Critic Architecture Matters: Dual vs. Unified Critics for Humanoid Loco-Manipulation

Mehmet Turan Yardımcı
Department of Computer Engineering
Çukurova University
Adana, Türkiye
mehmetturan2003@gmail.com

***Abstract*—Multi-objective reinforcement learning for humanoid robots must coordinate locomotion and manipulation within a single policy. A natural design choice is whether to use a single (unified) critic that estimates the combined value of all objectives, or separate (dual) critics with disjoint reward signals. We present a controlled comparison on the Unitree G1 humanoid (23 active DoF) in NVIDIA Isaac Lab, training loco-manipulation policies through a sequential curriculum spanning 13 levels from stationary reaching to walking with variable-orientation targets. In standardized evaluation, dual-critic policies reach targets 3.5× faster (6.5 vs. 22.6 simulation steps), achieve 2× higher throughput (14.3 vs. 7.0 validated reaches per 1,000 steps), and attain higher validated reach rates (65.2% vs. 53.8%) compared to the unified-critic policy. Notably, additional anti-gaming reward mechanisms provide no further improvement beyond the architectural change alone (60.9% vs. 65.2%). These results have direct implications for the emerging paradigm of RL fine-tuning of imitation-learned policies: when refining a pre-trained manipulation policy with RL, a unified critic risks suppressing the learned behavior through competing locomotion gradients. These findings demonstrate that critic architecture is a primary—and often overlooked—design choice in multi-objective humanoid RL, with greater impact than reward engineering on reaching efficiency.**

***Index Terms*—reinforcement learning, imitation learning, humanoid robots, critic architecture, loco-manipulation, fine-tuning**

I. Introduction

Humanoid loco-manipulation—the simultaneous coordination of walking and reaching—is a fundamental capability for robots operating in human environments. While imitation learning approaches such as ALOHA [1] and Mobile ALOHA [2] have demonstrated impressive manipulation, reinforcement learning (RL) offers a complementary paradigm. Consider how infants learn to walk: they begin by crawling—with no need to observe other crawlers—then progress to assisted walking, and eventually walk independently, in a process analogous to curriculum learning. This exploration-driven process stands in direct contrast to the imitation learning paradigm. Yet these paradigms are increasingly combined: recent work fine-tunes imitation-learned policies with RL to improve robustness and generalization beyond the demonstration distribution [16], [17]. In such hybrid pipelines, the critic architecture determines whether RL gradients *refine* or *overwrite* the pre-trained behavior—a question our findings directly address. However, RL introduces its own challenges, particularly when multiple objectives must be coordinated within a single policy.

A fundamental design choice in multi-objective RL is the critic architecture. A **unified critic** receives the concatenated observations from all objectives and estimates their combined value, while **dual critics** maintain separate value functions with disjoint reward signals. This choice is rarely discussed in the humanoid RL literature, where most works adopt one architecture without comparing alternatives.

In this paper, we present a controlled comparison of unified and dual critic architectures for humanoid loco-manipulation on the Unitree G1. Both architectures successfully learn reaching behavior, but we find substantial differences in *efficiency*: dual critics produce reaches that are 3.5× faster and achieve 2× higher throughput. Crucially, these differences are invisible to standard training metrics—reward values, curriculum progression, and reach counts appear similar across architectures. Only standardized evaluation with validated metrics reveals the gap.

The contributions of this paper are: (1) A Dual Actor-Critic framework with frozen-branch transfer for humanoid loco-manipulation on the Unitree G1 in NVIDIA Isaac Lab. (2) A controlled comparison showing dual critics achieve 3.5× faster reaching and 2× higher throughput than a unified critic, while additional anti-gaming reward mechanisms provide no further benefit. (3) A standardized three-way benchmark methodology demonstrating that standard training metrics fail to capture efficiency differences between architectures.

II. Related Work

Humanoid Locomotion and Loco-Manipulation. Radosavovic et al. [3] demonstrated zero-shot sim-to-real transfer of a locomotion policy on a full-size humanoid. He et al. [4] proposed HOVER for multi-mode whole-body control, and Sun et al. [5] introduced ULC, a unified controller for the G1 using sequential skill acquisition. Fu et al. [6] and Cheng et al. [7] demonstrated whole-body control for legged manipulation. These works adopt specific architectures without comparing alternatives. Our work isolates the critic architecture as a variable and measures its impact on reaching efficiency.

Curriculum Learning in RL. Curriculum learning [8] proposes training on progressively harder tasks. OpenAI's

Rubik's Cube work [9] used automatic domain randomization as an implicit curriculum, and Narvekar et al. [10] surveyed curriculum methods. ULC [5] employs sequential skill acquisition similar to ours. Our work reveals that curriculum progression metrics can mask efficiency differences between architectures.

Multi-Objective RL and Critic Design. Reward hacking [11]–[13] is well-documented, but typically attributed to reward design. Multi-critic approaches have been explored in multi-agent settings, but their impact on single-agent multi-objective robotics is less studied. Our work provides empirical evidence that separating critics by objective yields substantial efficiency gains in humanoid control, independent of reward design.

RL Fine-Tuning of Imitation-Learned Policies. A growing body of work uses RL to refine policies initially trained via imitation learning. Haldar et al. [16] showed that RL fine-tuning of IL policies improves robustness to distribution shift, while Luo et al. [17] demonstrated residual RL on top of diffusion-based IL policies. These approaches face a tension: the RL objective must improve specific behaviors without catastrophically forgetting the imitation-learned skills. This tension is structurally analogous to our multi-objective setting, where locomotion and manipulation compete for gradient signal through the critic. Our finding that dual critics prevent objective interference suggests they may similarly protect IL-learned behaviors during RL fine-tuning.

III. Method

A. System Overview

We train a Unitree G1 humanoid robot (23 active DoF) in NVIDIA Isaac Lab [14] with 4096 parallel environments. The robot uses 12 leg joints and 5 arm joints (right arm: shoulder pitch/roll/yaw, elbow pitch/roll), with wrist and hand joints fixed.

Training proceeds sequentially: Stages 1–3 train locomotion only; Stage 6 trains both branches jointly for loco-manipulation. Both branches use PPO [15] with learning rate 3×10^{-4} (cosine annealing), $\gamma = 0.99$, $\lambda = 0.95$, and clip ratio 0.2.

Unified Critic Architecture. Our initial design uses separate actors π_{loco} (57-dim obs $\rightarrow$ 12 leg actions) and π_{arm} (52-dim obs $\rightarrow$ 12 arm+finger actions), but a **single critic** V_{unified} that receives the concatenated 109-dimensional observation and estimates the combined value of locomotion and manipulation rewards.

Dual Actor-Critic Architecture. The revised design (Fig. 1) separates the system into fully independent branches. The *Locomotion branch* (actor π_{loco}, critic V_{loco}) observes proprioceptive state (57-dim: base velocities, projected gravity, leg joint states, velocity commands, gait phase) and outputs 12 leg joint targets. The *Arm branch* (actor π_{arm}, critic V_{arm}) observes arm-relevant state (52-dim: base motion, leg context, arm joint states, end-effector/target positions; 55-dim in S7 with 3 anti-gaming features) and outputs 5 arm joint residual actions. The two critics receive **disjoint reward signals**:

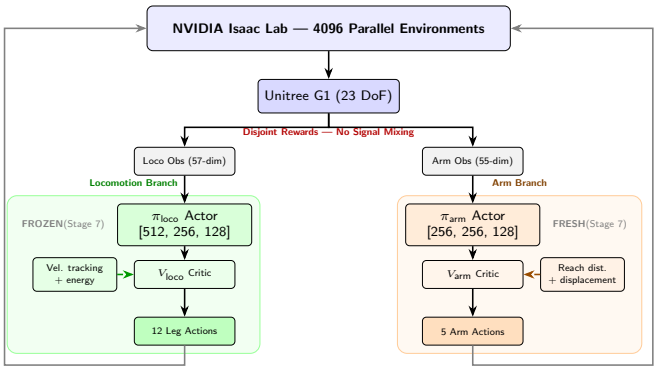

Fig. 1. Dual Actor-Critic architecture. The locomotion branch (green) and arm branch (orange) maintain separate critics with disjoint reward signals. The unified variant (not shown) concatenates both observation streams into a single 109-dim critic input.

locomotion critic on velocity tracking and balance, arm critic on reaching distance and displacement only.

B. Sequential Curriculum Design

The curriculum spans 13 levels (0–12) across four phases (Table I): stationary reaching (Levels 0–4), walking with reaching (5–6), fixed end-effector orientation (7–8), and variable orientation within an expanding cone (9–12, 20°–80°). Advancement requires sustained validated reach rate above a threshold.

TABLE I
CURRICULUM LEVELS FOR LOCO-MANIPULATION TRAINING.

Level	Phase	v_x (m/s)	Pos. Thresh.	Orient.
0–4	Stand+Reach	0.0	0.12→0.06m	None
5–6	Walk+Reach	0–0.3	0.06→0.05m	None
7–8	Fixed Orient	0–0.4	0.05m	Palm-down
9–12	Var. Orient	0–0.6	0.05→0.04m	20–80°

C. Anti-Gaming Reward Mechanisms

During development, we observed that proximity-based rewards combined with automatic target resampling could allow policies to accumulate spurious reach counts. To investigate whether reward engineering could further improve performance beyond the architectural change, we implemented five anti-gaming mechanisms in a third variant (Stage 7): (1) absolute workspace sampling with minimum distance constraints, (2) three-condition validated reach requiring position proximity, displacement, and time limit:

$$\text{valid} = (\|p_{ee} - p_t\| < \epsilon_{pos}) \wedge (\|p_{ee} - p_{ee}^0\| > d_{disp}) \wedge (t < t_{max}) \quad (1)$$

(3) movement-centric rewards with velocity-toward-target and stillness penalties, (4) validated reach rate for curriculum advancement, and (5) gaming detection heuristics. This variant uses the dual-critic architecture with a frozen locomotion branch and freshly initialized arm policy.

IV. Experiments

A. Setup

All experiments use NVIDIA Isaac Lab with 4096 parallel environments on an RTX 5070 Ti GPU (12GB VRAM), achieving approximately 17,000 steps/second. We compare three policies in a standardized benchmark with identical evaluation conditions (absolute target sampling, minimum target distance 0.12m, position threshold 0.06m, displacement threshold 0.10m, timeout 150 steps, 1 environment, deterministic actions):

- **S6u** (Unified Critic): Unified critic (109-dim), 52-dim arm obs, 12-dim arm action, Level 10/12
- **S6s** (Dual Critic): Dual Actor-Critic, 52-dim arm obs, 5-dim arm action, Level 12/12
- **S7** (Dual + Anti-Gaming): Dual Actor-Critic with anti-gaming reward mechanisms, frozen locomotion, fresh arm policy, Level 7/7

Each policy is evaluated over 3,000 steps in both standing and walking modes. Checkpoint loading is verified with bit-exact weight matching for shared locomotion parameters.

B. Results

Table II presents the three-way comparison. All three policies successfully reach targets, but with substantial efficiency differences.

TABLE II
STANDARDIZED BENCHMARK (3,000 STEPS PER MODE). S6U: UNIFIED CRITIC. S6S: DUAL CRITIC. S7: DUAL CRITIC + ANTI-GAMING.

Metric	S6u	S6s	S7
Critic architecture	Unified	Dual	Dual+AG
Curriculum level	10/12	12/12	7/7
Standing evaluation:			
Validated reach rate	53.8%	**65.2%**	60.9%
Position-only rate	59.0%	72.7%	71.9%
Avg. time-to-reach	22.6 steps	**6.5 steps**	5.8 steps
Validated / 1K steps	7.0	**14.3**	13.0
Timeout rate	41.0%	27.3%	28.1%
Avg. displacement	0.219m	0.206m	0.203m
Walking evaluation:			
Validated reach rate	51.1%	**63.1%**	56.2%
Avg. time-to-reach	16.4 steps	**7.0 steps**	5.9 steps
Validated / 1K steps	7.7	**13.7**	12.0
Timeout rate	37.8%	26.2%	28.1%
Arm action magnitude	1.22	2.54	3.04

The most striking difference is **reaching speed**: the unified-critic policy requires 22.6 simulation steps (0.45s) to reach a target on average, while the dual-critic policy requires only 6.5 steps (0.13s)—a 3.5× improvement. This directly translates to **throughput**: dual critics achieve 14.3 validated reaches per 1,000 steps compared to 7.0 for the unified critic, a 2× improvement.

The arm action magnitudes provide mechanistic insight: the unified critic produces actions with mean magnitude 1.22, roughly half that of the dual critics (2.54 and 3.04). This suggests the unified critic's value landscape partially suppresses arm actions—the arm moves, but with less commitment, resulting in slower convergence to targets and more timeouts.

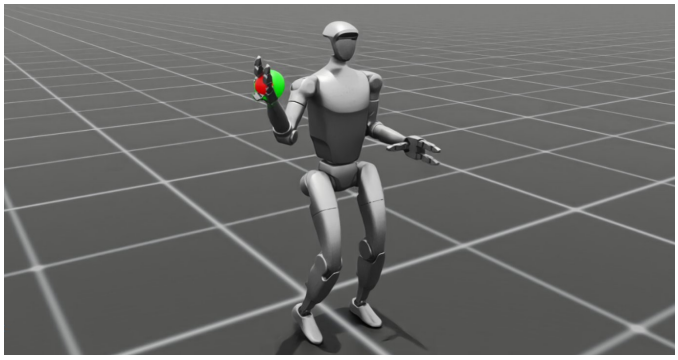

Fig. 2. Dual-critic policy (S6s) during play evaluation. The Unitree G1 reaches toward a target (green sphere) with the red marker indicating the end-effector tracking point.

Fig. 2 shows the trained dual-critic policy during play evaluation, demonstrating active reaching toward targets during locomotion.

C. Discussion

Three findings emerge from the benchmark:

(1) Critic architecture determines reaching efficiency. The unified-to-dual critic change produces the primary improvement: 3.5× faster reaching, 2× throughput, and 11 percentage points higher validated rate (65.2% vs. 53.8%). We attribute this to the unified critic's challenge of estimating combined value across competing objectives: locomotion reward dominates early training, partially suppressing arm action magnitudes and producing a more conservative reaching strategy.

(2) Anti-gaming reward mechanisms provide no additional benefit. The dual-critic policy without anti-gaming mechanisms (S6s, 65.2%) slightly outperforms the variant with five anti-gaming mechanisms (S7, 60.9%). The S7 policy trained for fewer iterations (reaching Level 7 vs. Level 12), which likely accounts for the small gap. This suggests that once the architectural bottleneck is resolved, additional reward engineering for anti-gaming is unnecessary.

(3) Training metrics mask efficiency differences. The unified-critic policy reached Level 10, accumulated 3.3M training reaches, and achieved reward 36.2—metrics that appear comparable to the dual-critic variants. The 3.5× speed difference and 2× throughput gap are invisible to these standard metrics and only emerge through standardized evaluation with time-to-reach and throughput measurements. This highlights the importance of evaluation metrics that go beyond reward and curriculum level.

(4) Implications for RL fine-tuning of imitation-learned policies. Our results suggest that critic architecture is particularly important in hybrid IL+RL pipelines. Consider a scenario where a manipulation policy is first learned from human demonstrations (e.g., via teleoperation), then fine-tuned with RL to improve precision or adapt to new objects. If a unified critic is used, the locomotion reward signal—which was not part of the original IL objective—can suppress the learned

arm behavior, analogous to the action magnitude suppression we observe in S6u (mean magnitude 1.22 vs. 2.54 for dual). A dual-critic architecture naturally isolates the RL fine-tuning signal to the relevant branch, preserving the imitation-learned behavior in other branches. This architectural choice may be critical for preventing catastrophic forgetting of IL-acquired skills during RL refinement, a key challenge identified in recent IL+RL work [16], [17].

Confounding factors. We note that S6u and S6s differ in arm action dimensionality (12 vs. 5 DoF) and critic architecture, while sharing identical observation dimensionality (52-dim arm obs). Future work will isolate these factors through controlled ablations. Nevertheless, the 5-DoF arm is a strict subset of the 12-DoF action space, making the critic architecture the most likely explanatory variable.

V. Conclusion and Future Work

We presented a controlled comparison of unified and dual critic architectures for humanoid loco-manipulation on the Unitree G1. Both architectures learn functional reaching, but dual critics achieve 3.5× faster reaching and 2× higher throughput. Additional anti-gaming reward mechanisms provide no further benefit beyond the architectural change. Our findings suggest that in multi-objective humanoid RL, critic architecture is a primary design choice with greater impact on efficiency than reward engineering.

Limitations include simulation-only evaluation, single-seed training, 5-DoF arm control, and the confounding between action dimensionality and critic architecture. Future work will conduct controlled ablations isolating action dimensionality from critic architecture, extend to 29-DoF dual-arm control with a Triple Actor-Critic, and investigate dual critics as a mechanism for preserving imitation-learned skills during RL fine-tuning—testing whether the gradient isolation we observe transfers to hybrid IL+RL training pipelines. We also plan to integrate vision-language models for hierarchical task specification and pursue sim-to-real transfer.

Acknowledgment

An AI assistant (Claude, Anthropic) was used for structuring the manuscript outline and refining academic prose. All technical content, experimental design, implementation, and results are solely the work of the author.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. RSS*, 2023.

[2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[3] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, 2024.

[4] T. He et al., "HOVER: Versatile neural whole-body controller for humanoid robots," *arXiv preprint arXiv:2410.21229*, 2024.

[5] 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.

[6] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Proc. CoRL*, 2022.

[7] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proc. RSS*, 2024.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009.

[9] OpenAI et al., "Solving Rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[10] S. Narvekar et al., "Curriculum learning for reinforcement learning domains: A framework and survey," *JMLR*, vol. 21, no. 181, pp. 1–50, 2020.

[11] D. Amodei et al., "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[12] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, "Defining and characterizing reward hacking," in *Proc. NeurIPS*, 2022.

[13] A. Pan, K. Bhatia, and J. Steinhardt, "The effects of reward misspecification: Mapping and mitigating misaligned models," in *Proc. ICLR*, 2022.

[14] M. Mittal et al., "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robot. Autom. Lett.*, vol. 8, no. 6, pp. 3740–3747, 2023.

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

[16] S. Haldar, J. Mathur, D. Bernstein, and L. Pinto, "Teach a robot to FISH: Versatile imitation from one minute of demonstrations," in *Proc. RSS*, 2023.

[17] J. Luo et al., "Serl: A software suite for sample-efficient robotic reinforcement learning," in *Proc. ICRA*, 2024.