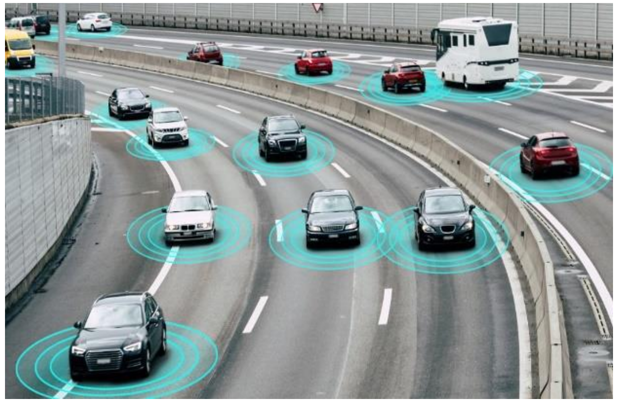
# Hierarchical Multi-Agent Skill Discovery

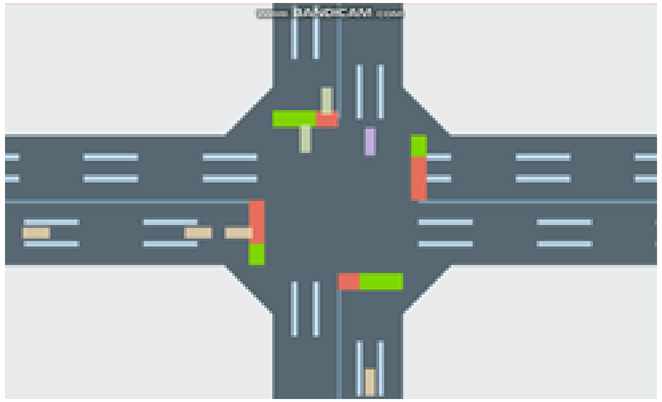
Mingyu Yang, Yaodong Yang, Zhenbo Lu, Wengang Zhou, Houqiang Li

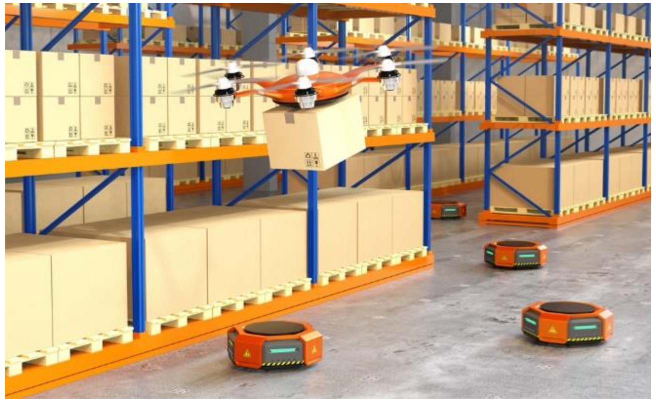
NeurIPS 2023

# Background

- MARL has recently shown remarkable potential in various real-world problems



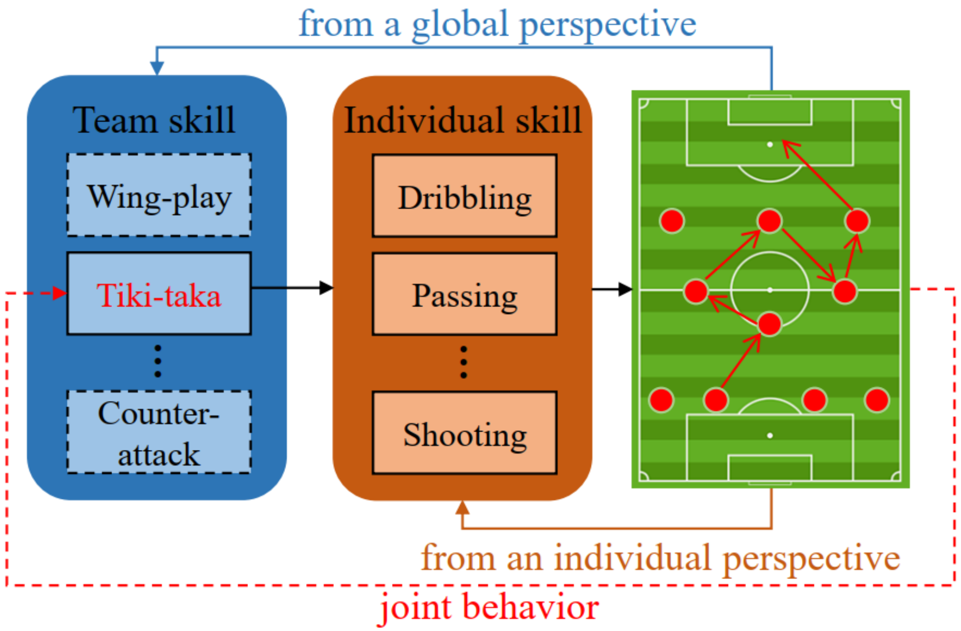




- However, current MARL algorithms (e.g., QMIX and MAPPO) typically reply on well-crafted team or individual rewards
- In this work, we focus on sparse reward multi-agent problems

# Motivation

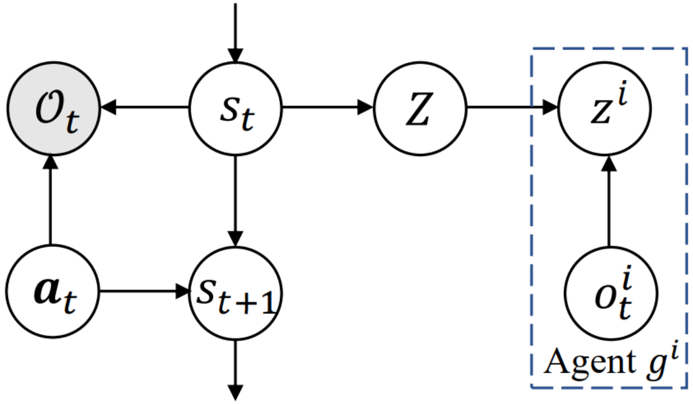
- Discover underlying skills within the multi-agent task and effectively combine these skills to achieve the final goal




- Two problems:
  - ➢ how to simultaneously learn team skill and individual skill
  - ➢ how to combine these skills to accomplish multi-agent tasks

# Multi-Agent Skill Discovery Problem Formulation

- We embed multi-agent skill discovery problem into a probabilistic graphical model




✓ $\mathcal{O}_t$ is a binary random variable, where $\mathcal{O}_t = 1$ denotes timestep $t$ is optimal, and $\mathcal{O}_t = 0$ indicates timestep $t$ is not optimal

✓ team skill $Z$ is conditioned on the global state $s_t$

✓ individual skill $z^i$ is conditioned on both the team skill $Z$ and agent $g^i$'s partial observation $o_t^i$

- We then perform structured variational inference to derive our objective

$$\log p(\mathcal{O}_{0:T}) \geq \mathbb{E}_{\tau\sim q(\tau)}\Bigg[\sum_{t=0}^{T}\Bigg(r(s_t, \boldsymbol{a}_t) + \underbrace{\log p(Z|s_t) + \sum\nolimits_{i=1}^{n} \log p(z^i|o_t^i, Z)}_{\text{diversity term}}$$

$$\underbrace{-\log q(Z|s_t) - \sum\nolimits_{i=1}^{n} \log q(z^i|o_t^i, Z)}_{\text{skill entropy term}} \underbrace{- \sum\nolimits_{i=1}^{n} \log q(a_t^i|o_t^i, z^i)}_{\text{action entropy term}}\Bigg)\Bigg],$$

# Method

- To optimize the derived lower bound, we utilize four approximate functions and integrate them into a two-level hierarchical structure
  - high-level skill coordinator

$$\pi_h(Z, z^{1:n} | s_t, \boldsymbol{o}_t) \rightarrow q(Z|s_t), q(z^i | o_t^i, Z)$$

  - low-level skill discoverer

$$\pi_l\left(a_t^i \middle| o_t^i, z^i\right) \rightarrow q\left(a_t^i \middle| o_t^i, z^i\right)$$

  - team skill discriminator

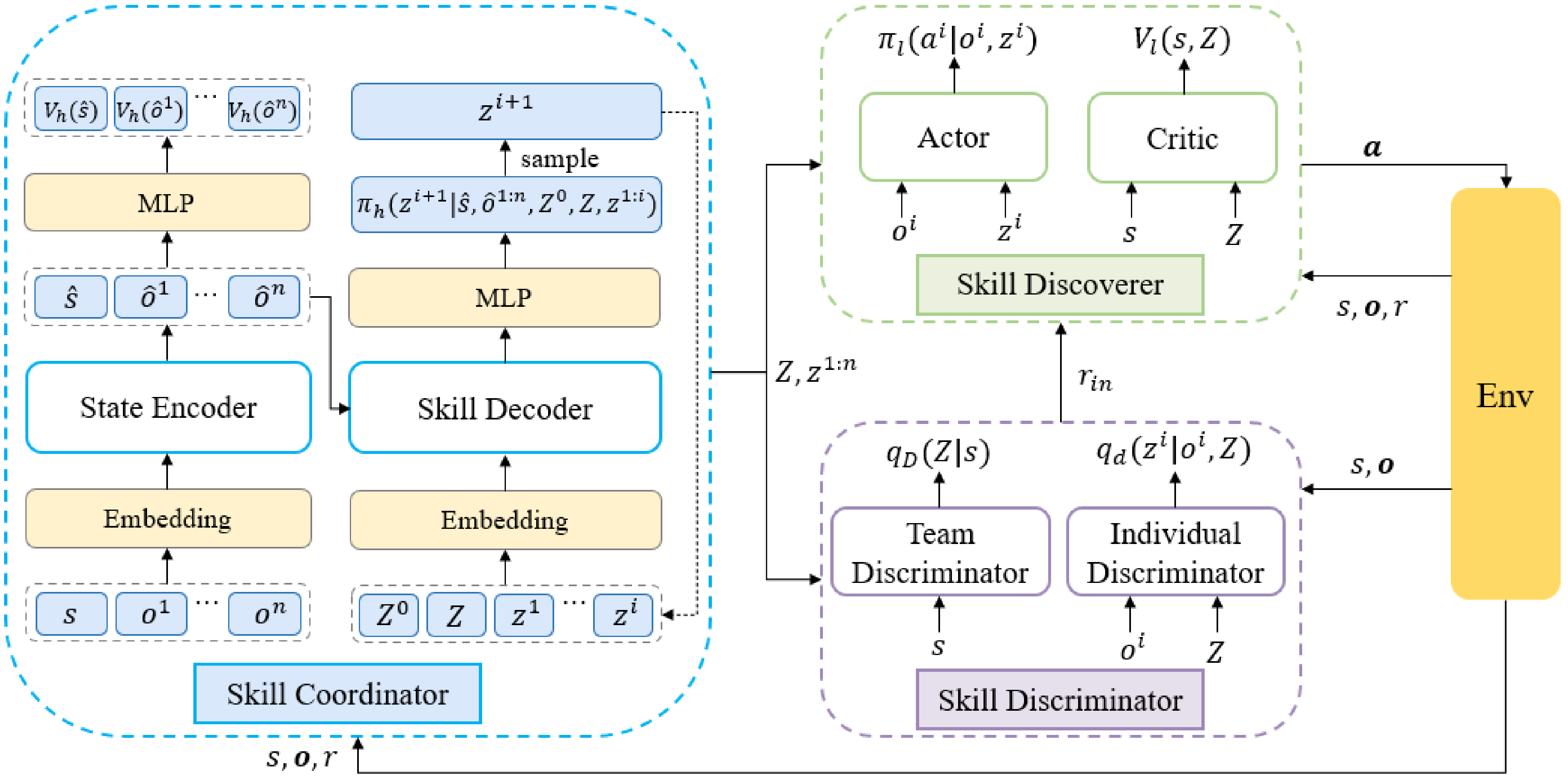
$$q_D(Z|s_t) \rightarrow p(Z|s_t)$$

  - individual skill discriminator

$$q_d\left(z^i \middle| o_t^i, Z\right) \rightarrow p\left(z^i \middle| o_t^i, Z\right)$$

# Method

- The Overall Framework



NEURAL INFORMATION PROCESSING SYSTEMS

# Method

- Overall Training
  - reward for high-level policy: $r_t^h = \sum_{p=0}^{k-1} r_{t+p}$
  - reward for low-level policy: $r_t^i = \lambda_e r_t + \lambda_D \log q_D(Z|s_{t+1}) + \lambda_d \log q_d(z^i|o_{t+1}^i, Z)$
  - we adopt the popular PPO objective to optimize both the high-level and low-level policy

$$\log p(\mathcal{O}_{0:T}) \geq \mathbb{E}_{\tau\sim q(\tau)}\left[\sum_{t=0}^{T}\Big(r(s_t, \boldsymbol{a}_t) + \underbrace{\log p(Z|s_t) + \sum\nolimits_{i=1}^{n} \log p(z^i|o_t^i, Z)}_{\text{diversity term}}\right.$$

used as extrinsic reward for both high-level and low-level policy

used as intrinsic reward for low-level policy

$$\left.\underbrace{-\log q(Z|s_t) - \sum\nolimits_{i=1}^{n} \log q(z^i|o_t^i, Z)}_{\text{skill entropy term}} \underbrace{- \sum\nolimits_{i=1}^{n} \log q(a_t^i|o_t^i, z^i)}_{\text{action entropy term}}\Big)\right],$$

entropy of high-level policy

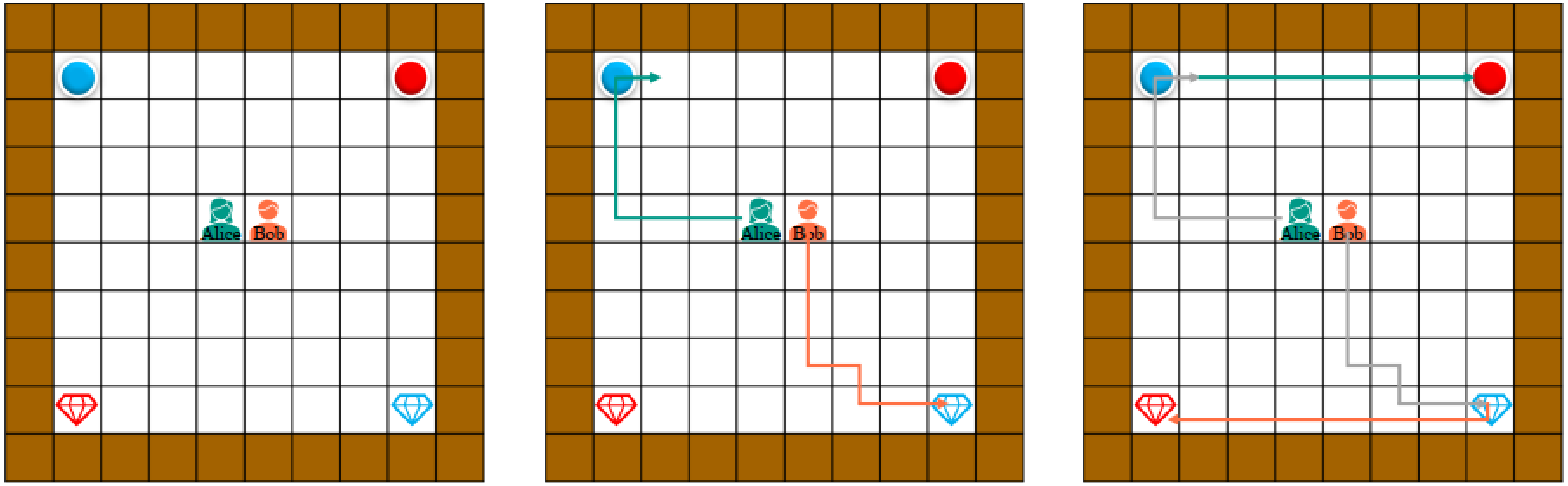
entropy of low-level policy

  - the skill discriminator is trained in a supervised manner with cross-entropy loss

$$\mathcal{L}_d(\phi_D, \phi_d) = -\mathbb{E}_{(s,Z)\sim\mathcal{D}}\left[\log q_D(Z|s)\right] - \sum_{i=1}^{n} \mathbb{E}_{(o^i,Z,z^i)\sim\mathcal{D}}\left[\log q_d(z^i|o^i, Z)\right],$$

# Experiments

- Case Study




(a) Alice_and_Bob (b) $Z = 0, z^1 = 3, z^2 = 2$ (c) $Z = 1, z^1 = 1, z^2 = 0$

➢ two team skills $Z = 0{,}1$ correspond to collecting the blue diamond and red diamond for the whole team

➢ four individual skills $z^i = 0, 1, 2, 3$ guide the individual agent to reach the red diamond, red button, blue diamond and blue button

# Experiments

- Performance on SMAC with 0-1 reward and Overcooked