VAIC: Vision-Guided Humanoid Agile Object Interaction Control via Decoupled Commands

Dongting Li[1,3*] **Qianyang Wu**[3*] **Xingyu Chen**[2] **Liang Li**[3] **Yuhang Lin**[3]
Sikai Wu[3] **Guoyao Zhang**[3] **Mingliang Zhou**[3] **Diyun Xiang**[3] **Qiang Zhang**[2]
Renjing Xu[2] **Jianzhu Ma**[1†]
[1] Tsinghua University [2] HKUST(Guangzhou) [3] Xiaomi Robotics Lab
*Equal Contribution † Corresponding author
https://vaic-humanoid.github.io/

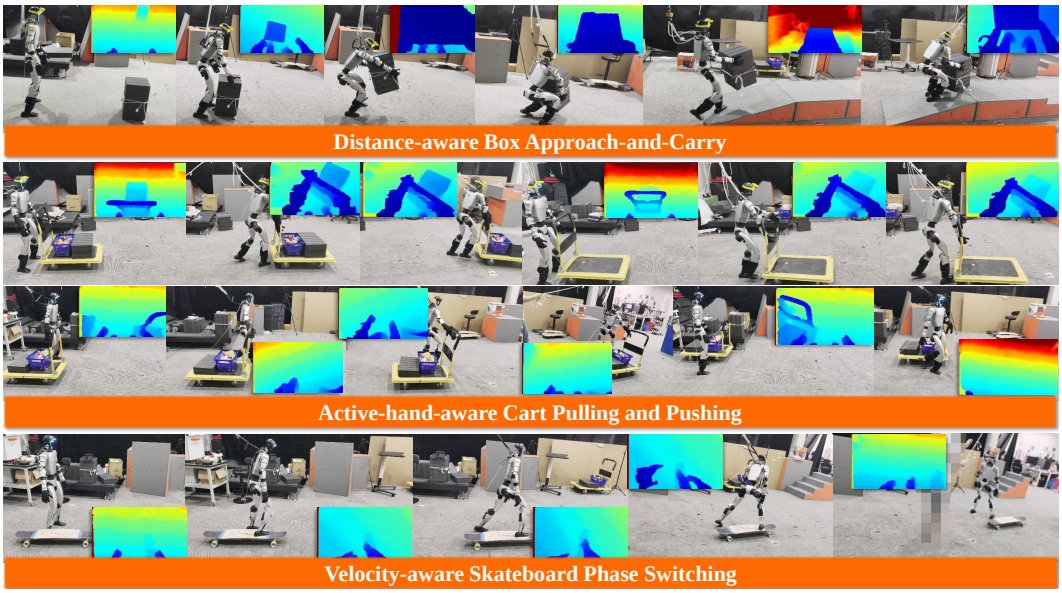

Figure 1: We propose VAIC, a unified vision-guided control framework that enables a humanoid robot to execute highly diverse and agile object interactions. By operating on onboard depth and decoupled user commands, VAIC successfully deploys dynamic tasks, such as distance-aware box carrying, underactuated cart interaction, and skateboarding.

Abstract: Humanoid robots hold immense potential for real-world assistance, yet agile interaction with objects in unstructured environments demands tightly coupled whole body coordination. Despite recent advancements, current controllers face a critical deployment gap. They rely heavily on dense reference trajectories and perfect state observability, which inherently limits physical generalization. We present Vision Guided Agile Interaction Control (VAIC), a unified framework that bridges this gap by operating exclusively on onboard depth, historical proprioception, and a decoupled user command interface. VAIC employs a two-stage distillation paradigm. First, a privileged teacher policy masters diverse interaction skills using precise object kinematics and exact environmental states. Second, a deployable student policy distills these capabilities by replacing full body tracking with velocity targets across multiple axes and an interaction indicator for each frame. The student utilizes a recurrent object adaptation module to implicitly infer unobservable object dynamics from raw depth streams and proprioception. Evaluations and real-world deployments on the humanoid robot demonstrate that a single VAIC policy successfully executes highly diverse dynamic tasks. These tasks include box carrying, cart interaction, and skateboarding, consistently outperforming baselines and advancing autonomous humanoid deployment.

Keywords: Whole-Body Control, Visual Policy, Humanoid-Object Interaction

1 Introduction

Humanoid robots are increasingly expected to operate in unstructured human environments, executing tasks such as carrying bulky objects over stairs, manipulating underactuated carts, and adapting to abrupt dynamic disturbances. Unlike pure locomotion, these tasks demand tightly coupled whole body coordination, where the dynamic state of the interacting object fundamentally alters the robot's balance and contact mechanics.

Recent advances in deep reinforcement learning and motion imitation have enabled increasingly capable humanoid-object interaction [1, 2, 3, 4]. Prior methods, such as HAIC [5], demonstrate that explicitly modeling coupled dynamics can enable agile interactions by tracking rich human demonstrations. However, deploying these capabilities in the physical world exposes a critical deployment gap characterized by two fundamental limitations. First, existing controllers typically demand dense, joint-level reference trajectories at runtime[6, 7, 2, 5]. This kinematic hand holding is highly impractical for real-world autonomy or teleoperation, where human operators can only provide high-level intent, such as a joystick velocity command [8, 9]. Second, these methods heavily rely on perfect state observability, such as the ground truth 6D pose of objects and precise terrain geometry [2, 10, 3, 11]. In physical deployment, the robot's downward view is frequently occluded by the objects it carries, and real time state estimation is plagued by severe sensor noise and blind spots.

We address both limitations with VAIC (Vision Guided Agile Interaction Control), a unified learning framework that bridges the gap between oracle-level simulation tracking and physical, perceptive deployment. The core insight of VAIC is to decouple high-level user intent from low-level interaction execution. Instead of forcing the deployable policy to track a dense reference motion, we condition it on a unified, intent-driven command interface comprising a per-frame velocity command for navigational intent and a per-frame interaction state specifying the manipulation phase. To overcome severe visual occlusion without losing critical interaction priors, VAIC employs a two-stage distillation paradigm. A privileged teacher policy first masters diverse interactive skills using exact object kinematics and terrain geometry. Subsequently, a deployable student policy distills these capabilities by learning to implicitly infer the missing privileged states, relying solely on temporal depth features, proprioceptive history, and the user commands via a recurrent Object Adaptation module.

This unified formulation allows a single policy to generalize across radically different dynamic profiles without task specific retraining. We validate VAIC across three highly challenging categories of real-world interaction: (1) **Box Carrying** across varied terrain (flat, slopes, and stairs) under severe visual occlusion; (2) **Cart Manipulation**, managing the non holonomic constraints and external forces of an underactuated payload; and (3) **Skateboarding**, maintaining balance under impulsive contact dynamics. Concretely, our primary contributions are three points:

1) We propose a Decoupled Command Interface for Interaction. We introduce a deployment friendly control formulation that replaces dense reference motion requirements with a high level command interface: multi axis velocity targets for locomotion intent and a per frame interaction state for phase control. This elegantly separates "where to go" from "how to interact," enabling intuitive teleoperation and autonomous navigation.

2) We develop an Occlusion Resistant Distillation method. To effectively utilize vision and handle imperfect observations, we introduce a depth recurrent Object Adaptation module. This module successfully distills privileged geometric and dynamic states from the teacher, allowing the robot to implicitly reconstruct unobservable object dynamics and execute precise interactions even when its exteroceptive vision is severely occluded or plagued by sensor noise.

3) We provide Unified Hardware Validation. We demonstrate that a single *ours* policy can generalize across diverse terrain sequences and seamlessly adapt to out-of-distribution object attributes. Extensive hardware deployments on the humanoid robot validate the system's robustness across car-

rying, pushing, pulling, and skateboarding behaviors, providing strong empirical evidence for its physical adaptability to unmodeled dynamic variations.

2 Related Work

2.1 Humanoid Whole-Body Control

Physics-based motion imitation has established the foundation for acquiring rich humanoid motor skills. DeepMimic and AMP demonstrated that reference-conditioned training and adversarial motion priors can produce natural and dynamic whole-body behaviors [12, 8]. Building on this foundation, recent general motion trackers and retargeting frameworks further improve robustness and deployability through universal tracking, morphology-aware retargeting, residual correction, and simulation-to-real alignment [13, 14, 15, 16, 17, 6, 3]. More recent efforts scale such policies toward behavior priors and foundation-style controllers, including Kimodo [18] for large-scale controllable motion generation and SONIC [7] for supersized motion tracking with larger models, datasets, and compute, alongside promptable or generalist whole-body control models [19, 20, 21, 22, 23]. However, these methods mainly focus on free-body motion skills, whereas our setting requires whole-body control under physical coupling with interactive objects.

2.2 Humanoid-Object Interaction

Learning to interact with objects requires jointly reasoning about whole-body contact, balance, and object dynamics. Large-scale human-object datasets provide diverse demonstrations for data-driven skill acquisition [24, 25, 26, 27, 28], while simulated Humanoid-Object Interaction frameworks leverage such data for motion imitation, generative synthesis, or task-tokenized priors [29, 30, 31, 32, 33]. For real robots, recent loco-manipulation systems demonstrate box carrying, force-adaptive manipulation, unified whole-body control, heterogeneous meta-control, and dynamics-aware agile interaction [34, 35, 36, 37, 5]. Athletic humanoid systems further extend object interaction to highly dynamic tasks such as tennis, badminton, soccer, and ping-pong [38, 39, 40, 41], while video-based imitation reduces manual motion engineering by extracting interactive priors from human demonstrations [2, 42, 10]. In contrast to methods that assume object states are available from simulation, motion capture, or external tracking[3, 1, 11, 43], VAIC targets interaction with onboard object perception.

2.3 Perceptive Humanoid Control

Depth cameras, stereo sensors, and LiDAR have been integrated into locomotion policies to traverse stairs, gaps, slopes, and stepping stones [44, 45, 46, 47, 48, 49], while perceptive parkour systems combine depth sensing with dynamic motion tracking or skill composition to negotiate challenging obstacles [50, 51, 52, 53]. In these works, perception is mainly used for foothold selection or obstacle avoidance in a static environment. Vision has also been introduced into humanoid-object interaction tasks. Some methods explore latent vision-language instructions, visual sim-to-real transfer, and end-to-end VLA architectures for loco-manipulation [54, 55, 56], while ego-vision contact planning uses forward-looking depth to anticipate contact geometry [57]. VisualMimic, LessMimic, Pro-HOI, and ULTRA use depth observations, perceptive object representations, or root-state guidance to support humanoid-object interaction [4, 58, 9, 59]. Different from these approaches, VAIC effectively combines vision and proprioceptive history to continuously infer the latent dynamics of the interacted object, enabling adaptive whole-body interaction with user-friendly command.

3 Methodology

Our goal is to train a unified whole-body controller capable of versatile object interactions in unstructured environments. As illustrated in Figure 2, VAIC bridges the gap between oracle simulation and real-world deployment through a specialized two-stage online distillation framework. By for-

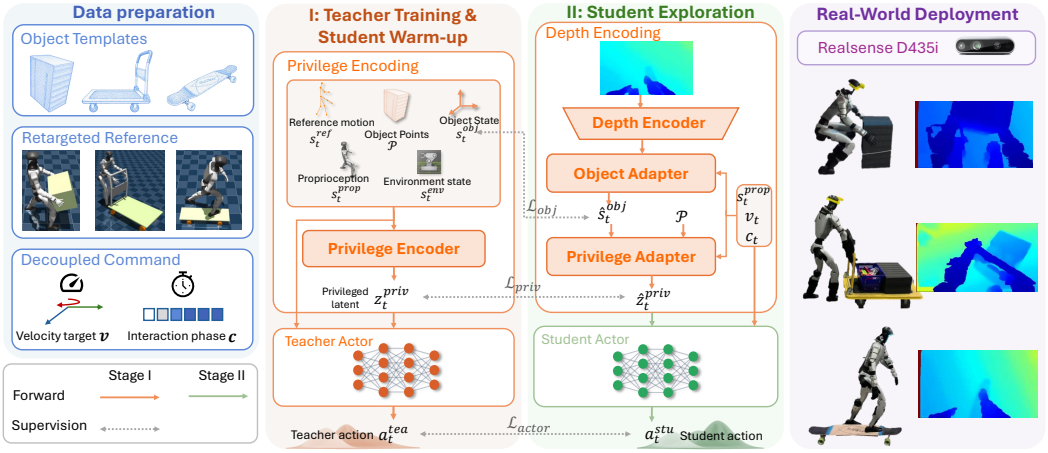

Figure 2: Overview of VAIC. The framework follows a two-stage distillation paradigm. A privileged teacher policy masters interaction skills using exact object states and reference motions. The deployable student policy strictly relies on onboard depth, proprioception, and a decoupled user command interface, distilling the teacher's capability via hierarchical alignment modules.

malizing a decoupled user command interface, we explicitly eliminate the restrictive reliance on dense reference trajectories during physical deployment.

3.1 Decoupled Command Interface

We utilize a comprehensive dataset of human-object interactions retargeted to the humanoid robot (data preparation and geometric object template $\mathcal{P}$ extraction are detailed in Appendix D).

Crucially, to facilitate intuitive teleoperation and autonomous deployment, we formalize a decoupled user command interface defined as $\mathbf{u}_t = (v_t, c_t)$. During the training phase, these commands are directly extracted from the retargeted reference motions to ensure temporal alignment with the demonstrated behaviors. This interface distills high-level task intent into two distinct components: (1) a multi-axis velocity target $v_t \in \mathbb{R}^3$, representing the navigational intent formatted as $[v_x, v_y, v_{\text{yaw}}]$, and (2) a per-frame binary indicator $c_t \in \{0, 1\}$ specifying the manipulation phase, where 0 denotes navigating or approaching, and 1 denotes active contact and interaction.

3.2 Stage 1: Teacher Training and Student Warm-up

In the first stage, we train a privileged teacher policy under an oracle simulation setting via reinforcement learning (PPO). As depicted in the Orange block of Figure 2, a Privilege Encoder compresses the multimodal privileged context into a dense privileged latent z_t^{priv}:

$$z_t^{priv} = f_{\text{enc\_priv}}(s_t^{ref}, s_t^{geo}, s_t^{prop}, s_t^{env}), \quad s_t^{geo} = \mathcal{T}(s_t^{obj}, \mathcal{P}) \tag{1}$$

Here, the transformed geometric feature s_t^{geo} explicitly maps the nominal template $\mathcal{P}$ into the robot's local frame using the ground-truth object state s_t^{obj} to provide exact spatial perception. The Teacher Actor then takes z_t^{priv} and the proprioceptive state s_t^{prop} to output the optimal joint action a_t^{tea}. Operating as an oracle motion tracker, the teacher can fully master the complex contact dynamics.

Concurrent Distillation: During the teacher's PPO exploration, we continuously utilize its rollout data to warm up the deployable student modules (Figure 2, Green block). The Object Adapter and Privilege Adapter are trained via supervised learning ($\mathcal{L}_{obj} = \|\hat{s}_t^{obj} - s_t^{obj}\|^2$, $\mathcal{L}_{priv} = \|\hat{z}_t^{priv} - z_t^{priv}\|^2$) to implicitly infer the unobservable object dynamics and align the latent representations. To maximize computational efficiency during this stage, the visual input is zero-injected, meaning the depth tensor is structurally maintained but populated entirely with zeros. Simultaneously, the Student Actor undergoes behavior cloning (BC) via $\mathcal{L}_{actor} = \|a_t^{stu} - a_t^{tea}\|^2$. Crucially, this distillation features an asymmetric input design: while the teacher tracks the dense reference

s_t^{ref}, the student actor is strictly conditioned on the corresponding decoupled commands (v_t, c_t) extracted from the dataset, completely removing its dependence on the reference kinematics.

3.3 Stage 2: Vision-Guided Student Exploration

While Stage 1 establishes a robust prior, purely off-policy distillation suffers from compounding errors when the student encounters out-of-distribution states during real-world execution. In Stage 2, the dense reference motion and perfect s_t^{obj} are completely removed from the deployable inputs. The Student Actor takes full control of environment exploration using PPO to optimize for task-completion and command-tracking rewards, while the preceding alignment modules continue to be trained via supervised learning using the exact states available during simulation rollouts.

During this active exploration phase, the hierarchical student modules operate as follows:

Object Adapter: A CNN-GRU Depth Encoder extracts temporal features from the noisy depth stream $\mathbf{d}_t$. The Object Adapter implicitly infers the dynamic object state:

$$\hat{s}_t^{obj} = f_{\text{obj\_adapt}}(z_t^{depth}, s_t^{prop}, v_t, c_t), \quad z_t^{depth} = f_{\text{enc\_depth}}(\mathbf{d}_t) \tag{2}$$

The binary indicator c_t serves as a critical mode-switching prior, allowing the recurrent network to anticipate changes in coupled object dynamics and contact forces during interaction transitions.

Privilege Adapter: Using $\hat{s}_t^{obj}$ and the geometric template $\mathcal{P}$, this adapter projects spatial occupancy priors into the robot's frame to seamlessly reconstruct the policy latent:

$$\hat{z}_t^{priv} = f_{\text{priv\_adapt}}(s_t^{prop}, \hat{s}_t^{geo}, v_t, c_t), \quad \hat{s}_t^{geo} = \mathcal{T}(\hat{s}_t^{obj}, \mathcal{P}) \tag{3}$$

Online Adaptation: Crucially, as the Student Actor actively explores and generates new trajectories, the Object Adapter and Privilege Adapter continue to be updated online using the student's own rollout data. This tight coupling prevents representation drift and strictly grounds the visual reconstruction in the student's actual physical interactions.

4 Experiments

4.1 Experimental Setup

4.1.1 Task Description

We evaluate VAIC on three challenging categories of whole-body interaction:

Box Interaction. The robot carries a bulky box across varying terrain profiles, including flat ground, slopes, stairs, and composite sequences. The box severely occludes the robot's downward view, requiring the policy to rely on depth-based implicit state reconstruction for terrain-aware locomotion.

Cart Interaction. The robot manipulates a wheeled cart via pulling and pushing motions. The cart's non-holonomic constraints and underactuated dynamics introduce time-varying external disturbances that actively affect the robot's balance.

Skateboard Interaction. The robot balances on and rides a skateboard, including high-speed gliding and dynamic dismount phases. This demands precise anticipation of the board's translational and rotational dynamics under impulsive forces.

4.1.2 Implementation Details

We conduct experiments on a humanoid robot equipped with an onboard Intel RealSense D435i camera. The policy is executed at 50 Hz, while the low-level joint controller runs at 200 Hz. Depth observations from the RealSense camera are captured at 60 Hz.

Data Collection. Human demonstrations were captured via an optical motion capture system tracking both the performer and the interactive objects. Raw kinematic data were retargeted to the humanoid robot and refined through a physics-based simulation pipeline to generate physically valid states with precise contact information.

Training. Both training stages were executed using PPO in Isaac Sim on a single NVIDIA RTX 4090 GPU. Stage 1 (Teacher) utilizes 4096 parallel environments; Stage 2 (Student Fine-tuning) utilizes 512 environments with real-time onboard camera rendering enabled. Stage 1 converges in approximately 5 hours. Stage 2 converges in approximately 15 hours.

4.1.3 Evaluation Metrics

All quantitative simulation comparisons are rigorously evaluated in MuJoCo. To provide a concise overview, metrics are averaged across all subtask sequences within each object category. We report the following three categories of metrics. (1) **Success Rate (SR):** An episode is successful if the robot maintains dynamic balance, executes the decoupled command (v_t, c_t), and avoids losing control of the object. (2) **Root State Metrics:** These include tracking errors for the humanoid root in position (E_{rpe}), orientation (E_{roe}), linear velocity (E_{rve}), angular velocity (E_{rave}), and linear acceleration (E_{rae}). (3) **Object State Metrics:** These comprise the corresponding tracking errors for the interactive object (E_{ope}, E_{ooe}, E_{ove}, E_{oave}, E_{oae}).

Table 1: Quantitative evaluation on **Box Interaction** (averaged across flat, slope, and stair terrains). VAIC leverages depth-based state prediction to effectively overcome severe downward visual occlusion caused by the carried box, achieving the highest success rate. Evaluation tracks both the humanoid root and the object state errors.

		Root State Metrics (↓)					Object State Metrics (↓)				
Method	**SR ↑**	E_{rpe}	E_{roe}	E_{rve}	E_{rave}	E_{rae}	E_{ope}	E_{ooe}	E_{ove}	E_{oave}	E_{oae}
PPO	0.0%	$1.83^{\pm 0.40}$	$0.84^{\pm 0.07}$	$0.51^{\pm 0.08}$	$0.43^{\pm 0.05}$	$0.55^{\pm 0.06}$	$0.94^{\pm 0.40}$	$0.89^{\pm 0.35}$	$0.27^{\pm 0.07}$	$0.40^{\pm 0.08}$	$0.43^{\pm 0.11}$
AMP	50.0%	$0.78^{\pm 0.22}$	$0.61^{\pm 0.30}$	$\mathbf{0.30}^{\pm 0.07}$	$0.42^{\pm 0.11}$	$\mathbf{0.42}^{\pm 0.08}$	$1.01^{\pm 0.46}$	$\mathbf{0.65}^{\pm 0.50}$	$0.27^{\pm 0.10}$	$0.25^{\pm 0.12}$	$\mathbf{0.27}^{\pm 0.09}$
VAIC	**66.7%**	$\mathbf{0.59}^{\pm 0.22}$	$\mathbf{0.51}^{\pm 0.14}$	$0.33^{\pm 0.05}$	$\mathbf{0.41}^{\pm 0.05}$	$0.53^{\pm 0.06}$	$\mathbf{0.62}^{\pm 0.09}$	$0.69^{\pm 0.38}$	$\mathbf{0.21}^{\pm 0.02}$	$\mathbf{0.25}^{\pm 0.09}$	$0.28^{\pm 0.03}$

Table 2: Quantitative evaluation on **Cart Interaction** (averaged across pull and push tasks). Underactuated cart dynamics cause standard imitation baselines to frequently fail. VAIC achieves superior success rates and minimal positional error, effectively anticipating the cart's physical feedback.

		Root State Metrics (↓)					Object State Metrics (↓)				
Method	**SR ↑**	E_{rpe}	E_{roe}	E_{rve}	E_{rave}	E_{rae}	E_{ope}	E_{ooe}	E_{ove}	E_{oave}	E_{oae}
PPO	41.7%	$1.01^{\pm 0.48}$	$0.49^{\pm 0.07}$	$0.38^{\pm 0.12}$	$0.32^{\pm 0.06}$	$0.32^{\pm 0.11}$	$0.83^{\pm 0.57}$	$0.04^{\pm 0.04}$	$0.31^{\pm 0.13}$	$0.02^{\pm 0.02}$	$\mathbf{0.20}^{\pm 0.05}$
AMP	58.3%	$0.61^{\pm 0.48}$	$0.54^{\pm 0.41}$	$0.23^{\pm 0.12}$	$\mathbf{0.24}^{\pm 0.07}$	$\mathbf{0.27}^{\pm 0.12}$	$0.76^{\pm 0.51}$	$\mathbf{0.03}^{\pm 0.03}$	$\mathbf{0.26}^{\pm 0.10}$	$\mathbf{0.02}^{\pm 0.02}$	$0.20^{\pm 0.06}$
VAIC	**91.7%**	$\mathbf{0.45}^{\pm 0.09}$	$\mathbf{0.33}^{\pm 0.13}$	$\mathbf{0.22}^{\pm 0.09}$	$0.26^{\pm 0.12}$	$0.30^{\pm 0.09}$	$\mathbf{0.61}^{\pm 0.12}$	$0.10^{\pm 0.12}$	$0.28^{\pm 0.11}$	$0.04^{\pm 0.04}$	$0.22^{\pm 0.06}$

Table 3: Quantitative evaluation on **Skateboard Interaction**. Impulsive contact forces disrupt the balance of pure kinematic trackers, whereas VAIC precisely tracks high-speed dynamics and achieves the highest success rate.

		Root State Metrics (↓)					Object State Metrics (↓)				
Method	**SR ↑**	E_{rpe}	E_{roe}	E_{rve}	E_{rave}	E_{rae}	E_{ope}	E_{ooe}	E_{ove}	E_{oave}	E_{oae}
PPO	0.0%	$2.09^{\pm 0.34}$	$0.99^{\pm 0.33}$	$0.50^{\pm 0.05}$	$0.39^{\pm 0.05}$	$0.45^{\pm 0.04}$	$1.75^{\pm 0.19}$	$0.12^{\pm 0.16}$	$0.39^{\pm 0.04}$	$0.05^{\pm 0.05}$	$\mathbf{0.26}^{\pm 0.03}$
AMP	12.5%	$1.30^{\pm 0.70}$	$0.98^{\pm 0.33}$	$0.40^{\pm 0.11}$	$0.44^{\pm 0.09}$	$0.40^{\pm 0.04}$	$1.20^{\pm 0.62}$	$0.19^{\pm 0.27}$	$0.36^{\pm 0.08}$	$0.07^{\pm 0.10}$	$0.28^{\pm 0.05}$
VAIC	**83.3%**	$\mathbf{0.62}^{\pm 0.12}$	$\mathbf{0.38}^{\pm 0.19}$	$\mathbf{0.29}^{\pm 0.05}$	$\mathbf{0.27}^{\pm 0.07}$	$\mathbf{0.35}^{\pm 0.07}$	$\mathbf{0.62}^{\pm 0.17}$	$\mathbf{0.07}^{\pm 0.06}$	$\mathbf{0.31}^{\pm 0.10}$	$\mathbf{0.02}^{\pm 0.02}$	$0.27^{\pm 0.04}$

4.1.4 Baselines

We compare VAIC in simulation against four representative whole body control and interaction baselines. (1) **PPO** [60]: A standard policy trained directly with dense reference motions, lacking the occlusion resistant distillation. (2) **AMP** [8]: An adversarial motion prior framework designed for kinematic style matching. (3) **PhysHSI** [1] and (4) **VisualMimic** [4]: Two state of the art interactive whole body controllers. The PPO and AMP baselines are retrained from scratch on our motion dataset using an identical reward structure to ensure a fair comparison.

4.2 Quantitative Evaluation

We present the quantitative simulation results across our three primary evaluation categories. Specifically, Table 1 details the performance metrics for the Box Interaction task across varied terrains. Table 2 reports the results for the Cart Interaction task under underactuated dynamics. Table 3

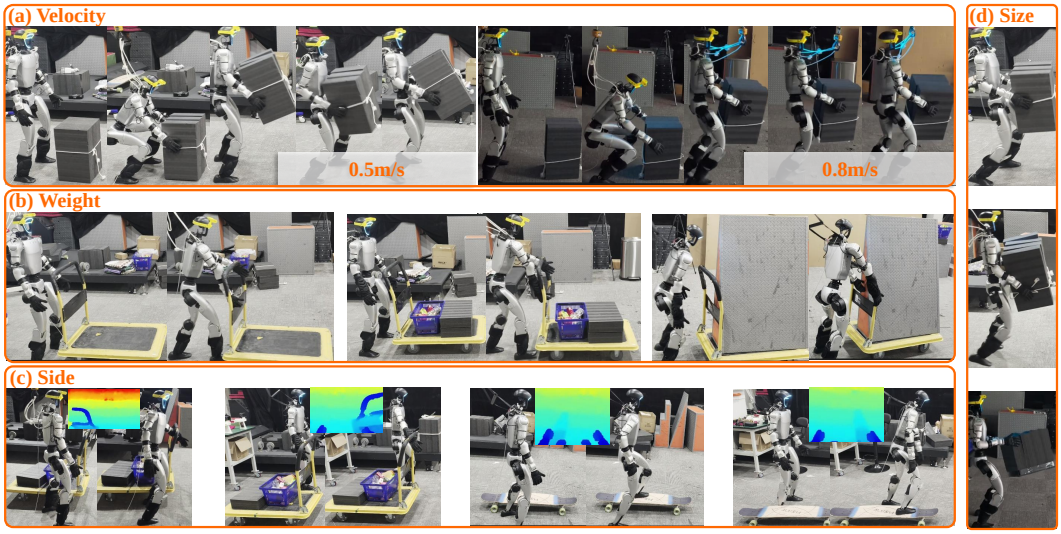

Figure 3: Real-world hardware generalization of VAIC to out-of-distribution object attributes and spatial conditions. The evaluations include (a) velocity command variations, (b) cart payload weight variations, (c) cart and skateboard position variations, and (d) box size variations. VAIC robustly adapts its whole-body posture to maintain stable interactions during physical deployment.

summarizes the evaluation on the Skateboard Interaction task involving high-speed impulsive contacts. By conditioning on the decoupled command interface rather than blindly tracking full-body reference kinematics, VAIC systematically outperforms all baselines. Pure kinematic trackers like PPO and AMP suffer catastrophic failure rates under dynamic disturbances due to their inability to implicitly estimate unobservable object states. In contrast, VAIC maintains superior balance and achieves the lowest tracking errors across almost all root and object metrics.

To further evaluate robustness against spatial initialization and object variations, we compare VAIC against PhysHSI and VisualMimic specifically on the Box Carry task. As detailed in Table 4, we evaluate across four configurations comprising combinations of two box sizes and two initial distances. While baseline interactive methods struggle with modified initializations, VAIC consistently achieves superior success rates, demonstrating a implicit understanding of coupled dynamics.

Table 4: Success rate comparison on the Box Carry task under varying initial conditions in MuJoCo. We evaluate physical robustness across combinations of two object sizes (Small, Large) and two initial interaction distances (Near, Far).

Method	Small-Near	Small-Far	Large-Near	Large-Far
PhysHSI	9/10	9/10	8/10	**9/10**
VisualMimic	-	-	**10/10**	8/10
VAIC	**10/10**	**10/10**	**10/10**	7/10

4.3 Generalization to Unseen Dynamics

While the aforementioned quantitative tracking evaluations were conducted in the MuJoCo simulator, the ultimate validation of our framework lies in physical deployment. A key advantage of replacing explicit kinematic tracking with intent-driven distillation is the emergence of robust physical adaptability to unmodeled variations on real hardware. As demonstrated in Figure 3, the deployed VAIC policy successfully handles out-of-distribution velocity commands, unmodeled payload weights, varying interaction sides, and different box sizes. The Object Adaptation module implicitly infers these altered physical relationships from history, allowing the humanoid robot to dynamically shift its center of gravity and adjust grasping postures in the real world.

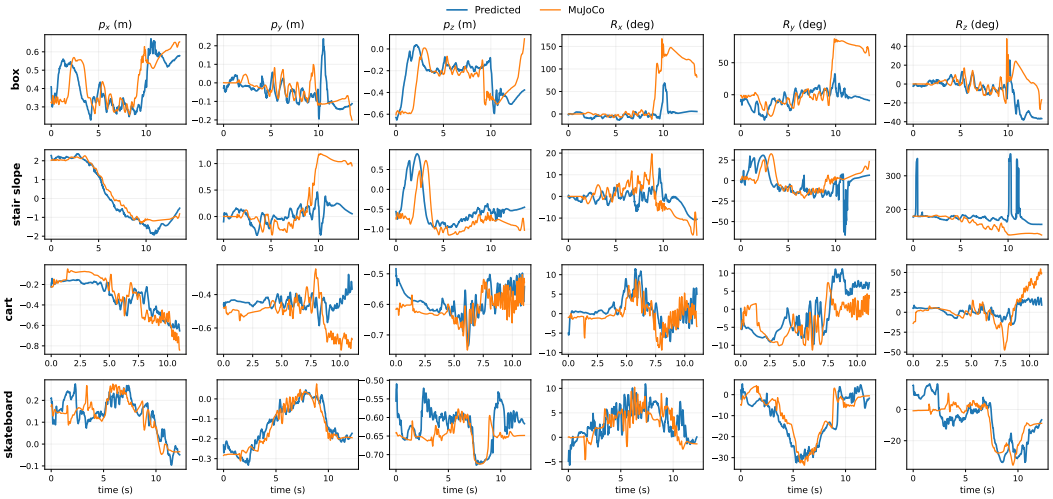

Figure 4: Comparison of the predicted object state from the VAIC adaptation module against MuJoCo ground truth. Conditioning the estimator on the binary interaction indicator c_t prevents representation collapse during high-speed dynamic phases.

4.4 Mechanism Analysis

We analyze the internal mechanism underlying this robustness by comparing the predicted object states $\hat{s}_t^{obj}$ from the Object Adaptation module of VAIC against the MuJoCo ground truth. As demonstrated in Figure 4, the depth-conditioned estimator tightly tracks the object's position and orientation even during severe self-occlusion (e.g., box carrying over slopes and stairs) and dynamic dismounts (e.g., pulling a cart and skateboarding). Crucially, conditioning the estimator on the binary interaction phase c_t significantly reduces temporal drift compared to purely proprioceptive alternatives.

5 Conclusion

In this paper, we presented VAIC (Vision-Guided Agile Interaction Control), a unified framework for humanoid whole-body loco-manipulation in unstructured environments. By introducing a high-level decoupled command interface and an asymmetric two-stage distillation paradigm, VAIC successfully eliminates the restrictive reliance on dense joint-level reference trajectories and perfect state observability during deployment. Facilitated by a recurrent Object Adaptation module, our policy effectively overcomes severe visual occlusion and implicitly predicts unmodeled coupled object dynamics. Extensive quantitative evaluations in simulation and real-world deployments on the humanoid robot across box carrying, cart manipulation, and skateboarding demonstrate the framework's superior balance, adaptability, and robust generalization, presenting a practical paradigm for autonomous humanoid deployment.

6 Limitation

While VAIC demonstrates robust generalization in physical deployments, several limitations remain. First, the depth-based Object Adaptation module is inherently susceptible to sensor noise and may fail when interacting with transparent or highly reflective objects. Second, although the recurrent architecture effectively infers unmodeled dynamics from proprioceptive history, extreme out-of-distribution physical variations (e.g., sudden and drastic mass shifts) can still induce representation collapse before the humanoid can adjust its balance. Finally, the decoupled command interface (v_t, c_t) is highly efficient for navigational intent and gross loco-manipulation, but it lacks the expressive granularity required for dexterous, fine-grained in-hand manipulation. Future work

will explore incorporating tactile feedback and richer semantic command spaces to address these constraints.

References

[1] H. Wang, W. Zhang, R. Yu, T. Huang, J. Ren, F. Jia, Z. Wang, X. Niu, X. Chen, J. Chen, et al. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.

[2] H. Weng, Y. Li, N. Sobanbabu, Z. Wang, Z. Luo, T. He, D. Ramanan, and G. Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025.

[3] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

[4] S. Yin, Y. Ze, H.-X. Yu, C. K. Liu, and J. Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

[5] D. Li, X. Chen, Q. Wu, B. Chen, S. Wu, H. Wu, G. Zhang, L. Li, M. Zhou, D. Xiang, J. Ma, Q. Zhang, and R. Xu. Haic: Humanoid agile object interaction control via dynamics-aware world model. *arXiv preprint arXiv:2602.11758*, 2026.

[6] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[7] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castañeda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[8] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20, 2021.

[9] Y. Lin, J. Shi, D. Wang, J. Kong, Y. Liu, C. Bai, and X. Li. Pro-hoi: Perceptive root-guided humanoid-object interaction. *arXiv preprint arXiv:2603.01126*, 2026.

[10] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[11] Y. Wang, Q. Zhao, Y. F. Lau, R. Yu, H. W. Tsui, Q. Chen, J. Wang, J. Pang, and P. Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. *arXiv preprint arXiv:2602.02473*, 2026.

[12] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

[13] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025.

[14] K. Yin, W. Zeng, K. Fan, M. Dai, Z. Wang, Q. Zhang, Z. Tian, J. Wang, J. Pang, and W. Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *arXiv preprint arXiv:2507.07356*, 2025.

[15] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

[16] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, 2025.

[17] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbabu, C. Pan, Z. Yi, G. Qu, K. Kitani, J. Hodgins, L. J. Fan, Y. Zhu, C. Liu, and G. Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[18] D. Rempe, M. Petrovich, Y. Yuan, H. Zhang, X. B. Peng, Y. Jiang, T. Wang, U. Iqbal, D. Minor, M. de Ruyter, et al. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026.

[19] W. Zeng, S. Lu, K. Yin, X. Niu, M. Dai, J. Wang, and J. Pang. Behavior foundation model for humanoid robots. *arXiv preprint arXiv:2509.13780*, 2025.

[20] M. Yuan, T. Yu, W. Ge, X. Yao, D. Li, H. Wang, J. Chen, X. Jin, B. Li, H. Chen, et al. Behavior foundation model: Towards next-generation whole-body control system of humanoid robots. *arXiv preprint arXiv:2506.20487*, 2025.

[21] Y. Li, Z. Luo, T. Zhang, C. Dai, A. Kanervisto, A. Tirinzoni, H. Weng, K. Kitani, M. Guzek, A. Touati, et al. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. *arXiv preprint arXiv:2511.04131*, 2025.

[22] N. Jiang, Z. He, W. Yu, L. Pang, Y. Li, H. Li, J. Cui, Y. Li, Y. Wang, Y. Zhu, et al. Uniact: Unified motion generation and action streaming for humanoid robots. *arXiv preprint arXiv:2512.24321*, 2025.

[23] Y. Shao, X. Huang, B. Zhang, Q. Liao, Y. Gao, Y. Chi, Z. Li, S. Shao, and K. Sreenath. Langwbc: Language-directed humanoid whole-body control via end-to-end learning. *arXiv preprint arXiv:2504.21738*, 2025.

[24] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2022.

[25] J. Zhang, H. Luo, H. Yang, X. Xu, Q. Wu, Y. Shi, J. Yu, L. Xu, and J. Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023.

[26] J. Lu, C.-H. P. Huang, U. Bhattacharya, Q. Huang, and Y. Zhou. Humoto: A 4d dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10886–10897, 2025.

[27] C. Zhao, J. Zhang, J. Du, Z. Shan, J. Wang, J. Yu, J. Wang, and L. Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–741, 2024.

[28] S. Xu, D. Li, Y. Zhang, X. Xu, Q. Long, Z. Wang, Y. Lu, S. Dong, H. Jiang, A. Gupta, Y.-X. Wang, and L.-Y. Gui. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[29] Y. Wang, J. Lin, A. Zeng, Z. Luo, J. Zhang, and L. Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

[30] S. Xu, H. Y. Ling, Y.-X. Wang, and L. Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[31] Y. Lin, Y. Xie, J. Xie, Y. Huang, R. Wang, J. Lv, Y. Ma, and X. Zuo. Simgenhoi: Physically realistic whole-body humanoid-object interaction via generative modeling and reinforcement learning. *arXiv preprint arXiv:2508.14120*, 2025.

[32] Q. Wu, Y. Shi, X. Huang, J. Yu, L. Xu, and J. Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024.

[33] L. Pan, Z. Yang, Z. Dou, W. Wang, B. Huang, B. Dai, T. Komura, and J. Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[34] J. Dao, H. Duan, and A. Fern. Sim-to-real learning for humanoid box loco-manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2024.

[35] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A.-a. Agha-mohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi. Falcon: Learning force-adaptive humanoid loco-manipulation. *arXiv preprint arXiv:2505.06776*, 2025.

[36] W. Sun, L. Feng, B. Cao, Y. Liu, Y. Jin, and Z. Xie. Ulc: A unified and fine-grained controller for humanoid loco-manipulation. *arXiv preprint arXiv:2507.06905*, 2025.

[37] L. Wei, X. Peng, R.-Z. Qiu, T. Huang, X. Cheng, and X. Wang. Hmc: Learning heterogeneous meta-control for contact-rich loco-manipulation. *arXiv preprint arXiv:2511.14756*, 2025.

[38] Z. Zhang, H. Lu, Y. Lian, Z. Chen, Y. Liu, C. Lin, H. Xue, Z. Zeng, Z. Qi, S. Zheng, et al. Learning athletic humanoid tennis skills from imperfect human motion data. *arXiv preprint arXiv:2603.12686*, 2026.

[39] Y. Chen, S. Dong, X. Ji, J. Sun, Z. Luo, L. Zhao, J. Zhang, W. Li, J. Ma, B. Xu, et al. Learning human-like badminton skills for humanoid robots. *arXiv preprint arXiv:2602.08370*, 2026.

[40] J. Kong, X. Liu, Y. Lin, J. Han, S. Schwertfeger, C. Bai, and X. Li. Learning soccer skills for humanoid robots: A progressive perception-action framework. *arXiv preprint arXiv:2602.05310*, 2026.

[41] J. Ren, Y. Li, K. Zhang, P. Fu, H. Jiang, Y. Pan, G. Zeng, T. Huang, W. Guo, P. Lu, et al. Smash: Mastering scalable whole-body skills for humanoid ping-pong with egocentric vision. *arXiv preprint arXiv:2604.01158*, 2026.

[42] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2025.

[43] T. Wu, X. Kong, Y. Chen, Q. Yu, H. Ye, J. Li, Y. Wang, and H. Dong. Sugar: A scalable human-video-driven generalizable humanoid loco-manipulation learning framework. *arXiv preprint arXiv:2605.20373*, 2026.

[44] A. Agarwal, A. Kumar, J. Malik, and D. Pathak. Legged locomotion in challenging terrains using egocentric vision. In *Conference on Robot Learning (CoRL)*, 2022.

[45] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*, 2024.

[46] J. Sun, G. Han, P. Sun, W. Zhao, J. Cao, J. Wang, Y. Guo, and Q. Zhang. Dpl: Depth-only perceptive humanoid locomotion via realistic depth synthesis and cross-attention terrain reconstruction. *arXiv preprint arXiv:2510.07152*, 2025.

[47] C. Han, S. He, Y. Cheng, L. Ye, and H. Liu. Prior: Perceptive learning for humanoid locomotion with reference gait priors. *arXiv preprint arXiv:2603.18979*, 2026.

[48] Y. Zhang, Y. Seo, J. Chen, Y. Yuan, K. Sreenath, P. Abbeel, C. Sferrazza, K. Liu, R. Duan, and G. Shi. Rpl: Learning robust humanoid perceptive locomotion on challenging terrains. *arXiv preprint arXiv:2602.03002*, 2026.

[49] W. Sun, Y. Su, L. Huang, A. Zhang, D. Wei, M. San, D. Tian, E. Cao, B. Cao, Y. Liu, et al. Now you see that: Learning end-to-end humanoid locomotion from raw pixels. *arXiv preprint arXiv:2602.06382*, 2026.

[50] Z. Zhuang, S. Yao, and H. Zhao. Humanoid parkour learning. In *Conference on Robot Learning (CoRL)*, 2024.

[51] Z. Wu, X. Huang, L. Yang, Y. Zhang, K. Sreenath, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferrazza, et al. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. *arXiv preprint arXiv:2602.15827*, 2026.

[52] Z. Zhuang, S. Zhu, M. Zhao, and H. Zhao. Deep whole-body parkour. *arXiv preprint arXiv:2601.07701*, 2026.

[53] S. Zhu, B. Ye, J. Wang, J. Chen, Z. Zhuang, L. Mou, R. Huang, and H. Zhao. Ttt-parkour: Rapid test-time training for perceptive robot parkour. *arXiv preprint arXiv:2602.02331*, 2026.

[54] H. Xue, X. Huang, D. Niu, Q. Liao, T. Kragerud, J. T. Gravdahl, X. B. Peng, G. Shi, T. Darrell, K. Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

[55] T. He, Z. Wang, H. Xue, Q. Ben, Z. Luo, W. Xiao, Y. Yuan, X. Da, F. Castañeda, S. Sastry, et al. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025.

[56] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025.

[57] H. Liu, Y. Gao, S. Teng, Y. Chi, Y. S. Shao, Z. Li, M. Ghaffari, and K. Sreenath. Ego-vision world model for humanoid contact planning. In *International Conference on Robotics and Automation (ICRA)*, 2026.

[58] Y. Lin, J. Cui, Y. Li, B. Jia, Y. Zhu, and S. Huang. Lessmimic: Long-horizon humanoid interaction with unified distance field representations. *arXiv preprint arXiv:2602.21723*, 2026.

[59] X. He, S. Xu, X. Li, R. Dong, L. Bian, Y.-X. Wang, and L.-Y. Gui. Ultra: Unified multimodal control for autonomous humanoid whole-body loco-manipulation. *arXiv preprint arXiv:2603.03279*, 2026.

[60] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

VAIC: Vision-Guided Humanoid Agile Object Interaction Control via Decoupled Commands

Appendix

In this appendix, we provide additional experimental setups and details:

1. **Demo Video.** A demonstration video including the real-world experiment is provided, as described in Sec. A.
2. **System Architecture and Parameters.** The observation and action spaces, alongside network architectures, are detailed in Sec. B.
3. **Experimental Details.** The reward formulation, training hyperparameters, real-world deployment setups, and mathematical definitions of our quantitative metrics are provided in Sec. C.
4. **Data Preparation and Nominal Templates.** Details regarding motion retargeting, object template generation, and the extraction of unified decoupled commands are presented in Sec. D.
5. **Ablation Study.** Additional ablation experiments analyzing the contribution of individual perceptive and temporal components are provided in Sec. E.

A Demo Video

In addition to the qualitative results presented in the main paper, we provide a supplementary video (`https://vaic-humanoid.github.io/static/videos/top.mp4`) for more detailed visualizations. The video comprehensively illustrates the efficacy of our approach, particularly highlighting robust real-world hardware deployment, real-time recording of depth-camera inputs, and the policy's generalization across various scenarios, object geometries, and weights.

B System Architecture and Parameters

B.1 Observation Space and Depth Processing

Our asymmetric two-stage distillation framework strictly divides the observation space. During Stage 1, the Teacher policy utilizes privileged ground-truth state information from the simulation (detailed in Table A) to master complex contact mechanics. Crucially, to ensure the teacher focuses entirely on mastering the underlying physical dynamics without being distracted by visual complexities, the depth camera is strictly disabled during this initial phase. To maximize computational efficiency during this stage, the visual input is zero-injected, meaning the depth tensor is structurally maintained but populated entirely with zeros. Beyond preventing early visual overfitting and bypassing costly rendering overhead, this strategy guarantees dimensional consistency across the observation spaces of both stages. This structural alignment is essential, as it allows the student policy to seamlessly inherit the teacher's network architecture and pre-trained weights, ensuring a smooth and stable distillation process in Stage 2.

The visual stream is only activated during Stage 2. Here, the Student policy replaces the dense privileged tracking inputs with a deployable observation space $o_t \in \mathcal{O}$, forcing the policy to ground its physical interactions in real-time perception. This deployable space strictly relies on:

- **Proprioceptive State (s_t^{prop}):** A 5-step history of joint positions, joint velocities, base angular velocity, and the projected gravity vector.
- **Decoupled Command (v_t, c_t):** A multi-axis velocity target $v_t \in \mathbb{R}^3$ and a binary interaction phase indicator $c_t \in \{0, 1\}$. To reduce the continuous observation space and mitigate policy sensitivity to minor joystick tremors, the velocity commands are quantized with a resolution of 0.2 units.
- **Exteroceptive Depth (d_t):** A temporal stream of downsampled depth images (64×36) captured by the onboard camera at 60 Hz.

Table A: Deployable Observation Space and Privileged Information for the VAIC framework.

Category	Observation Term	Noise (σ)	Description
Student	Proprioceptive History	0.015	History of joint positions, velocities, and gravity vector.
	Base Angular Velocity	0.05	Base angular velocity in the local frame (steps: [0]).
	Previous Actions	-	Joint target history from the previous 3 steps.
	Decoupled Command	-	Velocity target $\boldsymbol{v}_t$ and binary interaction phase indicator c_t.
	Depth Stream	-	Sequential depth images d_t processed by the CNN-GRU encoder.
	Obj. Geometric Prior	-	Canonical point cloud template $\mathcal{P}$ of the interactive object(s).
Teacher	Clean Proprioception	0.0	Noise-free history of all proprioceptive states.
	Root & Body Velocity	0.0	Linear velocity of the robot base and key bodies.
	Ref. Motion Tracking	0.0	Future reference root states, joint positions, and motion phase ϕ.
	Obj. Kinematic State	0.01	Exact relative pose and velocity of the interactive object(s).
	Obj. Geometric Prior	-	Canonical point cloud template $\mathcal{P}$ of the interactive object(s).
	Ref. Contact Pos.	0.01	Exact future reference contact positions on the object.
	Applied Forces	-	Real applied actions and torques.
	Dynamics Randomization	-	Mass, friction, restitution, scale, joint armature, and damping.

Depth Processing and Sim-to-Real Pipeline. To ensure robust physical deployment, we implement a rigorous depth processing pipeline that addresses sim-to-real gaps at both the sensor and signal levels. First, to simulate an Intel RealSense D435i camera mounted on the robot's torso (pitched downward at 48°), we inject extrinsic noise to the camera pose ($\sigma_{\text{pos}} = 0.01$ m, $\sigma_{\text{rot}} \approx 1^\circ$) to model installation errors, and intrinsic noise to the focal length (2%) and aperture (5%) via dynamic ray-casting resampling. Second, the raw simulated depth undergoes a sequence of signal-level augmentations: (1) *Random Frame Delay* of 0 to 4 steps mimicking hardware latency; (2) *Measurement Noise* via pixel-wise Gaussian perturbation ($\sigma = 0.05$ m); (3) *Episodic Bias Drift* $\mathcal{U}(-0.15, 0.15)$ m accounting for thermal drift; (4) *Pixel Dropout* ($p = 0.1$) simulating reflective artifacts; and (5) *Depth Quantization* into 100 discrete bins (equivalent to $\sim$2.9 cm accuracy) after clamping to a $[0.1, 3.0]$ m range and normalizing.

B.2 Action Space and Low-level Control

The policy outputs actions $a_t \in \mathbb{R}^{23}$ representing target joint position residuals Δq, such that the desired joint positions are $q_{\text{target}} = q_t + \Delta q$. Note that this dimension covers the whole-body joints excluding the 6 wrist degrees of freedom (3 per hand), as they are strictly maintained or passively compliant during the macroscopic interaction tasks. These targets are dispatched to a low-level proportional-derivative (PD) controller to compute the final joint torques τ:

$$\tau = k_p(q_{\text{target}} - q_t) - k_d\dot{q}_t \tag{4}$$

where k_p and k_d are the joint-specific stiffness and damping gains. This residual-style position control ensures compliant and stable physical interaction.

B.3 Network Architecture

Our framework employs a dual-recurrent architecture to systematically decouple visual feature extraction from privileged state inference. The network consists of three core components:

Temporal Depth Encoder: A CNN-GRU module processes the augmented depth stream $d_t \in \mathbb{R}^{1\times 36\times 64}$. The spatial features are extracted via a 3-layer Convolutional Neural Network (kernel size 5, stride 2, Mish activation), flattened, and compressed via a linear layer to a 64-D spatial embedding. A subsequent Gated Recurrent Unit (GRU, hidden dimension 64) accumulates these features temporally without burn-in, outputting a 64-D temporal depth feature $\boldsymbol{z}_t^{\text{depth}}$ followed by LayerNorm. This feature is strictly constrained for object state estimation and is not fed directly to the policy actor.

Object Adapter: This module explicitly predicts the interactive object's pose to decouple geometric perception from action generation. It first processes the proprioceptive observation and decoupled commands through an MLP to generate a 256-D embedding. This is concatenated with the 64-D temporal depth feature $\boldsymbol{z}_t^{\text{depth}}$ and passed through a subsequent MLP (hidden dimensions $[256, 256]$) to output the predicted object state $\hat{\boldsymbol{s}}_t^{\text{obj}}$. This branch is trained via supervised learning against the ground-truth object states from the simulation.

Privilege Adapter: To reconstruct the teacher's latent representation, this module fuses the proprioception, decoupled commands, and the geometrically transformed object point cloud feature $\hat{\boldsymbol{s}}_t^{\text{geo}}$. The inputs are pre-processed by an MLP and fed into a core GRU module (hidden dimension 256). To prevent gradient explosion and maintain temporal stability during long-horizon interactions, a burn-in strategy is applied where gradients for the first $T/4$ steps of the sequence are detached. The output utilizes a residual skip connection (summing the GRU output and the MLP pre-processing output) before a final linear layer yields the 256-D privileged latent $\hat{\boldsymbol{z}}_t^{\text{priv}}$ to drive the student actor.

C Experimental Details

C.1 Training Details

The complete reward formulation for the PPO training is detailed in Table B. It is constructed to balance precise human-motion imitation with physically viable foot constraints and interaction objectives. During the student fine-tuning phase, the dense motion tracking rewards are relaxed, and the policy prioritizes task-completion objectives aligned with the decoupled velocity commands.

The networks are optimized using Proximal Policy Optimization (PPO) with Generalized Advantage Estimation (GAE). Table C details the specific network dimensions, learning rates, and loss coefficients used during training.

C.2 Real-World Deployment and Sim-to-Real

We deploy the trained policies directly onto the physical humanoid robot. The inference pipeline operates onboard at 50 Hz, processing synchronized depth frames from the Intel RealSense D435i

Table B: Reward Functions for the VAIC Framework. During student exploration, the policy prioritizes decoupled command tracking and phase-conditioned interaction, while dense kinematic terms are relaxed into stylistic motion priors.

Term	Expression	Weight	Description
(a) Decoupled Command Tracking			
Linear Velocity	$\exp(-\|\boldsymbol{v}_{xy}^{\text{root}} - \boldsymbol{v}_{xy}^{\text{cmd}}\|^2/\sigma_v)$	1.0	Tracks the velocity command.
Yaw Velocity	$\exp(-\|\boldsymbol{\omega}_z^{\text{root}} - \boldsymbol{\omega}_z^{\text{cmd}}\|^2/\sigma_\omega)$	1.0	Tracks the yaw angular velocity command.
(b) Object Interaction Reward			
Object Position	$\exp(-\|\boldsymbol{p}_{\text{obj}} - \boldsymbol{p}_{\text{obj}}^{\text{ref}}\|^2/\sigma_p)$	1.0	Tracks global position of the interactive object(s).
Object Orientation	$\exp(-\|\boldsymbol{\theta}_{\text{obj}} \ominus \boldsymbol{\theta}_{\text{obj}}^{\text{ref}}\|^2/\sigma_\theta)$	1.0	Tracks global orientation of the interactive object(s).
Phase Contact	$c_t \cdot \mathbb{E}_{o,\,e}\left[\mathbb{I}_{o,e} \cdot r_{\text{pos}}^{o,e} \cdot r_{\text{force}}^{o,e}\right]$	1.0	Active contact is gated by interaction indicator c_t.
(c) Motion Imitation Prior (Relaxed)			
Joint Position	$\exp(-\|\boldsymbol{q} - \boldsymbol{q}^{\text{ref}}\|^2/\sigma_q)$	0.3	Acts as a stylistic motion prior.
Joint Velocity	$\exp(-\|\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}^{\text{ref}}\|^2/\sigma_{\dot{q}})$	0.3	Acts as a stylistic motion prior.
Upper Body Pos.	$\exp(-\|\boldsymbol{p}_{\text{up}} - \boldsymbol{p}_{\text{up}}^{\text{ref}}\|^2/\sigma_p)$	0.5	Soft tracking for arm posture.
Lower Body Pos.	$\exp(-\|\boldsymbol{p}_{\text{low}} - \boldsymbol{p}_{\text{low}}^{\text{ref}}\|^2/\sigma_p)$	0.5	Soft tracking for lower body.
Root Position	$\exp(-\|\boldsymbol{p}_{\text{root}} - \boldsymbol{p}_{\text{root}}^{\text{ref}}\|^2/\sigma_p)$	0.5	Prevent root position divergence.
Root Orientation	$\exp(-\|\boldsymbol{\theta}_{\text{root}} \ominus \boldsymbol{\theta}_{\text{root}}^{\text{ref}}\|^2/\sigma_\theta)$	0.5	Prevent root orientation divergence.
(d) Foot Constraints			
Feet Air Time	$\exp(\text{clip}(t_{\text{air}} - t_{\text{thr}})/\sigma) \cdot \mathbb{I}_{\text{step}}$	0.5	Encourages longer and more stable swing phases.
Feet Slip	$-\|\boldsymbol{v}_{\text{foot}}^{xy}\| \cdot \mathbb{I}_{\text{ground}}$	0.5	Penalizes sliding velocity when feet are grounded.
Feet Contact	$\exp(-\|\mathbb{I}_{\text{con}}^{\text{real}} - \mathbb{I}_{\text{con}}^{\text{ref}}\|^2/\sigma)$	0.5	Matches reference rhythmic contact states.
Feet Air Lift	$-\sum(h_{\text{foot}} < h_{\min}) \cdot \mathbb{I}_{\text{swing}}$	0.5	Penalizes tripping or dragging.
Impact Force	$-\|\boldsymbol{F}_{\text{impact}}\|^2$	1.0	Penalizes large and unsafe impact forces.
(e) Regularization			
Action Rate	$-\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|^2$	0.1	Penalizes rapid action changes to ensure smoothness.
Joint Velocity L2	$-\|\dot{\boldsymbol{q}}\|^2$	5e-4	Penalizes excessively high joint velocities (energy).
Joint Limits	$-\sum \text{clip}(\boldsymbol{q} - \boldsymbol{q}_{\text{limit}}, 0, \infty)$	10.0	Penalizes exceeding physical joint limits.
Torque Limits	$-\sum \text{clip}(\boldsymbol{\tau} - \boldsymbol{\tau}_{\text{limit}}, 0, \infty)$	0.01	Penalizes motor torque saturation.
Survival	1.0	1.0	Constant reward for avoiding early termination.

camera (captured at 60 Hz). The resulting joint position targets are dispatched to a low-level PD controller operating at 200 Hz. The specific stiffness (k_p) and damping (k_d) parameters configured for the humanoid robot hardware exactly follow the settings detailed in HAIC [5].

To bridge the sim-to-real gap, we subject the Isaac Sim environment to extensive domain randomization during training. Table D outlines the full distribution of randomized physical properties, including external push forces, altered mass scales, and modified joint friction, which ensure the policy remains robust to unmodeled physical perturbations during physical deployment.

Table C: Hyperparameters for Proximal Policy Optimization (PPO) and network architectures. The asymmetric Actor-Critic framework utilizes distinct loss coefficients to effectively balance relaxed motion imitation, privileged state distillation, and decoupled command tracking.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Optimizer	Adam	Number of Environments	4096
Rollout Steps (Horizon)	32	Mini-batches	8
Learning Epochs	3	Discount Factor (γ)	0.99
GAE Parameter (λ)	0.95	Clip Parameter (ϵ)	0.2
Entropy Coefficient	0.001	Max Gradient Norm	1.0
Desired KL	0.01	Learning Rate	3×10^{-4}
Initial Noise Std	1.0		
Loss Coefficients			
Value Loss (λ_{value})	1.0	Object Loss (λ_{obj})	1.0
Privileged Loss (λ_{priv})	1.0	Distillation Loss (λ_{distill})	1.0
Network Architecture			
Actor MLP Size	$[512, 256, 256]$	Critic MLP Size	$[512, 256, 128]$
Depth CNN Layers	3 Conv2d	Object Adapter MLP	$[256, 256]$
Depth GRU Hidden Dim	64	Privilege GRU Hidden Dim	256
Actor/Critic Activation	ELU	Adapter/CNN Activation	Mish

C.3 Evaluation Metrics

To comprehensively evaluate the interaction performance, we define two categories of metrics tracking the errors of both the humanoid root and the interactive object in the simulation environment.

Root State Metrics:

- **Root Position Error** (E_{rpe}, m): The Euclidean distance between the simulated and reference root positions.

$$E_{\text{rpe}} = \mathbb{E}\left[\left\|\boldsymbol{p}_t^{\text{root}} - \boldsymbol{p}_t^{\text{root,ref}}\right\|_2\right] \tag{5}$$

- **Root Orientation Error** (E_{roe}, rad): The angular distance between the simulated and reference root quaternions.

$$E_{\text{roe}} = \mathbb{E}\left[2\arccos\left(\left|\langle \boldsymbol{q}_t^{\text{root}}, \boldsymbol{q}_t^{\text{root,ref}}\rangle\right|\right)\right] \tag{6}$$

- **Root Linear Velocity Error** (E_{rve}, m/s): The Euclidean error of the root's linear velocity.

$$E_{\text{rve}} = \mathbb{E}\left[\left\|\boldsymbol{v}_t^{\text{root}} - \boldsymbol{v}_t^{\text{root,ref}}\right\|_2\right] \tag{7}$$

- **Root Angular Velocity Error** (E_{rave}, rad/s): The Euclidean error of the root's angular velocity vector.

$$E_{\text{rave}} = \mathbb{E}\left[\left\|\boldsymbol{\omega}_t^{\text{root}} - \boldsymbol{\omega}_t^{\text{root,ref}}\right\|_2\right] \tag{8}$$

- **Root Linear Acceleration Error** (E_{rae}, m/s^2): The Euclidean error of the root's linear acceleration.

$$E_{\text{rae}} = \mathbb{E}\left[\left\|\boldsymbol{a}_t^{\text{root}} - \boldsymbol{a}_t^{\text{root,ref}}\right\|_2\right] \tag{9}$$

Object State Metrics:

Table D: Domain Randomization Parameters for Robot and Objects.

Category	Parameter	Range / Distribution
Robot Dynamics		
Properties	Link Mass Scale	$\mathcal{U}(0.9, 1.1) \times$ default
	Center of Mass Offset	$\mathcal{U}(-0.02, 0.02)$ m
	Static Friction	$\mathcal{U}(0.3, 1.6)$
	Dynamic Friction	$\mathcal{U}(0.3, 1.2)$
Actuation	Joint Position Offset	$\mathcal{U}(-0.01, 0.01)$ rad
	Motor Stiffness Scale	$\mathcal{U}(0.9, 1.1)$
	Motor Damping Scale	$\mathcal{U}(0.9, 1.1)$
	Action Delay	$\mathcal{U}[40, 120]$ ms
Object Interaction		
Surface	Dynamic Friction	$\mathcal{U}(0.3, 0.8)$
	Static-to-Dynamic Ratio	$\mathcal{U}(1.0, 2.0)$
	Restitution	$\mathcal{U}(0.0, 0.2)$
Box	Mass	$\mathcal{U}(1.0, 2.0)$ kg
	Scale	$\mathcal{U}(0.5, 1.2)$
Cart	Body Mass	$\mathcal{U}(11.0, 13.0)$ kg
	Wheel Mass	$\mathcal{U}(0.2, 0.4)$ kg
	Wheel Joint Friction	$\mathcal{U}(0.01, 0.1)$ N·m
	Wheel Joint Damping	$\mathcal{U}(0.01, 0.1)$ N·m·s/rad
	Scale	$\mathcal{U}(0.9, 1.1)$
Skateboard	Body Mass	$\mathcal{U}(2.0, 5.0)$ kg
	Wheel Mass	$\mathcal{U}(0.1, 0.2)$ kg
	Wheel Armature	$\mathcal{U}(0.0,$ 1e-4$)$ kg·m^2
	Wheel Joint Damping	$\mathcal{U}(0.0,$ 1e-3$)$ N·m·s/rad
	Scale	$\mathcal{U}(0.8, 1.1)$
Slope / Stair	Scale	$\mathcal{U}(0.8, 1.2)$
External Perturbation		
Push	Push Force	$\mathcal{U}(0.2, 0.5) \times$ weight
	Push Min Interval	2s

- **Object Position Error** (E_{ope}, m): The Euclidean position error of the interactive object in global coordinates.

$$E_{\text{ope}} = \mathbb{E}\left[\left\|\boldsymbol{p}_t^{\text{obj}} - \boldsymbol{p}_t^{\text{obj,ref}}\right\|_2\right] \tag{10}$$

- **Object Orientation Error** (E_{ooe}, rad): The angular error of the object's orientation quaternion.

$$E_{\text{ooe}} = \mathbb{E}\left[2\arccos\left(\left|\langle \boldsymbol{q}_t^{\text{obj}}, \boldsymbol{q}_t^{\text{obj,ref}}\rangle\right|\right)\right] \tag{11}$$

- **Object Linear Velocity Error** (E_{ove}, m/s): The Euclidean error of the object's linear velocity.

$$E_{\text{ove}} = \mathbb{E}\left[\left\|\boldsymbol{v}_t^{\text{obj}} - \boldsymbol{v}_t^{\text{obj,ref}}\right\|_2\right] \tag{12}$$

- **Object Angular Velocity Error** (E_{oave}, rad/s): The Euclidean error of the object's angular velocity.

$$E_{\text{oave}} = \mathbb{E}\left[\left\|\boldsymbol{\omega}_t^{\text{obj}} - \boldsymbol{\omega}_t^{\text{obj,ref}}\right\|_2\right] \tag{13}$$

- **Object Linear Acceleration Error** (E_{oae}, m/s^2): The Euclidean error of the object's linear acceleration.

$$E_{\text{oae}} = \mathbb{E}\left[\left\|\boldsymbol{a}_t^{\text{obj}} - \boldsymbol{a}_t^{\text{obj,ref}}\right\|_2\right] \tag{14}$$

D Data Preparation and Nominal Templates

To generate high-fidelity physical states from raw human demonstrations, we utilize an optical motion capture system followed by a rigorous retargeting pipeline. First, the human skeleton kinematics are mapped to the humanoid robot's 29-DoF structure using the PoseLib framework. Second, an Isaac Gym digital twin is established where the robot and objects are kinematically driven to extract physically valid contact labels and generate a unified $(n+m)$-body state array.

We systematically extract unified decoupled commands directly from these reference motion trajectories. This extraction provides a dual benefit: it not only serves as the high-level task objective during training, but also establishes a highly standardized evaluation protocol. By conditioning the policy on these extracted commands during automated inference, we can record the generated kinematic trajectories and directly compute the tracking errors against the original reference motions, thereby facilitating rigorous, consistent, and reproducible benchmarking across all baselines.

Crucially for the deployable student policy, we extract a nominal point cloud template $\mathcal{P}$ for each interactive object. During Stage 2 exploration, the Privilege Adapter utilizes the implicitly inferred object state $\hat{s}_t^{\text{obj}}$ to project this canonical template $\mathcal{P}$ into the robot's local spatial frame via $\hat{s}_t^{\text{geo}} = \mathcal{T}(\hat{s}_t^{\text{obj}}, \mathcal{P})$, thereby reconstructing the dense spatial priors necessary for agile interaction without relying on privileged external tracking.

E Ablation Study

To evaluate the contribution of individual components within the VAIC framework, we conduct an ablation study on the high-speed Skateboard Interaction task, in which the skateboard is initialized on either side of the humanoid robot. As shown in Table E, we progressively remove key perceptive and temporal modules to assess their impact on dynamic balance and object state estimation.

- **w/o Depth:** Removing the exteroceptive depth input forces the policy to rely entirely on proprioception. This results in a catastrophic drop in success rate, as the robot cannot anticipate the initial contact side or spatial geometry of the impulsive contact.
- **w/o Object Adapter:** Disabling the implicit state inference module prevents the student policy from distilling the teacher's privileged geometric understanding, leading to significant divergence in object position (E_{ope}) and velocity (E_{ove}) tracking.
- **w/o GRU:** Removing the temporal recurrence degrades the policy's ability to maintain a consistent belief state of the object's latent dynamics during phase transitions, increasing root position error (E_{rpe}).

Table E: Quantitative ablation evaluation on **Skateboard Interaction**.

		Root State Metrics (↓)					**Object State Metrics (↓)**				
Method	**SR ↑**	E_{rpe}	E_{roe}	E_{rve}	E_{rave}	E_{rae}	E_{ope}	E_{ooe}	E_{ove}	E_{oave}	E_{oae}
w/o Depth	8.3%	$1.58^{\pm 0.41}$	$1.07^{\pm 0.17}$	$0.44^{\pm 0.07}$	$0.39^{\pm 0.05}$	$0.42^{\pm 0.03}$	$1.90^{\pm 0.27}$	$0.61^{\pm 0.91}$	$0.40^{\pm 0.04}$	$0.09^{\pm 0.12}$	$\mathbf{0.26}^{\pm 0.03}$
w/o Object Adapter	0.0%	$2.08^{\pm 0.65}$	$1.26^{\pm 0.19}$	$0.50^{\pm 0.09}$	$0.36^{\pm 0.02}$	$0.39^{\pm 0.02}$	$1.83^{\pm 0.14}$	$0.31^{\pm 0.51}$	$0.40^{\pm 0.03}$	$0.07^{\pm 0.06}$	$\mathbf{0.26}^{\pm 0.02}$
w/o GRU	33.3%	$1.35^{\pm 0.66}$	$0.95^{\pm 0.45}$	$0.38^{\pm 0.13}$	$0.35^{\pm 0.05}$	$0.37^{\pm 0.09}$	$1.46^{\pm 0.75}$	$0.74^{\pm 0.85}$	$0.36^{\pm 0.12}$	$0.11^{\pm 0.09}$	$0.27^{\pm 0.06}$
VAIC	**83.3%**	$\mathbf{0.62}^{\pm 0.12}$	$\mathbf{0.38}^{\pm 0.19}$	$\mathbf{0.29}^{\pm 0.05}$	$\mathbf{0.27}^{\pm 0.07}$	$\mathbf{0.35}^{\pm 0.07}$	$\mathbf{0.62}^{\pm 0.17}$	$\mathbf{0.07}^{\pm 0.06}$	$\mathbf{0.31}^{\pm 0.10}$	$\mathbf{0.02}^{\pm 0.02}$	$0.27^{\pm 0.04}$