Causality-Aware Efficient Exploration for Cooperative Multi-Agent Reinforcement Learning

Hongye Cao[1], **Tianpei Yang**[1,2*], **Fan Feng**[3], **Hammadi Rafik Ouariachi**[2], **Yali Du**[4]
Meng Fang[5], **Jing Huo**[1], **Yang Gao**[1,2]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[3]City University of Hong Kong, Hong Kong, China
[4]King's College London, London, UK
[5]University of Liverpool, Liverpool, UK

Abstract

Exploration is critical for cooperative multi agent reinforcement learning (MARL) to improve sample efficiency. However, existing intrinsic motivation based exploration strategies in MARL overlook the causal relationships among agents, global states, and rewards, suffering from interference by irrelevant factors and resulting in sample inefficiency. To address this issue, we propose Causality aware Efficient Exploration (CEE), a novel framework that enhances sample efficiency by inferring causal relationships between agents, global states with respect to rewards, thereby enabling causality guided exploration. Specifically, CEE operates through two components. First, CEE identifies causal relationships between global states and rewards, filtering out causally irrelevant state features that do not have a high impact on rewards to keep decision critical state information. Second, CEE discovers causal relationships between agents' behaviors and rewards to quantify each agent's contribution to collective performance. To achieve this, we introduce a causal entropy objective that promotes exploration aligned with decision critical aspects of the underlying causal structure. We provide comprehensive validation through experiments on 21 challenging tasks spanning SMAC, SMAC v2, and Google Research Football (GRF) environments. Our results demonstrate that CEE achieves superior performance in terms of sample efficiency and asymptotic performance compared to existing MARL methods.

Introduction

Cooperative multi-agent reinforcement learning (MARL) has achieved significant improvements in various real-world multi-agent systems, including autonomous driving (Kiran et al. 2022; Zheng and Gu 2025), multi-robot manipulation (Gu et al. 2023; Huang et al. 2025), and sensor networks (Lin et al. 2025; Xu, Zhong, and Wang 2020). However, existing approaches often suffer from sample inefficiency, with inadequate or suboptimal exploration strategies constituting one primary underlying cause (Hao et al. 2023).

*Corresponding to Tianpei Yang (tianpei.yang@nju.edu.cn).

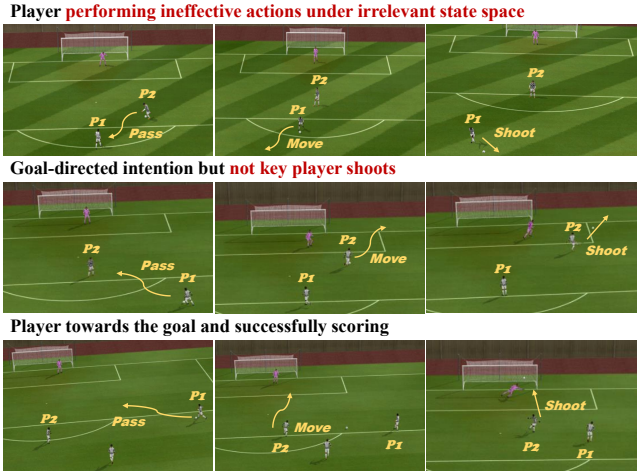

Figure 1: Example of a football task with three trajectories, aiming to score goals to achieve high reward returns.

A major challenge in MARL is to develop exploration strategies that are efficient in the presence of irrelevant or misleading information. In particular, a critical and underexplored question is how to identify and leverage the causal structure of the environment to guide effective exploration. For example, in cooperative football, only certain players and fields near the goal are essential for scoring, while other irrelevant factors may diminish exploration efficiency. Existing approaches focus on intrinsic motivation bonuses to promote exploration (Jaques et al. 2019; Chen, Huang, and Schneider 2024), but typically overlook the underlying causal relationships among agents, global states, and rewards, potentially leading to high exploration costs. Failure to filter out task-irrelevant state features could reduce efficiency, while neglecting the causal influence of each agent's actions make it hard to improve collaboration. Causal discovery can capture essential information by analyzing causal relationships between different factors, filtering out irrelevant information, and avoiding interference from spurious correlations (Spirtes, Glymour, and Scheines 2000; Pearl 2009; Cao et al. 2025b,a; Huang et al. 2022).

While causality-based exploration has proven effective in single-agent RL (Liu et al. 2024; Ji et al. 2024), its application to multi-agent tasks remains insufficiently unexplored.

In this work, we focus on identifying and exploiting causal relationships between agents, global states, with respect to rewards. There are a few causal MARL approaches (Liu et al. 2023; Wang et al. 2025) that primarily analyze causal relationships between actions and states. But, these works overlook the reward-guided causal relationships, where reward feedback serves as a critical signal during exploration. We aim to filter out irrelevant state features to mitigate interference from noisy information while enabling targeted exploration that prioritizes agents with higher reward contributions. As shown in Figure 1, a motivating football example illustrates three different trajectories. The first line shows players executing actions under irrelevant areas far from the goal, resulting in failure to achieve the objective. The second trajectory demonstrates that while players make progress toward the goal, the key shooting are not performed by the optimal player. The third trajectory reveals that the system achieves optimal task execution under critical state space and prioritizing the player with higher reward contribution for shooting.

Therefore, by leveraging the potential of causal discovery to tackle the issue of sample inefficiency, we propose Causality-aware Efficient Exploration (**CEE**), a novel framework that enhances sample efficiency by inferring causal relationships between agents, global states with respect to rewards. Specifically, **CEE** operates through a dual-stage causal analysis. First, **CEE** discovers causal relationships between global states and rewards, constructing a *state-reward causal matrix* that masks irrelevant state features that have small impact on rewards for filtering the mixing network input, thereby focusing exploration on decision-critical state information during centralized training process. Second, **CEE** identifies causal relationships between agents behaviors and global rewards by learning an *agent-reward causal matrix* that quantifies each agent's contribution to collective performance on rewards. Leveraging this learned *agent-reward causal matrix*, we design a causal entropy optimization objective that optimizes causally-informed exploration policy based on filtered state information. The main contributions of this work can be summarized as follows:

- We introduce **CEE**, a novel framework that prioritizes causal information through reward-guided exploration. To our knowledge, this is the first work to discover and leverage reward-guided causal relationships among agents and global states in cooperative MARL.
- **CEE** first learns a state-reward causal matrix to analyze causal relationships between global states and rewards, masking irrelevant state features to reduce noise during exploration. **CEE** then learns an agent-reward causal matrix to reweight different agents and incorporates a causal entropy objective that optimizes causally-informed exploration policy.
- We conduct extensive empirical validation across 21 challenging tasks in SMAC, SMAC-v2, and Google Research Football (GRF) environments. Our experimental results consistently demonstrate that **CEE** achieves superior performance in terms of sample efficiency, and asymptotic performance compared to existing methods.

Related Work

MARL. The centralized training with decentralized execution (CTDE) framework represents the most widely adopted learning paradigm for cooperative MARL, where agents leverage shared global state information during training while executing independent decisions during deployment (Kraemer and Banerjee 2016; Rashid et al. 2020). Under this framework, MARL approaches primarily involve two paradigms: policy-based and value-based methods. Policy-based approaches like MADDPG (Lowe et al. 2017), MAPPO (Yu et al. 2022), and COMA (Foerster et al. 2018) directly optimize policies using gradient methods. Value-based methods such as VDN (Sunehag et al. 2017), QMIX (Rashid et al. 2020), and QPLEX (Wang et al. 2020) focus on learning action-value Q functions while addressing the credit assignment problem through value decomposition. Despite significant progress, sample efficiency remains a critical challenge in MARL due to exponentially large joint action spaces and non-stationary dynamics caused by multiple agents. Current strategies (Chen, Huang, and Schneider 2024; Jianye et al. 2022) do not explicitly leverage causal relationships between agents, states, and rewards for exploration guidance to improve sample efficiency.

Causality in MARL. Existing causal MARL approaches focus on discovering causal relationships among agents to achieve interpretable decision-making and efficient exploration (Grimbly, Shock, and Pretorius 2021). Some works (Du et al. 2024; Pina, De Silva, and Artaud 2025) infer causal relationship among agents and design intrinsic rewards to encourage exploration and coordination. LAIES (Liu et al. 2023) uses causal models to calculate an agent's effect on the states and introduced intrinsic rewards to motivate lazy agents. Other approaches (Wang et al. 2025; Zhang et al. 2024) employ causal reward decomposition methods to achieve better credit assignment under online or offline settings.

Preliminaries

Dec-POMDP. A fully cooperative multi-agent task is usually modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP). Dec-POMDP is represented by a tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \mathcal{Z}, \mathcal{O}, \gamma \rangle$, where $\mathcal{N}$ is a finite set of n agents, and $\mathcal{S}$ is the state space. $\mathcal{A}$ represents the joint action space of all agents, and $\mathcal{A}^i$ is the local action space of agent $i \in \mathcal{N}$. At each time step t, each agent i receives its own observation $o_t^i \in \mathcal{Z}$ according to its observation function $\mathcal{O}^i(o_t^i|s_t) \in \mathcal{O}$ and chooses its local action $a_t^i \in \mathcal{A}^i$. After all agents select actions, the environment transits to the next state s_{t+1} according to the state transition function $\mathcal{P}(s_{t+1}|s_t, \boldsymbol{a_t})$ and all agents receives a joint reward r_t according to $\mathcal{R}(s_t, \boldsymbol{a_t})$, where $\boldsymbol{a_t} = (a_t^1, \ldots, a_t^n)$ denotes the joint action of all agents. Moreover, each agent learns its own policy $\pi^i(a^i|\tau^i) : \mathcal{T} \times \mathcal{A} \to [0, 1]$ which conditions on its action-observation history $\tau^i \in \mathcal{T}$.

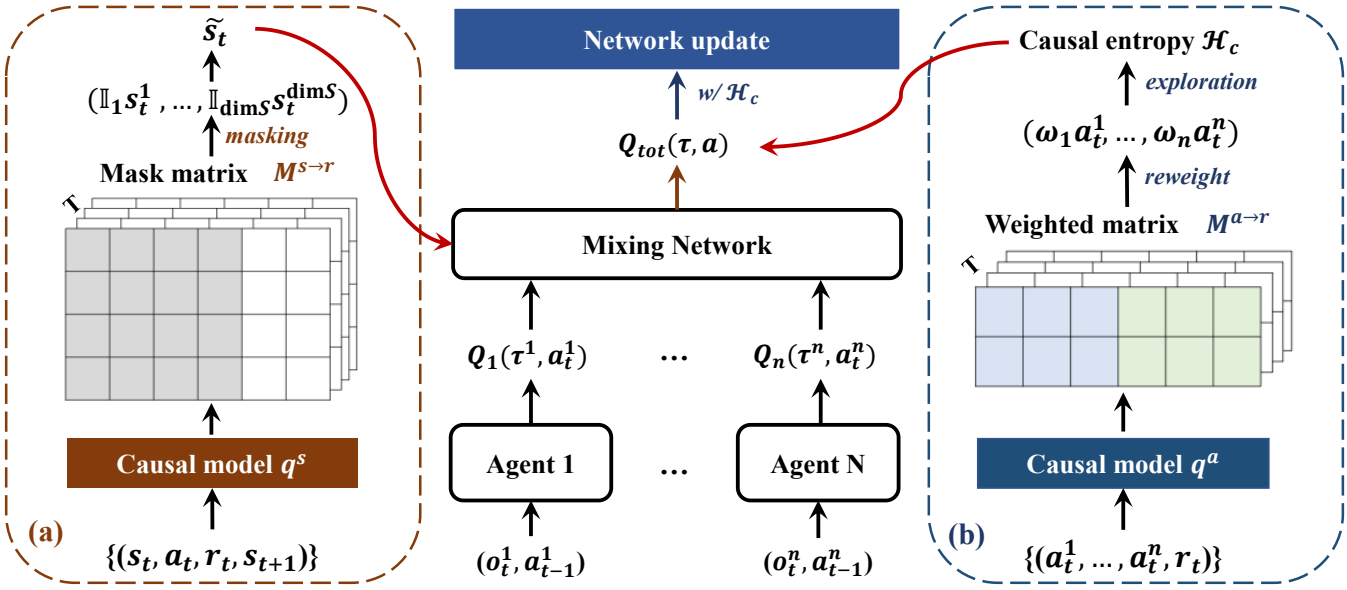

Figure 2: The framework of **CEE**. (a) The component of state-reward causal discovery to mask irrelevant state feature. (b) The component of agent-reward causal discovery for causal action exploration.

The goal of all agents is to learn an optimal joint policy $\pi^* = \langle \pi^{1,*}, \dots, \pi^{n,*} \rangle$ that maximizes the discounted cumulative reward $\mathbb{E}_{\pi,\mathcal{P}}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1]$ is the discount factor.

Value Decomposition Algorithms. Value decomposition algorithms aim to learn a factorized global action-value function, denoted as Q_{tot}, that can be decomposed into individual agents' decentralized utility functions Q_i. The Individual-Global-Max (IGM) principle is introduced to ensure optimal consistency between these two value functions, which is defined as follows:

$$\underset{\boldsymbol{a}}{\text{argmax}} Q_{tot}(\tau, \boldsymbol{a}) = \begin{pmatrix} \text{argmax}_{a^1} Q_1(\tau^1, a^1) \\ \text{argmax}_{a^2} Q_2(\tau^2, a^2) \\ \cdots \\ \text{argmax}_{a^n} Q_n(\tau^n, a^n) \end{pmatrix}, \quad (1)$$

where τ denotes the joint action-observation history of all agents and $\boldsymbol{a}$ represents the joint action. Specifically, QMIX adheres to this principle by enforcing a monotonicity constraint, which is achieved through a mixing network g that computes the global value function: $Q_{tot} = g(Q_1, \cdots, Q_n)$. The parameters of g are generated by hypernetworks conditioned on the global state s, and the weights of g are constrained to be non-negative to satisfy the monotonicity constraint. Our method builds upon the value decomposition architecture to achieve causality-aware efficient exploration in cooperative MARL.

Structural Causal Model. A Structural Causal Model (SCM) (Pearl 2009) is defined by a distribution over random variables, defined as $\mathcal{V} = \{s_t^1, \cdots, s_t^d, a_t^1, \cdots, a_t^n, r_t, s_{t+1}^1, \cdots, s_{t+1}^d\}$ and a Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a conditional distribution $\mathcal{P}(v_i|\text{PA}(v_i))$ for node $v_i \in \mathcal{V}$. Then the distribution can be specified as:

$$p(v_1, \dots, v_{|\mathcal{V}|}) = \prod_{i=1}^{|\mathcal{V}|} p(v_i|\text{PA}(v_i)), \quad (2)$$

where $\text{PA}(v_i)$ is the set of parents of the node v_i in graph $\mathcal{G}$. Notably, d represents the dimension of the global state s_t^d and n stands for the number of agent a_t^n in **CEE**.

Causal Structures in Dec-POMDP. To facilitate causal reasoning in MARL, we explicitly model the Dec-POMDP as with factored structures (Kearns and Koller 1999; Guestrin et al. 2003) and analyze the underlying causal structures between agents behaviors, global states, with respect to rewards. In this formulation, nodes represent variables, including the reward, dimensions of the global state, and the actions of agents, while directed edges denote the causal relationships among these variables within the environment dynamics and the reward generation process.

We leverage causal discovery methods to learn the structural causal graph $\mathcal{G}$, which can be represented by an adjacency matrix M capturing the directional influences among variables. Specifically, we focus on identifying the causal influence of the state and agents' actions on the reward, represented by structural vectors $M^{s\to r}$ and $M^{a\to r}$, respectively. Formally, the reward at time step t can be expressed as:

$$r_t = R(M^{s\to r} \odot s_t, M^{a\to r} \odot a_t, \epsilon_{r,t}), \quad (3)$$

where $\odot$ denotes the element-wise product. $M^{s\to r} \in \mathbb{R}^{|s|\times 1}$ and $M^{a\to r} \in \mathbb{R}^{|a|\times 1}$ are the adjacency matrices indicating the influence of current states and agents' actions on the reward, respectively, and $\epsilon_{r,t}$ represents i.i.d. Gaussian noise. Under the Markov condition and faithfulness assumption (Pearl 2009; Spirtes, Glymour, and Scheines 2001; Feng

et al. 2022), the structural vectors are identifiable. In this work, our objective is to discover and leverage these two causal matrices to encourage efficient exploration.

Approach

In this section, we introduce the proposed framework **CEE**, as shown in Figure 2. Under the value decomposition architecture, each agent network receives observation o_t^i and previous action a_{t-1}^i to compute the individual Q-value Q_i. All individual Q-values and the global state s_t are then fed into the mixing network to obtain the global Q-value Q_{tot}. **CEE** contains two components during this learning process: (a) state-reward causal discovery that masks causally irrelevant state features in the global state for the mixing network, reducing noisy information during exploration, and (b) agent-reward causal discovery that learns a causal matrix to reweight agents' contributions based on their causal relationships with rewards, thereby encouraging exploration through a causal entropy $\mathcal{H}_c$.

Causal State Filtering

As illustrated in Figure 1 (row 1), multi-agent systems often encounter some irrelevant state features that interfere with exploration. We aim to minimize interference from irrelevant state information. While reward signals constitute a fundamental component for guiding policy optimization; existing methods do not explicitly model the causal relationships between states and rewards. We examine the mixing network architecture, where the global state s is incorporated into the mixing network to provide comprehensive environmental context for computing the global $Q_{tot}(s,a) = \text{Mixer}(Q_1(o^1,a^1),\cdots,Q_n(o^n,a^n),s)$.

The global state s encompasses a substantial volume of policy-irrelevant information that bears no causal relationship to rewards. The indiscriminate incorporation of such extraneous state components into the mixing network introduces noise into the value function approximation process, thereby compromising learning efficiency. To address this limitation, we propose a principled approach that leverages causal analysis to discern the intrinsic relationships between state components and reward signals. By identifying and filtering causally-informed state information for the mixing network input, we aim to enhance the learning efficiency (Figure 2 (a)).

To discover causal relationships between global states and rewards in multi-agent systems, we first collect trajectories $\{(s_t,a_t,r_t,s_{t+1})\}$ during the training process. These trajectories capture the temporal evolution of global states and corresponding global rewards. We then employ an efficient algorithm DirectLiNGAM (Shimizu et al. 2011) to train a structural causal model q^s on collected trajectories, which estimates the causal influence between state dimensions and rewards while accounting for potential confounding effects. This analysis yields a causal matrix $M^{s\to r}\in\mathbb{R}^{N\times 1}$, where each element $M_i^{s\to r}$ quantifies the strength of the causal relationship between the i-th state dimension and reward.

Utilizing the learned causal matrix $M^{s\to r}$, we construct a binary mask M_s where each element is determined by:

$$M_s^i = \mathbb{I}[i\notin\mathcal{K}], i\in[1,\dim\mathcal{S}], \tag{4}$$

where $\dim\mathcal{S}$ denotes the state dimension, and $\mathcal{K}$ denotes the indices of the k state dimensions with the lowest causal values in $M^{s\to r}$, and k is the scale of causal masking, and $\mathbb{I}[\cdot]$ is the indicator function. The mixing network subsequently incorporates this masked state information:

$$Q_{tot}(s,a) = \text{Mixer}(Q_1(o^1,a^1),\cdots,Q_n(o^n,a^n),\tilde{s}), \tag{5}$$

where $\tilde{s} = s\odot M_s$. Through the application of the causal mask M_s, we ensure causally relevant state components contribute to the value function approximation, thereby enhancing the sample efficiency of the mixing network.

Causal Action Exploration

As illustrated in Figure 1 (row 2), after filtering out causal irrelevant state, we aim to mitigate excessive exploration by task-irrelevant agents with minimal reward influence while encouraging decision-making from contributory agents. Sufficiently considering the causal relationships between agents and rewards will effectively analyze the contribution of different agents at various stages of policy learning, thereby enabling more targeted exploration encouragement and facilitating credit assignment. Hence, we aim to design an agent-reward causality-aware encouraging optimization objective in MARL (Figure 2 (b)).

Specifically, we collect the actions of each agents and global rewards over multiple time steps, and input these data into the causal model q^a. We also utilize the DirectLiNGAM algorithm to identify the causal influence of each agent's actions on the reward, which is updated at a regular interval T. Then, we get a causal weighted matrix of $M^{a\to r}$. By leveraging this causal matrix, we can focus on the pivotal agent behaviors during learning process, potentially leading to more efficient exploration.

The learned weighted matrix is applied to reweight agents' actions $\{\omega_1 a_t^1,\cdots,\omega_n a_t^n\}$ within the joint policy, and incorporated into the mixing network optimization. Inspired from maximum entropy reinforcement learning (Haarnoja et al. 2018; Eysenbach and Levine 2021; Chen, Huang, and Schneider 2024), we propose a causality-aware entropy objective $\mathcal{H}_c$ for encouraging exploration to explore a wide range of agents that are causally informed, which is defined as follows:

$$\begin{aligned}&\mathcal{H}_c(\pi_c(a_t|s_t;M^{a\to r}))\\&=-\sum_{i=1}^{\mathcal{N}} M^{a\to r}\odot\pi_c(a_t^i|s_t)\log\pi_c(a_t^i|s_t),\end{aligned} \tag{6}$$

where π_c represents the joint policy, which is defined as $\pi_c(\cdot|s)\propto\text{softmax}(Q_{tot}(\cdot,s))$ in value-based MARL methods. By utilizing $\mathcal{H}_c$ instead of the standard policy entropy, we can prioritize the exploration of pivotal agents that are more likely to have significant causal effects on the reward.

Therefore, the current learning objective can be formulated as follows:

$$J(\pi_c) = \mathbb{E}_{\mu_0,\pi,\mathcal{P}}[\textstyle\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t)+\alpha\mathcal{H}_c(\pi_c(\cdot|s_t)))], \tag{7}$$

where α is the temperature parameter to control the degree of randomness, and μ_0 is the distribution of the initial state s_0. The modified value function is defined as follows:

$$V_{tot}(s) = \mathbb{E}_{a\sim\pi_c(\cdot|s)}[Q_{tot}(s,a) + \alpha\mathcal{H}_c(\pi_c(a|s))]. \quad (8)$$

Based on the causality-aware entropy, the Q-value could be computed iteratively by applying a modified Bellman operator as follows:

$$\mathcal{T}^{\pi_c}Q_{tot} = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}}[V_{tot}(s')]. \quad (9)$$

To optimize the mixing network, we employ TD(λ) (Schulman et al. 2015) for optimization, leveraging its balance between bias and variance for stable learning in multi-agent settings. The hyperparameter θ for the mixing network is updated using the following loss function:

$$\mathcal{L}_Q(\theta) = \frac{1}{2}(Q^{\theta}_{tot}(s_t,a_t) - \mathcal{T}^{\pi_c}_{\lambda}Q_{tot}(s_t,a_t))^2. \quad (10)$$

For value-based methods, we use the following loss function for updating the joint policy:

$$\mathcal{L}_{\pi_c}(\phi) = \frac{1}{2}\left(\sum_{i=1}^{n} f_i^{\phi}(Q_i(o_t^i,a_t^i),\tilde{s}_t) - Q_{tot}(s_t,a_t)\right), \quad (11)$$

where function f is an order-preserving transformation for maintaining the IGM principle, following (Chen, Huang, and Schneider 2024). This function is a component within the mixing network parameterized by hyperparameter ϕ.

Practical Implementations

We delineate the practical implementations of our proposed method driven by the aforementioned analysis. Algorithm 1 presents the detailed procedural steps. The implementation process consists of three primary phases: data collection, causal model learning, and network parameter optimization.

Initially, we employ policy π_c to collect interaction data from the environment (lines 2-3). The collected trajectories are stored in the replay buffer $\mathcal{D}$ to facilitate both causal model learning and network updates. Subsequently, we sample transition batches $\mathcal{D}_c$ from replay buffer $\mathcal{D}$ and learn two structural causal models to obtain causal matrices $M^{s\to r}$ and $M^{a\to r}$ (lines 4-10). The state-reward causal matrix $M^{s\to r}$ is employed to mask causally irrelevant state dimensions in the global state representation for mixing network input. Concurrently, the agent-reward causal matrix $M^{a\to r}$ is utilized to reweight agents' actions in causal entropy computation, thereby promoting causally-informed agents exploration. Finally, during each gradient step, we compute the loss function as specified in Eq. 10 and Eq. 11, and update the network hyperparameters accordingly (lines 11-16).

Experiments

Our experiments aim to address the following questions: (i) How does the performance of **CEE** compare to other cooperative MARL approaches in diverse tasks. (ii) Can **CEE**, through causal state filtering and casual action exploration, improve the sample efficiency? (iii) What are the effects of the components and computation burden in **CEE**?

Algorithm 1: Causality-Aware Efficient Exploration (**CEE**)

Input: hyperparameters θ, ϕ, replay buffer $\mathcal{D}$, causal update interval T, causal matrices $M^{s\to r}$, $M^{s\to r}$, policy π_c
Initiate: $\theta' \leftarrow \theta$, $\mathcal{D} \leftarrow \emptyset$

1: **for** each environment step t **do**
2: Collect data with π_c from environment
3: Add to replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1})\}$
4: **if** every T environment step **then**
5: Sample transitions $\mathcal{D}_c$ from replay buffer $\mathcal{D}$
6: Learn causal model with $\{(s_t, r_t)\}_{t=1}^{|\mathcal{D}_c|}$
7: Update causal mask matrix $M^{s\to r}$
8: Learn causal model with $\left\{(a_t^1, \cdots, a_t^n, r_t)\right\}_{t=1}^{|\mathcal{D}_c|}$
9: Update causal mask matrix $M^{a\to r}$
10: **end if**
11: **if** every gradient step **then**
12: calculate $\mathcal{L}_Q(\theta)$, $\mathcal{L}_{\pi_c}(\phi)$ via Eq. 10, Eq. 11
13: $\theta \leftarrow \theta - \lambda_Q \bigtriangledown_\theta \mathcal{L}_Q(\theta)$
14: $\phi \leftarrow \phi - \lambda_Q \bigtriangledown_\phi \mathcal{L}_{\pi_c}(\phi)$
15: $\theta' \leftarrow \omega\theta + (1-\omega)\theta'$
16: **end if**
17: **end for**

Experimental Setup

Benchmark and Implementation Details. We evaluate **CEE** on 4 SMAC (Samvelyan et al. 2019) tasks (2 hard and 2 super hard), 15 SMAC-v2 (Ellis et al. 2023) tasks, and 2 GRF (Kurach et al. 2020) tasks, covering scenarios ranging from hard to super hard, different fields of view, and from small-scale to large-scale settings with up to 14 cooperating agents, providing a comprehensive assessment of our method's performance. All codes and hyperparameter settings are based on the codebase of PyMARL2 (Hu et al. 2021). For evaluation, all experiments are carried out with five random seeds.

Baselines. We first select traditional CTDE methods as baselines, including VDN (Sunehag et al. 2017), QMIX (Rashid et al. 2020), QPLEX (Wang et al. 2020), and HPN-QMIX (Jianye et al. 2022). Moreover, we compare **CEE** with causal MARL method LAIES (Liu et al. 2023), exploration encouraged method Soft-QMIX (Chen, Huang, and Schneider 2024), and reconstruction-giuded method RGP (Qifan et al. 2025).

Main Results

SMAC. We first analyze the experimental results across 4 SMAC tasks, including both hard and super hard difficulty levels, as shown in Figure 3. The results demonstrate that **CEE** consistently achieves higher sample efficiency compared to both Soft-QMIX and QMIX across all tasks. Notably, in the challenging 3s5z vs 3s6z scenario, **CEE** exhibits superior performance in terms of both sample efficiency and asymptotic performance, validating the effectiveness of **CEE** in traditional MARL tasks.

SMAC-v2. To provide a more comprehensive analysis, we conduct experiments on the more challenging SMAC-v2 environment across protoss, terran and zerg scenarios. The

MAP	Setting	VDN	QMIX	QPLEX	HPN-QMIX	Soft-QMIX	RGP	CEE (Ours)
protoss	**5v5**	58.8±5.2	68.7±7.8	68.8±6.9	79.4±4.8	76.0±4.0	76.2±3.6	**80.1±3.7**
terran	**5v5**	65.1±6.6	65.6±5.9	59.4±9.1	78.0±5.2	72.0±4.0	75.0±5.9	**79.7±5.3**
zerg	**5v5**	51.6±7.1	62.7±5.3	65.6±9.4	67.4±7.0	71.8±4.0	61.4±7.1	**78.1±5.1**
protoss	**10v10**	43.8±9.2	55.4±4.6	62.5±6.2	78.1±6.3	71.9±3.1	-	**79.3±3.2**
terran	**10v10**	46.9±8.3	65.0±6.3	59.4±3.2	78.0±5.6	73.4±4.7	-	**81.5±6.2**
zerg	**10v10**	59.4±6.2	62.5±6.8	65.7±3.1	68.8±9.3	68.4±5.6	-	**74.8±4.1**
protoss_30°	**5v5**	7.3±5.2	7.2±1.6	9.4±3.1	8.8±3.6	11.1±1.4	10.1±5.0	**16.6±5.2**
terran_30°	**5v5**	2.5±2.0	5.2±2.0	9.3±6.2	4.4±2.8	9.4±9.3	9.4±1.2	**12.5±6.3**
zerg_30°	**5v5**	1.9±1.9	2.1±0.6	6.3±3.0	2.5±2.3	6.3±3.2	4.2±1.1	**7.8±1.6**
protoss_90°	**5v5**	18.0±3.9	27.6±8.2	48.9±8.3	47.5±5.8	50.9±10.6	37.2±6.5	**56.3±9.4**
terran_90°	**5v5**	21.1±8.6	41.9±5.3	46.9±12.4	49.5±7.4	53.1±13.2	46.3±4.1	**54.7±10.9**
zerg_90°	**5v5**	20.8±4.2	23.2±5.6	45.3±7.6	17.1±5.3	46.9±18.1	30.0±4.7	**48.8±7.2**
Average		33.1	40.6	45.6	48.3	50.9	-	**55.9**

Table 1: The eval win rates (%) on SMAC-v2 tasks, including 5v5, 10v10, 30°, and 90° field of view. We bold the best scores, and underline second-best results, ± is the standard deviation.

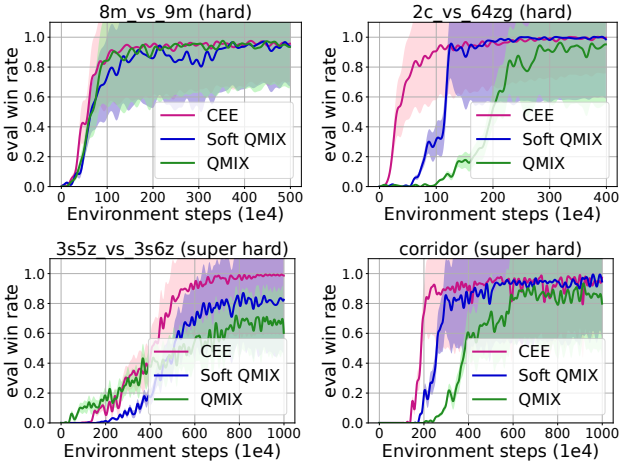

Figure 3: The learning curves on 4 SMAC tasks.

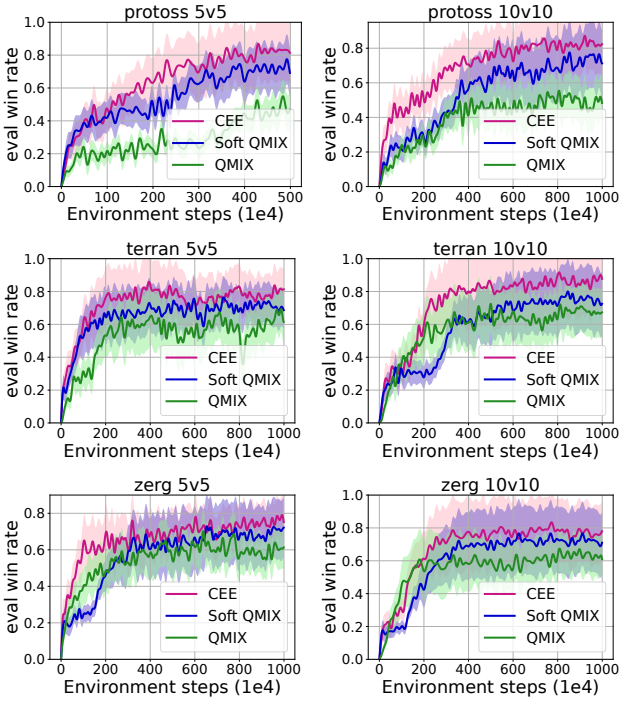

Figure 4: The learning curves on 6 SMAC-v2 tasks.

main results are presented in Table 1. In 6 tasks comprising 5v5 and 10v10 scenarios, **CEE** obtains the best win rates in all 6 tasks when compared against other baseline methods.

The advantages of **CEE** become even more pronounced under partially observable conditions, where efficient exploration becomes crucial for policy learning. In the highly constrained 30-degree field of view setting, **CEE** consistently outperforms all baselines across all three tasks, achieving win rates of 16.6%, 12.5%, and 7.8% for protoss, terran, and zerg respectively. These results represent significant improvements of 5.5%, 3.1%, and 1.5% over the second-best performing methods, demonstrating the framework's effectiveness in handling limited observability. Under the 90-degree observation range, **CEE** continues to demonstrate strong performance, achieving the best results in all 3 tasks with win rates of 56.3%, 54.7% and 48.8% respectively. Overall, CEE achieves the best performance in all 12 evaluated scenarios with an average win rate of 55.9%.

Furthermore, the comparative learning curves across 6 tasks shown in Figure 4 demonstrate that **CEE** exhibits higher sample efficiency compared to both Soft-QMIX and QMIX.

GRF. We further conduct experiments on 2 GRF tasks, comparing CEE against the causal MARL method LAIES, QMIX, and VDN, as shown in Figure 5. The results demonstrate that CEE achieves the best win rates of 63.6% and 70.6% across 2 tasks, respectively. Notably, CEE outper-

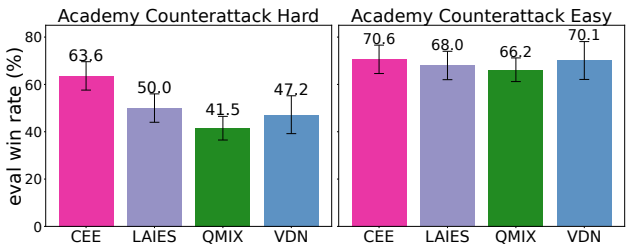

Figure 5: Results on 2 GRF tasks.

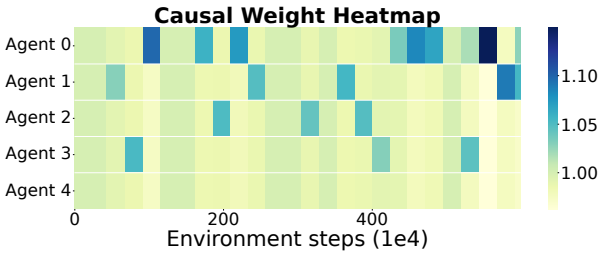

Figure 6: Causal weight heatmap of **CEE** in zerg 5v5 task.

forms the causal MARL method LAIES, which specifically addresses causal relationships in multi-agent settings, thereby highlighting the effectiveness of our proposed causal framework in multi-agent environments.

Property Analysis

Causal Weight Analysis. To better understand the causal influence during the learning process, we visualize the causal weights for the zerg 5v5 task in the SMAC-v2 environment, as shown in Figure 6. The analysis reveals that agents 0 and 1 maintain higher causal weights throughout the learning process, which encourages their exploration behavior. In contrast, agent 4 exhibits the lowest causal weight among all agents, indicating its relatively lower influence on the team's overall performance.

Ablation Study. To validate the effectiveness of two component in **CEE**, we conduct ablation experiments across 3 SMAC-v2 tasks, as shown in Figure 7. The results reveal that **CEE** without (w/o) causal state filtering (CS) achieves suboptimal performance compared to **CEE** w/o causal action exploration (CA) and the baseline QMIX method, demonstrating the significant impact of the causal action exploration module on model performance. Compared to the CS component, the CA module exhibits a more substantial influence on the overall performance. Notably, the presence of either component individually ensures that the model performance remains superior to QMIX method.

Computation Burden and Hyperparameter Analysis. We examine the computational burden of **CEE** across three SMAC-v2 tasks, as illustrated in Figure 8. The analysis demonstrates that on protoss and terran tasks, **CEE** exhibits elevated computational requirements compared to the other three baseline methods, yet the additional overhead remains

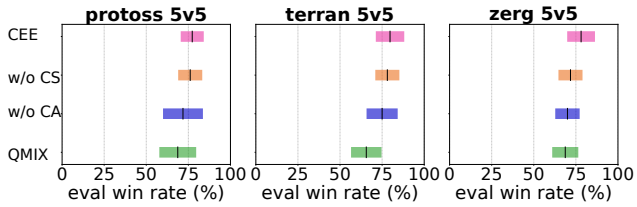

Figure 7: Ablation study on 3 SMAC-v2 tasks.

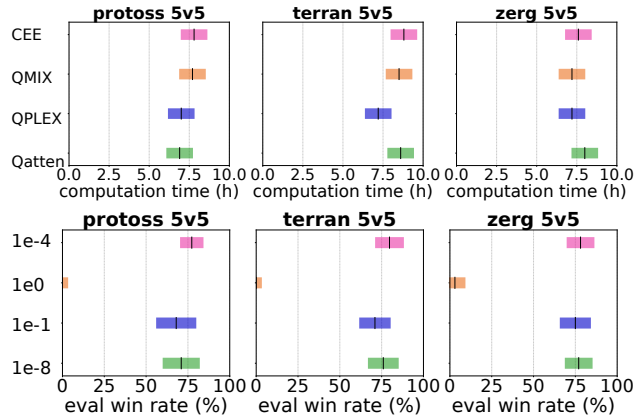

Figure 8: Computation time and hyperparameter analysis.

constrained within a one-hour interval. Conversely, we observe that on the zerg task, **CEE** manifests reduced computational demands relative to Qatten. These findings substantiate that while the integration of causal modules introduces computational overhead, this cost remains negligible and is readily justified by the considerable performance enhancements obtained. We conduct experiments to analyze the performance of **CEE** under different values of temperature factor α. The results demonstrate that the value of 1e-4 achieves the best performance across these 3 tasks. Moreover, under the values of 1e-1 and 1e-8, **CEE** still maintains above 70% win rates, which further validates that our method is robust.

Conclusion

In this work, we propose a causality-aware framework for efficient exploration in cooperative MARL, incorporating state-reward causal masking mechanisms for filtering causally-informed state feature and agent-reward causal entropy for causality-guided agents exploration. We conduct comprehensive experiments across 3 diverse environments, encompassing 21 tasks from SMAC, SMAC-v2, to GRF, demonstrating the superior performance and high sample efficiency of the proposed approach.

Limitation and Future Work. The primary limitation of our framework is the lack of explicit modeling of causal dependencies between agents and states. This issue may limit the framework's ability to capture the complex causal structure. Future work will incorporate agent-observation causality to improve sample efficiency and generalization in cooperative MARL.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 625B2085, Grant 62506157, Grant 62276128, Grant 62192783; in part by the Jiangsu Science and Technology Major Project BG2024031; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20243051; the Fundamental Research Funds for the Central Universities(14380128); in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Cao, H.; Feng, F.; Fang, M.; Dong, S.; Yang, T.; Huo, J.; and Gao, Y. 2025a. Towards Empowerment Gain through Causal Structure Learning in Model-Based RL. In *ICLR*.

Cao, H.; Feng, F.; Yang, T.; Huo, J.; and Gao, Y. 2025b. Causal Information Prioritization for Efficient Reinforcement Learning. In *The Thirteenth International Conference on Learning Representations*.

Chen, W.; Huang, S.; and Schneider, J. 2024. Soft-QMIX: Integrating Maximum Entropy For Monotonic Value Function Factorization. *arXiv preprint arXiv:2406.13930*.

Du, X.; Ye, Y.; Zhang, P.; Yang, Y.; Chen, M.; and Wang, T. 2024. Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17362–17370.

Ellis, B.; Cook, J.; Moalla, S.; Samvelyan, M.; Sun, M.; Mahajan, A.; Foerster, J.; and Whiteson, S. 2023. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 37567–37593.

Eysenbach, B.; and Levine, S. 2021. Maximum entropy RL (provably) solves some robust RL problems. *arXiv preprint arXiv:2103.06257*.

Feng, F.; Huang, B.; Zhang, K.; and Magliacane, S. 2022. Factored adaptation for non-stationary reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31957–31971.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Grimbly, S. J.; Shock, J.; and Pretorius, A. 2021. Causal multi-agent reinforcement learning: Review and open problems. *arXiv preprint arXiv:2111.06721*.

Gu, S.; Kuba, J. G.; Chen, Y.; Du, Y.; Yang, L.; Knoll, A.; and Yang, Y. 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319: 103905.

Guestrin, C.; Koller, D.; Parr, R.; and Venkataraman, S. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19: 399–468.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hao, J.; Yang, T.; Tang, H.; Bai, C.; Liu, J.; Meng, Z.; Liu, P.; and Wang, Z. 2023. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7): 8762–8782.

Hu, J.; Jiang, S.; Harding, S. A.; Wu, H.; and Liao, S.-w. 2021. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.03479*.

Huang, B.; Lu, C.; Leqi, L.; Hernández-Lobato, J. M.; Glymour, C.; Schölkopf, B.; and Zhang, K. 2022. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, 9260–9279. PMLR.

Huang, J.; Xu, Y.; Wang, Q.; Wang, Q. C.; Liang, X.; Wang, F.; Zhang, Z.; Wei, W.; Zhang, B.; Huang, L.; et al. 2025. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, 3040–3049. PMLR.

Ji, T.; Liang, Y.; Zeng, Y.; Luo, Y.; Xu, G.; Guo, J.; Zheng, R.; Huang, F.; Sun, F.; and Xu, H. 2024. ACE: Off-Policy Actor-Critic with Causality-Aware Entropy Regularization. In *Forty-first International Conference on Machine Learning*.

Jianye, H.; Hao, X.; Mao, H.; Wang, W.; Yang, Y.; Li, D.; Zheng, Y.; and Wang, Z. 2022. Boosting multiagent reinforcement learning via permutation invariant and permutation equivariant networks. In *The Eleventh International Conference on Learning Representations*.

Kearns, M.; and Koller, D. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, 740–747.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A. A. A.; Yogamani, S.; and Pérez, P. 2022. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 4909–4926.

Kraemer, L.; and Banerjee, B. 2016. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190: 82–94.

Kurach, K.; Raichuk, A.; Stanczyk, P.; Zajkac, M.; Bachem, O.; Espeholt, L.; Riquelme, C.; Vincent, D.; Michalski, M.; Bousquet, O.; et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4501–4510.

Lin, C.; Zhang, Y.; Han, G.; Lu, C.; Zhu, S.; and Wang, S. 2025. Multiple Autonomous Underwater Vehicles-Assisted Data Collection in 6G-Driven Underwater Wireless Networks Based on Software-Defined MARL. *IEEE Transactions on Intelligent Transportation Systems*, 1–12.

Liu, B.; Pu, Z.; Pan, Y.; Yi, J.; Liang, Y.; and Zhang, D. 2023. Lazy agents: A new perspective on solving sparse reward

problem in multi-agent reinforcement learning. In *International Conference on Machine Learning*, 21937–21950. PMLR.

Liu, Y.; Huang, B.; Zhu, Z.; Tian, H.; Gong, M.; Yu, Y.; and Zhang, K. 2024. Learning World Models with Identifiable Factorization. *Advances in Neural Information Processing Systems*, 36.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pina, R.; De Silva, V.; and Artaud, C. 2025. Discovering causality for efficient cooperation in multi-agent environments. *Neurocomputing*, 130358.

Qifan, L.; Shan, Y.; Liu, H.; Zhu, Z.; Long, T.; Zhang, W.; and Tian, Y. 2025. Reconstruction-Guided Policy: Enhancing Decision-Making through Agent-Wise State Consistency. In *The Thirteenth International Conference on Learning Representations*.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178): 1–51.

Samvelyan, M.; Rashid, T.; De Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvarinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; Bollen, K.; and Hoyer, P. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr): 1225–1248.

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, prediction, and search*. MIT press.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2020. Qplex: Duplex dueling multi-agent q-learning. *International Conference on Learning Representations*.

Wang, Z.; Du, Y.; Zhang, Y.; Fang, M.; and Huang, B. 2025. MACCA: Offline Multi-agent Reinforcement Learning with Causal Credit Assignment. *Transactions on Machine Learning Research*.

Xu, J.; Zhong, F.; and Wang, Y. 2020. Learning multi-agent coordination for enhancing target coverage in directional sensor networks. *Advances in Neural Information Processing Systems*, 33: 10053–10064.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35: 24611–24624.

Zhang, Y.; Du, Y.; Huang, B.; Fang, M.; and Pechenizkiy, M. 2024. A Causality-Inspired Spatial-Temporal Return Decomposition Approach for Multi-Agent Reinforcement Learning. In *NeurIPS 2024 Causal Representation Learning Workshop*.

Zheng, Z.; and Gu, S. 2025. Safe Multiagent Reinforcement Learning With Bilevel Optimization in Autonomous Driving. *IEEE Transactions on Artificial Intelligence*, 6(4): 829–842.