Celebrating Diversity in Shared Multi-Agent Reinforcement Learning

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang*, Chongjie Zhang*
Tsinghua University

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Ho-Bin Choi

2022. 03. 16

http://link.koreatech.ac.kr

LINK@KoreaTech

Laboratory of Intelligent Networks at KoreaTech

Abstract

◈ MARL's success is partly because of parameter sharing among agents

- However, such sharing may lead agents to behave similarly and limit their coordination capacity

◈ propose an information-theoretical regularization to maximize the mutual information between agents' identities and their trajectories, encouraging extensive exploration and diverse individualized behaviors

◈ incorporate agent-specific modules in the shared neural network architecture, which are regularized by L1-norm to promote learning sharing among agents while keeping necessary diversity.

Abstract


FRAME: 22
TIME: 0.079144
SCORE: 0 – 0
BALL OWNED TEAM: –1
BALL OWNED PLAYER: –1
REWARD 0.0000
CUM. REWARD: 0.0000


TIME: 0.082612
BALL OWNED TEAM: 0
BALL OWNED PLAYER: 8
REWARD 0.0000
CUM. REWARD: 0.0000


FRAME: 91
TIME: 0.079772
SCORE: 0 – 0
BALL OWNED TEAM: 0
BALL OWNED PLAYER: 1
REWARD 0.0000
CUM. REWARD: 0.0000

Introduction

◈ One central problem is that the joint action-observation space grows exponentially with the number of agents, which **imposes high demand on the scalability of learning algorithms**

◈ To address this scalability challenge, *policy decentralization with shared parameters* (PDSP) is widely used, where agents share their neural network weights

◈ To address this scalability challenge, policy decentralization with shared parameters (PDSP) is widely used, where agents share their neural network weights

Introduction

◈ Complex tasks typically require substantial exploration and diversified strategies among agents

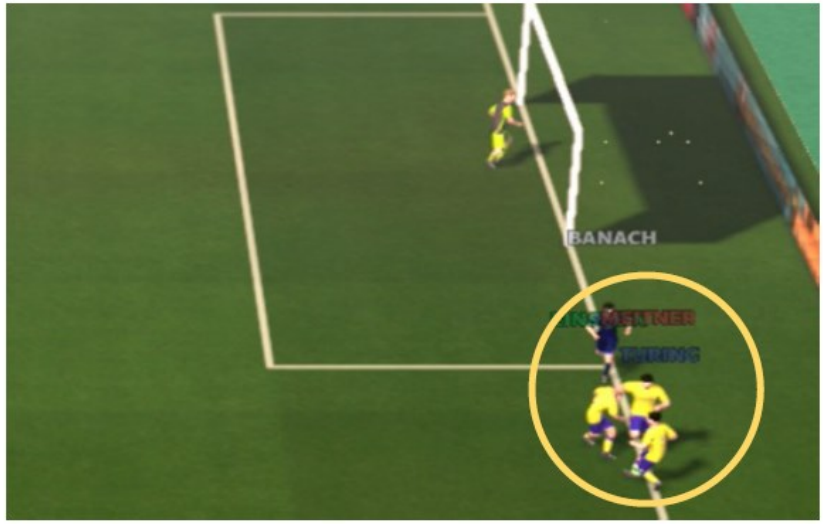
- When parameters are shared, agents tend to acquire homogeneous behaviors because they typically adopt similar actions under similar observations, preventing efficient exploration and the emergence of sophisticated cooperative policies

◈ Notably, sacrificing the merits of parameter sharing for diversity is also unfavorable

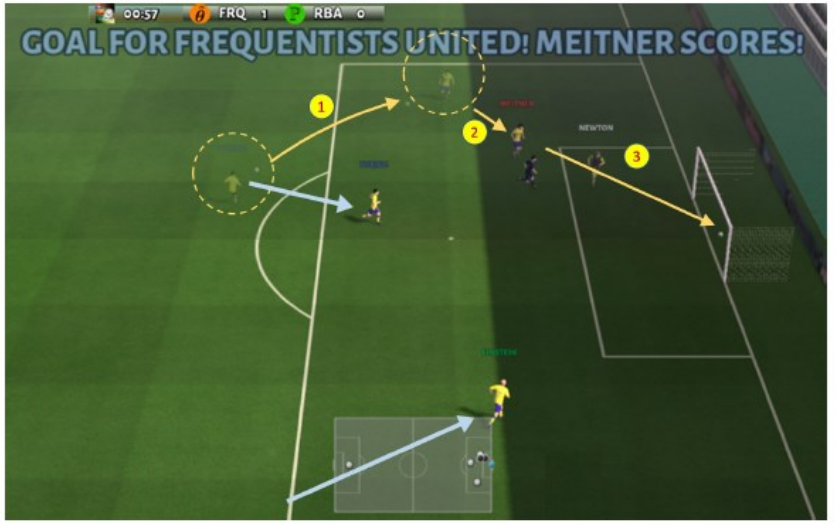
- Like humans, sharing necessary experience or understanding of tasks can broadly accelerate cooperation learning
- Without parameter sharing, agents search in a much larger parameter space, which may be wasteful because they do not need to behave differently all the time

◈ Therefore, the question is how to adaptively trade-off diversity and sharing

Introduction



(a) Parameter sharing: similar behaviors (competing for ball).



(b) Our approach: each agent has its responsibility to score.

Agents' history location → The goal's passing, dribbling and shooting → Off-the-ball moving

Figure 1: Shared parameters induce behaviors (**left**) and can hardly learn successful policies on the challenging Google Research Football task. Our method learns sophisticated cooperative strategies by **trading off diversity and sharing** (**right**).

Background

◈ Dec-POMDP, which is defined as a tuple $\mathcal{G} = \langle N, S, A, P, R, O, \Omega, n, \gamma \rangle$

- N is a finite set of n agents
- $s \in S$ is the true state of the environment
- A is the set of actions
- $\gamma \in [0, 1)$ is a discount factor
- At each time step, each agent $i \in N$
  - ✓ receives his own observation $o_i \in \Omega$ according to the observation function $O(s, i)$
  - ✓ selects an action $a_i \in A$, which results in a joint action vector $\boldsymbol{a}$
- $P(s'|s, \boldsymbol{a})$ is the transition function
- A global reward $r = R(s, \boldsymbol{a})$ is shared by all the agents
- Each agent has its own action-observation history $\tau_i \in \mathcal{T}_i \doteq (\Omega_i \times A)^*$

Method: 3.1 Identity-Aware Diversity

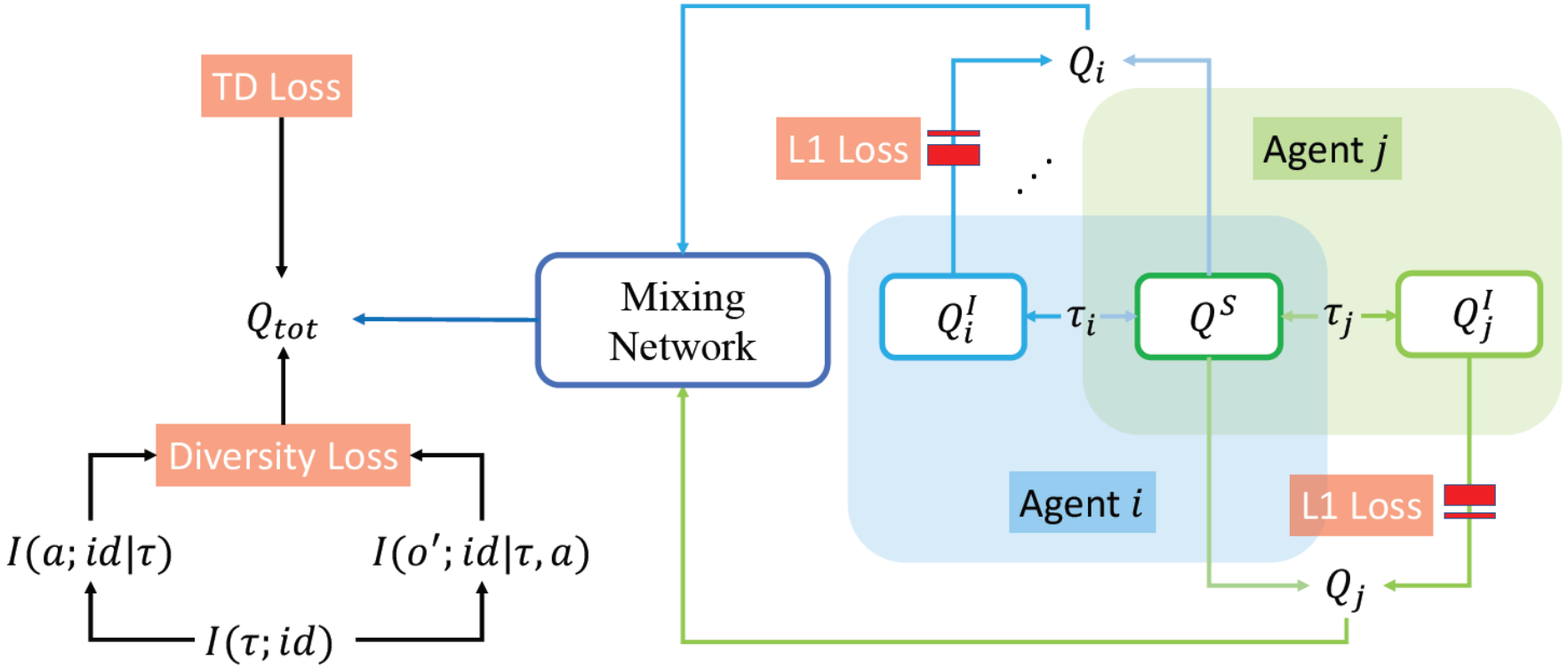


Figure 2: Schematics of our approach.

Information Theory

$$H(X) = -\sum_{x} p(x) \log p(x) = \mathbb{E}\left[\log \frac{1}{p(x)}\right]$$

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) log_2 p(x,y) - \sum_{x,y} p(x,y) log_2 p(x) - \sum_{x,y} p(x,y) log_2 p(y) \\ &= -H(X,Y) + H(X) + H(Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \end{aligned}$$

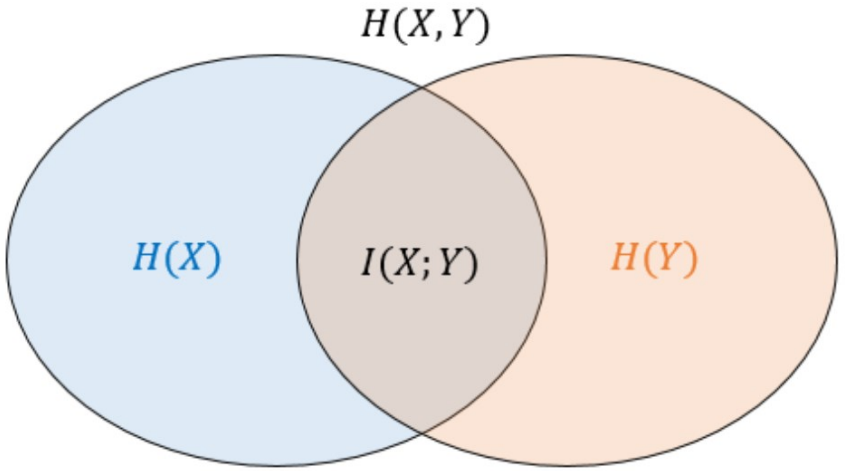


그림2 상호정보량

$$D_{KL}(P \parallel Q) = \sum_{i=0}^{n} p(x_i) log\left(\frac{p(x_i)}{q(x_i)}\right) \longrightarrow D_{KL}(P \parallel Q) = \sum_{i=0}^{n} p(x_i) log(p(x_i)) - \sum_{i=0}^{n} p(x_i) log(q(x_i))$$

Method: 3.1 Identity-Aware Diversity

◈ An information-theoretic objective for maximizing the mutual information between individual trajectory and agents' identity

$$I^{\boldsymbol{\pi}}(\tau_T; id) = H(\tau_T) - H(\tau_T|id) = E_{id,\tau_T\sim\boldsymbol{\pi}}\left[\log \frac{p(\tau_T|id)}{p(\tau_T)}\right], \tag{1}$$

$$p(\tau_T) \Rightarrow p(\tau_T) \text{ as } p(o_0)\prod_{t=0}^{T-1} p(a_t|\tau_t)p(o_{t+1}|\tau_t, a_t)$$

$$p(\tau_T|id) \Rightarrow p(o_0|id)\prod_{t=0}^{T-1} p(a_t|\tau_t, id)p(o_{t+1}|\tau_t, a_t, id)$$

$$I^{\boldsymbol{\pi}}(\tau_T; id) = E_{id,\tau}\Bigg[\underbrace{\log \frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}}_{③}\Bigg]. \tag{2}$$

Method: 3.1 Identity-Aware Diversity

◈ Term ① is determined by the environment, and we can ignore it when optimizing the mutual information

◈ The second term quantifies the information gain about agent's action selection when the identity is given, which measures action-aware diversity as $I(a; id|\tau)$

$$I^{\pi}(\tau_T; id) = E_{id,\tau}\Big[\underbrace{\log \frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}}_{③}\Big]. \quad (2)$$

Method: 3.1 Identity-Aware Diversity

◈ However, $p(a_t|\tau_t, id)$ is typically the distribution induced by ϵ-greedy, which only distinguishes the action with the highest possibility

◈ Therefore, directly optimizing this term conceals most information about the local Q-functions

◈ To solve this problem, we use the Boltzmann softmax distribution of local Q values to replace $p(a_t|\tau_t, id)$, which forms a lower bound of term ②:

$$E_{id,\tau}\left[\log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}\right] \geq E_{id,\tau}\left[\log \frac{\mathrm{SoftMax}(\frac{1}{\alpha}Q(a_t|\tau_t, id))}{p(a_t|\tau_t)}\right]. \tag{3}$$

◈ We maximize this lower bound to optimize Term ②

$$I^{\boldsymbol{\pi}}(\tau_T; id) = E_{id,\tau}\Bigg[\underbrace{\log \frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}}_{③}\Bigg]. \tag{2}$$

Method: 3.1 Identity-Aware Diversity

◈ Inspired by variational inference approaches, we derive and optimize a tractable lower bound for Term ③ at each time step by introducing a variational posterior estimator q_ϕ parameterized by ϕ:

$$E_{id,\tau}\left[\log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}\right] \geq E_{id,\tau}\left[\log \frac{q_\phi(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}\right], \tag{4}$$

◈ Intuitively, optimizing Eq. 4 encourages agents to have diverse observations that are distinguishable by agents' identification and thus measures observation-aware diversity as $I(o'; id|\tau, a)$

$$I^{\boldsymbol{\pi}}(\tau_T; id) = E_{id,\tau}\Bigg[\underbrace{\log \frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}}_{③}\Bigg]. \tag{2}$$

Method: 3.1 Identity-Aware Diversity

◈ To tighten the this lower bound, we minimize the KL divergence with respect to the parameters ϕ

◈ The gradient for updating ϕ is:

$$\begin{aligned}\nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi \mathbb{E}_{\tau,a,id}\left[D_{\mathrm{KL}}\left(p\left(\cdot|\tau,a,id\right)\|q_\phi\left(\cdot|\tau,a,id\right)\right)\right] = \nabla_\phi \mathbb{E}_{\tau,a,id,o'}\left[\log\frac{p\left(o'|\tau,a,id\right)}{q_\phi\left(o'|\tau,a,id\right)}\right] \\ &= -\mathbb{E}_{\tau,a,id,o'}\left[\nabla_\phi \log q_\phi\left(o'|\tau,a,id\right)\right].\end{aligned} \tag{5}$$

$$I^{\boldsymbol{\pi}}(\tau_T; id) = E_{id,\tau}\Bigg[\underbrace{\log\frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1}\log\frac{p(a_t|\tau_t,id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1}\log\frac{p(o_{t+1}|\tau_t,a_t,id)}{p(o_{t+1}|\tau_t,a_t)}}_{③}\Bigg]. \tag{2}$$

Method: 3.1 Identity-Aware Diversity

◈ Based on the lower bounds shown in Eq. 3 and Eq. 4, we introduce intrinsic rewards to optimise the information-theoretic objective (Eq. 1) for encouraging diverse behaviors:

$$r^I = E_{id}\left[\beta_2 D_{\mathrm{KL}}(\mathrm{SoftMax}(\beta_1 Q(\cdot|\tau_t, id))||p(\cdot|\tau_t)) + \beta_1 \log q_\phi(o_{t+1}|\tau_t, a_t, id) - \log p(o_{t+1}|\tau_t, a_t)\right]. \tag{6}$$

$$p\left(a_t|\tau_t\right) \approx \frac{1}{n}\sum_{id} \pi\left(a_t|\tau_t, id\right), \tag{13}$$

$$r^I = E_{id}\left[\beta_2 D_{\mathrm{KL}}(\mathrm{SoftMax}(\beta_1 Q(\cdot|\tau_t, id))||p(\cdot|\tau_t)) + \beta_1 \log q_{\eta_1}\left(id|o_{t+1}, \tau_t, a_t\right) - \log q_{\eta_2}\left(id|\tau_t, a_t\right)\right]. \tag{21}$$

$$I^{\boldsymbol{\pi}}(\tau_T; id) = E_{id,\tau}\Bigg[\underbrace{\log \frac{p(o_0|id)}{p(o_0)}}_{①} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(a_t|\tau_t, id)}{p(a_t|\tau_t)}}_{②} + \underbrace{\sum_{t=0}^{T-1} \log \frac{p(o_{t+1}|\tau_t, a_t, id)}{p(o_{t+1}|\tau_t, a_t)}}_{③}\Bigg]. \tag{2}$$

Method: 3.2 Action-Value Learning for Balancing Diversity and Sharing

◈ Defining experiences that need to be shared or exclusively learned is inefficient and usually can not generalize

◈ Therefore, we let agents adaptively decide whether to share experiences by decomposing Q_i as:

$$Q_i(a_i|\tau_i) = Q^S(a_i|\tau_i) + Q_i^I(a_i|\tau_i), \tag{7}$$

- Q_i^I: individual local Q-function
- Q^S: shared Q-function
- Q_i: local Q-function

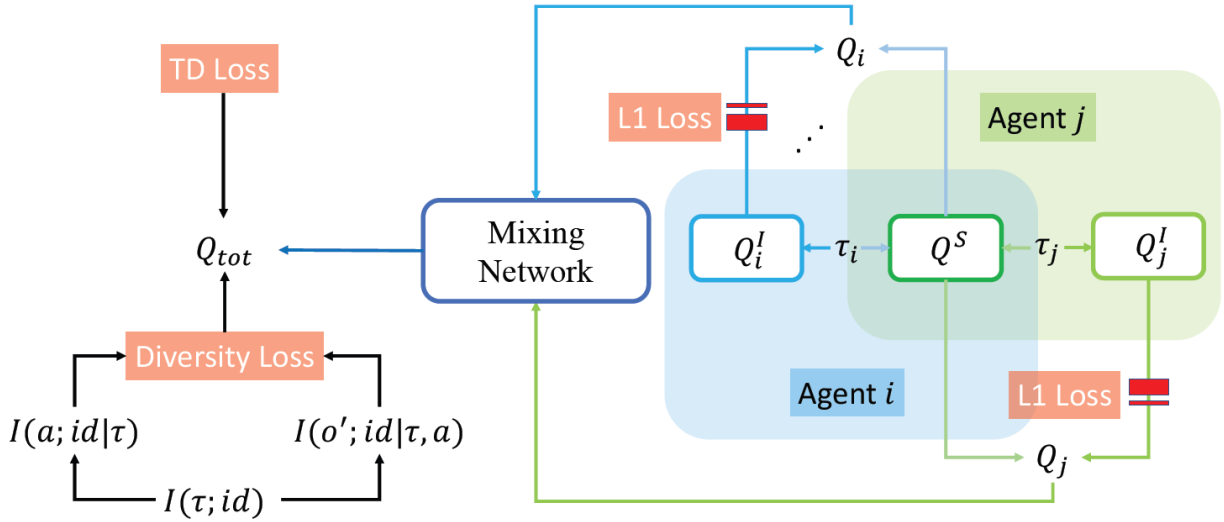


Figure 2: Schematics of our approach.

Method: 3.3 Overall Learning Objective

◈ Since the intrinsic rewards r^I inevitably involves the influence from all agents, we add r^I to environment rewards r^e and use the following TD loss:

$$\mathcal{L}_{TD}(\theta) = \left[r^e + \beta r^I + \gamma \max_{\boldsymbol{a}'} Q_{tot}\left(s', \boldsymbol{a}'; \theta^-\right) - Q_{tot}(s, \boldsymbol{a}; \theta)\right]^2, \tag{8}$$

◈ We use QPLEX to decompose Q_{tot} as mixing of local Q-functions Q_i and train the framework end-to-end by minimizing the loss:

$$\mathcal{L}(\theta) = \mathcal{L}_{TD}(\theta) + \lambda \sum_i \mathcal{L}_{L_1}(Q_i^I(\theta_i^I)), \tag{9}$$

- θ_i^I is the parameters of Q_i^I
- $\mathcal{L}_{L_1}(Q_i^I)$ is the L1 regularization term for independent Q-functions
- λ is a scaling factor

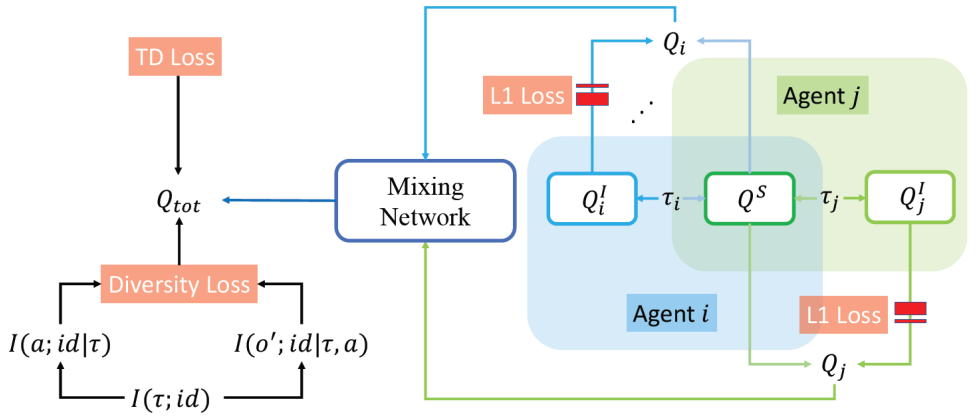
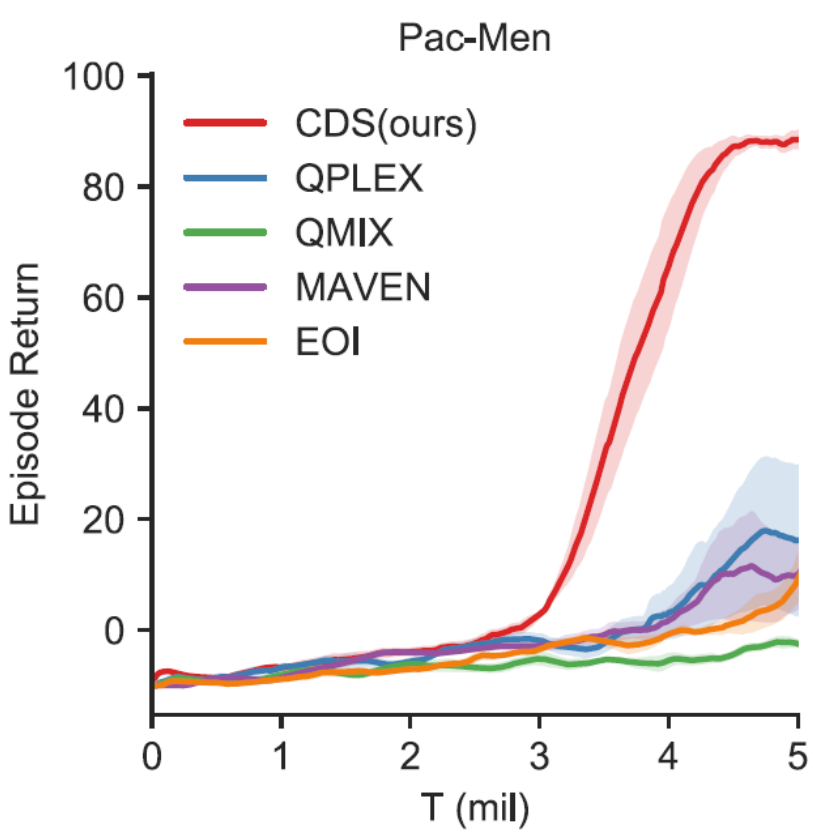
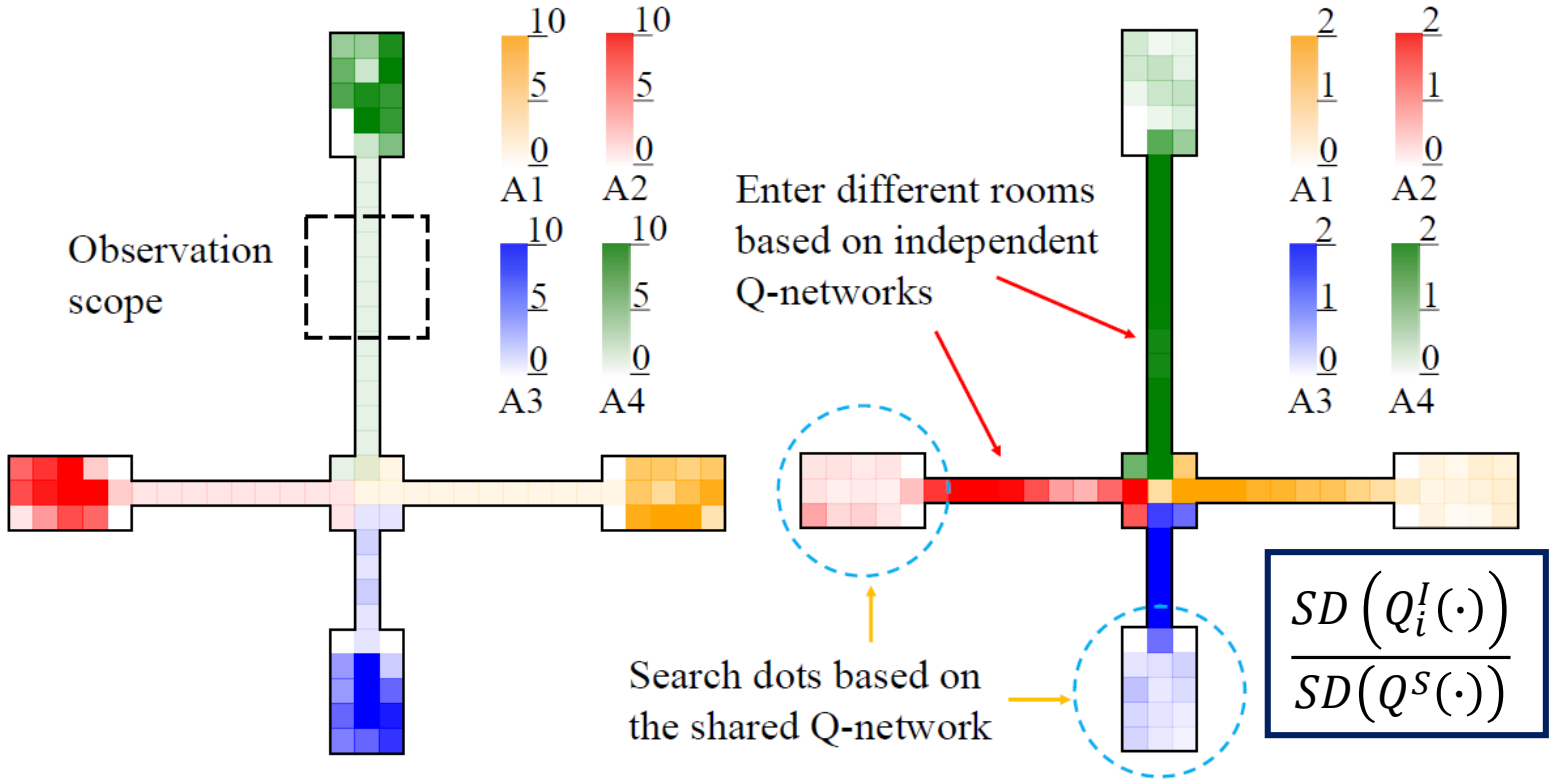


Figure 2: Schematics of our approach.

Case study: outperforming by being diverse only when necessary

A higher SD ratio indicates the independent Q-functions play a leading role, while a lower SD ratio indicates the shared Q-function's domination





(a) Comparison against baselines (b) Visitation heatmap (c) Behavior diversity heatmap

Figure 3: Why does our method work? The balance between identity-aware diversity and experience sharing encourages sophisticated strategies.

Experiments: Performance on Google Research Football (GRF)

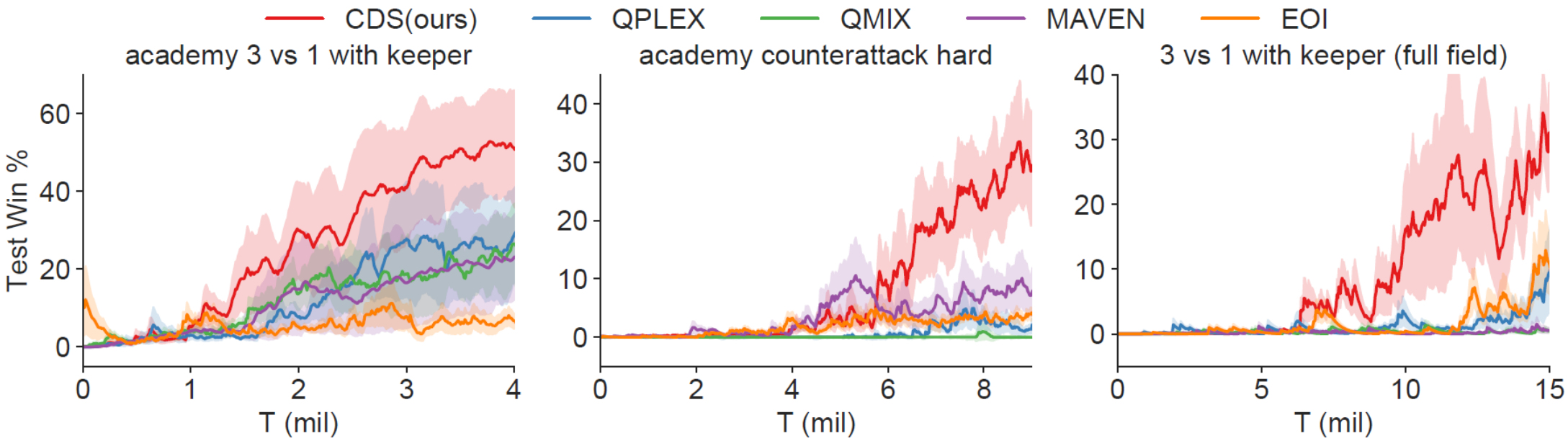


Figure 4: Comparison of our approach against baseline algorithms on Google Research Football.

Experiments: Performance on StarCraft II

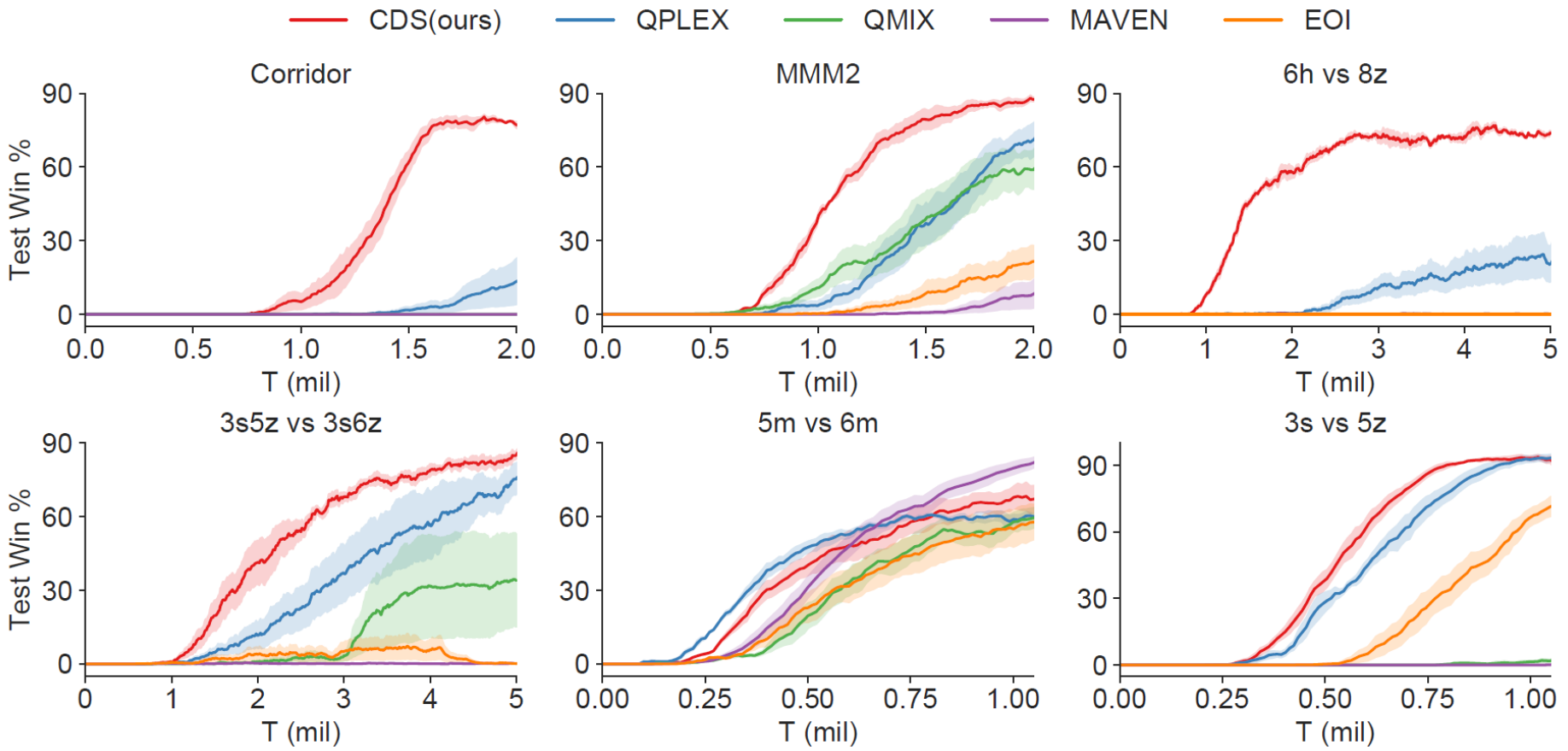


Figure 5: Comparison of our approach against baseline algorithms on four **super hard** SMAC maps: `corridor`, `MMM2`, `6h_vs_8z`, and `3s5z_vs_3s6z` and two **hard** SMAC maps: `5m_vs_6m` and `3s_vs_5z`.

Experiments: Ablations and Visualization

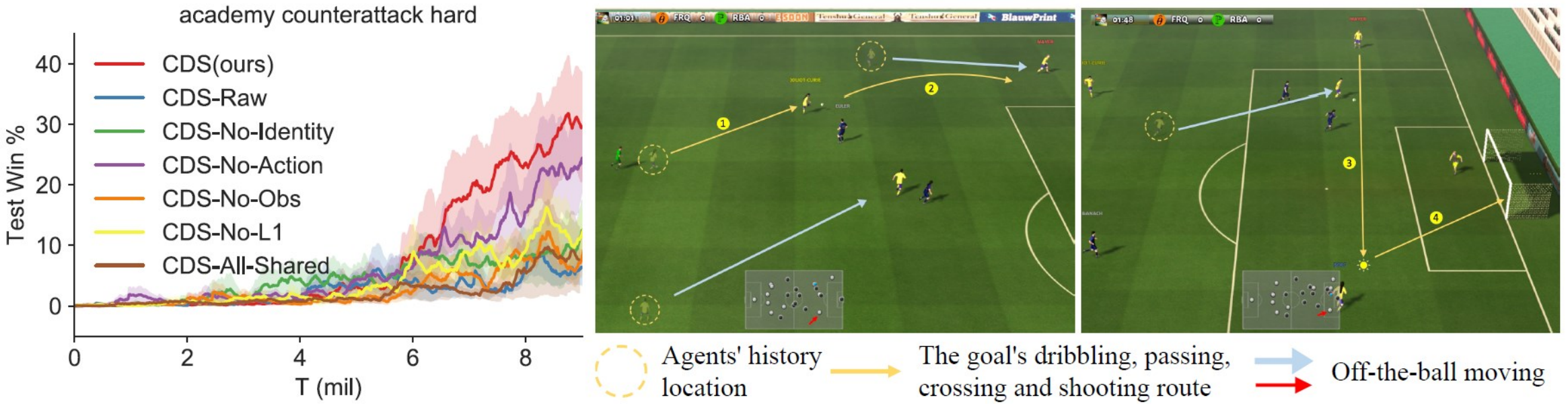


Figure 6: **Left.** Ablation studies on `academy_counterattack_hard`. **Right.** Visualization of trained policies, which achieve complex cooperation with impressive off-the-ball moving strategies.

Experiments: Ablations and Visualization

Figure 7: **Left.** Ablation studies in super hard map `corridor`. **Right.** Visualization of the final trained strategies, which achieves a hard-earned victory brought by the sacrifice of a warrior.

감 사 합 니 다