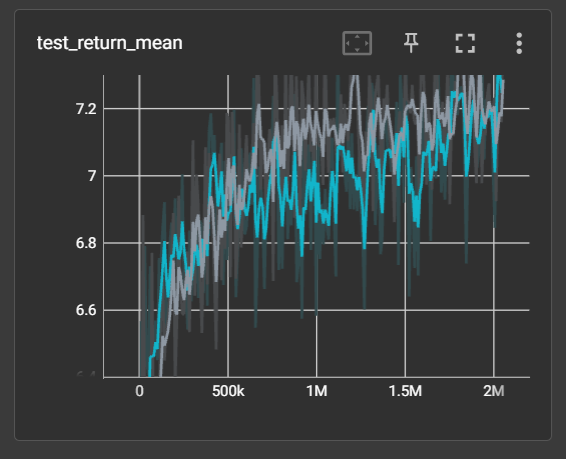
SMAC1选用地图：2m\_vs\_4m\_split

游戏规则设为击杀3个敌人就算赢。因为实际上做不到，故test\_battle\_won\_mean始终是0，用其他指标考量性能.

**Expert 1 & Expert 2:**





注：test\_return\_mean为采用reward shaping引导多样行为后的值，无reward shaping的实测结果：

Expert 1:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 6.0238

test\_ep\_length\_mean: 25.2500

Expert 2:

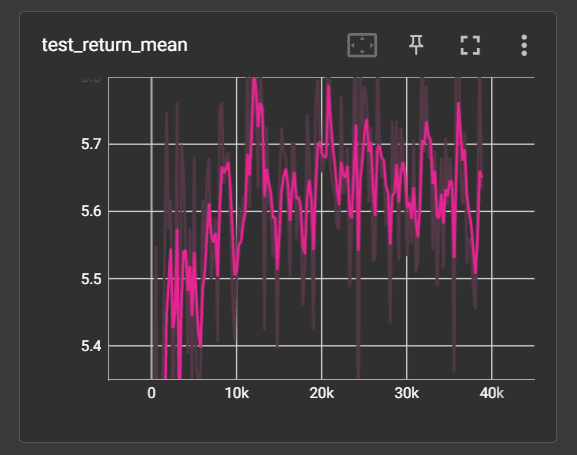
test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.8095

test\_ep\_length\_mean: 36.7500

**BC – 2 experts:**

****

****

最终测试结果：

Self-play：

test\_dead\_enemies\_mean: 1.8750

test\_return\_mean: 5.5536

test\_ep\_length\_mean: 32.7500

Cross-play:

BC x Expert 1:

test\_dead\_enemies\_mean: 1.5000

test\_return\_mean: 5.0893

test\_ep\_length\_mean: 22.6875

BC x Expert 2:

test\_dead\_enemies\_mean: 1.8750

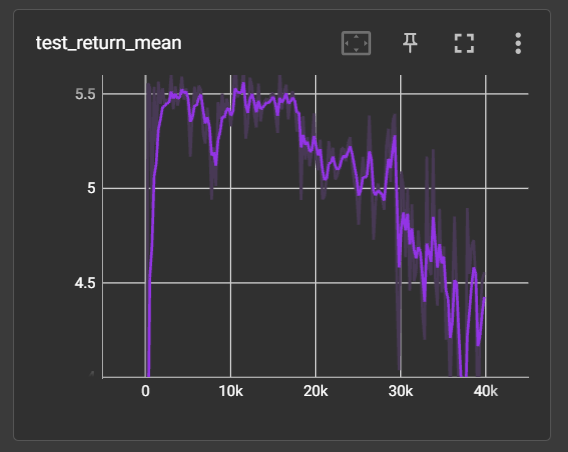
test\_return\_mean: 5.4107

test\_ep\_length\_mean: 35.6875

cross-play存在比较明显的性能退步，不能和两个expert都配合好。

**BC – 2 experts – with data id:**





达到的最高性能接近expert，但训练不稳定。

最终测试结果：

Self-play:

id=0:

test\_dead\_enemies\_mean: 1.9375

test\_return\_mean: 5.5923

test\_ep\_length\_mean: 25.0625

id=1:

test\_dead\_enemies\_mean: 1.9375

test\_return\_mean: 5.4673

test\_ep\_length\_mean: 80.7500

Cross-play:

BC x Expert 1, id=0

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.7024

test\_ep\_length\_mean: 39.3125

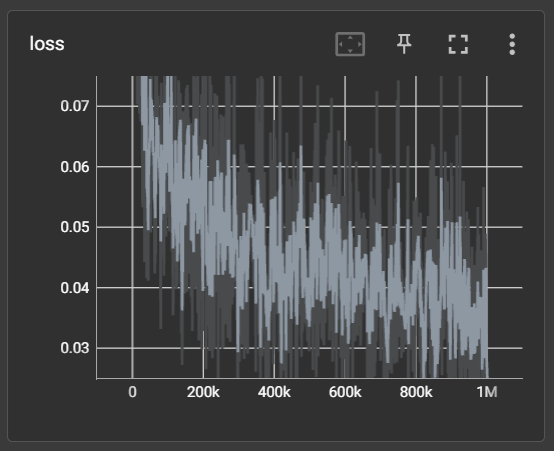
BC x Expert 2, id=1

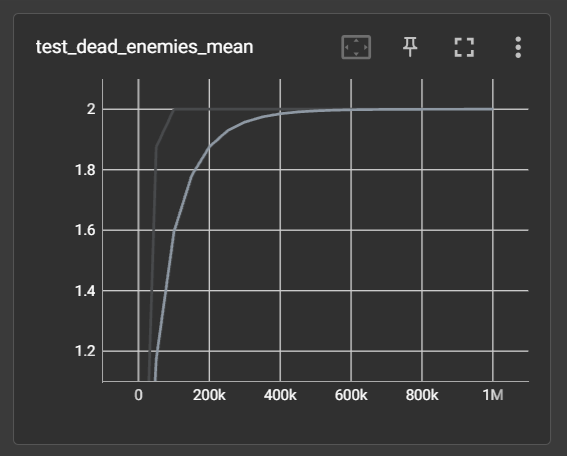
test\_dead\_enemies\_mean: 1.9375

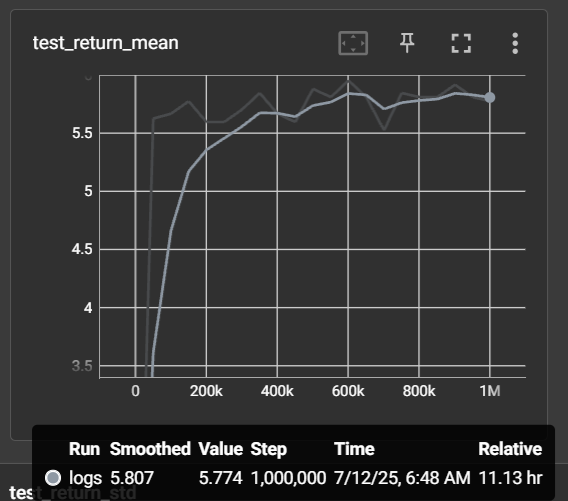
test\_return\_mean: 5.5744

test\_ep\_length\_mean: 49.8750

**MADiff – 2 experts （第一次实验）**:







最终测试结果：

Self-play:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.7381

test\_ep\_length\_mean: 40.2500

cross-play:

madiff x expert 1:

test\_dead\_enemies\_mean: 1.7500

test\_return\_mean: 5.3333

test\_ep\_length\_mean: 26.1250

madiff x expert 2:

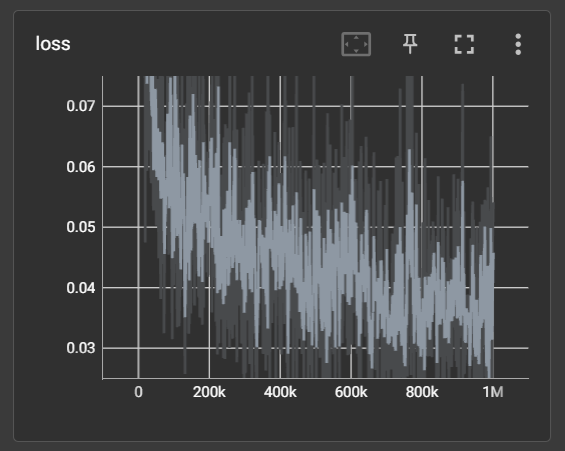
test\_dead\_enemies\_mean: 1.3750

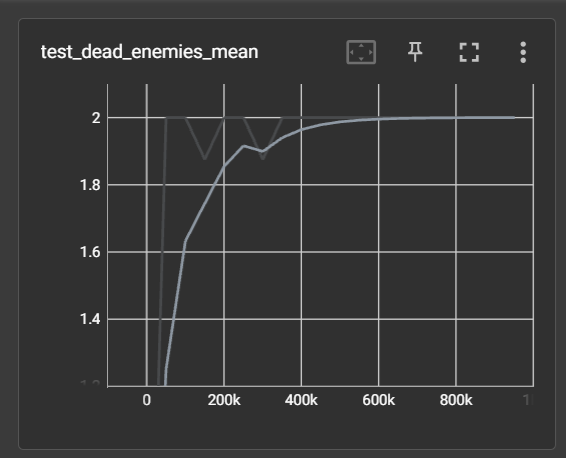
test\_return\_mean: 4.5298

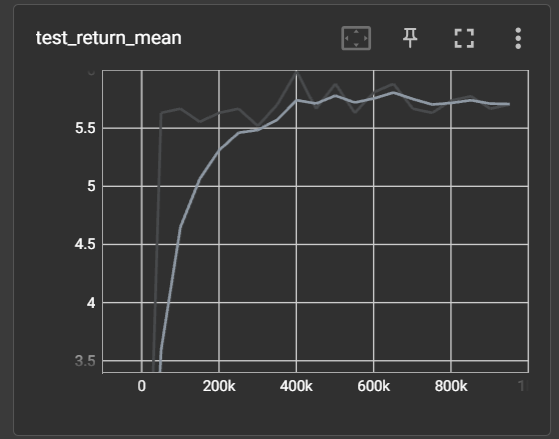
test\_ep\_length\_mean: 38.6250

实际cross为何反而远不如BC??

**MADiff – 2 experts (第二次实验)：**

****

****

****

Self-play:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.8810

test\_ep\_length\_mean: 25.0000

Cross-play:

madiff x expert 1:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.8810

test\_ep\_length\_mean: 26.6250

madiff x expert 2:

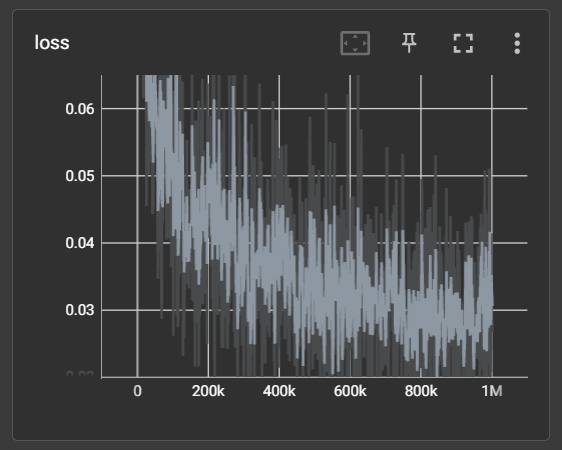
test\_dead\_enemies\_mean: 1.0000

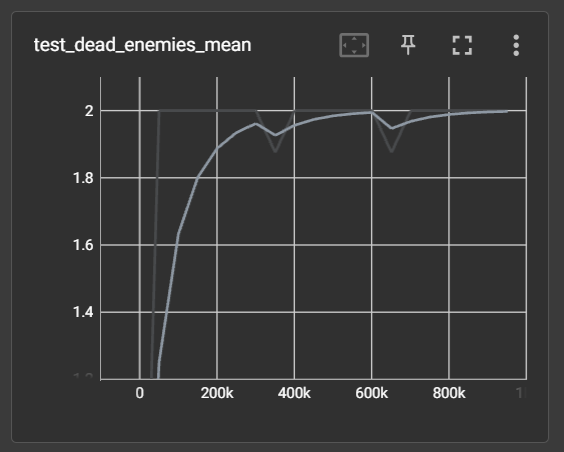
test\_return\_mean: 4.0833

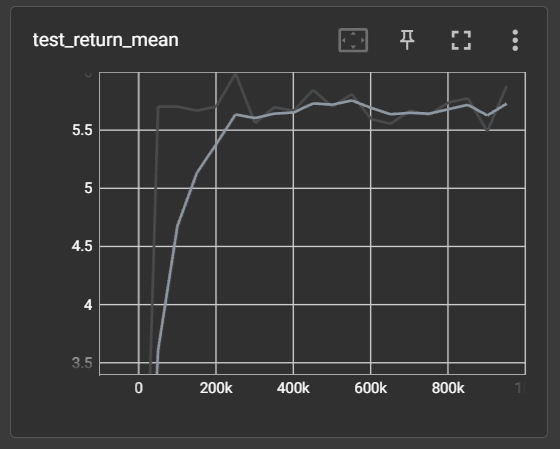
test\_ep\_length\_mean: 19.1250

注：两次实验区别仅在madiff的return\_scale超参数（分别为6, 9），用于指定对return进行归一化时预期的最大return，应设为与行为策略的return相近，在测试阶段也是以此值作为madiff模型的预期return输入。实验发现madiff与多expert配合时有较严重性能下降。

**Madiff – 2 experts – with data id:**







Self-play:

Data id=0:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.9167

test\_ep\_length\_mean: 32.2500

data id=1:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.5952

test\_ep\_length\_mean: 55.2500

cross-play:

madiff x expert 1, data id=0:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.8095

test\_ep\_length\_mean: 31.3750

madiff x expert 2, data id=1:

test\_dead\_enemies\_mean: 2.0000

test\_return\_mean: 5.5595

test\_ep\_length\_mean: 61.5000