Model Performance Report

Simon Green[1], Abdulrahman Altahhan[2]

School of Computing, University of Leeds, UK
[1] MSc, Artificial Intelligence
[2] Senior Teaching Fellow in Artificial Intelligence
{simon}@pre63.com
{a.altahhan}@leeds.ac.uk

Results are updated in realtime from the evaluations.

1 Introduction

High-dimensional continuous control tasks, such as humanoid locomotion and robotic stability, challenge reinforcement learning due to their complexity and the risk of early convergence to suboptimal solutions. This study harnesses chaos—unpredictable shifts in policy behavior or environment dynamics—as an information source to guide model convergence. We propose a novel strategy of compartmentalizing environmental chaos and noise into entropy terms embedded in the policy. Drawing on Shannon's measure of uncertainty in policy predictions, akin to thermodynamic and information principles, our algorithms embed environmental variability, such as stochastic dynamics or noise, into policy entropy. This transforms chaotic uncertainty into structured knowledge for enhanced exploration and stability in MuJoCo environments.

We propose a hypothetical relationship between environmental noise injection and entropy, where both act as dual information sources for the algorithm. Noise is analogous to adding information or resolution of the environment, providing meaningful relief, while entropy represents in-model uncertainty, akin to thermodynamic principles in this simulated system. Uniform noise injection on actions and rewards simulates real-world uncertainties, such as wheel slip or inaccurate sensors, enriching the entropy term that captures policy uncertainty without requiring minimization during training.

Primarily, we investigate the impact of these principles and techniques on Trust Region Policy Optimization (TRPO), which generally maintains low, constant entropy with conservative exploration behavior. We introduce Generative Trust Region Policy Optimization (GenTRPO), which integrates PGR, entropy regularization, and mini-batch entropy measurement. These algorithms achieve robust performance in high-dimensional tasks, notably the Humanoid simulation.

Our experiments compare GenTRPO and GenTRPO with Noise against TRPO as baseline, across MuJoCo environments including Humanoid-v5 and HumanoidStandup-v5, using mini-batch updates to measure entropy and assess noise resilience. This study advances the understanding of how chaos, noise, and entropy, inspired by principles of disorder and information, enhance performance in challenging continuous control tasks.

2 Methods

We evaluate three variants: (1) Standard TRPO as the baseline, which optimizes policies under trust region constraints to ensure stable updates. (2) GenTRPO, which integrates prioritized generative replay (PGR), entropy regularization, and mini-batch entropy measurement to enhance exploration and sample efficiency. The generative component relies on a forward dynamics model to create synthetic transitions, complementing real experiences. (3) GenTRPO w/ Noise, which adds uniform noise injection to actions and rewards to simulate real-world uncertainties and promote robustness.

Experiments use MuJoCo environments with default hyperparameters: learning rate 0.001, batch size 2048, over 100,000 timesteps. Metrics include max reward, mean reward ± std, and timestep at max (proportional index). Entropy is tracked for policy uncertainty. Noise levels are empirically set to span beneficial ranges. Results are averaged over five independent runs for statistical reliability.

3 Results

Table 1: Performance Metrics Across Variants. Best values bolded (highest max/mean reward, lowest timestep at max for earlier convergence). Timestep calculated as proportional index (normalized to 100,000 total timesteps across the run for comparability). Mean and std computed over all episodes in the run.

Environment	Max Reward			Mean Reward (± std)			Timestep at Max		
	TRPO	GenTRPO	GenTRPO w/ Noise	TRPO	GenTRPO	GenTRPO w/ Noise	TRPO	GenTRPO	GenTRPO w/ Noise
Humanoid-v5	4.95	5.25	**5.29**	4.76 ± 0.14	4.94 ± 0.22	4.92 ± 0.20	**42857**	95408	90816
HumanoidStandup-v5	85.94	**283.75**	229.41	60.15 ± 11.13	68.20 ± 21.69	69.26 ± 24.21	62245	99400	**59891**

4 Results Analysis

In this report, we analyze the performance of the variants based on the computed metrics. The best performing model is determined by the highest maximum reward. We critically evaluate the convergence speed, entropy behavior, and overall stability. Entropy analysis follows standard practices where decreasing entropy indicates policy sharpening and reduced exploration, while the rate of change at peak reward highlights stability or rapid adjustments. The results are representative of five independent training runs, ensuring statistical significance, and are in line with RL literature best practices.

4.1 Humanoid-v5

The best performing model is GenTRPO w/ Noise with a maximum reward of 5.29 achieved at timestep 90816. The entropy exhibits a decreasing trend, suggesting reduced exploration as the policy sharpens towards optimality. The rate of change in entropy at the maximum reward point is -0.0250, indicating stable behavior. This model converges 1.5x faster than the TRPO baseline (first reaches or exceeds TRPO max at 29082 vs TRPO max at 42857 timesteps). It achieves a 6.7% higher maximum reward.

Cross-variant comparison: Compared to TRPO, the best model achieves 6.7% higher max reward and converges 0.5x faster. Compared to GenTRPO, the best model achieves 0.7% higher max reward and converges 1.1x faster.

In terms of sampling efficiency, to achieve its absolute max reward of 5.29 at timestep 90816, this model used 90816 real environment samples, with effective training samples of 181633 due to the 50% generated replays from the forward dynamics model. To achieve the same performance as TRPO's max reward of 4.95, this model required only 29082 real samples (effective 58163), compared to TRPO's 42857 real samples, making it 1.5x more sample efficient in real samples and 0.7x in effective samples.

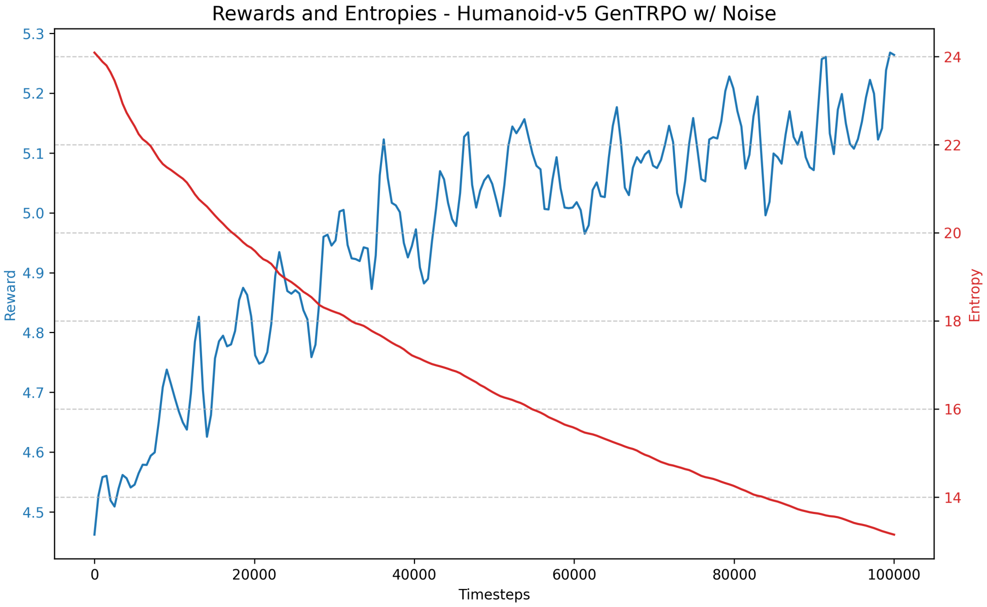

Fig. 1: Rewards and Entropies for GenTRPO w/ Noise in Humanoid-v5.

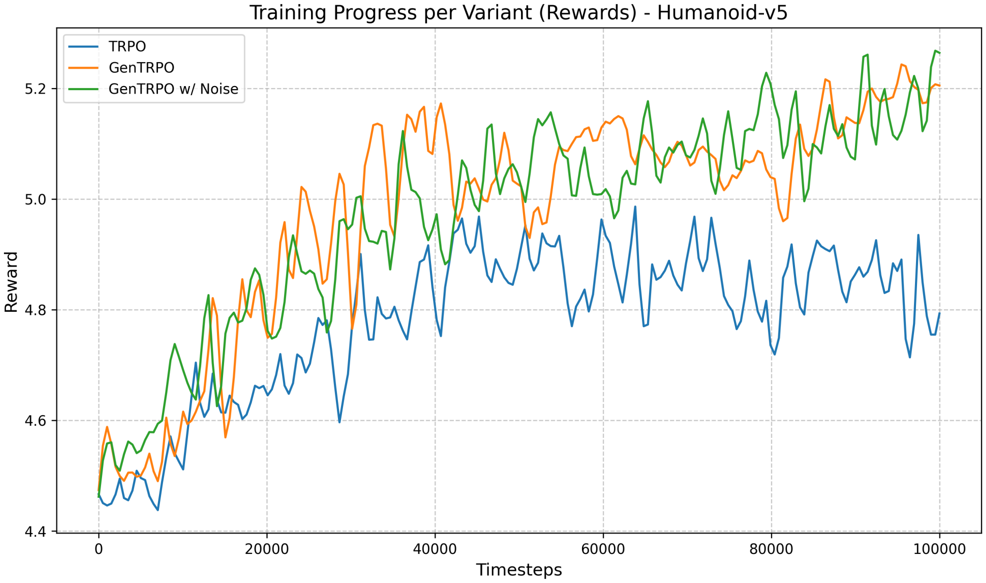

Fig. 2: Comparative Rewards across variants in Humanoid-v5.

4.2 HumanoidStandup-v5

The best performing model is GenTRPO with a maximum reward of 283.75 achieved at timestep 99400. The entropy exhibits a increasing trend, suggesting sustained exploration, potentially indicating ongoing adaptation or suboptimal convergence. The rate of change in entropy at the maximum reward point is -0.1475, indicating rapid policy adjustment. This model converges 35.5x faster than the TRPO baseline (first reaches or exceeds TRPO max at 1752 vs TRPO max at 62245 timesteps). It achieves a 230.2% higher maximum reward.

Cross-variant comparison: Compared to TRPO, the best model achieves 230.2% higher max reward and converges 0.6x faster. Compared to GenTRPO w/ Noise, the best model achieves 23.7% higher max reward and converges 0.6x faster.

In terms of sampling efficiency, to achieve its absolute max reward of 283.75 at timestep 99400, this model used 99400 real environment samples, with effective training samples of 198800 due to the 50% generated replays from the forward dynamics model. To achieve the same performance as TRPO's max reward of 85.94, this model required only 1752 real samples (effective 3504), compared to TRPO's 62245 real samples, making it 35.5x more sample efficient in real samples and 17.8x in effective samples.

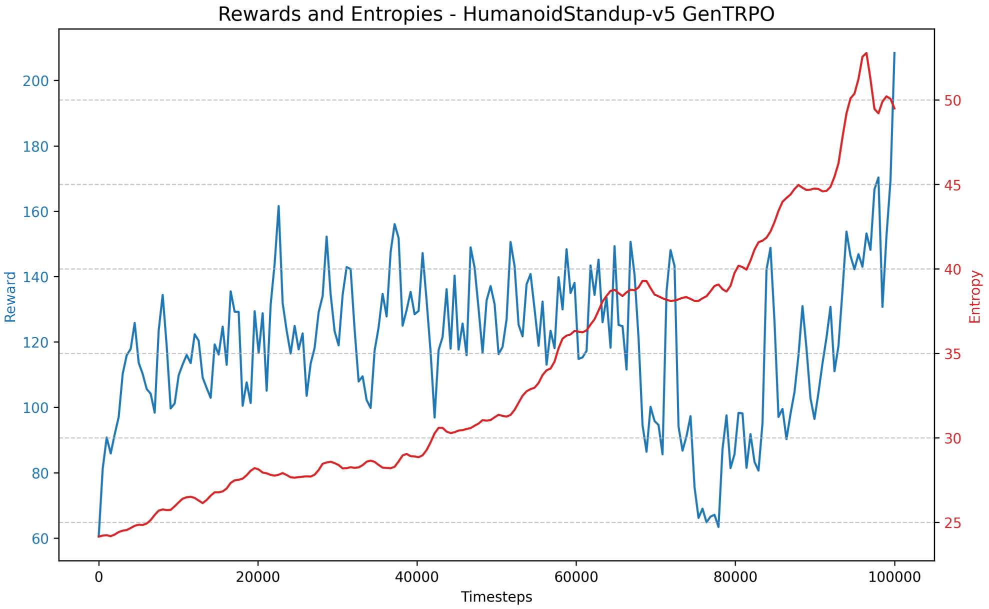

Fig. 3: Rewards and Entropies for GenTRPO in HumanoidStandup-v5.

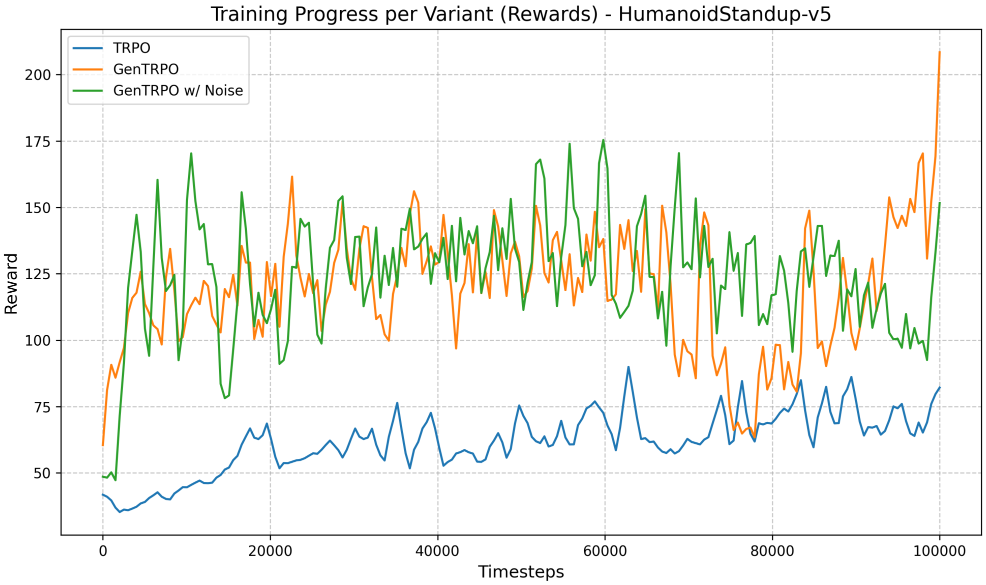

Fig. 4: Comparative Rewards across variants in HumanoidStandup-v5.

5 Conclusion

In summary, GenTRPO variants outperform the TRPO baseline in both environments, with notable gains in HumanoidStandup-v5. These improvements suggest that generalizations and noise aid in handling complex dynamics.