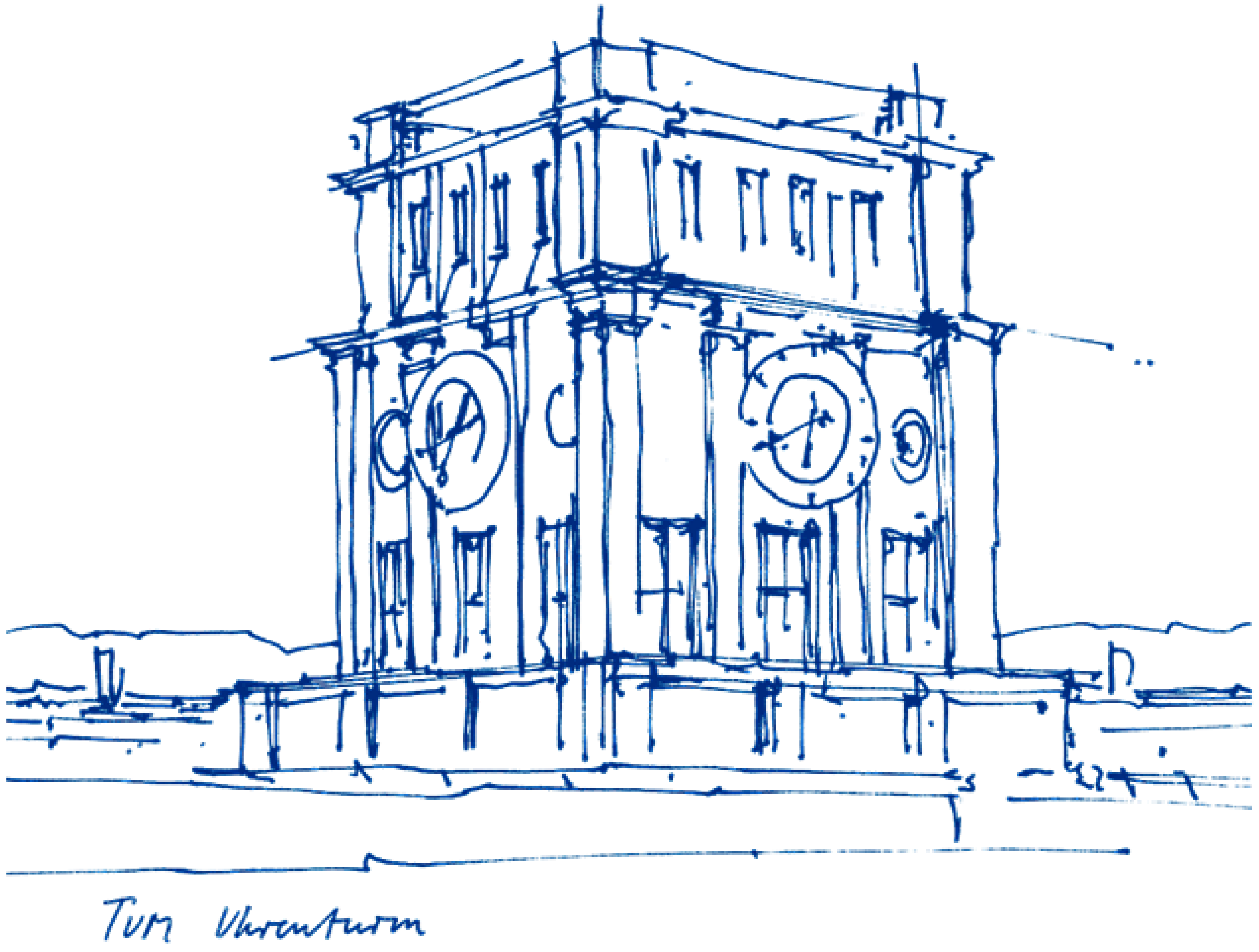
Factory Manipulation with Cooperative Multi-agent Reinforcement Learning

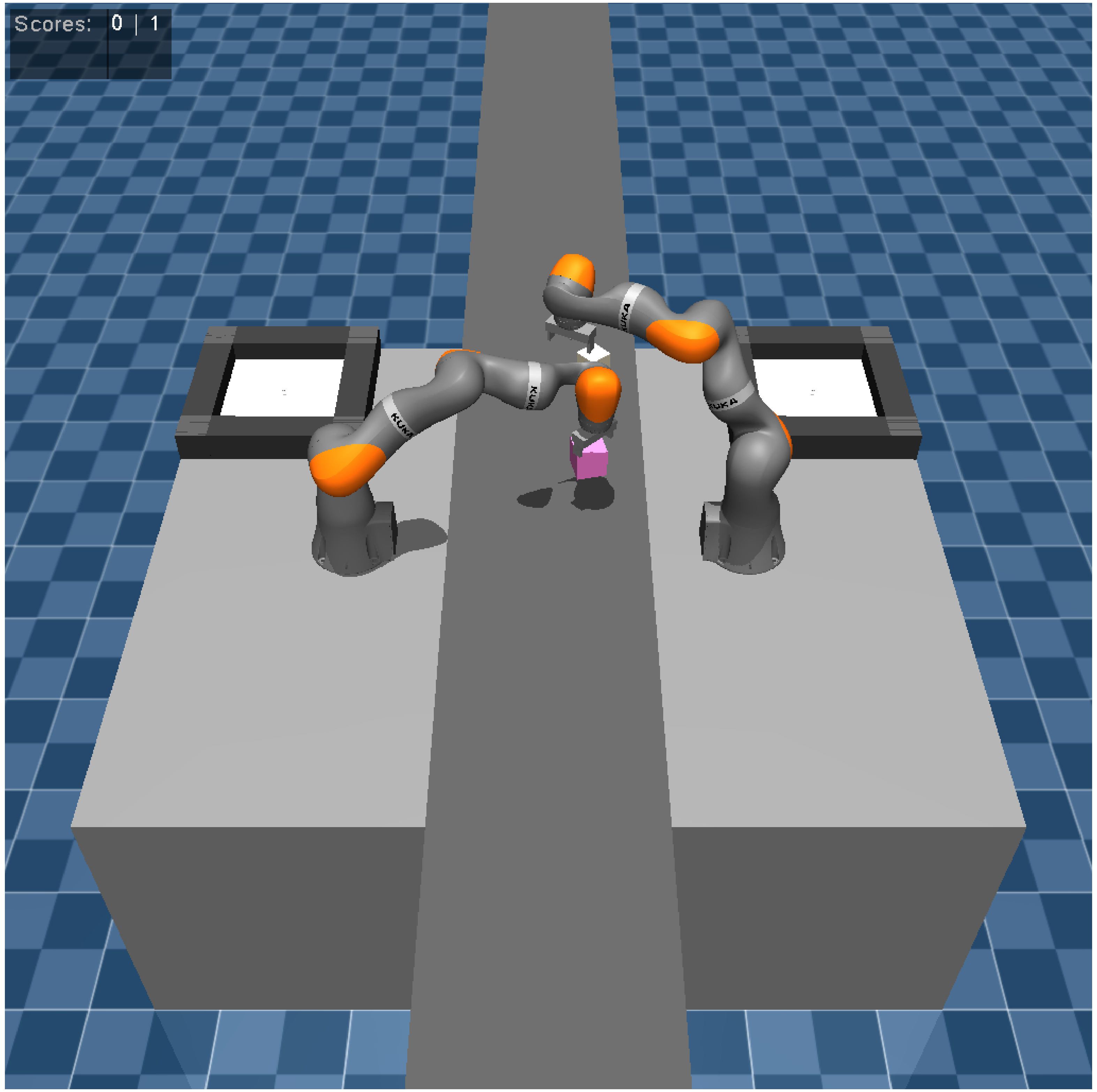
6 June 2024

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 Kalalıoğlu 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*
$\Longrightarrow$ 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*
$\Longrightarrow$ 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*
$\Longrightarrow$ true **multi-agent reinforcement learning** (due to partial observability)

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

$$\underbrace{\text{IK}}_{\text{1 arm}} \longrightarrow \underbrace{\text{IK} + \text{RL} \longrightarrow \text{RL} + \text{RL}}_{\text{2 arms}} \longrightarrow \underbrace{\text{MARL}}_{n\ \text{arms}}$$

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*
$\Longrightarrow$ true **multi-agent reinforcement learning** (due to partial observability)

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

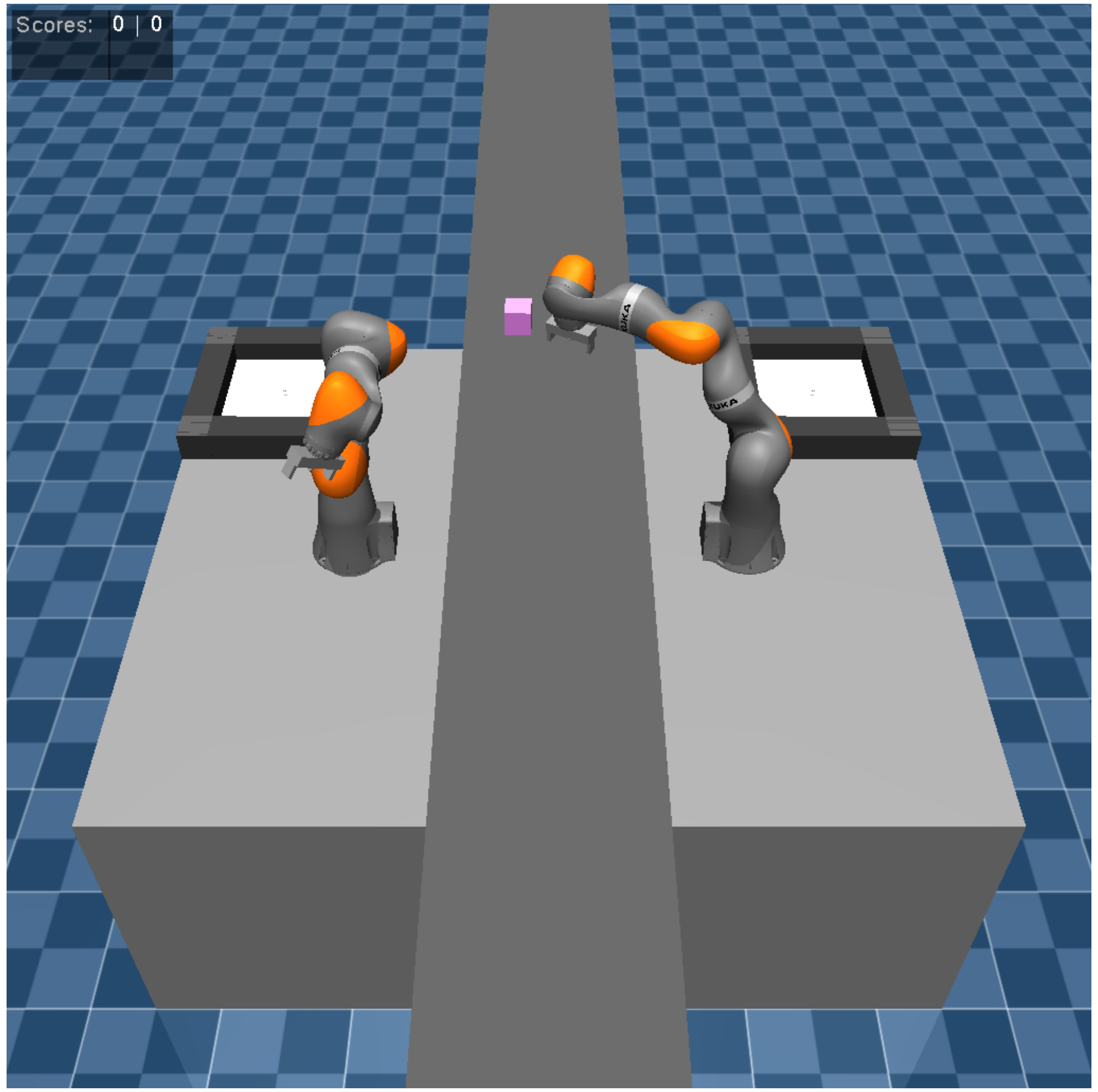
$$\underbrace{\text{IK}}_{\text{1 arm}} \longrightarrow \underbrace{\overset{\text{we're here!}}{\boxed{\text{IK} + \text{RL}}} \longrightarrow \text{RL} + \text{RL}}_{\text{2 arms}} \longrightarrow \underbrace{\text{MARL}}_{n\text{ arms}}$$

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



Reward Design

Which reward should we use?

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

Desirable Incentives:

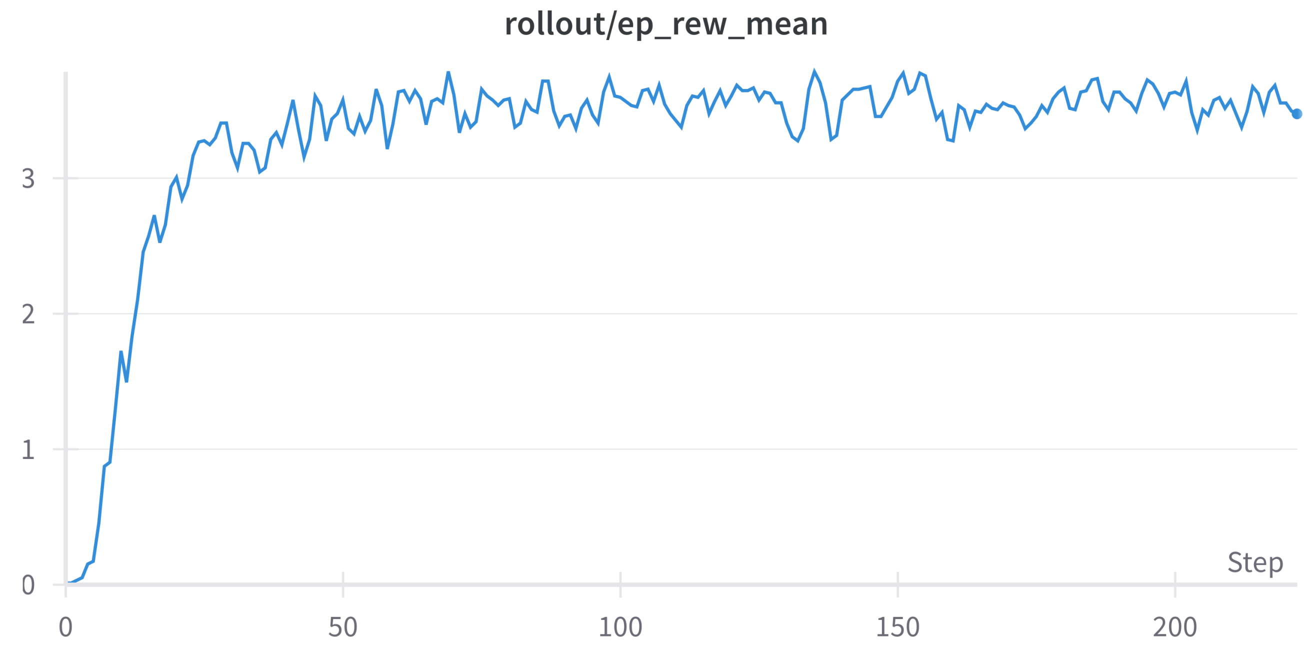
- I_0: Reward cubes put into basked
- I_1: Punish large deviation from base policy
- I_2: Reward vicinity to closest cube
- I_3: Punish distance to other robot arms
- I_4: Reward grasping while very close to cube
- I_5: Reward vicinity to basket with grasped cube
- I_6: Reward relaxing grasp over basket

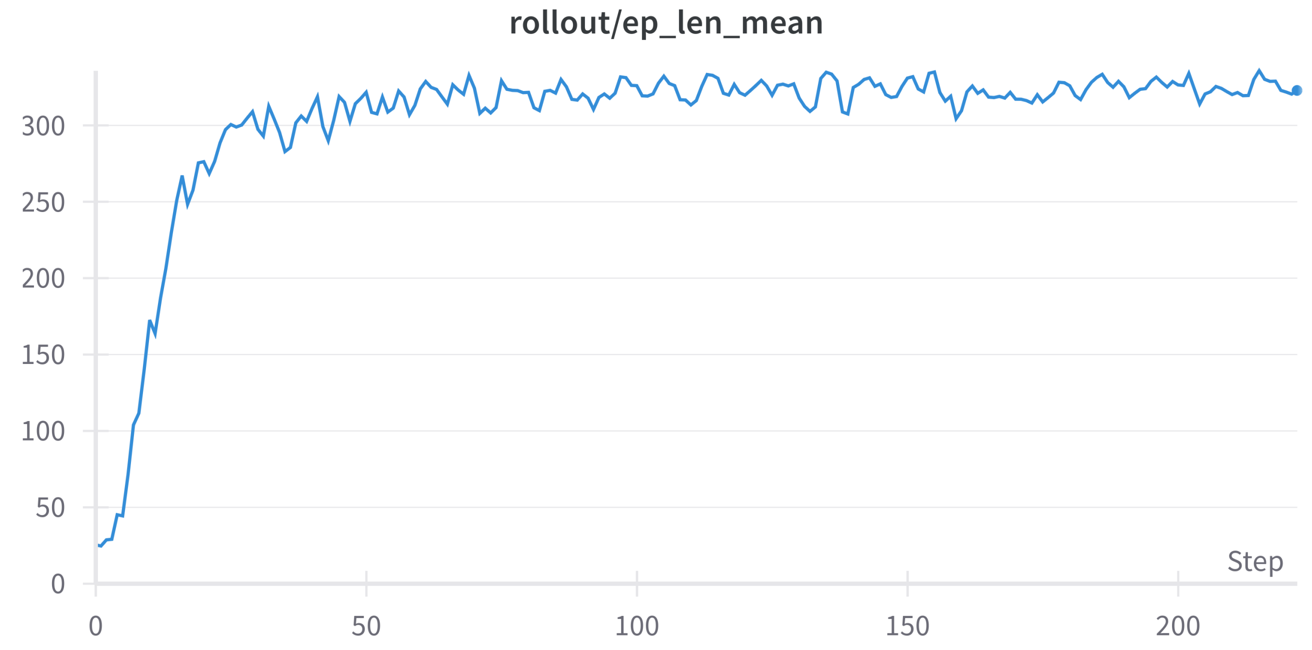
$$r = \sum_{i=0}^{6} \omega_i I_i$$

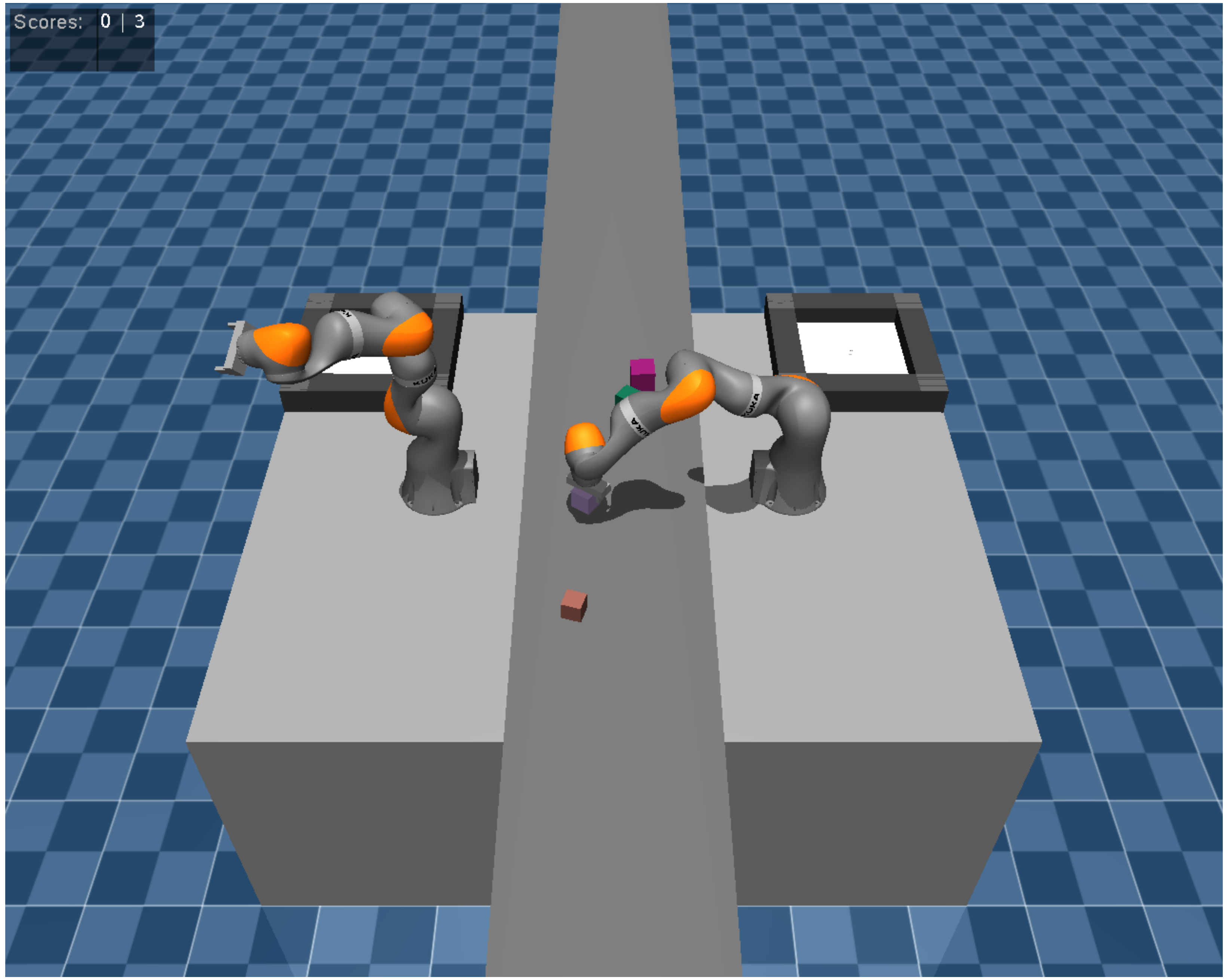
- goal: $\omega_0 >> \omega_i$ for $i \geq 1$
- ideally most $\omega_i = 0$

First Results

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







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

Next Steps

- explore denser reward augmentations
- use other algos/modifications for sparse case
- introduce RL for 2nd arm as well
- construct multi-agent env

TUMultuous

Discussion time!