Mid Project Progress Report

Colab Notebook:
https://colab.research.google.com/drive/1azu8OjyrxSsDT2tlJvU5ALy9CxYh3Qa5?usp=sharing

Github Repo: ShivomP/CSC-496-Final-Project

As a refresher, the goal of my project is to reproduce the standard PPO on ALE/Pong-v5 using Stable-Baseline3 as the baseline, then implement my own rendition of the PPO algorithm by changing 3 features and reporting the effects. At this point I have written all the code for the baseline and evaluation metrics that I planned and started to train the baseline. Sample efficiency is measured as the number of environment steps required to reach a target return threshold of 18 or more. I plan to run 5 random seeds for each implementation (baseline PPO, my PPO) to account for variance in reinforcement learning. So I got to run 1 iteration of the baseline on seed 42 at this point and have seemed to come across some troubles.The learning curve shows that captured training data represents only the initial 25,000 steps, during which the agent exhibited random play behavior (mean return: -20.70). The evaluation results demonstrate that learning occurred during the uncaptured portion of training (steps 25K-10M), with the final agent achieving a mean return of 6.95. This difference between early training metrics and final evaluation performance confirms that learning progression occurred but was not recorded due to episode buffer constraints. The evaluation data provides the valid measure of training outcome. You can also reference the colab notebook I have linked for further information regarding the training results. I plan to make a copy in my continued training as I feel I need to make some adjustments to my code. I am aware that I need to fix my callback to better capture the data for the learning curve and that will be the first adjustment I make before running seed 43. I also did try to increase the number of environments to speed up training but going from 8 to 16 only decreased time by roughly 1.5 hours. Any advice would be greatly appreciated.

Preliminary Results - Baseline PPO (Seed 42)
Total environment steps: 10,000,000
Wall-clock time: 9 hours 16 minutes

Evaluation Performance (20 episodes, deterministic policy):
Mean return: 6.95 ± 5.37
Median return: 7.50
Performance range: -10 to +15

Best episode: +15 (83% of target threshold)