ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Toward Human Preference Optimization for Vision-Language-Action Models: A Pilot Study on the Limits of Imitation Learning

Tae-Won Lee[1], Dongwook Kim[1]*
[1]*Department of Robotics and Mechatronics Engineering, DGIST*
{leetee95, dw_kim}@dgist.ac.kr *Corresponding Author

***Abstract*— Vision-Language-Action (VLA) models trained via imitation learning have achieved impressive results on robotic manipulation, yet their performance degrades significantly on complex, multi-step tasks. We evaluate NVIDIA GR00T N1.6, a state-of-the-art cross-embodiment VLA model, on the SimplerEnv benchmark to systematically identify where imitation learning falls short. Our results reveal a stark performance gap between simple single-step tasks (e.g., picking a can, 90.0%) and complex sequential tasks (e.g., placing an object in a closed drawer, 4.5%), suggesting that behavior cloning alone cannot capture the nuanced decision-making required for long-horizon manipulation. Based on these findings, we propose Human Preference Optimization (HPO) as a post-training strategy to bridge this gap — leveraging human trajectory rankings and reinforcement learning to refine VLA policies beyond what demonstration data alone can teach.**

Keywords*— *Vision-Language-Action Models, Imitation Learning, Human Preference Optimization, Reinforcement Learning, Robotic Manipulation

I. Introduction

State-of-the-art VLA models achieve near-perfect performance on simple grasping tasks, creating an illusion of solved manipulation — but complex, multi-step tasks remain far from reliable. While imitation learning through behavior cloning has been the dominant paradigm for training these models, it is fundamentally limited by the quality and coverage of demonstration data. The model can only reproduce what was demonstrated and cannot discover novel recovery strategies, explore alternative solutions, or improve beyond the demonstrator's skill level.

Reinforcement learning enables trial-and-error exploration that can uncover novel behaviors absent from any dataset — but it requires a well-defined reward function, which is notoriously difficult to design for complex manipulation tasks. Human Preference Optimization (HPO) bridges this gap: humans provide the reward signal by ranking robot trajectories, and RL provides the optimization mechanism to act on it. This paradigm has transformed large language models (e.g., RLHF in ChatGPT [1]) but has not yet been systematically applied to VLA models for robotic manipulation.

In this work, we first establish a rigorous empirical baseline to quantify where imitation learning fails, and then propose HPO as a targeted post-training strategy for the identified failure modes.

II. Background

A. Imitation Learning Limitations.

Behavior cloning learns a mapping from observations to actions using expert demonstrations. It suffers from distribution shift — compounding errors when the robot encounters unseen states — lacks recovery behaviors, and cannot optimize for objectives beyond mimicking the demonstrator [2].

B. Human Preference Optimization

Instead of learning from demonstrations alone, HPO trains a reward model from human pairwise comparisons of robot trajectories. The policy is then refined via RL to maximize this learned reward. This enables optimization for qualities that are hard to specify programmatically: motion smoothness, efficiency, safety, and task robustness [3].

III. Experimental Setup

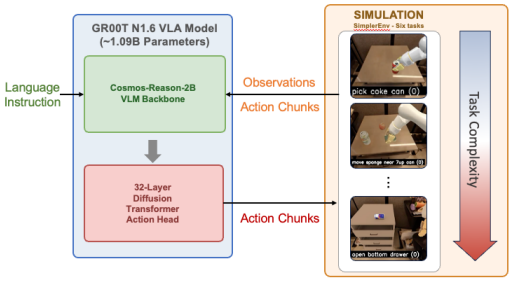

Fig. 1. VLA closed-loop evaluation pipeline. GR00T N1.6 receives observations and returns action chunks. Six SimplerEnv tasks are evaluated in order of increasing complexity, from single-object grasping to multi-step sequential manipulation.

We evaluate the pretrained GR00T N1.6-fractal checkpoint on 6 SimplerEnv Fractal tasks using a Google Robot embodiment. Tasks are ordered by increasing complexity: (1) Pick coke can; (2) Pick object; (3) Move near; (4) Close drawer; (5) Open drawer; (6) Place in closed drawer — a sequential task requiring multi-step reasoning. The evaluation uses a client-server architecture with ZeroMQ communication. The simulation client runs 10 parallel environments, 200 episodes per task, n_action_steps=1, max_episode_steps=300. Hardware: NVIDIA RTX PRO 5000 Blackwell (48GB), Ubuntu 24.04. Table 1 summarizes the six tasks and their characteristics.

IV. Results And Analysis

Table I summarizes the closed-loop success rate and average episode time.

Task	Avg. Success Rate	Avg. Time (s)
Pick coke can	0.900	2.18
Pick object	0.875	2.70
Move near	0.851	2.70
Close drawer	0.435	10.63
Open drawer	0.100	15.25
Place in closed drawer	0.045	436.50

Table. 1. Closed-loop success rate and average episode time on SimplerEnv Fractal (Google Robot). GR00T N1.6 pretrained checkpoint, 200 episodes per task.

As task complexity increases from simple grasping to sequential multi-step manipulation, Fig. 2 shows a clear degradation pattern in both success rate and episode time. Performance drops from 90% on simple grasping to 4.5% on sequential multi-step tasks. The evaluated model is trained entirely via behavior cloning. Therefore, the observed degradation directly reflects the limitations of imitation learning as a training paradigm. The average episode time further supports this: simple tasks complete in under 3 seconds, while complex tasks approach the maximum step timeout, indicating the policy fails to make meaningful progress.

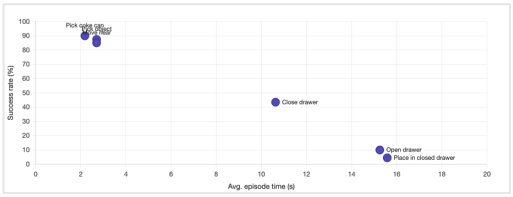

Fig. 2. Success rate vs. average episode time across 6 manipulation tasks. Simple grasping tasks (top-left) achieve high success with fast episodes, while complex sequential tasks (bottom-right) show both low success and longer episodes due to max-step timeouts on failure.

Fig. 3 illustrates the contrast between a simple task (pick coke can) and a complex task (open bottom drawer), where the latter demands articulated object interaction that is rarely covered in demonstration data.

Fig. 3. Example simulation environments from SimplerEnv Fractal. (Left) Pick coke can — a simple single-object grasping task achieving 90% success rate. (Right) Open bottom drawer — a complex articulated manipulation task achieving 10% success rate. The stark visual and performance contrast illustrates the challenge gap that imitation learning alone cannot bridge.

Three failure modes emerge:

No recovery behavior. Recovery states are absent from training data. RL-based refinement enables the policy to explore and discover recovery strategies via PPO or DPO.

Distribution shift. Errors compound over long horizons. Human trajectory ranking provides learning signal even from failed episodes.

No quality optimization. The model mimics demonstrations but does not learn why certain trajectories are better. A learned reward model captures preferences that binary success/failure cannot express.

These failure modes are not addressable by simply scaling up demonstration data — they require a fundamentally different optimization signal, which Human Preference Optimization can provide.

V. Proposed HPO Pipeline

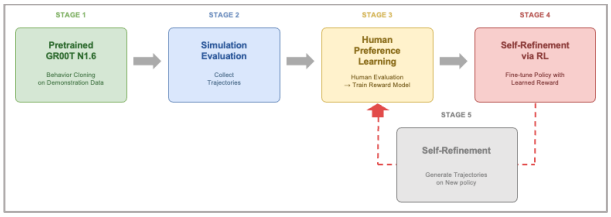

Fig. 4. Proposed Human Preference Optimization pipeline. The pretrained GR00T N1.6 policy (Stage 1) generates rollout trajectories in simulation (Stage 2). Human evaluators rank trajectory pairs to train a reward model (Stage 3), which guides RL-based policy fine-tuning (Stage 4). The refined policy generates new trajectories for iterative self-refinement (Stage 5), progressively improving performance on complex tasks.

VI. Discussion And Future Work

Self-refinement alternative. Explore whether VLA models can self-evaluate rollout quality using the VLM backbone's reasoning capability as an automatic reward signal.

RL integration. Combining self-refinement with RL for progressive autonomous improvement of manipulation skills.

Sim-to-Real transfer. Validate that HPO improvements in simulation transfer to physical robot execution.\

Limitation. This work evaluates a single VLA model on a single benchmark. Future work will extend to additional models (e.g., OpenVLA, Octo) and benchmarks (LIBERO, RoboCasa) to verify that the identified failure modes generalize across architectures and task domains.

References

[1] L. Ouyang et al., "Training language models to follow instructions with human feedback," NeurIPS, 2022.

[2] S. Ross et al., "A reduction of imitation learning and structured prediction to no-regret online learning," AISTATS, 2011.

[3] P. Christiano et al., "Deep reinforcement learning from human preferences," NeurIPS, 2017.

[4] NVIDIA, "GR00T N1.6," 2025. github.com/NVIDIA/Isaac-GR00T

[5] X. Li et al., "SimplerEnv," arXiv:2405.05941, 2024