

Language-Guided Humanoid Loco-Manipulation via Vision-Language-Action Models

Vardhan Dongre^{*}, Sunny Deshpande^{*}, Het Patel^{*}

Intro to Humanoid Robotics ECE598JK

University of Illinois Urbana-Champaign

Urbana, IL, USA

{vdongre2, sunnynd2, hcp4}@illinois.edu

Abstract—Vision-Language-Action (VLA) models have emerged as a promising paradigm for enabling humanoid robots to perform complex manipulation tasks by grounding natural language instructions in visual perception. In this work, we investigate the impact of replacing these disparate encoders with a unified Contrastive Language-Image Pre-training (CLIP) model within the HumanVLA architecture. We employ a Teacher-Student training pipeline where a privileged teacher policy, trained via reinforcement learning, supervises a student policy that operates solely on egocentric vision and language instructions. We evaluate our approach on the Human-in-the-Room (HITR) dataset using the Isaac Gym simulation environment. Our experiments demonstrate that utilizing a frozen CLIP encoder improves the student policy’s generalization to unseen rearrangement tasks, achieving a success rate of 62.1% compared to the 60.2% baseline. Conversely, we find that fine-tuning the CLIP encoders within this imitation learning framework leads to optimization instability and performance collapse, highlighting the effectiveness of leveraging robust, pre-trained multimodal representations for humanoid loco-manipulation.

I. INTRODUCTION

Recent progress in embodied artificial intelligence has enabled agents to perform increasingly complex manipulation and rearrangement tasks by grounding natural language instructions in visual perception and motor control. In particular, vision–language–action (VLA) models have emerged as a powerful paradigm for humanoid control, allowing agents to interpret free-form instructions, perceive their environment through egocentric sensors, and execute coordinated whole-body actions. Despite these advances, a central challenge remains: learning robust and generalizable policies from limited demonstrations without access to privileged state information at test time.

HumanVLA addresses this challenge through a teacher–student framework, where a privileged teacher policy is trained using reinforcement learning with full state access, and a student policy learns to imitate the teacher using only partial observations, including egocentric images, language instructions, proprioception, and previous actions. This approach achieves strong performance on long-horizon humanoid rearrangement tasks, demonstrating that high-

quality visuomotor policies can be distilled from expert teachers into deployable student agents.

While the HumanVLA framework establishes an effective training pipeline, its student model relies on separate, frozen perception modules: an ImageNet-pretrained EfficientNet for vision and a BERT-based encoder for language. These encoders are trained independently and lack explicit alignment between visual and linguistic representations. As a result, the student policy must implicitly learn to fuse heterogeneous embeddings into a coherent task representation, which may limit its ability to generalize to unseen object configurations, instructions, or visual conditions.

In this project, we investigate whether joint vision–language representations can improve the performance and generalization of HumanVLA-style student policies. Specifically, we replace the original vision and language encoders with a CLIP-based model that embeds images and text into a shared multimodal space. CLIP has been shown to provide robust cross-modal alignment across diverse visual and linguistic concepts, making it a promising candidate for instruction-conditioned control. Beyond encoder choice, we also examine the role of fine-tuning versus freezing multimodal representations during imitation learning, assessing whether limited domain adaptation of the visual encoder can further benefit humanoid control.

We conduct a systematic study comparing the original HumanVLA student with multiple CLIP-based variants, including frozen and partially fine-tuned encoders, while keeping the teacher, training procedure, and evaluation protocol unchanged. Our evaluation focuses on success rate, final placement precision, and execution time on both training and held-out rearrangement tasks in the HITR environment. Through these experiments, we aim to better understand how multimodal representation learning choices impact the downstream performance of vision–language–action policies for embodied manipulation.

II. RELATED WORK

a) Vision-Language-Action (VLA) Models.: The recent wave of Vision–Language–Action (VLA) models has significantly advanced the integration of perception, language

^{*}These authors contributed equally to this work.

understanding, and control for embodied agents. Models such as GR00T [2], HumanVLA [3], RT-1 [4], RT-2 [5] PaLM-E [6], OpenVLA [7], SmolVLA [8] demonstrate that large-scale multimodal pretraining can transfer web knowledge to physical manipulation domains, enabling robots to execute high-level natural language commands like “pick up the sponge and clean the spill”. These models couple transformer-based encoders with action decoders trained on paired image–text–action data, grounding linguistic concepts to motor policies. However, their generalization to unseen object configurations, multi-step reasoning tasks, and long-horizon planning remains limited, particularly when extended to humanoid morphologies with full-body actuation.

b) Simulation Environments.: The Human-In-The-Room (HITR) [2] dataset provides a specialized benchmark for humanoid loco-manipulation, featuring diverse room environments—such as bedrooms, living rooms, warehouses, and kitchens which are populated with interactive objects. While broader frameworks like the BEHAVIOR-1K benchmark [9] and OmniGibson offer extensive object libraries and scenes for general embodied AI, HITR specifically targets the unique challenges of whole-body control and language grounding in human-centric spaces.

Our project builds on these lines of work by shifting the focus from policy training to systematic evaluation. Rather than proposing new architectures or benchmarks, we investigate how existing humanoid VLA models reason about semantics, adapt to unseen configurations, and maintain task consistency when facing visual clutter and ambiguous instructions. This perspective complements current VLA development by identifying where reasoning failures occur

c) Learning in Humanoid Robots.: Currently, learning-based loco-manipulation is heavily reliant on simulation with many components of contact being simplified [10]. When applying to the real world, strategies such as domain randomization, reward shaping, or reference-based guidance are used to handle contact uncertainties. Imitation Learning (IL) is also used currently to collect real autonomous humanoid data. Various skills such as fall recovery and kicking are handled typically by Hierarchical Reinforcement Learning (RL), but with limited scalability [10].

Research on creating a layered learning architecture is being done, with four layers; Human Layer for learning human-like skills from IL, Hybrid Skill Layer for data-model fusion (IL, RL, and Model-Based Control), Representation Layer for learning structured latent skill spaces, and Foundation Model for VLA-capabilities on the humanoid robot [10].

III. METHOD

A. Teacher–Student Framework

We adopt a teacher–student learning framework to train humanoid vision–language–action policies for long-horizon rearrangement tasks. A privileged teacher policy is first trained using reinforcement learning with access to full environment state, including object poses and task-relevant metadata unavailable at test time. The teacher operates in the HITR

humanoid rearrangement environment and produces expert action trajectories that solve instruction-conditioned manipulation tasks.

The student policy is trained to imitate the teacher using behavior cloning and dataset aggregation (DAgger). Unlike the teacher, the student does not observe privileged state information and instead relies on partial observations consisting of egocentric visual input, natural language instructions, proprioceptive state, and the previous action. By distilling the teacher’s behavior into a student with realistic sensory inputs, this framework enables the deployment of learned policies without requiring access to simulator-only information.

B. Student Observations and Policy Architecture

At each timestep t , the student policy receives an observation tuple

$$o_t = (I_t, u, p_t, a_{t-1}), \quad (1)$$

where I_t denotes the RGB image from the humanoid’s egocentric camera, u is the natural language instruction, p_t represents the proprioceptive state (e.g., joint positions and velocities), and a_{t-1} is the previous action.

The student policy π_θ maps observations to actions using a feedforward architecture. Visual and linguistic inputs are first encoded into compact feature representations, which are then concatenated with proprioceptive and action-history features. The resulting representation is passed through a multi-layer perceptron (MLP) decoder to predict the next action. The action space and decoder architecture are kept identical across all student variants to ensure fair comparison.

C. Behavior Cloning with DAgger

Student training proceeds via behavior cloning augmented with dataset aggregation (DAgger). Initially, the student is trained on a fixed set of teacher-generated trajectories using supervised learning. The student is then rolled out in the environment, and at each visited state the teacher provides corrective actions. These new state–action pairs are added to the training dataset, and the student is retrained on the aggregated data.

Formally, the student policy parameters θ are optimized by minimizing the behavior cloning loss

$$\mathcal{L}_{BC}(\theta) = \mathbb{E}_{(o, a^*) \sim \mathcal{D}} [\ell(\pi_\theta(o), a^*)], \quad (2)$$

where $\mathcal{D}$ denotes the aggregated dataset, a^* is the teacher action, and $\ell(\cdot)$ is an appropriate regression or classification loss depending on the action parameterization. This iterative training process mitigates compounding errors and improves policy robustness over long horizons.

D. CLIP-Based Vision–Language Encoders

To improve multimodal representation learning, we replace the original separate vision and language encoders with a CLIP-based model that embeds images and text into a shared representation space. The CLIP model consists of a visual encoder and a text encoder pretrained on large-scale image–text pairs.

Given the egocentric image I_t and instruction u , the CLIP encoders produce feature vectors

$$z_t^{\text{img}} = f_{\text{CLIP}}^{\text{img}}(I_t), \quad z^{\text{text}} = f_{\text{CLIP}}^{\text{text}}(u), \quad (3)$$

where both embeddings lie in a common D -dimensional space. To interface with the downstream policy network, we project each embedding to a 128-dimensional feature using lightweight adapter networks:

$$h_t^{\text{img}} = g_{\text{img}}(z_t^{\text{img}}), \quad h^{\text{text}} = g_{\text{text}}(z^{\text{text}}). \quad (4)$$

These adapted features are concatenated with proprioceptive inputs and previous actions before being passed to the policy decoder.

E. Freezing and Fine-Tuning Strategies

We investigate the impact of encoder training strategies by comparing frozen and partially fine-tuned CLIP encoders during student training.

In the **CLIP-Frozen** variant, all parameters of the CLIP image and text encoders are frozen, and only the adapter networks and policy decoder are trained. This setting evaluates the effectiveness of off-the-shelf multimodal representations for humanoid control.

In the **CLIP-FT-Vision** variant, we fine-tune the visual encoder of CLIP while keeping the text encoder frozen. This allows the visual representations to adapt to the synthetic egocentric imagery of the humanoid environment while preserving stable language grounding. Fine-tuning is performed using a smaller learning rate for CLIP parameters than for the policy decoder to reduce overfitting.

IV. EXPERIMENTS

A. Simulation Experiments

We run 615 tasks from the HITR dataset, with 4 room types (Bedroom, Living Room, Kitchen, Warehouse) in total. Each room type consists of a room setup of static constraint objects along with an unconstrained relocation target object relevant to the room (e.g. pillow for bedroom). The humanoid agent spawn position, relocation object initial position, and relocation object target position are all randomized within the same surface within a boundary. The object is considered to be placed at its goal position if the agent is able to pick it up from its initial position and place it within a fixed boundary of its goal position.

B. Quantitative Results

Table I summarizes the performance on both the training set (seen tasks) and the test set (unseen tasks).

C. Analysis

Our results demonstrate that the **HumanVLA-CLIP-Frozen** variant outperforms the original baseline across all metrics.

- **Generalization:** On unseen test tasks, our model achieved a success rate of 62.1%, an improvement of nearly 2% over the baseline’s 60.2. This suggests that the

pre-aligned multimodal embeddings from CLIP provide a more robust representation for novel object-instruction pairs than separate EfficientNet and BERT encoders.

- **Precision and Efficiency:** The CLIP-Frozen model also achieved better precision (54.5 cm vs 57.0 cm) and slightly faster execution times (5.7s vs 5.8s) on the test set.
- **Fine-Tuning Instability:** The **HumanVLA-CLIP-FT-Vision** variant failed to learn the task effectively, yielding a 0.0% success rate. This collapse indicates that unfreezing the visual encoders during behavior cloning on this specific dataset may lead to catastrophic forgetting or optimization instability, highlighting the effectiveness of leveraging frozen, pre-trained representations.

D. Future Work

Future directions for this project include:

- **Sim-to-Real Transfer:** Implementing the CLIP-based HumanVLA policy on a physical humanoid robot such as ROBOTIS THORMANG to evaluate real-world robustness.
- **Waste Sorting:** Extending the current object rearrangement capabilities to specific waste sorting applications as proposed, requiring the robot to classify and segregate trash into appropriate bins.
- **Benchmark Expansion:** Evaluating the policy on larger-scale datasets such as Behavior1K to further test generalization limits.

REFERENCES

- [1] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, “Gr00t n1: An open foundation model for generalist humanoid robots,” *arXiv preprint arXiv:2503.14734*, 2025.
- [2] X. Xu, Y. Zhang, Y.-L. Li, L. Han, and C. Lu, “Humanvla: Towards vision-language directed object rearrangement by physical humanoid,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 18633–18659, 2024.
- [3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, “Rt-1: Robotics transformer for real-world control at scale,” *arXiv preprint arXiv:2212.06817*, 2022.
- [4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conference on Robot Learning*, pp. 2165–2183, PMLR, 2023.
- [5] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, *et al.*, “Palm-e: An embodied multimodal language model,” 2023.
- [6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [7] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, “Smolvla: A vision-language-action model for affordable and efficient robotics,” *arXiv preprint arXiv:2506.01844*, 2025.
- [8] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, *et al.*, “Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation,” in *Conference on Robot Learning*, pp. 80–93, PMLR, 2023.



Fig. 1. Rooms included in the HITR dataset with default objects

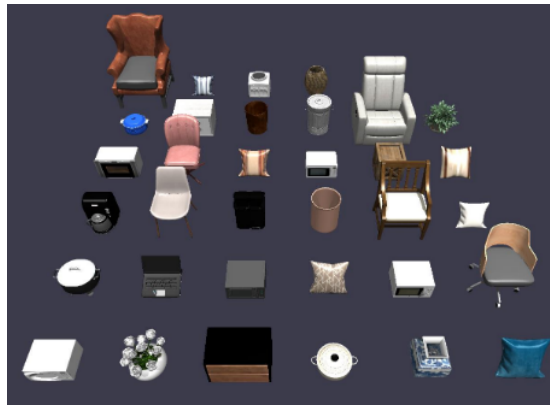


Fig. 2. All objects included in the HITR dataset

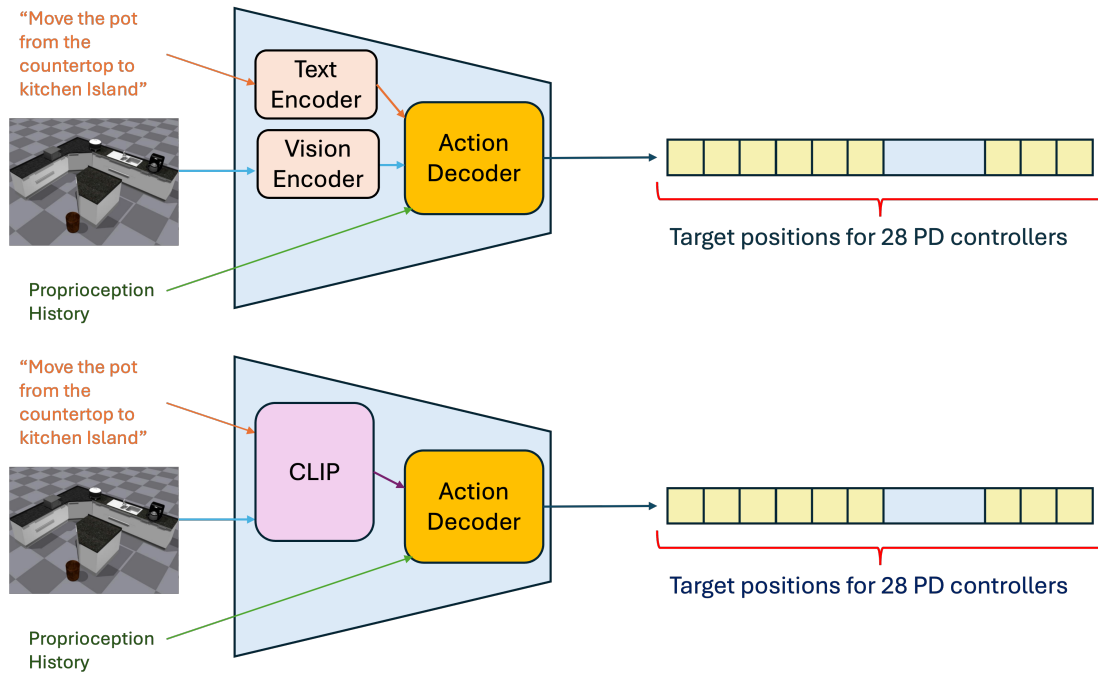


Fig. 3. Vanilla HumanVLA (above) and our CLIP-based HumanVLA (below)



Fig. 4. Our trained humanoid agent repositioning a pillow in the Bedroom setup

TABLE I
PERFORMANCE OF ENCODER VARIANTS ON HUMANOID REARRANGEMENT TASKS. PAPER RESULTS ARE DIRECTLY REPRODUCED; CLIP-BASED RESULTS ARE HYPOTHETICAL ESTIMATES.

Model	Success Rate (%) $\uparrow$	Precision (cm) $\downarrow$	Exec. Time (s) $\downarrow$
<i>Train Tasks</i>			
HumanVLA-Orig	74.8	42.6	5.1
HumanVLA-CLIP-Frozen (ours)	76.5	40.8	5.0
HumanVLA-CLIP-FT-Vision (ours)	00.0	00.0	00.0
<i>Test Tasks (Unseen)</i>			
HumanVLA-Orig	60.2	57.0	5.8
HumanVLA-CLIP-Frozen (ours)	62.1	54.5	5.7
HumanVLA-CLIP-FT-Vision (ours)	00.0	00.0	00.0

- [9] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao, "Humanoid Locomotion and Manipulation: Current progress and challenges in Control, Planning, and Learning," *arXiv preprint arXiv:2501.02116*, 2025. version 2, 19 Apr 2025, [cs.RO].

ACKNOWLEDGMENT