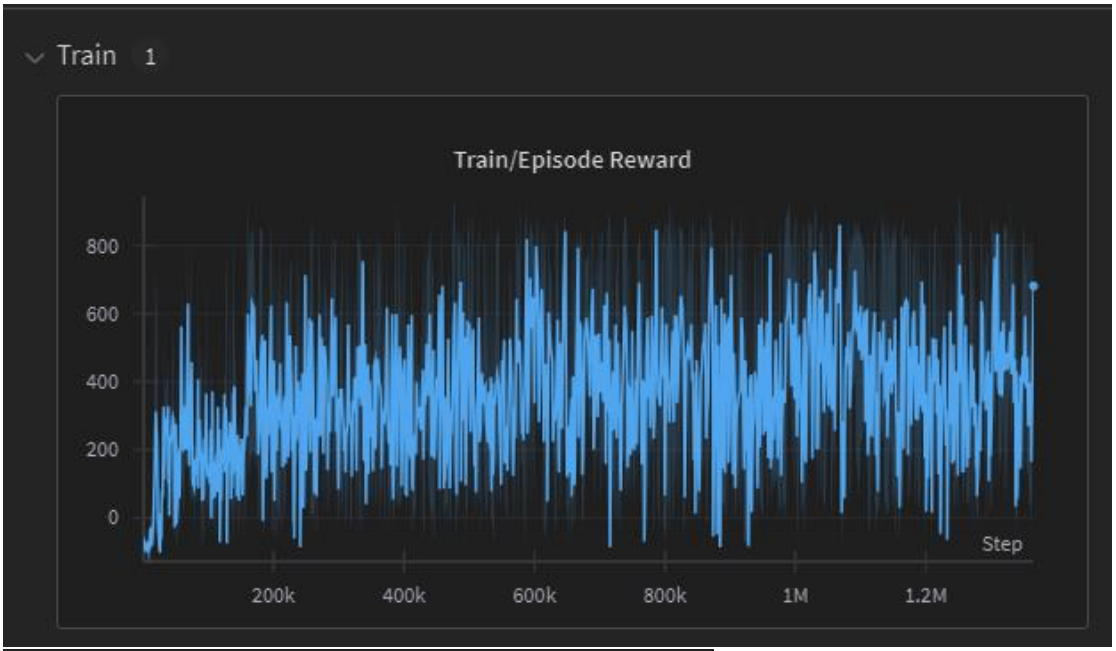
RL Topic HW3

1. Training curve and testing result of TD3:

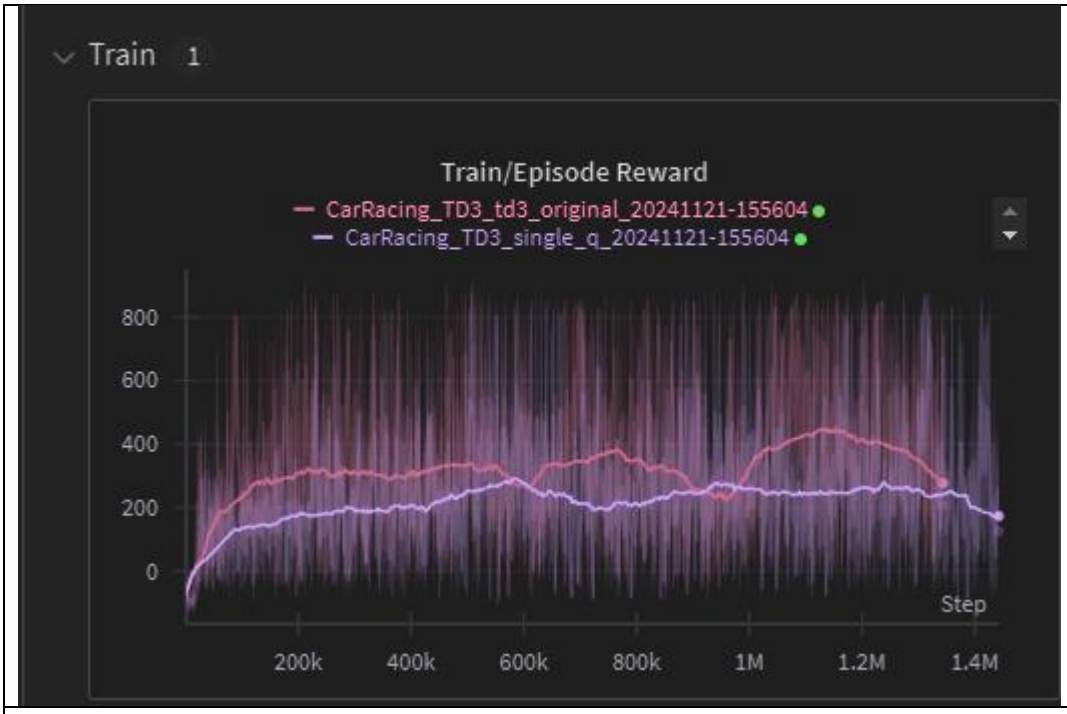



```
==========================================
Evaluating...
Episode: 1 Length: 252 Total reward: 258.19
Episode: 2 Length: 344 Total reward: 411.36
Episode: 3 Length: 146 Total reward: 132.90
Episode: 4 Length: 169 Total reward: 170.28
Episode: 5 Length: 999 Total reward: 789.29
Episode: 6 Length: 465 Total reward: 628.14
Episode: 7 Length: 999 Total reward: 812.21
Episode: 8 Length: 449 Total reward: 526.43
Episode: 9 Length: 294 Total reward: 329.23
Episode: 10 Length: 337 Total reward: 444.77
average score: 450.27996950458044
==========================================
```

2. Questions:

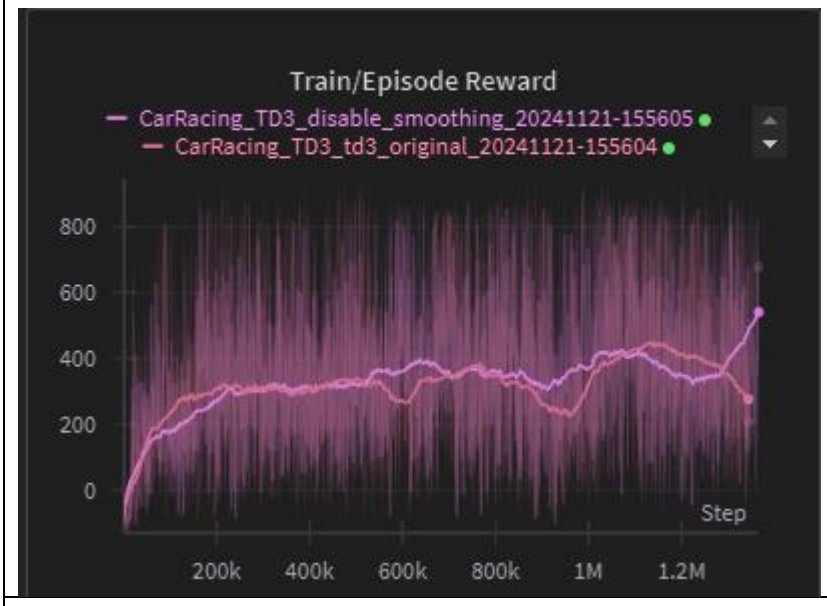
a. Single vs Double Q network

Twin Q-Networks in TD3 improve performance by reducing overestimation bias through the use of the minimum Q-value from two networks, leading to more accurate value estimates and stable learning. In contrast, a Single Q-Network is simpler but prone to overestimation, which can result in suboptimal policies and less stable training. Twin Q-networks trade increased computational cost for better reliability and performance.



b. Disable target policy smoothing

Enabling target policy smoothing in TD3 improves stability by adding noise to target actions, making the Q-network less sensitive to small action errors and reducing overfitting to deterministic policies. Disabling it can lead to less robust policies, as the Q-network may overfit to specific actions, making training less stable in noisy environments.



c. Disable delayed update

Delayed update steps in TD3 improve stability by updating the policy and target networks less frequently than the Q-networks. This prevents the policy from being overly influenced by noisy or unstable Q-value estimates. Without delayed updates, the policy might overfit

to fluctuating Q-values, leading to instability and poorer performance. Delayed updates trade slower updates for more robust and reliable learning.