Autonomous Partner Selection for Cooperative Multi-Agent Reinforcement Learning

Rui Tang, Biao Luo*, Yongzheng Cui

School of Automation, Central South University, Changsha 410083, China
ruitang02@csu.edu.cn, biao.luo@hotmail.com, yongzhengcui@csu.edu.cn

Abstract

In cooperative Multi-Agent Reinforcement Learning (MARL), the subgroup-wise learning is employed to assign sub-tasks to agents towards the enhancement of team collaboration. However, the present work is dependent on manually defined allocation criteria, which hinders its capacity to adapt to environmental changes promptly, and also relaxes communication restrictions, thereby constraining the application of algorithms in a range of fields. In order to address these issues, the Autonomous Partner Selection (APS) framework is proposed, which offers an implicit grouping mechanism in an autonomous way. Each agent is capable of autonomously selecting cooperative partners and integrating their own observations with those of partners to harmonise the cooperative behaviour during the training stage. With a view to strictly restricting communication, the intention encoder is trained through information distillation, which enables agents to selectively take more cooperative actions based solely on local observations. Meanwhile, in order to circumvent potential conflicts engendered by homogenization behaviour, we employ a contrastive learning strategy to the cooperative intention generated by agents, thereby ensuring that the behavioural tendencies exhibited by different individuals remain as diverse as possible. Finally, extensive comparative experiments on the StarCraft Multi-Agent Challenge and Google Research Football are conducted. The results demonstrate that APS exhibits superior performance in comparison to the state-of-the-art algorithms across a range of tasks, and agents can adapt their grouping strategies in accordance with the environment to facilitate enhanced cooperation.

Introduction

Despite the extensive implementation of cooperative MARL algorithms in domains such as robotics (Ou et al. 2024; Gu et al. 2023; Liang, Chang, and Pan 2023), video games (Ye, Chen, and Zhang 2020), autonomous driving (Zheng and Gu 2024), and related fields, numerous challenges persist, including scalability, non-stationarity (Papoudakis et al. 2019). The Centralized Training and Decentralized Execution (CTDE) (Gupta, Egorov, and Kochenderfer 2017; Oliehoek and Amato 2016) framework has been proposed as a solution to these issues. Agents are capable of acquiring global information during training and can select actions solely based on their local observations when in execution. Centralized training ensures that agents can perceive changes from the surrounding environment and other agents in time when training, thereby solving the problem of non-stationarity. Decentralized execution can avoid searching in large-scale joint action spaces, reduce computational complexity, and alleviate problems caused by scalability.

Under the CTDE paradigm, value-based algorithms have demonstrated efficacy across numerous benchmarks (Iqbal et al. 2021). However, in complex tasks, the incomplete utilisation of information and the tendency to fall into local optima can still greatly hinder team performance. A significant amount of work has been conducted to address these challenges. Subgroup-wise learning, in which agents are assigned to subgroups or different roles, attracts wide attention. A meticulous division of labour or guidance can facilitate the selective utilisation of observation information by each subgroup to complete sub-tasks and thereby enhance the collaborative performance at the team level. Typically, such methods rely on manually selected criteria or contributions to divide the agents. However, it is not always possible to adjust these standards and indicators in a timely manner according to changes in the complex environment, which in turn limits the performance. Concurrently, certain methodologies employ a relaxation of communication restrictions, utilising the guidance derived from global information as inputs for agents. Such an arrangement constitutes a violation of the CTDE paradigm, consequently diminishing the potential for practical real-world applications.

With recent developments in deep reinforcement learning, the Gated Recurrent Unit (GRU) (Cho et al. 2014) is frequently employed in the construction of agent networks, a consequence of its highly effective capacity to process temporal information. The hidden states in GRU retain both the information previously encountered and that currently observed, thus providing a method for addressing immediacy. It has also been determined that individuals who have been exposed to analogous information tend to demonstrate greater proficiency in communication and cooperation. Indeed, shared experiences contribute to establishing implicit consensus among agents, thereby facilitating more precise prediction of each other's behaviour and alignment of inten-

*Corresponding author.

tions, without the need for explicit communication. In light of these premises, we posit that agents possess the capacity to autonomously select partners for collaboration, driven by their own experiences and the information they have acquired. Meanwhile, in order to avoid additional inputs, the agent should be confined to extracting information from local observations.

To achieve these functions, a novel subgroup collaboration framework for MARL, APS, is proposed. First, to permit agents the autonomy to select partners independently, the attention mechanism is employed to process their experience information and promote harmonious cooperation. However, such a process would necessitate access to global information, a requirement that conflicts with the CTDE paradigm. To this end, an intention coder is trained through information distillation, enabling the extraction of key collaboration clues from local observations in decentralized decision-making settings. Finally, in order to circumvent the conflict engendered by homogeneous cooperation tendencies among agents, we employ contrastive learning to ensure the diversity of cooperation tendencies, thereby promoting orderly cooperation. We finally implemented the APS framework on QMIX (Rashid et al. 2020b) and conducted a series of comparisons on StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al. 2019) and Google Research Football (GRF) (Kurach et al. 2020). The main contributions of this paper are:

- We propose a novel subgroup framework for MARL that allows agents to autonomously select cooperative partners and achieve implicit grouping dynamically in complex environments, achieving efficient cooperation.
- We propose an attention matching mechanism that facilitates the selection and integration of information from the same subgroup units and harmonizes the cooperative behaviour. Furthermore, information distillation ensures that agents can also extract cooperative information from local observations under the CTDE framework.
- To circumvent potential conflicts arising from homogeneity, emphasis is placed on the disparities in tendencies between different agents, thereby facilitating orderly cooperation among subgroups.
- We conducted a series of comparative experiments between APS and several state-of-the-art (SOTA) methods in complex and sparse environments, thereby validating the efficient performance. In addition, we also designed ablation experiments to demonstrate and analyse the impact of each module in APS on overall performance.

Related Work

CTDE in MARL

Agents can only select actions grounded in local observations and gain access to supplementary global information for training purposes within the framework of the CTDE paradigm (Gupta, Egorov, and Kochenderfer 2017; Oliehoek and Amato 2016), currently the mainstream of MARL. This paradigm circumvents the necessity for searching in the joint action space, thereby mitigating the issue of scalability. Furthermore, it enables agents to discern changes in other agents, thus addressing the problem of non-stationarity.

Owing to CTDE, a considerable number of MARL algorithms have demonstrated favourable performance. One class of these methods is formed around value decomposition, namely value-based algorithms. VDN (Sunehag et al. 2018) decomposes the joint value function into the linear sum of individual value functions. As an extension, the IGM principle is proposed as a means of ensuring consistency between the global optimal action and the local optimal action (Son et al. 2019). QMIX (Rashid et al. 2020b) implements a monotonic constraint based on the principle and achieves favourable performance on SMAC (Samvelyan et al. 2019). QPLEX (Wang et al. 2021), WQMIX (Rashid et al. 2020a), Qatten (Yang et al. 2020) further expand the expressive capacity of mixing networks. In addition, there exist methods designed to enhance the performance of the value-based MARL algorithm with regard to exploration (Sun, Lee, and Lee 2021; Chen et al. 2023) and sample efficiency (Zheng et al. 2021; Na, Seo, and Moon 2024).

The other type is policy-based methods, which generally incorporate several distributed actors as well as a centralized critic, wherein the critic leverages global information for training. MADDPG (Lowe et al. 2017) incorporates all actions taken by agents as inputs for the critic during training, enabling the consideration of interactions between agents within the policy optimization. MAPPO (Yu et al. 2022) implements PPO (Schulman et al. 2017) in multi-agent systems and demonstrates that on-policy policy-based methods can also perform well in different environments (Mordatch and Abbeel 2018; Samvelyan et al. 2019) under the CTDE framework. COMA (Foerster et al. 2018), SIC-MA (Chen et al. 2022), and PMIC (Li et al. 2022) improve the performance of the policy-based algorithm on several tasks (Mordatch and Abbeel 2018; de Witt et al. 2020) through their respective mechanisms. Nevertheless, both value-based and policy-based methods continue to be vulnerable to the issue of local optima and require a substantial number of interactions to achieve satisfactory performance.

Subgroup-wise Learning in MARL

As the CTDE framework remains deficient when it comes to complex environments, subgroup-wise learning has attracted attention for its potential to enhance performance with greater refinement. In multi-agent systems, the primary objective is typically concerned with maximising the collective benefits of the team. The guidance aimed at fostering cooperation at the subgroup or role-wise level can be regarded as a subdivision of the original tasks, which can result in the emergence of more sophisticated techniques and endeavours (Lee, Yang, and Lim 2019; Iqbal, Costales, and Sha 2022). Group divisions (Phan et al. 2021) and the assignment of identities to different roles (Li et al. 2021) can be considered as methods of subgroup-wise learning. The rationale pertains to the grouping of agents and the implicit or explicit assignment of distinct subtasks to different subgroups. The latter is a special form, whereby each agent is divided into a group.

Existing work (Liu et al. 2021; Shao et al. 2022; Li et al.

2020) generally classifies agents into different subgroups or roles based on their competencies, contributions, relations, and other factors. For instance, ACORM (Hu et al. 2024) segments agents into predetermined groups based on their hidden states, and attains behavioural heterogeneity and knowledge transfer by either increasing the discrepancy between different groups or reducing the discrepancy within the same group. Gomarl (Zang et al. 2023) performs automatic and dynamic grouping of agents based on their contributions for a certain time interval, and subsequently assigns them to the most appropriate subgroups, without the need for human intervention. A hierarchical control is also introduced for agents in the same group to specialize in similar policies. GACG (Duan, Lu, and Xuan 2024) derives the connections between agent pairs based on local observations and group-level dependencies from behaviour patterns, thereby promoting decision-making by a message-passing mechanism.

While promising results have been achieved, these methods are contingent on human definition or information-passing between agents, thereby loosing the constraints. Moreover, the criterion employed for grouping presents a challenge in terms of timely adjustment in response to environmental changes, which may lead to suboptimal cooperation. To this end, we propose a direct and effective MARL algorithm with the following improvements: 1)our work does not involve the manual designation of the division of labour that different agents should engage in; rather, it enables agents to autonomously select suitable partners in order to adapt to changing circumstances and complex environments; 2)our method does not necessitate any additional communication protocols and adheres strictly to the CTDE framework, a characteristic that renders it applicable across a wide range of systems.

Preliminary

Problem Setup

A cooperative MARL task can be modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), described as a tuple $< N, S, A, O, P, r, \gamma >$. $N = \{1, \cdots, n\}$ is the set of n agents and $s \in S$ is the global state of the environment. At each time step t, each agent i recieves a local observation $o_i^t \in O$ and selects an action $a_i^t \in A_i$ according to its own policy $\pi_i(\cdot|o_i^t)$ to form a joint action $\mathbf{a}^t \in (A_1 \times \cdots \times A_n) \equiv A^N$. This induces a transition to the next state s^{t+1} according to the state transition function $P(s^{t+1}|s^t, \mathbf{a}^t)$. All agents share the same reward function $r(s^t, \mathbf{a}^t)$ and $\gamma \in [0, 1]$ is the discount factor. Each agent learns its policy to jointly maximize the discounted reward $R^t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$.

CTDE Paradigm

The CTDE paradigm permits agents to learn a joint value function $Q_{tot}(\cdot; \theta_\omega) = f_{\theta_\omega}([Q_i(o_i, a_i; \theta)]|_{i=1,\cdots,n}, s)$ by accessing the global state s and local value functions of all agents during the training process. However, an agent is only constrained to selecting actions based on its own observation information during execution. Reasonable credit assignment provides agents with the capacity to recognise the contribution of their actions to the team and adjust their respective policies accordingly. The parameters of mixing networks θ_ω and local value networks θ are adjusted by minimizing the Temporal-Difference (TD) loss:

$$L_{TD}(\theta) = E_D[(y_{tot} - Q_{tot}(\mathbf{a}, s; \theta_\omega))^2], \tag{1}$$

where $y_{tot} = r + \gamma \max_{\mathbf{a}'} Q_{tot}(\mathbf{a}', s'; \overline{\theta_\omega})$ is the target value and $\overline{\theta_\omega}$ is the parameter of the target mixing network.

Method

In this section, we represent a novel cooperative MARL algorithm, namely APS. A prerequisite for enhancing team performance is to facilitate the advancement of agents towards more sophisticated cooperation. Existing methods tend to categorise agents or allocate roles based on predetermined criteria, such as the contribution of agents to the team. These methods typically depend on artificially designed criteria, which are unable to respond promptly to changes in the relationships between agents within complex and dynamic environments. Consequently, they are deficient in their capacity to adjust groupings in a timely manner, which limits cooperative performance. Nevertheless, a more efficacious approach would be to permit the agent to select its own collaborators independently. Agents possess the capacity to select actions based on local observations and past experience, which inherently encompasses the ability to cooperate with specific agents through the selection of actions. The overall architecture is detailed in Figure. 1.

Cooperative Information Distillation

In order to achieve autonomous selection of cooperative entities, the employment of attention mechanisms is considered. Specifically, the hidden states of the agent networks serve as the sources of queries and keys. This is premised on the concept that the hidden state can retain past memory information, thereby ensuring the consistency of the agents' choices. Therefore, for all agents in the team, the respective hidden states are concatenated to obtain a joint vector $h^t = [h_1^t, h_2^t, \cdots, h_n^t]^T \in R^{n \times d_h}$, where $h_n^t \in R^{d_h}$ is the hidden state of agent n at time step t. Subsequently, the hidden embedding vector $\overline{h}^t = f_\zeta(h^t) \in R^{n \times d_e}$ that have been produced by the hidden encoder $h_\zeta(\cdot) : R^{d_h} \to R^{d_e}$ are transformed into queries $Q \in R^{n \times d_e}$ and keys $K \in R^{n \times d_e}$:

$$\begin{bmatrix} Q \\ K \end{bmatrix} = \overline{h}^t \begin{bmatrix} W_q \\ W_k \end{bmatrix}. \tag{2}$$

where $W_q \in R^{d_e \times d_e}$ and $W_k \in R^{d_e \times d_e}$ are learnable parameter matrices.

In terms of value, it is envisaged to contain sufficient information to support the agent in achieving cooperation that is beneficial to the team. Therefore, following the procedure for obtaining the query and key, the local observations from all agents are concatenated into a joint observation vector $o^t \in R^{n \times d_o}$, and a parameter matrix $W_v \in R^{d_e \times d_e}$ is also used to obtain the values $V = \overline{o}^t \cdot W_v \in R^{n \times d_e}$, where $\overline{o}^t = f_\xi(o^t) \in R^{n \times d_e}$ is the observation embedding vector generated from the input encoder $f_\xi(\cdot) : R^{d_o} \to R^{d_e}$.

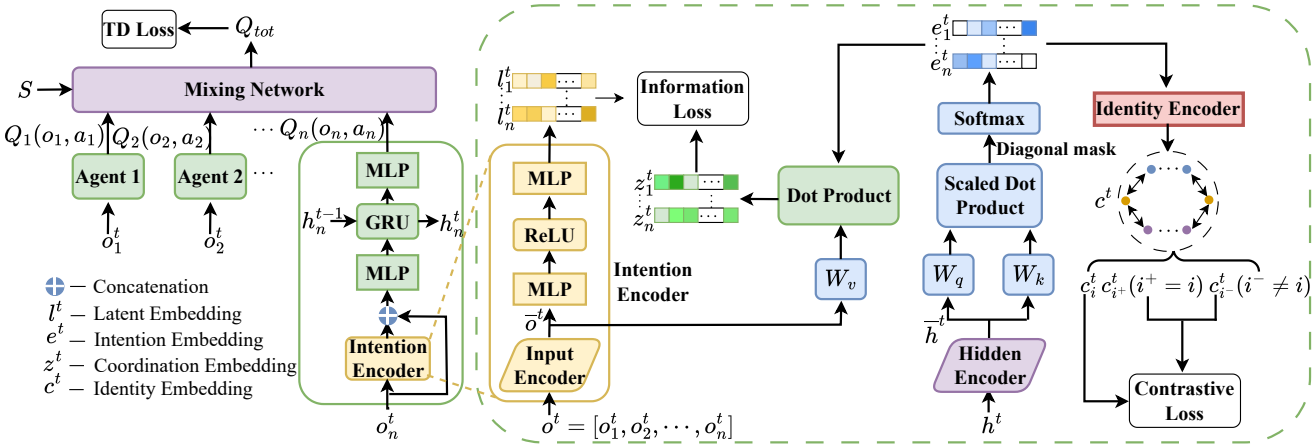

Figure 1: Overall architecture of APS. The intention encoder extracts possible cooperative information from local observations and uses it as input, along with the original observations, to obtain local Q values. The purple box on the right represents the training process. The hidden states serve as the sources of the query and key in the attention mechanism, and the observation embeddings processed by the input encoder serve as the value. The intention encoder is trained to guide agents to focus on relevant information from other agents by information distillation. Furthermore, to circumvent potential conflicts, a contrastive loss is incorporated to achieve disparities in the tendencies of diverse agents.

The query and key are then multiplied in order to calculate the correlation. Agents with analogous experiential backgrounds are more likely to reach a consensus and achieve effective cooperation when presented with the same situation. To encourage agents to consider collaborating with each other, we mask the associations between themselves after dot producting the query and key. The resulting matrix is then sent to softmax to obtain the agents' intention embedding matrix $e^t \in R^{n\times n}$. Each row represents an agent's cooperative intention towards other agents, while the diagonal position is masked as 0 to indicate that an agent only focuses on other agents:

$$e^t = softmax(mask(\frac{QK^T}{\sqrt{d_e}})), \tag{3}$$

where $\sqrt{d_e}$ is the scaling coefficient, and the mask operation is implemented by assigning a $-\infty$. The intention embedding matrix is then dot producted by the values obtained from joint local observations to yield the possible coordination embedding matrix $z^t = e^t \cdot V \in R^{n\times d_e}$. Each row of z^t represents the potential cooperative interactions between a specific agent and other agents.

The coordination matrix obtained in the aforementioned process contains cooperative information that exceeds the local observations. However, it is worth noting that cooperative information is obtained from the observations of multiple agents, and each agent's own observations may contain certain elements of cooperative information. To facilitate the extraction of cooperative information from local observations, it is imperative to concentrate the agent's attention on potential cues while refraining from the introduction of extra information. Towards this end, efforts are directed towards leveraging comprehensive cooperative information to distill the partial cooperative information contained within local observations, which is referred to as information distillation. Therefore, an intention encoder is designed that consists of two MLPs and an activation function, namely the ReLU function. $l^t = \mathbf{f}_{(\phi,\xi)}(o^t) = f_\phi(f_\xi(o^t)) \in R^{n\times d_e}$ is the latent embedding matrix, where $f_\phi(\cdot) : R^{d_e} \to R^{d_e}$ is the network employed by the intent encoder $\mathbf{f}_{(\phi,\xi)}(\cdot)$ to process $\overline{o}^t$ parametered by ϕ. To enable the encoder to discover collaborative interactions from local observations, the following loss function is employed in the training process:

$$L_{info}(\phi,\zeta,\xi) = E_t[KL(\mathbf{f}_{(\phi,\xi)}(l^t|o^t))||A(z^t|f_\zeta(h^t), f_\xi(o^t))], \tag{4}$$

where $KL(\cdot)$ is the Kullback-Leible divergence function, $A(\cdot)$ is the attention function. As the KL divergence is minimised, the discrepancy of probability distribution between l^t and z^t can be diminished. To this end, the cooperative information distribution of local observation extraction is conducive to approximating the ideal cooperative information distribution under the guidance of global information, thereby enabling agents to extract meaningful cooperative information and realise effective cooperation by relying solely on local observation.

Behavioral Tendency Differentiation

In order to save parameter space and improve training efficiency, parameter sharing is a widely used technique in MARL. However, this may result in a tendency towards homogenisation among agents, which is not favourable for cooperation. In a similar manner, when several agents exhibit a homogeneous tendency towards cooperation, conflicts may emerge, exerting an adverse effect on team performance.

To circumvent potential conflicts, the concept of contrastive learning is employed to differentiate the cooperative tendencies exhibited by different agents. An identity encoder $f_\psi(\cdot) : R^{d_e} \to R^{d_e}$ composed of two fully connected layers has been designed to extract identity information and

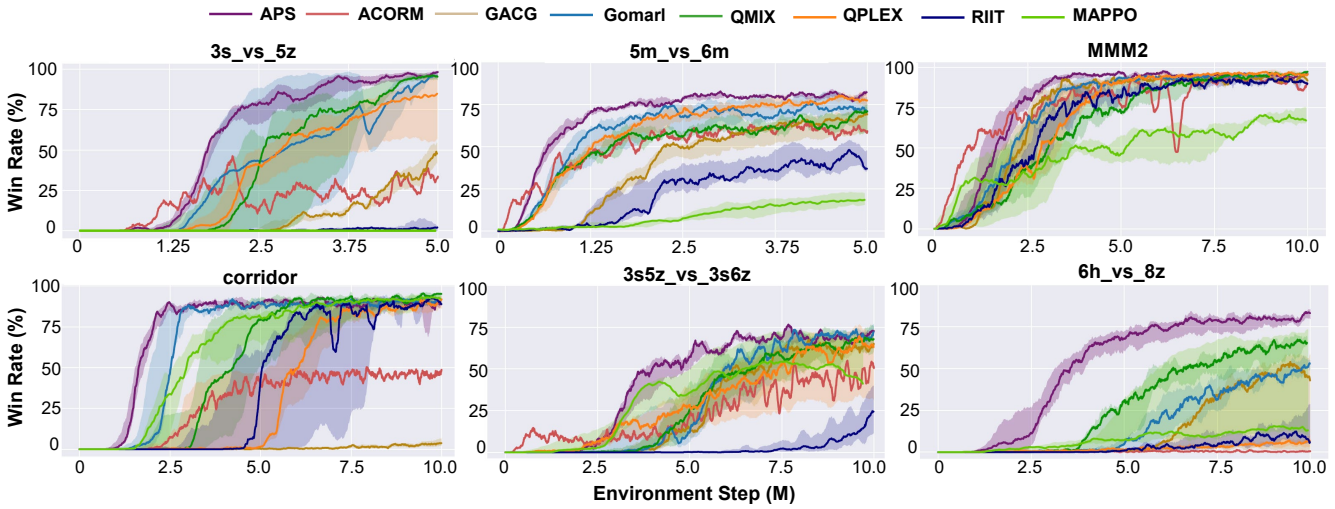

Figure 2: Comparisons of test win rate on 2 Hard SMAC maps: 3s_vs_5z, 5m_vs_6m, and 4 Super-Hard SMAC maps: MMM2, corridor, 3s5z_vs_3s6z, and 6h_vs_8z. The win rate curve is the median result of five distinct random seeds. 25% - 75% quartile is shown shaded.

avoid interference caused by diagonal 0 elements. The intention embedding matrix is processed by the identity encoder to yield the identity embedding matrix $c^t \in R^{n \times d_e}$, where each row represents the identity of a specific agent.

For agent i, the identity embedding of itself is regarded as a positive sample c^t_{i+}, whilst those of other agents are regarded as negative samples $c^t_{i-} \in c^t / c^t_i$. The maximization of the exponential of positive sample product results in an increase in the gap between different samples, thereby achieving intention heterogeneity among agents:

$$X(a,b) = exp(a^T \cdot b),$$
$$L_{contra}(\zeta,\psi) = E_{i,t}[-\log \frac{X(c^t_i, c^t_{i+})}{X(c^t_i, c^t_{i+}) + X(c^t_i, c^t_{i-})}]. \quad (5)$$

Overall Objective

It is possible to train an intention encoder that can mine possible cooperative information from local observations by incorporating both L_{info} and L_{contra}. The observation information processed by the intention encoder is concatenated with the original observations and transmitted as input to the subsequent networks in order to obtain the local Q value:

$$Q_i(o_i, \cdot) = f_\theta(o_i, \mathbf{f}_{(\phi,\zeta)}(o_i)), \quad (6)$$

each agent then selects an action based on ϵ greedy principle to obtain $Q_i(o_i, a_i)$. Finally, in a manner analogous to QMIX, the TD loss is calculated using Q_{tot} obtained through a mixing network. The overall training objective is as follows:

$$L(\theta,\phi,\zeta,\xi,\psi) = E_D[L_{TD} + \lambda_i \cdot L_{info} + \lambda_c \cdot L_{contra}], \quad (7)$$

where λ_i and λ_c are coefficients, D is the replay buffer.

Experiment

To demonstrate the effectiveness of the proposed method, experiments are conducted on several benchmarks, and a comparison is made with some SOTA methods: GACG (Duan, Lu, and Xuan 2024), ACORM (Hu et al. 2024), Gomarl (Zang et al. 2023), MAPPO (Yu et al. 2022), RIIT (Hu et al. 2023), QPLEX (Wang et al. 2021), and QMIX (Rashid et al. 2020b). GACG guides agents to group seamlessly through inferring cooperation graphs; ACORM groups agents based on their agent embeddings to achieve efficient cooperation; Gomarl achieves effective cooperation through an automatic grouping strategy; MAPPO and RIIT are policy-based MARL algorithms that have been validated as effective in multiple environments, while QMIX and QPLEX are value-based algorithms also validated as effective.

Performance

As illustrated in Figure. 2, we present the results of the experiment conducted on the SMAC environments. It is evident that APS has attained performance that surpasses that of the SOTA algorithms on almost all maps. On maps exhibiting relatively modest environmental dynamics, such as MMM2, the SOTA method can achieve favourable results. However, in the 6h_vs_8z map, the agents must learn to focus fire on enemies to achieve victory. Selecting the focus object based on temporal changes imposes considerable demands, thereby hindering the SOTA algorithms' efficacy in guiding the agents towards efficient cooperation, consequently leading to unsatisfactory performance. However, through the appropriate training of intention encoders, APS can successfully guide the agent to select suitable partners based on the surroundings and focus on the opponents.

Figure. 3 presents the experimental results conducted on GRF. As a sparse reward environment, GRF presents a chal-

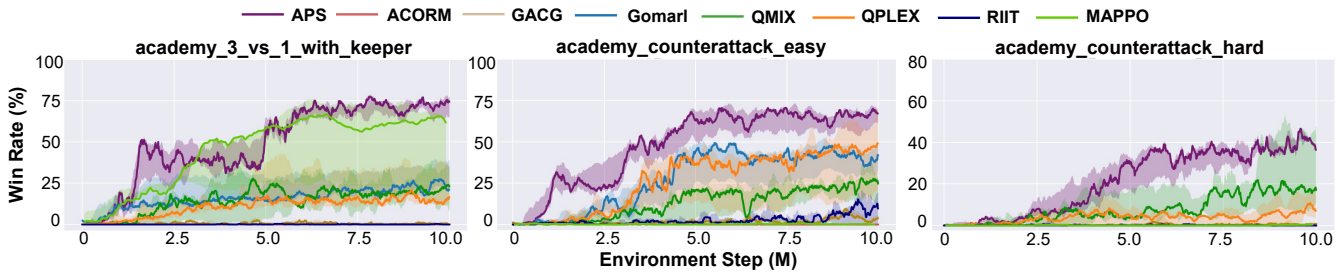

Figure 3: Comparisons of test win rate on 3 GRF maps: academy_3_vs_1_keeper, academy_counterattack_easy, academy_counterattack_hard.

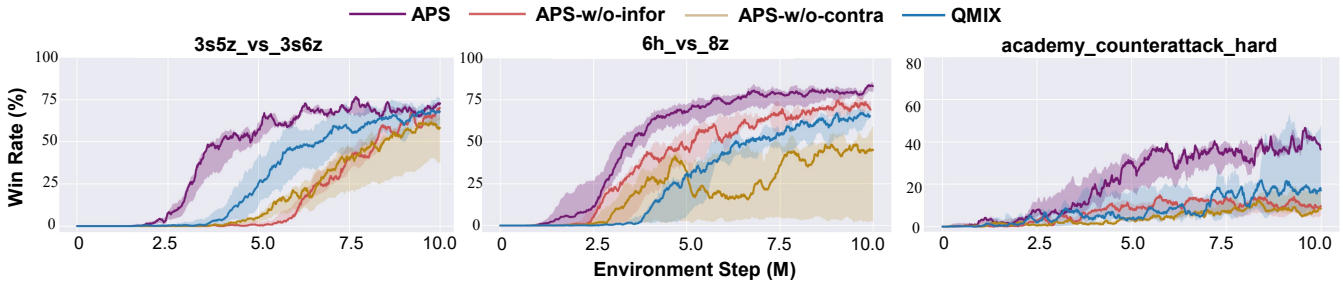

Figure 4: Ablation experiment results on information distillation and tendency differentiation.

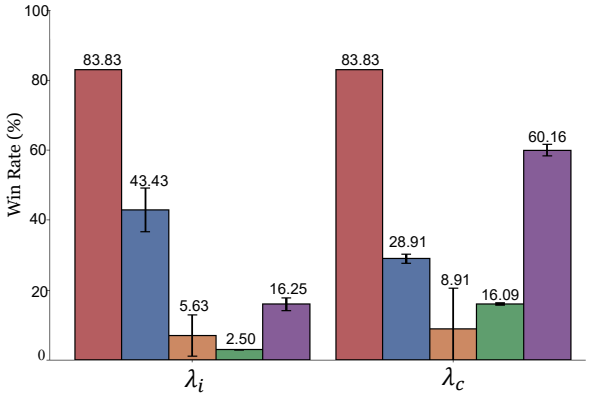

Figure 5: Sensitivity analysis results on hyperparameters λ_i and λ_c. From left to right, the figure displays the test win rate of five random seeds of APS on the 6h_vs_8z when $\lambda_i = 0.1, 0.3, 0.5, 0.7, 0.9$ and when $\lambda_c = 0.01, 0.03, 0.05, 0.07, 0.09$.

lenge to agents in terms of determining the benefits that individual actions bring to the team, thereby increasing the level of difficulty. Consequently, some SOTA methods demonstrated poor performance across all three testing maps. However, due to the capacity of APS to direct agents in the discovery of cooperative objects from local observations, even in the absence of an environmental reward in the intermediate step, the agent will endeavour to cooperate with other agents and attain favourable performance, particularly in academy_counterattack_hard. These results suggest that APS possesses effective collaborative abilities, even in sparse reward environments.

Hyperparameter Sensitivity

In order to analyse the sensitivity of APS to λ_i and λ_c, two groups of experiments were conducted on the map of 6h_vs_8z. For the first group, the value of parameter λ_c is fixed at 0.01, and the values of parameter λ_i are 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. The second group fixed the parameter $\lambda_i = 0.1$ to evaluate the performance of APS when λ_c was set to 0.01, 0.03, 0.05, 0.07, and 0.09.

The results depicted in Figure. 5 demonstrate that as λ_i and λ_c are increased in a gradual manner, the performance will undergo a decline. In the event of $\lambda_i = 0.9$ or $\lambda_c = 0.09$, the relative performance will be enhanced; however, it remains inferior to the original optimum performance. This finding indicates that maintaining adequate levels of information distillation and tendency differentiation is imperative. An excess of attention paid to the accuracy of information distillation or the differentiation level of tendency will divert the agents' focus from the team cooperation performance to the secondary objectives.

Ablation and Visualization Analysis

We sought to demonstrate the efficacy of information distillation and tendency differentiation for APS by conducting ablation experiments. To this end, we refer to the method of removing information distillation as APS-w/o-infor and that of excluding tendency differentiation as APS-w/o-contra, respectively. The experimental results are displayed in Figure. 4. As is apparent on the maps of SMAC and GRF, APS

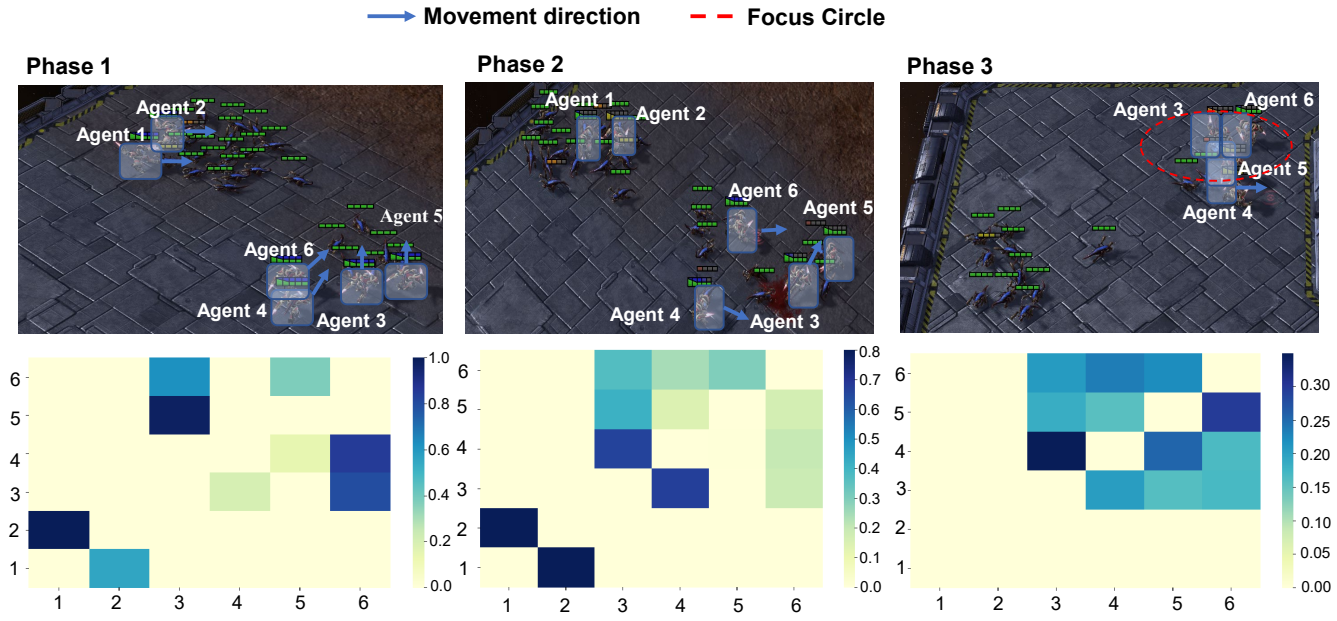

Figure 6: A visual analysis of APS's performance on Corridor. The blue line denotes the direction of movement of the agent, and all agents within the same red circle are participating in focusing fire on enemies. The upper row of the figure displays the battle screenshot, whilst the lower row shows the corresponding attention heat map. Each row in the heat map represents the extent of attention an agent pays to other peers.

continues to demonstrate superior performance. This finding suggests that both information distillation and tendency differentiation play a beneficial role in enhancing performance. Furthermore, it is noteworthy that the elimination of tendency differentiation across the majority of maps would lead to an inferior performance in comparison to QMIX, as a consequence of the conflicts engendered by homogenization tendencies. The variant algorithm APS-w/o-infor will cause the agents to pay attention to the noise in their local observations, thereby impeding cooperation.

Further, to illustrate the efficacy of APS in coordination, a visualisation analysis was conducted on the corridor in SMAC. As demonstrated in Figure. 6, the findings reveal that APS can effectively direct agents to achieve efficient cooperation and overcome enemy forces in response to environmental changes. At the start of the battle, agents 1 and 2 proceed towards the top left corner of the map, thus drawing the majority of enemy forces to that area. The agents 3-6 employ a strategy of distraction, whereby they attract the enemy's attention by repeatedly moving, thus creating an opportunity for their teammates to launch an attack. This coordination method has the potential to both defeat the enemy and maintain the health of the player to a considerable degree. In phase 3, agents 1 and 2 are eliminated by the enemy, but agents 3-6 are still healthy and able to attack the remaining enemy. Specifically, Agent 4 is responsible for attracting enemy fire (illustrated in the battle screenshot). Concurrently, the attention heat map reveals that agent 4 predominantly directs its focus towards teammates 3, 5, and 6, while these teammates reciprocate with a comparable level of attention, thereby establishing a transient coalition to collectively eliminate the residual adversaries. This is an empirical manifestation of the autonomous selection of partners by agents. By means of a strategic exploitation of local observations, the agent identifies potential partners with great acumen and takes appropriate actions to promote cooperative behaviour, thereby achieving temporary partnerships.

Conclusion

This article proposes a novel learning framework, APS, to address the limitations of delayed adjustment and reliance on additional information in the subgroup-wise learning methods for MARL. It provides agents with assistance in the autonomous selection of cooperative partners, thereby facilitating their adaptation to complex environmental changes and the achievement of efficient cooperation. The experimental results robustly validate that APS exhibits a substantially superior collaborative capability in comparison to other SOTA algorithms, irrespective of scenarios that demand intricate micro-operations or those characterised by the scarcity of reward signals. Concurrently, ablation experiments suggest that information distillation and tendency differences can contribute to enhanced team performance. In future research, the scalability of APS in large agent systems promises to be a fruitful avenue. We believe that the principles underpinning APS hold significant promise for addressing complex real-world applications, such as coordinating drone swarms and managing dynamic resource allocation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62373375 and U2341216, and the Science and Technology Innovation Program of Hunan Province under Grant 2024RC1011.

References

Chen, L.; Guo, H.; Du, Y.; et al. 2022. Signal instructed coordination in cooperative multi-agent reinforcement learning. In *Distributed Artificial Intelligence: Third International Conference*, 185–205. Springer.

Chen, Z.; Luo, B.; Hu, T.; et al. 2023. LJIR: Learning joint-action intrinsic reward in cooperative multi-agent reinforcement learning. *Neural Networks*, 167: 450–459.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

de Witt, C. S.; Peng, B.; Kamienny, P.-A.; et al. 2020. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*.

Duan, W.; Lu, J.; and Xuan, J. 2024. Group-Aware Coordination Graph for Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 3926–3934.

Foerster, J.; Farquhar, G.; Afouras, T.; et al. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 363, 2974–2982.

Gu, S.; Kuba, J. G.; Chen, Y.; et al. 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319: 103905.

Gupta, J. K.; Egorov, M.; and Kochenderfer, M. 2017. Cooperative multi-agent control using deep reinforcement learning. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, volume 10642, 66–83.

Hu, J.; Wang, S.; Jiang, S.; et al. 2023. Rethinking the Implementation Tricks and Monotonicity Constraint in Cooperative Multi-agent Reinforcement Learning. In *Proc. International Conference on Learning Representations*.

Hu, Z.; Zhang, Z.; Li, H.; et al. 2024. Attention-Guided Contrastive Role Representations for Multi-agent Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Iqbal, S.; Costales, R.; and Sha, F. 2022. Alma: Hierarchical learning for composite multi-agent tasks. *Advances in Neural Information Processing Systems*, 35: 7155–7166.

Iqbal, S.; De Witt, C. A. S.; Peng, B.; et al. 2021. Randomized entity-wise factorization for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 4596–4606. PMLR.

Kurach, K.; Raichuk, A.; Stańczyk, P.; et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4501–4510.

Lee, Y.; Yang, J.; and Lim, J. J. 2019. Learning to coordinate manipulation skills via skill behavior diversification. In *International Conference on Learning Representations*.

Li, C.; Wang, T.; Wu, C.; et al. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 3991–4002.

Li, P.; Tang, H.; Yang, T.; et al. 2022. PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration. In *International Conference on Machine Learning*, 12979–12997. PMLR.

Li, S.; Gupta, J. K.; Morales, P.; et al. 2020. Deep implicit coordination graphs for multi-agent reinforcement learning. *arXiv preprint arXiv:2006.11438*.

Liang, H.; Chang, Z.; and Pan, Y. 2023. Dual-event-triggered intelligence security control for multiagent systems against DoS attacks with applications in mobile robot systems. *IEEE Transactions on Artificial Intelligence*, 5(2): 916–924.

Liu, B.; Liu, Q.; Stone, P.; et al. 2021. Coach-player multi-agent reinforcement learning for dynamic team composition. In *International Conference on Machine Learning*, 6860–6870. PMLR.

Lowe, R.; Wu, Y. I.; Tamar, A.; et al. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30: 6382–6393.

Mordatch, I.; and Abbeel, P. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1495–1502.

Na, H.; Seo, Y.; and Moon, I.-c. 2024. Efficient Episodic Memory Utilization of Cooperative Multi-Agent Reinforcement Learning. In *Proc. International Conference on Learning Representations*.

Oliehoek, F. A.; and Amato, C. 2016. *A concise introduction to decentralized POMDPs*, volume 1.

Ou, W.; Luo, B.; Xu, X.; et al. 2024. Reinforcement learned multi–agent cooperative navigation in hybrid environment with relational graph learning. *IEEE Transactions on Artificial Intelligence*, 6: 25–36.

Papoudakis, G.; Christianos, F.; Rahman, A.; et al. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*.

Phan, T.; Ritz, F.; Belzner, L.; et al. 2021. Vast: Value function factorization with variable agent sub-teams. *Advances in Neural Information Processing Systems*, 34: 24018–24032.

Rashid, T.; Farquhar, G.; Peng, B.; et al. 2020a. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 10199–10210.

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

Samvelyan, M.; Rashid, T.; Schroeder de Witt, C.; et al. 2019. The StarCraft Multi-Agent Challenge. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, 2186–2188.

Schulman, J.; Wolski, F.; Dhariwal, P.; et al. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, J.; Lou, Z.; Zhang, H.; et al. 2022. Self-organized group for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 5711–5723.

Son, K.; Kim, D.; Kang, W. J.; et al. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proc. International Conference on Machine Learning*, volume 97, 5887–5896.

Sun, W.-F.; Lee, C.-K.; and Lee, C.-Y. 2021. DFAC framework: Factorizing the value function via quantile mixture for multi-agent distributional Q-learning. In *International Conference on Machine Learning*, 9945–9954. PMLR.

Sunehag, P.; Lever, G.; Gruslys, A.; et al. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, 2085–2087.

Wang, J.; Ren, Z.; Liu, T.; et al. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *Proc. International Conference on Learning Representations*.

Yang, Y.; Hao, J.; Liao, B.; et al. 2020. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*.

Ye, D.; Chen, G.; and Zhang, W. 2020. Towards Playing Full MOBA Games with Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, 621–632.

Yu, C.; Velu, A.; Vinitsky, E.; et al. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624.

Zang, Y.; He, J.; Li, K.; et al. 2023. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 46105–46121.

Zheng, L.; Chen, J.; Wang, J.; et al. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34: 3757–3769.

Zheng, Z.; and Gu, S. 2024. Safe multi-agent reinforcement learning with bilevel optimization in autonomous driving. *IEEE Transactions on Artificial Intelligence*, 6: 829–842.