# 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 | Variant | Max Reward | Mean Reward (± std) | Timestep at Max |
|---|---|---|---|---|
| Humanoid-v5 | TRPO | 4.95 | $4.76 \pm 0.14$ | **42857** |
| Humanoid-v5 | GenTRPO (Noise=0) | 5.25 | $4.94 \pm 0.22$ | 95408 |
| Humanoid-v5 | GenTRPO | **5.29** | $4.92 \pm 0.20$ | 90816 |
| HumanoidStandup-v5 | TRPO | 85.94 | $60.15 \pm 11.13$ | 62245 |
| HumanoidStandup-v5 | GenTRPO (Noise=0) | **283.75** | $68.20 \pm 21.69$ | 99400 |
| HumanoidStandup-v5 | GenTRPO | 229.41 | $69.26 \pm 24.21$ | **59896** |

# 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 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.

The slope of the reward curve at the end of the run is 0.0099. The reward is mildly increasing at the end, indicating ongoing but gradual learning.

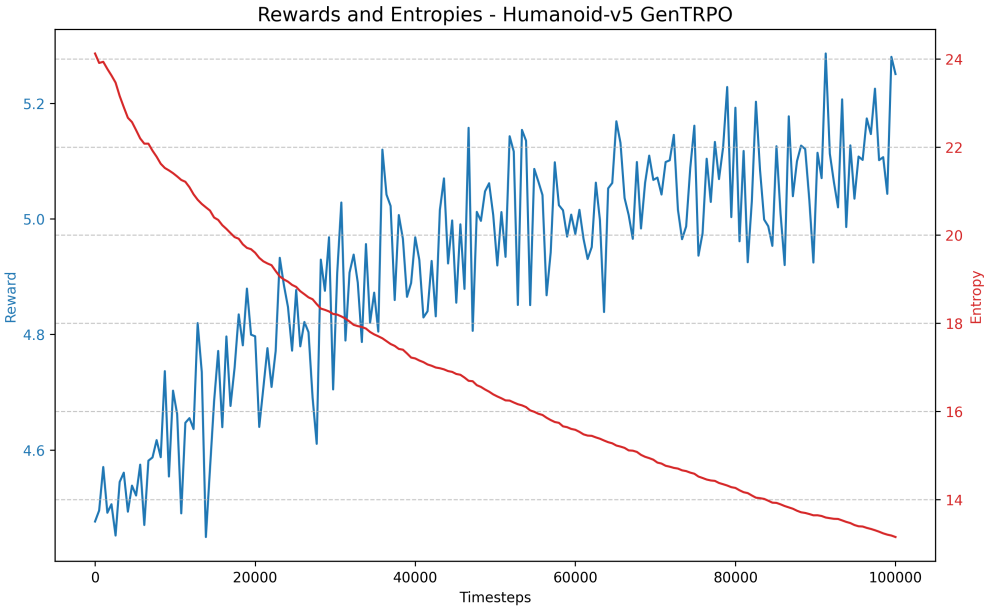

Fig. 1: Rewards and Entropies for GenTRPO in Humanoid-v5.

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 (Noise=0), the best model achieves 0.7% higher max reward and converges 1.1x faster.

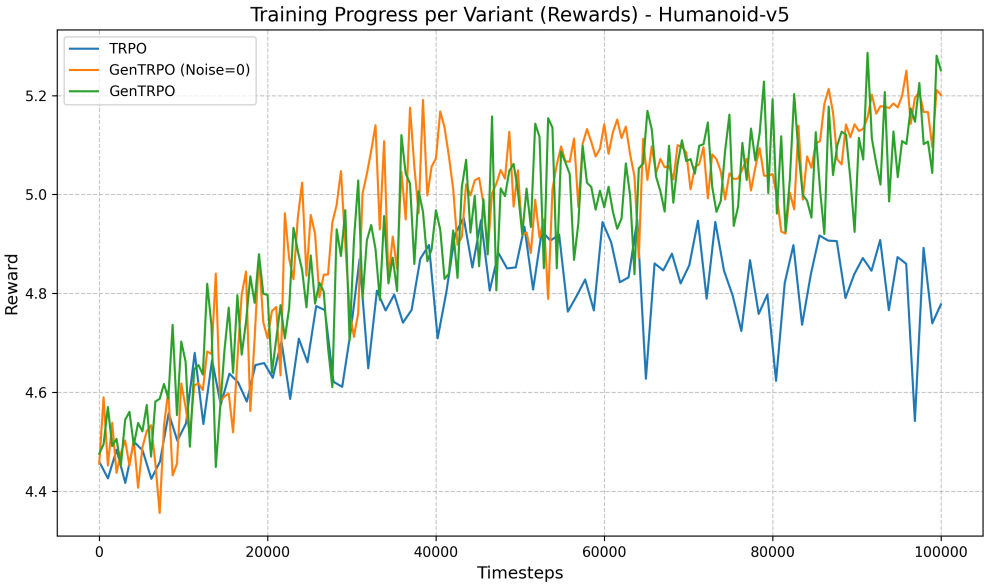


Fig. 2: Comparative Rewards across variants in Humanoid-v5.

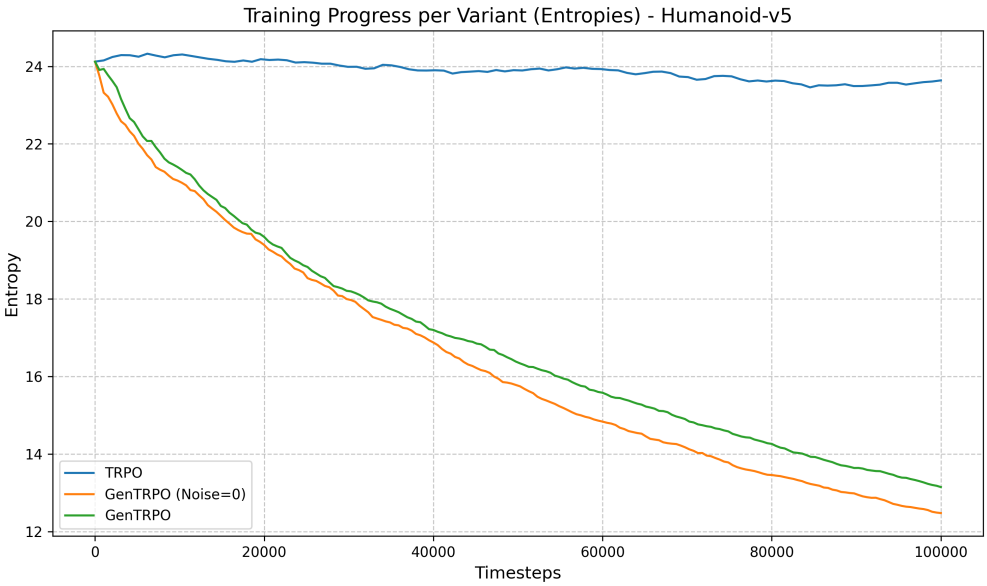


Fig. 3: Comparative Entropies across variants in Humanoid-v5.

In terms of sampling efficiency, to achieve its absolute max reward of 5.29 at timestep 90816, this model used 90816 real environment samples. To achieve the same performance as TRPO's max reward of 4.95, this model required only 29082 real samples, compared to TRPO's 42857 real samples, making it 1.5x more sample efficient in real samples.

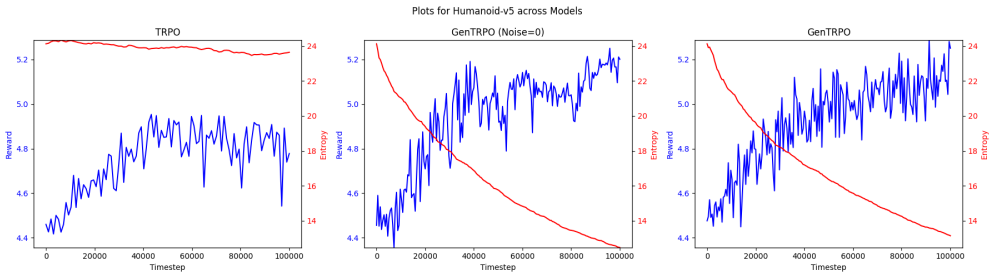


Fig. 4: Grid of plots for Humanoid-v5 across all models.

## 4.2 HumanoidStandup-v5

The best performing model is GenTRPO (Noise=0) 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.

The slope of the reward curve at the end of the run is 5.6054. This sharp upward trajectory suggests that the model is still improving and may achieve even higher performance with additional training timesteps.

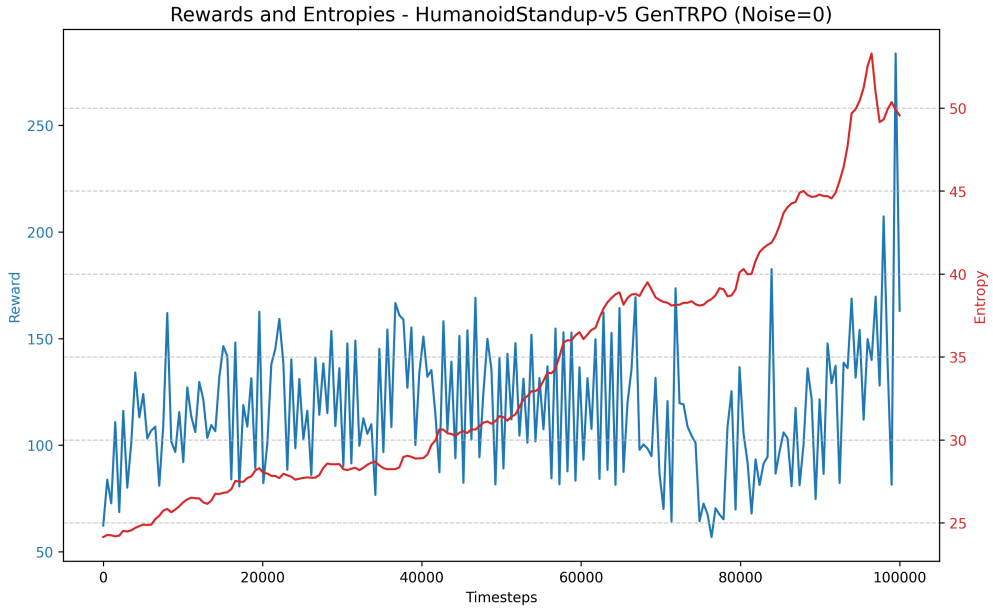


Fig. 5: Rewards and Entropies for GenTRPO (Noise=0) in HumanoidStandup-v5.

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, the best model achieves 23.7% higher max reward and converges 0.6x faster.

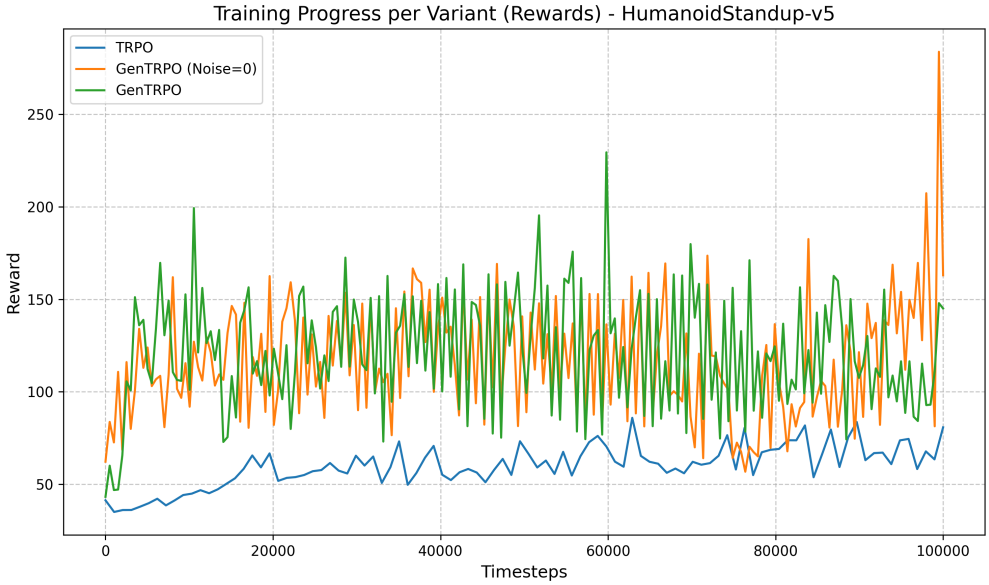


Fig. 6: Comparative Rewards across variants in HumanoidStandup-v5.

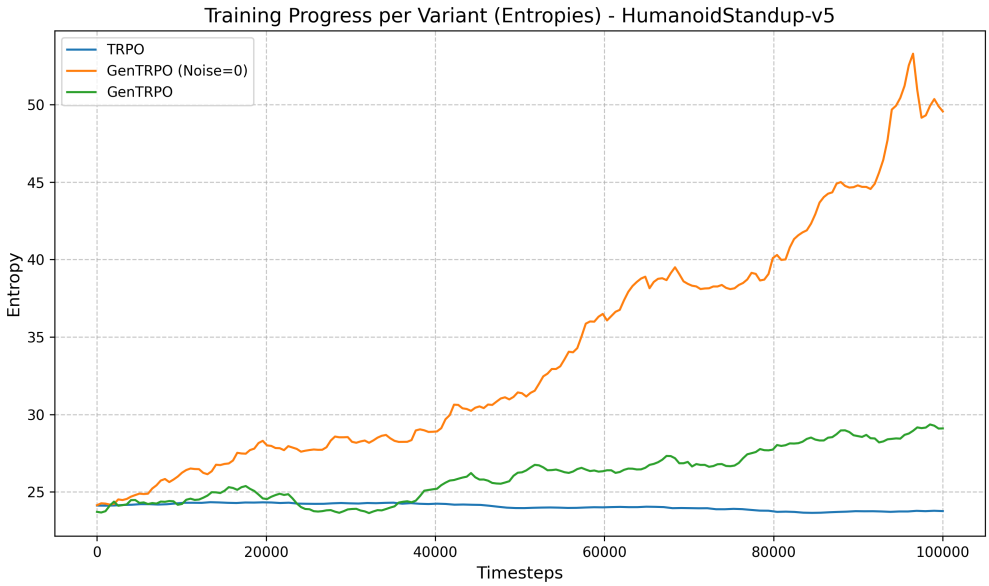


Fig. 7: Comparative Entropies across variants in HumanoidStandup-v5.

In terms of sampling efficiency, to achieve its absolute max reward of 283.75 at timestep 99400, this model used 99400 real environment samples. To achieve the same performance as TRPO's max reward of 85.94, this model required only 1752 real samples, compared to TRPO's 62245 real samples, making it 35.5x more sample efficient in real samples.

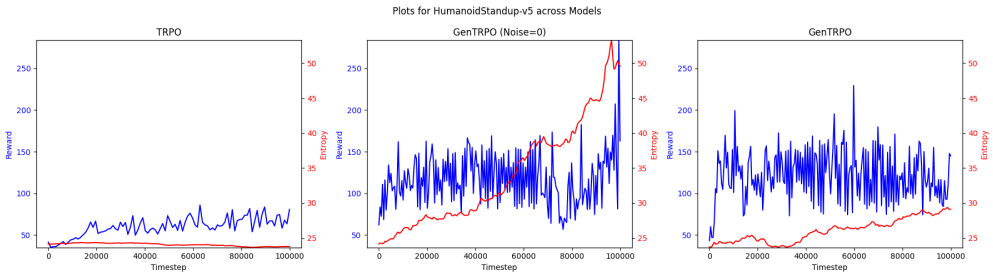


Fig. 8: Grid of plots for HumanoidStandup-v5 across all models.

# 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.

# References

1. Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017.
2. Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration, 2018.
3. Sahar Roostaie and Mohammad Mehdi Ebadzadeh. Entrpo: Trust region policy optimization method with entropy regularization, 2021.
4. John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017.
5. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
6. Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards, 2020.
7. Renhao Wang, Kevin Frans, Pieter Abbeel, Sergey Levine, and Alexei A. Efros. Prioritized generative replay, 2024.

# A Annex: Supplementary Plots

This annex provides supplementary plots for reference. Each plot is described below, focusing on its content and purpose.

## A.1 Individual Rewards and Entropies Plots

The following plots display the reward and entropy curves for individual model variants in each environment. These graphs illustrate the progression of rewards and entropies over training timesteps for a specific model and environment combination.

The plot for TRPO in Humanoid-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

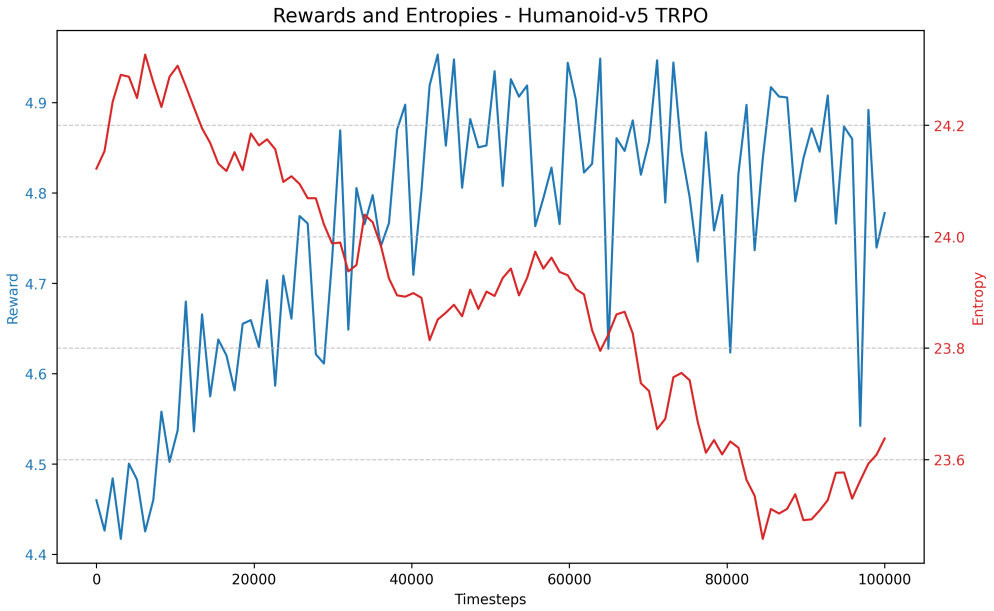


Fig. 9: Rewards and entropies over timesteps for TRPO in Humanoid-v5.

The plot for GenTRPO (Noise=0) in Humanoid-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

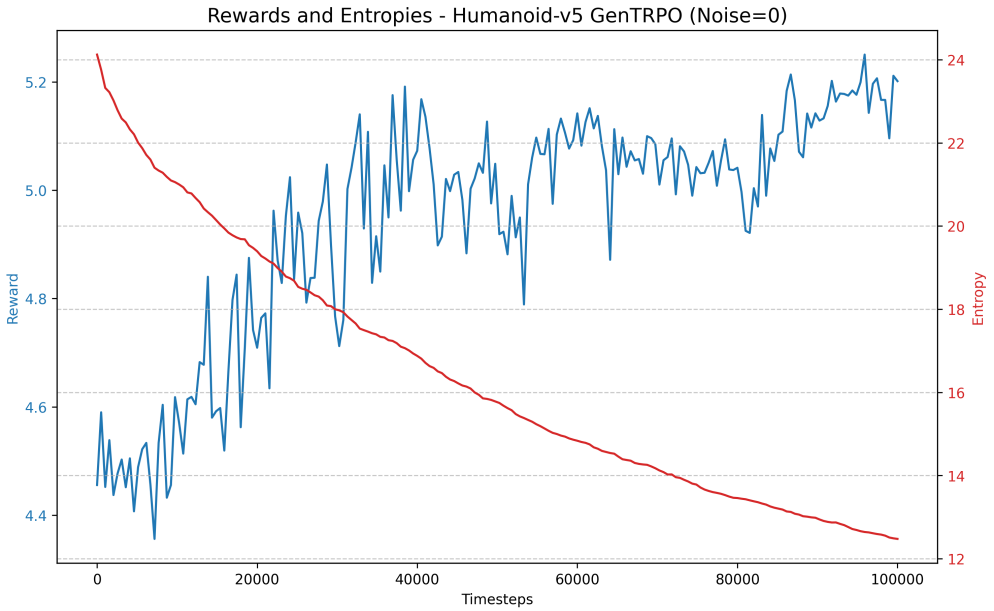


Fig. 10: Rewards and entropies over timesteps for GenTRPO (Noise=0) in Humanoid-v5.

The plot for GenTRPO in Humanoid-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

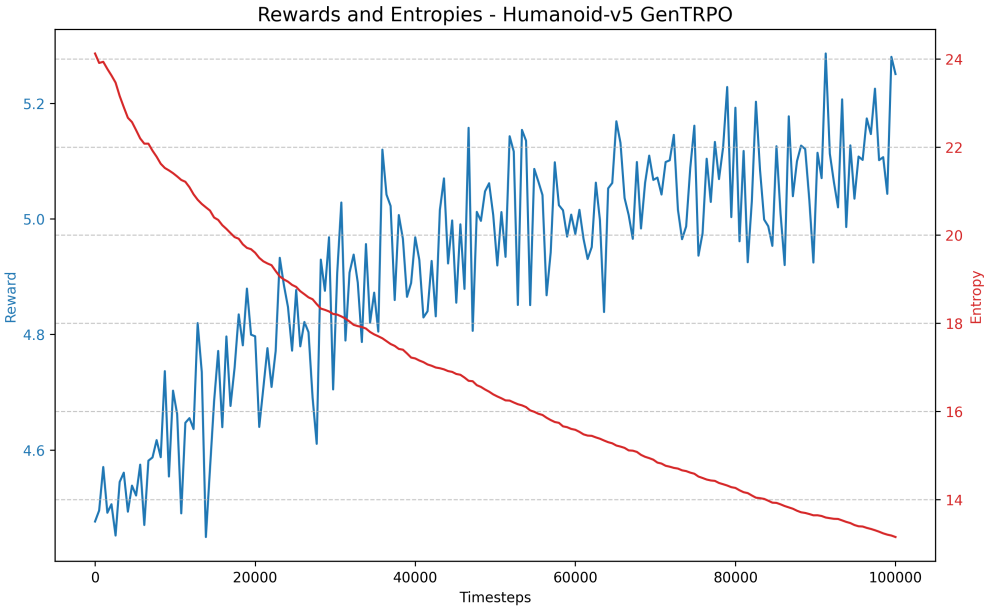


Fig. 11: Rewards and entropies over timesteps for GenTRPO in Humanoid-v5.

The plot for TRPO in HumanoidStandup-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

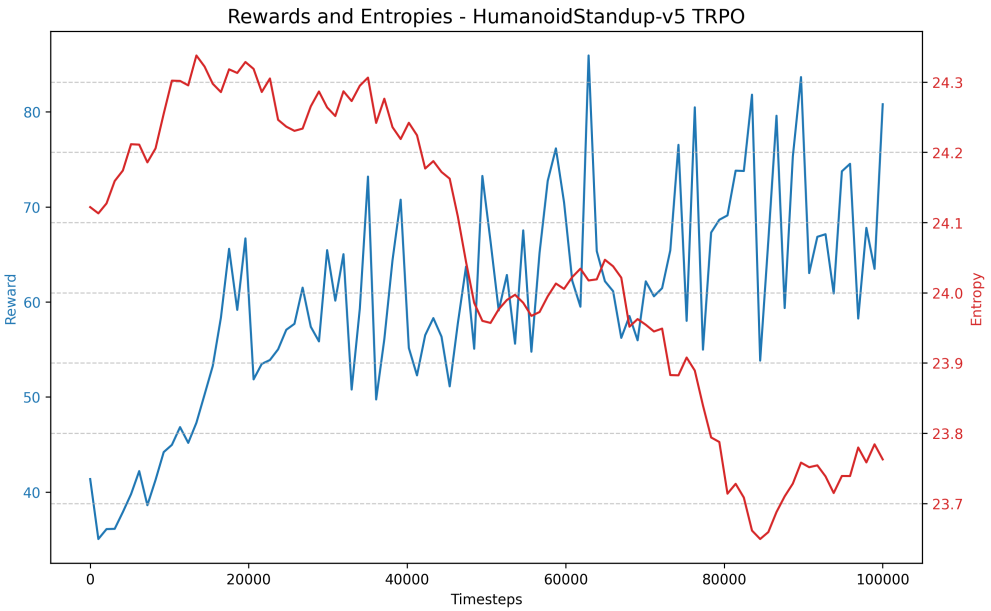


Fig. 12: Rewards and entropies over timesteps for TRPO in HumanoidStandup-v5.

The plot for GenTRPO (Noise=0) in HumanoidStandup-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

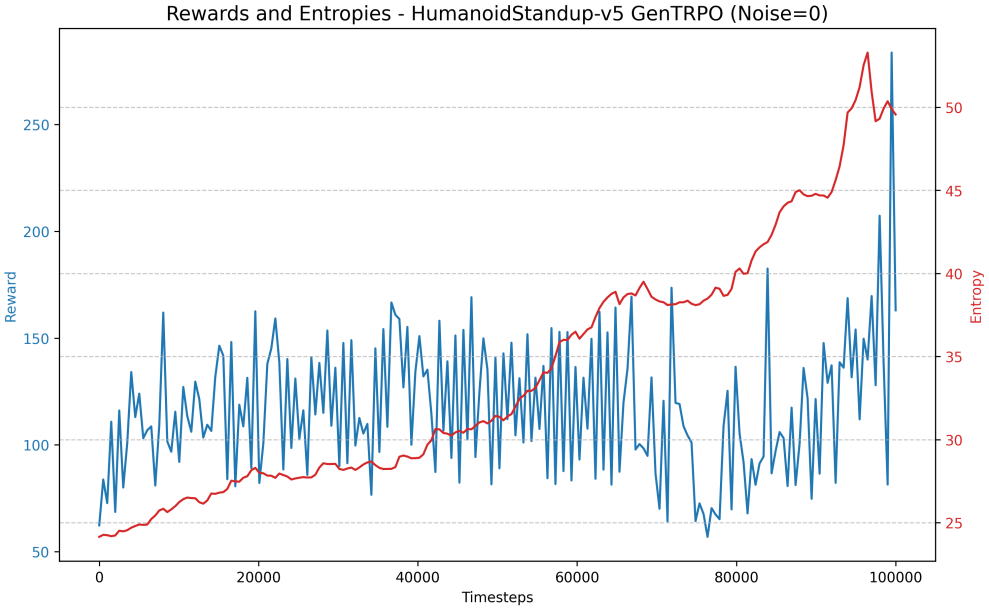


Fig. 13: Rewards and entropies over timesteps for GenTRPO (Noise=0) in HumanoidStandup-v5.

The plot for GenTRPO in HumanoidStandup-v5 shows the reward values (typically on one axis) and entropy values (on another axis) as functions of training timesteps.

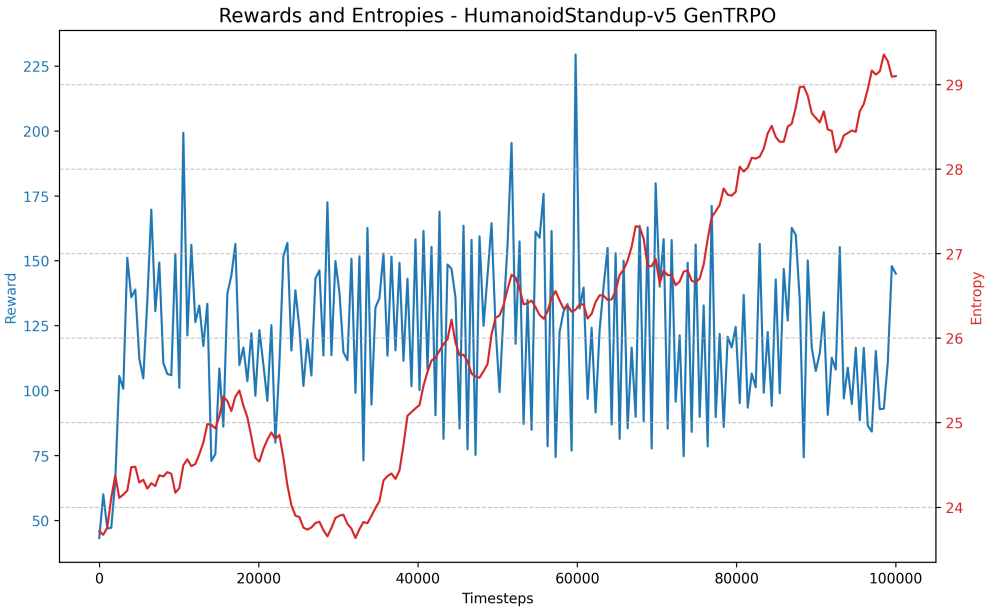


Fig. 14: Rewards and entropies over timesteps for GenTRPO in HumanoidStandup-v5.

## A.2 Comparative Rewards Plots

These plots compare the reward curves across all model variants for a specific environment. They allow for visual comparison of how different variants perform in terms of rewards over the training period.

The comparative rewards plot for Humanoid-v5 aggregates the reward curves from all variants, enabling side-by-side evaluation.

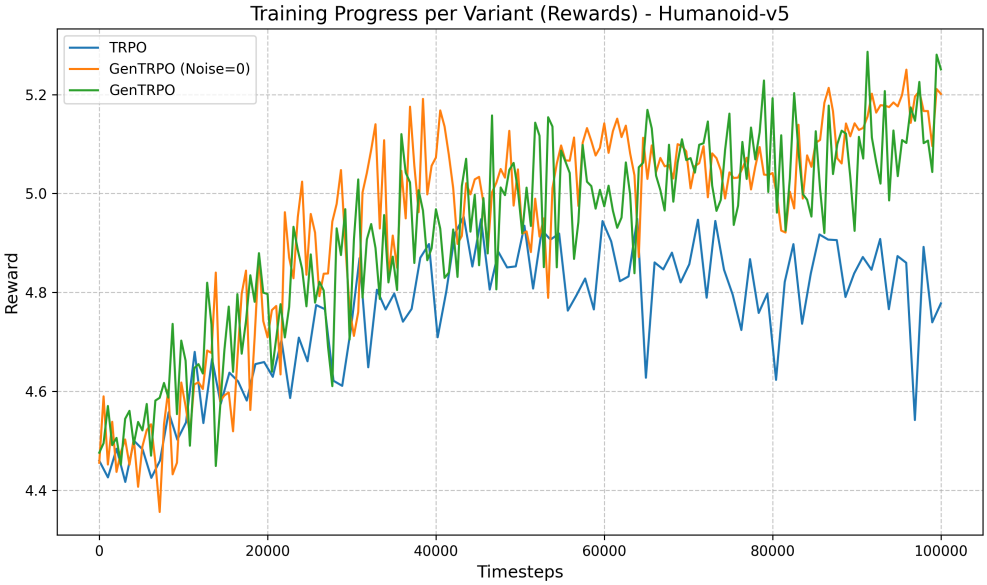


Fig. 15: Comparative rewards over timesteps across all variants in Humanoid-v5.

The comparative rewards plot for HumanoidStandup-v5 aggregates the reward curves from all variants, enabling side-by-side evaluation.

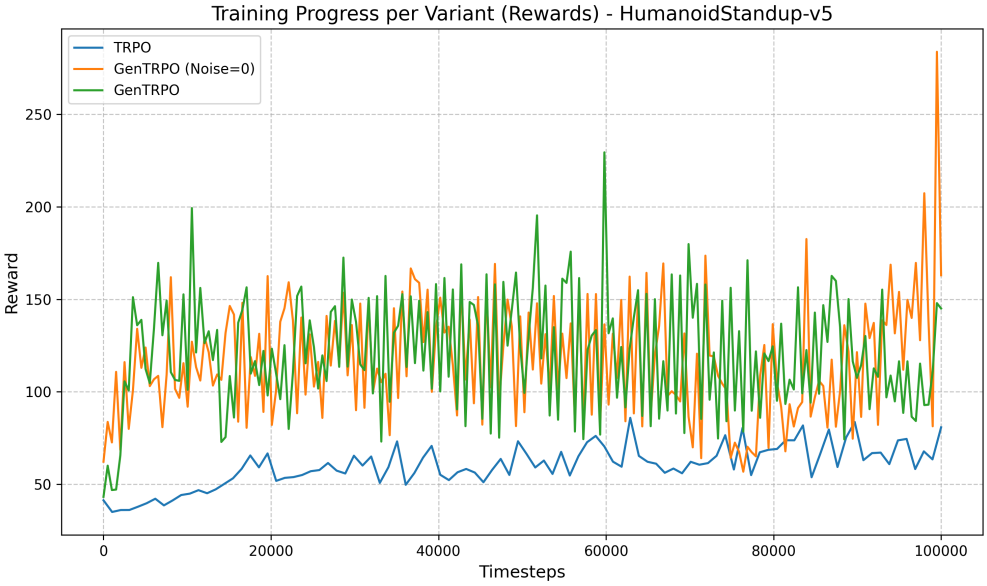


Fig. 16: Comparative rewards over timesteps across all variants in HumanoidStandup-v5.

## A.3 Comparative Entropies Plots

Similar to the comparative rewards, these plots show the entropy curves across all model variants for each environment, highlighting differences in exploration behavior.

The comparative entropies plot for Humanoid-v5 aggregates the entropy curves from all variants.

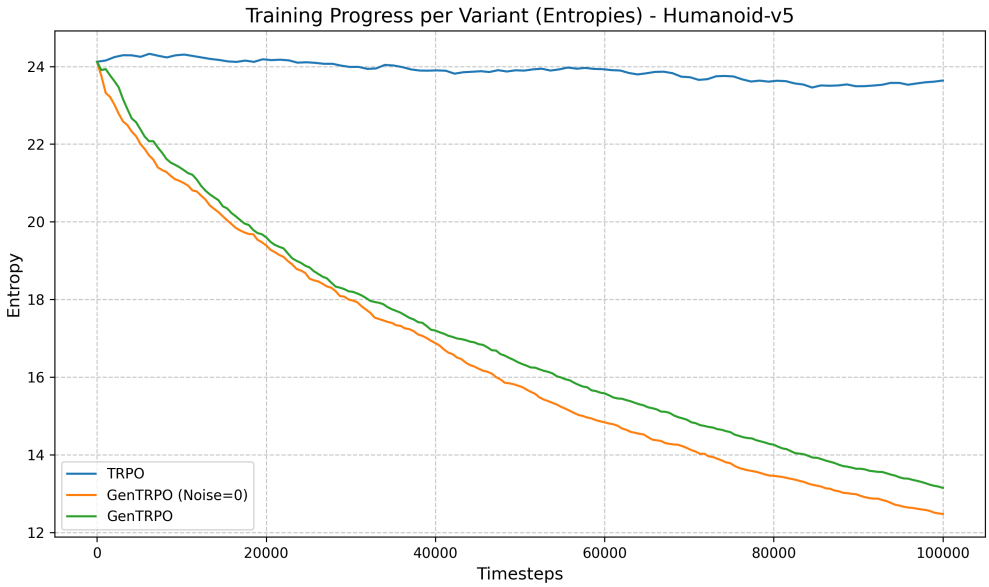


Fig. 17: Comparative entropies over timesteps across all variants in Humanoid-v5.

The comparative entropies plot for HumanoidStandup-v5 aggregates the entropy curves from all variants.

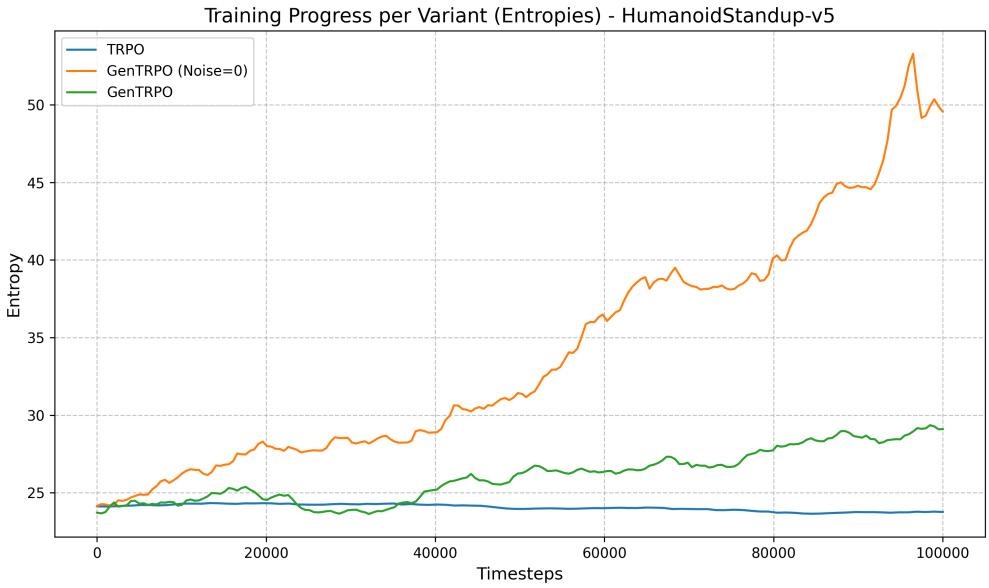


Fig. 18: Comparative entropies over timesteps across all variants in HumanoidStandup-v5.

## A.4 Grid Plots per Model

These grid plots compile the rewards and entropies for a single model across all environments, providing a consolidated view per model.

The grid plot for TRPO displays rewards and entropies across different environments in a grid format.

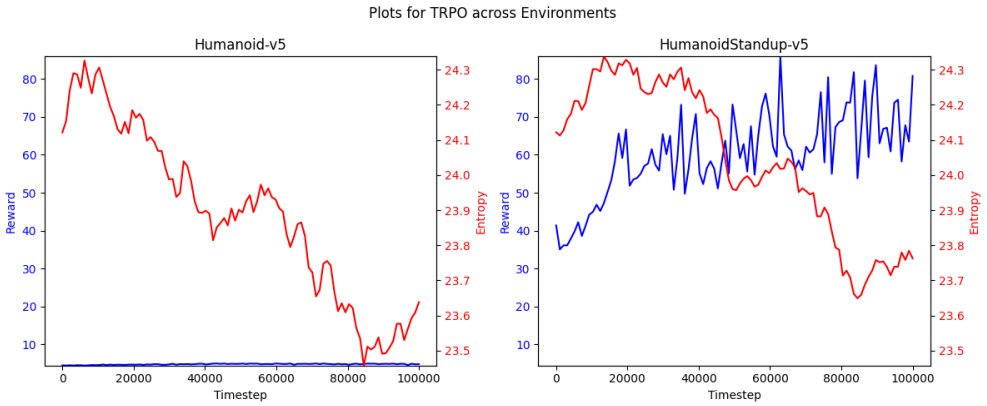


Fig. 19: Grid of plots for TRPO across all environments.

The grid plot for GenTRPO (Noise=0) displays rewards and entropies across different environments in a grid format.

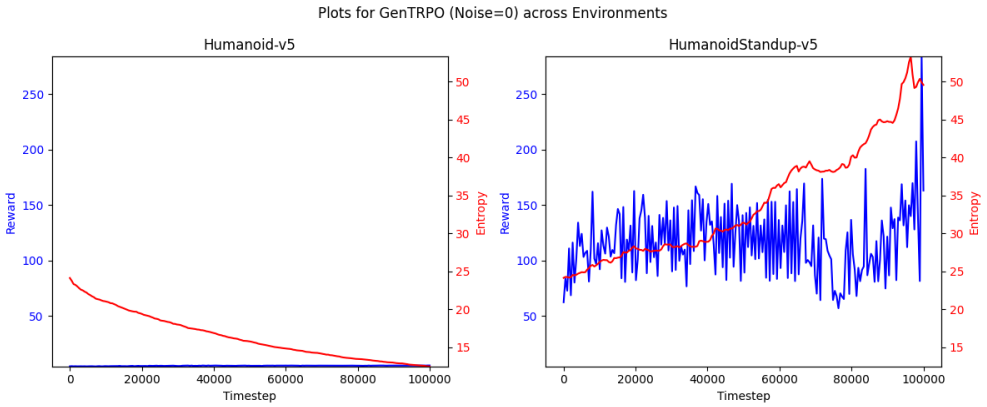


Fig. 20: Grid of plots for GenTRPO (Noise=0) across all environments.

The grid plot for GenTRPO displays rewards and entropies across different environments in a grid format.

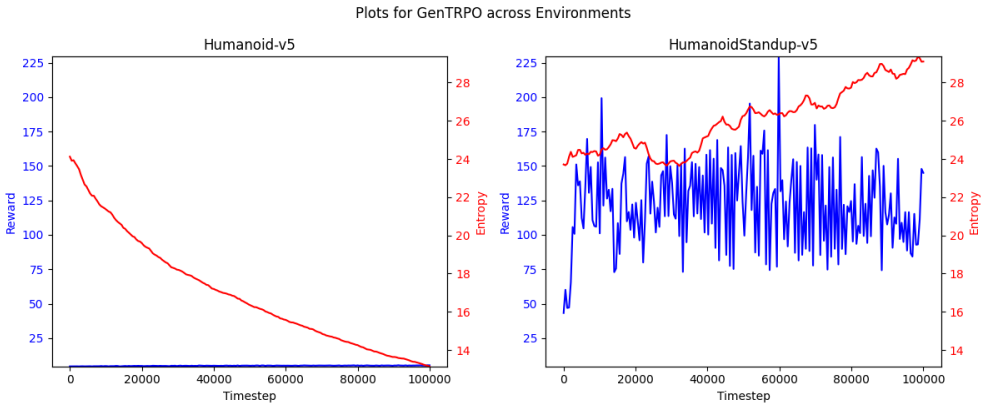


Fig. 21: Grid of plots for GenTRPO across all environments.

## A.5 Grid Plots per Environment

These grid plots compile the rewards and entropies for a single environment across all models, offering a per-environment overview.

The grid plot for Humanoid-v5 displays rewards and entropies across different models in a grid format.

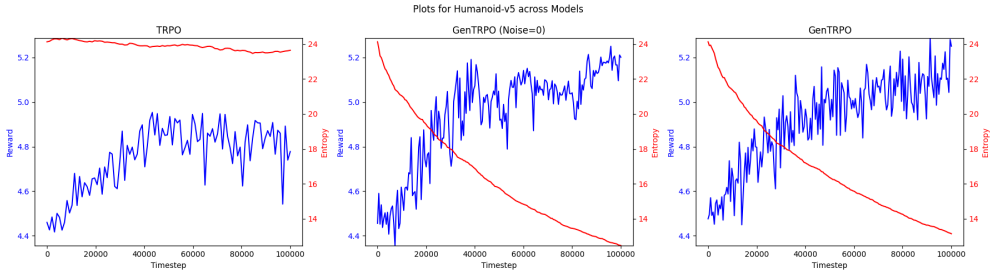


Fig. 22: Grid of plots for Humanoid-v5 across all models.

The grid plot for HumanoidStandup-v5 displays rewards and entropies across different models in a grid format.

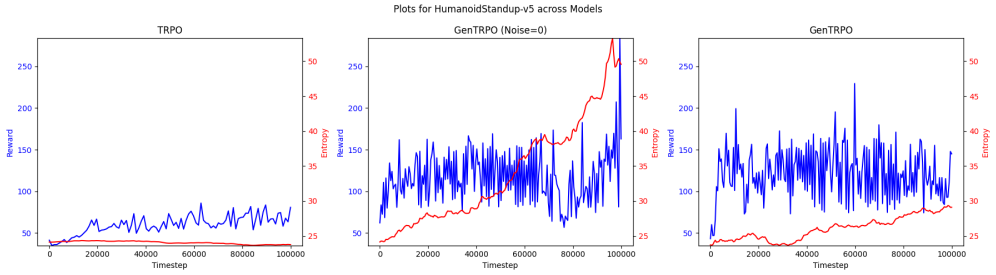


Fig. 23: Grid of plots for HumanoidStandup-v5 across all models.

## A.6 Detailed Metrics Table

The following table presents detailed metrics for all model-environment pairs, including maximum reward, mean reward with standard deviation, timestep at maximum reward, end slope of the reward curve, entropy trend, and entropy rate at maximum reward point.

Table 2: Detailed Metrics for All Model-Environment Pairs

| Environment | Variant | Max Reward | Mean Reward (± std) | Timestep at Max | End Slope | Entropy Trend | Entropy Rate at Max |
|---|---|---|---|---|---|---|---|
| Humanoid-v5 | TRPO | 4.95 | $4.76 \pm 0.14$ | 42857 | -0.0143 | decreasing | 0.0246 |
| Humanoid-v5 | GenTRPO (Noise=0) | 5.25 | $4.94 \pm 0.22$ | 95408 | -0.0038 | decreasing | -0.0147 |
| Humanoid-v5 | GenTRPO | 5.29 | $4.92 \pm 0.20$ | 90816 | 0.0099 | decreasing | -0.0250 |
| HumanoidStandup-v5 | TRPO | 85.94 | $60.15 \pm 11.13$ | 62245 | 0.8030 | decreasing | -0.0076 |
| HumanoidStandup-v5 | GenTRPO (Noise=0) | 283.75 | $68.20 \pm 21.69$ | 99400 | 5.6054 | increasing | -0.1475 |
| HumanoidStandup-v5 | GenTRPO | 229.41 | $69.26 \pm 24.21$ | 59896 | -1.9069 | increasing | -0.0008 |