TMAE: Learning Targeted Multi-Agent Exploration via Causal Inference

Chuxiong Sun[1], Dunqi Yao[1,2], Rui Wang[1,3], Wenwen Qiang[1], Changwen Zheng[1], Jiangmeng Li[1†]

[1]National Key Laboratory of Space Integrated Information System, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]National Key Laboratory of Complex System Modeling and Simulation Technology
{chuxiong2016, dunqi2021, wangrui, qiangwenwen, changwen, jiangmeng2019}@iscas.ac.cn

Abstract

Exploration in sparse-reward tasks remains a fundamental challenge in multi-agent reinforcement learning (MARL) due to complex inter-agent interactions and the expansive exploration space. To address this issue, we propose Targeted Multi-Agent Exploration (TMAE), a novel framework that uncovers the causal relationships between the state space and the reward function, thereby reducing the exploration space and enabling more targeted exploration. Specifically, we construct a structural causal model (SCM) to model the causality between sub-state variables and sparse rewards, providing a robust analytical foundation for subsequent causal inference. Through counterfactual causal intervention, TMAE identifies the most critical subspaces for discovering rare but pivotal events while filtering out confounders. By incorporating these causal insights into the exploration process, TMAE prioritizes subspaces with stronger causal effects on sparse rewards, significantly enhancing exploration efficiency. We evaluate TMAE on a range of MARL benchmarks featuring sparse rewards, consistently demonstrating superior exploration efficiency compared to state-of-the-art methods. Furthermore, visualized causal insights derived from TMAE reveal its ability to effectively capture intricate dependencies and priorities in targeted exploration, showcasing strong alignment with prior domain knowledge.

Introduction

Multi-agent reinforcement learning (MARL) has emerged as a crucial field due to its ability to address a wide range of complex real-world problems, including applications in Game AI (Peng et al. 2017), Traffic Control (Xu et al. 2021), and Network Routing (Ye, Zhang, and Yang 2015). Its notable success is predominantly attributed to the Centralized Training and Decentralized Execution (CTDE) paradigm (Lowe et al. 2017; Rashid et al. 2018; Yu et al. 2022; Sun et al. 2021, 2024b,a, 2025; Li et al. 2024c, 2025), which effectively alleviates non-stationarity and preserves policy scalability. However, exploration remains a significant obstacle in MARL, especially in sparse reward tasks.

To tackle this challenge, recent approaches have focused on augmenting extrinsic rewards with intrinsic rewards derived from the global state. Such intrinsic rewards are specifically designed to guide agents toward influential behaviors (Wang et al. 2019; Li et al. 2024a), novel states (Zheng et al. 2021; Xu et al. 2023a; Zhang et al. 2023; Zang et al. 2025), and diverse trajectories (Mahajan et al. 2019; Li et al. 2021; Bettini, Kortvelesy, and Prorok 2024; Yang et al. 2023). However, as the dimensionality of the joint exploration space grows exponentially with the number of agents, global intrinsic rewards often fail to manage the large search space and complex agent interactions effectively.

In practice, although the state space may be enormous, sparse rewards and rare events often depend on only a small subset of the state space. For example, in a multi-dimensional state space, the reward for unlocking a door might depend on just one dimension—whether the agents possess the key—demonstrating how certain subspaces are critical for successfully exploring sparse rewards. If task-relevant prior knowledge (i.e., identifying the subspaces on which sparse rewards depend) is available, exploration efficiency can be significantly improved (Liu et al. 2021). However, acquiring such knowledge in real-world scenarios is often challenging or impractical. Hence, existing methods often rely on black-box techniques or approximate, task-agnostic priors as substitutes. For example, CMAE (Liu et al. 2021) employs dimensionality reduction techniques to project high-dimensional states into a lower-dimensional latent space. FOX (Jo et al. 2024) explores the formation space defined by observation differences rather than the entire state space. Furthermore, SAME (Xu et al. 2023b) assumes that sub-state spaces with higher uncertainty are more relevant to the reward function and encourage agents to explore uncertain subspaces. While these approaches can reduce the exploration space, such black-box techniques or one-size-fits-all priors fail to capture the true reward structure needed to accelerate exploration effectively.

In this work, we propose Targeted Multi-Agent Exploration (TMAE), a novel framework designed to automatically extract task-relevant prior knowledge from past exploration data by studying the causal relationships between the state space and sparse rewards. Concretely, we decompose the state space and construct a structural causal model (SCM) that links each sub-state variable to the sparse reward. We then utilize powerful causal inference tools—counterfactual causal intervention—to pinpoint how each subspace influences the sparse reward. This discovered causal knowledge is subsequently injected into the

exploration process, driving agents to concentrate on the most influential subspaces. By reducing the exploration space and focusing on key events, TMAE significantly improves exploration efficiency. We evaluate TMAE on various MARL environments, including Google Research Football (GRF) (Kurach et al. 2020) and the StarCraft Multi-agent Challenge (SMAC) (Samvelyan et al. 2019), under sparse-reward settings. Compared with state-of-the-art multi-agent exploration algorithms such as FOX (Jo et al. 2024), ICES (Li et al. 2024b), CDS (Li et al. 2021), COIN (Li et al. 2024a) and SMMAE (Zhang et al. 2023), our method demonstrates superior exploration efficiency and performance.

To the best of our knowledge, TMAE is the first approach to investigate the causal relationship between sparse rewards and sub-state spaces, thereby effectively uncovering task-relevant exploration knowledge and enabling efficient, targeted multi-agent exploration.

Related works

Multi-agent exploration

Intrinsic rewards, such as curiosity (Burda et al. 2018; Badia et al. 2020) and diversity (Eysenbach et al. 2018), are metrics used to assess the significance of explored samples in RL and have proven highly effective in single-agent environments. However, in the MARL domain (Lowe et al. 2017; Sunehag et al. 2017; Rashid et al. 2018; Wang et al. 2021; Yu et al. 2022), additional challenges emerge from partial observability and intricate inter-agent relationships. Therefore, a critical issue in multi-agent exploration is the application of intrinsic rewards to either global states or local observations (i.e., exploring from either the global perspective or the individual perspectives of multiple agents). This perspective leads us to categorize existing methods for multi-agent exploration into global exploration and self-exploration.

Self-exploration incorporates intrinsic objectives at the individual agent level. Concretely, EMC (Zheng et al. 2021) employs prediction errors of individual Q-values as intrinsic rewards. SMMAE (Zhang et al. 2023) utilizes state marginal matching to expand the exploration space of each agent. ADER (Kim and Sung 2023) evaluates the necessary degree of exploration for each agent and determines the optimal target entropy to drive maximum entropy exploration. Social influence (Jaques et al. 2019) measures how an agent's actions can affect the actions of others, thereby capturing the dependencies between their exploration policies. EITI and EDTI (Wang et al. 2019) quantify the influence of one agent's behavior on the transition dynamics and the expected returns of other agents. CDS (Li et al. 2021) promotes diverse individualized behaviors by leveraging mutual information between agents' identities and their local trajectories.

Global exploration seeks to encourage comprehensive exploration of the state space. For instance, MAVEN (Mahajan et al. 2019) maximizes the mutual information between global trajectories and latent variables to generate a variety of global behaviors. HMASD (Yang et al. 2023) advances a skill discovery approach that develops team-wide skills, fostering diverse trajectories from both an individual and global perspective. MAGIC (Chen et al. 2022) adopts a goal-oriented, multi-stage model to improve goal cognition, helping agents grasp tasks at a goal level and enabling cooperative global exploration. Moreover, other methodologies (Iqbal and Sha 2019; Chitnis et al. 2020) create heuristic intrinsic rewards to incentivize desirable collective behaviors in multi-agent systems.

In conclusion, self-exploration provides a straightforward and scalable approach but may lead to less coordinated exploration behaviors among agents. In contrast, global exploration can promote cooperative behaviors, yet it faces challenges such as exploration space explosion, particularly in complex tasks involving a large number of agents. *In this work, we aim to leverage the strengths of both approaches, promoting efficient cooperative exploration while ensuring scalability.* Our study is closely aligned with efforts that focus on reducing exploration space (Liu et al. 2021; Xu et al. 2023a; Jo et al. 2024). However, existing methods such as CMAE (Liu et al. 2021) and FOX (Jo et al. 2024) focus only on using techniques such as projection and formation to map the high-dimensional state space to a low-dimensional latent space. These naive projections lack semantic information, proving inefficient in complex MARL tasks. SAME (Xu et al. 2023a) uses the principle of optimism in the face of uncertainty (Strehl and Littman 2008) to assess the importance of each subspace, but this is only a prior-based approximation and does not fundamentally explore the impact of different subspaces on return or exploration efficiency. To the best of our knowledge, TMAE represents the first attempt to directly study the relationship and influence of exploring different subspaces.

Causal inference in RL

Our work is also related to approaches that incorporate causal inference(Wang et al. 2024; Song et al. 2024; Zhang et al. 2024) into reinforcement learning (RL) (Zeng et al. 2024). Notably, (Liu et al. 2023) represents an early attempt to leverage causal inference in multi-agent RL (MARL) tasks with sparse rewards by computing the causal effect of actions on states to mitigate the lazy agent issue. Despite its effectiveness, (Liu et al. 2023) relies on a strong assumption: it requires prior knowledge to identify the critical parts of the state space and then focuses on how agents' behaviors affect these important state components. However, acquiring such prior knowledge is inherently difficult, and discovering essential state components is itself a fundamental challenge in sparse-reward scenarios. Consequently, we propose TMAE, which aims to reduce the exploration space and improve exploration efficiency by uncovering the causal links between states and rewards.

Background

Problem Setting

In this work, we consider sparse reward multi-agent tasks modeled by Decentralized Partially Observable Markov Decision Process (Dec-POMDP). Dec-POMDP is typically defined by a tuple $G = (N, S, O, A, \mathbb{O}, P, R, \gamma)$. In this formulation: $N = (agent_1, ..., agent_n)$ depicts the collec-

tive of agents, n denotes the number of agents. S encompasses global states, offering a comprehensive environmental overview. O refers to the accessible local observations. A signifies a set of available actions. $\mathbb{O}$ refers to the observation function, which describes how agents perceive the environment based on the global state. P acts as the transition function, illustrating environmental dynamics. R is a reward function contingent on global states and joint actions. γ represents the discount factor. The fundamental goal of multi-agent exploration is to motivate agents to collaboratively engage in novel behaviors, steering the environment towards novel global states.

Causal Inference

Causal Inference (PEARL 1988) provides a theoretical framework to understand and quantify cause-and-effect relationships in complex systems. It employs tools such as SCM and probabilistic reasoning to model these relationships. In causal inference, uppercase letters, such as X, denote random variables, while lowercase letters, such as x, represent specific outcomes of those random variables.

Intervention (Pearl 1993) is crucial in causal inference, denoted as $do(X = x)$ or $do(x)$. Intervention enables the identification of direct causal effects between variables without being influenced by confounding factors. A causal effect can be expressed as $P(y|do(X = x))$, where y denotes a potential outcome of Y and x is the intervened value of X, representing the direct causal effect of setting X to x on Y.

Counterfactual Intervention (Pearl 1994) is another essential tool in causal inference, allowing object variables to take counterfactual values with minimal influence on the system, thus preserving the causal relationship of other factors. It enables the estimation of probability distributions under counterfactual situations. Counterfactual probability can be denoted as $P(Y_{X=x} = y)$, describing probability of $Y = y$ had X been x. In this paper, for simplicity, we abbreviate the counterfactual probabilities $P(R_{A=a^*} = r)$ as $P_{a^*}(r)$.

METHODS	SUCCESS	SUCCESSFUL EPISODES	WIN RATE
QMIX	N	0	0%
CDS	Y	12	0%
COIN	N	0	0%
EMC	Y	153	5.2%
ICES	Y	233	3.2%
SMMAE	Y	72	0%
RND	Y	201	4.8%

Table 1: Statistical information by training 60000 time steps on *push box*. Success indicates whether the agents have finished the task while exploring. Successful episodes indicate how many episodes have the agents finished the task while exploring. Win rate indicates the test won rate after training.

Methodology

Revisiting challenge in multi-agent exploration

As illustrated in Table 1, we observe an intriguing phenomenon: certain exploration algorithms can collect a small amount of successful exploration data in tasks with sparse rewards by designing intrinsic rewards that encourage agents to identify influential behaviors, novel states, and diverse trajectories. However, they fail to effectively utilize these successful demonstrations to learn exploration knowledge, ultimately hindering the formation of stable exploration policies. Therefore, we pose a more fundamental question:

Can we glean insights from both successful and unsuccessful exploration data to enhance subsequent exploration efficiency?

Capturing sparse reward dependencies from successful exploration data

To enhance exploration efficiency, we model the exploration process as two alternating steps. The first step involves standard intrinsic motivated exploration, where samples are collected through interactions with the environment. The second step analyzes previously explored data to determine factors contributing to successful or unsuccessful outcomes. Specifically, this step identifies the importance of each subspace within the state space concerning sparse reward signals. These importance weights subsequently guide targeted exploration, converting successful exploration experiences into actionable exploration knowledge. However, effectively extracting knowledge from explored data remains a key challenge.

Fortunately, causal inference provides a theoretical framework that quantifies cause-and-effect relationships. To glean insights from both successful and unsuccessful past experiences, we first decompose the global state space S into distinct sub-state space, $S = (S^1, ..., S^M)$. This decomposition enables more granular analysis of which regions in the state space lead to successful or failed exploration. Then, we propose a SCM to aid in this analysis. As shown in Fig. 1, H represents historical state trajectory, in order to analyze the causal effect of a specific S^i on sparse rewards, we decompose the state space within the SCM into two components S^i and S^{-i}, where S^{-i} represents the remaining sub-state space. Then, we provide the formal definition of the **Average Causal Effect (ACE)** of a sub-state space on reward.

Given sub-state variable S^i and reward variable R, the probability of $R = r$ with intervening $S^i = s^i_\dagger$, denoted as $P(r|do(S^i = s^i_\dagger))$, describes the impact of setting S^i to a specific value $s^i_\dagger$ on the distribution of R. The Average Causal Effect (ACE) of S^i on R, denoted as $ACE(S^i, R)$, quantifies the expected difference in reward caused by interventions on the sub-state across its potential values. It is formally defined as follows:

$$ACE(S^i, R) = \frac{1}{N_{S^i}} \sum_{s^i} [D_{KL}(P(r|do(S^i = s^i))||P(r))]. \tag{1}$$

where N_{S^i} is the number of sub-state in subspace i. According to the definition of ACE, we need to compute

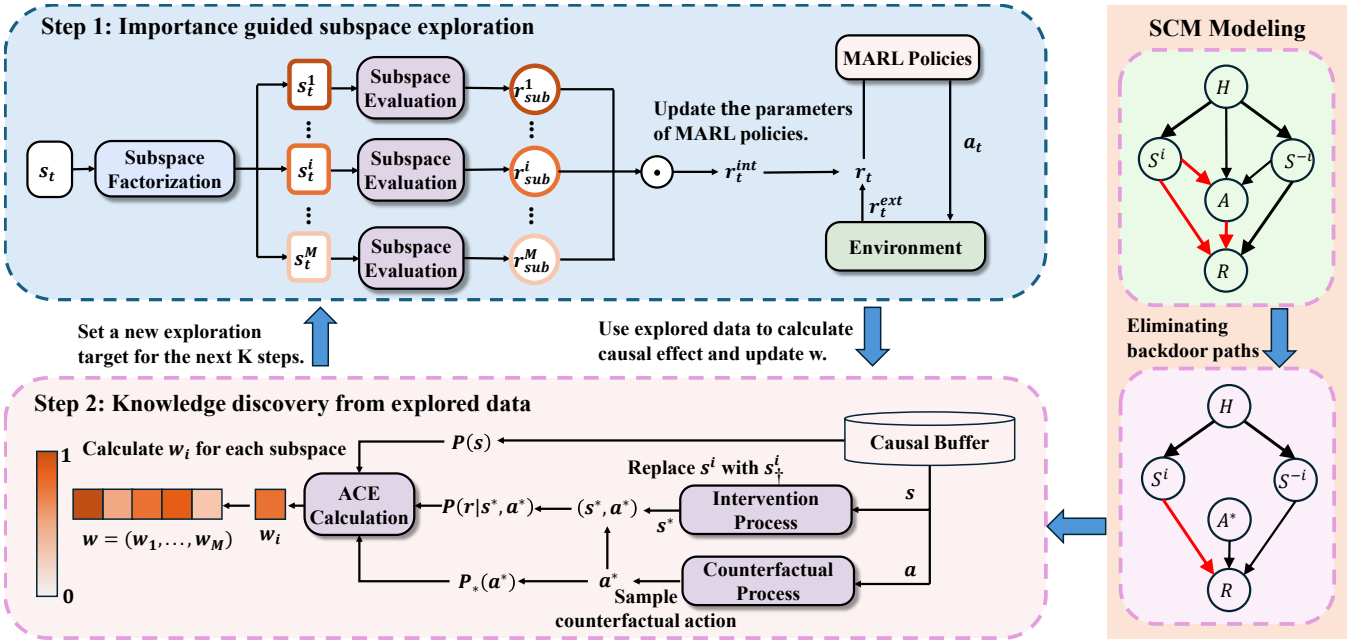

Figure 1: Framework of TMAE. We first construct an SCM by eliminating backdoor paths in the traditional Dec-POMDP SCM to analyze causal effects on obtaining sparse rewards. Then, we compute the Average Causal Effect of each subspace S^i on reward R, denoted as ACE(S^i, R), to quantify sub-state space importance. Finally, the derived importance values guide subsequent exploration, where darker border colors of each r^i_{sub} (upper left) indicate higher exploration priority.

$P(r|do(S^i = s^i)$ for each potential sub-state s^i. However, in the original SCM of Dec-POMDP, backdoor paths such as $S^i \leftarrow H \rightarrow A \rightarrow R$ exist between S^i and R. These paths introduce indirect causal effects from H and S^{-i} into the computation, leading to confounding. Typically, backdoor or frontdoor adjustment criteria are employed to block the influence of such paths. However, in the original SCM, H is highly complex and typically unobservable, making it impossible to apply these criteria directly. Consequently, an alternative computational formula is required that satisfies the following conditions:

- No Interventional Probabilities: The formula should avoid requiring direct access to interventional probabilities.
- No Unobservable Variables: The computation must not rely on unobservable variable H.

To address these challenges, we reapply the chain rule of probability to decompose the total probability and derive the following expression:

$$\begin{aligned} &P(r|do(S^i = s^i_\dagger)) \\ &= \sum_{a,s^{-i}} P(r|a, s^i_\dagger, s^{-i}) \sum_h P(a|s^i_\dagger, s^{-i}, h) P(s^{-i}, h). \end{aligned} \tag{2}$$

where $s^i_\dagger$ represents the intervention value of variable S^i. However, computing the term $\sum_h P(a \mid s^i_\dagger, s^{-i}, h), P(s^{-i}, h)$ is still non-trivial because it still relies on variable H.

By performing counterfactual interventions on A, we can effectively eliminate dependence on H. This is because counterfactual interventions only modify the causal relationships related to the object, leaving unrelated causal relationships unchanged. For example, counterfactual intervention $A = a^*$ will change the indirect causal effect of $S^i \rightarrow A \rightarrow R$, while the direct causal effect of $S^i \rightarrow R$ will stay unchanged. Thus, we can modify the SCM and preserve the direct causal effect of $S^i \rightarrow R$ by performing a counterfactual intervention on action distribution. For simplicity, we replace the original action distribution with a fixed prior distribution, $P_*(a^*)$, which is independent of H, S^i, and S^{-i}. This modification removes the dependencies between A and all of its causes in the SCM, as illustrated in Fig.1. Since A^* is the only object of counterfactual intervention, the distributions of other variables remain unchanged. This allows the derivation of counterfactual intervention probability as follows:

$$\begin{aligned} &P(r|do(S^i = s^i_\dagger)) \doteq \sum_{a^*} P_*(a^*) P_{a^*}(r|do(S^i = s^i_\dagger)) \\ &= \sum_{a^*} P_*(a^*) \sum_{s^{-i}} P(r|a^*, s^i_\dagger, s^{-i}) \sum_h P(s^{-i}, h) \\ &= \sum_{a^*} P_*(a^*) \sum_{s^{-i}} P(r|a^*, s^i_\dagger, s^{-i}) P(s^{-i}) \\ &= \sum_{a^*} P_*(a^*) \sum_s P(r|a^*, s^*) P(s). \end{aligned} \tag{3}$$

where $P_*(a^*)$ is the counterfactual action distribution, $P_{a^*}(r|do(S^i = s^i_\dagger)$ represents counterfactual intervention probability, s^* represents the modified state replacing the i-th component of global state with $s^i_\dagger$. At this point, the de-

pendency on H has been eliminated, allowing us to compute the causal effect of S^i on R using only observable information. We have $P_*(a^*)$ and $P(s)$ can be extracted by count-based methods such as Hash-Count (Tang et al. 2017). Subsequently, we train a reward model to approximate the reward distribution. Using modified states s^* and counterfactual intervened actions a^* as inputs to the reward model, we estimate $P(r|a^*, s^*)$, the distribution necessary for computing the causal effect.

Leveraging learned knowledge to drive target exploration

As shown in Fig.1, the high-level motivation of TMAE is to first leverage causal inference to uncover the impact of each subspace on rewards and then use these causal insights as subsequent exploration targets. These insights guide agents to focus their exploration on critical subspaces, reducing the overall exploration space and enhancing efficiency. After analyzing the relationship between sparse rewards and sub-state spaces, it is crucial to develop an efficient mechanism to integrate these causal insights into the exploration process. Hence, we propose a simple yet effective sub-space exploration framework.

First, we model the intrinsic reward for each sub-state. Given a global state s_t and its decomposed sub-states, $(s_t^1, ..., s_t^M)$, the subspace intrinsic reward can be formulated as follows:

$$r_t^i = f(s_t^i) \tag{4}$$

where s_t^i represents the i-th sub-state and f denotes the subspace intrinsic reward function. In this work, we employ Hash-Count (Tang et al. 2017) as the method for evaluating intrinsic rewards in each sub-state for two main reasons: at first, state-of-the-art multi-agent intrinsic reward methods, such as EMC (Zheng et al. 2021) and CDS (Li et al. 2021), are primarily designed based on global states. It remains unclear how these methods could be effectively incorporated into subspace exploration, and their performance in such settings is uncertain. Furthermore, a simple intrinsic reward mechanism, such as Hash-Count, is sufficient to identify exploration-promising data.

Then, we propose to utilize the ACE of the sub-state space S^i on the reward R, which encapsulates the dependency between the reward function and the subspace, to model the priority of exploring S^i. A subspace with a stronger causal effect on the sparse reward should be given higher priority in subsequent exploration. The exploration target and priority can be formally defined as follows:

$$w_i = ACE(S^i, R) \tag{5}$$

Then, we incorporate causal knowledge by employing a simple weighted sum of subspace intrinsic rewards. This process can be formally defined as follows:

$$r_t^{int} = \sum_i w_i r_t^i \tag{6}$$

It is worth noting that TMAE can be integrated with any knowledge injection method. However, we find that the proposed simple approach is both efficient and effective for the tasks considered in this study. Furthermore, to ensure stable exploration, we perform causal inference every K steps to obtain a new weight vector w.

Experiments

In this section, we evaluate TMAE across various benchmarks to address the following questions:

Q1. Can TMAE efficiently explore sparse-reward multi-agent tasks, and how does its exploration efficiency compare to other state-of-the-art methods?

Q2. How does each core component of TMAE contribute to its overall performance?

Q3. What valuable knowledge and insights has TMAE discovered through the exploration process?

Experimental Setup

Benchmarks. To investigate **Q1**, we evaluate TMAE on two challenging MARL benchmarks: Google Research Football (GRF) (Kurach et al. 2020) and StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al. 2019). Specifically, for GRF, we use three scenarios: academy_3_vs_1_with_keeper, academy_corner, and academy_counterattack_hard. For SMAC, we use five scenarios: 8m_vs_9m, 5m_vs_6m, MMM2, 6h_vs_8z, and 3s5z_vs_3s6z. We configured all test scenarios under a sparse-reward setting, presenting a more challenging exploration task. These configurations reduce the frequency of reward signals, thereby significantly increasing the difficulty of exploration. In SMAC, agents receive rewards only when units on either side are eliminated or when the episode terminates in victory or defeat. In GRF, rewards are provided solely upon winning or losing the match. The environmental settings are consistently applied across all experiments, including those for TMAE, baseline methods, and ablation studies.

Baselines. To evaluate TMAE's exploration efficiency, we compare it against a diverse set of state-of-the-art multi-agent exploration methods, including FOX (Jo et al. 2024), ICES (Li et al. 2024b), CDS (Li et al. 2021), COIN (Li et al. 2024a) and SMMAE (Zhang et al. 2023). These baselines encompass both self-exploration and global exploration strategies, ranging from intrinsic reward-based approaches—such as curiosity, diversity, and influence—to methods specifically designed to reduce the exploration space. By comparing TMAE with these advanced techniques, we aim to provide a comprehensive analysis of TMAE's overall performance and its exploration efficiency.

Exploration Performance and Efficiency

We begin our analysis by comparing the performance of TMAE with various baselines across multiple benchmark scenarios. As depicted in Fig.2, TMAE outperforms state-of-the-art baselines across all scenarios, showing a substantial advantage in exploration efficiency in six of the most challenging exploration scenarios and achieving comparable performance in two simpler scenarios.

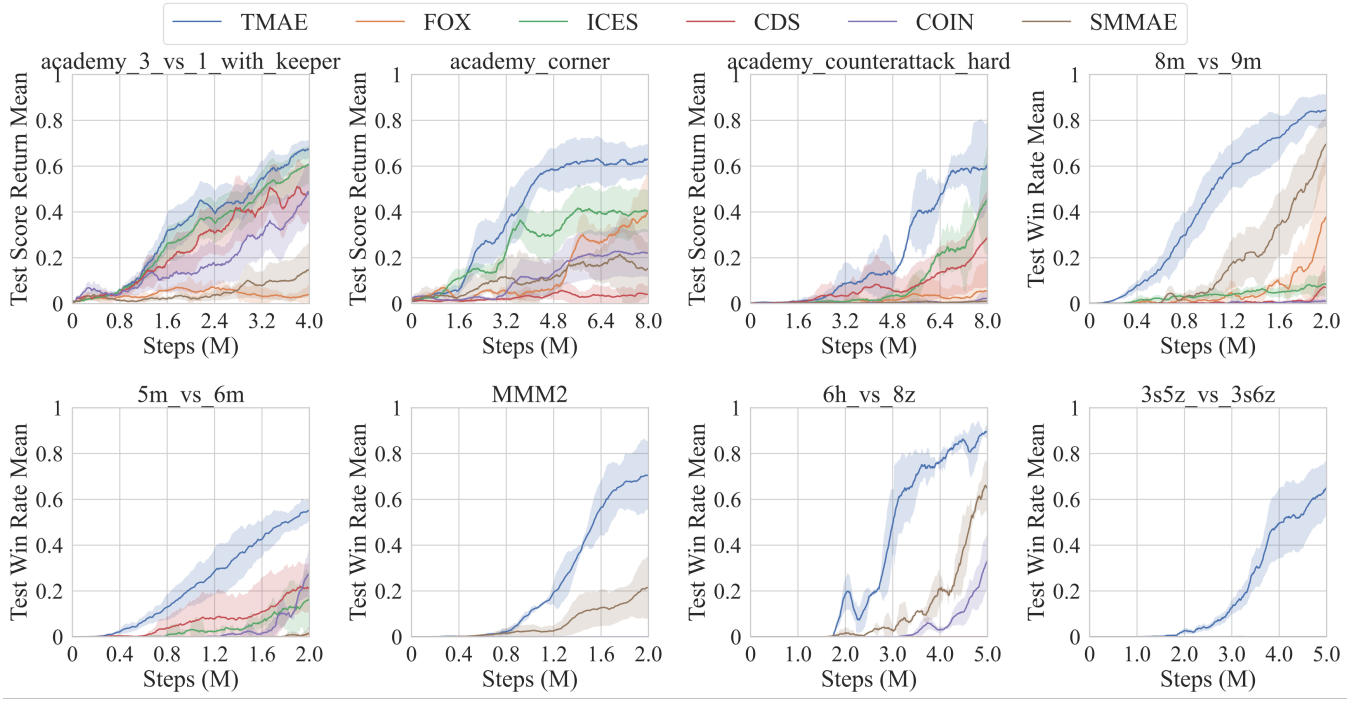

Figure 2: Performance across multiple benchmarks: All results are reported as the median performance over five random seeds.

In the GRF benchmark, many baseline methods converge to local optima after a certain number of training steps and struggle to maintain consistent exploration across different scenarios. In contrast, TMAE demonstrates robust exploration performance across all three GRF scenarios and achieves superior exploration efficiency. This success is attributed to TMAE's ability to uncover the causal relationships between sparse rewards and sub-state spaces, enabling it to effectively focus on critical information from the environment while filtering out exploration confounders.

In the SMAC scenarios, TMAE consistently outperforms all baselines by a large margin in terms of both cooperative performance and exploration efficiency. Specifically, under the sparse reward setting, most baseline methods achieve only limited efficient exploration. For instance, SMMAE performs well in simpler tasks like 8m_vs_9m and 6h_vs_8z but fails to explore efficiently in more complex scenarios. In contrast, TMAE consistently succeeds in exploration across all tasks, surpassing the baselines in exploration efficiency. Notably, in the 3s5z_vs_3s6z scenario, where all baseline methods struggle, TMAE still maintains high exploration efficiency, underscoring its superior capability in tackling complex sparse-reward tasks. This consistent outperformance highlights that TMAE can effectively extract task-specific exploration priors, driving targeted exploration across diverse multi-agent environments.

Ablation Studies

To investigate **Q2** and understand the contribution of each component to TMAE's superior performance, we designed ablation experiments focusing on variants:

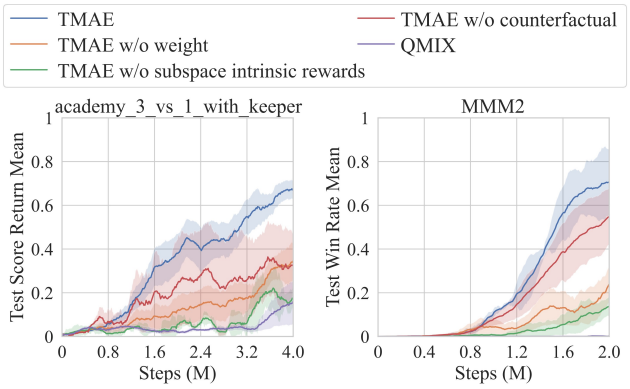

Figure 3: Ablation Study.

- **TMAE:** This represents the complete method proposed in our work.
- **TMAE w/o weight:** This variant omits the causal subspace weighting mechanism. The final intrinsic reward is obtained by equally summing intrinsic rewards of all subspaces.
- **TMAE w/o subspace intrinsic rewards:** This variant further eliminates subspace exploration, utilizing RND directly on global states to encourage global exploration.
- **TMAE w/o counterfactual:** This variant removes the counterfactual intervention on actions and instead uses actual actions as inputs to the reward model.
- **QMIX:** This serves as our baseline for comparison, re-

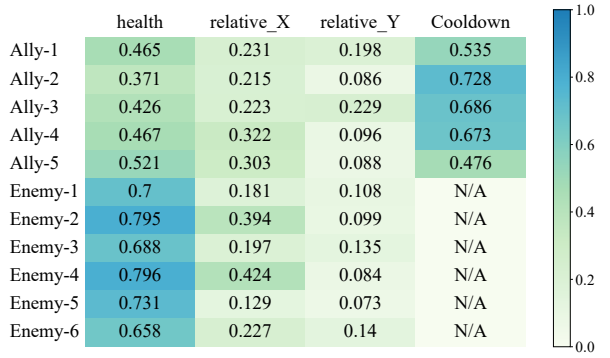

Figure 4: Visualization of the discovered causal weights w.

flecting the core functionality without the enhancements introduced in TMAE.

As illustrated in Fig. 3, we evaluate each variant of TMAE in academy_3_vs_1_with_keeper scenario from GRF and MMM2 scenario from SMAC. The results demonstrate that TMAE's superior performance is primarily attributed to its causal inference–motivated targeted exploration. Specifically, compared to the original QMIX, which employs ϵ-greedy as its exploration strategy, both the global exploration variant (TMAE w/o subspace intrinsic rewards) and the naive subspace exploration variant (TMAE w/o weight) achieve moderate performance improvements. This highlights the effectiveness of incorporating intrinsic rewards into MARL. However, in challenging sparse-reward tasks with high-dimensional state spaces, relying solely on global intrinsic rewards fails to achieve efficient exploration. By introducing causal weights as exploration targets, both TMAE and TMAE w/o counterfactual significantly outperform these variants as the causal insights can effectively reduce exploration space and improve exploration efficiency. Upon further analysis, we observe that TMAE w/o counterfactual exhibits training instability in the GRF scenarios, with larger performance variance. This instability likely arises because the causal effect estimated by this variant is influenced by confounders from historical information, which compromises the accuracy of causal inference. In contrast, TMAE effectively mitigates this issue, achieving stable performance improvements. This highlights the importance of counterfactual processing in eliminating confounders from historical information, which is crucial for accurate causal effect estimation and efficient exploration.

Visualizations of learned subspace weights

To understand the valuable knowledge and insights discovered by TMAE, we visualize the causal insights for each subspace after 2 million training steps in the 5m_vs_6m scenario of SMAC, where the state space consists of the health, position, and cooldown status of both allied and enemy agents. As illustrated in Fig. 4, we visualize the exploration priority derived from causal discovery for each subspace, where darker colors indicate higher exploration priority.

We observe that TMAE can uncover unique exploration priorities for each subspace within the complex state space, aligning well with domain knowledge and the task's reward structure. Specifically, among all subspaces, enemy health and ally cooldown exhibit the highest priorities. This observation aligns with the objectives in SMAC under a sparse reward setting, where exploring enemy health and ally cooldown contributes to eliminating enemy units. In this sparse reward setting, defeating an enemy unit yields a positive reward. Furthermore, the success of tasks in SMAC is defined by the sparse reward triggered upon defeating all enemy units, which demonstrates that TMAE successfully captures the dependencies of the sparse reward. These findings highlight the importance of these subspaces in guiding exploration targets.

Additionally, ally health also exhibits relatively high priority, as exploring ally health influences the sparse reward: an ally's death results in a penalty reward. Moreover, we observe that the priority for exploring ally health is lower than that of enemy health, which is consistent with the reward structure. Specifically, defeating an enemy unit grants a reward of $+10$, whereas losing an ally incurs a smaller penalty of -5. This further validates the alignment of TMAE's exploration priorities with the underlying reward structure in SMAC.

Another interesting finding is that the **causal weights** for the x-coordinates are higher than those for the y-coordinates, indicating that our method encourages agents to move along the x-axis. After reviewing the environmental settings and game mechanics, we find that in the 5m_vs_6m scenario, the initial positioning determines that movement along the x-axis allows agents to locate enemies more easily, thereby indirectly influencing the reward distribution. Hence, the causal insights guide agents to prioritize exploration along the x-axis.

Overall, designing exploration priorities for the state space based on environmental settings and reward structures is a highly complex task. Even constructing the corresponding prior knowledge requires extensive hyperparameter tuning and experimentation. TMAE automatically extracts relevant knowledge from past data by leveraging the expected difference in the reward distribution caused by interventions on the sub-state, which is of great significance for enabling targeted exploration and improving exploration efficiency.

Conclusion

In this work, we delve into the challenge of exploration in multi-agent tasks with sparse rewards. To fully leverage knowledge from previously explored data and thereby accelerate subsequent exploration, we propose TMAE, a framework that analyzes the causal relationship between the state space and sparse rewards. Specifically, we decompose the state space and intervene on each state component, using the ACE to quantify its influence on the reward distribution and its role in discovering sparse rewards. By modeling the ACE between subspaces and the reward function, we establish the exploration priority for each subspace. These causal insights are then integrated into a subspace exploration framework, guiding subsequent exploration tasks and enabling targeted and efficient exploration.

Acknowledgements

This work is supported by the National Natural Science Foundation of China No. 62406313, 2023 Special Research Assistant Grant Project of the Chinese Academy of Sciences.

References

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, Z. D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; and Blundell, C. 2020. Never Give Up: Learning Directed Exploration Strategies. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bettini, M.; Kortvelesy, R.; and Prorok, A. 2024. Controlling Behavioral Diversity in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2405.15054*.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Chen, X.; Liu, X.; Zhang, S.; Ding, B.; and Li, K. 2022. Goal Consistency: An Effective Multi-Agent Cooperative Method for Multistage Tasks. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 172–178. ijcai.org.

Chitnis, R.; Tulsiani, S.; Gupta, S.; and Gupta, A. 2020. Intrinsic Motivation for Encouraging Synergistic Behavior. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Iqbal, S.; and Sha, F. 2019. Coordinated Exploration via Intrinsic Rewards for Multi-Agent Reinforcement Learning. *CoRR*, abs/1905.12127.

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.

Jo, Y.; Lee, S.; Yeom, J.; and Han, S. 2024. FoX: Formation-aware exploration in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12985–12994.

Kim, W.; and Sung, Y. 2023. An Adaptive Entropy-Regularization Framework for Multi-Agent Reinforcement Learning. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 16829–16852. PMLR.

Kurach, K.; Raichuk, A.; Stańczyk, P.; Zajac, 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.

Li, C.; Wang, T.; Wu, C.; Zhao, Q.; Yang, J.; and Zhang, C. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 3991–4002.

Li, J.; Kuang, K.; Wang, B.; Li, X.; Wu, F.; Xiao, J.; and Chen, L. 2024a. Two heads are better than one: a simple exploration framework for efficient multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Li, X.; Liu, Z.; Chen, S.; and Zhang, J. 2024b. Individual Contributions as Intrinsic Exploration Scaffolds for Multi-agent Reinforcement Learning. *CoRR*, abs/2405.18110.

Li, Z.; Wu, L.; Su, K.; Wu, W.; Jing, Y.; Wu, T.; Duan, W.; Yue, X.; Tong, X.; and Han, Y. 2024c. Coordination as inference in multi-agent reinforcement learning. *Neural Networks*, 172: 106101.

Li, Z.; Zhao, W.; Wu, L.; and Pajarinen, J. 2025. AgentMixer: Multi-Agent Correlated Policy Factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18611–18619.

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, I.-J.; Jain, U.; Yeh, R. A.; and Schwing, A. 2021. Cooperative exploration for multi-agent deep reinforcement learning. In *International conference on machine learning*, 6826–6836. PMLR.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Abbeel, O. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, 6379–6390.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in Neural Information Processing Systems*, 32.

PEARL, J. 1988. Probabilistic Reasoning in Intelligent Systems; Network of Plausible Inference. *Morgan Kaufmann, 1988*.

Pearl, J. 1993. [Bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3): 266–269.

Pearl, J. 1994. A probabilistic calculus of actions. In *Uncertainty in artificial intelligence*, 454–462. Elsevier.

Peng, P.; Wen, Y.; Yang, Y.; Yuan, Q.; Tang, Z.; Long, H.; and Wang, J. 2017. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*.

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*.

Song, Z.; Zhao, S.; Zhang, X.; Li, J.; Zheng, C.; and Qiang, W. 2024. Learning Invariant Causal Mechanism from Vision-Language Models. *CoRR*, abs/2405.15289.

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based Interval Estimation for Markov Decision Processes. *J. Comput. Syst. Sci.*, 74(8): 1309–1331.

Sun, C.; He, P.; Ji, Q.; Zang, Z.; Li, J.; Wang, R.; and Wang, W. 2024a. M2i2: Learning efficient multi-agent communication via masked state modeling and intention inference. *arXiv preprint arXiv:2501.00312*.

Sun, C.; He, P.; Wang, R.; and Zheng, C. 2025. Revisiting Communication Efficiency in Multi-Agent Reinforcement Learning from the Dimensional Analysis Perspective. In Das, S.; Nowé, A.; and Vorobeychik, Y., eds., *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, 1977–1986. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

Sun, C.; Wu, B.; Wang, R.; Hu, X.; Yang, X.; and Cong, C. 2021. Intrinsic Motivated Multi-Agent Communication. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, 1668–1670. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

Sun, C.; Zang, Z.; Li, J.; Li, J.; Xu, X.; Wang, R.; and Zheng, C. 2024b. T2mac: Targeted and trusted multi-agent communication through selective engagement and evidence-driven integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15154–15163.

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*.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*.

Wang, J.; Zhao, S.; Qiang, W.; Li, J.; Zheng, C.; Sun, F.; and Xiong, H. 2024. Towards the Causal Complete Cause of Multi-Modal Representation Learning. *arXiv preprint arXiv:2407.14058*.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2019. Influence-based multi-agent exploration. *arXiv preprint arXiv:1910.05512*.

Xu, B.; Wang, Y.; Wang, Z.; Jia, H.; and Lu, Z. 2021. Hierarchically and cooperatively learning traffic signal control. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 669–677.

Xu, P.; Zhang, J.; Yin, Q.; Yu, C.; Yang, Y.; and Huang, K. 2023a. Subspace-aware exploration for sparse-reward multi-agent tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11717–11725.

Xu, P.; Zhang, J.; Yin, Q.; Yu, C.; Yang, Y.; and Huang, K. 2023b. Subspace-aware exploration for sparse-reward multi-agent tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11717–11725.

Yang, M.; Yang, Y.; Lu, Z.; Zhou, W.; and Li, H. 2023. Hierarchical Multi-Agent Skill Discovery. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ye, D.; Zhang, M.; and Yang, Y. 2015. A multi-agent framework for packet routing in wireless sensor networks. *sensors*, 15(5): 10026–10047.

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.

Zang, Z.; Sun, C.; Liu, L.; Sun, F.; and Zheng, C. 2025. Loss of Plasticity: A New Perspective on Solving Multi-Agent Exploration for Sparse Reward Tasks. In Das, S.; Nowé, A.; and Vorobeychik, Y., eds., *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, 2299–2308. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

Zeng, Y.; Cai, R.; Sun, F.; Huang, L.; and Hao, Z. 2024. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, S.; Cao, J.; Yuan, L.; Yu, Y.; and Zhan, D. 2023. Self-Motivated Multi-Agent Exploration. In Agmon, N.; An, B.; Ricci, A.; and Yeoh, W., eds., *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, 476–484. ACM.

Zhang, Y.; Li, J.; Liu, L.; and Qiang, W. 2024. Rethinking misalignment in vision-language model adaptation from a causal perspective. *Advances in Neural Information Processing Systems*, 37: 39224–39248.

Zheng, L.; Chen, J.; Wang, J.; He, J.; Hu, Y.; Chen, Y.; Fan, C.; Gao, Y.; and Zhang, C. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34: 3757–3769.