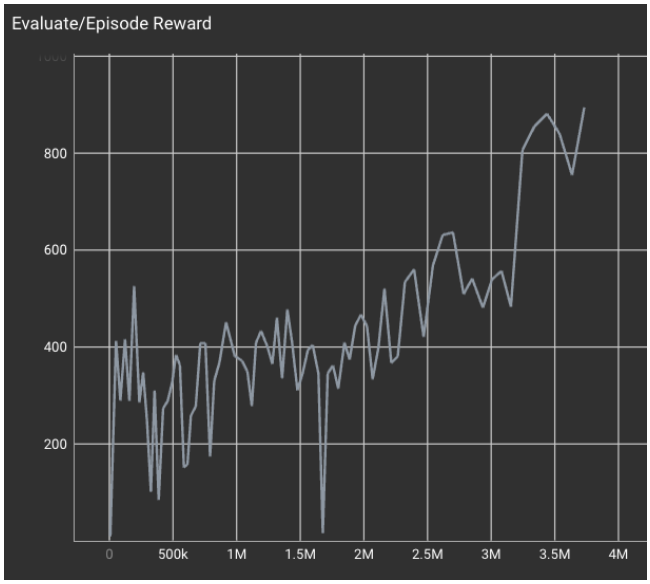
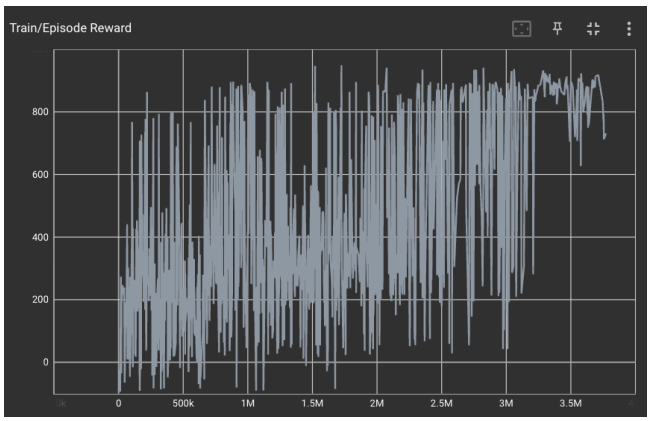
Lab4_311352004_童政瑜_report

1. Experimental Results





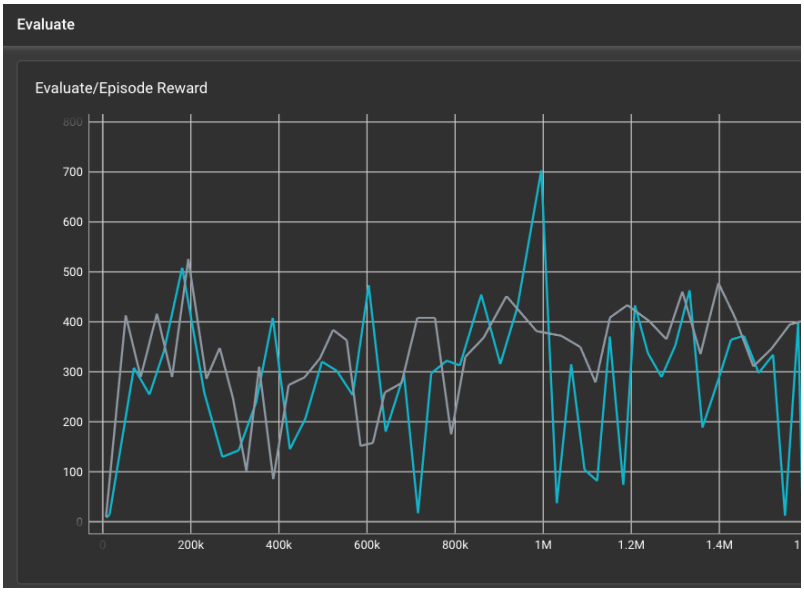

```
================================================
Evaluating...
Episode: 1  Length: 999 Total reward: 892.65
Episode: 2  Length: 999 Total reward: 882.21
Episode: 3  Length: 999 Total reward: 854.40
Episode: 4  Length: 746 Total reward: 925.30
Episode: 5  Length: 999 Total reward: 876.59
Episode: 6  Length: 999 Total reward: 876.74
Episode: 7  Length: 999 Total reward: 886.11
Episode: 8  Length: 793 Total reward: 920.60
Episode: 9  Length: 999 Total reward: 866.44
Episode: 10 Length: 999 Total reward: 832.86
average score: 881.3899928407151
================================================
```

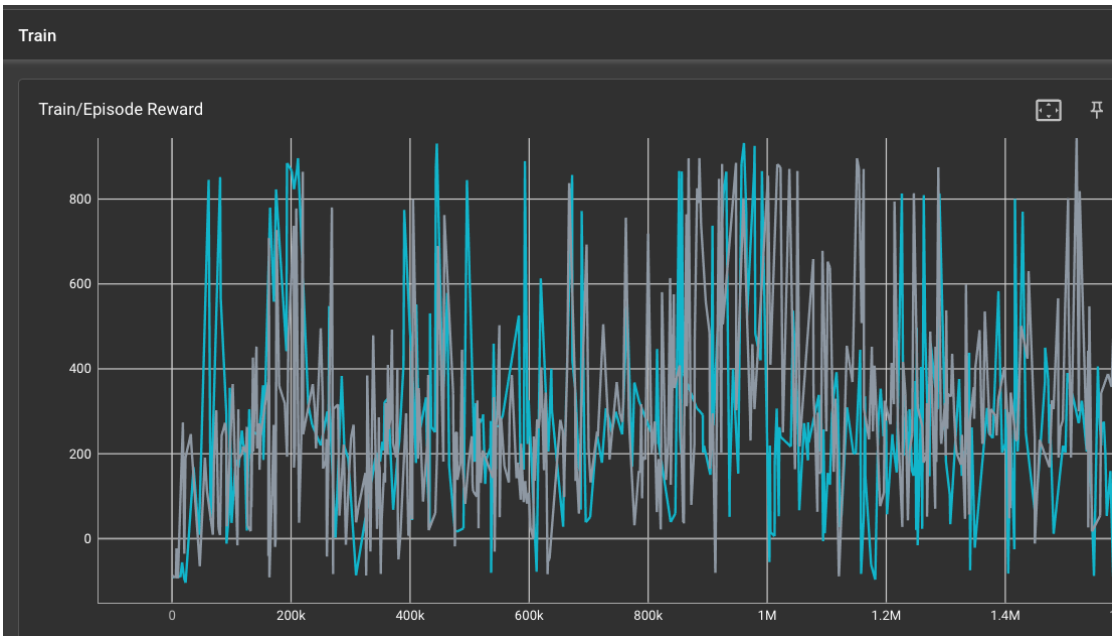
2. Bonus

(1) Screenshot of Tensorboard training curve and compare the performance of using twin Q-networks and single Q-networks in TD3, and explain.

Twin Q-networks : grey line

Single Q-network : blue line

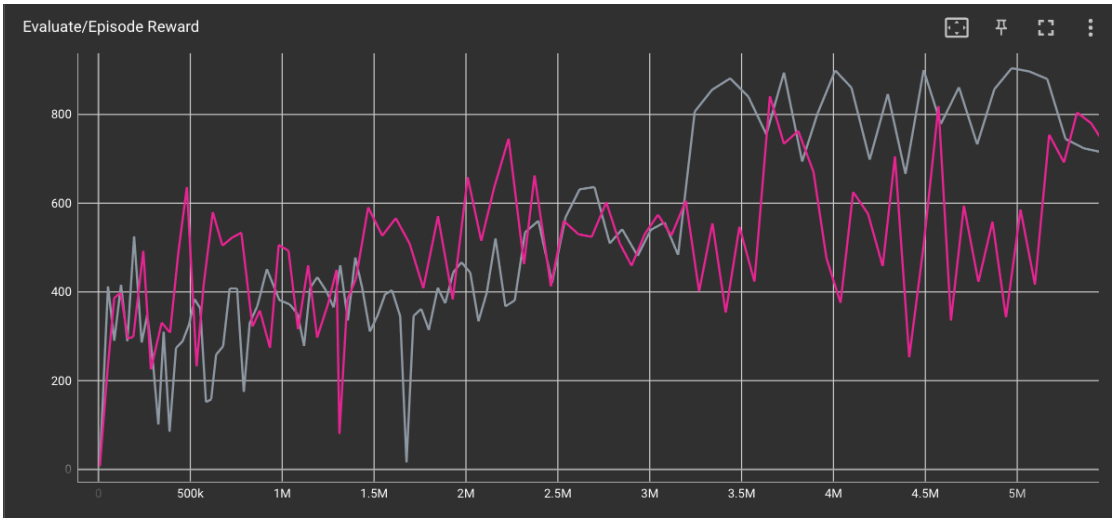




Single Q-Network : The training curve of the single Q-network has significant variability, with several spikes in performance. The trend seems to increase over time, which indicates that the model is learning, but the high variance suggests that the learning process might be unstable or that the model may be overestimating the Q-values.

Twin Q-Networks : In comparison, the training curve for the TD3 strategy using twin Q-networks also shows an increasing trend over time but with seemingly less variance than the single Q-network. The use of twin Q-networks in TD3 aims to mitigate the overestimation of Q-values by taking the minimum estimate from two separate Q-networks. This usually results in a more stable learning process, which seems to be indicated by the smoother curve.

(2) Screenshot of Tensorboard training curve and compare the impact of enabling and disabling target policy smoothing in TD3, and explain.
Enable smoothing : grey line
Disable smoothing : pink line



With Smoothing : When smoothing is enabled, the target policy is expected to have less variance in its performance, leading to more stable learning. It helps the algorithm to generalize better to different states by reducing

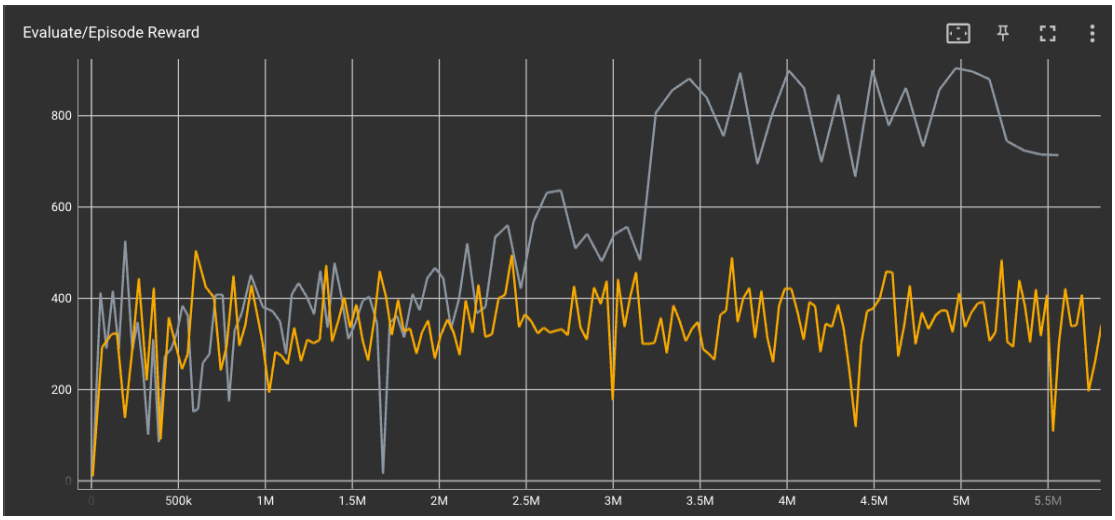
overfitting to the sampled data.

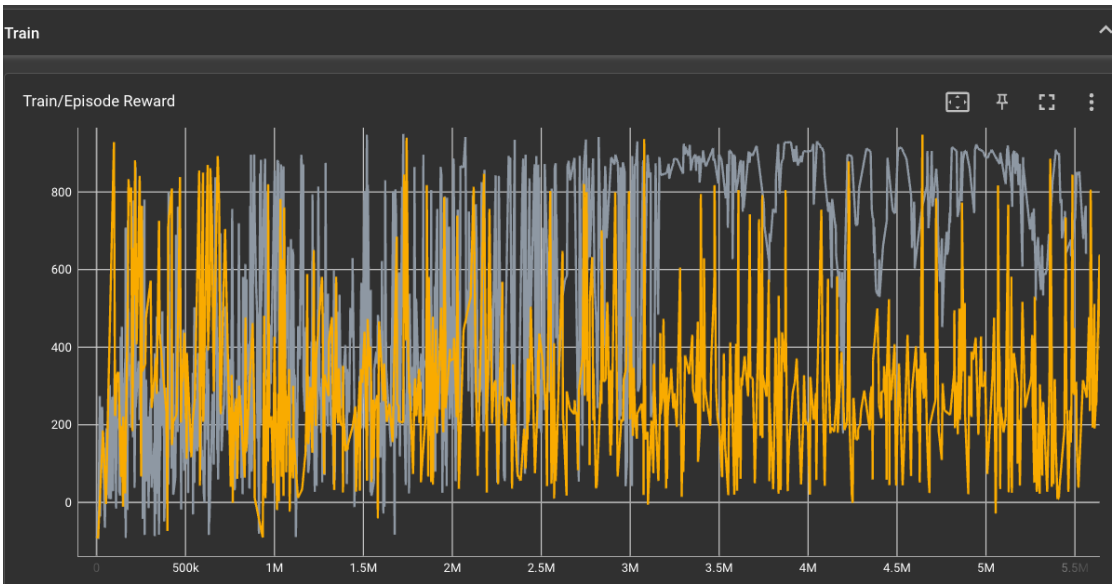
Without Smoothing : Disabling smoothing can lead to a more erratic learning process with higher variance in performance. It can cause the policy to be more shortsighted, focusing on immediate rewards that it has overestimated rather than a more consistent long-term strategy.

(3) Screenshot of Tensorboard training curve and compare the impact of delayed update steps and compare the results, and explain.

Delayed : grey line

Non-Delayed : yellow line





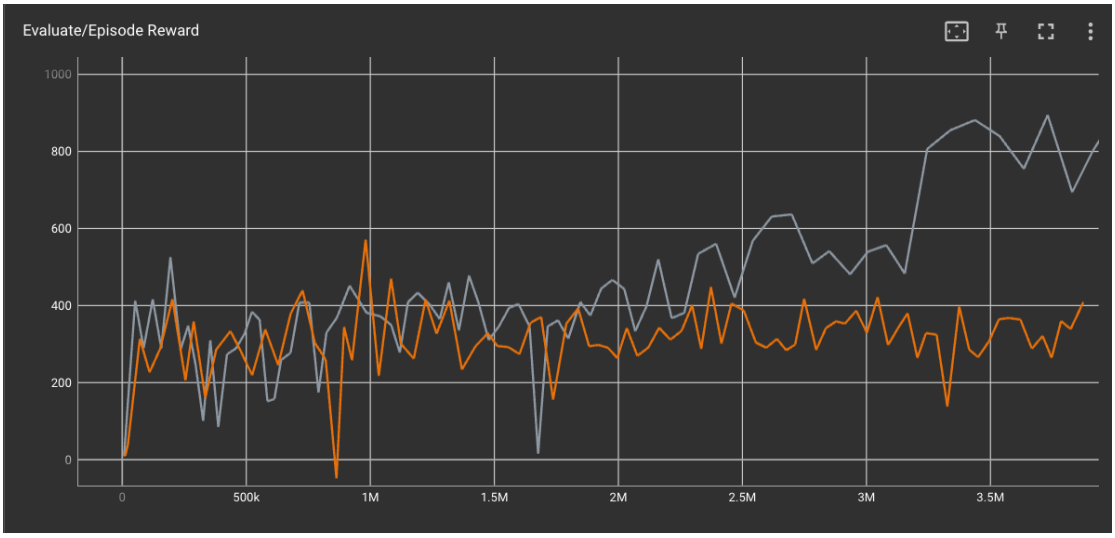
Non-Delayed Update : In both the training and evaluation graphs, the yellow line exhibits a greater degree of fluctuation compared to the grey line. This suggests higher variability in performance and possibly a more reactive policy that might be more prone to overfitting or being influenced by noisy estimates of the Q-values.

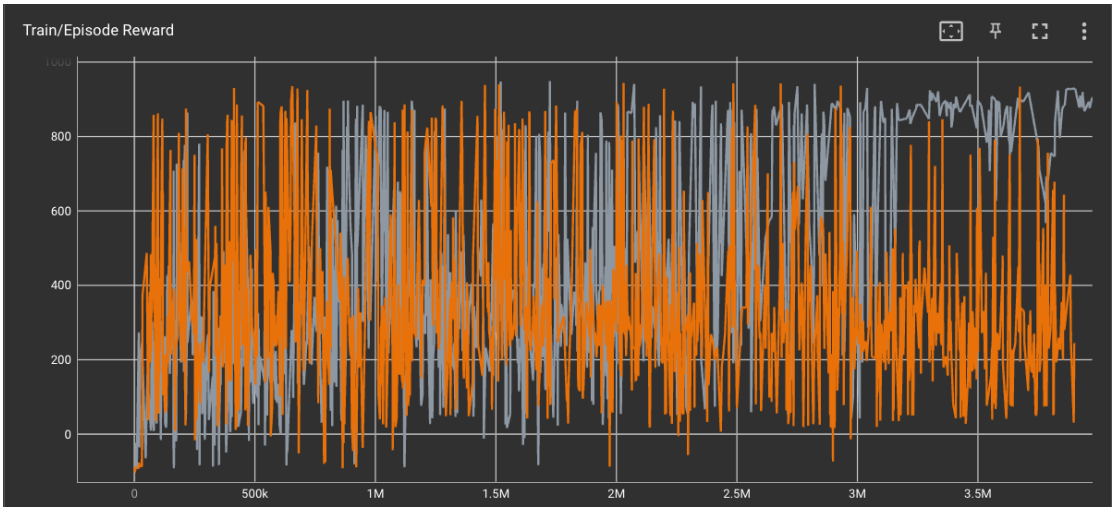
Delayed Update : In TD3, policy updates are delayed (less frequent than value updates) to reduce variance and overestimation. The grey line appears to be smoother with less severe drops in performance, indicating a more stable learning process and a policy that potentially generalizes better.

(4) Screenshot of Tensorboard training curve and compare the effects of adding different levels of action noise (exploration noise) in TD3, and explain.

Using Gaussian noise : grey line

Using OU noise : orange line





OU Noise : OU noise is typically used in scenarios where we want to simulate more realistic physical processes, such as control tasks where momentum is important, because it has temporal correlations. The orange line, representing the OU noise, shows more variability in the training curve. This might indicate that OU noise is driving the policy to explore more diverse parts of the state-action space, which can be beneficial in complex environments but might also introduce more variance in the learning process.

Gaussian Noise : Gaussian noise is a more conventional choice for exploration in reinforcement learning tasks. It is uncorrelated, meaning it does not take previous actions into account, which can make exploration more random. The grey line in the graphs appears to be smoother compared to the orange line, suggesting that Gaussian noise may lead to a more stable exploration process, although potentially less targeted than OU noise in terms of exploiting the physical correlations in the environment.