Adaptive Learning in Motion:
Harnessing Cloud-Based AI for a Smart Soccer Ball's Real-Time Navigation

City University of Seattle
School of Technology & Computing
TEAM A

Verónica Elze
ElzeVeronica@CityUniversity.edu
Master of Artificial Intelligence

Vidhyalakshmi Amarnath
AmarnathVidhyalaksh@CityUniversity.edu
Master of Artificial Intelligence

Honorine Ndom Ndzah
NdomndzahHonorine@CityUniversity.edu
Master of Computer Science

Jayasudha Kalamegam
KalamegamJayasudha@CityUniversity.edu
Master of Computer Science

Abstract

This project aims to develop a Smart Soccer Ball, a compact, autonomous system that utilizes reinforcement learning, real-time sensor technology, and cloud-based AI platforms to navigate dynamic environments. Leveraging AWS SageMaker for RL model training and SimSpace Weaver for scalable 3D simulations, the soccer ball will be trained to adapt to both static and dynamic obstacles within a puzzle maze. The project will be conducted in two phases: Phase 1 focuses on developing a Minimum Viable Product with basic maze navigation and static obstacles, while Phase 2 (contingent on efficient progress) will explore advanced RL techniques, dynamic environments, and sophisticated sensors for complex obstacle handling. This endeavor provides an opportunity for hands-on AI and robotics experience, offering insights into low-cost, adaptable autonomous systems and smart sports technology. By demonstrating adaptive, real-time decision-making in a compact device, the project highlights the potential of applying advanced AI to sensor-driven robotics.

Keywords: smart soccer ball, reinforcement learning, AWS SageMaker, autonomous navigation, adaptive robotics, dynamic environments

1. INTRODUCTION

Problem to Solve

One of the key challenges we face is navigating dynamic environments with an autonomous system. Although autonomous navigation has been explored in various fields, it remains complex when environments constantly change. Reinforcement learning (RL) models, though powerful, often struggle to generalize across varying conditions and adjust to real-time changes (Kober et al., 2013). Traditional RL approaches, which rely heavily on trial and error, can be inefficient in unpredictable environments.

This is where model-based reinforcement learning (MBRL) steps in, enhancing decision-making by allowing systems to anticipate and adapt to changes. MBRL has proven effective in sensor-driven robotics, addressing real-time navigation challenges (Polydoros & Nalpantidis, 2017). By incorporating MBRL, we aim to improve the adaptability and efficiency of the Smart Soccer Ball, enabling it to tackle both static and dynamic obstacles.

Motivation

Teaching an autonomous agent to navigate dynamic environments presents exciting challenges with real-world implications. Our goal is to develop a Smart Soccer Ball capable of learning how to navigate a puzzle maze. RL models show promise but often struggle to adapt to ever-changing environments (Kober et al., 2013).

MBRL enhances decision-making and control by planning ahead, making it particularly useful in environments that require real-time adaptability (Polydoros & Nalpantidis, 2017). This method will allow our soccer ball to handle dynamic obstacles effectively.

There's also growing interest in low-cost autonomous systems in industries like smart sports technology, where AI-driven devices enhance performance (Zhang, 2022). By integrating sensors and leveraging AWS SageMaker for real-time decision-making, we aim to demonstrate how small-scale AI systems can efficiently tackle complex tasks (Malekzadeh et al., 2018). Plus, let's be honest, it's undeniably entertaining to see a soccer ball outwit a maze.

Usefulness/Beneficiaries

This project demonstrates practical applications for industries and education alike. By leveraging RL, we show how autonomous systems can adapt to dynamic environments, a challenge for traditional models (Kober et al., 2013).

For industries like smart sports, MBRL offers valuable insights into developing low-cost systems that respond to environmental changes. It improves decision-making by anticipating rather than reacting (Polydoros & Nalpantidis, 2017).

Educationally, this project provides hands-on experience with sensor integration and AI platforms like AWS SageMaker, enabling real-time decision-making (Malekzadeh et al., 2018). These skills are essential as AI and robotics continue to grow.

Beyond sports, this project contributes to autonomous navigation research, showing how small-scale systems can handle complex tasks. Watching a soccer ball master a maze? Definitely a bonus (Zhang, 2022).

2. LITERATURE REVIEW

Autonomous systems capable of adapting to dynamic environments are increasingly relevant across various domains, including robotics, artificial intelligence (AI), and sports technology. The Smart Soccer Ball project aims to develop a compact, sensor-driven device that utilizes reinforcement learning (RL) to navigate mazes and dynamic obstacles. This literature review explores foundational research on similar projects, identifies applicable methodologies, and highlights how these insights contribute to the development of this project.

Reinforcement Learning in Robotics

Reinforcement learning has emerged as a transformative approach for robotics, enabling systems to learn optimal policies through trial and error. Kober et al. (2013) provide a comprehensive survey of RL applications in robotics, emphasizing its utility in environments requiring adaptive decision-making. Their research highlights challenges such as designing appropriate reward structures and balancing exploration versus exploitation—critical elements in the Smart Soccer Ball's ability to learn efficient navigation strategies. Unlike traditional rule-based systems, RL models dynamically adjust to new scenarios, offering flexibility that aligns with the project's goals.

The Smart Soccer Ball project specifically benefits from model-based reinforcement learning (MBRL). As Polydoros and Nalpantidis (2017) explain, MBRL enhances a system's ability to predict and respond to environmental changes. By leveraging these predictive capabilities, the soccer ball can navigate dynamic mazes more effectively than model-free methods allow. This focus on adaptability differentiates the project from similar applications of RL in static environments.

Cloud Computing and AI Scalability

The integration of cloud-based platforms like AWS SageMaker is another pivotal aspect of the Smart Soccer Ball project. SageMaker provides a scalable environment for developing, training, and deploying RL models, offering computational resources that significantly reduce the time and cost of experimentation (Amazon Web Services, 2021). This capability contrasts with earlier robotics projects that relied heavily on on-premises hardware, which limited scalability.

Tripuraneni and Song (2019) underscore the role of cloud infrastructure in accelerating AI model development and deployment. Their work demonstrates how platforms like SageMaker streamline workflows by enabling distributed computing and iterative model tuning. This scalability is critical for the Smart Soccer Ball, where rapid experimentation with various RL algorithms—such as proximal policy optimization (PPO) and soft actor-critic (SAC)—is essential for optimizing navigation strategies (Schulman et al., 2017; Haarnoja et al., 2018).

Autonomous Navigation and Sensor Integration

Sensor-driven systems form the backbone of many autonomous robotics projects. Malekzadeh et al. (2018) show how integrating accelerometers and proximity sensors enhances navigation accuracy in compact robots. Their findings are directly applicable to the Smart Soccer Ball's design, where sensors will enable the ball to detect and avoid obstacles. Unlike projects that rely solely on vision-based systems, this project incorporates multiple sensor modalities to improve real-time decision-making.

The use of LIDAR as a potential upgrade aligns with research by Malla and Dholakiya (2022), who demonstrate its effectiveness in mapping dynamic environments. While LIDAR is not part of the initial phase of the Smart Soccer Ball project, its inclusion in future iterations could significantly enhance spatial awareness and obstacle detection capabilities. This step would elevate the system's performance to meet the challenges of complex, real-world scenarios.

Comparative Analysis with Similar Projects

Zhang (2022) explores AI applications in soccer training, employing genetic algorithms for optimizing movement paths. While genetic algorithms differ from RL, Zhang's work underscores the value of AI in dynamic navigation. The Smart Soccer Ball project builds on these insights by applying RL techniques that offer greater flexibility and adaptability. Unlike genetic algorithms, which evolve solutions over multiple iterations, RL models learn in real time, making them more suited for the dynamic obstacle scenarios envisioned for this project.

Another comparable effort is Shah et al.'s (2020) AirSim framework, which uses high-fidelity simulations for training autonomous vehicles. The Smart Soccer Ball leverages similar principles by incorporating AWS SimSpace Weaver for dynamic environment simulations. However, while AirSim focuses on vehicles with high computational demands, this project emphasizes compact, cost-effective robotics.

Integration and Contribution

The Smart Soccer Ball project synthesizes advances in reinforcement learning, cloud-based AI, and sensor-driven robotics to create an autonomous system with real-world applications. By leveraging model-based RL, scalable cloud platforms like AWS SageMaker, and advanced sensor technologies, the project addresses challenges in navigating dynamic environments. Compared to related work, the project stands out for its emphasis on cost-effective design and adaptability, demonstrating the potential of AI in sports and robotics.

This literature review highlights the foundational research and methodologies that guide the Smart Soccer Ball's development, positioning it as a pioneering effort in the intersection of autonomous systems and sports technology.

3. METHODOLOGY

To develop a reinforcement learning (RL) solution for navigating a simulated maze environment, we have focused on building and validating the foundational components locally. This approach ensures a controlled development environment before transitioning to cloud-based deployment and scaling.

Environment Setup and Customization

Using the MiniGrid library, we created a grid-based maze environment tailored for RL tasks. A custom observation wrapper was implemented to preprocess and flatten the multi-dimensional observations into a single vector, streamlining compatibility with RL algorithms. Discrete actions, including movements in cardinal directions and interactions with the environment, were defined to enable efficient navigation by the agent. All customization and validation of the environment have been carried out in a local environment, ensuring that the setup is robust before integration with external systems.

Agent Training

The Proximal Policy Optimization (PPO) algorithm was used to train the agent, leveraging the Stable Baselines3 framework. Training was executed locally, utilizing the DummyVecEnv wrapper to vectorize the environment and enable efficient agent-environment interactions. This local setup allowed iterative testing and debugging in a controlled environment.

Training progress was monitored with TensorBoard, which provided real-time insights into key metrics such as rewards, policy loss, and training duration. By logging these metrics, we validated the agent's learning progress and fine-tuned parameters, such as learning rate and policy architecture, for improved performance. The agent's speed in reaching the green square goal showed a notable improvement, decreasing from an initial navigation time of approximately 106 seconds to just 2 seconds by the end of training.

Evaluation and Validation

Our local setup also supported detailed evaluation and debugging of the environment and agent behavior. Observations were validated to ensure they were processed correctly, and actions were tested to confirm expected outcomes. Debugging efforts focused on resolving compatibility issues, such as observation flattening errors and reward signal inconsistencies. The validation process demonstrated that the agent could learn and adapt within the customized environment.

This locally built foundation ensures that all components are well-tested and functional before deploying to AWS for further scalability and performance enhancements. By prioritizing local implementation, we have established a reliable framework to transition into more complex, cloud-based simulations.

Tools

Our Smart Soccer Ball project will leverage a combination of hardware, software, and cloud-based tools to facilitate reinforcement learning (RL) and real-world agent interactions. These tools are integral to developing, testing, and deploying our adaptive navigation system.

Hardware (to be implemented)

- Sphero Mini App-Enabled Robotic Ball: Will serves as the physical embodiment of our RL agent, translating digital decision-making into tangible movements.
- Mobile Device: Will be used to control and monitor the Sphero Mini through the Sphero Edu App, ensuring seamless interaction during testing phases.

Software Tools

- Python Programming Language: Provides the foundation for RL model development, cloud integration, and environment customization.
- MiniGrid Library: A minimalistic, grid-based environment used for creating and customizing maze scenarios for training our RL agent.
- Stable Baselines3 Framework: Facilitates the implementation of the Proximal Policy Optimization (PPO) algorithm, chosen for its efficiency and reliability in RL tasks.
- OpenAI Gym: Allows preliminary testing in a simulated environment to refine RL strategies before deployment.

Cloud Tools (to be implemented)

- AWS SageMaker: Will handle the training and fine-tuning of RL models, leveraging its scalable infrastructure to accelerate development cycles.
- AWS SimSpace Weaver: Will simulate dynamic, three-dimensional puzzle mazes for testing agent adaptability and optimizing navigation strategies.

Development and Monitoring Tools

- TensorBoard: Enables real-time visualization of training metrics such as rewards, loss, and policy performance.
- GitHub: Will ensures version control and collaborative tracking of code and model iterations.

By integrating these tools, the project will achieve a balance between virtual simulation and real-world testing, fostering a robust development environment for adaptive RL models.

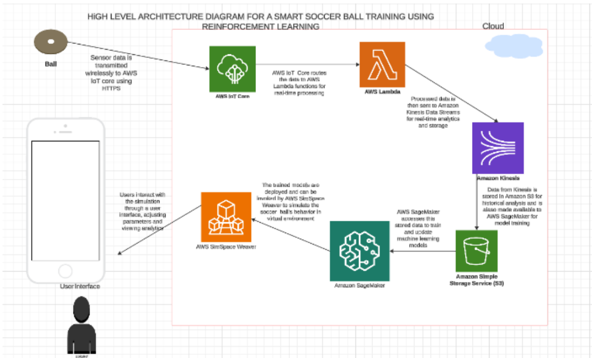


Figure 1 Architecture Diagram

High-Level Architecture Overview

The high-level architecture diagram (Figure 2, see also Appendix A) illustrates the proposed integration of hardware, cloud-based AI tools, and simulation environments in the Smart Soccer Ball project. This architecture enables seamless data flow, real-time analytics, and reinforcement learning model training. Key components include:

- AWS IoT Core: Facilitates wireless data transmission from the soccer ball's sensors to the cloud.
- AWS Lambda: Processes sensor data in real-time, triggering downstream workflows.
- Amazon Kinesis and S3: Streams processed data for real-time and historical analytics, stored for subsequent use.
- AWS SageMaker: Trains and refines reinforcement learning models using collected data.
- AWS SimSpace Weaver: Simulates dynamic environments, allowing the soccer ball to learn and adapt in a virtual 3D maze.

This architecture will leverage cloud computing for scalability, ensuring efficient training and evaluation of the soccer ball's navigation capabilities.

4. ACKNOWLEDGMENTS

I would like to acknowledge the assistance of OpenAI's ChatGPT tool in improving the verbiage, flow, and transitions of this paper.

5. REFERENCES

The following references include a combination of sources recommended by ChatGPT, with the majority published or updated within the last five years. All citations adhere to APA (American Psychological Association) guidelines.

Scholarly Sources

These sources provide foundational research and insights relevant to the project.

Amazon Web Services. (2021). *AWS SageMaker: Developer Guide*. Retrieved from https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2961–2970. https://arxiv.org/abs/1801.01290

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research, 32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

Malla, A., & Dholakiya, M. (2022). Autonomous navigation in unstructured environments using LIDAR-based reinforcement learning. *International Journal of Advanced Robotic Systems, 19*(1), 1–12. https://doi.org/10.1177/17298814221075422

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Polydoros, A. S., & Nalpantidis, L. (2017). Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems, 86*(2), 153–173. https://doi.org/10.1007/s10846-017-0570-0

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv Preprint*. https://arxiv.org/abs/1707.06347

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2020). AirSim: High-fidelity visual and physical simulation for autonomous vehicles. *Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Retrieved from https://github.com/microsoft/AirSim

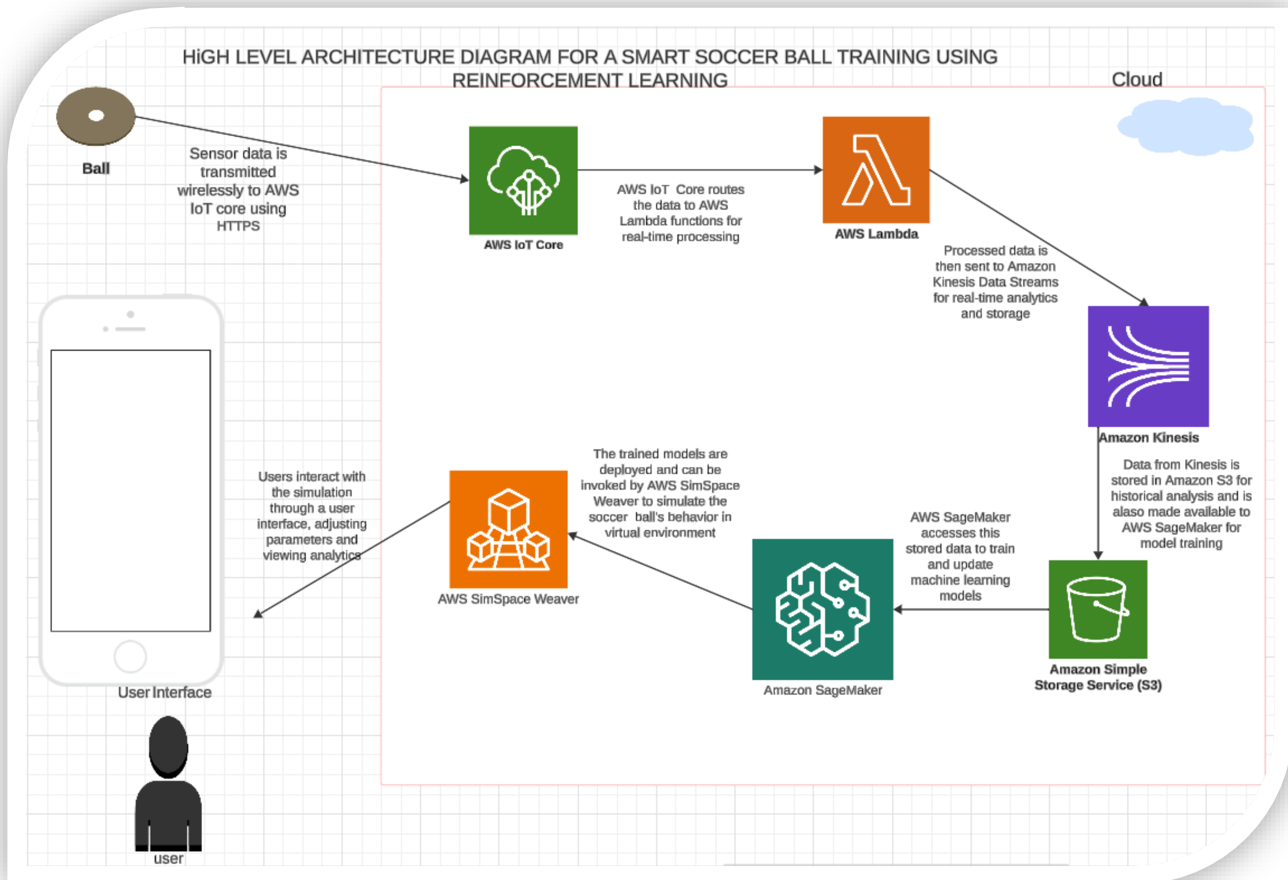
Tripuraneni, V., & Song, M. (2019). Cloud infrastructure for scalable machine learning: A comprehensive study. *Journal of Cloud Computing, 8*(1), 1–18. https://doi.org/10.1186/s13677-019-0124-9

Zhang, Y. (2022). Genetic algorithms in sports: Enhancing soccer strategy through artificial intelligence. *Computers in Sport Science, 11*(3), 215–231. https://doi.org/10.1016/j.coss.2022.07.003

APPENDIX A

Architecture

This appendix presents the high-level architecture diagram outlining the integration of hardware, software, and cloud components in the Smart Soccer Ball project.



APPENDIX B

Local Environment Initialization, Agent Interaction, and Episode Completion Output
This appendix includes visualizations from TensorBoard logs, showcasing key training metrics such as reward progression and policy loss over time.

```
C:\Users\MissV\OneDrive\Documents\Education\CityU\2024FallQ4\AI620\Code> python environment.py
Custom observation shape: (1, 193) Observation space: Box(-inf, inf, (193,), float32)
Action space: Discrete(7) C:\Users\MissV\OneDrive\Documents\Education\CityU\2024FallQ4\AI620\Code\venv\Lib\site-packages\stable_baselines3\common\vec_env\base_vec_env.py:243: UserWarning: You tried to call render() but no render_mode was passed to the env constructor. warnings.warn("You tried to call render() but no render_mode was passed to the env constructor.")
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
```

```
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
```

```
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
```

```
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [3] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [2] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [6] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [1] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [4] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [0] Reward: [0.], Done: [False], Info: [{'TimeLimit.truncated': False}]
Action taken: [5] Reward: [0.], Done: [ True], Info: [{'episode': {'r': 0.0, 'l': 256, 't': 0.157253}, 'TimeLimit.truncated': True,
'terminal_observation': array([ 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 1., 0.,
0., 10., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 2., 5., 0., 2., 5., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1.,
0., 0., 2., 5., 0., 2., 5., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 2., 5., 0., 2., 5., 0., 1., 0., 0., 1., 0., 0., 1.,
0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 2., 5., 0., 2., 5., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 2., 5., 0., 2.,
5., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 1., 0., 0., 8., 1., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2., 5., 0., 2.,
5., 0., 2., 5., 0., 2., 5., 0., 0.], dtype=float32)}] Episode complete.
```

APPENDIX C

Agent Training
This appendix provides a still image of the environment during training in human render mode, illustrating the agent's perspective within the grid-world maze.

```
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.9        |
|    ep_rew_mean          | 0.955       |
| time/                   |             |
|    fps                  | 723         |
|    iterations           | 54          |
|    time_elapsed         | 76          |
|    total_timesteps      | 55296       |
| train/                  |             |
|    approx_kl            | 0.055399723 |
|    clip_fraction        | 0.155       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.167      |
|    explained_variance   | 0.661       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0264     |
|    n_updates            | 530         |
|    policy_gradient_loss | -0.0259     |
|    value_loss           | 0.000623    |
-----------------------------------------
----------------------------------------
| rollout/                |            |
|    ep_len_mean          | 12.2       |
|    ep_rew_mean          | 0.957      |
| time/                   |            |
|    fps                  | 723        |
|    iterations           | 55         |
|    time_elapsed         | 77         |
|    total_timesteps      | 56320      |
| train/                  |            |
|    approx_kl            | 0.27952486 |
|    clip_fraction        | 0.146      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.129     |
|    explained_variance   | 0.52       |
|    learning_rate        | 0.0003     |
|    loss                 | -0.0221    |
|    n_updates            | 540        |
|    policy_gradient_loss | -0.0178    |
|    value_loss           | 0.000592   |
----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 11.9        |
|    ep_rew_mean          | 0.958       |
| time/                   |             |
|    fps                  | 720         |
|    iterations           | 56          |
|    time_elapsed         | 79          |
|    total_timesteps      | 57344       |
| train/                  |             |
|    approx_kl            | 0.023819925 |
|    clip_fraction        | 0.24        |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.217      |
|    explained_variance   | 0.833       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0315     |
|    n_updates            | 550         |
|    policy_gradient_loss | -0.0296     |
|    value_loss           | 0.000192    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.3        |
|    ep_rew_mean          | 0.957       |
| time/                   |             |
|    fps                  | 720         |
|    iterations           | 57          |
|    time_elapsed         | 81          |
|    total_timesteps      | 58368       |
| train/                  |             |
|    approx_kl            | 0.073045135 |
|    clip_fraction        | 0.135       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.154      |
|    explained_variance   | 0.843       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.04       |
|    n_updates            | 560         |
|    policy_gradient_loss | -0.0231     |
|    value_loss           | 0.000152    |
-----------------------------------------
----------------------------------------
| rollout/                |            |
|    ep_len_mean          | 11.6       |
|    ep_rew_mean          | 0.959      |
| time/                   |            |
|    fps                  | 720        |
|    iterations           | 58         |
|    time_elapsed         | 82         |
|    total_timesteps      | 59392      |
| train/                  |            |
|    approx_kl            | 0.06047496 |
|    clip_fraction        | 0.109      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.131     |
|    explained_variance   | 0.589      |
|    learning_rate        | 0.0003     |
|    loss                 | 0.00682    |
|    n_updates            | 570        |
|    policy_gradient_loss | -0.0187    |
|    value_loss           | 0.000467   |
----------------------------------------
Eval num_timesteps=60000, episode_reward=0.96 +/- 0.00
Episode length: 11.00 +/- 0.00
-----------------------------------------
| eval/                   |             |
|    mean_ep_length       | 11          |
|    mean_reward          | 0.961       |
| time/                   |             |
|    total_timesteps      | 60000       |
| train/                  |             |
|    approx_kl            | 0.017942129 |
|    clip_fraction        | 0.043       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.094      |
|    explained_variance   | 0.852       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0165     |
|    n_updates            | 580         |
|    policy_gradient_loss | -0.0123     |
|    value_loss           | 0.000154    |
-----------------------------------------
--------------------------------
| rollout/           |         |
|    ep_len_mean     | 13      |
|    ep_rew_mean     | 0.954   |
| time/              |         |
|    fps             | 718     |
|    iterations      | 59      |
|    time_elapsed    | 84      |
|    total_timesteps | 60416   |
--------------------------------
-------------------------------------
```

```
| rollout/                |             |
|    ep_len_mean          | 12          |
|    ep_rew_mean          | 0.958       |
| time/                   |             |
|    fps                  | 718         |
|    iterations           | 60          |
|    time_elapsed         | 85          |
|    total_timesteps      | 61440       |
| train/                  |             |
|    approx_kl            | 0.0436577   |
|    clip_fraction        | 0.262       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.248      |
|    explained_variance   | 0.632       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0092     |
|    n_updates            | 590         |
|    policy_gradient_loss | -0.0319     |
|    value_loss           | 0.000525    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 11.3        |
|    ep_rew_mean          | 0.96        |
| time/                   |             |
|    fps                  | 719         |
|    iterations           | 61          |
|    time_elapsed         | 86          |
|    total_timesteps      | 62464       |
| train/                  |             |
|    approx_kl            | 0.054237213 |
|    clip_fraction        | 0.0599      |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.106      |
|    explained_variance   | 0.818       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.019      |
|    n_updates            | 600         |
|    policy_gradient_loss | -0.0169     |
|    value_loss           | 0.000258    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 14.7        |
|    ep_rew_mean          | 0.948       |
| time/                   |             |
|    fps                  | 718         |
|    iterations           | 62          |
|    time_elapsed         | 88          |
|    total_timesteps      | 63488       |
| train/                  |             |
|    approx_kl            | 0.08230614  |
|    clip_fraction        | 0.37        |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.3        |
|    explained_variance   | 0.951       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0369     |
|    n_updates            | 610         |
|    policy_gradient_loss | 0.023       |
|    value_loss           | 2.75e-05    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 15.1        |
|    ep_rew_mean          | 0.947       |
| time/                   |             |
|    fps                  | 718         |
|    iterations           | 63          |
|    time_elapsed         | 89          |
|    total_timesteps      | 64512       |
| train/                  |             |
|    approx_kl            | 0.03069654  |
|    clip_fraction        | 0.202       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.277      |
|    explained_variance   | 0.306       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0343     |
|    n_updates            | 620         |
|    policy_gradient_loss | -0.0198     |
|    value_loss           | 0.000805    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 15          |
|    ep_rew_mean          | 0.947       |
| time/                   |             |
|    fps                  | 719         |
|    iterations           | 64          |
|    time_elapsed         | 91          |
|    total_timesteps      | 65536       |
| train/                  |             |
|    approx_kl            | 0.03838364  |
|    clip_fraction        | 0.217       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.263      |
|    explained_variance   | 0.667       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.073      |
|    n_updates            | 630         |
|    policy_gradient_loss | -0.0305     |
|    value_loss           | 0.000518    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.2        |
|    ep_rew_mean          | 0.957       |
| time/                   |             |
|    fps                  | 718         |
|    iterations           | 65          |
|    time_elapsed         | 92          |
|    total_timesteps      | 66560       |
| train/                  |             |
|    approx_kl            | 0.19488329  |
|    clip_fraction        | 0.347       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.282      |
|    explained_variance   | 0.54        |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0969     |
|    n_updates            | 640         |
|    policy_gradient_loss | -0.0641     |
|    value_loss           | 0.00129     |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.6        |
|    ep_rew_mean          | 0.956       |
| time/                   |             |
|    fps                  | 718         |
|    iterations           | 66          |
|    time_elapsed         | 94          |
|    total_timesteps      | 67584       |
| train/                  |             |
|    approx_kl            | 0.12408055  |
|    clip_fraction        | 0.23        |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.255      |
|    explained_variance   | 0.888       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0465     |
|    n_updates            | 650         |
|    policy_gradient_loss | 0.0242      |
|    value_loss           | 4.94e-05    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 11.7        |
|    ep_rew_mean          | 0.959       |
| time/                   |             |
|    fps                  | 717         |
```

```
|    iterations           | 67          |
|    time_elapsed         | 95          |
|    total_timesteps      | 68608       |
| train/                  |             |
|    approx_kl            | 0.025122331 |
|    clip_fraction        | 0.173       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.206      |
|    explained_variance   | 0.771       |
|    learning_rate        | 0.0003      |
|    loss                 | 0.0226      |
|    n_updates            | 660         |
|    policy_gradient_loss | -0.0263     |
|    value_loss           | 0.000238    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 11.9        |
|    ep_rew_mean          | 0.958       |
| time/                   |             |
|    fps                  | 717         |
|    iterations           | 68          |
|    time_elapsed         | 97          |
|    total_timesteps      | 69632       |
| train/                  |             |
|    approx_kl            | 0.06556613  |
|    clip_fraction        | 0.141       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.162      |
|    explained_variance   | 0.95        |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0489     |
|    n_updates            | 670         |
|    policy_gradient_loss | -0.00874    |
|    value_loss           | 4.17e-05    |
-----------------------------------------
Eval num_timesteps=70000, episode_reward=0.96 +/- 0.00
Episode length: 11.00 +/- 0.00
-----------------------------------------
| eval/                   |             |
|    mean_ep_length       | 11          |
|    mean_reward          | 0.961       |
| time/                   |             |
|    total_timesteps      | 70000       |
| train/                  |             |
|    approx_kl            | 0.062489584 |
|    clip_fraction        | 0.0564      |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.0847     |
|    explained_variance   | 0.778       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0217     |
|    n_updates            | 680         |
|    policy_gradient_loss | -0.0199     |
|    value_loss           | 0.000264    |
-----------------------------------------
---------------------------------
| rollout/           |          |
|    ep_len_mean     | 11.1     |
|    ep_rew_mean     | 0.961    |
| time/              |          |
|    fps             | 716      |
|    iterations      | 69       |
|    time_elapsed    | 98       |
|    total_timesteps | 70656    |
---------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 13.7        |
|    ep_rew_mean          | 0.952       |
| time/                   |             |
|    fps                  | 715         |
|    iterations           | 70          |
|    time_elapsed         | 100         |
|    total_timesteps      | 71680       |
| train/                  |             |
|    approx_kl            | 0.05116283  |
|    clip_fraction        | 0.119       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.131      |
|    explained_variance   | 0.985       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0654     |
|    n_updates            | 690         |
|    policy_gradient_loss | -0.0128     |
|    value_loss           | 9.36e-06    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 13.2        |
|    ep_rew_mean          | 0.954       |
| time/                   |             |
|    fps                  | 715         |
|    iterations           | 71          |
|    time_elapsed         | 101         |
|    total_timesteps      | 72704       |
| train/                  |             |
|    approx_kl            | 0.039104126 |
|    clip_fraction        | 0.158       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.24       |
|    explained_variance   | 0.178       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0193     |
|    n_updates            | 700         |
|    policy_gradient_loss | -0.00839    |
|    value_loss           | 0.000591    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.7        |
|    ep_rew_mean          | 0.956       |
| time/                   |             |
|    fps                  | 715         |
|    iterations           | 72          |
|    time_elapsed         | 103         |
|    total_timesteps      | 73728       |
| train/                  |             |
|    approx_kl            | 0.058678307 |
|    clip_fraction        | 0.166       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.236      |
|    explained_variance   | 0.761       |
|    learning_rate        | 0.0003      |
|    loss                 | 0.0961      |
|    n_updates            | 710         |
|    policy_gradient_loss | -0.0254     |
|    value_loss           | 0.000299    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 14.2        |
|    ep_rew_mean          | 0.95        |
| time/                   |             |
|    fps                  | 714         |
|    iterations           | 73          |
|    time_elapsed         | 104         |
|    total_timesteps      | 74752       |
| train/                  |             |
|    approx_kl            | 0.042911883 |
|    clip_fraction        | 0.147       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.215      |
|    explained_variance   | 0.893       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0327     |
|    n_updates            | 720         |
|    policy_gradient_loss | -0.0165     |
|    value_loss           | 0.000142    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
```

```
|    ep_len_mean          | 14.2        |
|    ep_rew_mean          | 0.95        |
| time/                   |             |
|    fps                  | 710         |
|    iterations           | 74          |
|    time_elapsed         | 106         |
|    total_timesteps      | 75776       |
| train/                  |             |
|    approx_kl            | 0.02610638  |
|    clip_fraction        | 0.215       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.269      |
|    explained_variance   | 0.556       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0248     |
|    n_updates            | 730         |
|    policy_gradient_loss | -0.0269     |
|    value_loss           | 0.000532    |
---------------------------------------
---------------------------------------
| rollout/                |             |
|    ep_len_mean          | 13          |
|    ep_rew_mean          | 0.954       |
| time/                   |             |
|    fps                  | 710         |
|    iterations           | 75          |
|    time_elapsed         | 108         |
|    total_timesteps      | 76800       |
| train/                  |             |
|    approx_kl            | 0.04112082  |
|    clip_fraction        | 0.249       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.29       |
|    explained_variance   | 0.645       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0464     |
|    n_updates            | 740         |
|    policy_gradient_loss | -0.0367     |
|    value_loss           | 0.000662    |
---------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.1        |
|    ep_rew_mean          | 0.958       |
| time/                   |             |
|    fps                  | 711         |
|    iterations           | 76          |
|    time_elapsed         | 109         |
|    total_timesteps      | 77824       |
| train/                  |             |
|    approx_kl            | 0.058698382 |
|    clip_fraction        | 0.252       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.229      |
|    explained_variance   | 0.883       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0224     |
|    n_updates            | 750         |
|    policy_gradient_loss | -0.0408     |
|    value_loss           | 0.000145    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.5        |
|    ep_rew_mean          | 0.956       |
| time/                   |             |
|    fps                  | 711         |
|    iterations           | 77          |
|    time_elapsed         | 110         |
|    total_timesteps      | 78848       |
| train/                  |             |
|    approx_kl            | 0.070883654 |
|    clip_fraction        | 0.226       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.171      |
|    explained_variance   | 0.918       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0669     |
|    n_updates            | 760         |
|    policy_gradient_loss | -0.0423     |
|    value_loss           | 0.000216    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.2        |
|    ep_rew_mean          | 0.957       |
| time/                   |             |
|    fps                  | 712         |
|    iterations           | 78          |
|    time_elapsed         | 112         |
|    total_timesteps      | 79872       |
| train/                  |             |
|    approx_kl            | 0.017188694 |
|    clip_fraction        | 0.0901      |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.171      |
|    explained_variance   | 0.66        |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0169     |
|    n_updates            | 770         |
|    policy_gradient_loss | -0.0119     |
|    value_loss           | 0.000467    |
-----------------------------------------
Eval num_timesteps=80000, episode_reward=0.96 +/- 0.00
Episode length: 11.00 +/- 0.00
---------------------------------------
| eval/                   |            |
|    mean_ep_length       | 11         |
|    mean_reward          | 0.961      |
| time/                   |            |
|    total_timesteps      | 80000      |
| train/                  |            |
|    approx_kl            | 0.04313474 |
|    clip_fraction        | 0.107      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.213     |
|    explained_variance   | 0.829      |
|    learning_rate        | 0.0003     |
|    loss                 | 0.00571    |
|    n_updates            | 780        |
|    policy_gradient_loss | -0.0187    |
|    value_loss           | 0.000195   |
---------------------------------------
---------------------------------
| rollout/           |          |
|    ep_len_mean     | 12       |
|    ep_rew_mean     | 0.958    |
| time/              |          |
|    fps             | 712      |
|    iterations      | 79       |
|    time_elapsed    | 113      |
|    total_timesteps | 80896    |
---------------------------------
---------------------------------------
| rollout/                |            |
|    ep_len_mean          | 12.8       |
|    ep_rew_mean          | 0.955      |
| time/                   |            |
|    fps                  | 711        |
|    iterations           | 80         |
|    time_elapsed         | 115        |
|    total_timesteps      | 81920      |
| train/                  |            |
|    approx_kl            | 0.06471266 |
|    clip_fraction        | 0.314      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.294     |
|    explained_variance   | 0.938      |
|    learning_rate        | 0.0003     |
|    loss                 | -0.0486    |
|    n_updates            | 790        |
|    policy_gradient_loss | -0.0427    |
```

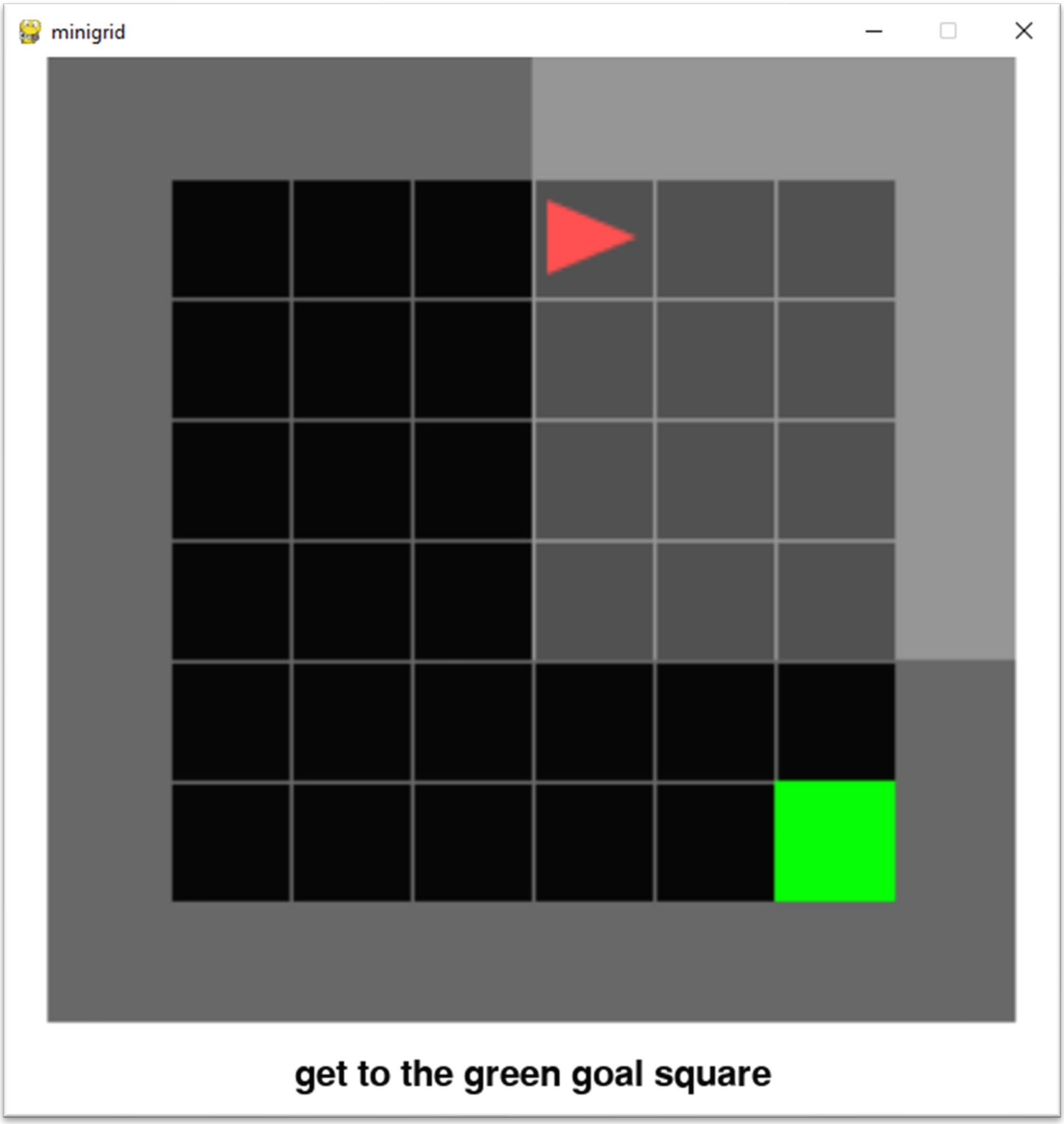
```
| value_loss | 5.16e-05 |
---------------------------------------
---------------------------------------
| rollout/ | |
| ep_len_mean | 11.8 |
| ep_rew_mean | 0.958 |
| time/ | |
| fps | 711 |
| iterations | 81 |
| time_elapsed | 116 |
| total_timesteps | 82944 |
| train/ | |
| approx_kl | 0.06405732 |
| clip_fraction | 0.211 |
| clip_range | 0.2 |
| entropy_loss | -0.176 |
| explained_variance | 0.795 |
| learning_rate | 0.0003 |
| loss | -0.0243 |
| n_updates | 800 |
| policy_gradient_loss | -0.0288 |
| value_loss | 0.000221 |
---------------------------------------
----------------------------------------
| rollout/ | |
| ep_len_mean | 11.6 |
| ep_rew_mean | 0.959 |
| time/ | |
| fps | 710 |
| iterations | 82 |
| time_elapsed | 118 |
| total_timesteps | 83968 |
| train/ | |
| approx_kl | 0.061312027 |
| clip_fraction | 0.0476 |
| clip_range | 0.2 |
| entropy_loss | -0.0518 |
| explained_variance | 0.881 |
| learning_rate | 0.0003 |
| loss | -0.0393 |
| n_updates | 810 |
| policy_gradient_loss | -0.023 |
| value_loss | 0.000125 |
----------------------------------------
----------------------------------------
| rollout/ | |
| ep_len_mean | 12.2 |
| ep_rew_mean | 0.957 |
| time/ | |
| fps | 708 |
| iterations | 83 |
| time_elapsed | 119 |
| total_timesteps | 84992 |
| train/ | |
| approx_kl | 0.039377097 |
| clip_fraction | 0.181 |
| clip_range | 0.2 |
| entropy_loss | -0.117 |
| explained_variance | 0.814 |
| learning_rate | 0.0003 |
| loss | -0.0468 |
| n_updates | 820 |
| policy_gradient_loss | 0.022 |
| value_loss | 0.000187 |
----------------------------------------
----------------------------------------
| rollout/ | |
| ep_len_mean | 11.8 |
| ep_rew_mean | 0.958 |
| time/ | |
| fps | 708 |
| iterations | 84 |
| time_elapsed | 121 |
| total_timesteps | 86016 |
| train/ | |
| approx_kl | 0.022505693 |
| clip_fraction | 0.109 |
| clip_range | 0.2 |
| entropy_loss | -0.158 |
| explained_variance | 0.671 |
| learning_rate | 0.0003 |
| loss | -0.0147 |
| n_updates | 830 |
| policy_gradient_loss | -0.0127 |
| value_loss | 0.000268 |
----------------------------------------
----------------------------------------
| rollout/ | |
| ep_len_mean | 15.4 |
| ep_rew_mean | 0.946 |
| time/ | |
| fps | 707 |
| iterations | 85 |
| time_elapsed | 123 |
| total_timesteps | 87040 |
| train/ | |
| approx_kl | 0.11171889 |
| clip_fraction | 0.142 |
| clip_range | 0.2 |
| entropy_loss | -0.131 |
| explained_variance | 0.914 |
| learning_rate | 0.0003 |
| loss | -0.0735 |
| n_updates | 840 |
| policy_gradient_loss | -0.0356 |
| value_loss | 6.9e-05 |
----------------------------------------
----------------------------------------
| rollout/ | |
| ep_len_mean | 14.4 |
| ep_rew_mean | 0.95 |
| time/ | |
| fps | 706 |
| iterations | 86 |
| time_elapsed | 124 |
| total_timesteps | 88064 |
| train/ | |
| approx_kl | 0.05634898 |
| clip_fraction | 0.322 |
| clip_range | 0.2 |
| entropy_loss | -0.321 |
| explained_variance | 0.337 |
| learning_rate | 0.0003 |
| loss | -0.0435 |
| n_updates | 850 |
| policy_gradient_loss | -0.03 |
| value_loss | 0.00163 |
----------------------------------------
--------------------------------------
| rollout/ | |
| ep_len_mean | 11.6 |
| ep_rew_mean | 0.959 |
| time/ | |
| fps | 706 |
| iterations | 87 |
| time_elapsed | 126 |
| total_timesteps | 89088 |
| train/ | |
| approx_kl | 0.1144023 |
| clip_fraction | 0.191 |
| clip_range | 0.2 |
| entropy_loss | -0.188 |
| explained_variance | 0.632 |
| learning_rate | 0.0003 |
| loss | -0.0365 |
| n_updates | 860 |
| policy_gradient_loss | -0.0271 |
| value_loss | 0.000624 |
--------------------------------------
Eval num_timesteps=90000, episode_reward=0.96 +/- 0.00
Episode length: 11.00 +/- 0.00
----------------------------------------
```

```
| eval/                   |             |
|    mean_ep_length       | 11          |
|    mean_reward          | 0.961       |
| time/                   |             |
|    total_timesteps      | 90000       |
| train/                  |             |
|    approx_kl            | 0.065284915 |
|    clip_fraction        | 0.063       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.0858     |
|    explained_variance   | 0.89        |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0338     |
|    n_updates            | 870         |
|    policy_gradient_loss | 0.0276      |
|    value_loss           | 0.000171    |
-----------------------------------------
---------------------------------
| rollout/           |          |
|    ep_len_mean     | 12.8     |
|    ep_rew_mean     | 0.955    |
| time/              |          |
|    fps             | 705      |
|    iterations      | 88       |
|    time_elapsed    | 127      |
|    total_timesteps | 90112    |
---------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 14.6        |
|    ep_rew_mean          | 0.949       |
| time/                   |             |
|    fps                  | 704         |
|    iterations           | 89          |
|    time_elapsed         | 129         |
|    total_timesteps      | 91136       |
| train/                  |             |
|    approx_kl            | 0.096123055 |
|    clip_fraction        | 0.196       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.191      |
|    explained_variance   | 0.633       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0416     |
|    n_updates            | 880         |
|    policy_gradient_loss | -0.0249     |
|    value_loss           | 0.000407    |
-----------------------------------------
----------------------------------------
| rollout/                |            |
|    ep_len_mean          | 11.1       |
|    ep_rew_mean          | 0.961      |
| time/                   |            |
|    fps                  | 704        |
|    iterations           | 90         |
|    time_elapsed         | 130        |
|    total_timesteps      | 92160      |
| train/                  |            |
|    approx_kl            | 0.19162205 |
|    clip_fraction        | 0.145      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.15      |
|    explained_variance   | 0.256      |
|    learning_rate        | 0.0003     |
|    loss                 | 0.0508     |
|    n_updates            | 890        |
|    policy_gradient_loss | -0.0202    |
|    value_loss           | 0.00348    |
----------------------------------------
----------------------------------------
| rollout/                |            |
|    ep_len_mean          | 12.3       |
|    ep_rew_mean          | 0.957      |
| time/                   |            |
|    fps                  | 703        |
|    iterations           | 91         |
|    time_elapsed         | 132        |
|    total_timesteps      | 93184      |
| train/                  |            |
|    approx_kl            | 0.06836694 |
|    clip_fraction        | 0.347      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.264     |
|    explained_variance   | 0.581      |
|    learning_rate        | 0.0003     |
|    loss                 | -0.0699    |
|    n_updates            | 900        |
|    policy_gradient_loss | 0.066      |
|    value_loss           | 0.000154   |
----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.4        |
|    ep_rew_mean          | 0.957       |
| time/                   |             |
|    fps                  | 703         |
|    iterations           | 92          |
|    time_elapsed         | 133         |
|    total_timesteps      | 94208       |
| train/                  |             |
|    approx_kl            | 0.020714343 |
|    clip_fraction        | 0.161       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.218      |
|    explained_variance   | 0.581       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.01       |
|    n_updates            | 910         |
|    policy_gradient_loss | -0.00672    |
|    value_loss           | 0.000254    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 12.3        |
|    ep_rew_mean          | 0.957       |
| time/                   |             |
|    fps                  | 703         |
|    iterations           | 93          |
|    time_elapsed         | 135         |
|    total_timesteps      | 95232       |
| train/                  |             |
|    approx_kl            | 0.037652392 |
|    clip_fraction        | 0.184       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.199      |
|    explained_variance   | 0.896       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.0423     |
|    n_updates            | 920         |
|    policy_gradient_loss | -0.0175     |
|    value_loss           | 0.000112    |
-----------------------------------------
-----------------------------------------
| rollout/                |             |
|    ep_len_mean          | 11.2        |
|    ep_rew_mean          | 0.961       |
| time/                   |             |
|    fps                  | 702         |
|    iterations           | 94          |
|    time_elapsed         | 136         |
|    total_timesteps      | 96256       |
| train/                  |             |
|    approx_kl            | 0.080989845 |
|    clip_fraction        | 0.205       |
|    clip_range           | 0.2         |
|    entropy_loss         | -0.133      |
|    explained_variance   | 0.918       |
|    learning_rate        | 0.0003      |
|    loss                 | -0.05       |
|    n_updates            | 930         |
|    policy_gradient_loss | -0.0212     |
|    value_loss           | 8.25e-05    |
```

```
-----------------------------------------
-----------------------------------------
| rollout/                |            |
|    ep_len_mean          | 21         |
|    ep_rew_mean          | 0.923      |
| time/                   |            |
|    fps                  | 703        |
|    iterations           | 95         |
|    time_elapsed         | 138        |
|    total_timesteps      | 97280      |
| train/                  |            |
|    approx_kl            | 0.13105541 |
|    clip_fraction        | 0.316      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.254     |
|    explained_variance   | 0.971      |
|    learning_rate        | 0.0003     |
|    loss                 | -0.0671    |
|    n_updates            | 940        |
|    policy_gradient_loss | 0.197      |
|    value_loss           | 1.27e-05   |
-----------------------------------------
----------------------------------------
| rollout/                |           |
|    ep_len_mean          | 27.9      |
|    ep_rew_mean          | 0.897     |
| time/                   |           |
|    fps                  | 703       |
|    iterations           | 96        |
|    time_elapsed         | 139       |
|    total_timesteps      | 98304     |
| train/                  |           |
|    approx_kl            | 0.1051268 |
|    clip_fraction        | 0.194     |
|    clip_range           | 0.2       |
|    entropy_loss         | -0.144    |
|    explained_variance   | -0.119    |
|    learning_rate        | 0.0003    |
|    loss                 | -0.0391   |
|    n_updates            | 950       |
|    policy_gradient_loss | 0.162     |
|    value_loss           | 0.00478   |
----------------------------------------
----------------------------------------
| rollout/                |           |
|    ep_len_mean          | 15.9      |
|    ep_rew_mean          | 0.943     |
| time/                   |           |
|    fps                  | 704       |
|    iterations           | 97        |
|    time_elapsed         | 140       |
|    total_timesteps      | 99328     |
| train/                  |           |
|    approx_kl            | 0.5337625 |
|    clip_fraction        | 0.366     |
|    clip_range           | 0.2       |
|    explained_variance   | -0.0747   |
|    learning_rate        | 0.0003    |
|    learning_rate        | 0.0003    |
|    loss                 | -0.0265   |
|    n_updates            | 960       |
|    policy_gradient_loss | 0.308     |
|    value_loss           | 0.00451   |
----------------------------------------
Eval num_timesteps=100000, episode_reward=0.96 +/- 0.00
Episode length: 11.00 +/- 0.00
----------------------------------------
| eval/                   |            |
|    mean_ep_length       | 11         |
|    mean_reward          | 0.961      |
| time/                   |            |
|    total_timesteps      | 100000     |
| train/                  |            |
|    approx_kl            | 0.13666381 |
|    clip_fraction        | 0.168      |
|    clip_range           | 0.2        |
|    entropy_loss         | -0.155     |
|    explained_variance   | -0.427     |
|    learning_rate        | 0.0003     |
|    loss                 | -0.0696    |
|    n_updates            | 970        |
|    policy_gradient_loss | 0.0638     |
|    value_loss           | 0.00239    |
----------------------------------------
---------------------------------
| rollout/           |          |
|    ep_len_mean     | 13.1     |
|    ep_rew_mean     | 0.954    |
| time/              |          |
|    fps             | 705      |
|    iterations      | 98       |
|    time_elapsed    | 142      |
|    total_timesteps | 100352   |
---------------------------------
Training complete.
Model saved as ppo_minigrid_model.
Testing trained agent...
```

APPENDIX D

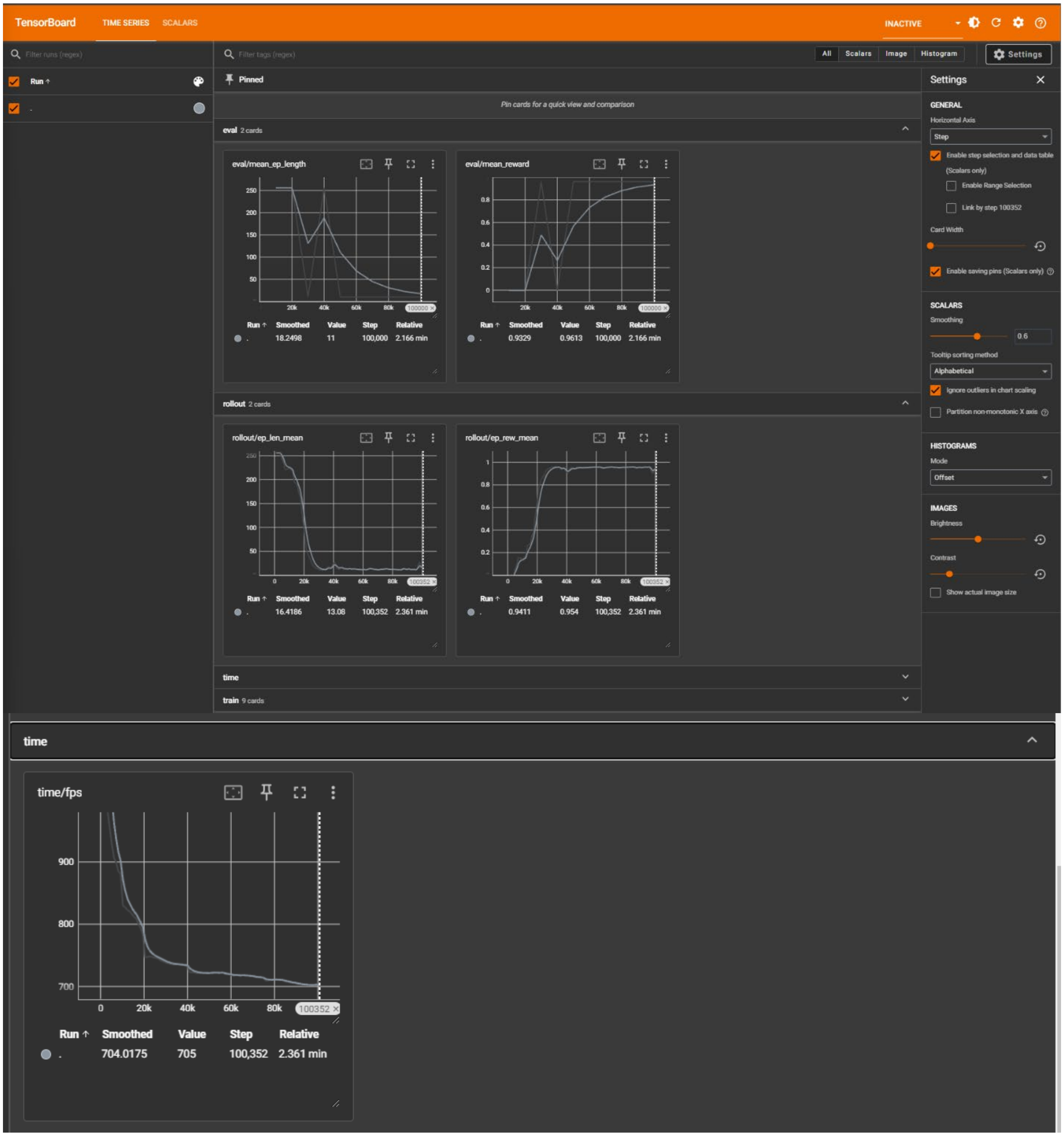
Render Mode = Human
This still image, captured during training in human render mode, illustrates the agent (red triangle) navigating a grid environment toward the green goal square as per the specified task.

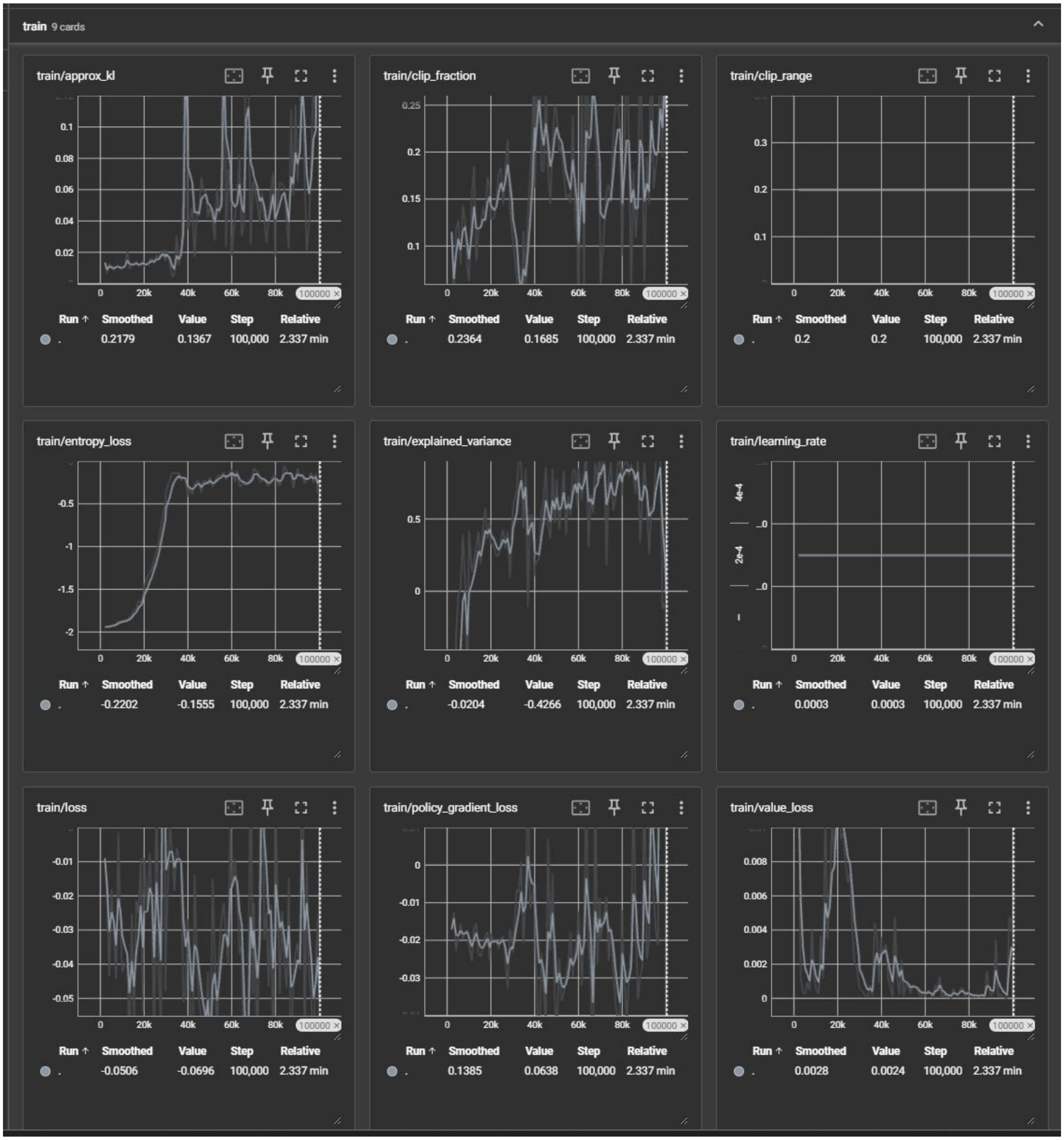


APPENDIX E

TensorBoard Logs – 1st Visualized Training Metrics

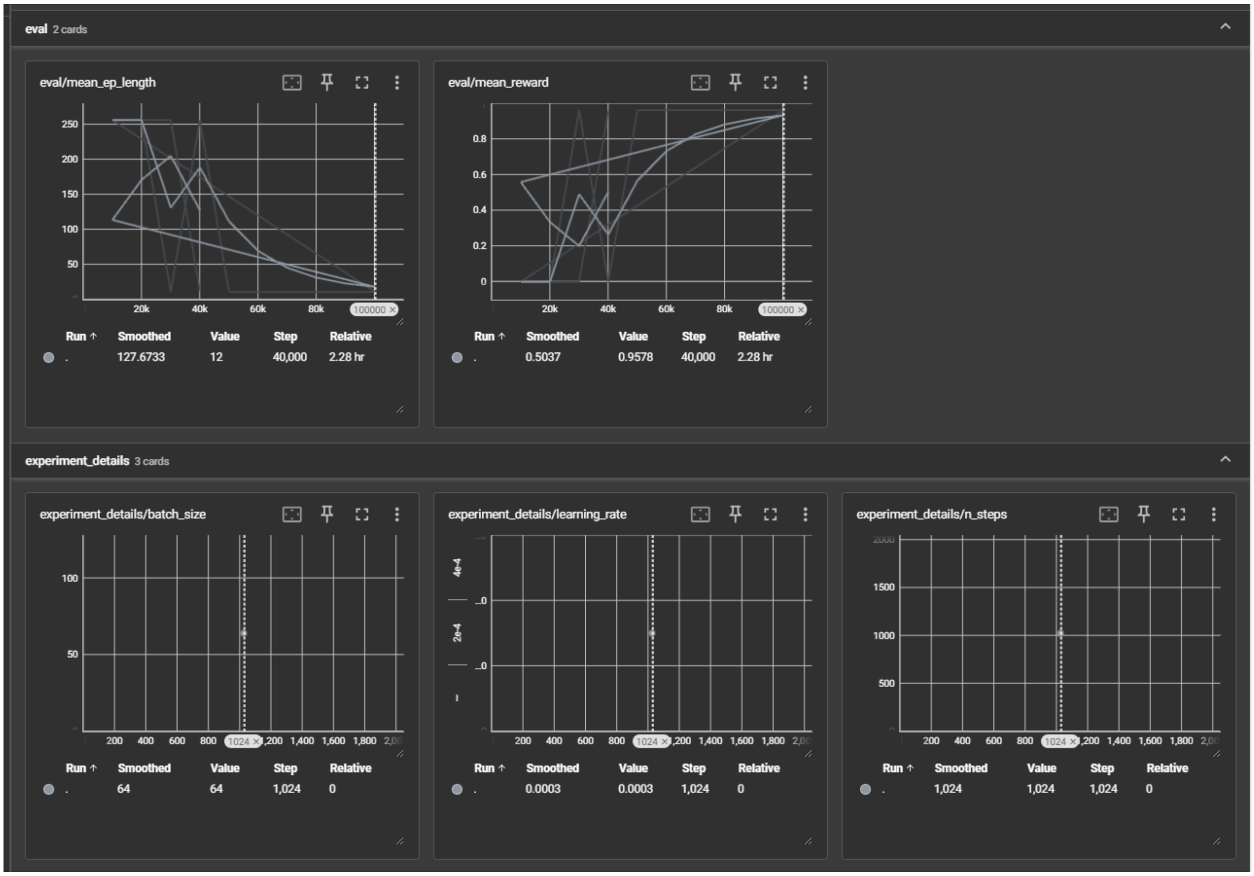
This appendix contains detailed logs from the initial evaluation phase, highlighting the agent's performance metrics, episode outcomes, and termination conditions during validation runs.

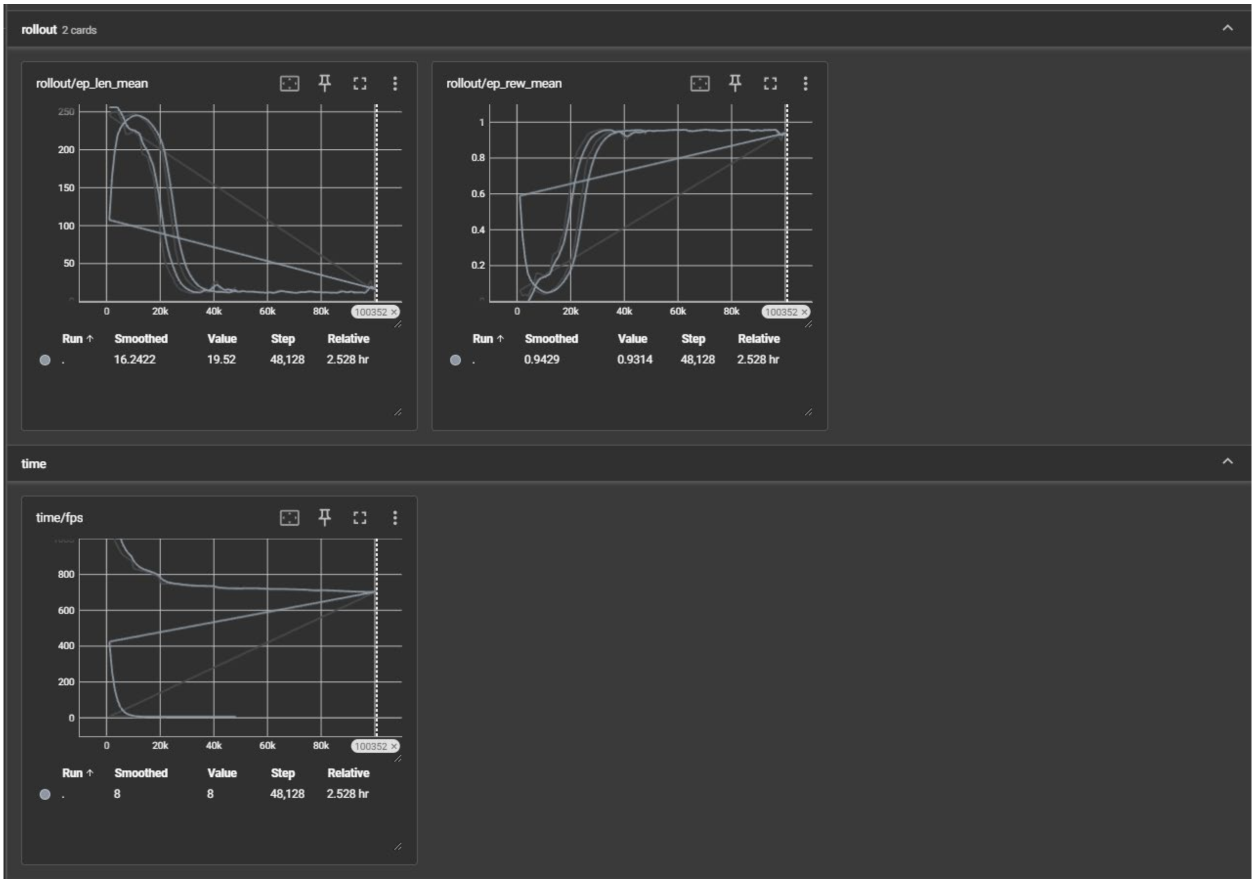


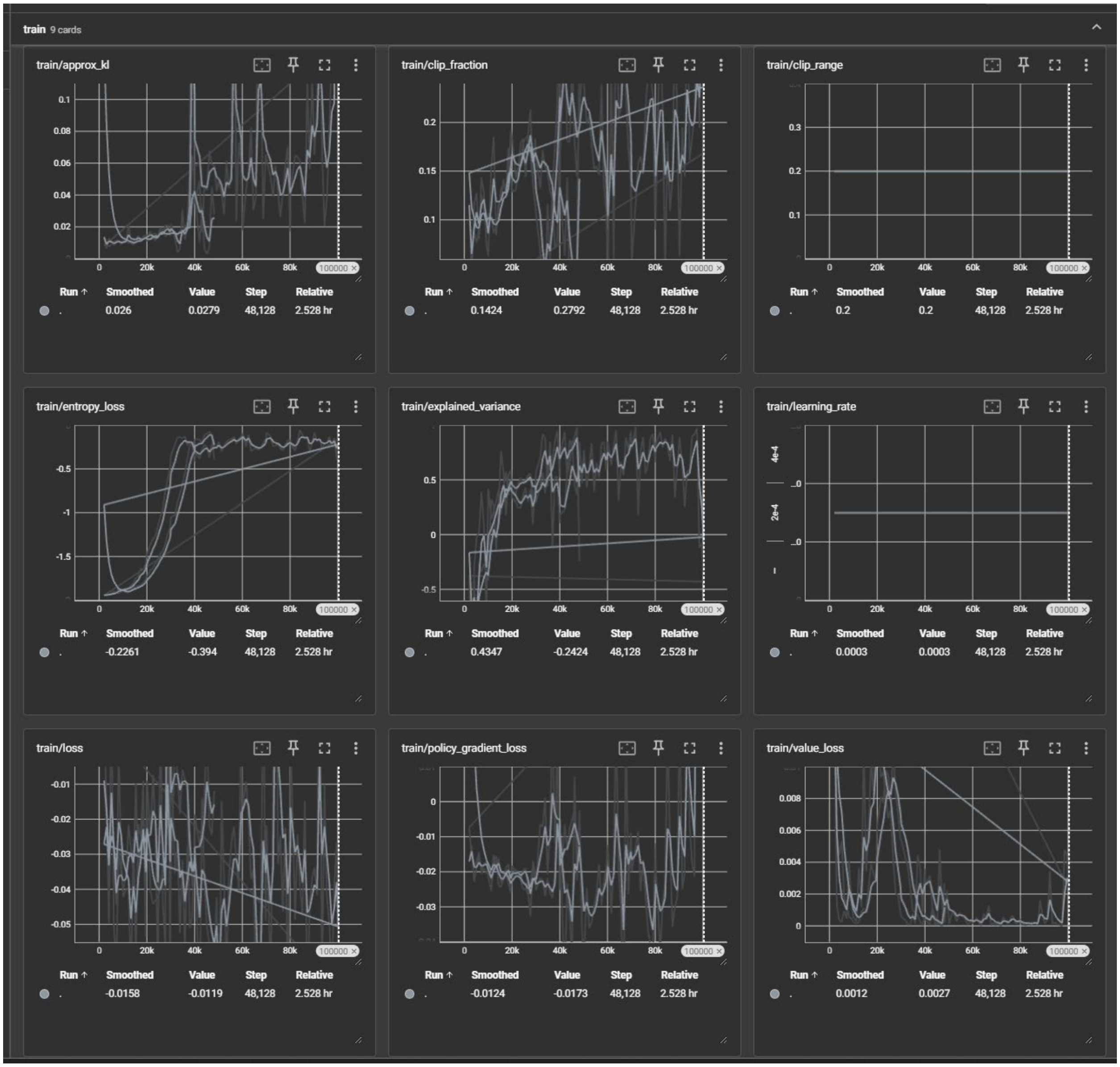

train 9 cards
train/approx_kl
Run ↑ Smoothed Value Step Relative
. 0.2179 0.1367 100,000 2.337 min
train/clip_fraction
Run ↑ Smoothed Value Step Relative
. 0.2364 0.1685 100,000 2.337 min
train/clip_range
Run ↑ Smoothed Value Step Relative
. 0.2 0.2 100,000 2.337 min
train/entropy_loss
Run ↑ Smoothed Value Step Relative
. -0.2202 -0.1555 100,000 2.337 min
train/explained_variance
Run ↑ Smoothed Value Step Relative
. -0.0204 -0.4266 100,000 2.337 min
train/learning_rate
Run ↑ Smoothed Value Step Relative
. 0.0003 0.0003 100,000 2.337 min
train/loss
Run ↑ Smoothed Value Step Relative
. -0.0506 -0.0696 100,000 2.337 min
train/policy_gradient_loss
Run ↑ Smoothed Value Step Relative
. 0.1385 0.0638 100,000 2.337 min
train/value_loss
Run ↑ Smoothed Value Step Relative
. 0.0028 0.0024 100,000 2.337 min

APPENDIX F

TensorBoard Logs – 2nd Visualized Training Metrics + Additional Details
This appendix provides comprehensive logs from the training runs, documenting the agent's progress, action sequences, rewards, and termination conditions to support the analysis of training performance.




rollout 2 cards
rollout/ep_len_mean
Run ↑ Smoothed Value Step Relative
. 16.2422 19.52 48,128 2.528 hr
rollout/ep_rew_mean
Run ↑ Smoothed Value Step Relative
. 0.9429 0.9314 48,128 2.528 hr
time
time/fps
Run ↑ Smoothed Value Step Relative
. 8 8 48,128 2.528 hr


train 9 cards
train/approx_kl
Run ↑ Smoothed Value Step Relative
. 0.026 0.0279 48,128 2.528 hr
train/clip_fraction
Run ↑ Smoothed Value Step Relative
. 0.1424 0.2792 48,128 2.528 hr
train/clip_range
Run ↑ Smoothed Value Step Relative
. 0.2 0.2 48,128 2.528 hr
train/entropy_loss
Run ↑ Smoothed Value Step Relative
. -0.2261 -0.394 48,128 2.528 hr
train/explained_variance
Run ↑ Smoothed Value Step Relative
. 0.4347 -0.2424 48,128 2.528 hr
train/learning_rate
Run ↑ Smoothed Value Step Relative
. 0.0003 0.0003 48,128 2.528 hr
train/loss
Run ↑ Smoothed Value Step Relative
. -0.0158 -0.0119 48,128 2.528 hr
train/policy_gradient_loss
Run ↑ Smoothed Value Step Relative
. -0.0124 -0.0173 48,128 2.528 hr
train/value_loss
Run ↑ Smoothed Value Step Relative
. 0.0012 0.0027 48,128 2.528 hr