Navigating the Flappy Bird Challenge with Reinforcement Learning: Insights from TD3 and PPO

Adrianus Jonathan Engelbracht, Alexander Hartung, Tabea Runzheimer

***Abstract*— This study explores the application of reinforcement learning algorithms —TD3 and PPO— in mastering the dynamics of Flappy Bird. Using extensive hyperparameter trials, we optimized TD3 and PPO to enhance agent performance. These insights highlight the necessity of hyperparameter tuning for improved navigation and performance. Our comparative analysis provides key insights into adapting RL frameworks for dynamic environments.**

I. INTRODUCTION

In recent years, the field of reinforcement learning (RL) has gained considerable attention due to its promising potential for teaching agents to perform complex tasks through trial and error. Among the many environments used to train RL agents, games serve as a particularly popular testbed due to their well-defined rules and challenging dynamics. One such game that has captured the interest of the research community is Flappy Bird. Despite its seemingly simple mechanics, mastering Flappy Bird requires sophisticated strategies to navigate through an unpredictable series of obstacles. In this work, we explore the application of two reinforcement learning algorithms - Twin Delayed Deep Deterministic Policy Gradient (TD3) and Proximal Policy Optimization (PPO) - to mastering the Flappy Bird environment.

II. METHODOLOGY

Our research focuses on using an established Flappy Bird library [1] to train agents using different RL algorithms. The aim is to explore and understand the intricacies of different RL strategies in this complex environment. We took the following approach:

1) TD3 and PPO Hyperparameter Trials using stable-baselines3 PPO [2]
2) Development of a custom PPO agent
3) Comparative Analysis

The primary phase of the investigation concentrated on the TD3 and PPO algorithm developed using the stable-baselines3 library [2]. A series of comprehensive trials was conducted to explore a diverse range of hyperparameter configurations. This process involved systematic experimentation with various parameters such as learning rates, batch sizes, and exploration strategies. The primary objective of this stage of the investigation was to make refinements to the performance of the algorithm, with a view to optimizing the agents' capacity to negotiate the Flappy Bird environment with proficiency. In this process, we experimented systematically with four key hyperparameters. First, we varied the number of timesteps used for training the agents, testing values of 50.000, 100.000, 200.000, and 500.000. This allowed us to assess the impact of different training durations on the agents' learning and performance. Next, we examined different learning rates by exploring values of 0.0001, 0.001, 0.005, and 0.01. This experimentation aimed to determine the influence of the rate at which the optimizer updates the policy on the algorithm's convergence efficiency and the quality of the learned strategy. We also considered the batch size, trying out sizes of 32, 64, 128, and 256. By doing so, we aimed to understand how the volume of data processed at each update affected the stability and accuracy of the learning process. Lastly, we varied the discount factor (gamma) with values of 0.95, 0.98, and 0.99. This helped us explore how the algorithm weighs future rewards against immediate outcomes, impacting long-term agent performance. Each possible combination of these hyperparameters was tested to comprehensively understand their effects on both algorithm's performance.

Following the preliminary evaluation of TD3 and PPO using stable-baselines3, the subsequent phase of the investigation centred on the development of a custom implementation of the PPO algorithm. The primary motivation behind this endeavour was to gain finer control over the learning process, allowing for deeper insights into the training dynamics and potential optimisations beyond the existing implementations. Throughout the development process, extensive debugging and performance validation were conducted to ensure correct implementation. This process entailed verifying loss values, monitoring gradients, and assessing the behaviour of policy updates in accordance with expectations. Subsequent to the implementation of the core PPO framework, the custom agent was trained on the Flappy Bird environment, and a comparison was made between its performance and that of the stable-baselines3 implementations.

The final step in the research was a comprehensive comparative analysis of all three solutions: the stable baselines TD3 and PPO, and the custom PPO. The objective of this analysis was to evaluate their respective strengths and weaknesses within the complex dynamics of the Flappy Bird game.

Authors are listed in alphabetical order and contributed equally to this work.
Fulda University of Applied Sciences, Fulda, Germany. Emails: adrianus-jonathan.engelbrecht@ai.hs-fulda.de, alexander.hartung@ai.hs-fulda.de, tabea.runzheimer@ai.hs-fulda.de

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma	Avg. Passed Pipes
714.8	713.87	322.75	1.0	0.15	200000	0.001	64	0.98	134
317.21	315.18	168.11	1.0	0.27	500000	0.0001	32	0.98	84
381.55	385.04	241.98	1.0	0.36	500000	0.001	32	0.95	76
276.22	275.44	147.54	1.0	0.18	500000	0.001	128	0.95	38
116.82	116.89	4.11	1.0	0.60	100000	0.001	256	0.95	6.25

TABLE I
TOP 5 MODELS USING TD3 WITH STABLE-BASLINES3

III. RESULTS

A. TD3 Hyperparameter Trials

In the present investigation, the performance of a range of models trained using different hyperparameter configurations was evaluated. The top five models of TD3 shown in Table I exhibited distinct characteristics in terms of average reward, convergence speed, stability, sample efficiency, and the balance between exploration and exploitation. The full results can be seen in Table IV.

The present investigation examined various models trained with different hyperparameter configurations, detailed in Table I. The model that demonstrated the highest level of performance achieved an average reward of 714.8, as illustrated in the first row of the table, with an impressive convergence speed of approximately 713.9. The model exhibited remarkable stability, with a score of 322.75, and demonstrated high sample efficiency. The model's exploration versus exploitation metric was 0.1488, indicating a well-balanced strategy between trying novel approaches and optimizing known successful ones.

As illustrated in the subsequent rows of Table I, variations in hyperparameters such as learning rates, number of steps, and batch sizes manifested nuanced impacts on model performance. It is noteworthy that the majority of models maintained a sample efficiency score of 1.0, indicating effective utilisation of training data. However, variations in the exploration vs. exploitation metric suggested different strategic advantages in handling the Flappy Bird environment's complexity. It is also important to note that a reward of approximately 101 indicates that the Flappy Bird agent was unable to pass the first pipe, highlighting a critical failure point in certain configurations and reinforcing the necessity of optimizing hyperparameters for improved navigation and performance.

The hyperparameter trials for the TD3 algorithm revealed that configurations with 200,000 training steps generally produced the highest average rewards. Furthermore, the most effective learning rate and batch size across the top models was determined to be 0.001 and 64, respectively. Among the gamma values that were tested, 0.98 was found to be the optimal setting, as it resulted in the highest reward being achieved by the model. While all models exhibited perfect sample efficiency (1.0), the exploration-exploitation balance demonstrated variability, with the most optimal models exhibiting values ranging from 0.15 to 0.18.

B. PPO Hyperparameter Trials

In the present investigation, the performance of a range of PPO models trained using different hyperparameter configurations was evaluated. The top five models of PPO, as illustrated in Table II, showed distinct characteristics in terms of average reward, convergence speed, stability, sample efficiency, and balance between exploration and exploitation. The complete set of results can be found in Table V.

As demonstrated in Table II, notable performance variations were exhibited by the models trained with the PPO algorithm, depending on the hyperparameter settings. The model that demonstrated the highest level of performance achieved an average reward of 911.99, with a convergence speed of 909.81, as can be seen in the first row of the table. The model also demonstrated stability, with a score of 206.99, and exhibited perfect sample efficiency (1.0). The exploration vs. exploitation value of 0.09 suggested a well-balanced strategy, favoring exploitation over exploration.

As demonstrated in the subsequent rows of Table 1, adjustments to hyperparameters such as learning rates, batch sizes, and gamma values had subtle yet significant effects on performance. The top five models demonstrated high stability and sample efficiency, suggesting that the training processes were well-tuned. The average reward across the top five models ranged from 870.95 to 911.99, suggesting the potential for further optimization within these configurations. The findings underscore the significance of meticulously calibrating hyperparameters to attain a harmonious equilibrium between the maximization of reward and the efficient exploration of the search space.

The key insights from the PPO hyperparameter trials indicate that configurations with 100,000 steps consistently outperformed those with larger step counts in terms of average reward. Among the various learning rates tested, a value of 0.001 proved to be more effective than higher values like 0.005. Additionally, higher gamma values (0.99) were associated with the best average rewards, though they did not always lead to the highest stability. Regarding batch size, a batch size of 128 appeared to be the most effective across different configurations.

C. Custom PPO Agent

As a result of the development process of our custom PPO agent, it was observed that, after completing approximately 30,000 steps, the agent began to unlearn, resulting in a decrease in reward values to zero. This pattern prompted

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma	Avg. Passed Pipes
911.99	909.81	206.99	1.00	0.09	100000	0.001	64	0.99	22.51
891.89	900.14	239.85	1.00	0.08	500000	0.0001	128	0.95	21.94
884.68	885.74	229.68	1.00	0.08	200000	0.001	128	0.98	21.76
883.38	880.38	227.16	1.00	0.09	500000	0.0001	64	0.95	21.68
870.95	876.33	228.91	1.00	0.09	100000	0.005	32	0.98	21.36

TABLE II
TOP 5 MODELS USING PPO WITH STABLE-BASLINES3

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma	Avg. Passed Pipes
1635.96	1646.50	1503.63	1.0	0.11	30000	0,00003	128	0.9999	41.61
626.73	624.63	592.16	1.0	0.10	30000	0,00001	64	0.9999	14.38
406.06	403.92	383.41	1.0	0.10	30000	0,00001	32	0.95	8.39
138.5	139.16	49.52	1.0	0.10	30000	0,00001	256	0.9999	1.19
117.79	117.96	24.23	1.0	0.11	30000	0,00001	128	0.98	0.54

TABLE III
TOP 5 MODELS USING CUSTOM PPO

a deviation from the hyperparameters previously utilised in stable-baselines3 implementations of TD3 and PPO, with the objective of addressing this issue and optimising performance. Table III summarizes the top 5 results we obtained from our experiments.

With an increased learning rate of 0.00003, the agent achieved the highest average reward of 1635.96, indicating a positive impact on convergence speed and stability. Conversely, larger batch sizes, such as 256, resulted in lower performance metrics, highlighting a trade-off between computational efficiency and the quality of learning. The gamma value of 0.9999 was found to be consistently superior to lower values, emphasising the significance of this parameter in attaining a higher cumulative reward through enhanced future reward consideration.

The full results can be found in the Appendix Table VI.

D. Comparative Analysis

A comparison of the top-performing models from TD3 and PPO reveals significant variations in their hyperparameter configurations and performance characteristics. It was evident that both algorithms demonstrated optimal efficacy at 100,000 to 200,000 training steps, with a learning rate of 0.001 being a prevalent feature across the leading trials. However, a notable distinction emerged in the batch size employed by the two algorithms. While PPO models demonstrated a tendency to utilise batch sizes of 128, TD3 predominantly adopted a smaller batch size of 64. This observation suggests that TD3 and PPO may employ divergent strategies to ensure training stability.

In relation to gamma, the PPO models demonstrated optimal performance with a higher gamma of 0.99, indicating a pronounced emphasis on long-term rewards. In contrast, the TD3 models exhibited superior performance with a slightly lower gamma of 0.98, suggesting a prioritisation of immediate reward. The exploration vs. exploitation parameter exhibited a mean value of approximately 0.09 in the PPO model, indicative of a marginally more exploratory behaviour in comparison to the TD3 model, which registered values in the range of 0.15 to 0.18.

In the context of the study, both algorithms exhibited perfect sample efficiency (1.0) in their top models, thereby emphasising the efficient use of training data. A comparative analysis revealed that PPO exhibited superior stability and reward in the top results, a phenomenon that may be attributed to its higher gamma and larger batch size, while TD3 showed a more consistent performance due to its different balance between exploration and exploitation.

A notable observation is the high variance in the number of pipes passed within the top TD3 trials. The top TD3 model demonstrated a remarkable ability to pass 134 pipes, while the PPO models exhibited a significantly lower average, with the highest achieving 22.51. This finding suggests that TD3 may possess enhanced navigation and obstacle avoidance capabilities within the environment. In contrast, PPO, despite its higher stability and reward, appears to encounter greater challenges in navigating the environment effectively.

The custom PPO agent exhibited superior performance with an average reward of 1635.96, which is notably higher than the stable-baselines3 PPO (911.99) and TD3 (714.8). This finding indicates that the custom hyperparameter optimization significantly enhanced the agent's capacity to attain higher rewards within a relatively brief timeframe of 30,000 steps.

A pivotal factor contributing to the efficacy of the custom PPO was the optimisation of its hyperparameters, which included a reduced learning rate (0.00003) and a gamma value of 0.9999. This configuration fostered both the accumulation of high rewards and the preservation of stability. In contrast, the stable-baselines3 implementations, which employed higher learning rates, may have encountered difficulties in ensuring long-term policy stability.

The exploration vs. exploitation balance was consistent across all agents, with the custom PPO effectively managing minor exploration adjustments to sustain high performance.

IV. CONCLUSION

In conclusion, our exploration of hyperparameter adjustments for a custom PPO agent in the context of Flappy Bird yielded enhancements over standard TD3 and PPO implementations from stable-baselines3. While the custom PPO demonstrated enhanced average rewards and training efficiency, these results must be interpreted with the understanding that they are early indicators rather than definitive advancements. The modest gains in performance highlight the importance of further investigation into hyperparameter tuning and its effect on learning dynamics. It is recommended that future research efforts concentrate on conducting more extensive testing and validation in a range of environments. This will allow for a more comprehensive understanding and validation of these initial observations.

REFERENCES

[1] *Flappy-bird-gym: An OpenAI gym environment for the Flappy Bird game.* [Online]. Available: https://github.com/Talendar/flappy-bird-gym (visited on 03/25/2025).

[2] *Stable-baselines3: Pytorch version of Stable Baselines, implementations of reinforcement learning algorithms.* [Online]. Available: https://github.com/DLR-RM/stable-baselines3 (visited on 03/26/2025).

[3] *Adrianus252/RL_project_flappybird_group42*. [Online]. Available: https://github.com/Adrianus252/RL_Project_FlappyBird_Group42 (visited on 03/31/2025).

[4] *Overview - RL Flappy Bird Group 42 - Obsidian Publish*, en. [Online]. Available: https://publish.obsidian.md/rl-flappybird-group42/1.+Introduction/Overview (visited on 03/31/2025).

APPENDIX

TABLE IV: Full Results of Training and Testing with TD3

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.23954816	100000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.0005685458	200000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.0026012054	50000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.19642796	100000	0.0001	256	0.98
101.0	101.0	0.0	1.0	0.17876239	200000	0.0001	256	0.98
101.0	101.0	0.0	1.0	0.19400172	50000	0.0001	256	0.98
101.0	101.0	0.0	1.0	0.23392361	100000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.29140437	200000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.21453488	50000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.37993836	100000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.058753524	50000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.068148255	100000	0.001	32	0.98
101.0	101.0	0.0	1.0	0.0	200000	0.001	32	0.98
32.0	32.0	0.0	1.0	0.0	50000	0.001	32	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.001	32	0.99
32.0	32.0	0.0	1.0	0.0	50000	0.001	32	0.99
117.18	117.19166666666668	0.8646386528486913	1.0	0.2074504	100000	0.001	64	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.001	64	0.95
101.0	101.0	0.0	1.0	1.1256122e-17	50000	0.001	64	0.95
101.0	101.0	0.0	1.0	0.42841285	100000	0.001	64	0.98
714.8	713.8687500000001	322.7518241621571	1.0	0.14884533	200000	0.001	64	0.98
32.0	32.0	0.0	1.0	0.0	50000	0.001	64	0.98
101.0	101.0	0.0	1.0	0.0049105044	100000	0.001	64	0.99
32.0	32.0	0.0	1.0	0.0	200000	0.001	64	0.99
102.51	102.51041666666669	1.431747184386964	1.0	0.5801723	50000	0.001	64	0.99
101.0	101.0	0.0	1.0	0.0	100000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.014882878	50000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.29360515	100000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.15584816	200000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.44449162	50000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.0021266676	100000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.0025465796	200000	0.0001	32	0.98

Continued on next page

TABLE IV: (Continued)

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.046380397	50000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.001	128	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.001	128	0.99
101.0	101.0	0.0	1.0	0.0	50000	0.001	128	0.99
116.82	116.88541666666669	4.109452518280264	1.0	0.5951551	100000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.5767296	200000	0.001	256	0.95
32.0	32.0	0.0	1.0	0.0	50000	0.001	256	0.95
106.64	106.62499999999999	6.766860424155355	1.0	0.5454867	100000	0.001	256	0.98
32.0	32.0	0.0	1.0	0.0	200000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.0	50000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.025589874	100000	0.001	256	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.001	256	0.99
101.0	101.0	0.0	1.0	0.43733478	50000	0.001	256	0.99
101.0	101.0	0.0	1.0	0.0	100000	0.005	32	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.005	32	0.95
101.0	101.0	0.0	1.0	0.0	50000	0.005	32	0.95
32.0	32.0	0.0	1.0	0.0	100000	0.005	32	0.98
32.0	32.0	0.0	1.0	0.0	200000	0.005	32	0.98
101.0	101.0	0.0	1.0	0.0	50000	0.005	32	0.98
32.0	32.0	0.0	1.0	0.0	100000	0.005	32	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.005	32	0.99
101.0	101.0	0.0	1.0	0.0	50000	0.005	32	0.99
32.0	32.0	0.0	1.0	0.0	100000	0.005	64	0.95
32.0	32.0	0.0	1.0	0.0	200000	0.005	64	0.95
32.0	32.0	0.0	1.0	0.0	50000	0.005	64	0.95
32.0	32.0	0.0	1.0	0.0	100000	0.005	64	0.98
32.0	32.0	0.0	1.0	0.0	200000	0.005	64	0.98
101.0	101.0	0.0	1.0	0.0	50000	0.005	64	0.98
32.0	32.0	0.0	1.0	0.0	100000	0.005	64	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.005	64	0.99
101.0	101.0	0.0	1.0	0.0	50000	0.005	64	0.99
101.0	101.0	0.0	1.0	0.0014096616	100000	0.0001	32	0.99
101.0	101.0	0.0	1.0	0.058297433	200000	0.0001	32	0.99
101.0	101.0	0.0	1.0	0.34099022	50000	0.0001	32	0.99
32.0	32.0	0.0	1.0	0.0	100000	0.005	128	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.005	128	0.95
32.0	32.0	0.0	1.0	0.0	50000	0.005	128	0.95

Continued on next page

TABLE IV: (Continued)

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.0	100000	0.005	128	0.98
32.0	32.0	0.0	1.0	0.0	200000	0.005	128	0.98
32.0	32.0	0.0	1.0	0.0	50000	0.005	128	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.005	128	0.99
32.0	32.0	0.0	1.0	0.0	200000	0.005	128	0.99
101.0	101.0	0.0	1.0	0.0	50000	0.005	128	0.99
32.0	32.0	0.0	1.0	0.0	100000	0.005	256	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.005	256	0.95
101.0	101.0	0.0	1.0	0.0	50000	0.005	256	0.95
101.0	101.0	0.0	1.0	0.0	100000	0.005	256	0.98
101.0	101.0	0.0	1.0	0.0	200000	0.005	256	0.98
32.0	32.0	0.0	1.0	0.0	50000	0.005	256	0.98
32.0	32.0	0.0	1.0	0.0	100000	0.005	256	0.99
32.0	32.0	0.0	1.0	0.0	200000	0.005	256	0.99
101.0	101.0	0.0	1.0	0.0	50000	0.005	256	0.99
32.0	32.0	0.0	1.0	0.0	100000	0.01	32	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.01	32	0.95
32.0	32.0	0.0	1.0	0.0	50000	0.01	32	0.95
32.0	32.0	0.0	1.0	0.0	100000	0.01	32	0.98
101.0	101.0	0.0	1.0	0.0	200000	0.01	32	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.01	32	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.01	32	0.99
32.0	32.0	0.0	1.0	0.0	100000	0.01	64	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.01	64	0.95
101.0	101.0	0.0	1.0	0.0041162176	100000	0.0001	64	0.95
101.0	101.0	0.0	1.0	0.21682757	200000	0.0001	64	0.95
101.0	101.0	0.0	1.0	0.01065852	50000	0.0001	64	0.95
32.0	32.0	0.0	1.0	0.0	100000	0.01	64	0.98
101.0	101.0	0.0	1.0	0.0	200000	0.01	64	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.01	64	0.99
32.0	32.0	0.0	1.0	0.0	200000	0.01	64	0.99
101.0	101.0	0.0	1.0	0.0	100000	0.01	128	0.95
32.0	32.0	0.0	1.0	0.0	200000	0.01	128	0.95
32.0	32.0	0.0	1.0	0.0	100000	0.01	128	0.98
32.0	32.0	0.0	1.0	0.0	200000	0.01	128	0.98
32.0	32.0	0.0	1.0	0.0	100000	0.01	128	0.99
32.0	32.0	0.0	1.0	0.0	200000	0.01	128	0.99

Continued on next page

TABLE IV: (Continued)

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma
32.0	32.0	0.0	1.0	0.0	100000	0.01	256	0.95
101.0	101.0	0.0	1.0	0.0	200000	0.01	256	0.95
101.0	101.0	0.0	1.0	0.0	100000	0.01	256	0.98
101.0	101.0	0.0	1.0	0.0	200000	0.01	256	0.98
101.0	101.0	0.0	1.0	0.0	100000	0.01	256	0.99
101.0	101.0	0.0	1.0	0.0	200000	0.01	256	0.99
101.0	101.0	0.0	1.0	0.013765709	200000	0.0001	32	0.95
101.66	101.66041666666666	1.4779715829473852	1.0	0.49349087	500000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.0037594032	200000	0.0001	32	0.98
317.21	315.175	168.10974362005314	1.0	0.27305958	500000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.26061955	100000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.15567426	200000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.2577756	50000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.0011972742	500000	0.0001	32	0.99
101.0	101.0	0.0	1.0	0.4264543	500000	0.0001	64	0.95
101.0	101.0	0.0	1.0	0.5235426	500000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.16878958	500000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.0018354427	500000	0.0001	128	0.95
101.0	101.0	0.0	1.0	0.27324846	500000	0.0001	128	0.98
102.26	102.26250000000003	2.2874439883852893	1.0	0.66430426	500000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.04448829	500000	0.0001	256	0.95
101.0	101.0	0.0	1.0	0.0660648	500000	0.0001	256	0.98
101.0	101.0	0.0	1.0	0.0	500000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.13174301	100000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.09643787	200000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.5710811	50000	0.0001	64	0.99
381.55	385.0375000000001	241.9769152212665	1.0	0.36345157	500000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.0	500000	0.001	32	0.98
32.0	32.0	0.0	1.0	0.0	500000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.0	500000	0.001	64	0.95
101.0	101.0	0.0	1.0	0.0	500000	0.001	64	0.98
101.0	101.0	0.0	1.0	1.2691278e-15	500000	0.001	64	0.99
276.22	275.4375	147.5449477278026	1.0	0.18036482	500000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.56372374	500000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.0	500000	0.001	128	0.99
101.0	101.0	0.0	1.0	0.0	500000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.29914907	100000	0.0001	128	0.95

Continued on next page

TABLE IV: (Continued)

Average Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.15908673	200000	0.0001	128	0.95
101.0	101.0	0.0	1.0	0.46206355	50000	0.0001	128	0.95
101.0	101.0	0.0	1.0	0.6312196	500000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.01546663	100000	0.0001	128	0.98
101.0	101.0	0.0	1.0	0.4661112	200000	0.0001	128	0.98
101.0	101.0	0.0	1.0	0.36904824	50000	0.0001	128	0.98
101.0	101.0	0.0	1.0	0.2513304	100000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.51649725	200000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.043544043	50000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.13085295	100000	0.0001	256	0.95
101.0	101.0	0.0	1.0	0.00066719955	200000	0.0001	256	0.95
101.0	101.0	0.0	1.0	0.10265379	50000	0.0001	256	0.95

TABLE V: Full Results of Training and Testing with PPO with Stable-Baslines3

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	32	0.95
812.6	806.5124999999998	323.45047	0.9999999699556208	0.09508175316214572	19.75	500000	0.0001	32	0.95
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	32	0.95
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	256	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.0001	256	0.98
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.0001	256	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	256	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.08902852661503774	0.0	500000	0.0001	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	256	0.99
101.0	101.0	0.0	1.0	0.22815410253896673	0.0	100000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.21737672777178704	0.0	200000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.23639839231447893	0.0	500000	0.001	32	0.95
129.49	129.86875	50.865803	1.000000042421531	0.05020530212173093	1.37	50000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.2090657778649152	0.0	100000	0.001	32	0.98
101.0	101.0	0.0	1.0	0.21610038231545928	0.0	200000	0.001	32	0.98
665.35	665.4687499999999	305.92642	0.9999999633064365	0.07653727149046367	15.68	500000	0.001	32	0.98
101.0	101.0	0.0	1.0	0.16239192236055291	0.0	50000	0.001	32	0.98
101.0	101.0	0.0	1.0	0.23084599549063814	0.0	100000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.19484756396431724	0.0	200000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.15014410351926283	0.0	50000	0.001	32	0.99
101.0	101.0	0.0	1.0	0.03424566218998138	0.0	100000	0.001	64	0.95
101.0	101.0	0.0	1.0	0.19817860994020187	0.0	200000	0.001	64	0.95
101.0	101.0	0.0	1.0	0.2012234094696599	0.0	500000	0.001	64	0.95
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.001	64	0.95
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.001	64	0.98
101.0	101.0	0.0	1.0	0.06302911479266737	0.0	200000	0.001	64	0.98
101.0	101.0	0.0	1.0	0.11028526615037741	0.0	500000	0.001	64	0.98
99.12	99.15416666666668	2.9505253	1.0000000277096652	0.1562036097346594	0.0	50000	0.001	64	0.98
911.99	909.8145833333334	206.99017	0.9999999892919604	0.08809114421006185	22.51	100000	0.001	64	0.99
101.0	101.0	0.0	1.0	0.009001078325654346	0.0	200000	0.001	64	0.99

Continued on next page

TABLE V: (Continued)

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.1358239388295265	0.0	500000	0.001	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.001	64	0.99
101.61	101.62916666666666	0.8111104	1.0000000060068062	0.20189385855737568	0.0	100000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.14751298892265466	0.0	500000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.001	128	0.95
101.0	101.0	0.0	1.0	0.0893147730614646	0.0	100000	0.001	128	0.98
884.68	885.7437500000001	229.6795	0.9999999917210531	0.08294436643723685	21.76	200000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.19051857661013624	0.0	500000	0.001	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.001	128	0.98
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.1187922752671307	0.0	200000	0.001	64	0.98
101.0	101.0	0.0	1.0	0.22839721595921975	0.0	500000	0.0001	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.21718262915400444	0.0	100000	0.001	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.001	128	0.99
781.18	775.0708333333331	333.9562	0.99999999062416	0.09860452939298676	18.84	500000	0.001	128	0.99
101.0	101.0	0.0	1.0	0.2014527987452211	0.0	50000	0.001	128	0.99
101.0	101.0	0.0	1.0	0.16781099892167436	0.0	100000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.15901382217429663	0.0	200000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.05051465542593862	0.0	500000	0.001	256	0.95
32.69	32.71875	6.8654127	0.999999957990486	0.0008116851289089307	0.0	50000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.009799039309871582	0.0	200000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.09705911185177926	0.0	500000	0.001	256	0.98
72.71	73.25625	33.936497	0.9999999874085086	0.03173388427758309	0.0	50000	0.001	256	0.98
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.001	256	0.99
74.78	74.11874999999999	33.491665	0.9999999836760748	0.010500931281246939	0.0	200000	0.001	256	0.99
101.0	101.0	0.0	1.0	0.08015880796000394	0.0	500000	0.001	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.001	256	0.99
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.005	32	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.005	32	0.95
101.0	101.0	0.0	1.0	0.188414861288109	0.0	500000	0.005	32	0.95
101.0	101.0	0.0	1.0	0.22630722478188411	0.0	50000	0.005	32	0.95

Continued on next page

TABLE V: (Continued)

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
870.95	876.3312500000001	228.9071	1.0000000140157659	0.08936215192978958	21.36	100000	0.005	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.005	32	0.98
67.45	67.84374999999999	31.128563	0.9999999547552585	0.023887451590080214	0.0	500000	0.005	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.005	32	0.99
328.51	326.66458333333327	181.18486	1.000000029727025	0.07378423839943546	6.17	200000	0.005	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.005	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.005	32	0.99
404.93	405.4583333333334	203.25966	0.999999981912383	0.11595142270154543	8.42	100000	0.005	64	0.95
207.79	209.43333333333328	81.82999	0.9999999676891709	0.06696935644904273	2.86	200000	0.005	64	0.95
101.0	101.0	0.0	1.0	0.23777864915204389	0.0	500000	0.005	64	0.95
101.0	101.0	0.0	1.0	0.1451014606411136	0.0	50000	0.005	64	0.95
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.005	64	0.98
101.0	101.0	0.0	1.0	0.21522399764728942	0.0	200000	0.005	64	0.98
602.49	598.8395833333333	299.4375	0.9999999837912248	0.1506848459319278	13.87	500000	0.005	64	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	64	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.005	64	0.99
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.005	64	0.99
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.005	64	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	32	0.99
856.2	859.308333333333	254.15318	1.0000000142572194	0.09047996279890251	20.89	500000	0.0001	32	0.98
101.0	101.0	0.0	1.0	0.12229781393980982	0.0	500000	0.0001	32	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	32	0.99
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.005	128	0.95
101.0	101.0	0.0	1.0	0.08613861386138615	0.0	200000	0.005	128	0.95
163.13	163.36249999999998	32.86903	1.0000000299320329	0.16113494767144587	1.75	500000	0.005	128	0.95
53.39	52.70000000000001	31.912033	0.9999999885680546	0.023274188805019122	0.0	50000	0.005	128	0.95
101.0	101.0	0.0	1.0	0.08501715518086464	0.0	100000	0.005	128	0.98
374.64	371.50000000000006	232.2598	1.000000039100036	0.17731982511652283	7.64	200000	0.005	128	0.98
410.7	415.9666666666667	267.27325	1.0000000297225013	0.11914195798272485	8.45	500000	0.005	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	128	0.98
32.0	32.0	0.0	1.0	0.00181640625	0.0	100000	0.005	128	0.99
401.17	384.74375000000003	429.8993	1.0000000334714318	0.03945579210078576	9.24	200000	0.005	128	0.99

Continued on next page

TABLE V: (Continued)

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.005	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	128	0.99
101.0	101.0	0.0	1.0	0.009142240956768943	0.0	100000	0.005	256	0.95
101.0	101.0	0.0	1.0	0.03043623174198608	0.0	200000	0.005	256	0.95
809.07	802.55625	262.58887	1.000000009052639	0.09087794503376374	19.56	500000	0.005	256	0.95
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	256	0.95
61.67	62.33125000000001	34.16023	0.9999999703088261	0.10723201061017303	0.0	100000	0.005	256	0.98
101.0	101.0	0.0	1.0	0.06893637878639348	0.0	200000	0.005	256	0.98
101.0	101.0	0.0	1.0	0.24312322321341046	0.0	500000	0.005	256	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	256	0.98
101.0	101.0	0.0	1.0	0.225728850112734	0.0	100000	0.005	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.005	256	0.99
101.0	101.0	0.0	1.0	0.13760219586315064	0.0	500000	0.005	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.005	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.01	32	0.95
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	32	0.95
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.01	32	0.95
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.01	32	0.95
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.01	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	32	0.98
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.01	32	0.98
101.0	101.0	0.0	1.0	0.13610038231545926	0.0	50000	0.01	32	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.01	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.01	32	0.99
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	32	0.99
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.01	32	0.99
32.0	32.0	0.0	1.0	0.000908203125	0.0	100000	0.01	64	0.95
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	64	0.95
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	64	0.95
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.01	64	0.95
114.1	113.65416666666668	17.672861	0.9999999866268282	0.20206559645082112	0.34	100000	0.0001	64	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	64	0.95
883.38	880.3750000000001	227.1594	1.0000000055274203	0.0888037328046784	21.68	500000	0.0001	64	0.95
101.0	101.0	0.0	1.0	0.23822762474267226	0.0	500000	0.001	32	0.95
101.0	101.0	0.0	1.0	0.11073424174100578	0.0	50000	0.0001	64	0.95

Continued on next page

TABLE V: (Continued)

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.01	64	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	64	0.98
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	64	0.98
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.01	64	0.98
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.01	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.01	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.01	64	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.01	64	0.99
103.06	103.00000000000001	3.719731	0.9999999763108262	0.24508579985100126	0.11	100000	0.01	128	0.95
681.05	682.6250000000001	332.95297	0.99999998207616	0.07858522424837598	16.0	200000	0.01	128	0.95
101.0	101.0	0.0	1.0	0.04607195373002646	0.0	500000	0.01	128	0.95
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.01	128	0.95
101.0	101.0	0.0	1.0	0.22315067150279383	0.0	100000	0.01	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	128	0.98
80.49	80.21249999999999	38.85421	0.9999999734596787	0.039779584967341894	0.28	50000	0.01	128	0.98
101.0	101.0	0.0	1.0	0.18833447701205772	0.0	100000	0.01	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.01	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.01	128	0.99
101.0	101.0	0.0	1.0	0.1844191745907264	0.0	100000	0.01	256	0.95
683.46	675.7708333333334	303.3	1.0000000321491473	0.0800973712931779	16.03	200000	0.01	256	0.95
372.28	370.0895833333334	229.39018	0.9999999967210081	0.07843894644240328	7.35	500000	0.01	256	0.95
101.0	101.0	0.0	1.0	0.08745809234388789	0.0	50000	0.01	256	0.95
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.01	256	0.98
101.0	101.0	0.0	1.0	0.21826095480835211	0.0	200000	0.01	256	0.98
101.74	101.74583333333332	1.4044214	0.9999999790030424	0.24338695464532367	0.0	500000	0.01	256	0.98
87.0	86.91041666666666	29.283443	1.0	0.04143179887056077	0.0	50000	0.01	256	0.98
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.01	256	0.99
101.0	101.0	0.0	1.0	0.2450191157729634	0.0	200000	0.01	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	500000	0.01	256	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.01	256	0.99
101.0	101.0	0.0	1.0	0.060386236643466325	0.0	100000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	64	0.98
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.0001	64	0.98

Continued on next page

TABLE V: (Continued)

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passes Pipes	Steps	Learning Rate	Batch Size	Gamma
101.0	101.0	0.0	1.0	0.002709538280560729	0.0	500000	0.001	128	0.95
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	64	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.1355063229095187	0.0	100000	0.001	256	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	500000	0.0001	64	0.99
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	64	0.99
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	128	0.95
103.79	103.79791666666665	2.7578797	1.000000008820959	0.1943457790155624	0.02	200000	0.0001	128	0.95
891.89	900.135416666667	239.84727	1.0000000164240406	0.08087562505911411	21.94	500000	0.0001	128	0.95
57.53	57.58750000000003	33.3135	0.999999978781451	0.021594505134300564	0.0	50000	0.0001	128	0.95
101.0	101.0	0.0	1.0	0.0	0.0	100000	0.0001	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	200000	0.0001	128	0.98
99.82	99.82083333333333	2.1371944	0.9999999969427392	0.2037343482762574	0.0	500000	0.0001	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	50000	0.0001	128	0.98
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.12655229879423588	0.0	200000	0.0001	128	0.99
650.93	646.5166666666668	308.73306	0.9999999887480702	0.06689025627372512	15.1	500000	0.0001	128	0.99
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.0001	128	0.99
32.0	32.0	0.0	1.0	0.0	0.0	100000	0.0001	256	0.95
101.0	101.0	0.0	1.0	0.0	0.0	200000	0.0001	256	0.95
805.26	807.9166666666666	309.57492	1.0000000121272943	0.07630119511690087	19.5	500000	0.0001	256	0.95
101.0	101.0	0.0	1.0	0.0	0.0	50000	0.0001	256	0.95

TABLE VI: Full Results of Training and Testing with Custom PPO

Avg. Reward	Convergence Speed	Stability	Sample Efficiency	Exploration vs. Exploitation	Avg. Passed Pipes	Steps	Learning Rate	Batch Size	Gamma
1635.96	1646.4958333333332	1503.6332060712148	1.0	0.11137928358890846	41.61	30000	0,00003	128	0.9999
626.73	624.6270833333334	592.1552136897893	1.0	0.09985551459054282	14.38	30000	0,00001	64	0.9999
406.06	403.9166666666667	383.4105846217603	1.0	0.10169708025597102	8.39	30000	0,00001	32	0.95
138.5	139.16458333333335	49.51737068948633	1.0	0.10019025444323633	1.19	30000	0,00001	256	0.9999
117.79	117.95624999999997	24.228204638396136	1.0	0.11485876033942143	0.54	30000	0,00001	128	0.98