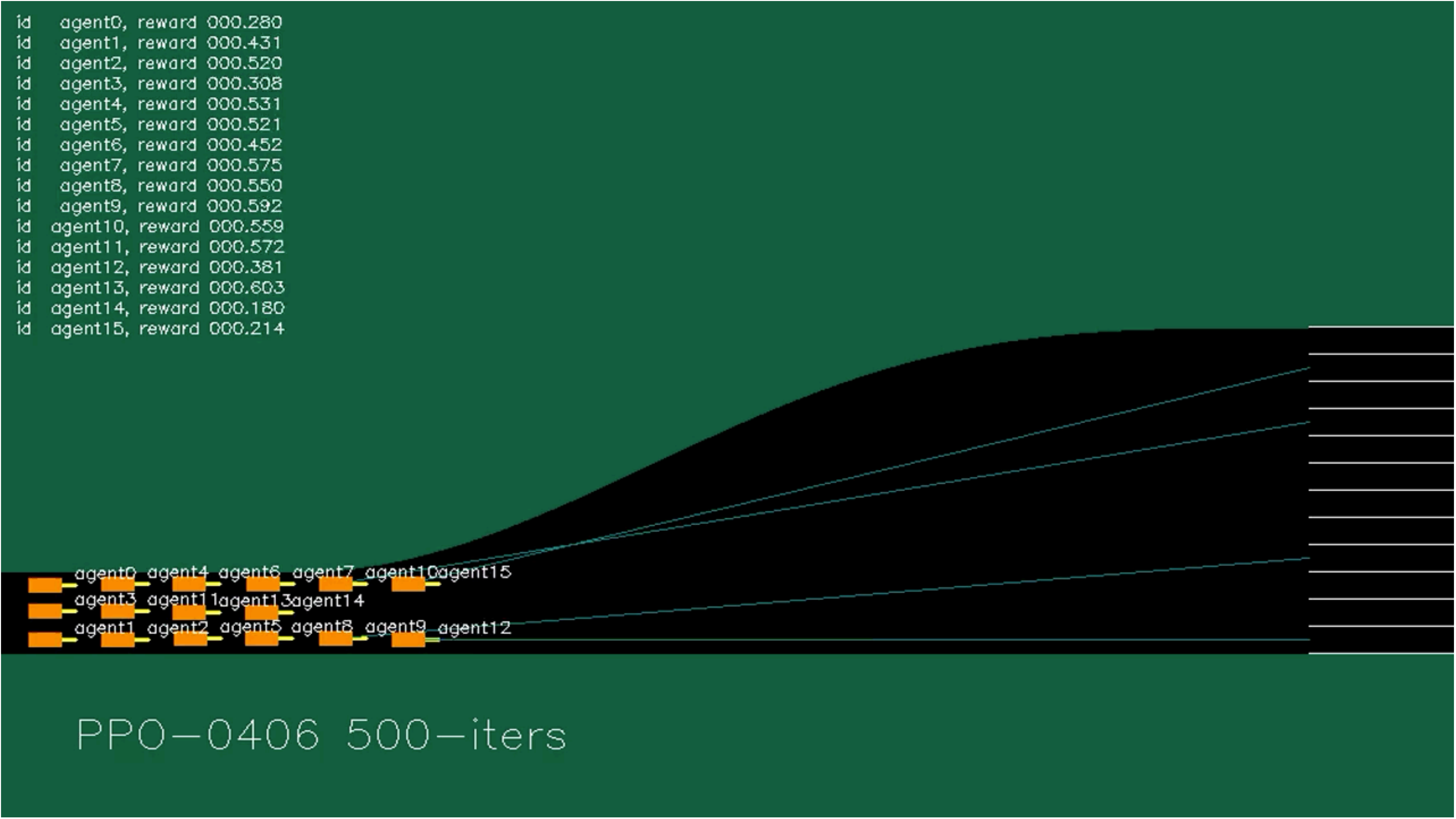
4

Performance of small batch training

Metric	**Value**
iteration	500
— performance —	
agent collision	0.3075
boundary col. rate	0.0275
gate collision rate	0.1917
pass rate	0.4669
success rate	0.0756
episode reward	826.8
single reward	51.68
episode length	227.6
— configuration —	
num gpus	1
num workers	10
num env per worker	2
train batch size	4000

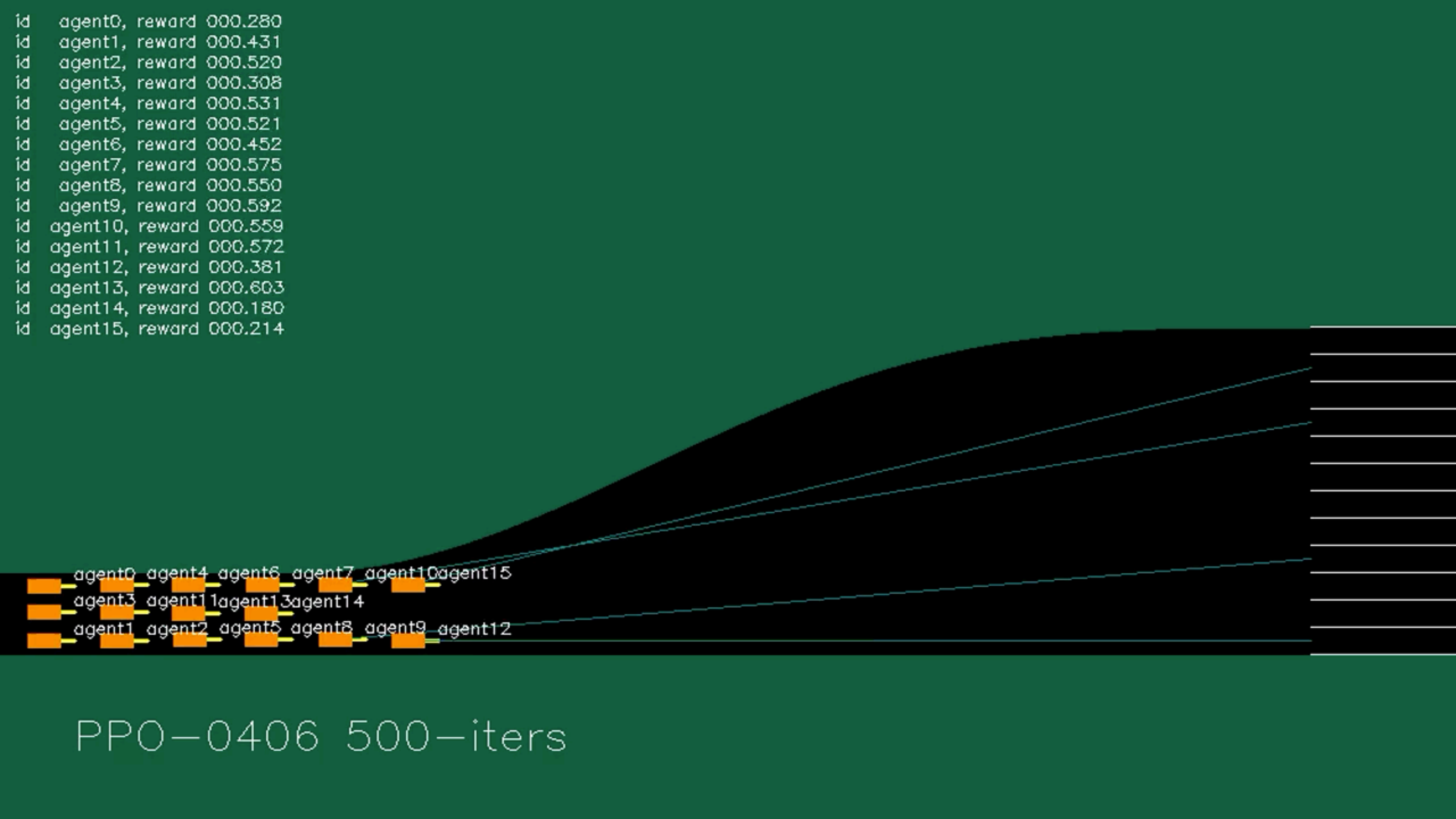


```
id   agent0, reward 000.280
id   agent1, reward 000.431
id   agent2, reward 000.520
id   agent3, reward 000.308
id   agent4, reward 000.531
id   agent5, reward 000.521
id   agent6, reward 000.452
id   agent7, reward 000.575
id   agent8, reward 000.550
id   agent9, reward 000.592
id  agent10, reward 000.559
id  agent11, reward 000.572
id  agent12, reward 000.381
id  agent13, reward 000.603
id  agent14, reward 000.180
id  agent15, reward 000.214
```

agent0 agent4 agent6 agent7 agent10agent15
agent3 agent11agent13agent14
agent1 agent2 agent5 agent8 agent9 agent12

PPO-0406 500-iters

Waiting and observing behaviors emerge! But still easy to collide.



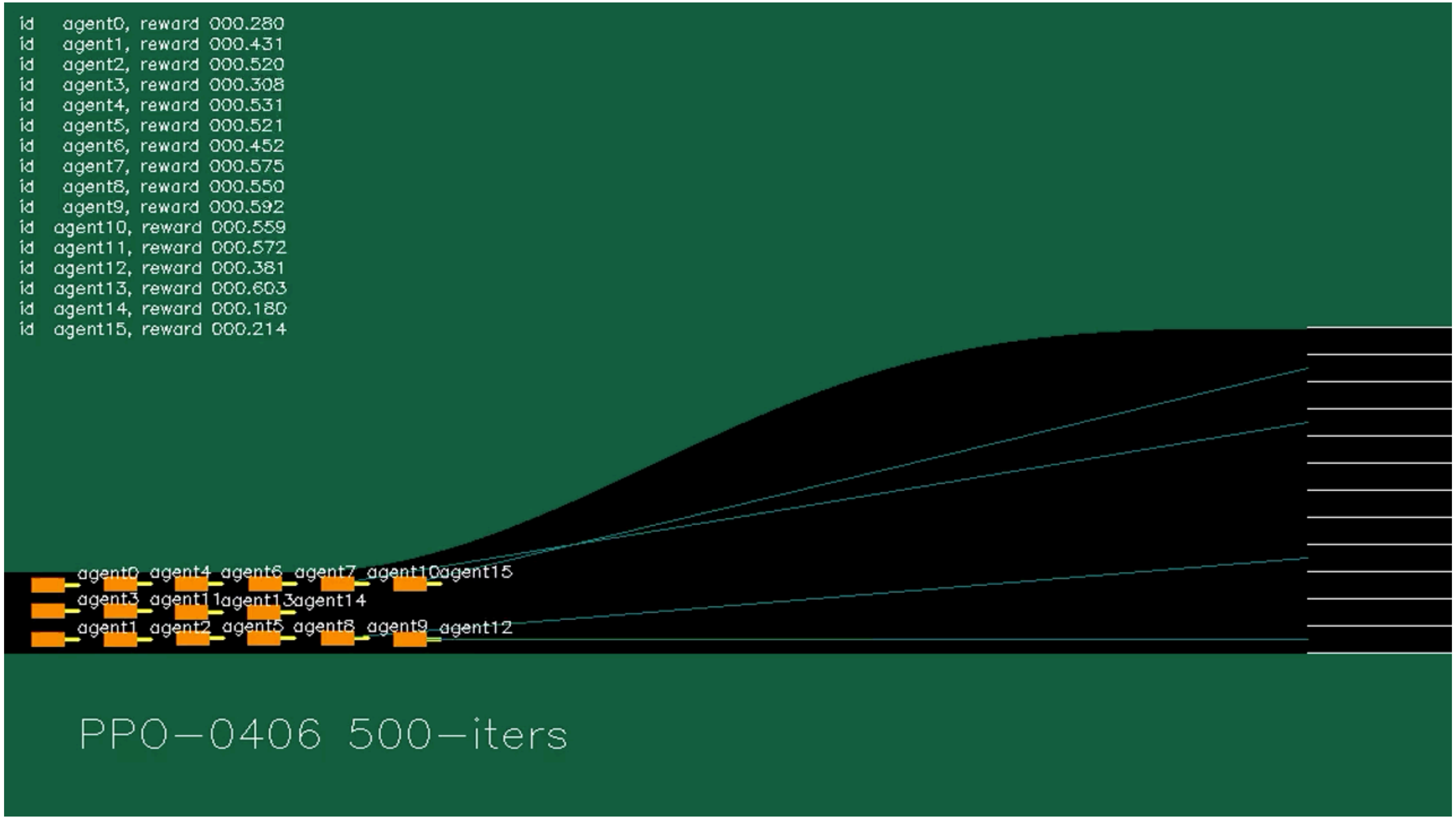
```
id   agent0, reward 000.280
id   agent1, reward 000.431
id   agent2, reward 000.520
id   agent3, reward 000.308
id   agent4, reward 000.531
id   agent5, reward 000.521
id   agent6, reward 000.452
id   agent7, reward 000.575
id   agent8, reward 000.550
id   agent9, reward 000.592
id  agent10, reward 000.559
id  agent11, reward 000.572
id  agent12, reward 000.381
id  agent13, reward 000.603
id  agent14, reward 000.180
id  agent15, reward 000.214
```

agent0 agent4 agent6 agent7 agent10agent15
agent3 agent11agent13agent14
agent1 agent2 agent5 agent8 agent9 agent12

PPO-0406 500-iters

Performance of small batch training

Waiting and observing behaviors emerge! But still easy to collide.



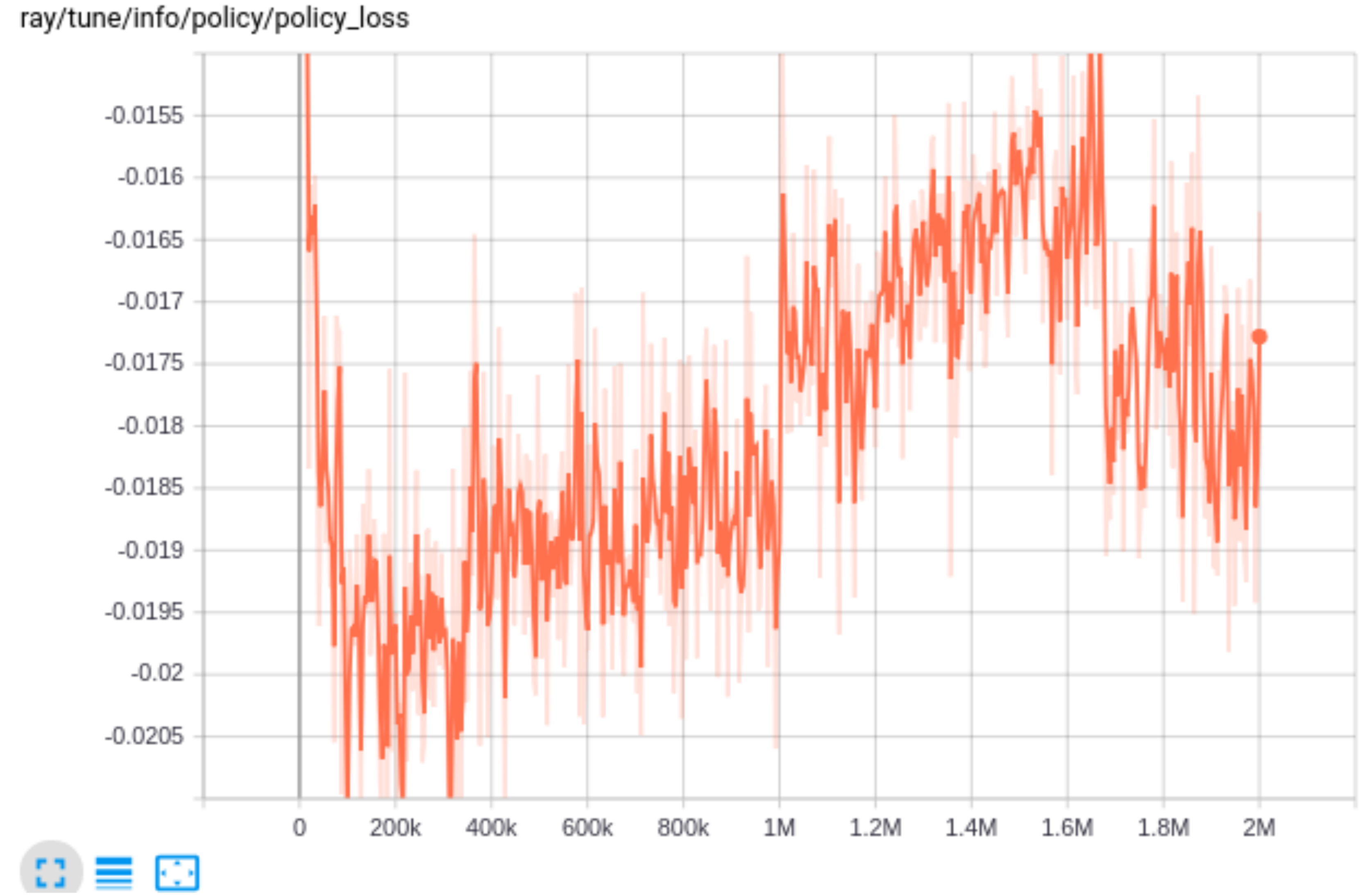
```
id   agent0, reward 000.280
id   agent1, reward 000.431
id   agent2, reward 000.520
id   agent3, reward 000.308
id   agent4, reward 000.531
id   agent5, reward 000.521
id   agent6, reward 000.452
id   agent7, reward 000.575
id   agent8, reward 000.550
id   agent9, reward 000.592
id  agent10, reward 000.559
id  agent11, reward 000.572
id  agent12, reward 000.381
id  agent13, reward 000.603
id  agent14, reward 000.180
id  agent15, reward 000.214
```

agent0 agent4 agent6 agent7 agent10agent15
agent3 agent11agent13agent14
agent1 agent2 agent5 agent8 agent9 agent12
PPO−0406 500−iters

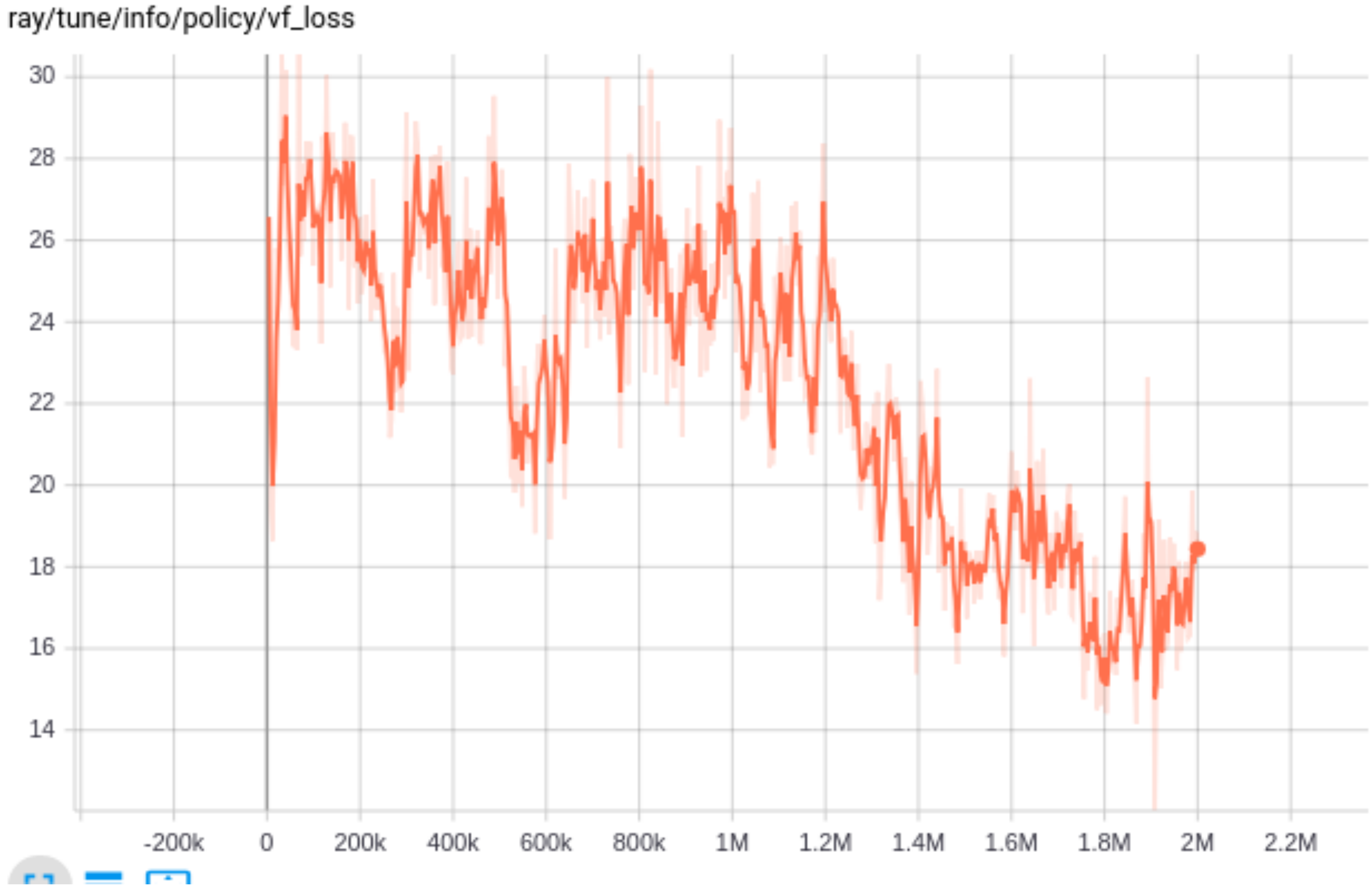
Metric	Value
iteration	500
— performance —	
agent collision	0.3075
boundary col. rate	0.0275
gate collision rate	0.1917
pass rate	0.4669
success rate	0.0756
episode reward	826.8
single reward	51.68
episode length	227.6
— configuration —	
num gpus	1
num workers	10
num env per worker	2
train batch size	4000

Discussion of small batch training

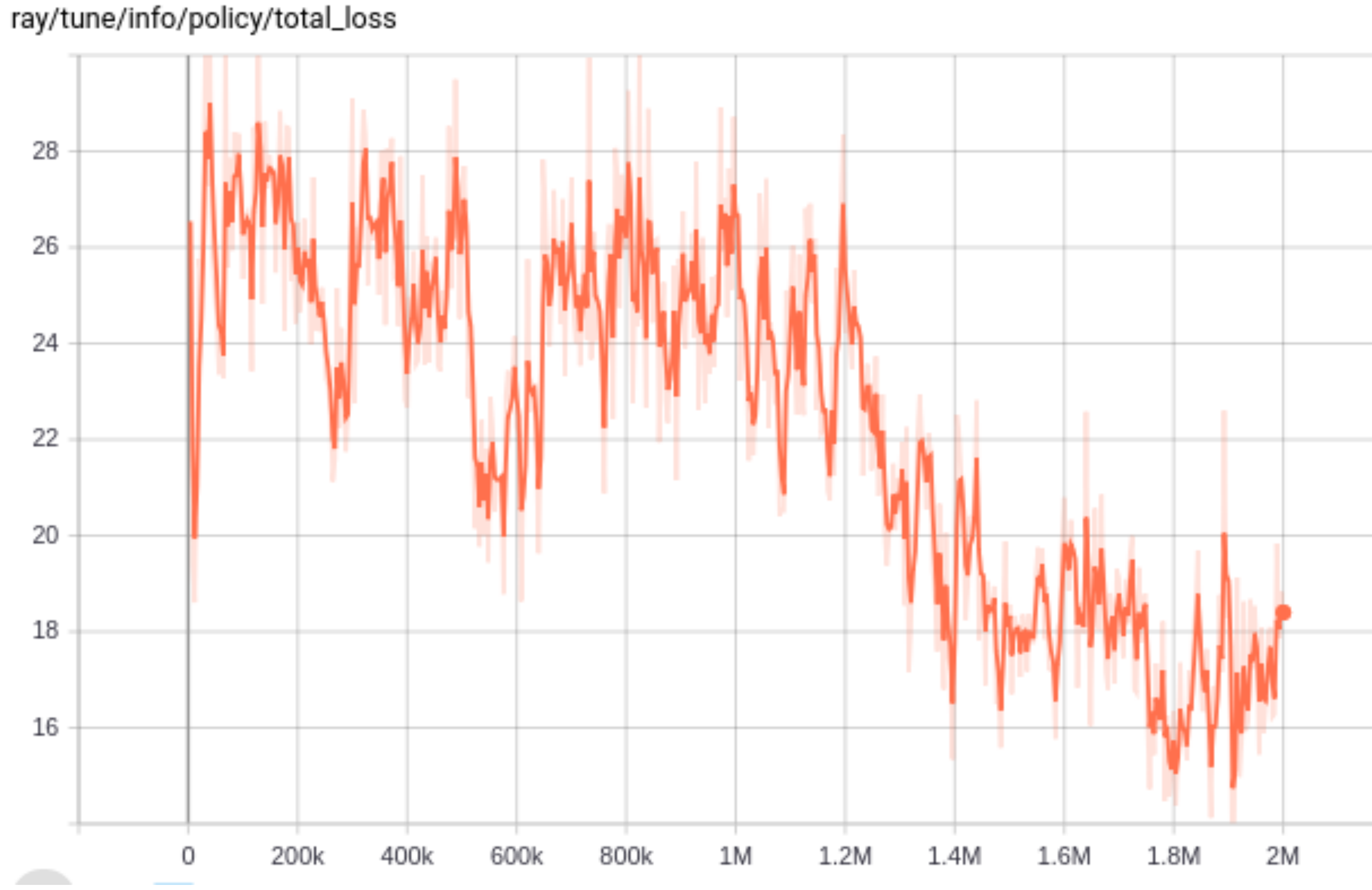
Policy Loss



Value Loss



Total Loss



- Though policy loss increasing, value loss predominates the learning, due to greater magnitude.