Transferable Capability Vectors in Parametric Space for Vision-Language-Action Models. *CoRR*, 2026.

StarVLA Community. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *CoRR*, 2026.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulation. In *Proc. of ICLR*, 2025.

Unitree Robotics. UnifoLM-WBT-Dataset. Hugging Face, 2026. Dataset.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *CoRR*, 2025.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-Horizon Imitation Learning by Watching Human Play. In *Proc. of CoRL*, pp. 201–221, 2023a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *Proc. of CVPR*, pp. 14549–14560, 2023b.

Taowen Wang, Zikang Xie, Bin Yang, Yunheng Wang, Zizhao Yuan, Yuetong Fang, Yixiao Feng, Yichi Wang, Xingyu Chen, Haodong Chen, Qiwei Wu, Weisheng Xu, Lihan Chen, Lusong Li, Zecui Zeng, and Renjing Xu. HumanoidArena: Benchmarking Egocentric Hierarchical Whole-body Learning. *CoRR*, 2026a.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified Vision-Language-Action Model. In *Proc. of ICLR*, 2026b.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, Sheng Zang, Weiduo Yuan, Marco Pavone, Di Huang, and Yue Wang. Ψ_0: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation. *CoRR*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World Action Models are Zero-shot Policies. *CoRR*, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do World Action Models Need Test-time Future Imagination? *CoRR*, 2026.

Tao Zhang, Song Xia, Ye Wang, and Qin Jin. EasyMimic: A Low-Cost Framework for Robot Imitation Learning from Human Videos. *CoRR*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proc. of RSS*, 2023.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Campagnolo Guizilini, and Yue Wang. Humanoid Everyday: A Comprehensive Robotic Dataset for Open-World Humanoid Manipulation. In *Proc. of ICRA 2026 Workshop*, 2026.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. MotionWAM: Towards Foundation World Action Models for Real-Time Humanoid Loco-Manipulation. *CoRR*, 2026a.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data. *CoRR*, 2026b.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, pp. 2165–2183, 2023.

A Implementation Details

All HuGE variants use pretrained Qwen3-VL-2B-Instruct (Qwen Team, 2025) as the vision-language backbone. Under full-parameter fine-tuning, we jointly update the backbone, projection layers, and action modules, while keeping the SONIC controller fixed. For distributed training, we use DeepSpeed (Rasley et al., 2020; Rajbhandari et al., 2020) on a single node equipped with eight NVIDIA 4090 GPUs, each with 48 GB of memory. With this setup, Stage 1 and Stage 2 training take approximately 38 and 12 hours, respectively. Detailed configurations are provided in Table 5.

Table 5: Training hyper-parameters of HuGE.

Configuration	Stage 1	Stage 2	Stage 3	Configuration	Stage 1	Stage 2	Stage 3
VLM Learning Rate	0.00001			Projector Learning Rate	0.000025		
Action Head Learning Rate	0.0001			Learning Schedule	Cosine Decay		
Warm-up Ratio	5	2	1	Numerical Precision	BFloat16		
Global Batch Size	256 / 192 (OFT-Style only)			Epoch	1	1	20k (Step)
Sample Stride	10	5	1	Optimizer	AdamW (Loshchilov & Hutter, 2019)		

All training stages share the same image preprocessing pipeline. Specifically, RGB frames are downsampled when necessary to limit the longer side to 512 pixels while preserving the aspect ratio, padded with black pixels to form a centered square, and resized to 256×256 using bicubic interpolation. During training, we randomly crop a 243×243 region, corresponding to approximately 95% of each spatial dimension, and resize the crop back to 256×256 using bicubic interpolation. Additionally, we apply brightness, contrast, and saturation jitter sequentially, with factors sampled uniformly from $[0.7, 1.3]$, $[0.6, 1.4]$, and $[0.5, 1.5]$, respectively. At inference, we replace random cropping with center cropping and disable color jitter. Finally, the resulting RGB images are passed to the Qwen3-VL processor to construct the visual model inputs.

Across the three real-world configurations, the robot demonstrations remain fixed. For the two configurations augmented with egocentric demonstrations, raw and visually adapted videos share the same action, state, and language supervision. In both augmented configurations, robot and egocentric sample starts receive equal expected sampling probability. Additionally, all three configurations share normalization statistics computed from the preceding training data and both downstream robot and egocentric demonstrations.

B PROPRIOCEPTIVE STATE REPRESENTATION

State layout. Each HUGE variant conditions action-chunk prediction on a single 34-dimensional G1 state at the starting index t:

$$s_t = \left[(q_t^{\text{body}})^\top,\ h_t^L,\ h_t^R,\ (g_t^{\text{body}})^\top\right]^\top \in \mathbb{R}^{34}, \tag{1}$$

where $q_t^{\text{body}} \in \mathbb{R}^{29}$ contains body joint positions in radians, $h_t^L, h_t^R \in \{0, 1\}$ encode left and right hand states (0 for open and 1 for closed), and $g_t^{\text{body}} \in \mathbb{R}^3$ is the unit gravity direction in the body frame. The corresponding dimension order and units are listed in Table 6.

Table 6: The 34-dimensional G1 proprioceptive state used by HUGE. Joint fields give angular positions in radians, and hand states use 0 for open and 1 for closed. Gravity fields give the dimensionless x, y, and z components of the unit gravity direction in the body frame.

Index	Field	Meaning	Unit	Index	Field	Meaning	Unit
0	`left_hip_pitch_joint`	Left hip pitch	rad	17	`left_shoulder_yaw_joint`	Left shoulder yaw	rad
1	`left_hip_roll_joint`	Left hip roll	rad	18	`left_elbow_joint`	Left elbow	rad
2	`left_hip_yaw_joint`	Left hip yaw	rad	19	`left_wrist_roll_joint`	Left wrist roll	rad
3	`left_knee_joint`	Left knee	rad	20	`left_wrist_pitch_joint`	Left wrist pitch	rad
4	`left_ankle_pitch_joint`	Left ankle pitch	rad	21	`left_wrist_yaw_joint`	Left wrist yaw	rad
5	`left_ankle_roll_joint`	Left ankle roll	rad	22	`right_shoulder_pitch_joint`	Right shoulder pitch	rad
6	`right_hip_pitch_joint`	Right hip pitch	rad	23	`right_shoulder_roll_joint`	Right shoulder roll	rad
7	`right_hip_roll_joint`	Right hip roll	rad	24	`right_shoulder_yaw_joint`	Right shoulder yaw	rad
8	`right_hip_yaw_joint`	Right hip yaw	rad	25	`right_elbow_joint`	Right elbow	rad
9	`right_knee_joint`	Right knee	rad	26	`right_wrist_roll_joint`	Right wrist roll	rad
10	`right_ankle_pitch_joint`	Right ankle pitch	rad	27	`right_wrist_pitch_joint`	Right wrist pitch	rad
11	`right_ankle_roll_joint`	Right ankle roll	rad	28	`right_wrist_yaw_joint`	Right wrist yaw	rad
12	`waist_yaw_joint`	Waist yaw	rad	29	`left_hand`	Left hand state	Binary
13	`waist_roll_joint`	Waist roll	rad	30	`right_hand`	Right hand state	Binary
14	`waist_pitch_joint`	Waist pitch	rad	31	`projected_gravity[0]`	Gravity direction, x	Scalar
15	`left_shoulder_pitch_joint`	Left shoulder pitch	rad	32	`projected_gravity[1]`	Gravity direction, y	Scalar
16	`left_shoulder_roll_joint`	Left shoulder roll	rad	33	`projected_gravity[2]`	Gravity direction, z	Scalar

Normalization. We clip and scale each state dimension using the 1st and 99th percentile bounds, ℓ_i and u_i:

$$\hat{s}_{t,i} = 2\frac{\text{clip}(s_{t,i}, \ell_i, u_i) - \ell_i}{u_i - \ell_i + 10^{-8}} - 1. \tag{2}$$

These bounds are computed per dimension from pooled training frames across cumulative stages. Specifically, Stage 1 uses all Stage 1 training data, Stage 2 pools all training data from Stages 1–2, and Stage 3 pools all training data from Stages 1–3. At inference, we reuse the Stage 3 normalization statistics saved with the checkpoint.

C TASK INSTRUCTIONS

Table 7 lists the language instructions used for the seven HumanoidArena tasks (Wang et al., 2026a) and the real-world cola pick-and-drop task.

Table 7: Task instructions for stage 3.

Setting	Task	Task instruction
HumanoidArena	Football	Kick the football into the goal.
	DoubleDesk	Put the hammer from the right table into the basket on the left table.
	P&PBox	Move the box from the table onto the shelf.
	OpenDoor	Open the door.
	SitSofa	Sit on the sofa.
	Boxing	Strike the green markers on the punching bag.
	VisNavi	Avoid obstacles and move to the yellow marked area.
Real world	Cola pick-and-drop	Walk forward, grab the cola and throw into the trash bin