Multi-Agent RL

Zhi Wang

Nanjing University, China

2025-04-27

Contents

Reinforcement Learning

- Supervised Learning
  - (input, label)
- Unsupervised Learning
  - (input)
- **Reinforcement Learning**
  - sequential decision-making

- Computer Vision
  - Input: image pixels
- Natural Language Processing
  - Input: sentences
- **Reinforcement Learning**
  - Input: states

The Era of RL

- Video games: Human-level control through DRL, Nature 2015 (视频游戏)
- AlphaGo, Nature 2016;   AlphaGo Zero, Nature 2017 (围棋)
- AlphaStar in StarCraft II, Nature 2019 (星际争霸II)
- DRL for legged robots, Science Robotics 2019 (机器人学习)
- Superhuman AI for multiplayer poker, Science 2019 (德州扑克，多人非完全信息博弈)
- Discovering faster matrix multiplication algorithms, Nature 2022 (矩阵相乘算法发现，基础数学)
- Magnetic control of tokamak plasmas, Nature 2022 (可控核聚变控制)
- Outracing champion Gran Turismo drivers, Nature 2022 (赛车模拟控制)
- Safety validation of autonomous vehicles, Nature 2023 (无人驾驶安全验证)
- Faster sorting algorithms discovering, Nature 2023 (排序算法发现，基础信息科学)
- Champion-level drone racing, Nature 2023 (无人机竞速)
- Mastering diverse control tasks through world models, Nature 2025
- ……

The Dilemma of RL

RL = Artificial General Intelligence (AGI)?

Yet?

The Dilemma of RL

Transformers

Attention is **all you need**
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent
… **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …
☆ Save 99 Cite Cited by 176805 Related articles All 73 versions

Vision Transformers

[PDF] An **image** is **worth** 16x16 words: Transformers for **image** recognition at scale
A Dosovitskiy, L Beyer, A Kolesnikov… - arXiv preprint arXiv …, 2020 - arxiv.org
… directly to **images**, with the fewest possible modifications. To do so, we split an **image** into patches … only to small-resolution **images**, while we handle medium-resolution **images** as well. …
☆ Save 99 Cite Cited by 60296 Related articles All 21 versions

Decision Transformers

Decision transformer: Reinforcement learning via sequence modeling
L Chen, K Lu, A Rajeswaran, K Lee… - Advances in neural …, 2021 - proceedings.neurips.cc
… of the **Transformer** architecture, and associated advances in language modeling such as GPT-x and BERT. In particular, we present **Decision Transformer**, … , **Decision Transformer** simply …
☆ Save 99 Cite Cited by 1942 Related articles All 13 versions

Artificial General Intelligence (AGI)

ChatGPT (Generative Pre-Training)

Next-token prediction

Transformer

Architecture

Enter text:

The dog eats the apples.

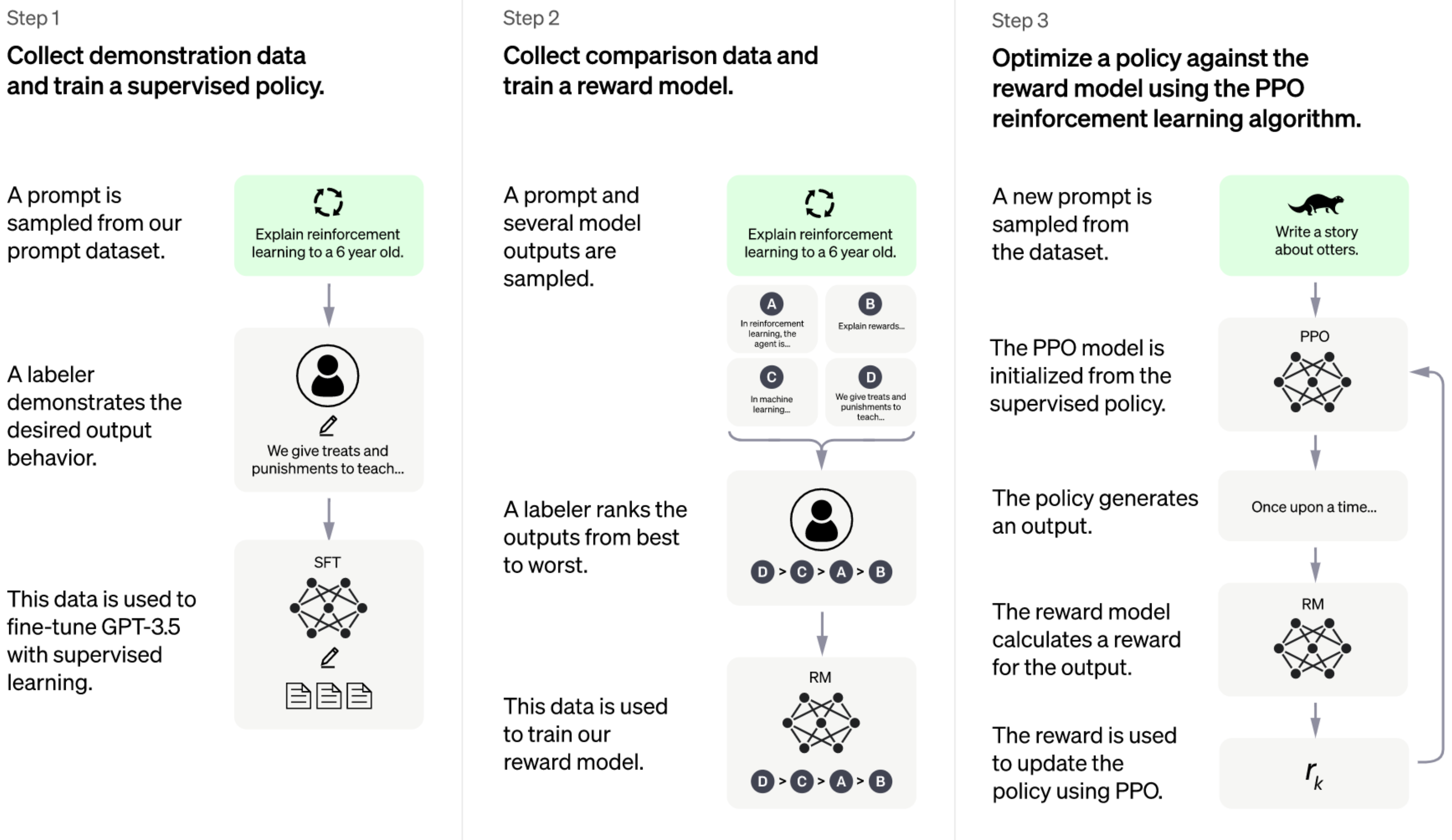
The	dog	eats	the	apples	.
464	3290	25365	262	22514	13

464 3290 25365 262 22514 13

Self-supervised learning

Algorithm

RL in ChatGPT



RL: Fine-tuning in Step 3, playing an auxiliary role

The Dilemma of RL

- Computer Vision
  - Input: **image pixels**
- Natural Language Processing
  - Input: **sentences**
- **Reinforcement Learning**
  - Input: **states, (states, actions)**

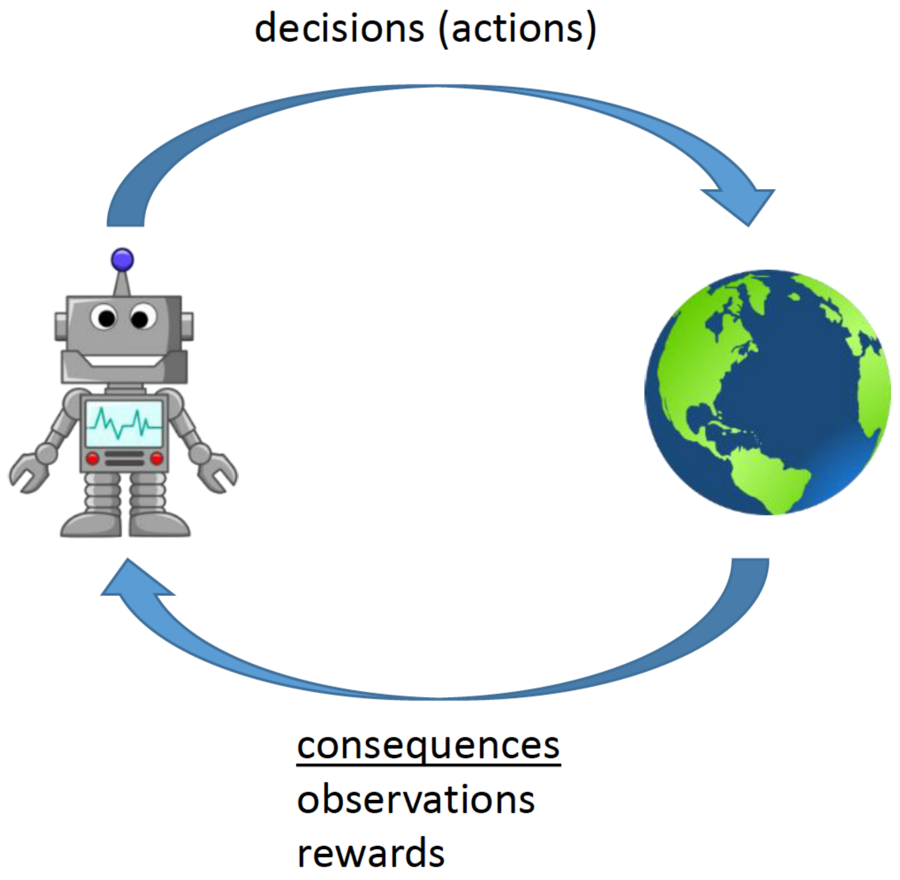
Semantics

not aligned

The Dilemma of RL

Data

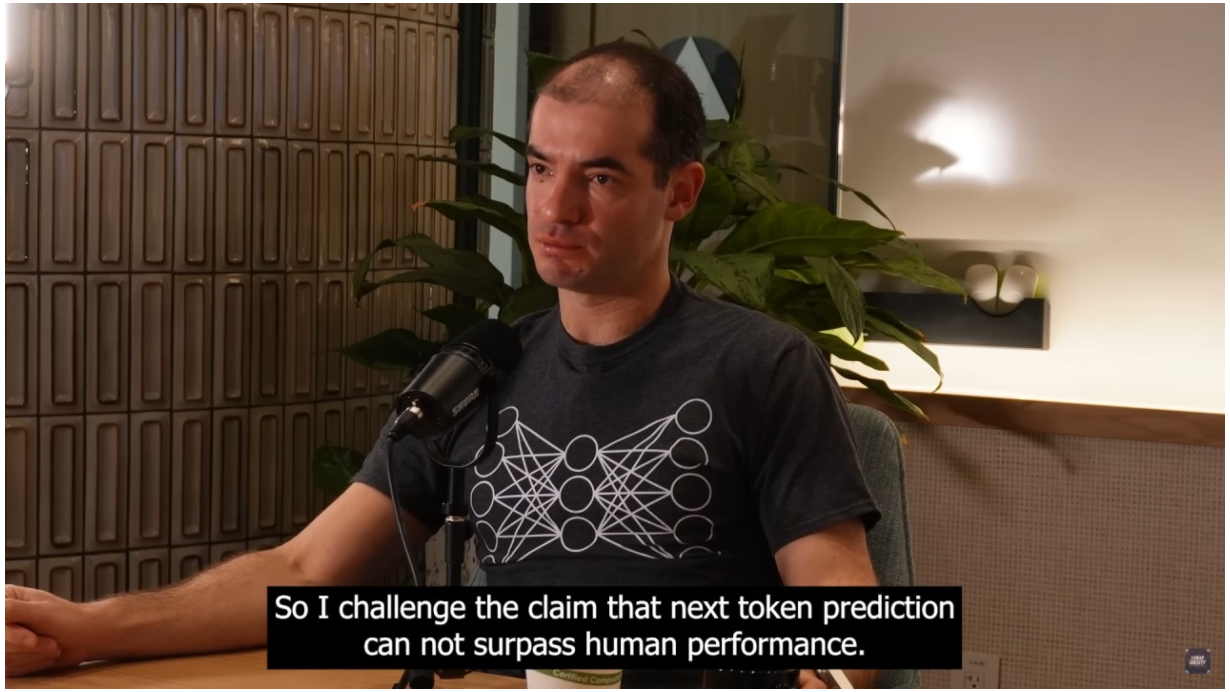
From online interactions



Confidence in RL

Supervised learning

Maybe imitating the intelligence within existing data?



Confidence in RL

Supervised learning

Maybe imitating the intelligence within existing data?

Reinforcement learning

Can surpass the intelligence within existing data definitely

Confidence in RL

LLM: From Pre-Training to Post-Training

Pre-training will end

-- by Ilya Sutskever

@NeurIPS 2025

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

Confidence in RL

LLM: From Pre-Training to Post-Training

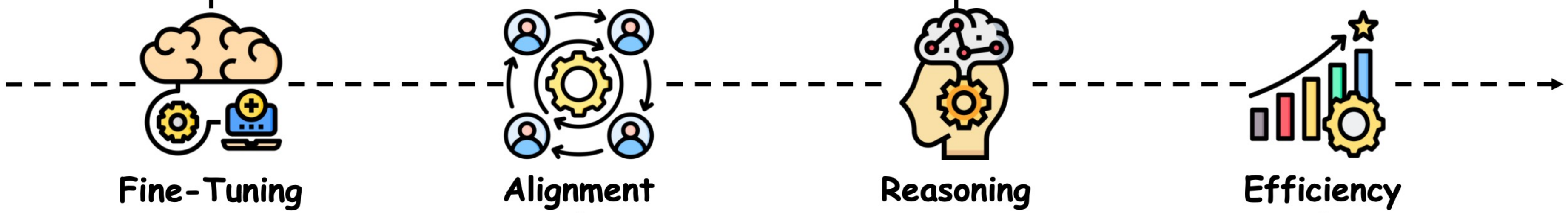
Reasoning, inference

What comes next?

- “Agents”??
- “Synthetic data”
- Inference time compute ~ O1

- Supervised Fine-Tuning
- Reinforcement Fine-Tuning

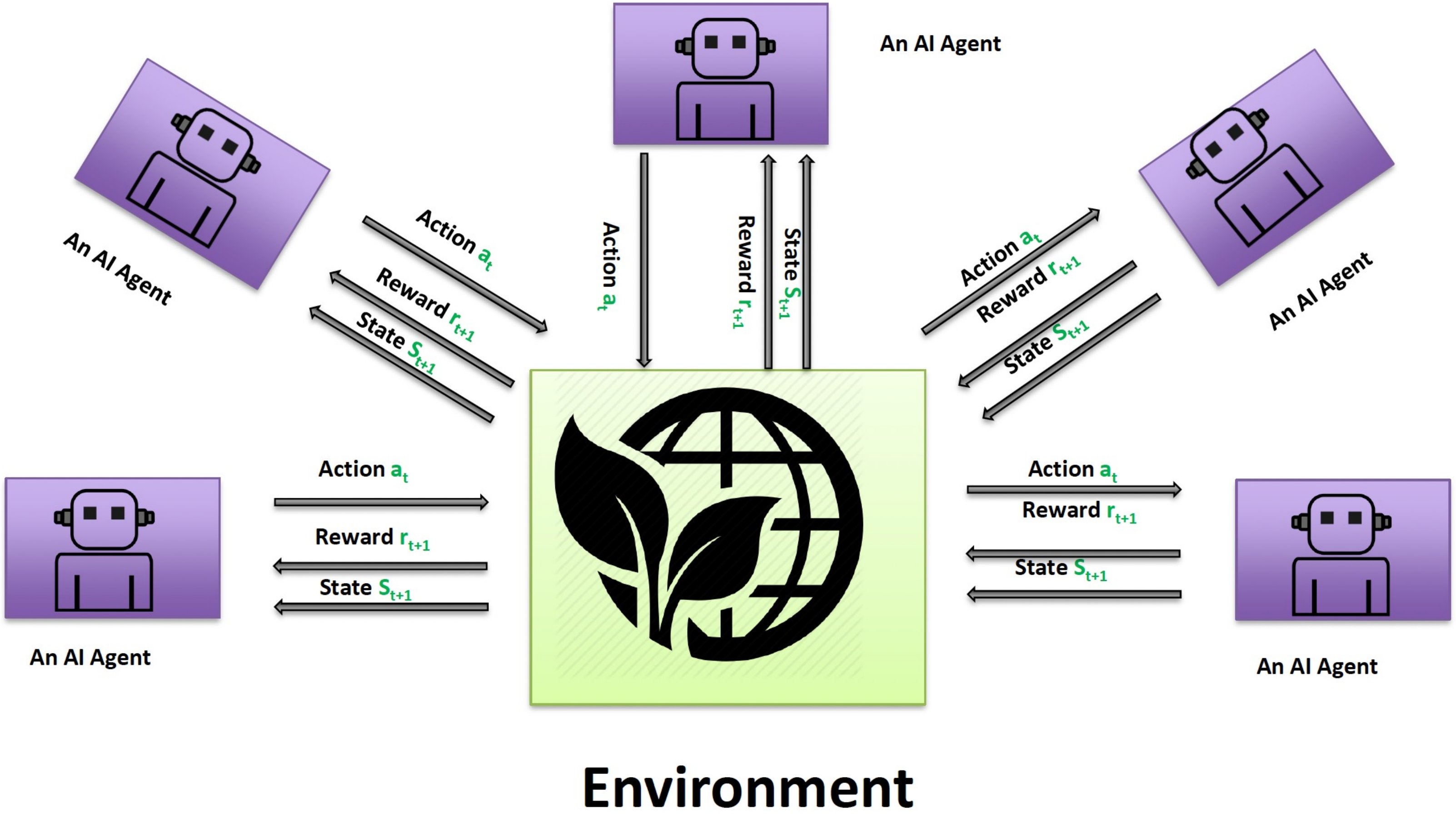
- Self-Refine for Reasoning
- Reinforcement Learning for Reasoning



- Reinforcement Learning with Human Feedback
- Direct Preference Optimization
- Group Relative Policy Optimization

- Model Compression
- Parameter-Efficient Fine-Tuning
- Knowledge Distillation

Multi-Agent RL



A set of autonomous agents that share a common environment

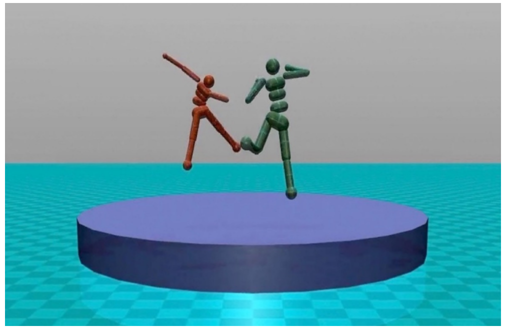
Difficulty in MARL

- **MARL is fundamentally difficult**
  - since agents not only interact with the environment but also with each other

- **If use single-agent Q-learning by considering other agents as a part of the environment**
  - Such a setting breaks the theoretical convergence guarantees and makes the learning unstable
  - i.e., the changes in strategy of one agent would affect the strategies of other agents and vice versa

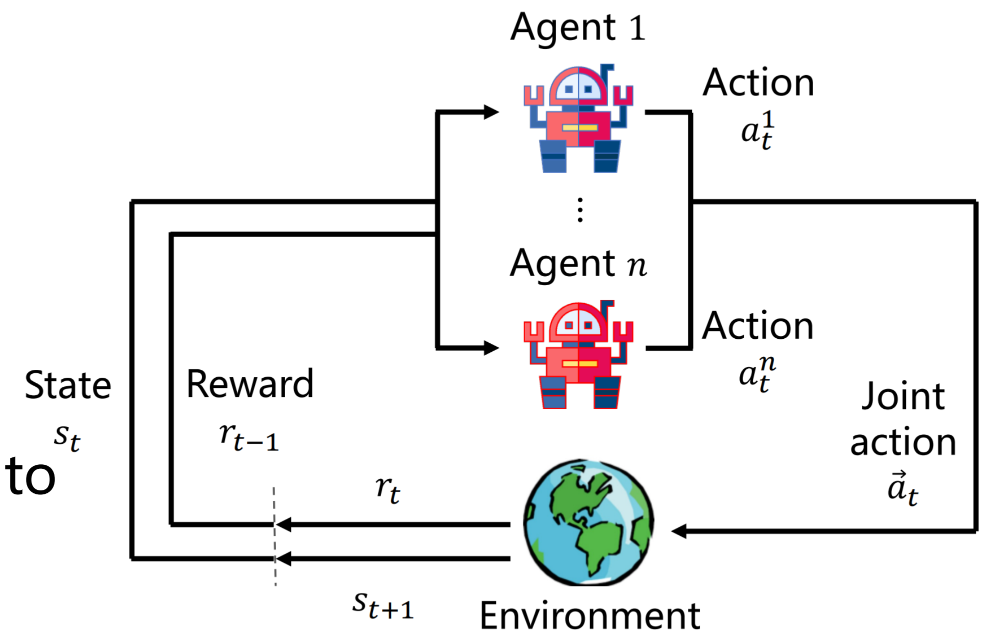
Types of Multi-agent Systems

- **Cooperative**
  - Maximizing a shared team reward
  - Coordination problems
- **Competitive**
  - Self-interested: maximizing an individual opposite reward
  - Zero-sum games
  - Minimax equilibria
- **Mixed**
  - Self-interested with different individual rewards (not opposite)
  - General-sum games



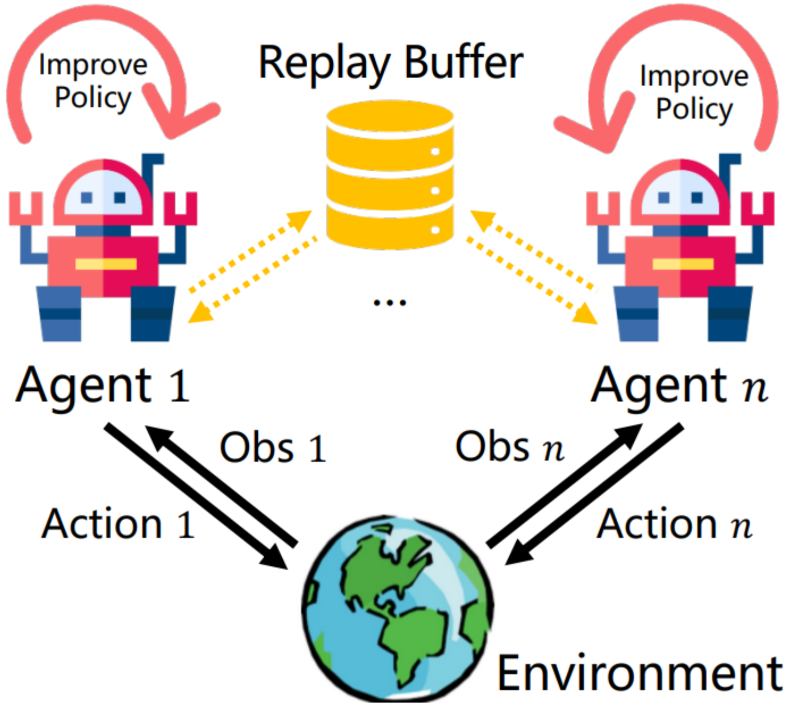
Cooperative Multi-Agent MDP

- **Assume agents can observe the global state**
  - Agent: $i \in I = \{1,2,\ldots,N\}$
  - State: $s \in S$
  - Action: $a_i \in A$, joint action $a =< a_1,\ldots,a_n >\in A^N$
  - Transition function: $\mathrm{P}(s'|s,\boldsymbol{a})$
  - Reward: $r(s,\boldsymbol{a})$
  - Agent $i's$ policy $\pi_i(s): S \rightarrow A$
  - Objective: finding a joint policy $\boldsymbol{\pi} =< \pi_1,\ldots,\pi_n >$ to maximize expected return: $\mathrm{R} = \sum_{t=0}^{\infty}\gamma^t r_t$
  - Value funtion: $Q^{\pi}(s,\boldsymbol{a}) = \mathbb{E}[R|s_0 = s, \boldsymbol{a_0} = \boldsymbol{a}, \boldsymbol{\pi}]$
  - With optimal $Q^*(s,\boldsymbol{a})$, optimal $\boldsymbol{\pi}^*(\boldsymbol{s}) = argmax_a\, Q^*(s,\boldsymbol{a})$



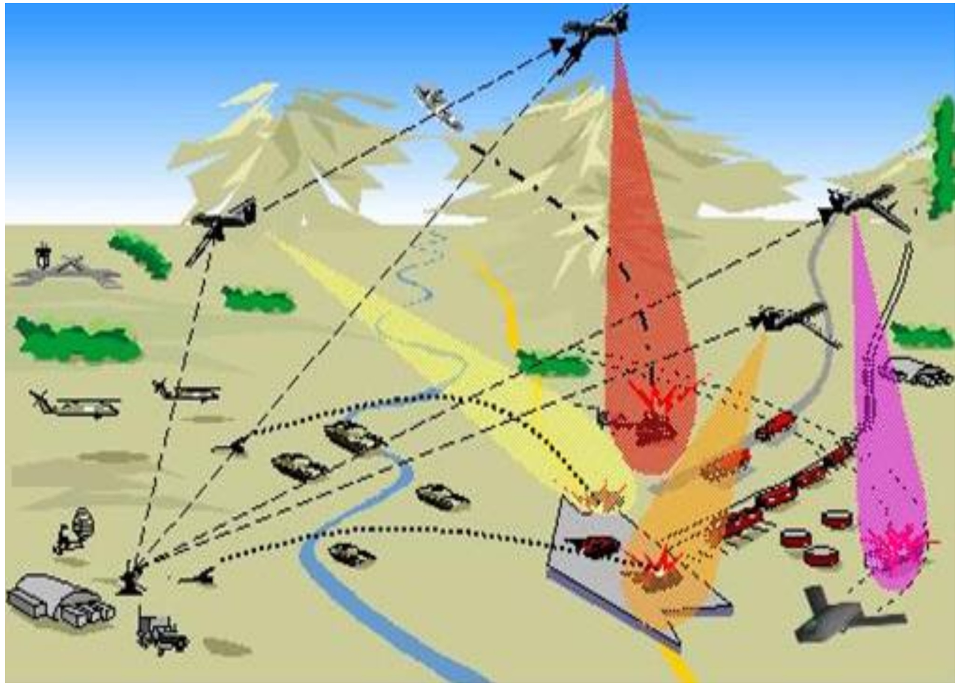
Decentralized Partially Observable MDP

- **Agent can not observe the global states**
  - Observation: $o_i \in \Omega$
  - Observation function: $o_i \in \Omega \sim O(s, i)$
- **Decentralized policy for agent i :**
  - $\pi_i(\tau_i): T \to A$
  - Action-observation history: $\tau_i \in T = (\Omega \times A)^*$
- **Communication and sensory constraints**
  - Decentralized execution

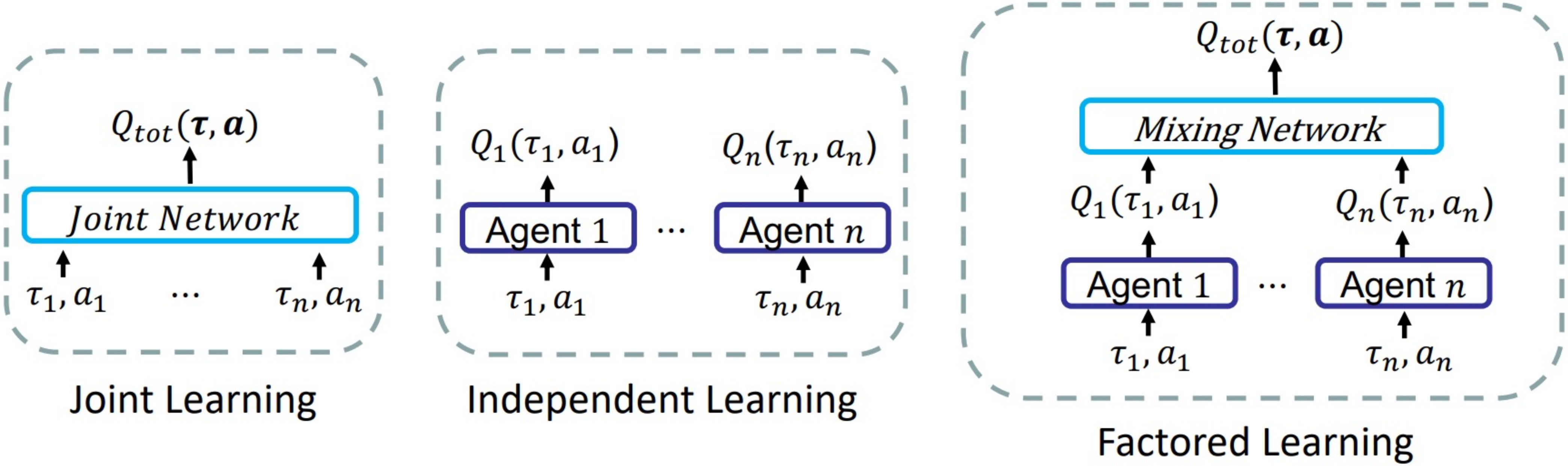


Challenges of Cooperative MARL

- **Scalability**
  - Curse of dimensionality
- **Multi-Agent Credit Assignment**
  - Each agent’s contribution to the team
- **Learning Efficiency**
  - Requiring extensive interactions
- **Limited Observability**
  - Sensory constraints
- **Exploration**
  - An exponential joint policy space



MARL Paradigms



- Optimality

- Scalability

- Scalability

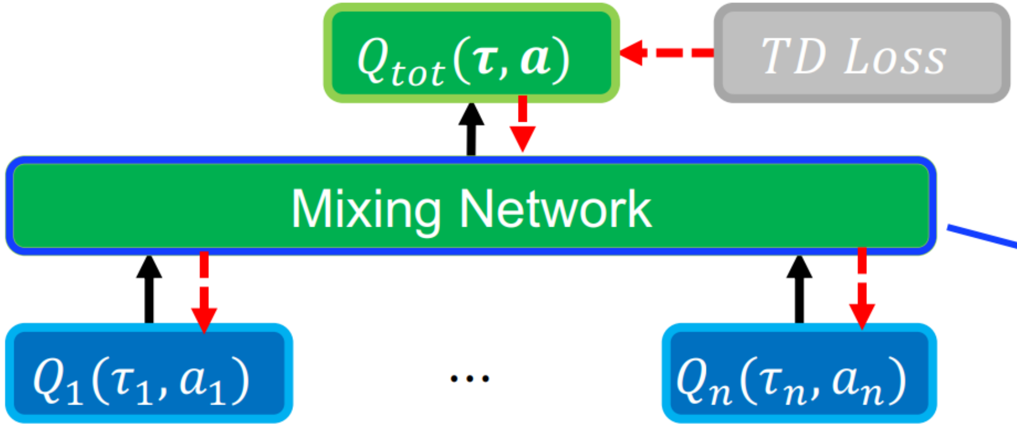
- Suboptimality
- Non-stationarity
- Credit assignment

- Scalable centralized training
- Decentralized execution
- Implicit credit assignment
- Optimality & Convergence

- Centralized training

Factored Value Functions for MARL

- Scalable centralized training with decentralized execution



$$L(\theta) = \mathbb{E}[(r + \gamma V(\boldsymbol{\tau}'; \theta^-) - Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}; \theta))^2]$$

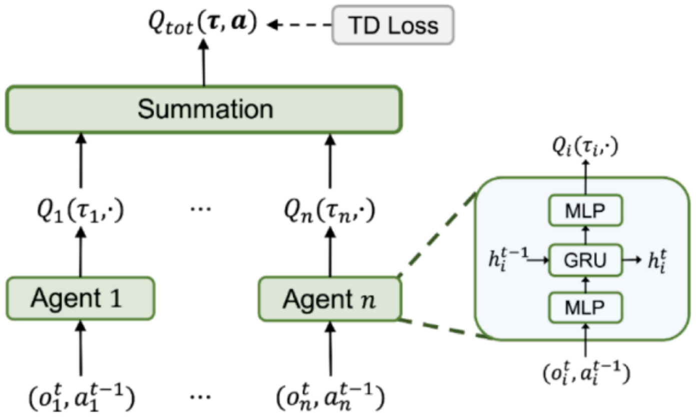
$$\boxed{V(\boldsymbol{\tau}'; \theta^-) = \max_{\boldsymbol{a}'} Q_{tot}(\boldsymbol{\tau}', \boldsymbol{a}'; \theta^-)}$$

$$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = f(Q_1(\tau_1, a_1), \ldots, Q_n(\tau_n, a_n))$$

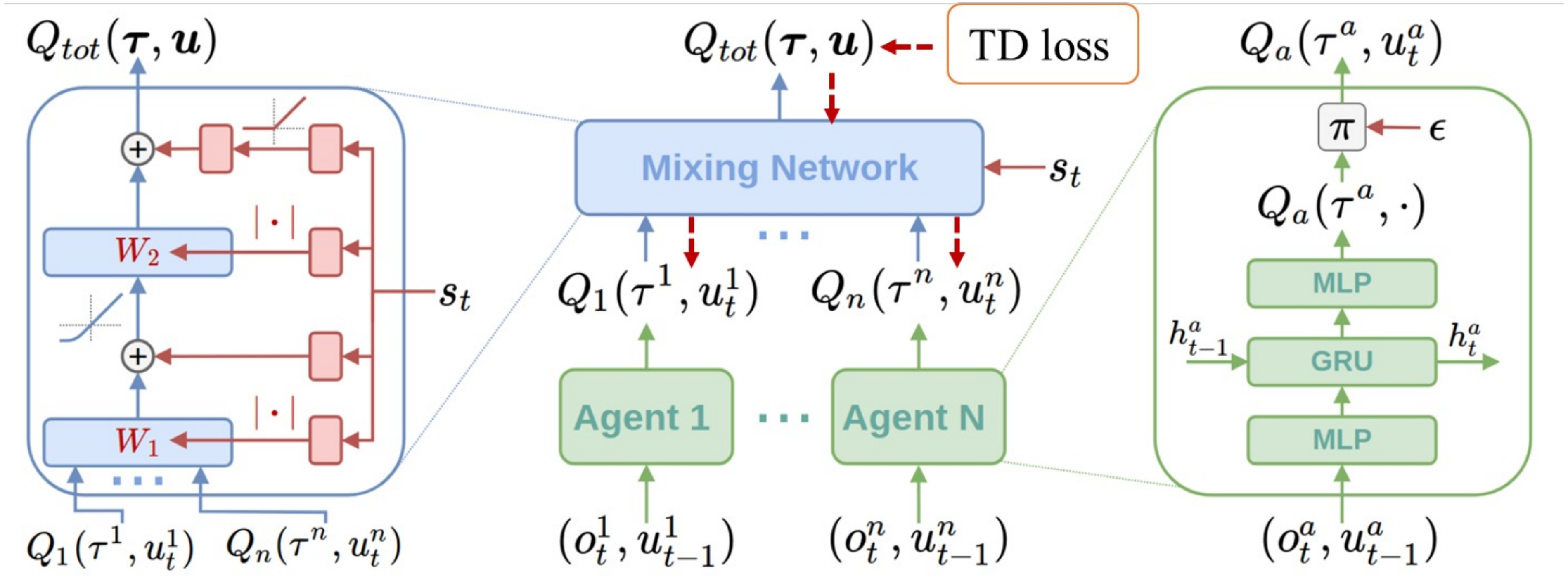
- Individual-Global Maximization (IGM) Constraint
  - $\operatorname{argmax}_{\boldsymbol{a}} Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \left(\operatorname{argmax}_{a_1} Q_1(\tau_1, a_1), \ldots, \operatorname{argmax}_{a_n} Q_n(\tau_n, a_n)\right)$
  - Consistent greedy action selection between joint and individuals

Factored Value Functions for MARL

- Linear Mixing: $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_i Q_i(\tau_i, a_i)$ [Sunehag et. al., 2017]
- Satisfying IGM Constraint
  - $\underset{\boldsymbol{a}}{\operatorname{argmax}}\, Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \begin{pmatrix} \operatorname{argmax}_{a_1} Q_1(\tau_1, a_1) \\ \cdots \\ \operatorname{argmax}_{a_n} Q_n(\tau_n, a_n) \end{pmatrix}$
- No parameters in the mixing network
- No specific reward for each agent
- Implicit credit assignment through gradient backpropagation



Centralized Training and Decentralized Execution



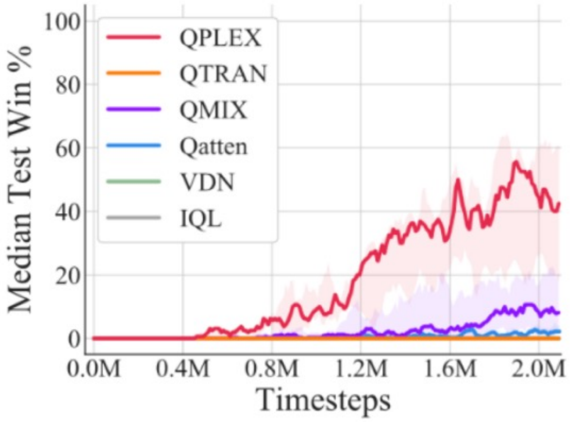
- **Individual-Global Maximiztion (IGM) Constraint**
  - $argmax_{\boldsymbol{a}}\, Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = (argmax_{a_1}\, Q_1(\tau_1, a_1), \ldots, argmax_{a_n}\, Q_n(\tau_n, a_n))$
  - Consistent greedy action selection between joint and individuals

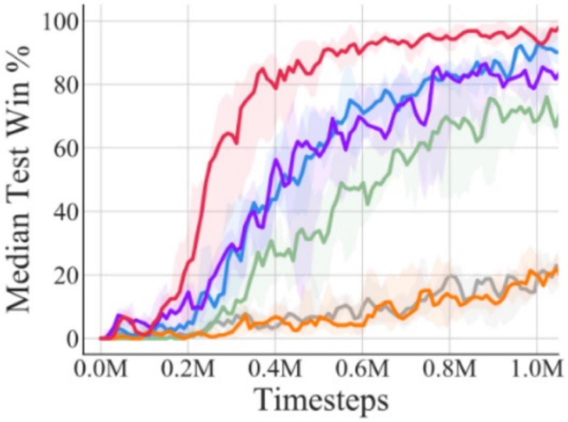
Environment

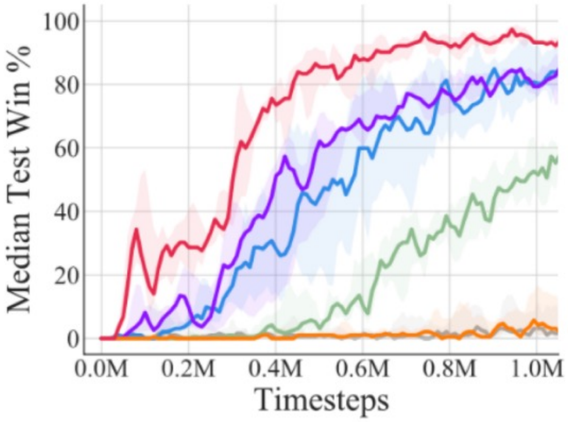
StarCraft Multi-Agent Challenge (SMAC)

- GRF
- MA-Mujoco
- MOBA

Online Learning Performance







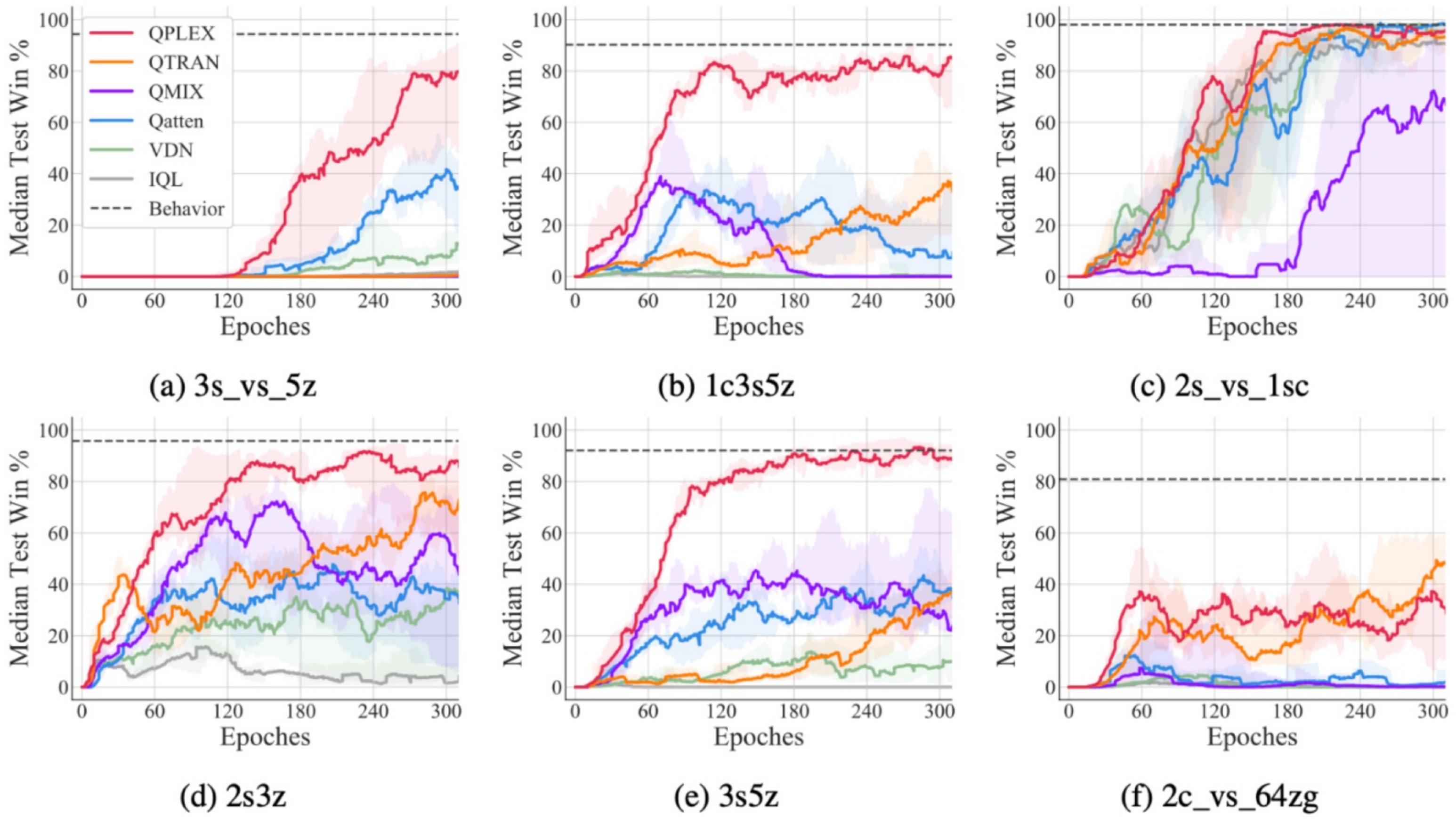
(a) 5s10z

(b) 1c3s5z

(c) 3s5z

Data-Driven Offline MARL Learning

Data collected by a behavior policy learned by QMIX



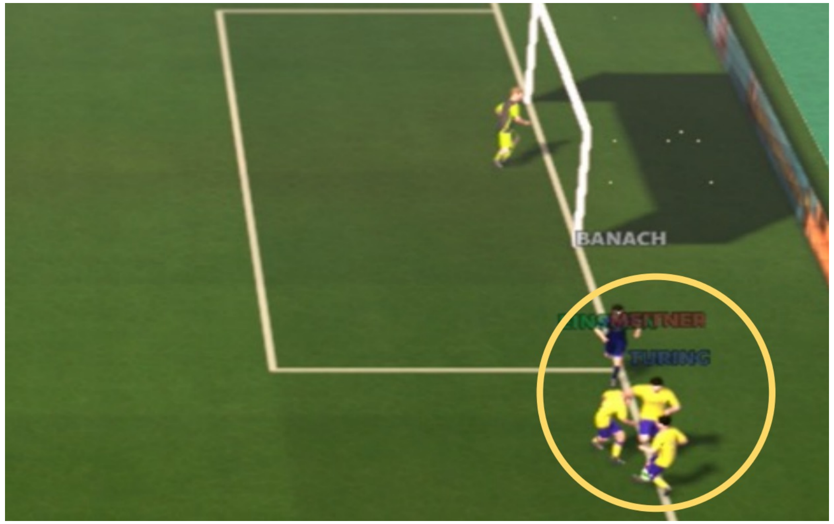
MARL with Factored Policies

- COMA: Counterfactual Multi-Agent Policy Gradients [Foerster et. al, 2017]
- MADDPG: Multi-agent actor-critic for mixed cooperative-competitive environments [Lowe et. al, 2017]
- MAPPO: Multi-Agent PPO [Yu et. al, 2021]
- HATRPO/HAPPO: Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning [Kuba et. al, 2021]
- T-PPO: Towards Global Optimality in Cooperative MARL with Sequential Transformation [Ye et. al, 2021]

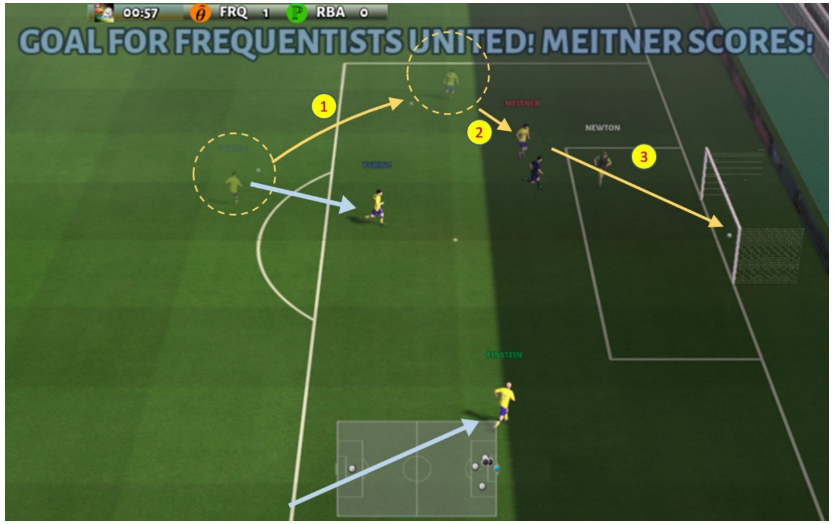
Contents

Limitation: Shared learning among agents

- **Parameter sharing is critical for deep MARL methods**
- **However, agents tend to acquire homogeneous behaviors**
- **Dynamic sharing with diversity is essential for practical tasks**



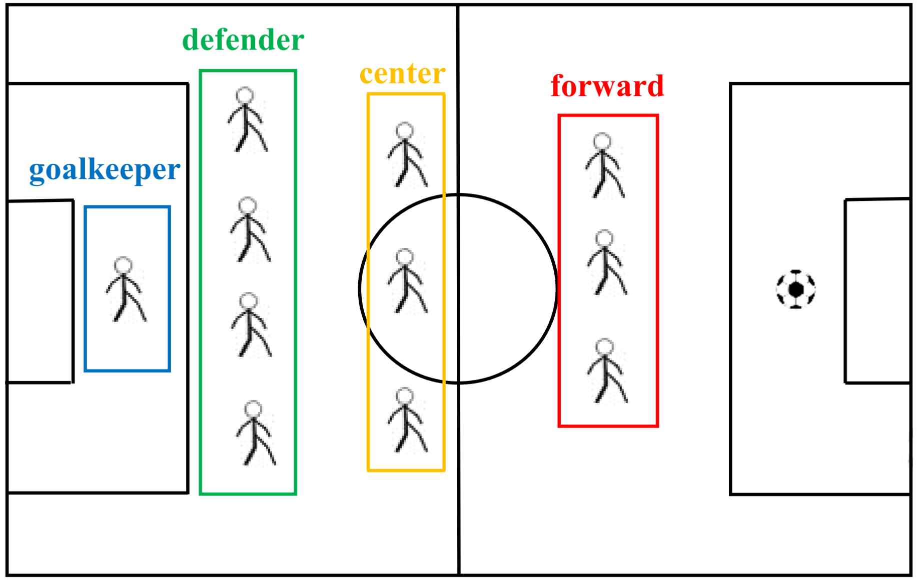
Similar behaviors (competing for ball)



Each agent has its responsibility to score

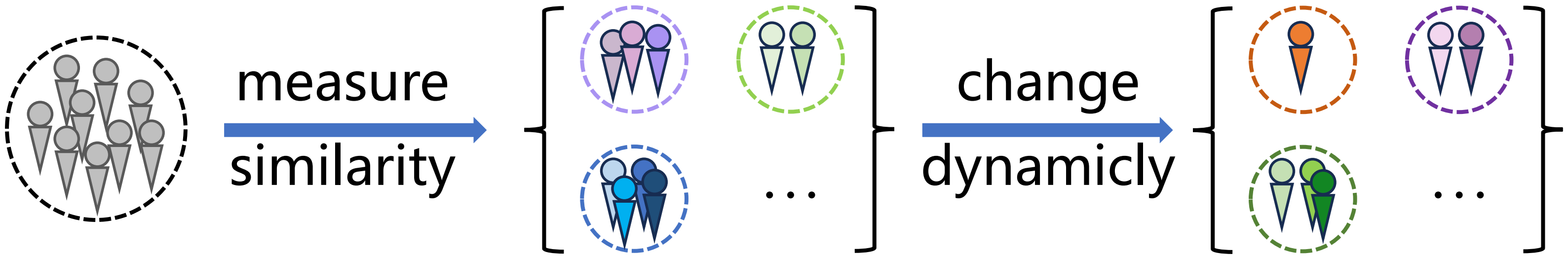
MARL with Representation Learning

- **Agents with similar role have similar policies and share their learning**
  - Roles ⟷ Subtasks ⟷ Skills
  - An example of subtask assignment in football: forward, center, defender, goalkeeper



- **Benefits of task decomposition and Role (Subtask or Skill) assignment**
  - Agent can change its roles in different situations ➡ Tackles agent homogenization
  - Agent learn policy conditioned on their roles ➡ Facilitates efficient knowledge transfer

Challenge of Role Representation Learning

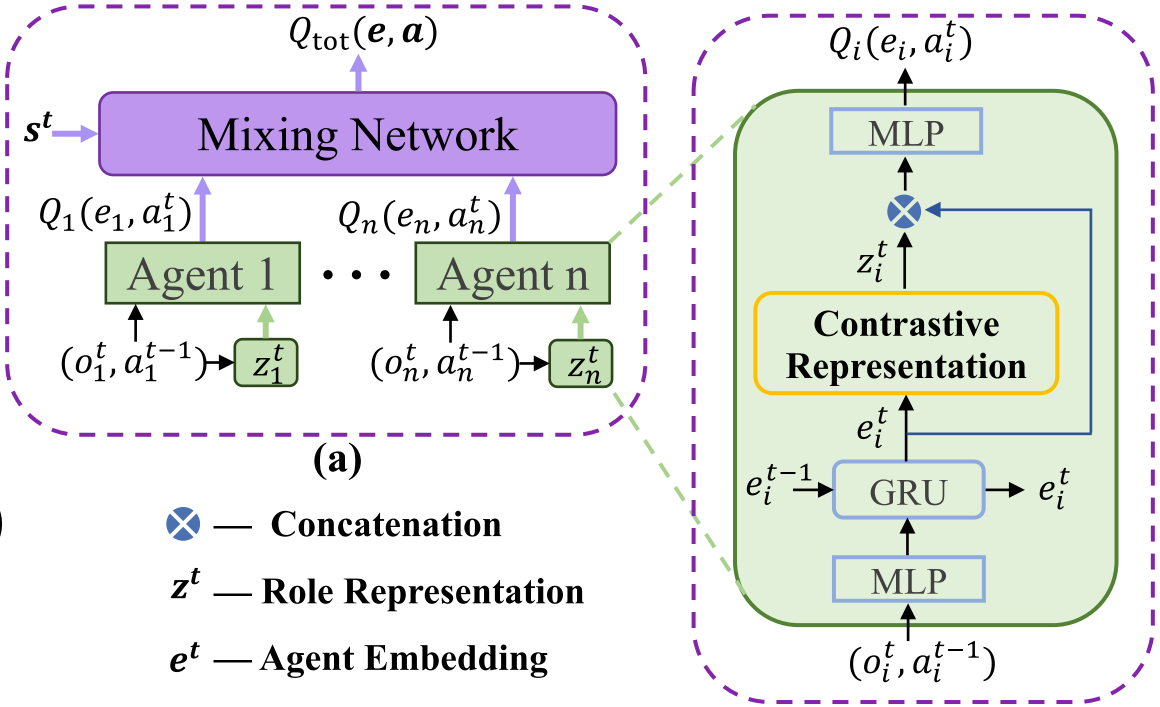


- ☒ **How to measure the similarity of the agents?**
- ☒ **How to define role representation?**
- ☒ **How to achieve the knowledge transfer?**
- ☒ **How to change the role dynamicly?**

Our method: ACORM

(i) Contrastive Role Representation

- **Learn agent embedding**
  - Extract complex agent behaviors from trajectory as $e_i^t = f_\Phi(o_i^t, a_i^{t-1}, e_i^{t-1})$
- **Learn role representation**
  - Reinforce role representation($z^t \sim f_\theta(z^t | e^t)$) through contrastive learning



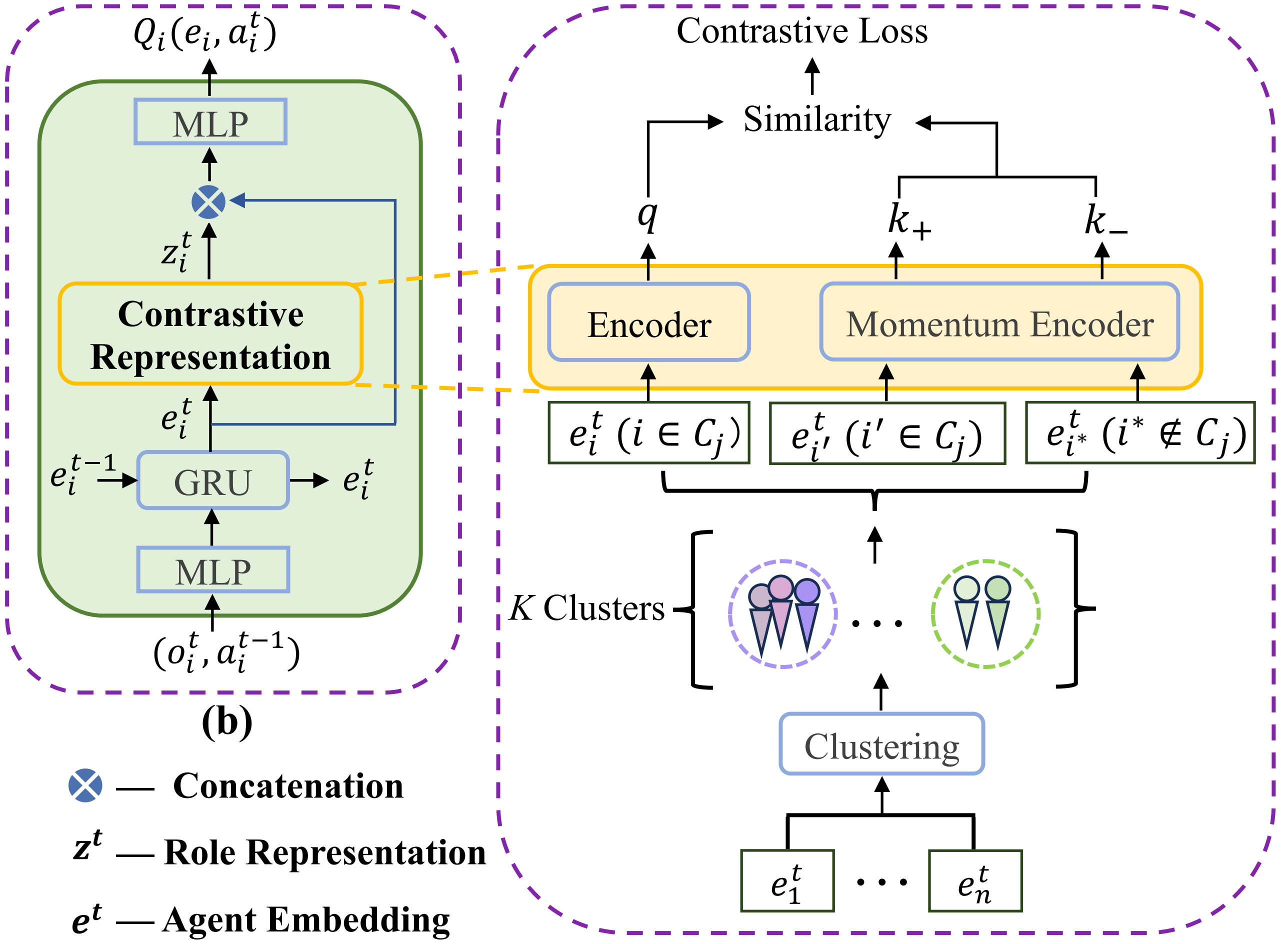
i) Contrastive Role Representation

- **Negative pairs generation**
  - Cluster the agent embedding
  - the same cluster set as positive keys
  - The different clusters set as negative
- **Calculate contrative loss**
  - InfoNCE loss is rearranged as

$$\mathcal{L}_K = -\log \frac{\exp(q^\top W k_+)}{\exp(q^\top W k_+) + \exp(q^\top W k_-)}$$

- **Update momentum encoder**

$$\theta_k \leftarrow \beta\theta_k + (1-\beta)\theta_q$$



ii) Attention-Guided Role Coordination

■ Attention mechanism to enhance coordination

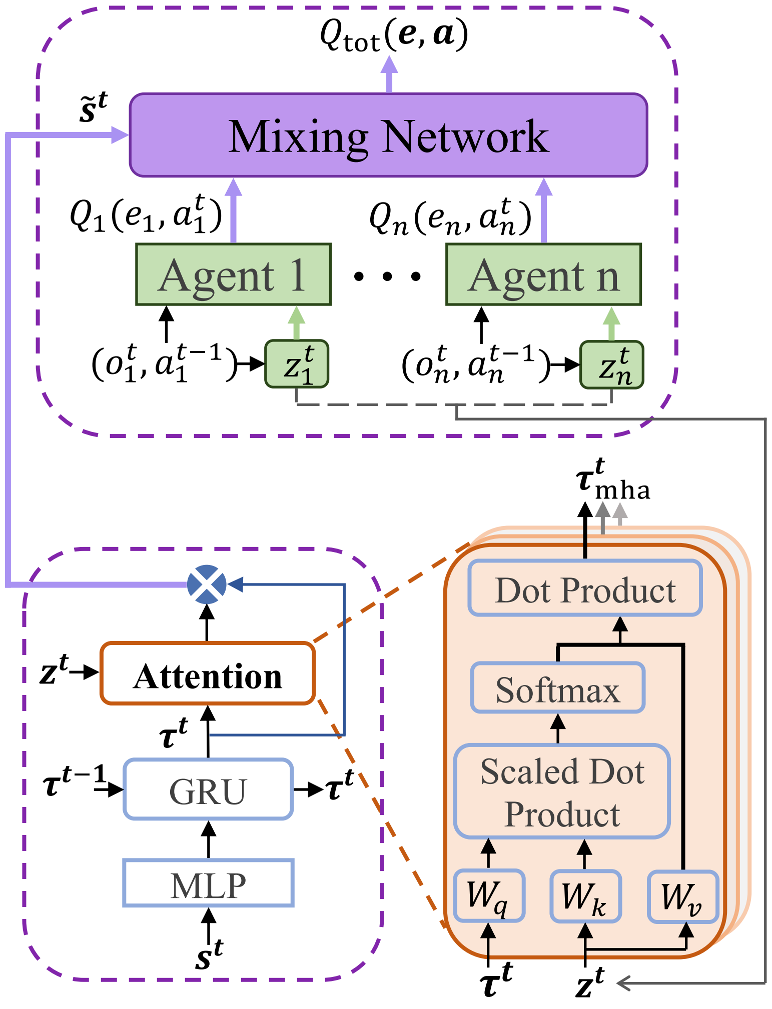
- Set state embedding as the query, role representation as the key and value
- calculate a weighted combination of role representations as:

$$\boldsymbol{\tau}_{\text{atten}} = \sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} \alpha_i \cdot z_i W^V$$

- The attention weight α_i computes as:

$$\alpha_i = \frac{\exp\left(\frac{1}{\sqrt{d_k}} \cdot \boldsymbol{\tau} W^Q \cdot (z_i W^K)^\top\right)}{\sum_{j=1}^{n} \exp\left(\frac{1}{\sqrt{d_k}} \cdot \boldsymbol{\tau} W^Q \cdot (z_j W^K)^\top\right)}$$

- Obtain the aggregated output as: $\boldsymbol{\tau}_{\text{mha}} = (\boldsymbol{\tau}_{\text{atten}}^1, \ldots, \boldsymbol{\tau}_{\text{atten}}^H) W^O$



Attention-guided **CO**ntrastive Role Representations for **M**ARL (**ACORM**)

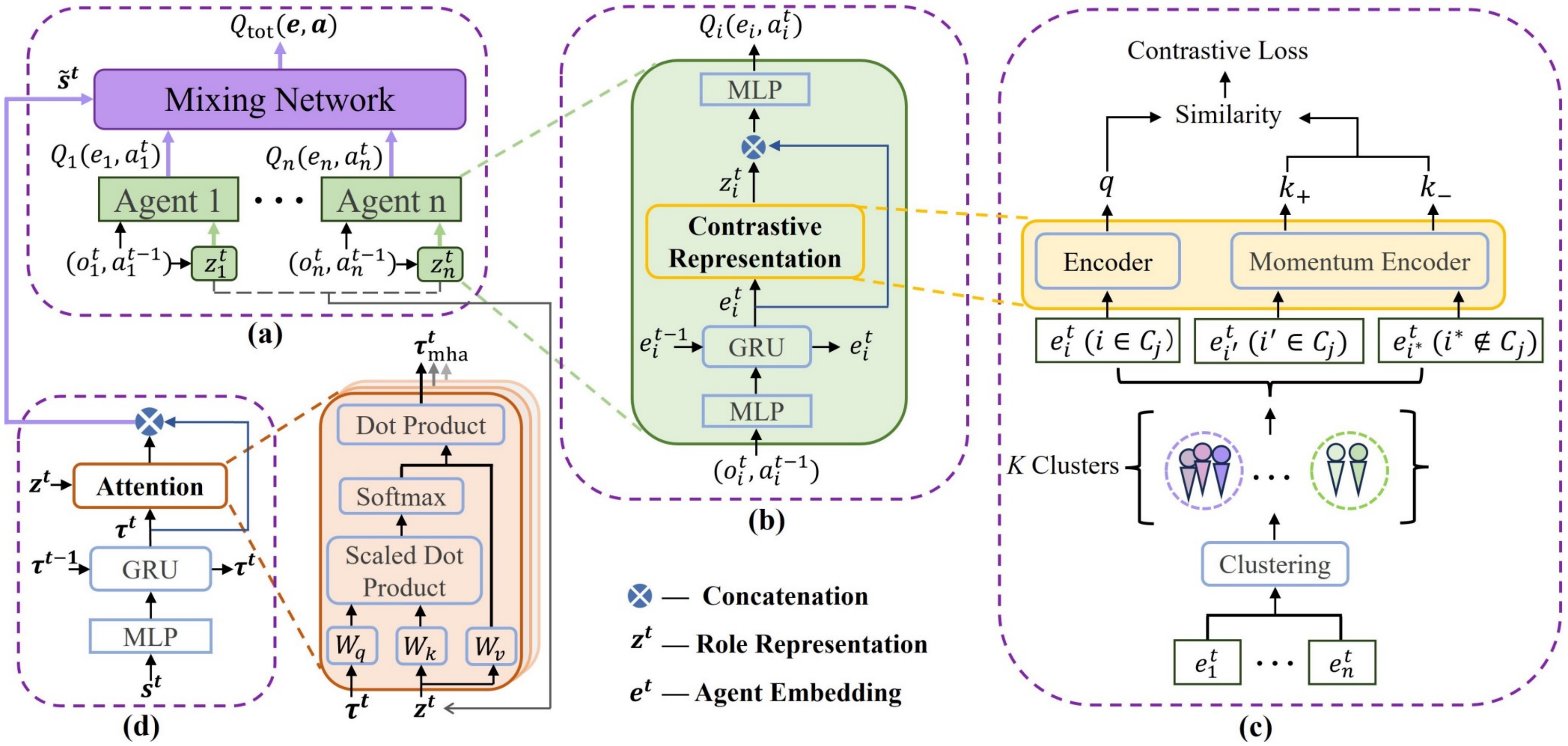
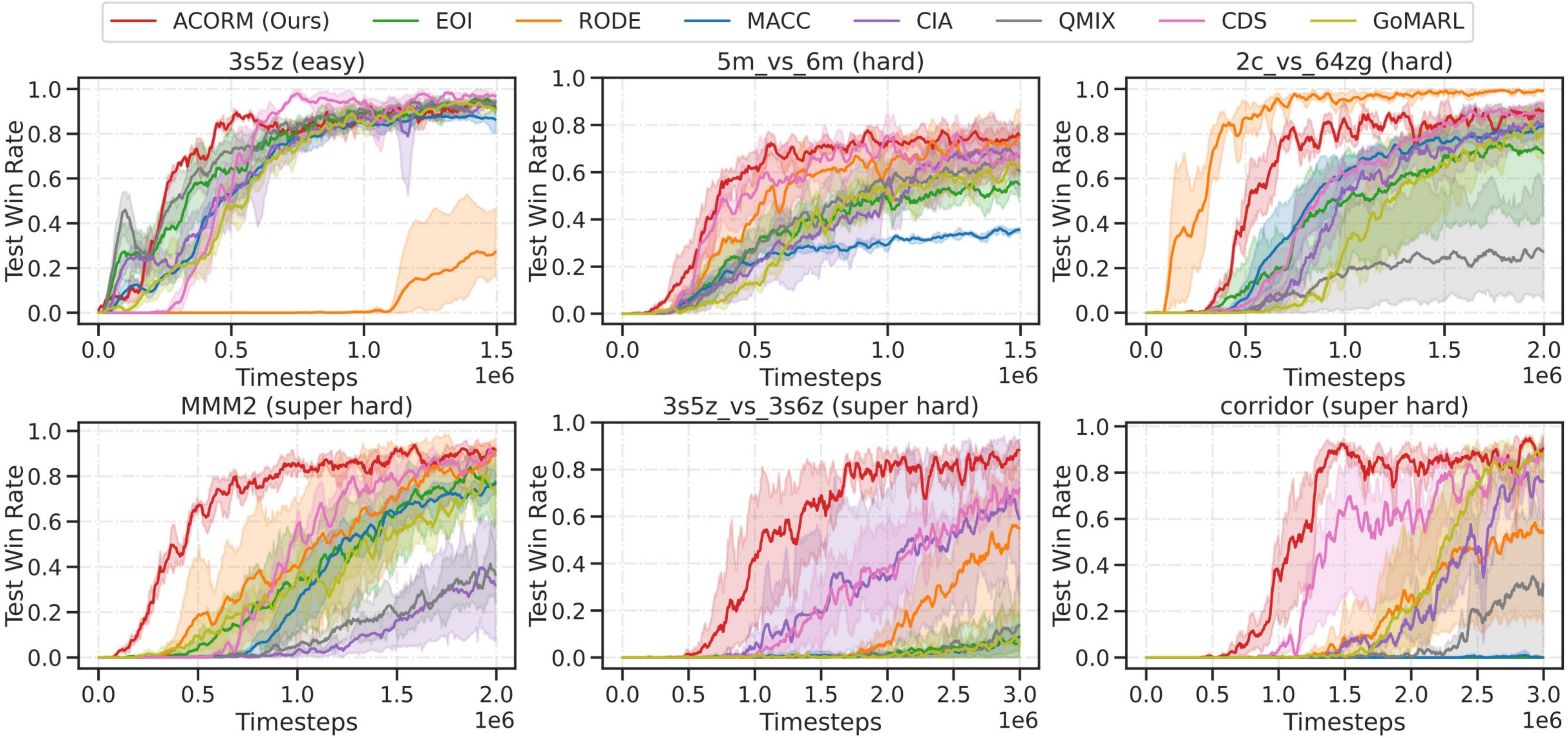


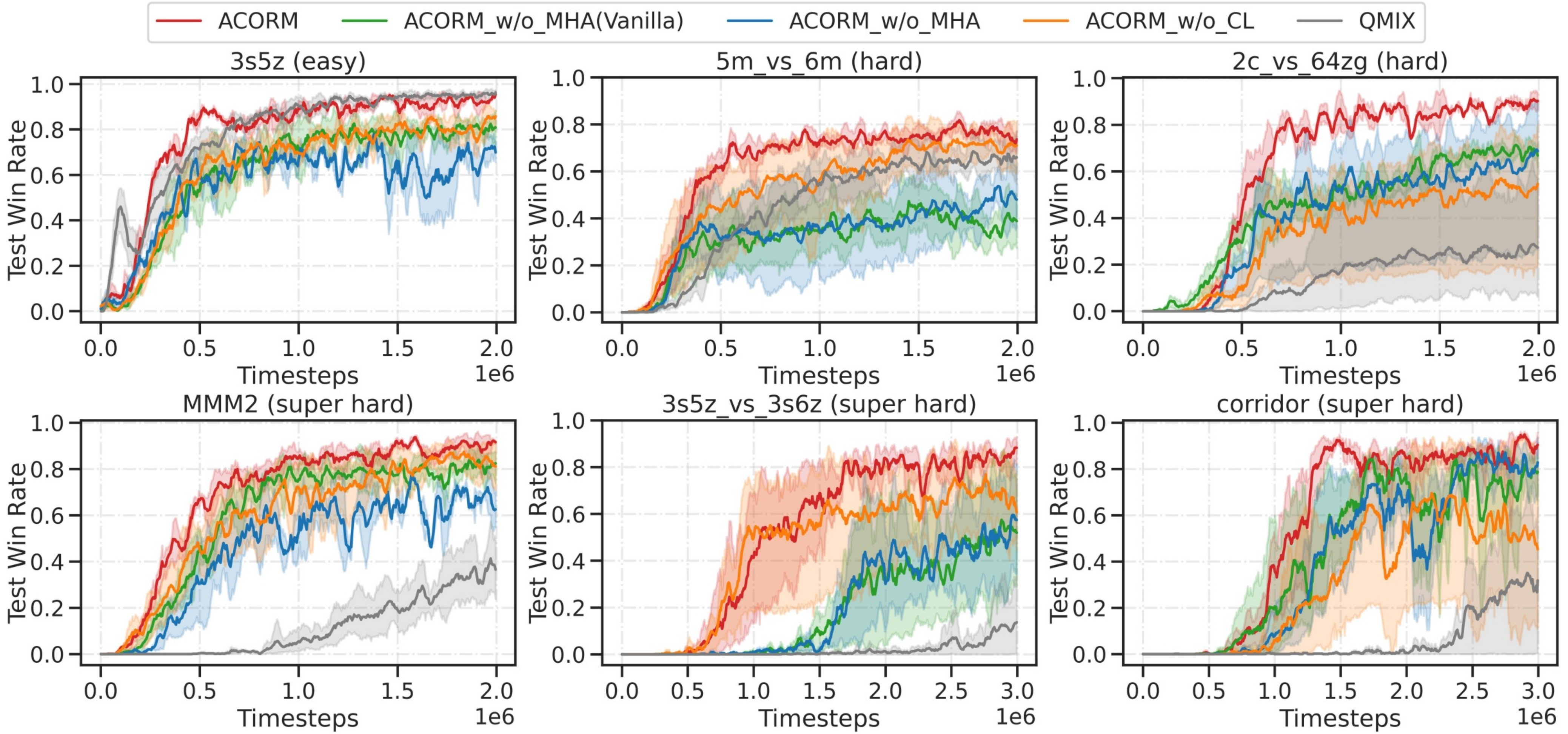
Figure 1: The ACORM framework based on QMIX. (a) The overall architecture. (b) The structure of shared individual Q-network. (c) The detail of contrastive role representation learning, where z_i is the query q, and $z_{i'}/z_{i^*}$ are positive/negative keys k_+/k_-. (d) The attention module that incorporates learned role representations into the mixing network's input for better value decomposition.

- Zican Hu, Zhi Wang*, et al., Attention-guided contrastive role representations for MARL, ICLR 2024.

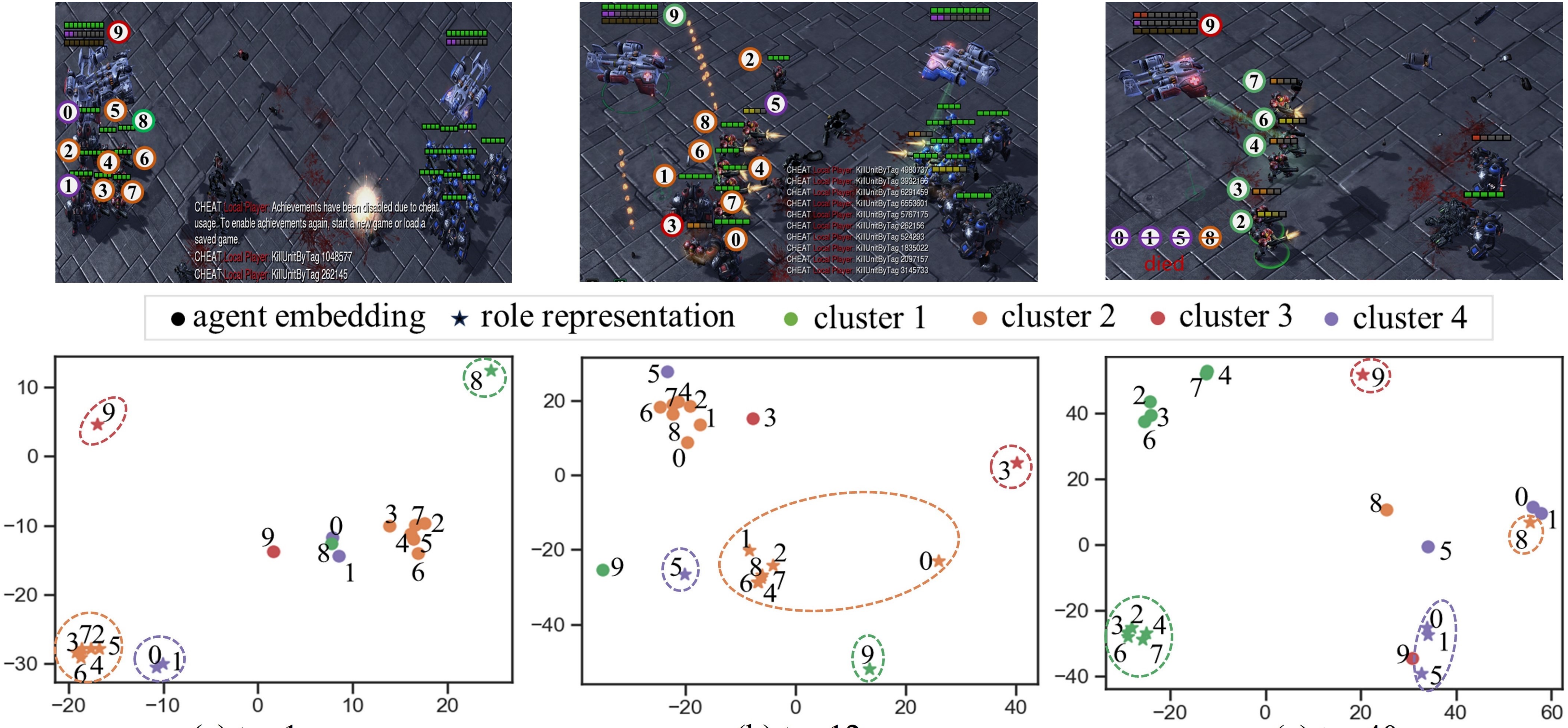
Performance on SMAC



Ablation study



Visualize role representations



(a) t = 1 (b) t = 12 (c) t = 40

Visualize attention mechanism

(a) t = 4

(b) t = 10

(c) t = 36

Conclusions

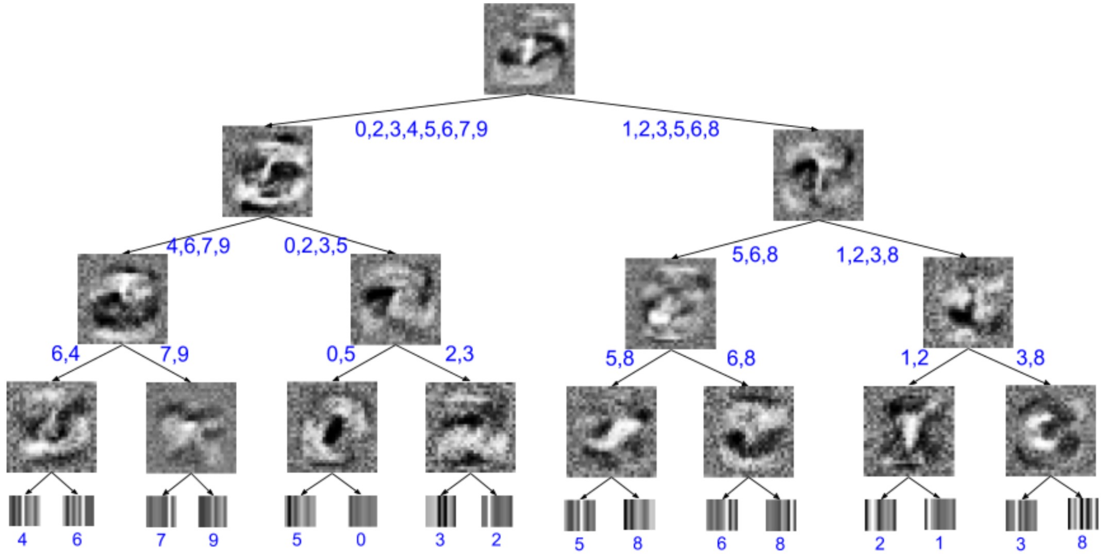
- A general role representation learning framework (分工)
- Leverage role representations to realize more expressive credit assignment (协作)
- Tackle agent homogenization and facilitate efficient knowledge transfer

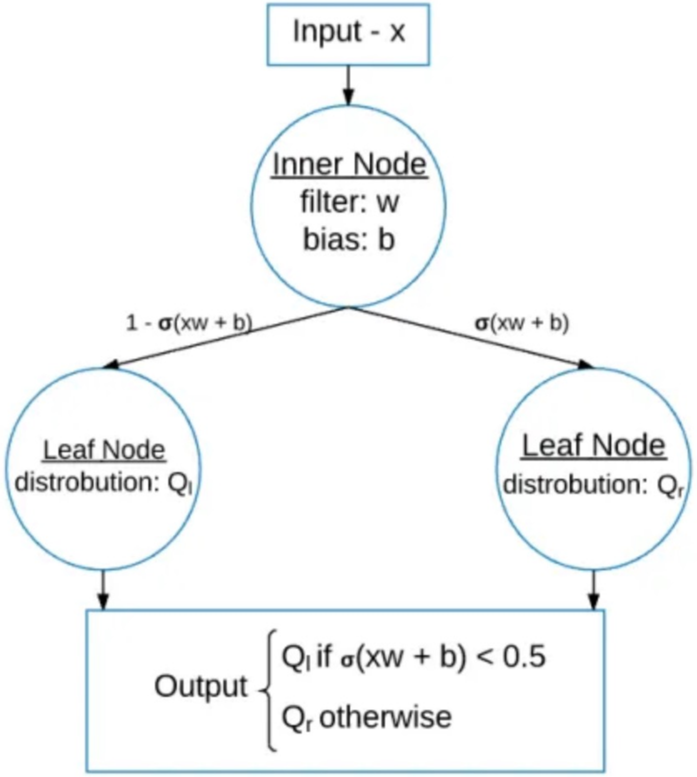
Paper: https://openreview.net/forum?id=LWmuPfEYhH Code: https://github.com/NJU-RL/ACORM

Contents

Soft Decision Trees

- Decision trees have long been valued for their simplicity and interpretability
  - mimic human decision-making processes by splitting data into branches at binary decision points, making them intuitive to understand and explain
- The term “soft” decision trees extends this concept further
  - incorporate elements of neural networks to enhance flexibility and performance





A soft binary decision tree with a single inner node and two leaf nodes

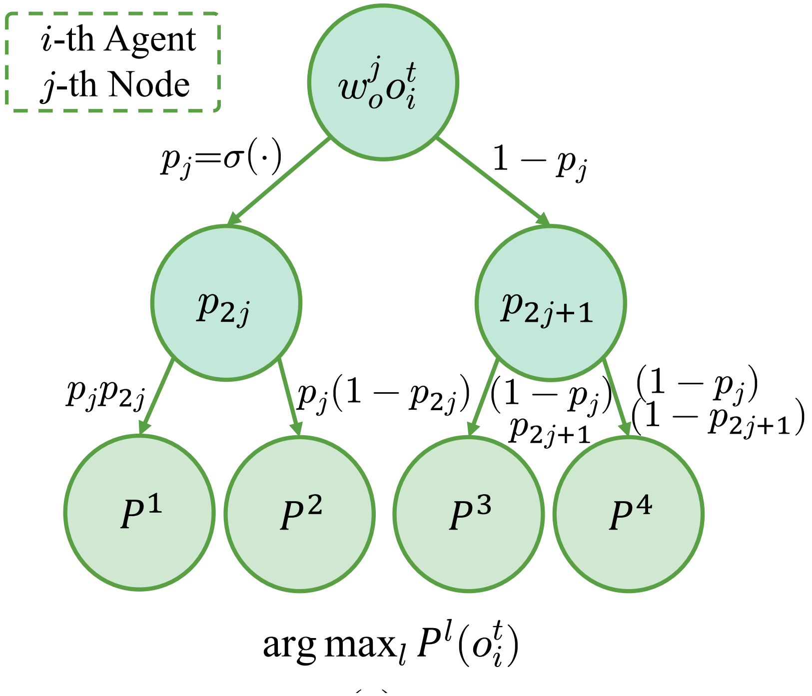
- Hinton et al., Distilling a Neural Network Into a Soft Decision Tree, 2017.

Soft Decision Trees for MARL

■ Motivation

➢ Soft decision trees for the Q-function

- Differentiable structure, soft decision boundaries
- Good representation ability
- Good interpretability for decision problems



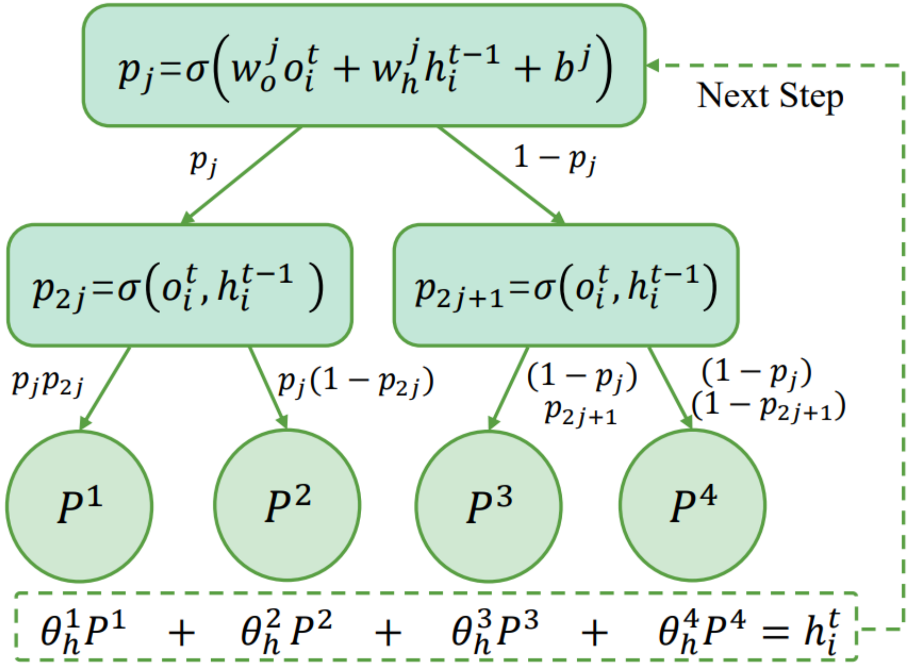
Recurrent Soft Decision Trees

■ Key insight

- Incorporate history information akin to RNNs

$$h_i^t = \sum_{l \in LeafNodes} P^l(o_i^t, h_i^{t-1})\,\theta_h^l$$

$$Q_i(\tau_i, \cdot) = h_i^t w_q$$



Ensemble Recurrent Soft Decision Trees

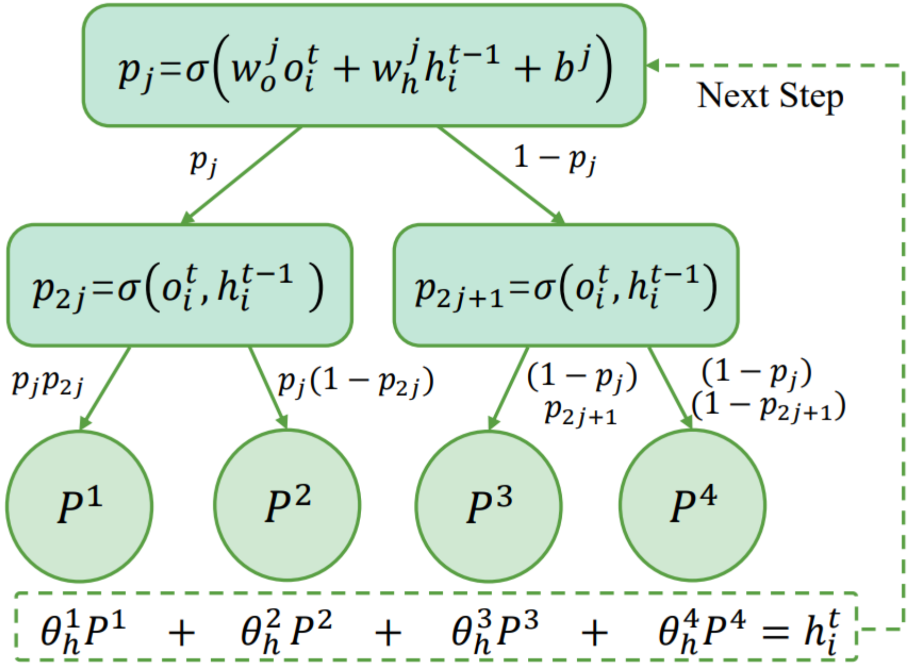
Key insight

- Increase representation power
- Ensure interpretability
- Linear ensembling

$$h_i^t = \sum_{l \in LeafNodes} P^l(o_i^t, h_i^{t-1})\,\theta_h^l \qquad Q_i(\tau_i, \cdot) = h_i^t\, w_q$$

$$h_i^t = \left[h_{i,(1)}^t, h_{i,(2)}^t, \cdots, h_{i,(H)}^t\right]$$

$$Q_i(\tau_i, \cdot) = h_{i,(1)}^t\, w_{q,(1)} + h_{i,(2)}^t\, w_{q,(2)} + \dots + h_{i,(H)}^t\, w_{q,(H)}$$



Mixing Tree Architecture

■ Key insight

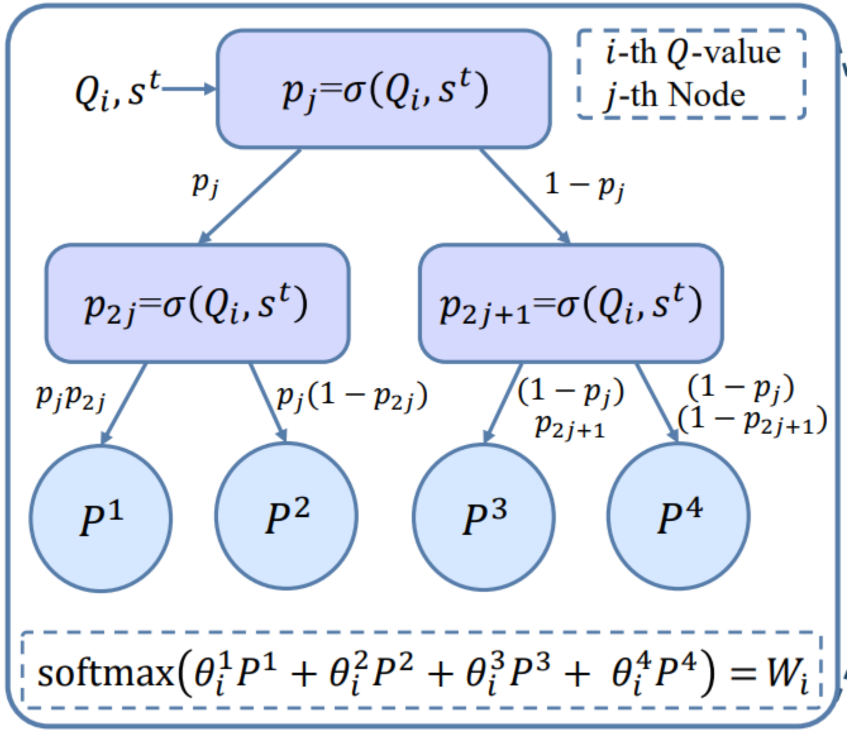
- Value decomposition using soft decision tress

$$p_j(Q_i, s^t) = \sigma(w_q^j Q_i + w_s^j s^t + b^j)$$

$$\phi_i = \sum_{l \in LeafNodes} P^l(Q_i, s^t)\, \theta_i^l,$$

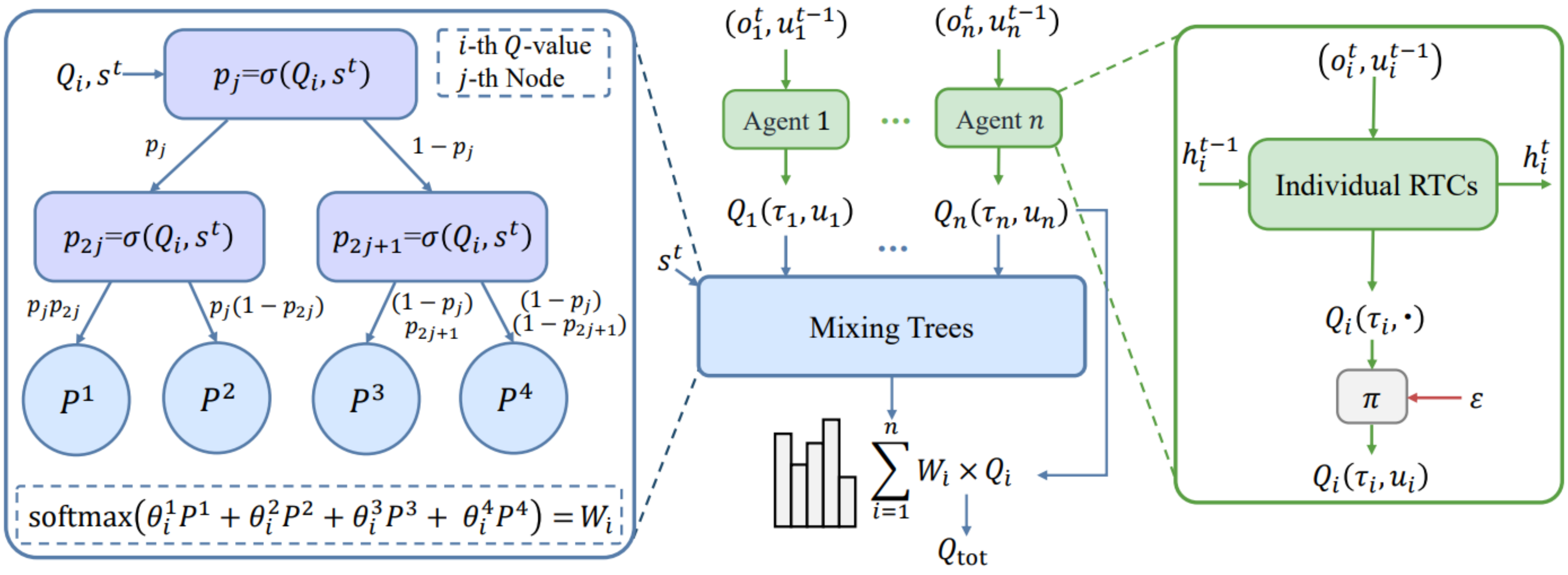
$$W_i = \frac{exp(\sum_{k=1}^{H} \phi_{i,(k)}\, w_{\phi,(k)})}{\sum_{i=1}^{n} exp(\sum_{k=1}^{H} \phi_{i,(k)}\, w_{\phi,(k)})},$$

$$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) \approx \sum_{i=1}^{n} W_i Q_i(\tau_i, u_i)$$



MIXing Recurrent soft decision Trees (MIXRTs)

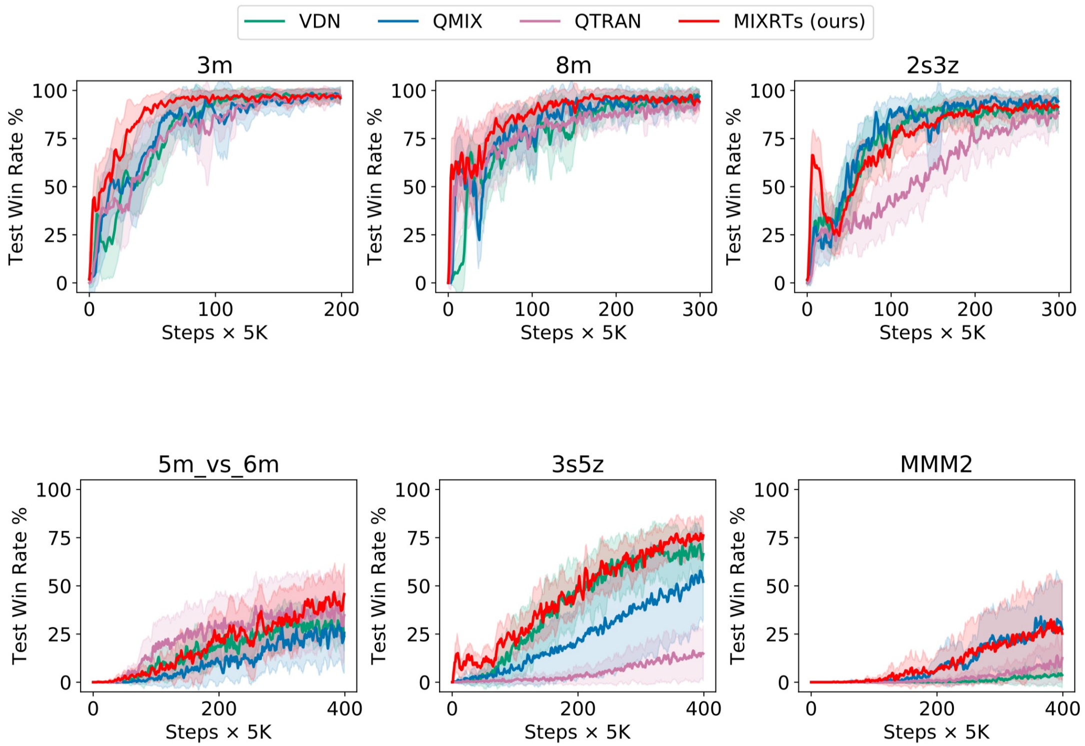
■ Put it together



- Zichuan Liu, Zhi Wang*, et al., MIXRTs: Toward Interpretable Multi-Agent RL via Mixing Recurrent Soft Decision Trees, TPAMI, 2025.

Primary Results: Performance

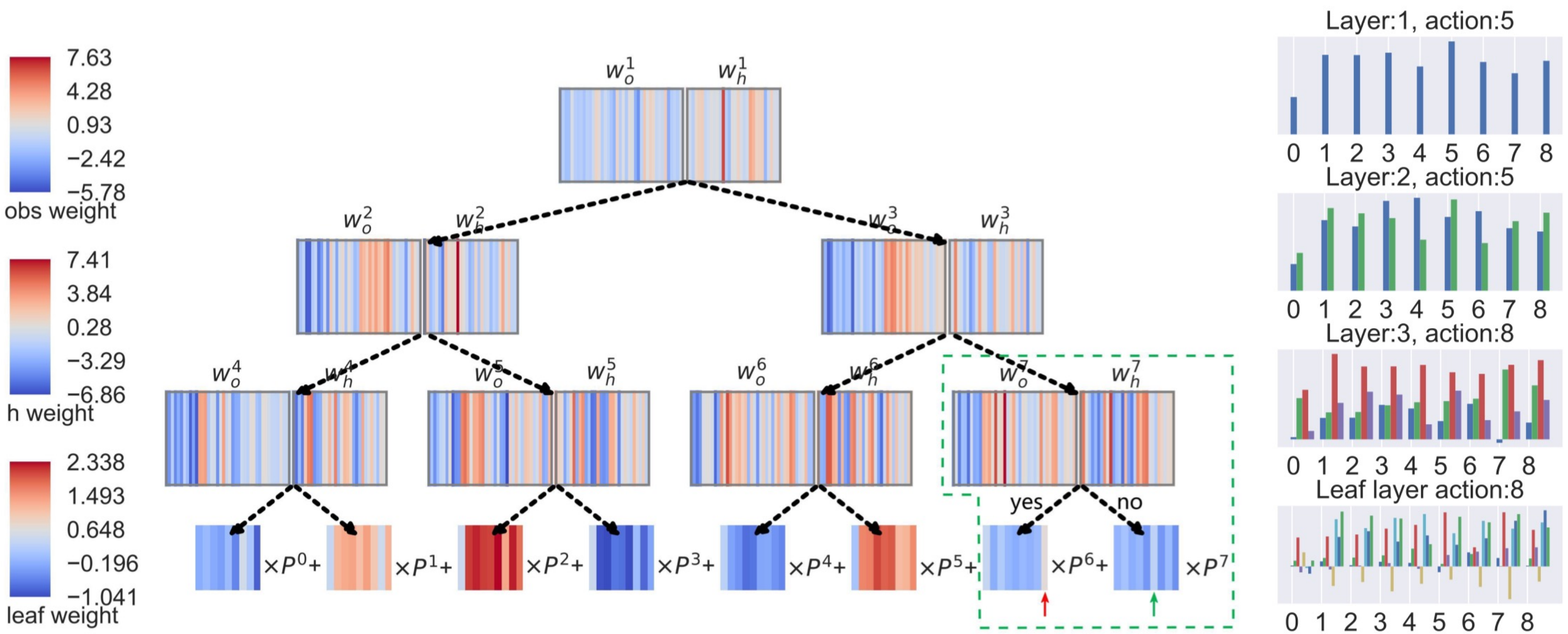
- 在简单场景上的性能
  - 快速掌握简单任务
  - 模型取得较高的胜率
- 在困难/超困难场景上的性能
  - 实现具有竞争力的性能
  - 在学习过程中更加稳定



- 模型参数比较
  - 线性模型、参数较少

Method	3m	2s3z	5m_vs_6m	3s5z	8m_vs_9m	MMM2	6h_vs_8z
VDN	28, 297	31, 883	30, 412	35, 534	32, 911	39, 250	32, 206
QMIX	37, 738	62, 892	55, 789	111, 951	96, 304	173, 651	72, 847
QTRAN	70, 911	84, 437	80, 518	101, 492	94, 513	120, 320	88, 436
MIXRTs (ours)	20, 880	34, 448	28, 752	48, 560	38, 592	62, 736	35, 440

Primary Results: Interpretability



(a) (b)

- 解释软决策树的结构
  - 更红的颜色意味着获得更高的特征权重
  - 在绿色方框中，位置17、14和29处有着更红的颜色，表示敌人是否可见、自身的生命值等特征，易发现智能体更倾向于攻击敌人

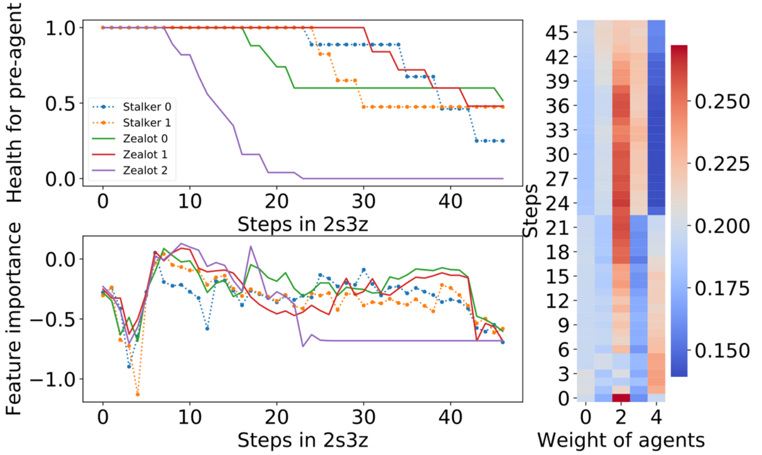
Primary Results: Interpretability

■ 案例分析

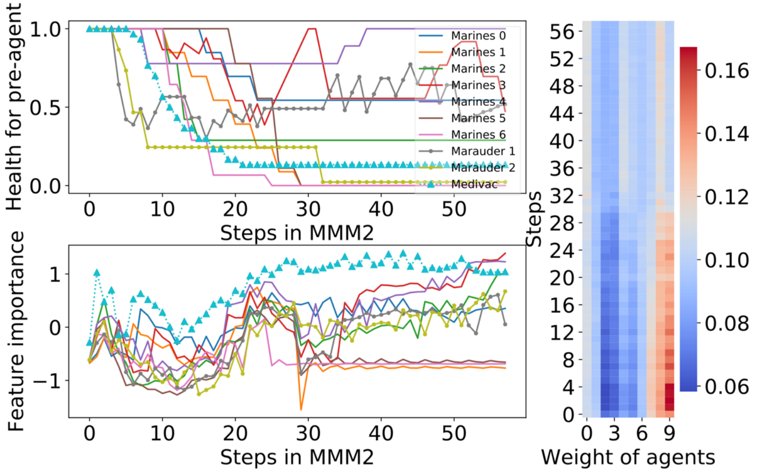
- ➢ 对特征的重要度解释
  - ➢ 特征重要度的定义

$$I(o_i^t) = \sum_j P^j(o_i^t, h_i^{t-1}) w_o^j$$

  - ➢ 特征重要度在某一回合中：以健康属性为例
    - ➢ 相同的智能体有着相同趋势
    - ➢ 死亡的智能体具有较低的重要度
    - ➢ 医疗船在战斗中具有较高的重要度
    - ➢ …….



(a) Map 2s3z



(b) Map MMM2

Thank You.

Zhi Wang (王志)

https://heyuanmingong.github.io

Email: zhiwang@nju.edu.cn

Nanjing University, China

2025-04-27