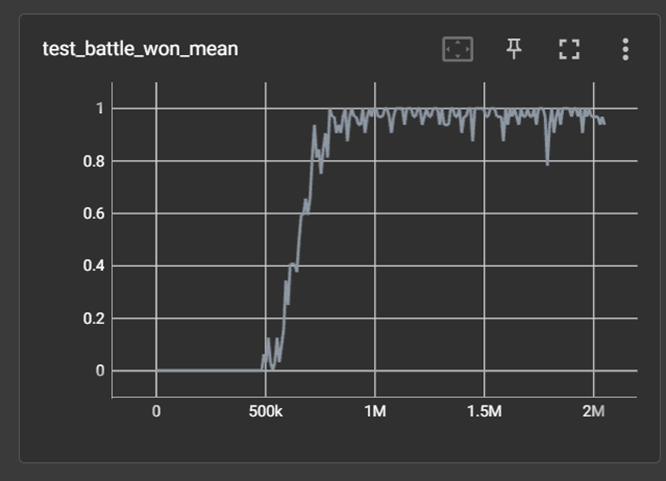
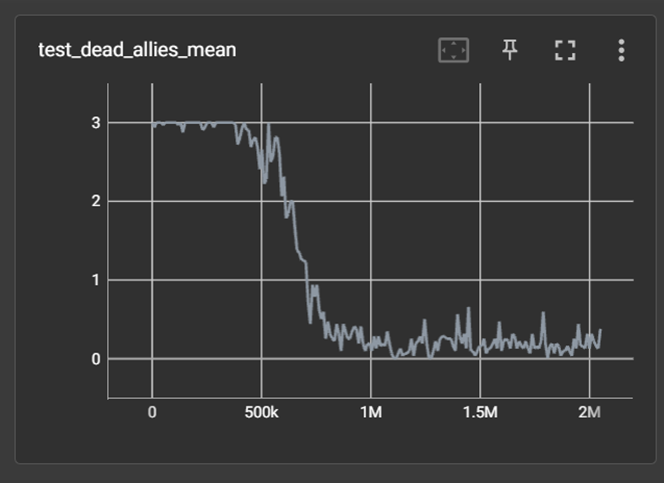
SMAC2 选用地图：3m\_vs\_6z

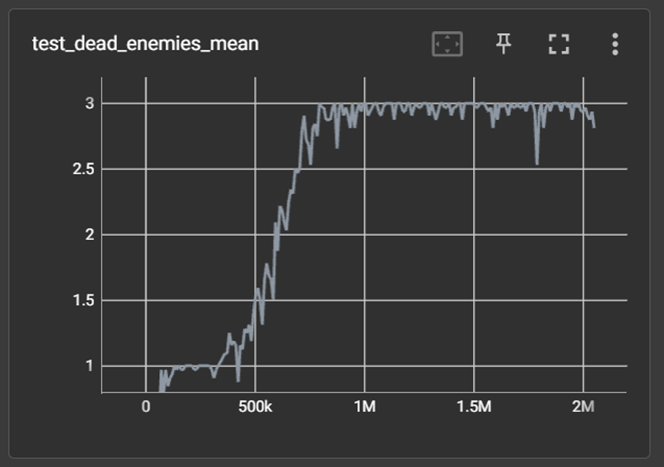
**######单行为策略：**

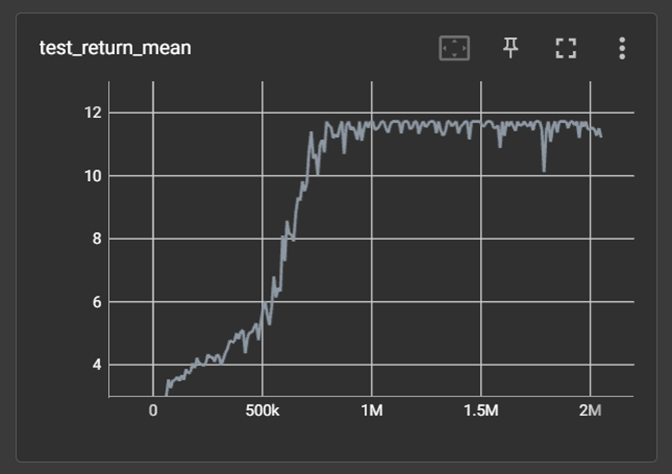
**expert**

此为没有在环境中进行reward shaping，用QMIX在线训练得到的策略，收集了4000 episodes离线数据。后面多行为策略训练中，expert 1, expert 2分别进行不同的reward shaping，偏好攻击enemy 0~2和enemy 3~5，各收集4000 episodes。



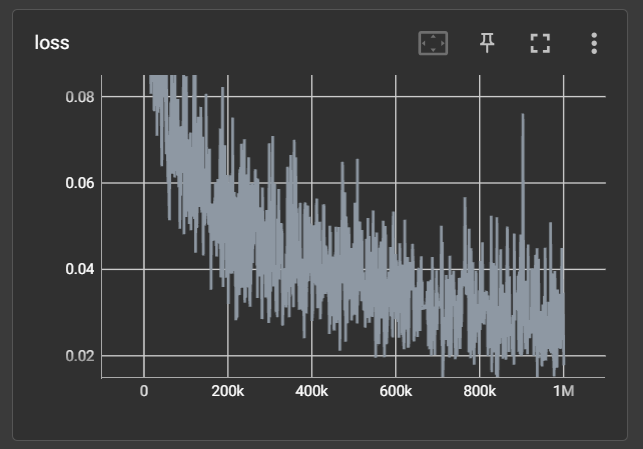


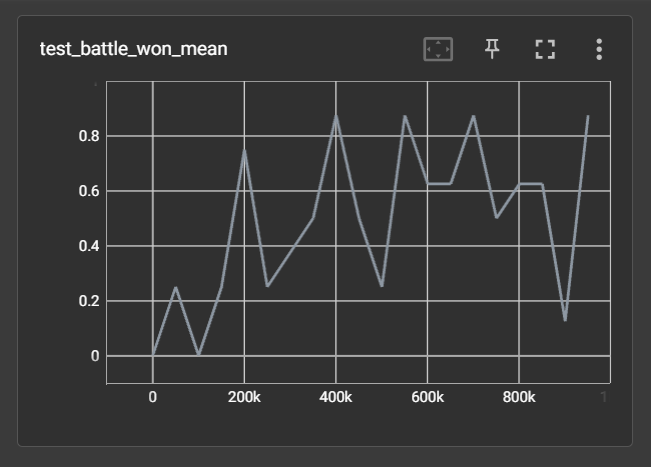


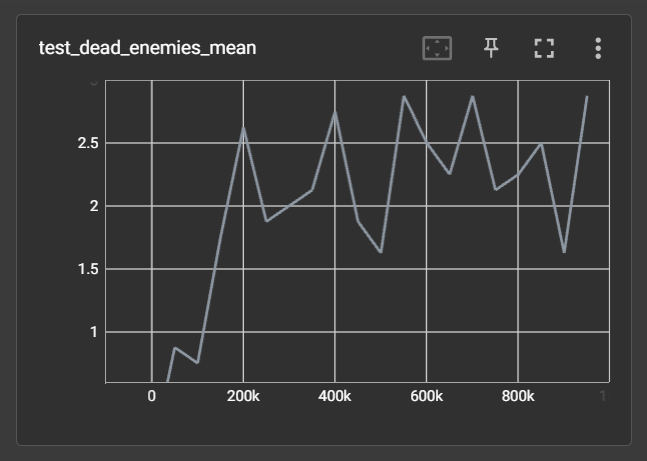


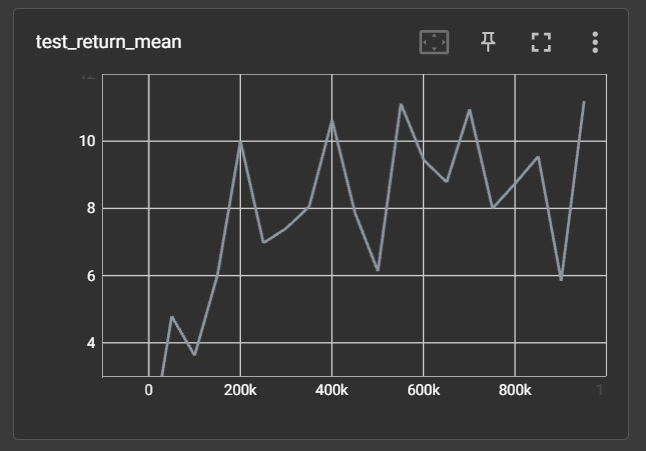
**madiff - 1 expert**

在expert生成的数据集上离线训练，没有reward shaping。



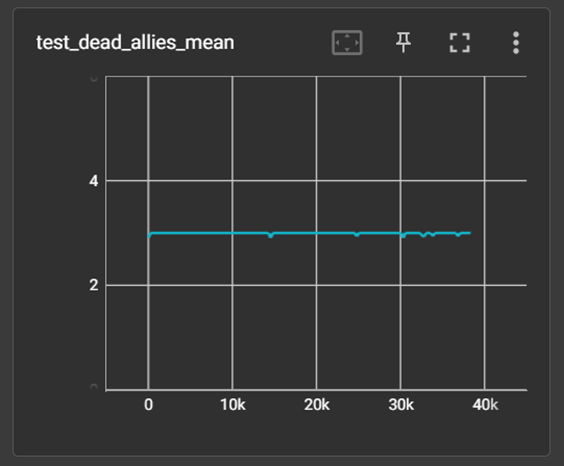


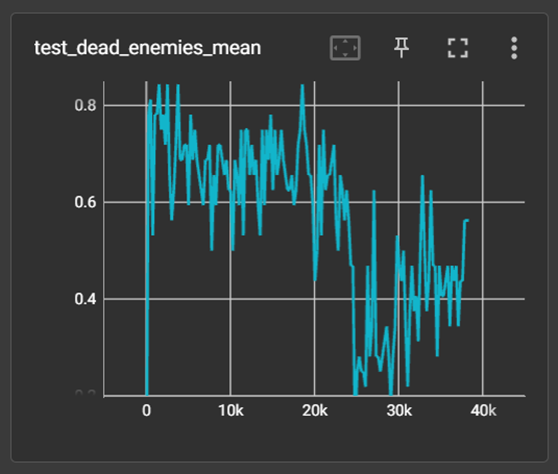


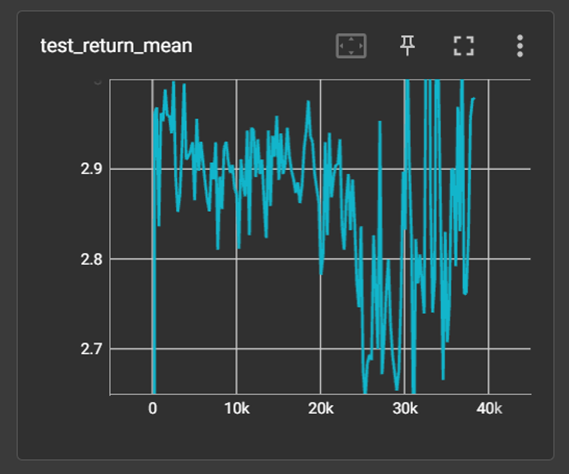


**BC**

BC基本上不work。battle\_won\_mean一直是0.







注：

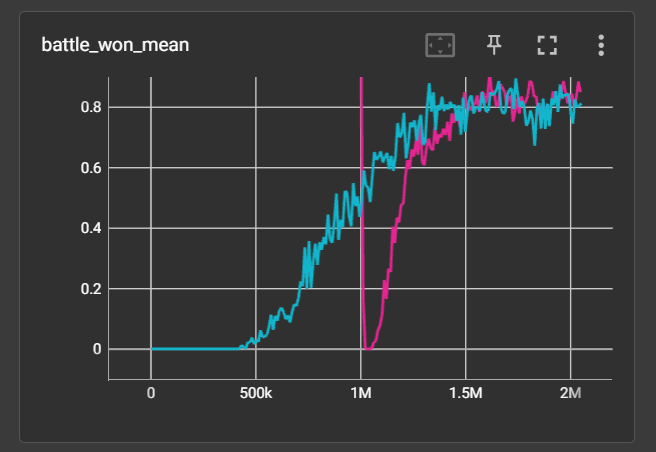
除此之外进行了双行为策略、加上数据id的BC训练，各指标与此基本相同。

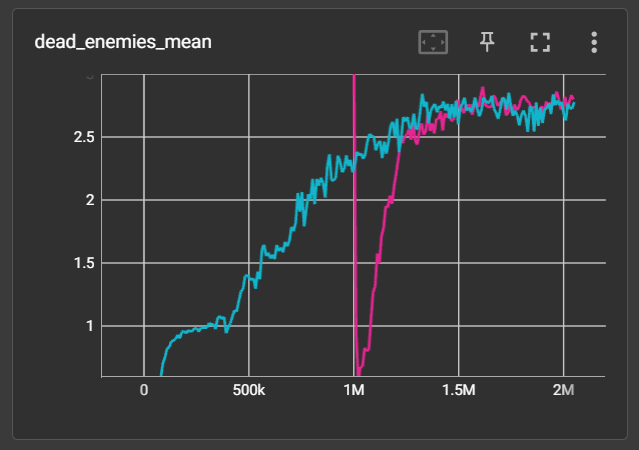
另外还用QMIX\_CQL分别进行了单行为策略、无数据id的训练，和双行为策略、有数据id的训练，前者能恢复大部分性能（test\_return\_mean稳定于9），后者仍然不work（test\_return\_mean从6降到4以下）。

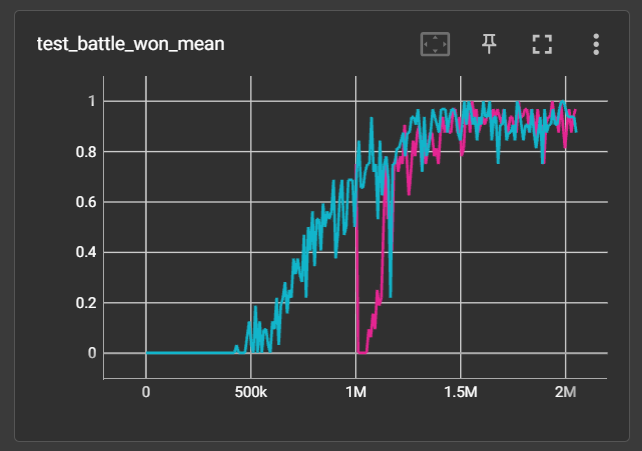
**######多行为策略：**

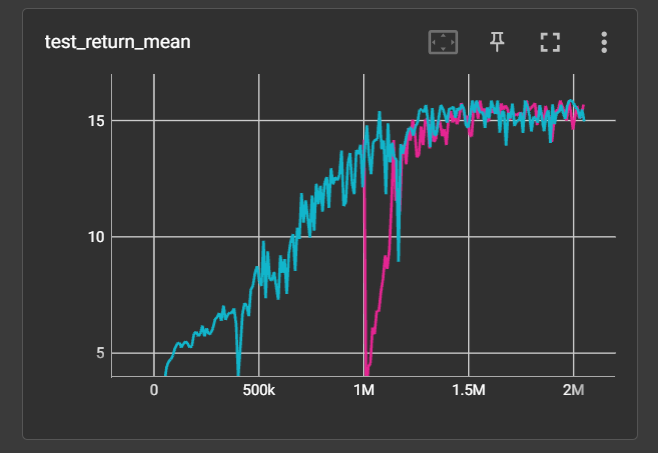
**expert 1 & expert 2**

（expert 2训练中途有打断恢复）



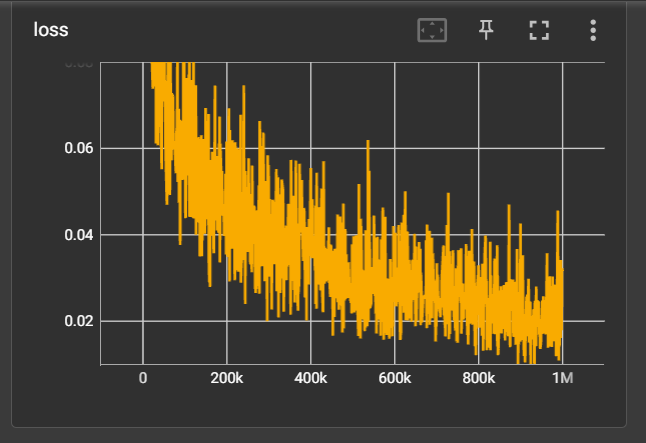


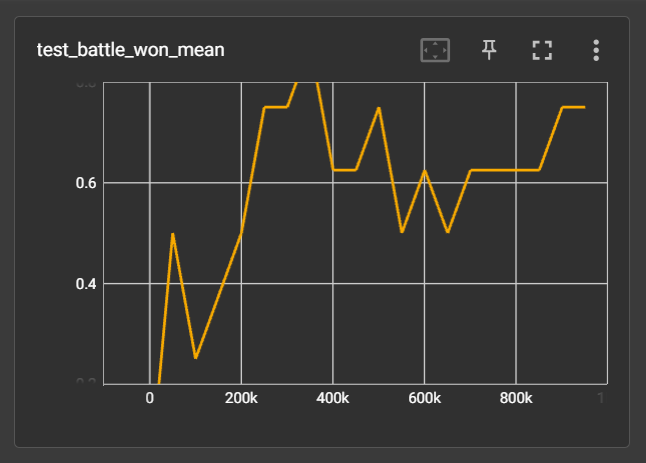


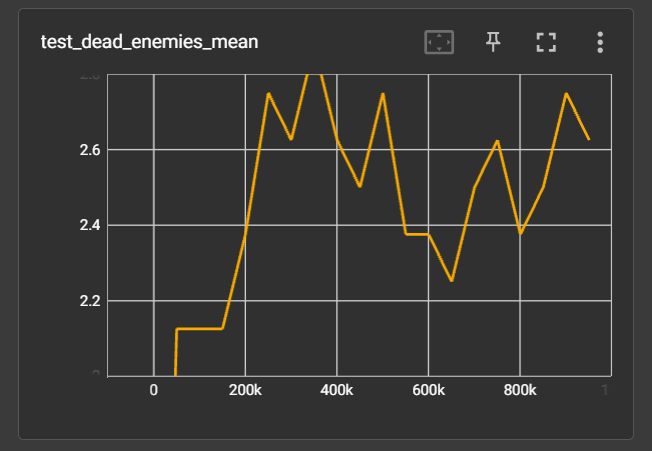


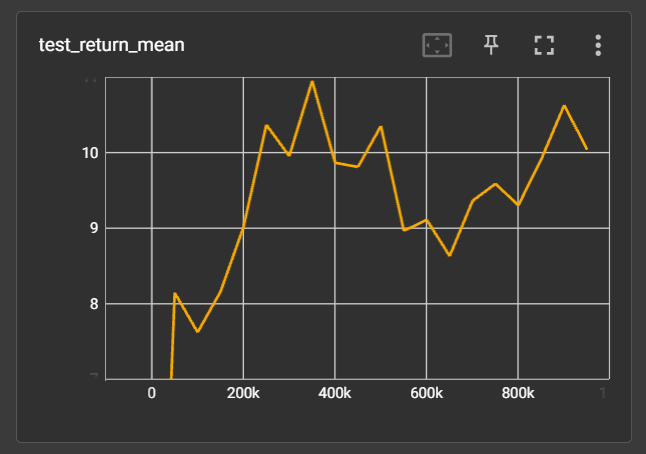
注：由于reward shaping对偏好的3个敌人击打奖励提升1.5倍，非偏好的敌人击打奖励降为0.5倍，故expert最终得到的reward为约15，高于无reward shaping时的约12。

**madiff - 2 experts**









训练得到的最终模型的self-play与expert 1, 2的self-play对比，以及与expert 1, 2的cross-play结果如下：

Self-play结果：  
​  
Madiff:  
​  
test\_dead\_enemies\_mean: 2.0000  
​  
test\_return\_mean: 7.4526  
​  
test\_ep\_length\_mean: 86.3750  
​  
Expert 1:  
​  
test\_dead\_enemies\_mean: 3.0000  
​  
test\_return\_mean: 11.72413793  
​  
test\_ep\_length\_mean: 82.5000  
​  
Expert 2:  
​  
test\_dead\_enemies\_mean: 3.0000  
​  
test\_return\_mean: 11.72413793  
​  
test\_ep\_length\_mean: 78.8750  
​  
   
​  
Cross-play结果：(ego x tm，ego控制1个agent，tm控制2个)  
​  
Madiff x Expert 1:  
​  
test\_dead\_enemies\_mean: 2.2500  
​  
test\_return\_mean: 9.3815  
​  
test\_ep\_length\_mean: 70.3750  
​  
Madiff x Expert 2:  
​  
test\_dead\_enemies\_mean: 2.3750  
​  
test\_return\_mean: 9.3182  
​  
test\_ep\_length\_mean: 94.2500

（所有测试保证8次重复取平均）

打印敌人轨迹发现cross-play时madiff agent确实会适应不同expert tm的进攻策略：expert 1偏好攻打0~2号敌人，madiff agent就会陪它打完这些敌人；对于expert 2则是相反。

例如，与expert 2 cross-play的一个episode打印敌人坐标轨迹如下（共6个敌人，每相邻两列为一个敌人x, y坐标，nan表示敌人已死亡），依次杀死了敌人4, 5, 3.

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 0.8234 0.4847 0.6775 0.7878

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 0.8234 0.4847 0.6775 0.7878

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 0.8234 0.4847 0.6775 0.7878

0.3204 0.7878 0.2095 0.4847 0.3203 0.1834 0.6644 0.1834 0.8234 0.4847 0.6775 0.7878

…

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan 0.6775 0.7878

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan 0.6775 0.7878

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan 0.6775 0.7878

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan 0.6775 0.7878

...

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan nan nan

0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan nan nan

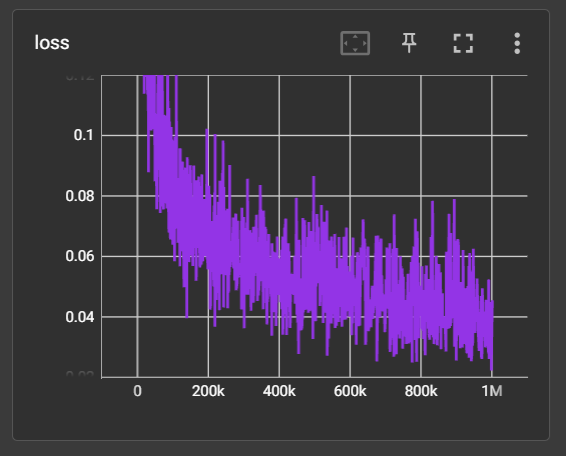
0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan nan nan

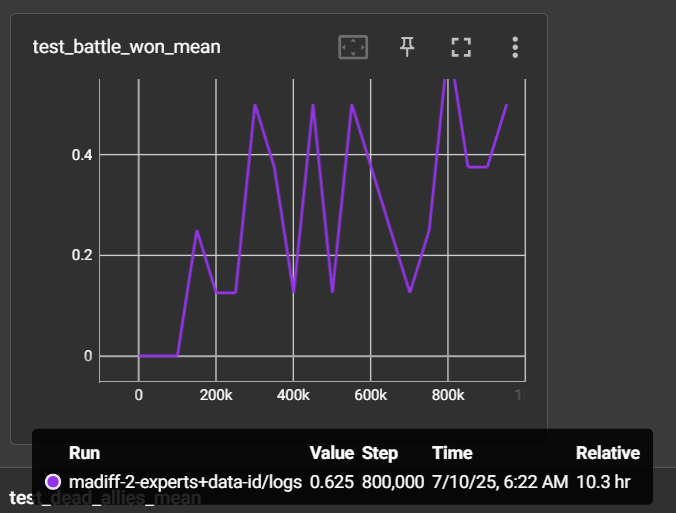
0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 0.6644 0.1834 nan nan nan nan

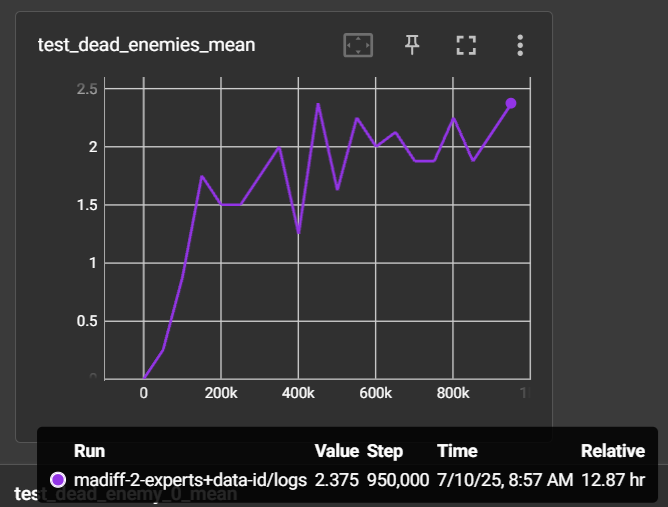
...

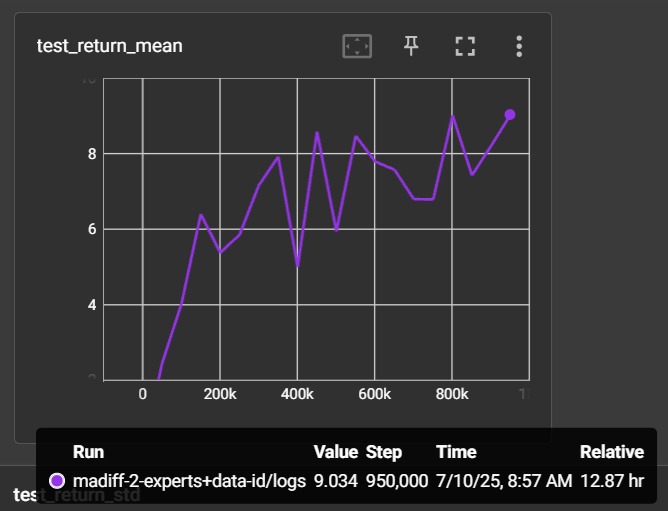
0.3204 0.7878 0.1787 0.4847 0.3203 0.1834 nan nan nan nan nan nan

**madiff - 2 experts - with data id**









训练得到的最终模型的self-play以及与expert 1, 2的cross-play结果如下：

Self-play结果：  
设置id=0：  
test\_dead\_enemies\_mean: 2.1250  
test\_return\_mean: 8.4267  
test\_ep\_length\_mean: 82.0000  
设置id=1：  
test\_dead\_enemies\_mean: 2.6250  
test\_return\_mean: 10.2284  
test\_ep\_length\_mean: 73.6250  
​  
Cross-play结果：(ego x tm，ego控制1个agent，tm控制2个)  
madiff x expert 1，设置id=0：  
test\_dead\_enemies\_mean: 3.0000  
test\_return\_mean: 11.7241  
test\_ep\_length\_mean: 81.1250  
madiff x expert 2，设置id=1：  
test\_dead\_enemies\_mean: 3.0000  
test\_return\_mean: 11.7241  
test\_ep\_length\_mean: 78.3750

相比无data id有提升，cross-play几乎完全恢复了专家策略的性能。