Dual Self-Awareness Value Decomposition Framework without Individual Global Max for Cooperative MARL

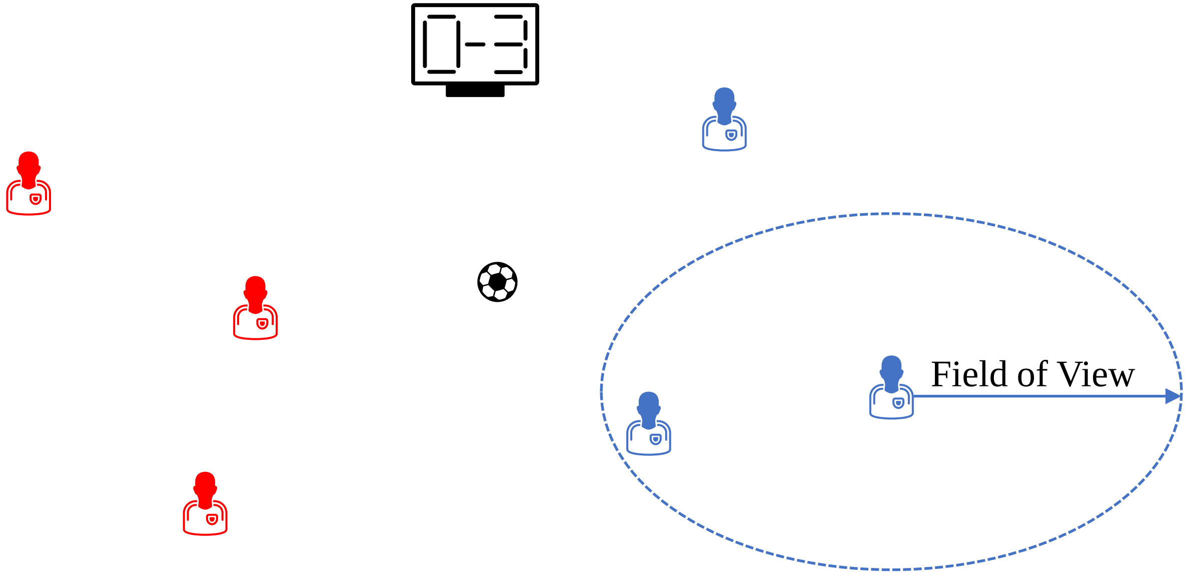
Zhiwei Xu, Bin Zhang, Dapeng Li, Guangchong Zhou, Zeren Zhang, Guoliang Fan

Institute of Automation, Chinese Academy of Sciences

School of Artificial Intelligence, University of Chinese Academy of Sciences

Dec-POMDP

- A Dec-POMDP is a "Decentralized Partially Observable Markov Decision Process"
- It's a bunch of "agents" that are working together for a common reward
- Each agent only takes a local observation of the environment
- A fully cooperative multi-agent task can be described as a tuple $\langle S, U, p, r, Z, O, n, \gamma \rangle$

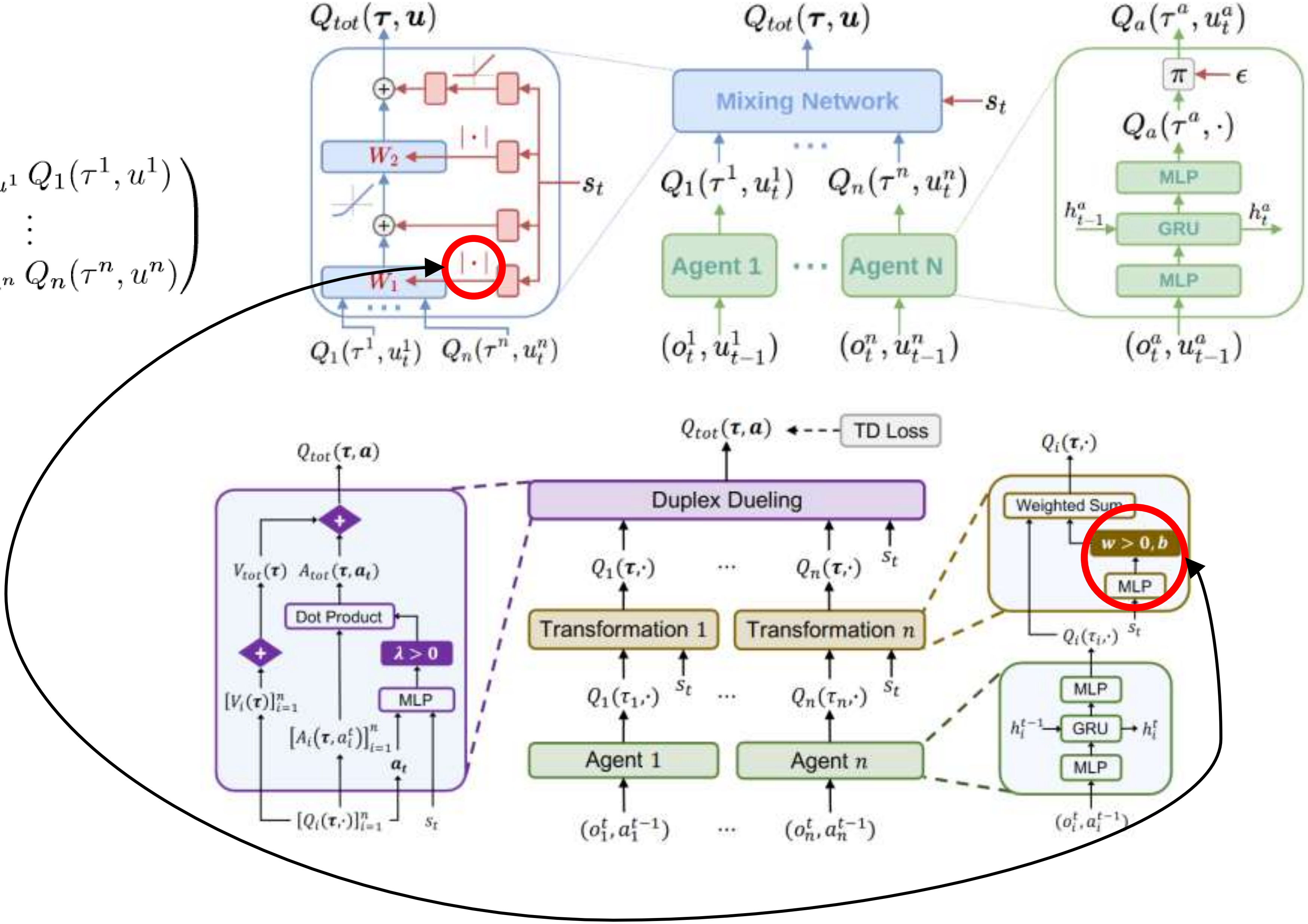


Motivation

Individual Global Max:

$$\underset{\mathbf{u}}{\operatorname{argmax}}\, Q_{tot}(\boldsymbol{\tau},\mathbf{u}) = \begin{pmatrix} \operatorname{argmax}_{u^1} Q_1(\tau^1,u^1) \\ \vdots \\ \operatorname{argmax}_{u^n} Q_n(\tau^n,u^n) \end{pmatrix}$$

Absolute Operator



$Q_{tot}(\boldsymbol{\tau},\boldsymbol{u})$
W_2
$|\cdot|$
W_1
s_t
$Q_1(\tau^1,u_t^1)$ $Q_n(\tau^n,u_t^n)$
$Q_{tot}(\boldsymbol{\tau},\boldsymbol{u})$
Mixing Network
s_t
$Q_1(\tau^1,u_t^1)$ $Q_n(\tau^n,u_t^n)$
Agent 1
Agent N
(o_t^1,u_{t-1}^1) (o_t^n,u_{t-1}^n)
$Q_a(\tau^a,u_t^a)$
π
ϵ
$Q_a(\tau^a,\cdot)$
MLP
h_{t-1}^a
GRU
h_t^a
MLP
(o_t^a,u_{t-1}^a)
$Q_{tot}(\boldsymbol{\tau},\boldsymbol{a})$
TD Loss
$Q_{tot}(\boldsymbol{\tau},\boldsymbol{a})$
$V_{tot}(\boldsymbol{\tau})$ $A_{tot}(\boldsymbol{\tau},\boldsymbol{a}_t)$
Dot Product
$\lambda > 0$
MLP
$[V_i(\boldsymbol{\tau})]_{i=1}^n$
$[A_i(\boldsymbol{\tau},a_i^t)]_{i=1}^n$
$\boldsymbol{a}_t$
$[Q_i(\boldsymbol{\tau},\cdot)]_{i=1}^n$
s_t
Duplex Dueling
$Q_1(\boldsymbol{\tau},\cdot)$ ... $Q_n(\boldsymbol{\tau},\cdot)$ s_t
Transformation 1
Transformation n
$Q_1(\tau_1,\cdot)$ s_t ... $Q_n(\tau_n,\cdot)$ s_t
Agent 1
Agent n
(o_1^t,a_1^{t-1}) ... (o_n^t,a_n^{t-1})
$Q_i(\boldsymbol{\tau},\cdot)$
Weighted Sum
$w > 0, b$
MLP
$Q_i(\tau_i,\cdot)$ s_t
MLP
h_i^{t-1}
GRU
h_i^t
MLP
(o_i^t,a_i^{t-1})

Motivation

Hard to obtain Max Q —— (IGM-Free)

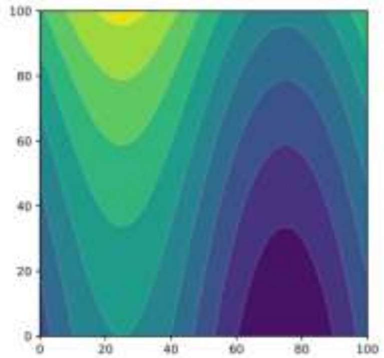
OR ?

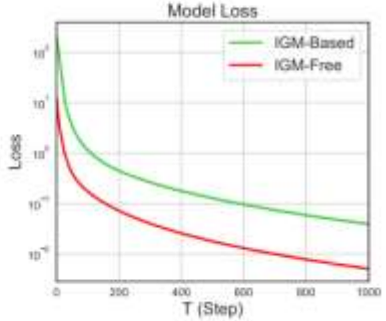
Restrict the set of functions that can be represented
——(IGM-Based)

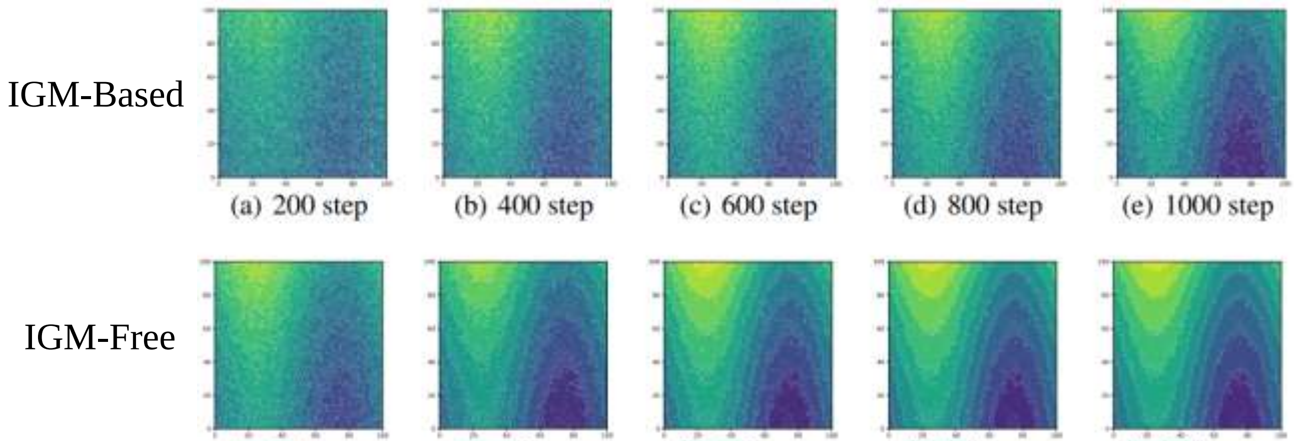
Example

$$Q_i(u^i) = \frac{u^i}{100}, \quad \forall i \in \{1,2\}$$

$$R(u^1,u^2) = \sin\left(2\pi Q_1(u^1)\right) + \exp\left(Q_2(u^2)\right)$$



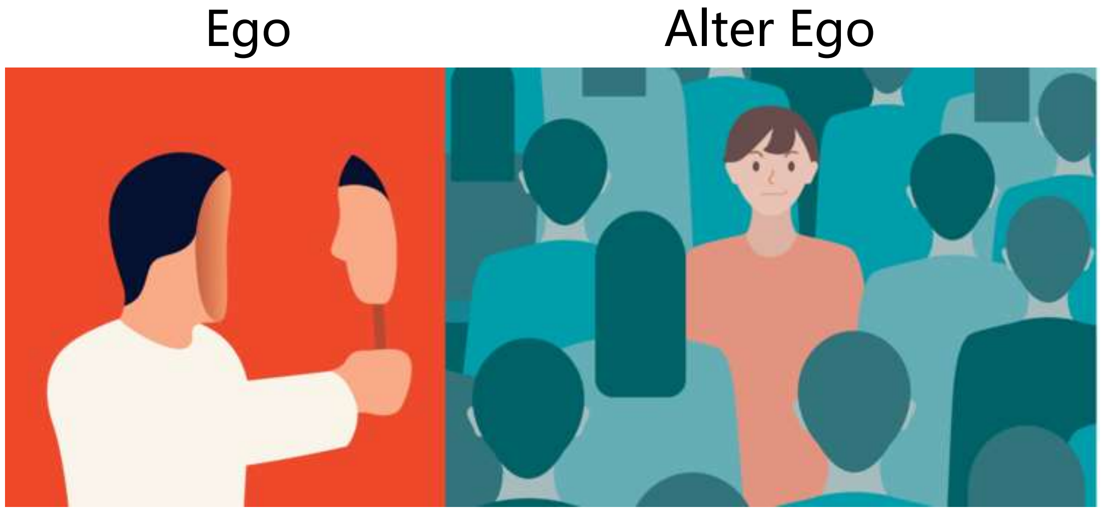


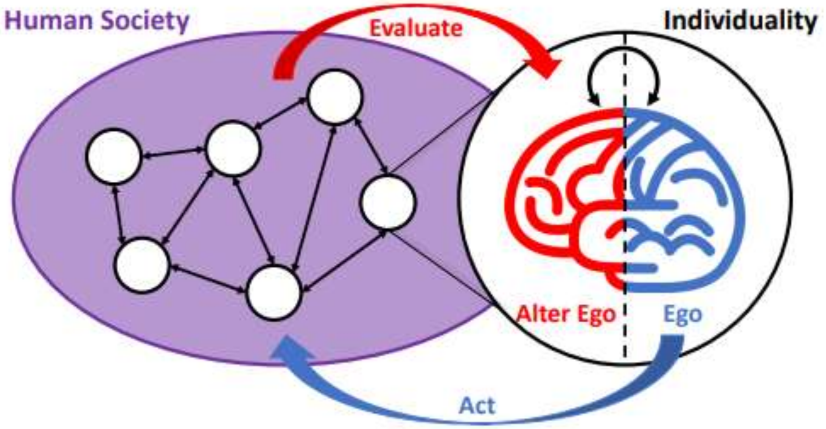


(a) 200 step (b) 400 step (c) 600 step (d) 800 step (e) 1000 step

(f) 200 step (g) 400 step (h) 600 step (i) 800 step (j) 1000 step

Dual Self-Awareness Hypothesis

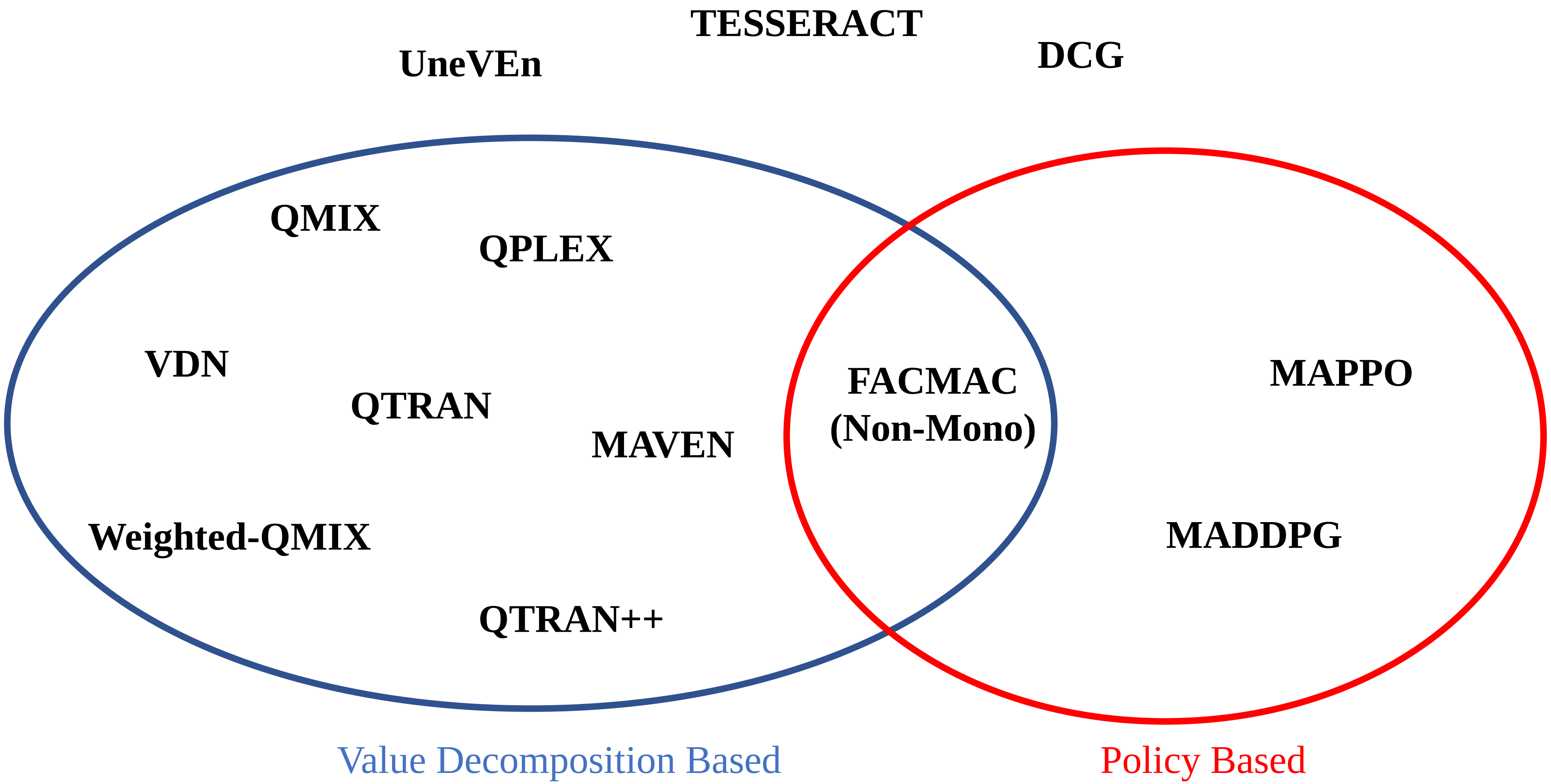




There are concepts of **ego** and **alter ego** in psychology. The ego usually refers to the conscious part of the individual, and Freud considered the ego to be the executive of the personality. Some people believe that an alter ego pertains to a different version of oneself from the authentic self. Others define the alter ego in more detail as the evaluation of the self by others.

Enlightened by these psychological concepts, we propose a novel MARL algorithm, **D**ual self-**A**wareness **V**alue d**E**composition (DAVE).

Related Work



TESSERACT

UneVEn

DCG

QMIX

QPLEX

VDN

QTRAN

MAVEN

Weighted-QMIX

QTRAN++

FACMAC
(Non-Mono)

MAPPO

MADDPG

Value Decomposition Based

Policy Based

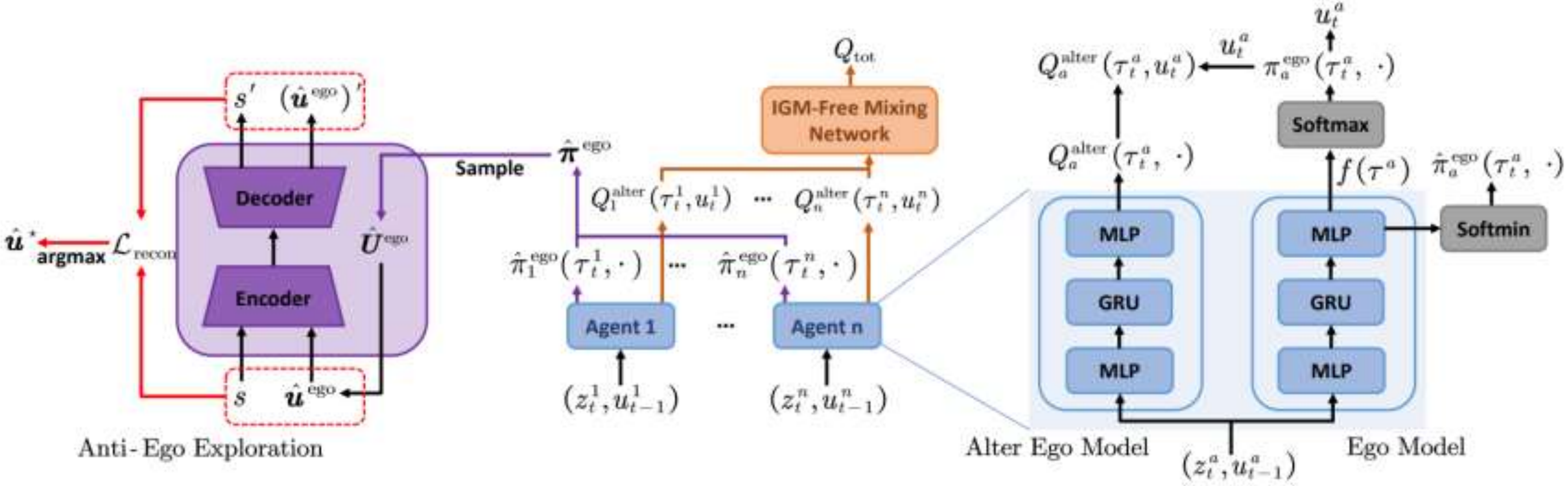
Dual Self-Awareness Framework

The objective of the IGM-free value decomposition method is as follows:

$$\arg\max_{\boldsymbol{\pi}} Q_{\text{tot}}(s, \boldsymbol{u}),$$

Without the IGM assumption, it is **NP-hard** because it cannot be solved and verified in polynomial time. Therefore, in our proposed dual self-awareness framework, each agent has an additional policy network to assist the value function network to find the action corresponding to the optimal joint policy.

Ego Policy Model & **Alter Ego Value Function Model**



s' $(\hat{\boldsymbol{u}}^{\text{ego}})'$
Decoder
Encoder
s $\hat{\boldsymbol{u}}^{\text{ego}}$
$\hat{\boldsymbol{u}}^*$
argmax
$\mathcal{L}_{\text{recon}}$
$\hat{\boldsymbol{U}}^{\text{ego}}$
Sample
$\hat{\boldsymbol{\pi}}^{\text{ego}}$
Anti-Ego Exploration
Q_{tot}
IGM-Free Mixing Network
$Q_1^{\text{alter}}(\tau_t^1, u_t^1)$ ··· $Q_n^{\text{alter}}(\tau_t^n, u_t^n)$
$\hat{\pi}_1^{\text{ego}}(\tau_t^1, \cdot)$ ··· $\hat{\pi}_n^{\text{ego}}(\tau_t^n, \cdot)$
Agent 1
···
Agent n
(z_t^1, u_{t-1}^1)
(z_t^n, u_{t-1}^n)
u_t^a
$Q_a^{\text{alter}}(\tau_t^a, u_t^a)$
u_t^a
$\pi_a^{\text{ego}}(\tau_t^a, \cdot)$
Softmax
$Q_a^{\text{alter}}(\tau_t^a, \cdot)$
$f(\tau^a)$
$\hat{\pi}_a^{\text{ego}}(\tau_t^a, \cdot)$
Softmin
MLP
GRU
MLP
MLP
GRU
MLP
Alter Ego Model
Ego Model
(z_t^a, u_{t-1}^a)

Dual Self-Awareness Framework

So we modify the objective of the cooperative multi-agent value decomposition problem from previous equation to

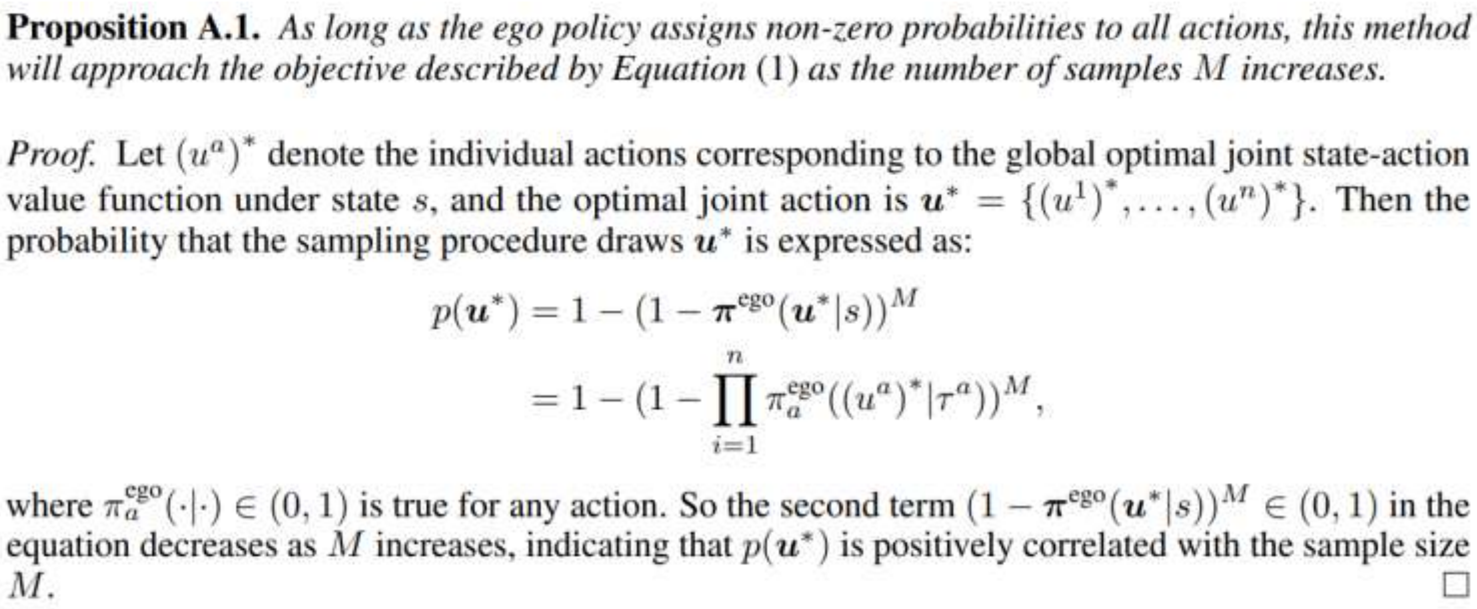
$$\arg\max_{\boldsymbol{\pi}^{\text{ego}}} Q_{\text{tot}}^{\text{alter}}(s, \boldsymbol{u}^{\text{ego}}), \qquad \text{s.t.} \quad \boldsymbol{u}^{\text{ego}} \in \boldsymbol{U}^{\text{ego}},$$

where $\boldsymbol{U}^{\text{ego}} := \{\boldsymbol{u}_i^{\text{ego}} \sim \boldsymbol{\pi}^{\text{ego}}(s)\}_{i=1}^{M}$ and M is the number of samples.

Define

$$\boldsymbol{u}^{\star} = \arg\max_{\boldsymbol{u}^{\text{ego}}} Q_{\text{tot}}^{\text{alter}}(s, \boldsymbol{u}^{\text{ego}})$$

The loss function for the joint ego policy can be written as: $\mathcal{L}_{\text{ego}} = -\log \boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u}^{\star} \mid s).$



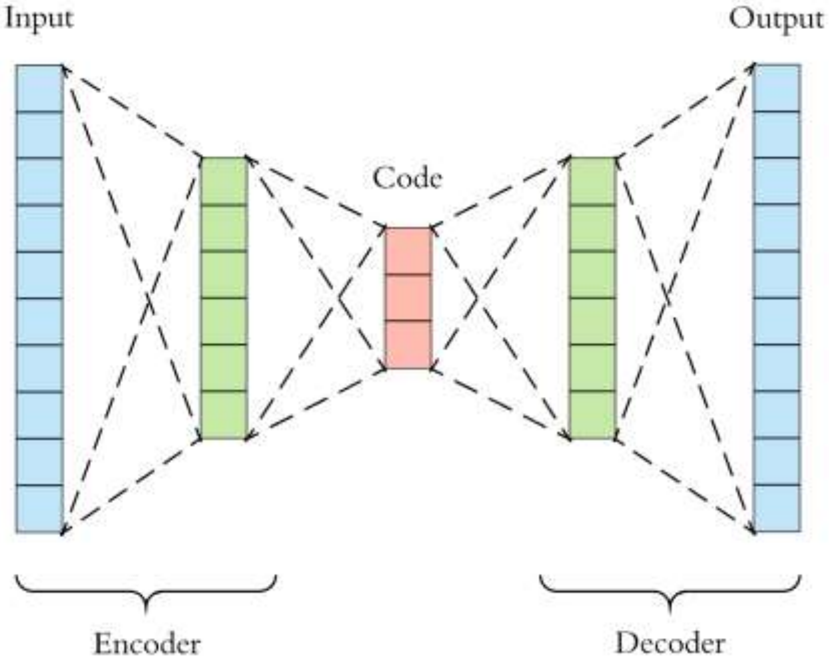
Proposition A.1. *As long as the ego policy assigns non-zero probabilities to all actions, this method will approach the objective described by Equation* (1) *as the number of samples M increases.*

Proof. Let $(u^a)^*$ denote the individual actions corresponding to the global optimal joint state-action value function under state s, and the optimal joint action is $\boldsymbol{u}^* = \{(u^1)^*, \ldots, (u^n)^*\}$. Then the probability that the sampling procedure draws $\boldsymbol{u}^*$ is expressed as:

$$\begin{aligned} p(\boldsymbol{u}^*) &= 1 - (1 - \boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u}^*|s))^M \\ &= 1 - (1 - \prod_{i=1}^{n} \pi_a^{\text{ego}}((u^a)^*|\tau^a))^M, \end{aligned}$$

where $\pi_a^{\text{ego}}(\cdot|\cdot) \in (0,1)$ is true for any action. So the second term $(1 - \boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u}^*|s))^M \in (0,1)$ in the equation decreases as M increases, indicating that $p(\boldsymbol{u}^*)$ is positively correlated with the sample size M. □

Anti-Ego Exploration



Our method leverages the fact that the auto-encoder cannot effectively encode novel data and accurately reconstruct it.

The relationship between the ego and anti-ego policy of each agent is as follows:

$$\pi_a^{\text{ego}}(\tau^a) = \text{Softmax}(f(\tau^a)),$$
$$\hat{\pi}_a^{\text{ego}}(\tau^a) = \text{Softmin}(f(\tau^a)),$$

Use Autoencoder to filter out the full-explored state-action pairs.

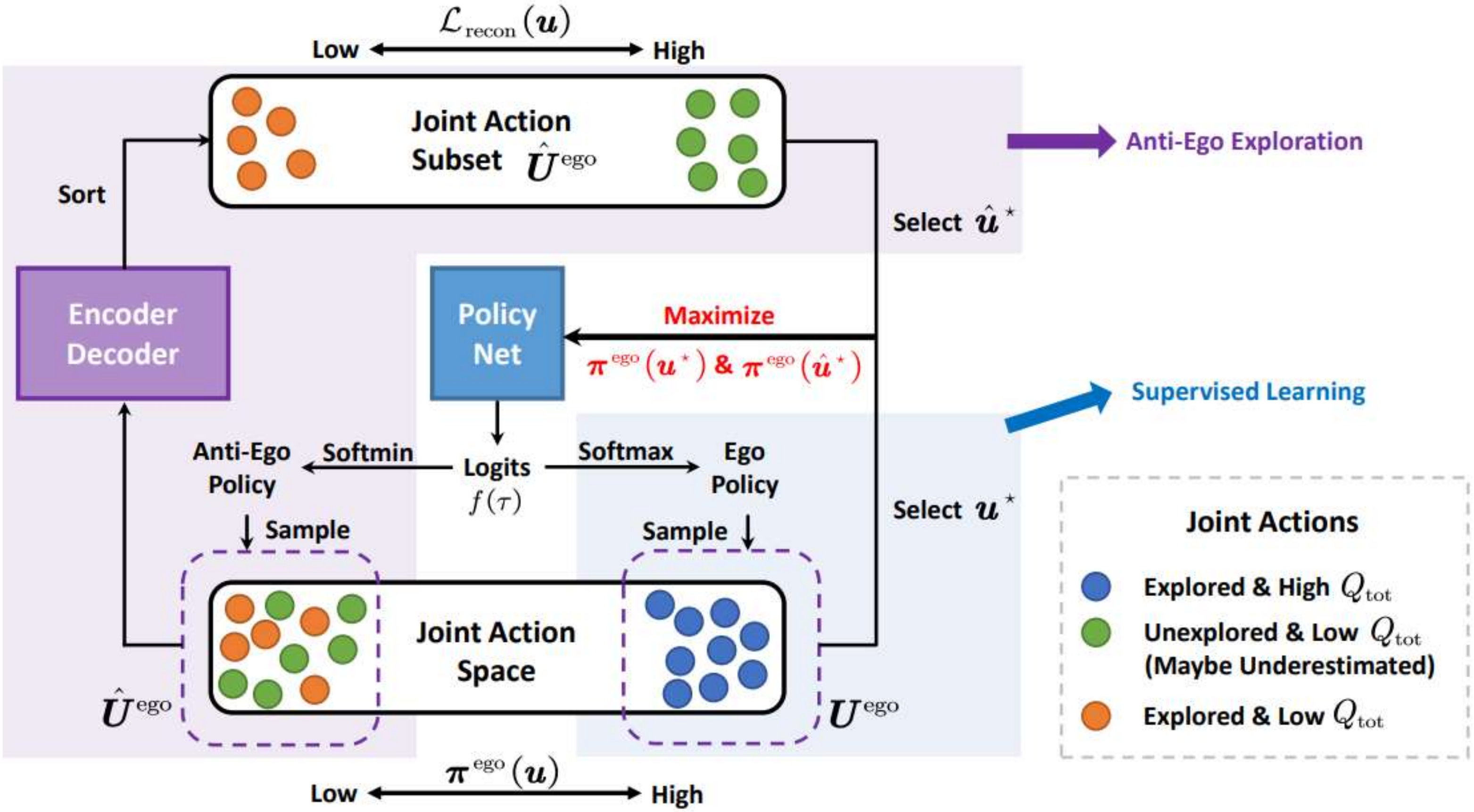
$$\mathcal{L}_{\text{recon}}(s, \boldsymbol{u}) = \text{MSE}(s, s') + \sum_{a=1}^{n} \text{CE}(u^a, (u^a)'),$$

$$\hat{\boldsymbol{u}}^\star = \arg\max_{\hat{\boldsymbol{u}}^{\text{ego}}} \mathcal{L}_{\text{recon}}(s, \hat{\boldsymbol{u}}^{\text{ego}}),$$

where $\hat{\boldsymbol{u}}^{\text{ego}} \in \hat{\boldsymbol{U}}^{\text{ego}} := \{\hat{\boldsymbol{u}}_i^{\text{ego}} \sim \hat{\boldsymbol{\pi}}^{\text{ego}}(s)\}_{i=1}^{M}$

$$\mathcal{L}_{\text{ego}} = -\left(\log \boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u}^\star \mid s) + \lambda \log \boldsymbol{\pi}^{\text{ego}}(\hat{\boldsymbol{u}}^\star \mid s)\right)$$

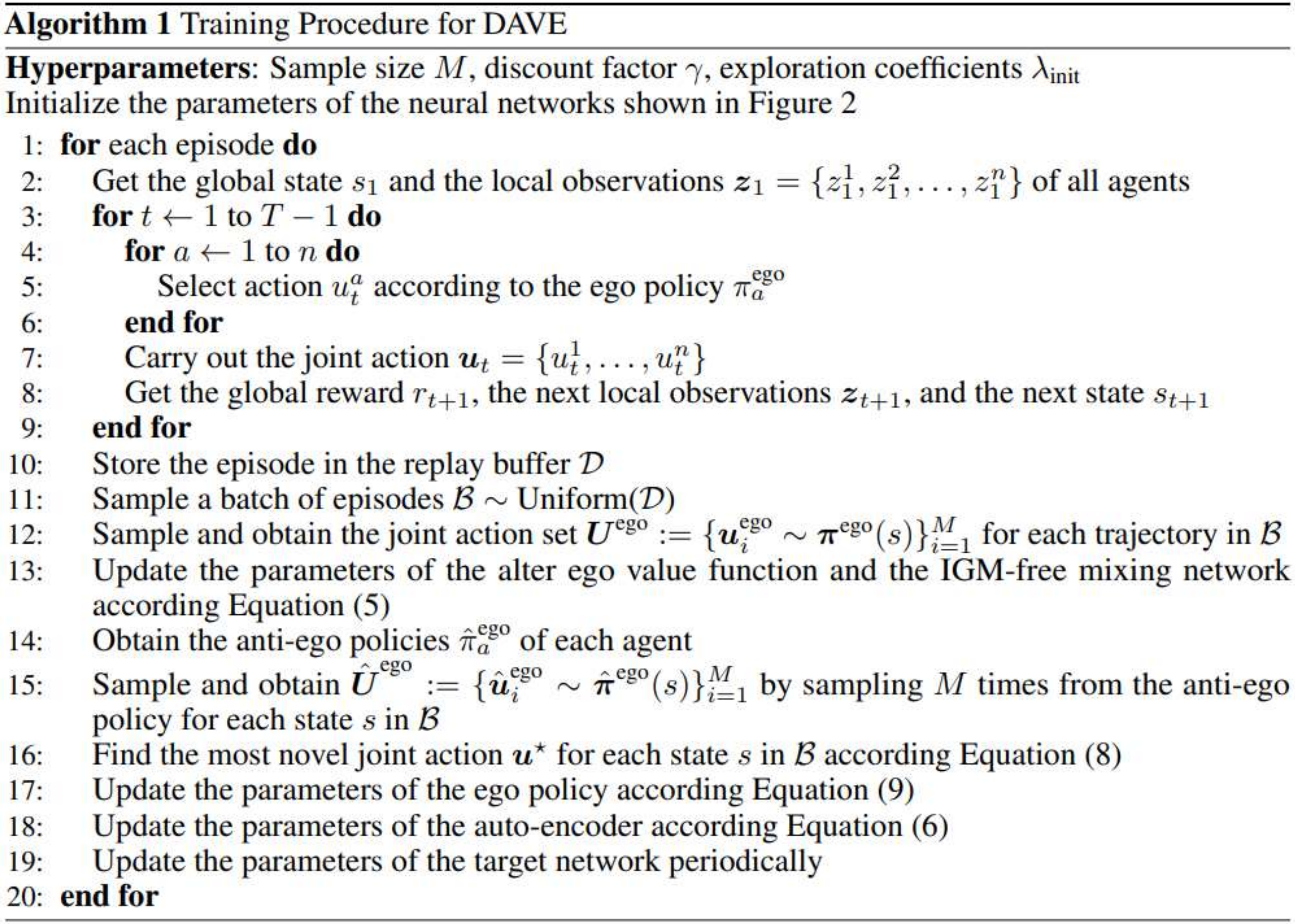
Details of the Ego Policy Update



$\mathcal{L}_{\text{recon}}(\boldsymbol{u})$
Low
High
Joint Action Subset $\hat{\boldsymbol{U}}^{\text{ego}}$
Anti-Ego Exploration
Sort
Select $\hat{\boldsymbol{u}}^\star$
Encoder Decoder
Policy Net
Maximize
$\boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u}^\star)$ & $\boldsymbol{\pi}^{\text{ego}}(\hat{\boldsymbol{u}}^\star)$
Supervised Learning
Anti-Ego Policy
Softmin
Logits $f(\tau)$
Softmax
Ego Policy
Select $\boldsymbol{u}^\star$
Sample
Sample
Joint Action Space
$\hat{\boldsymbol{U}}^{\text{ego}}$
$\boldsymbol{U}^{\text{ego}}$
$\boldsymbol{\pi}^{\text{ego}}(\boldsymbol{u})$
Low
High
Joint Actions
● Explored & High Q_{tot}
● Unexplored & Low Q_{tot} (Maybe Underestimated)
● Explored & Low Q_{tot}

Figure 9: Diagram of the update of ego policy, where $\boldsymbol{U}^{\text{ego}} := \{\boldsymbol{u}_i^{\text{ego}} \sim \boldsymbol{\pi}^{\text{ego}}(s)\}_{i=1}^{M}$ and $\hat{\boldsymbol{U}}^{\text{ego}} := \{\hat{\boldsymbol{u}}_i^{\text{ego}} \sim \hat{\boldsymbol{\pi}}^{\text{ego}}(s)\}_{i=1}^{M}$. Dots indicate different joint actions. The selected actions are $\boldsymbol{u}^\star = \underset{\boldsymbol{u}^{\text{ego}}}{\arg\max}\, Q_{\text{tot}}^{\text{alter}}(s, \boldsymbol{u}^{\text{ego}})$ and $\hat{\boldsymbol{u}}^\star = \underset{\hat{\boldsymbol{u}}^{\text{ego}}}{\arg\max}\, \mathcal{L}_{\text{recon}}(s, \hat{\boldsymbol{u}}^{\text{ego}})$.

Algorithmic Description



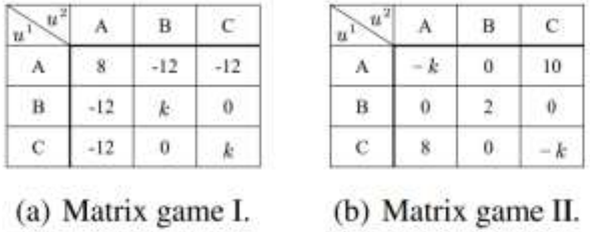
Algorithm 1 Training Procedure for DAVE

Hyperparameters: Sample size M, discount factor γ, exploration coefficients λ_{init}

Initialize the parameters of the neural networks shown in Figure 2

```
for each episode do
    Get the global state $s_1$ and the local observations $\boldsymbol{z}_1 = \{z_1^1, z_1^2, \ldots, z_1^n\}$ of all agents
    for $t \leftarrow 1$ to $T - 1$ do
        for $a \leftarrow 1$ to $n$ do
            Select action $u_t^a$ according to the ego policy $\pi_a^{\text{ego}}$
        end for
        Carry out the joint action $\boldsymbol{u}_t = \{u_t^1, \ldots, u_t^n\}$
        Get the global reward $r_{t+1}$, the next local observations $\boldsymbol{z}_{t+1}$, and the next state $s_{t+1}$
    end for
    Store the episode in the replay buffer $\mathcal{D}$
    Sample a batch of episodes $\mathcal{B} \sim \text{Uniform}(\mathcal{D})$
    Sample and obtain the joint action set $\boldsymbol{U}^{\text{ego}} := \{\boldsymbol{u}_i^{\text{ego}} \sim \boldsymbol{\pi}^{\text{ego}}(s)\}_{i=1}^{M}$ for each trajectory in $\mathcal{B}$
    Update the parameters of the alter ego value function and the IGM-free mixing network according Equation (5)
    Obtain the anti-ego policies $\hat{\pi}_a^{\text{ego}}$ of each agent
    Sample and obtain $\hat{\boldsymbol{U}}^{\text{ego}} := \{\hat{\boldsymbol{u}}_i^{\text{ego}} \sim \hat{\boldsymbol{\pi}}^{\text{ego}}(s)\}_{i=1}^{M}$ by sampling $M$ times from the anti-ego policy for each state $s$ in $\mathcal{B}$
    Find the most novel joint action $\boldsymbol{u}^{\star}$ for each state $s$ in $\mathcal{B}$ according Equation (8)
    Update the parameters of the ego policy according Equation (9)
    Update the parameters of the auto-encoder according Equation (6)
    Update the parameters of the target network periodically
end for
```

Experiment (Matrix Game)



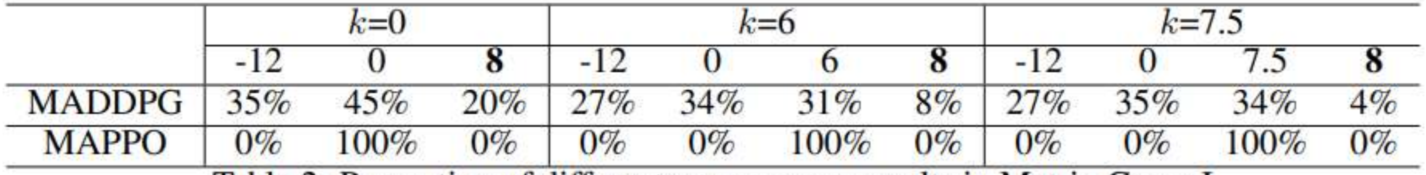
u^1 \ u^2	A	B	C
A	8	-12	-12
B	-12	k	0
C	-12	0	k

(a) Matrix game I.

u^1 \ u^2	A	B	C
A	$-k$	0	10
B	0	2	0
C	8	0	$-k$

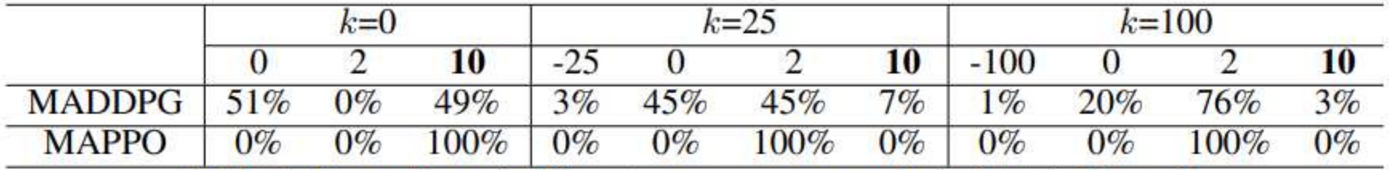
(b) Matrix game II.

Figure 3: Payoffs of the two matrix games.



	k=0			k=6				k=7.5			
	-12	0	**8**	-12	0	6	**8**	-12	0	7.5	**8**
MADDPG	35%	45%	20%	27%	34%	31%	8%	27%	35%	34%	4%
MAPPO	0%	100%	0%	0%	0%	100%	0%	0%	0%	100%	0%

Table 2: Proportion of different convergence results in Matrix Game I.



	k=0			k=25				k=100			
	0	2	**10**	-25	0	2	**10**	-100	0	2	**10**
MADDPG	51%	0%	49%	3%	45%	45%	7%	1%	20%	76%	3%
MAPPO	0%	0%	100%	0%	0%	100%	0%	0%	0%	100%	0%

Table 3: Proportion of different convergence results in Matrix Game II.

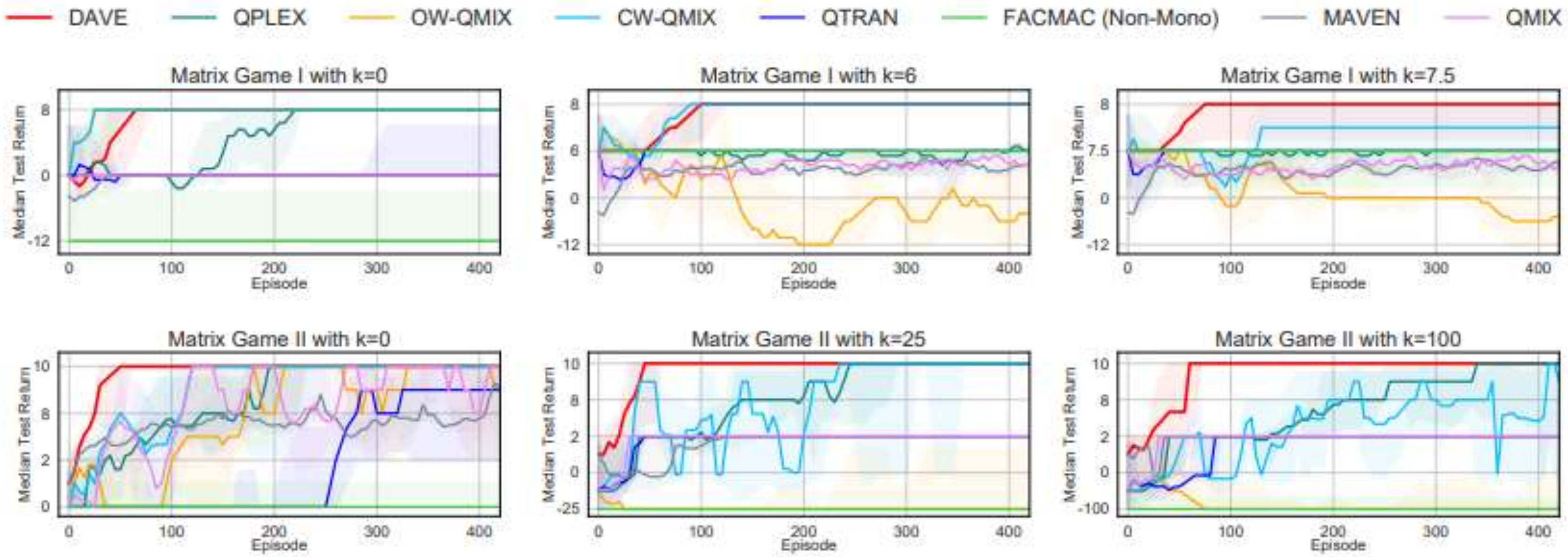


Figure 4: The learning curves of DAVE and other baselines on the matrix games. Note that the ordinates are non-uniform.

Experiment (Matrix Game)

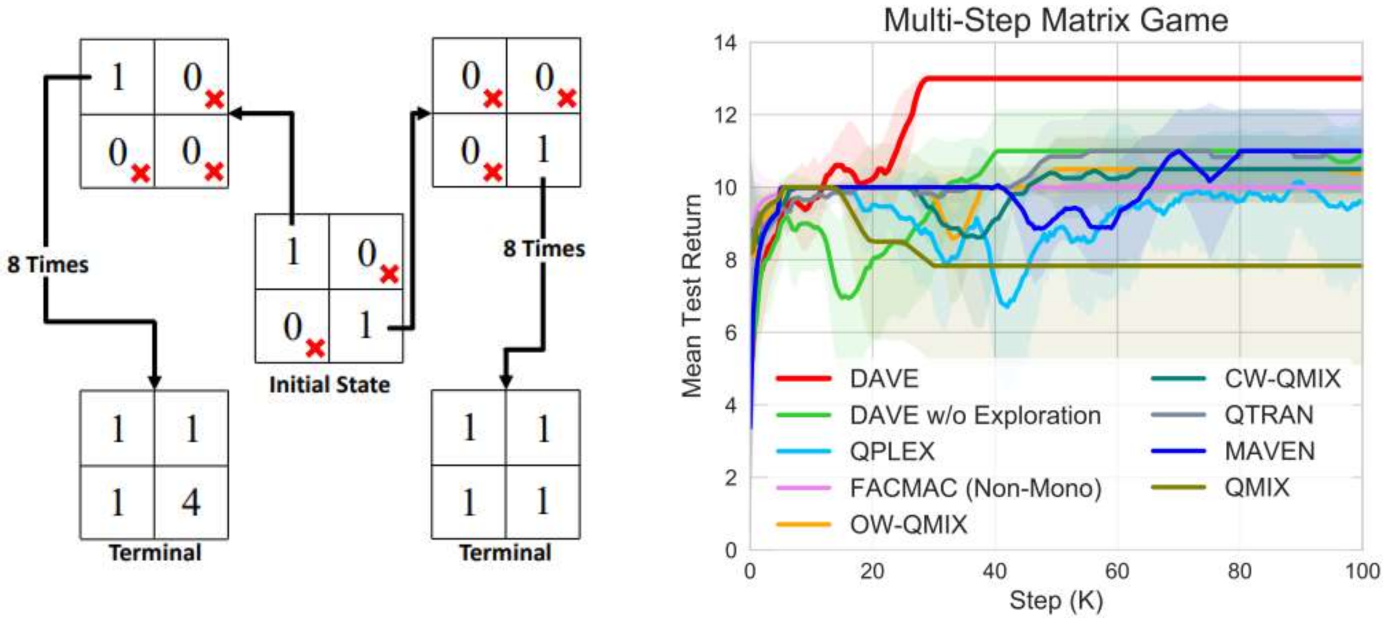


Figure 5: **Left**: Illustrations of the multi-step matrix game. **Right**: Performance over time on the multi-step matrix game.

Experiment (SMAC)

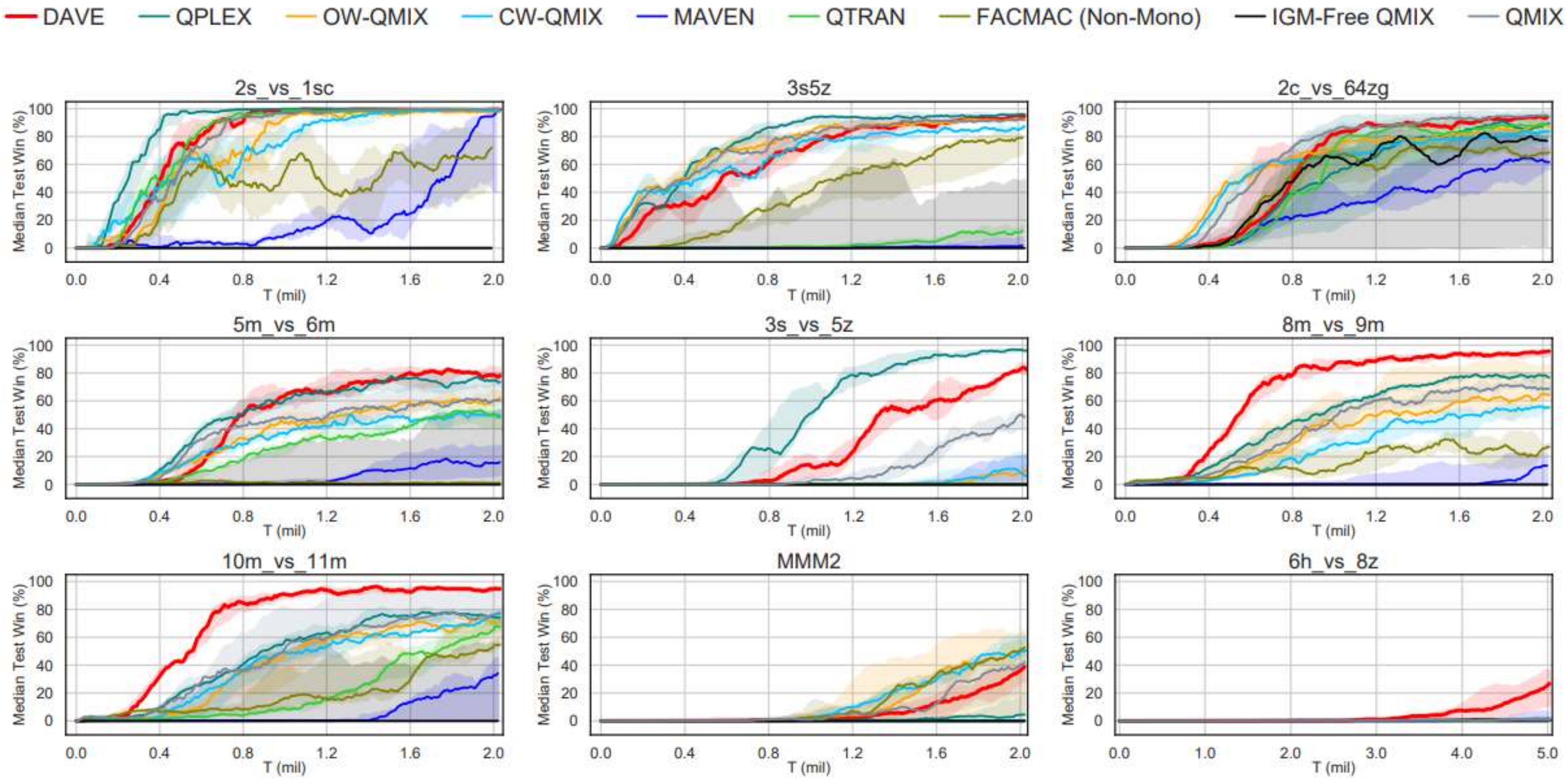


Figure 6: Performance comparison with baselines in different scenarios.

Experiment (SMACv2)

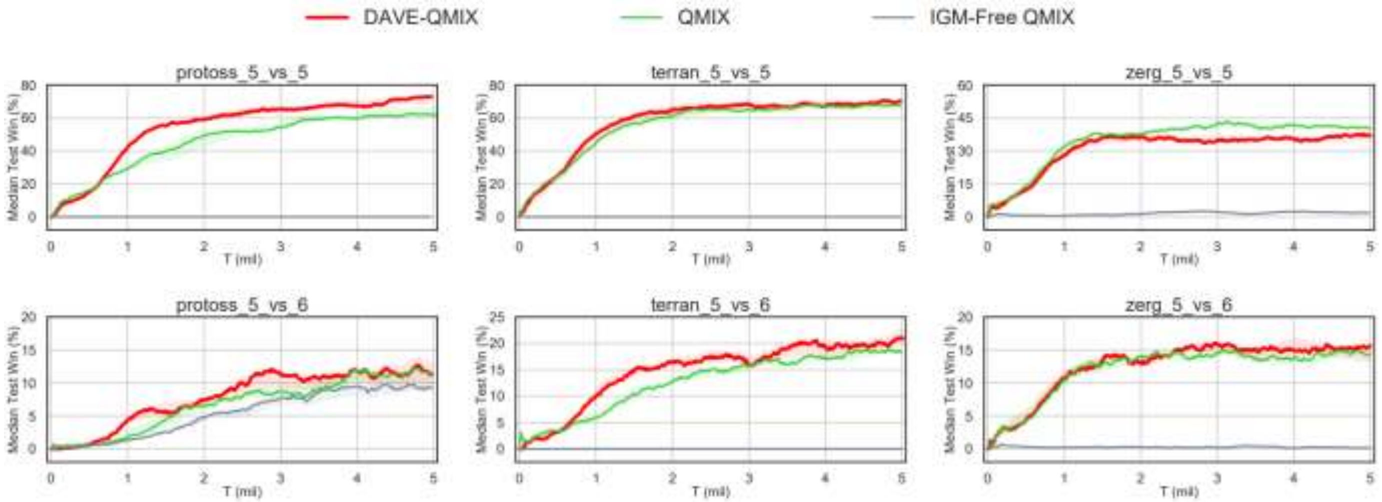


Figure 13: Comparisons of median win rate for variants of QMIX on SMACv2.

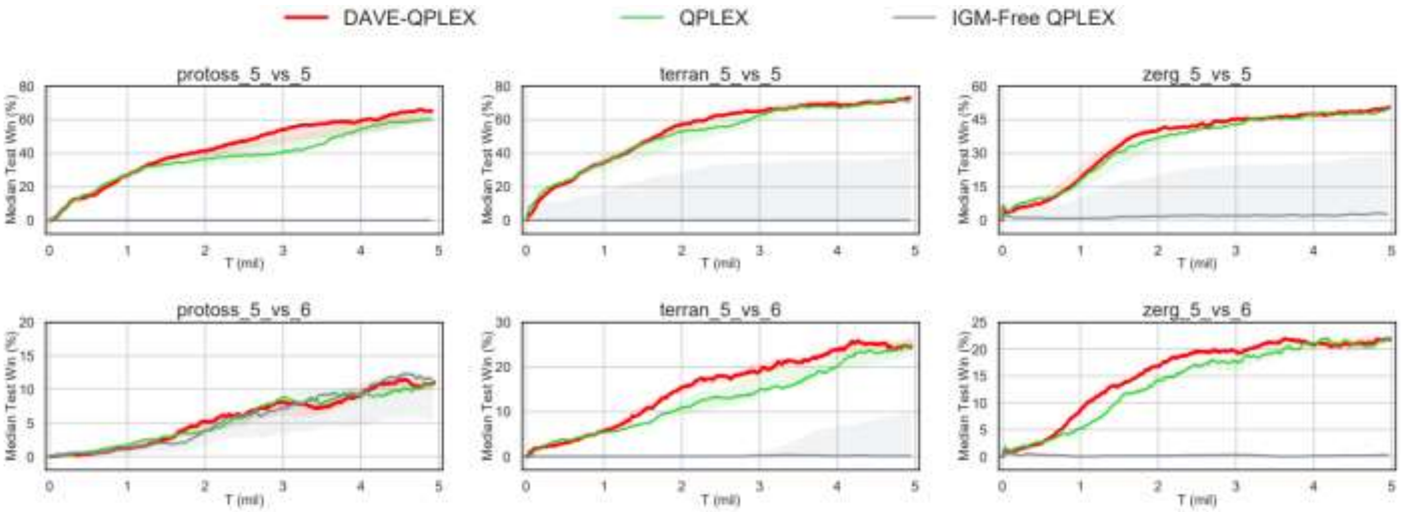
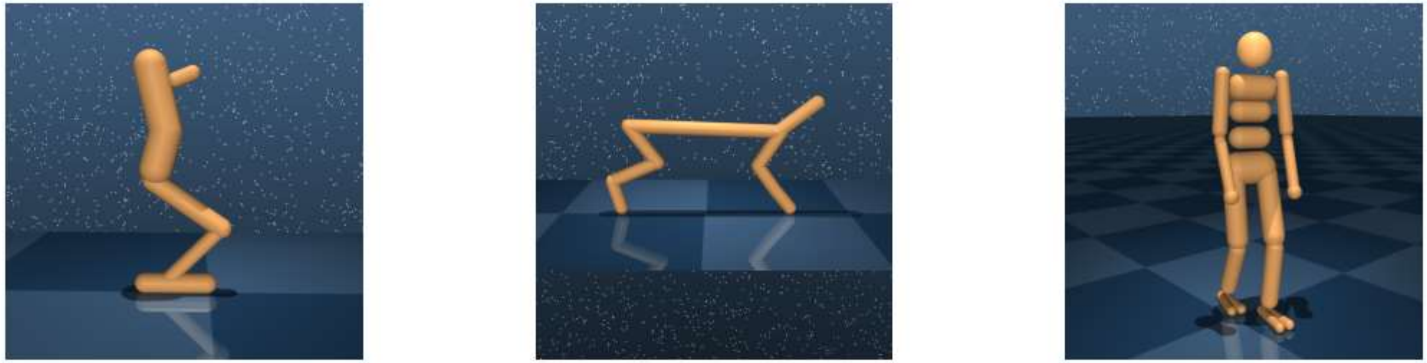


Figure 14: Comparisons of median win rate for variants of QPLEX on SMACv2.

Experiment (MA-MuJoCo)



(a) Hopper (3) (b) HalfCheetah (6) (c) Humanoid (17)

Figure 16: Illustration of benchmark tasks in Multi-Agent MuJoCo.

Contains

$$31^{17} \approx 2.26 \times 10^{25}$$

joint actions

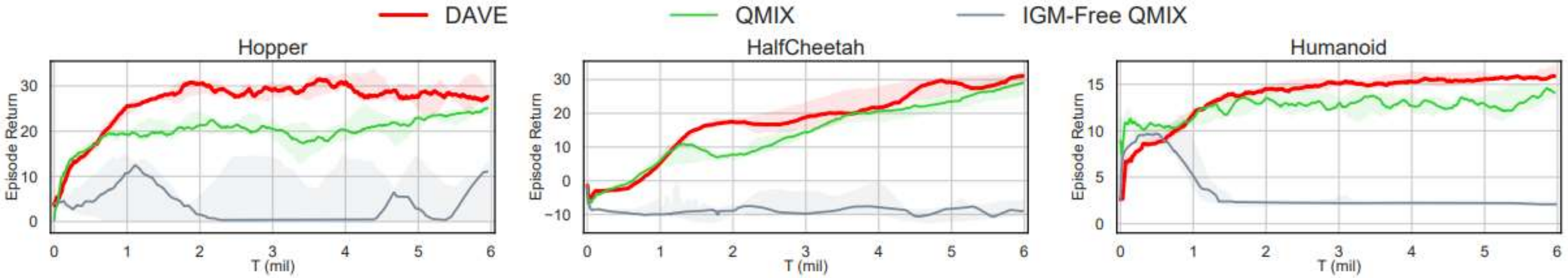


Figure 17: Median episode return on different MA-MuJoCo tasks. Note that only QMIX follows the IGM assumption.

Experiment

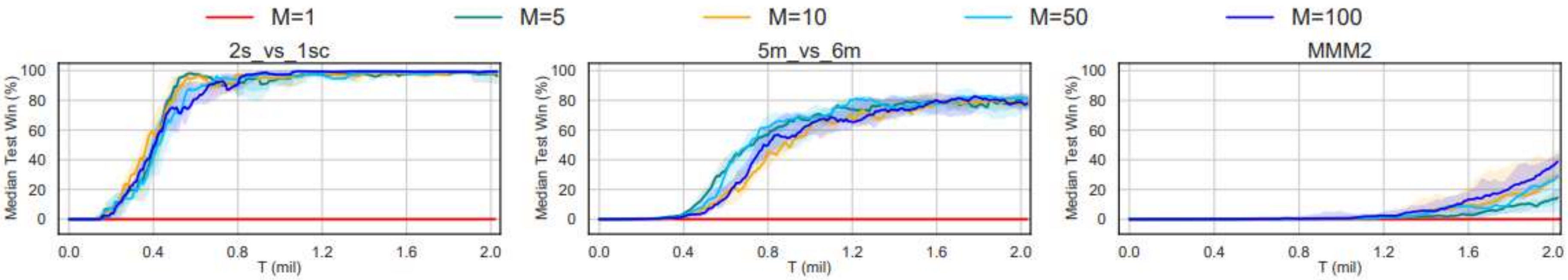


Figure 7: Influence of the sample size M for DAVE.

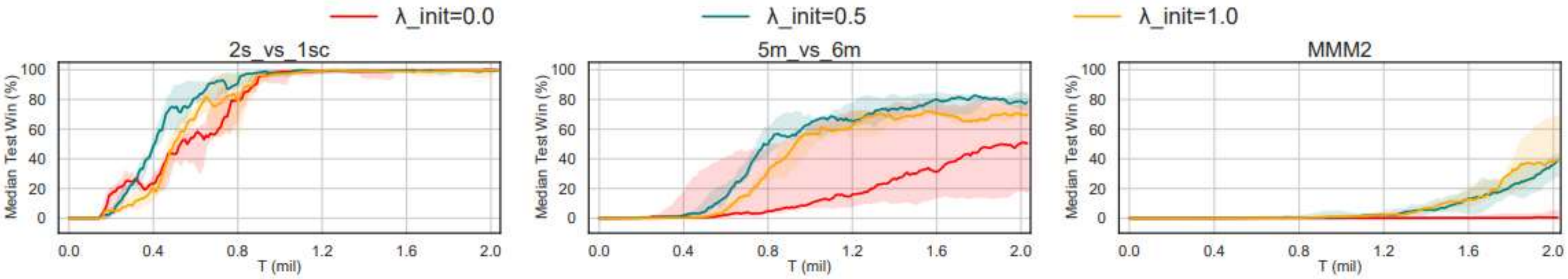


Figure 8: Results of DAVE with different λ_{init}. The action space for the three scenarios gradually increases from left to right.

Summary

Pros

- The first multi-agent value decomposition method that completely abandons IGM
- Can be applied to most IGM-based value decomposition methods and turn them into IGM-free ones
- Can achieve desirable performance in various cooperative tasks, including non-monotonic and complex tasks
- Use Autoencoder to avoid the algorithm becoming stuck in a local optimum

Cons

- Choice of λ
- May be hard to solve the tasks with large action spaces