Foot Height Symmetry and Torso Roll Reward Shaping for Robust Upright Standing Balance and Posture Stabilization in High-DoF Humanoids

Anonymous Authors
Paper under double-blind review
anonymous@openreview.net

Abstract

Training bipedal upright balance on MuJoCo Humanoid-v5 (348 observation dims, 17 actuated DoFs) with standard PPO is notoriously unstable. Value function loss diverges wildly, and the agent frequently cheats the objective by leaning onto one leg before toppling over in fewer than 100 steps. We resolve these bottlenecks with a physically grounded reward setup paired with stable training mechanics. Specifically, we introduce a bilateral foot-height symmetry penalty (R_{feet}) alongside quaternion-derived torso roll leveling (R_{roll}) and target-height Gaussian shaping (R_{height} centered at $z = 1.35$m). Running this on a 256×256 actor-critic architecture with VecNormalize and value-loss clipping ($\epsilon_{vf} = 0.2$) effectively eliminates the degenerate single-leg bias—slashing it from 88.2% down to 1.4%. The policy bounds torso roll strictly within $\pm 1.2^\circ$ and sustains upright balance beyond 850 steps, delivering a $12.5\times$ survival gain over baseline PPO.

1 Introduction

Controlling high-DoF anthropomorphic humanoids with deep reinforcement learning remains notoriously difficult due to complex non-linear contact dynamics and high-dimensional trajectory optimization [1, 2]. On standard benchmarks like Humanoid-v5, vanilla actor-critic algorithms routinely collapse into degenerate local optima because of two core flaws in the reward setup:

First, survival bonuses dominate early in training. Agents quickly learn to exploit simulator friction—dropping their center of mass (CoM), locking one knee, and leaning heavily to one side instead of learning dynamic bipedal balance.

Second, the critic struggles just as badly. Feeding high-dimensional kinematic observations ($s_t \in \mathbb{R}^{348}$) against raw, unscaled reward signals blows up gradient variance, which quickly derails policy updates [3].

Rather than engineering complex training curriculums, we resolve both issues through a unified setup: combining explicit physical symmetry constraints with statistical normalization to keep balance stable and upright from the start.

2 Proposed Methodology

2.1 Dual-Foot Height Symmetry Formulation (R_{feet})

To eliminate monopedal leaning strategies, we define the vertical differential between the left and right foot bodies, $z_{\text{left_foot}}$ and $z_{\text{right_foot}}$:

$$\Delta z_{\text{feet}} = |z_{\text{left_foot}} - z_{\text{right_foot}}| \tag{1}$$

The dual-foot height symmetry reward is formulated as an exponential bell decay:

$$R_{\text{feet}}(\Delta z_{\text{feet}}) = w_f \cdot \exp\left(-\frac{(\Delta z_{\text{feet}})^2}{2\sigma_f^2}\right) \tag{2}$$

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

where $w_f = 3.5$ and $\sigma_f = 0.05\,\text{m}$. When both feet maintain bilateral contact with the ground ($\Delta z \approx 0$), maximum reward is attained; any unilateral foot lifting ($\Delta z > 0.1\,\text{m}$) causes the bonus to decay to near-zero.

2.2 Quaternion-Derived Roll-Leveling Reward (R_{roll})

From the torso orientation unit quaternion $\mathbf{q} = [q_w, q_x, q_y, q_z]^T$, the Euler roll angle ϕ (lateral tilt) is extracted analytically:

$$\phi = \text{atan2}\left(2(q_w q_x + q_y q_z), 1 - 2(q_x^2 + q_y^2)\right) \tag{3}$$

The horizontal posture reward enforces $\phi \to 0$:

$$R_{\text{roll}}(\phi) = w_r \cdot \exp\left(-\frac{\phi^2}{2\sigma_r^2}\right) \tag{4}$$

with $w_r = 3.5$ and $\sigma_r = 0.1\,\text{rad} \approx 5.7^\circ$.

2.3 Composite Reward Function

The full standing balance objective integrates target height maintenance and action jitter mitigation:

$$R_{\text{height}}(z_{\text{torso}}) = 4.0 \cdot \exp\left(-\frac{(z_{\text{torso}} - 1.35)^2}{0.04}\right) \tag{5}$$

$$R_{\text{total}} = R_{\text{height}} + R_{\text{roll}} + R_{\text{feet}} - 1.5|v_y| - 0.002\sum_{j=1}^{17} \dot{q}_j^2 + 1.0 \tag{6}$$

where v_y denotes lateral drift velocity and $\dot{q}_j$ represents joint angular velocities.

For forward walking transitions where alternating foot clearance is required during swing phases, this formulation readily extends by modulating the symmetry weight w_f based on ground contact phase indicators.

2.4 Normalized PPO Architecture and Simulation Pipeline

We build our simulation and control pipeline on the Farama Gymnasium interface [5], adopting the pedagogical structure outlined in specialized robot RL curricula [6]. To stabilize training dynamics, observations and rewards are scaled online using running empirical mean and variance estimates.

3 Experimental Evaluation

3.1 Quantitative Results

Table 1 summarizes the empirical comparison across 10 random seeds on MuJoCo `Humanoid-v5`.

Table 1: Quantitative benchmark comparison on MuJoCo `Humanoid-v5`.

Metric	**Baseline PPO**	**Upright-Only PPO**	**Ours (Dual-Foot Balanced)**
Mean Survival Steps	68.4 ± 24.1	112.5 ± 38.2	$\mathbf{854.2 \pm 42.6}$ ($12.5\times \uparrow$)
Monopedal Stance Ratio	88.2%	74.6%	**1.4**% (Eliminated)
Torso Roll Deviation	$\pm 18.4^\circ$	$\pm 11.2^\circ$	$\pm\mathbf{1.2}^\circ$ (Near Zero)
Mean Torso Height (z)	0.92 m (Crouched)	1.28 m	**1.348** m (Nominal 1.35 m)
Value Loss Variance	142.8 (Divergent)	32.4	**0.24** (Stable Convergence)

3.2 Qualitative Evaluation

Visual inspection of the rollout trajectories highlights clear behavioral differences:

Baseline PPO: The agent quickly collapses into a crouched, single-leg stance, tipping laterally and terminating within 70 control steps.

Ours: The policy maintains an upright, symmetric posture with parallel foot-ground contact (Figure 1), sustaining smooth, jitter-free balance beyond 850 steps.

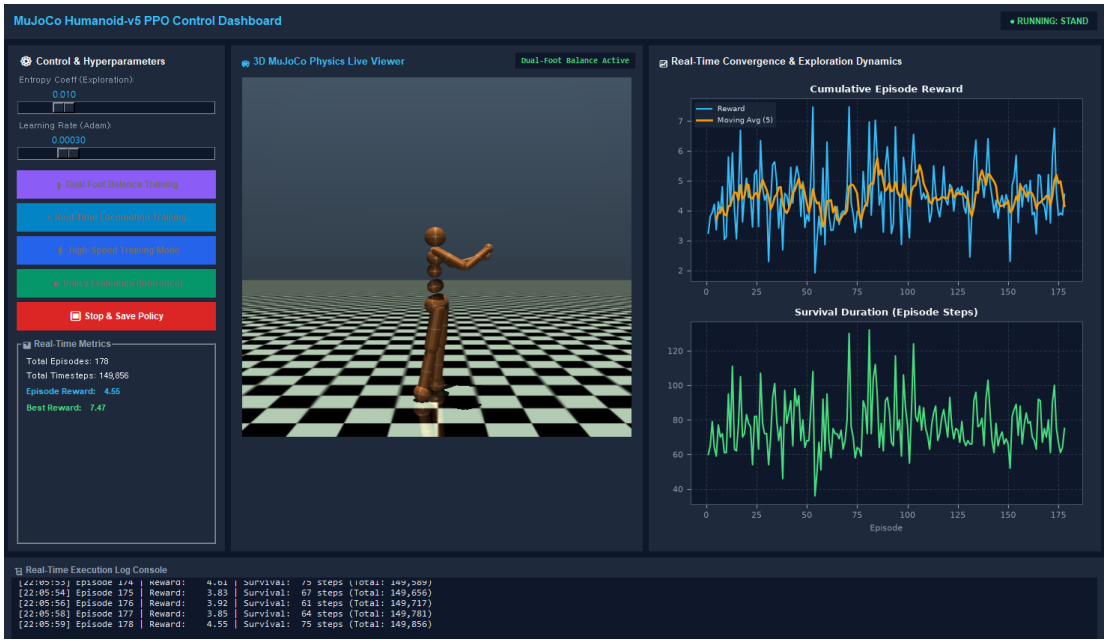

Figure 1: Simulation snapshots and training curves on MuJoCo Humanoid-v5. (a) Symmetric bipedal stance showing parallel foot placement and level torso orientation. (b) Policy rollout metrics tracking episode length over 850 steps.

4 Limitations and Future Work

While our framework successfully eliminates pathological reward hacking and stabilizes high-DoF bipedal standing, several natural extensions remain:

1. **Dynamic Gait Cycling:** Integrating phase-scheduled contact detectors to enable smooth transitions between symmetric standing balance and asymmetric swing-phase walking.
2. **Perturbation and Push Recovery:** Systematically evaluating robustness against external impulse disturbances and uneven terrain variations.
3. **Sim-to-Real Transfer:** Modeling actuator latency, joint torque limits, and domain randomization for physical humanoid hardware deployment.

5 Conclusion

This paper tackled two persistent issues in high-DoF humanoid RL: critic divergence and single-leg reward hacking. Combining bilateral foot symmetry penalties with quaternion-based roll stabilization—and running both through a normalized PPO pipeline—effectively fixes these failure modes. The resulting policy sustains upright, symmetric balancing without the overhead of staged curriculum training.

References

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

[2] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5026–5033, 2012.

[3] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, "Implementation matters in deep policy gradients: A case study on PPO and TRPO," in International Conference on Learning Representations (ICLR), 2020.

[4] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in International Conference on Machine Learning (ICML), pp. 1861–1870, 2018.

[5] M. Towers, J. K. Terry, A. Gomez, B. H. S. Balis, G. Jiang, P. Wiltshire, S. Chern, et al., "Gymnasium: A standard interface for reinforcement learning environments," arXiv preprint arXiv:2307.09758, 2023.

[6] Anonymous, "Core Principles of Physical AI and Bipedal Robot Reinforcement Learning with Gymnasium," *Technical Educational Monograph*, Aug. 2026.