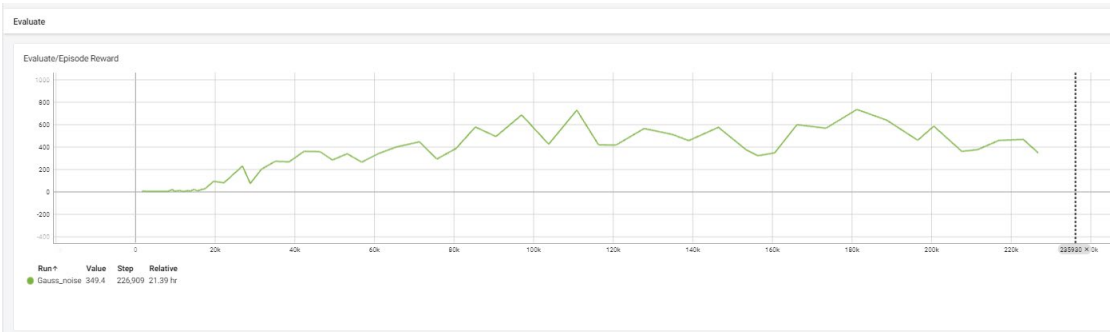
# Lab 4: Twin Delayed DDPG (TD3)

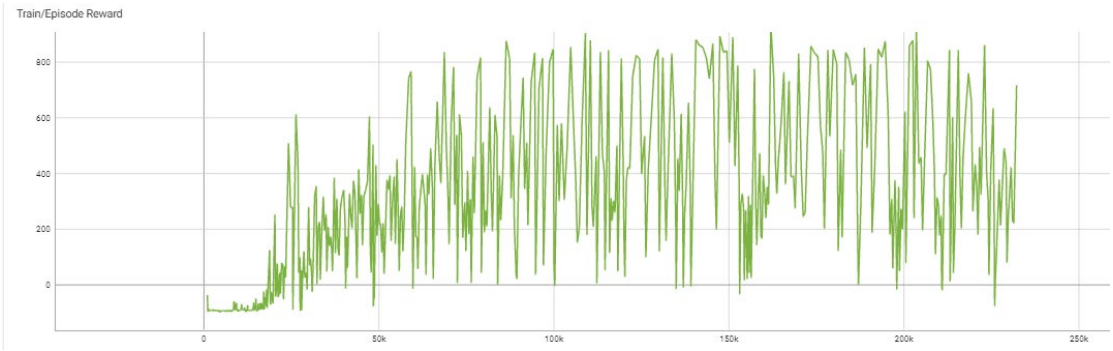
學生: 陳澤昕

學號 : 311356003

## Experimental Results (30%)

- Training curve:

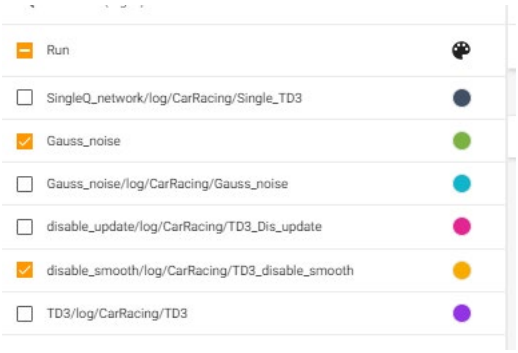


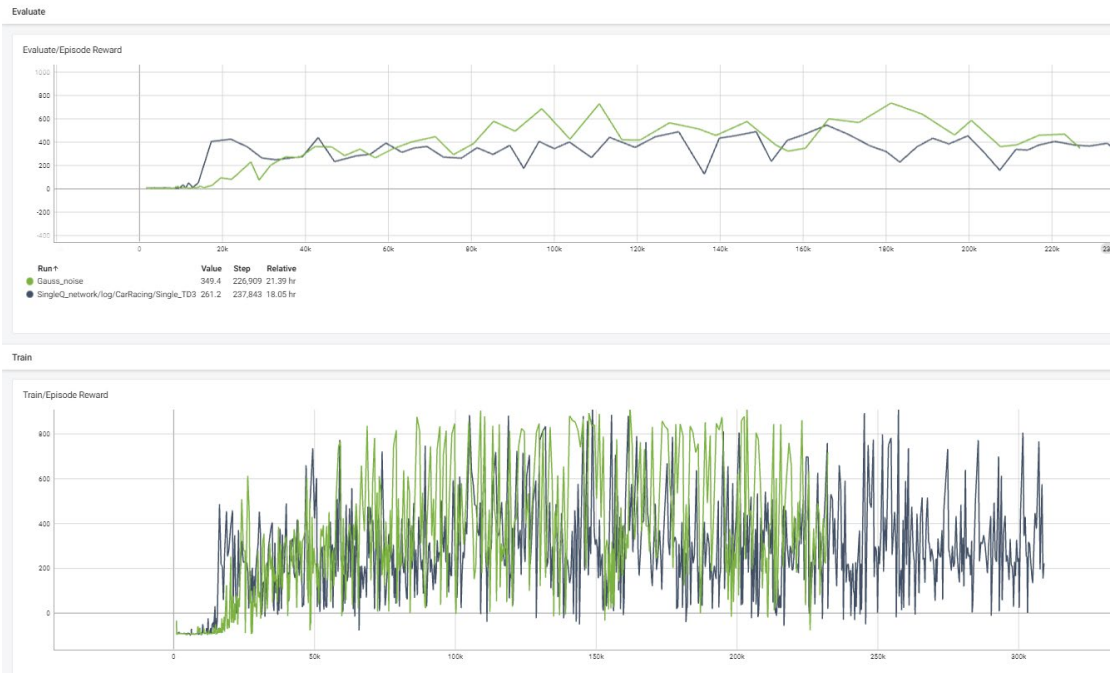

- Testing results (10 games):

```
Evaluating...
Episode: 1  Length: 682 Total reward: 691.42
Episode: 2  Length: 999 Total reward: 851.39
Episode: 3  Length: 999 Total reward: 870.80
Episode: 4  Length: 699 Total reward: 579.07
Episode: 5  Length: 999 Total reward: 843.46
Episode: 6  Length: 999 Total reward: 869.07
Episode: 7  Length: 479 Total reward: 538.67
Episode: 8  Length: 912 Total reward: 876.23
Episode: 9  Length: 790 Total reward: 835.40
Episode: 10 Length: 453 Total reward: 346.58
average score: 730.2085092645808
```

**Experimental Results and Discussion of bonus parts (Impact of Twin Q-Networks, Target Policy Smoothing, Delayed Policy Update Mechanism, Action Noise Injection) (bonus) (30%)**




(1) Screenshot of Tensorboard training curve and compare the performance of using twin Q-networks and single Q-networks in TD3, and explain (5%).




The twin Q-networks improve upon single Q-networks by reducing the overestimation bias. This is achieved by using two separate networks to independently estimate Q-values, and then taking the minimum of these two estimates. This approach leads to more accurate and stable value estimations, enhancing the overall performance and reliability of the learning process in complex environments.

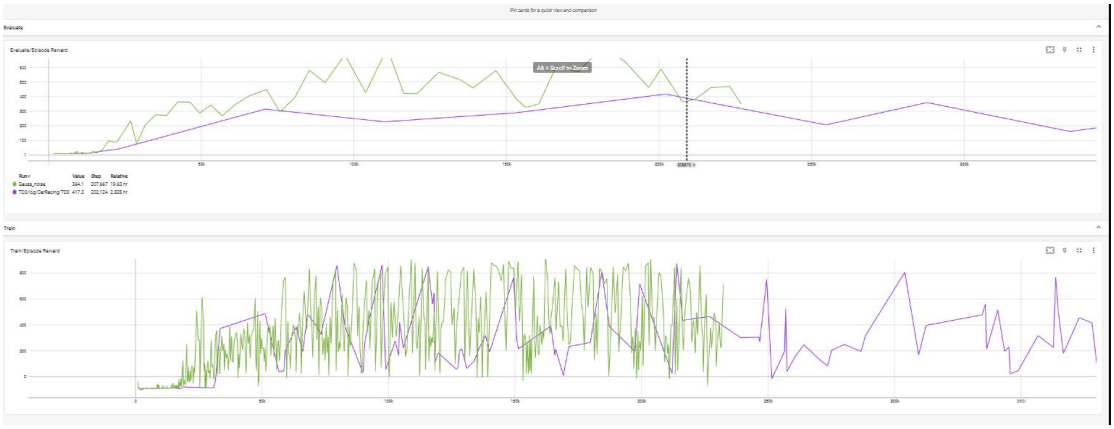
(2) Screenshot of Tensorboard training curve and compare the impact of enabling and disabling target policy smoothing in TD3, and explain (5%).


Target policy smoothing in TD3 improves the stability and robustness of the learning process by preventing the policy from exploiting small errors in the Q-function estimation. This leads to more accurate and reliable policy development, especially in complex and noisy environments.

(3) Screenshot of Tensorboard training curve and compare the impact of delayed update steps and compare the results, and explain (5%).


Delayed update steps in TD3 improve the stability of learning by reducing the frequency of policy updates, which helps in mitigating the risk of overfitting to recent experiences and ensures more reliable policy improvement over time.

(4) Screenshot of Tensorboard training curve and compare the effects of adding different levels of action noise (exploration noise) in TD3, and explain (5%).




In TD3, the difference between OU (Ornstein-Uhlenbeck) noise and Gaussian noise lies in their characteristics: OU noise is temporally correlated, providing consistent, smoother exploration, whereas Gaussian noise is independent at each time step, offering more varied and random exploration.

(5) Screenshot of Tensorboard training curve and compare your reward function with the original one and explain why your reward function works better. (10%).