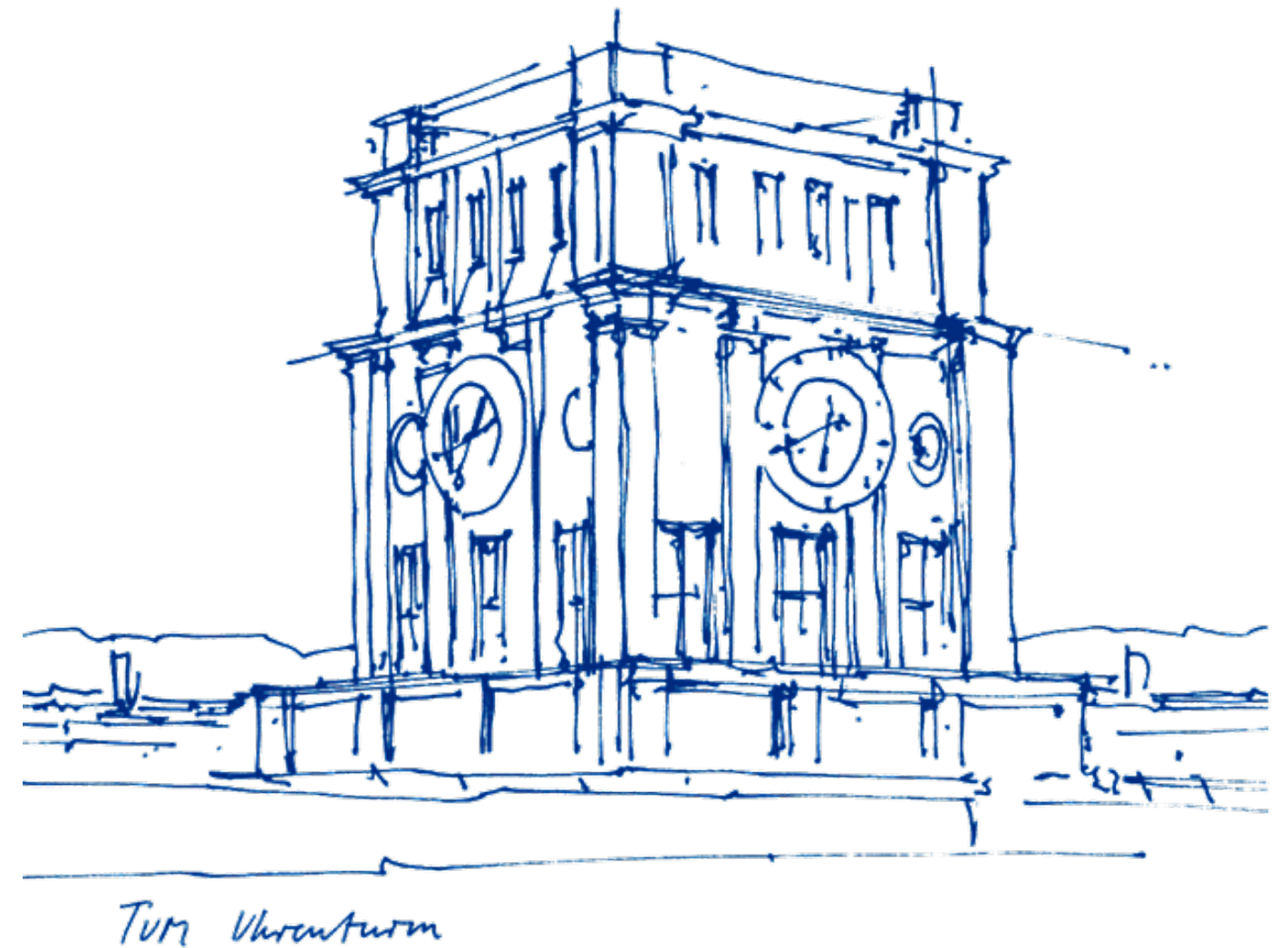




# Factory Manipulation with Cooperative Multi-agent Reinforcement Learning

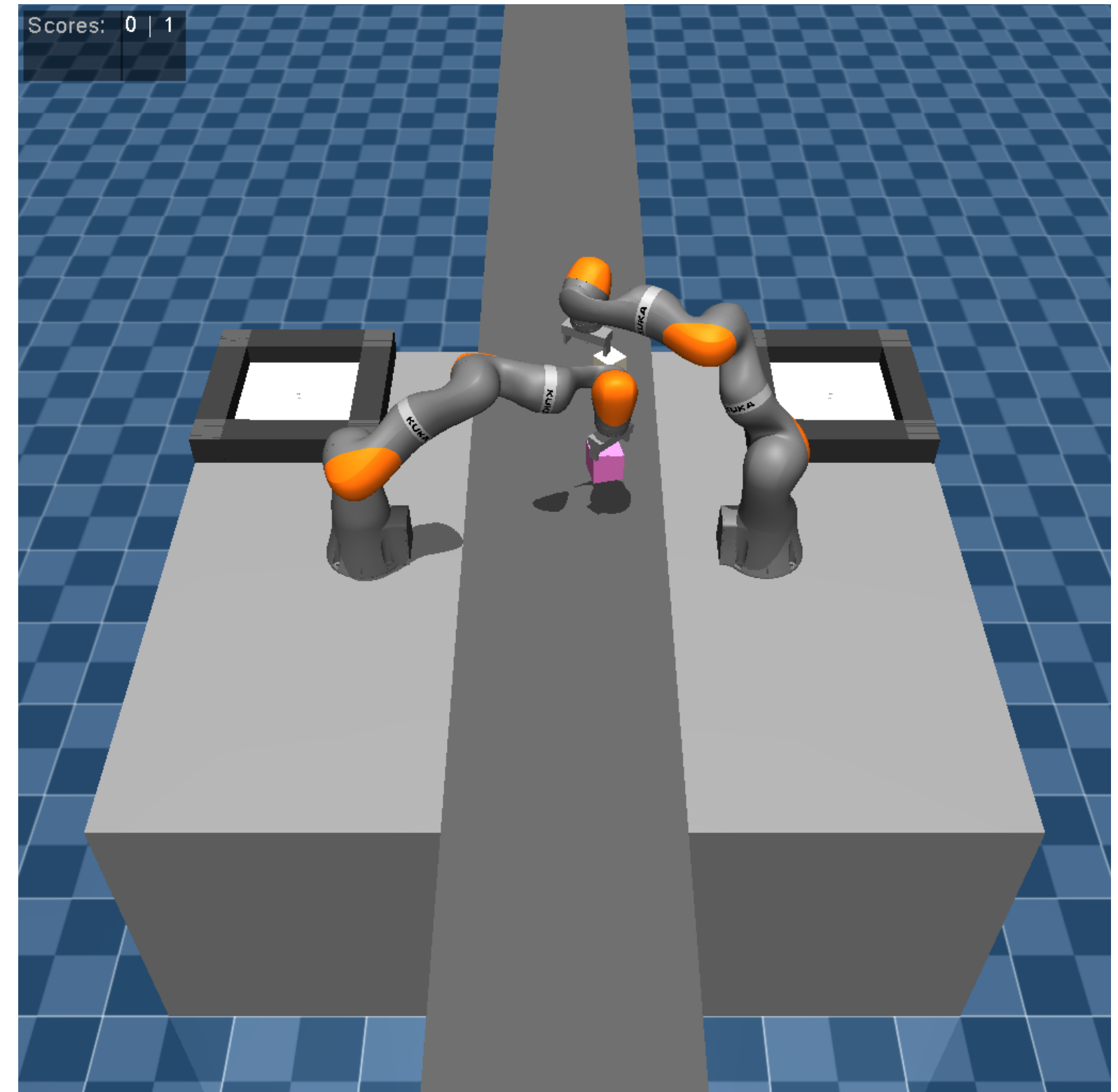
6 June 2024

Nikolas Kirschstein & Kassian Köck  
(Team 7)



# Problem Setting

- Gym env based on MuJoCo (Howell et al. 2022)
  - 2 or more robot arms (8 DOF each)
  - basket in reach for each arm
  - conveyor belt with increasing speed
  - cubes transported on conveyor belt
  - score = number cubes in baskets
- episode ends if either:
  - arm hits the env (incl. other arms)
  - cube is missed by all arms
- conventional pre-programming-based approaches too inflexible and tedious  
→ use of MARL (cp. Pérez-Francisco et al. 1998, Bozma and Kalalioğlu 2012, Yu et al. 2017, Han et al. 2020)



# Our Goal



Multiple robot arms cooperating to maximise efficiency in factory manipulation task  
(PnP along conveyor belt as representative and important special case)

# Our Goal



Multiple robot arms cooperating to maximise efficiency in factory manipulation task  
(PnP along conveyor belt as representative and important special case)

**Final boss:** Many robot arms, on both sides of the belt and also from the ceiling  
*s.t. communication cost too high to broadcast joint states in real time*  
 $\implies$  true **multi-agent reinforcement learning** (due to partial observability)

# Our Goal



Multiple robot arms cooperating to maximise efficiency in factory manipulation task  
(PnP along conveyor belt as representative and important special case)

**Final boss:** Many robot arms, on both sides of the belt and also from the ceiling  
*s.t. communication cost too high to broadcast joint states in real time*  
 $\implies$  true **multi-agent reinforcement learning** (due to partial observability)

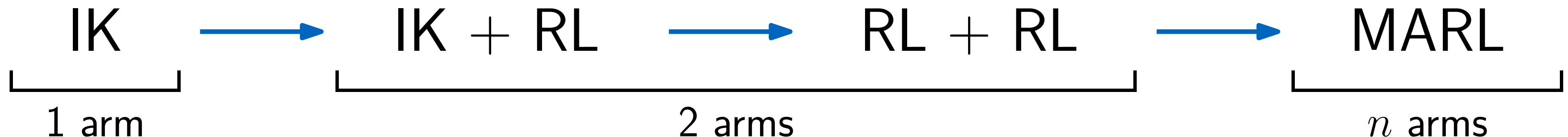
- simplification: only learn *deviation* from IK base policy
- roadmap: incremental approach

# Our Goal

Multiple robot arms cooperating to maximise efficiency in factory manipulation task  
(PnP along conveyor belt as representative and important special case)

**Final boss:** Many robot arms, on both sides of the belt and also from the ceiling  
*s.t. communication cost too high to broadcast joint states in real time*  
 $\Rightarrow$  true **multi-agent reinforcement learning** (due to partial observability)

- simplification: only learn *deviation* from IK base policy
- roadmap: incremental approach

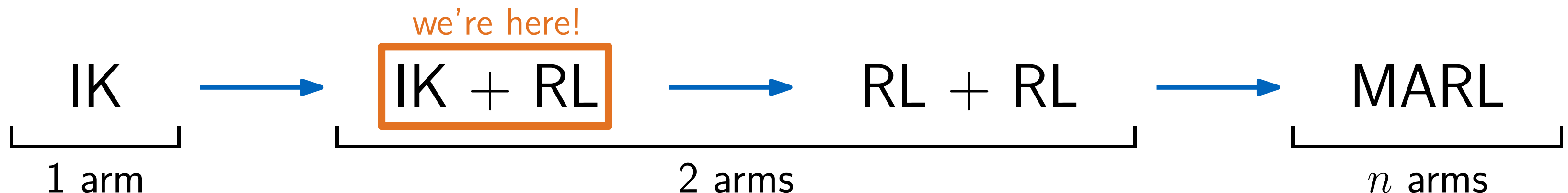


# Our Goal

Multiple robot arms cooperating to maximise efficiency in factory manipulation task  
(PnP along conveyor belt as representative and important special case)

**Final boss:** Many robot arms, on both sides of the belt and also from the ceiling  
*s.t. communication cost too high to broadcast joint states in real time*  
 $\Rightarrow$  true **multi-agent reinforcement learning** (due to partial observability)

- simplification: only learn *deviation* from IK base policy
- roadmap: incremental approach

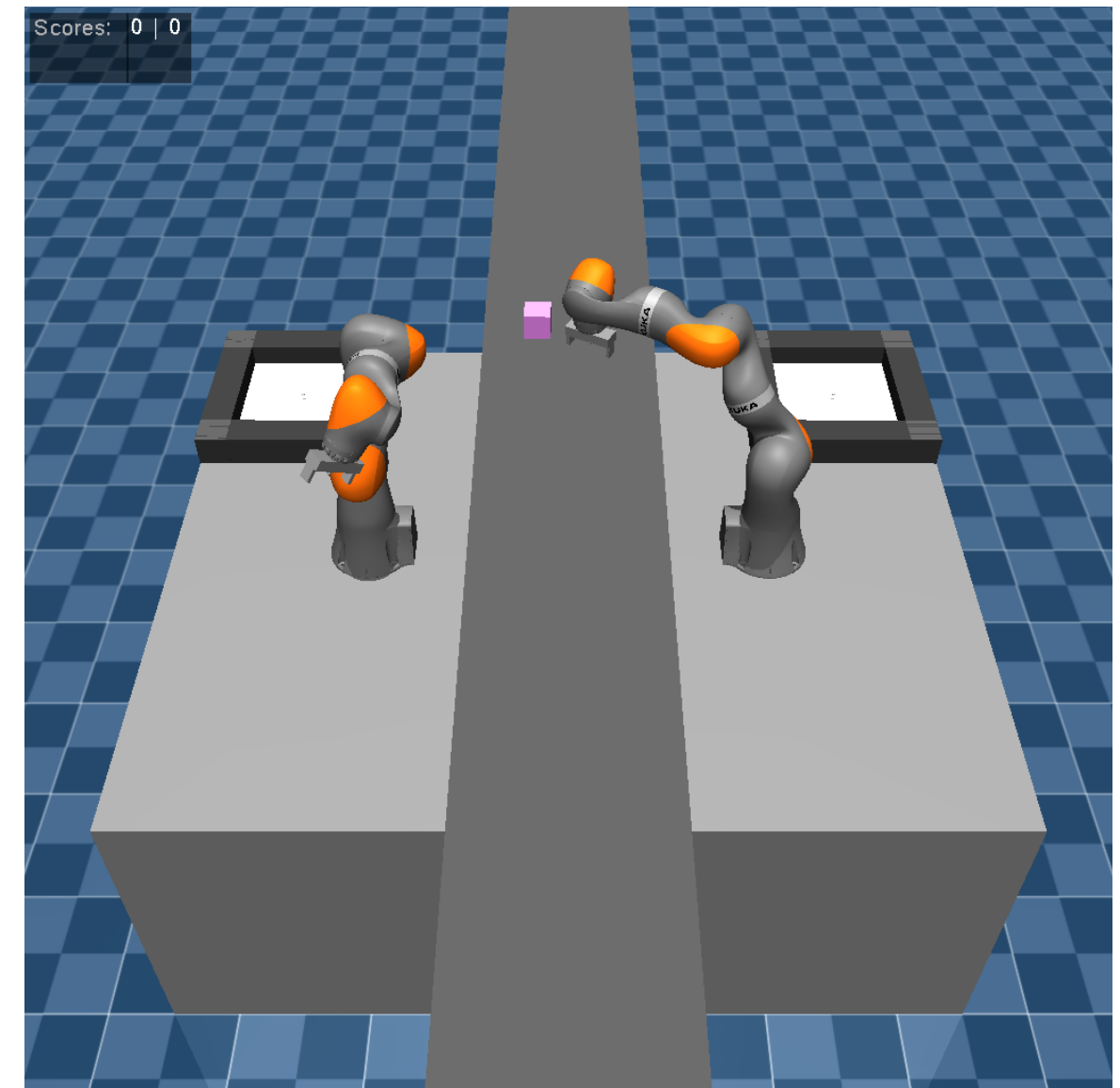


# Reward Design

Which reward should we use?

1st intention: reward of **1** if **block is thrown in the basket**, else 0

- highly sparse reward  
→ **learning very hard**
- nearly random behaviour overpowers base policy





Which reward should we use?

2nd intention: reward increases **monotonically** with **progress to target**

Desirable Incentives:

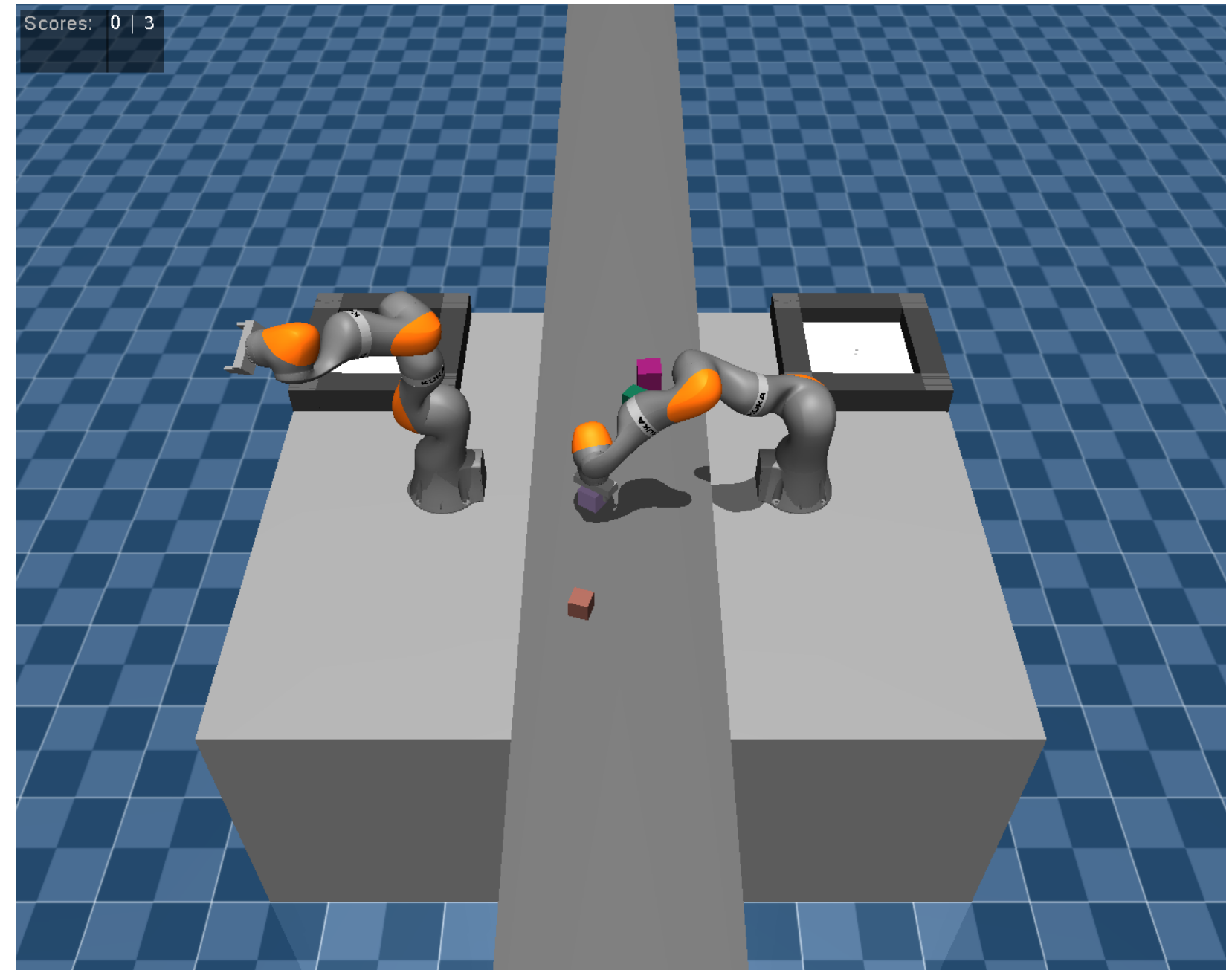
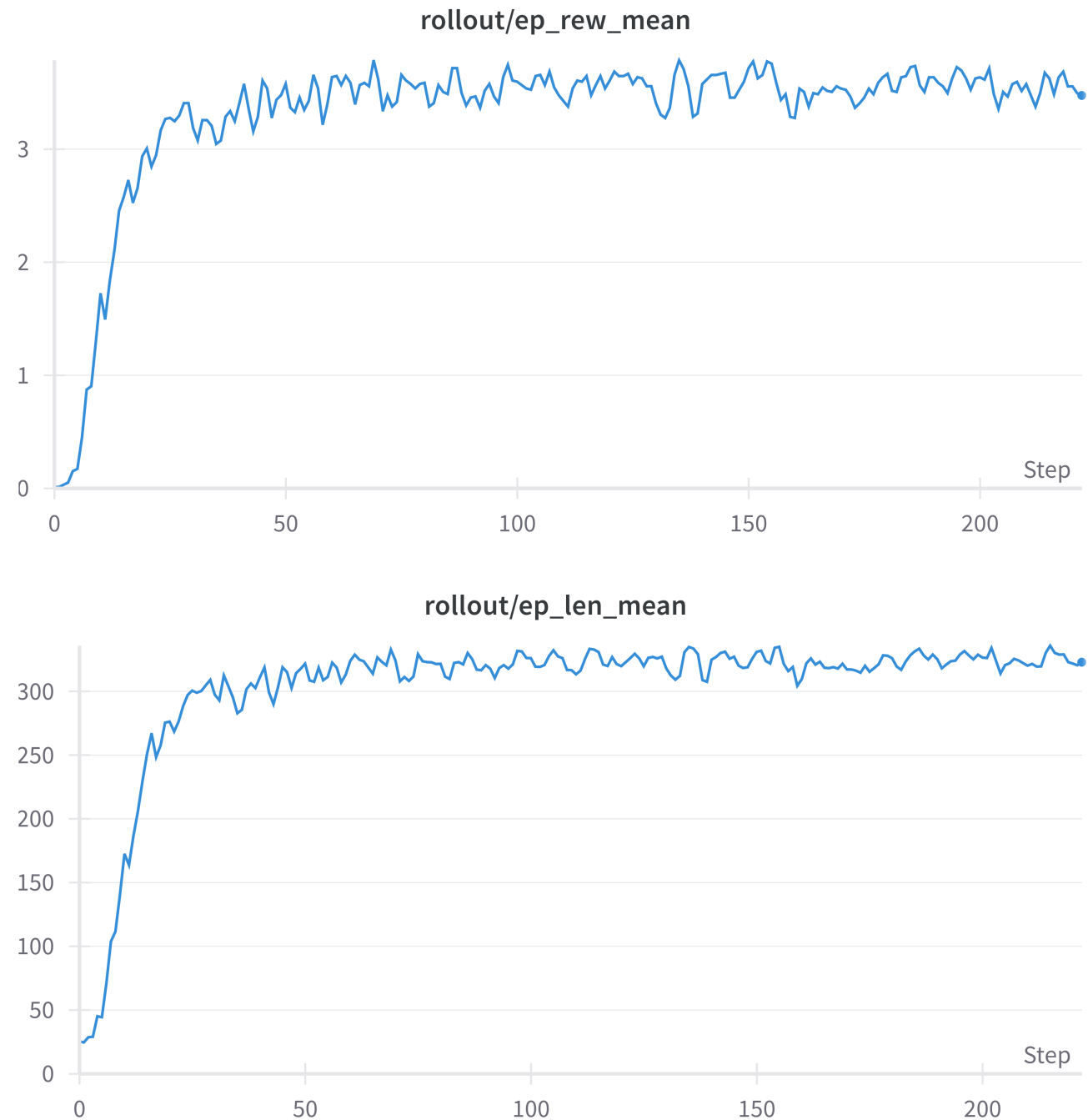
- $I_0$ : Reward cubes put into basket
- $I_1$ : Punish large deviation from base policy
- $I_2$ : Reward vicinity to closest cube
- $I_3$ : Punish distance to other robot arms
- $I_4$ : Reward grasping while very close to cube
- $I_5$ : Reward vicinity to basket with grasped cube
- $I_6$ : Reward relaxing grasp over basket

$$r = \sum_{i=0}^6 \omega_i I_i$$

- goal:  $\omega_0 \gg \omega_i$  for  $i \geq 1$
- ideally most  $\omega_i = 0$

# First Results

PPO for  $\omega_0, \omega_1 > 0$  and  $\omega_2, \dots, \omega_6 = 0$



learnt RL policy steers IK base policy away from other arm (undesired)

# Next Steps



- explore denser reward augmentations
- employ RL for 2nd arm as well
- construct multi-agent env



Discussion time!

