

From General Relation Patterns to Task-Specific Decision-Making in Continual Multi-Agent Coordination

Chang Yao^{1,2}, Youfang Lin^{1,2}, Shoucheng Song^{1,2}, Hao Wu^{1,2}, Yuqing Ma³
Sheng Han^{1,2} and Kai Lv^{1,2*}

¹School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China

²Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing, China

³Institute of Artificial Intelligence, Beihang University, Beijing, China

{yaochang,yflin,insis_songsc,wuhao_,shhan,lvkai}@bjtu.edu.cn, mayuqing@buaa.edu.cn

Abstract

Continual Multi-Agent Reinforcement Learning (Co-MARL) requires agents to address catastrophic forgetting issues while learning new coordination policies with the dynamics team. In this paper, we delve into the core of Co-MARL, namely Relation Patterns, which refer to agents’ general understanding of interactions. In addition to generality, relation patterns exhibit task-specificity when mapped to different action spaces. To this end, we propose a novel method called General Relation Patterns-Guided Task-specific Decision-Maker (RPG). In RPG, agents extract relation patterns from dynamic observation spaces using a relation capturer. These task-agnostic relation patterns are then mapped to different action spaces via a task-specific decision-maker generated by a conditional hypernetwork. To combat forgetting, we further introduce regularization items on both the relation capturer and the conditional hypernetwork. Results on SMAC and LBF demonstrate that RPG effectively prevents catastrophic forgetting when learning new tasks and achieves zero-shot generalization to unseen tasks.

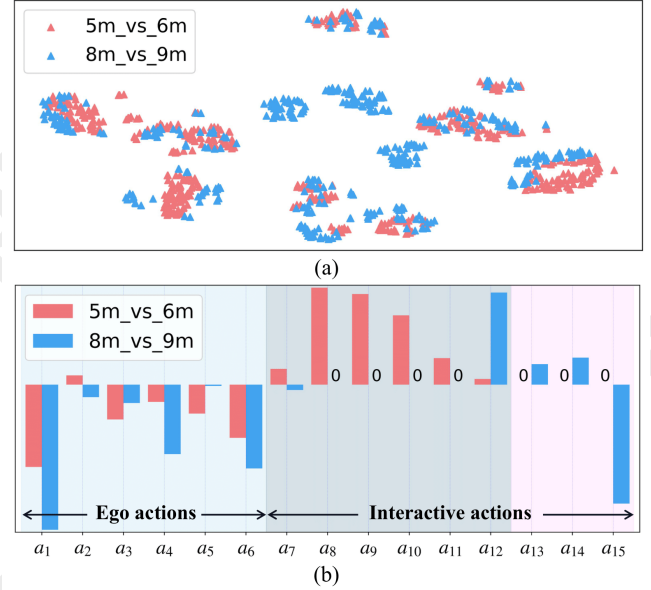


Figure 1: (a) Clustering visualization of relation patterns; (b) Agent’s action selection in the similarity relation patterns. The purple background represents the action space unique to the 8m_vs_9m.

1 Introduction

Collaborative Multi-Agent Reinforcement Learning (MARL) [Oroojlooy and Hajinezhad, 2023] focuses on how multiple agents can coordinate to achieve global goals. Traditional research mainly concentrates on enhancing agents’ cooperation abilities in stable environments [Rashid *et al.*, 2020b; Yu *et al.*, 2022]. However, in the real world, agents often face constant changes in diverse tasks [Kiran *et al.*, 2021; Yu *et al.*, 2021; Zhang *et al.*, 2024; Wang *et al.*, 2024a; Wang *et al.*, 2024b]. Particularly, in continual coordination scenarios, agents may lose the ability to perform well in previously learned tasks when adapting to new ones, a phenomenon known as catastrophic forgetting [Robins, 1995] in Continual Learning (CL).

CL seeks to exploit both the shared and the specific properties across different tasks. The properties reflect the essence

of continual learning across various domains. For instance, in the class-incremental learning domain, models should understand the overlap or complementarity between new and old classes in feature space, leveraging learned knowledge to adapt to new tasks [Ji *et al.*, 2023]. In the multimodal CL domain, models learn sample-specific and invariant features to construct representations that are broadly adaptable, to prevent catastrophic forgetting [Zhang *et al.*, 2023]. Compared to class-incremental and multimodal CL, the challenges in Continual Multi-Agent Reinforcement Learning (Co-MARL) are more complex due to the uncertain scale of entities. Thus, it is also crucial to figure out the essence of Co-MARL. In this paper, we argue that the essence of continual multi-agent coordination lies in **Relation Patterns** between tasks.

A **relation** represents the influence of an observable entity on an agent’s decision-making. Specifically, the relation between the agent and each entity is determined by the combination of the entity’s importance to the agent and the en-

*Corresponding author: Kai Lv.

tity’s features. In this paper, a **relation pattern** is formed by integrating the relations of multiple entities. Through experiments, we show that common relation patterns exist across distinct tasks. In Figure 1(a), we apply t-SNE [Van der Maaten and Hinton, 2008] to analyze relation patterns in two cooperative tasks. The red and blue triangles representing the two tasks are tightly clustered, indicating strong similarity in the relation patterns between the two tasks. As the entity scale increases from task 5m_vs.6m to task 8m_vs.9m in SMAC [Samvelyan *et al.*, 2019], some blue triangles form additional isolated clusters. This reflects that in a more complex observation space, agents can form more relation patterns.

Although the clustered relation patterns exhibit similarity across different tasks, the value of similar relation patterns on decision-making can vary significantly depending on the specific task context. For example, in the 8m_vs.9m, agents can win a game by capturing various “focus-fire” patterns (e.g., 4-on-2 or 3-on-2). However, in the 5m_vs.6m, the same “focus-fire” patterns may lead to a loss when applying the same policies. This indicates that identical relation patterns result in different returns across tasks. We validate this preference difference in Figure 1(b), where the agents demonstrate noticeable differences in their decision-making preferences (Q-value distribution) across two action spaces.

To **capture** and **exploit** these general relation patterns, we propose RPG(General Relation Patterns-Guided Task-specific Decision-Maker). 1) In **capturing** the relation patterns, we design a scalable relation capturer that allows agents to measure interactions with other entities from their first-person perspective, forming general relation patterns. We further add an anti-forgetting regularization to retain key relation patterns across tasks with varying entity numbers. 2) In **exploiting** the relation patterns, a task-specific decision-maker is introduced, enabling the agent to respond to dynamic threats and adapt to new collaborative teams. We use a conditioned hypernetwork to generate decision-makers for each task, mapping similar relation patterns to diverse action spaces for flexible adaptation. To ensure the stability of each decision-maker, we add a hypernetwork loss term that prevents forgetting prior cooperation while enabling fast adaptation to new teams. The contributions are as follows:

- We capture the essence of Co-MARL via relation patterns, linking dynamic observation and action spaces.
- We introduce a scalable Relation Capturer that helps agents extract relation patterns, preventing catastrophic forgetting and enhancing cross-task collaboration.
- We inject plasticity into decision-makers through the hypernetwork, enabling flexible decisions based on relation patterns in dynamic environments.
- RPG outperforms existing baselines in continual tasks, excelling on benchmarks like SMAC[Samvelyan *et al.*, 2019] and LBF[Papoudakis *et al.*, 2020].

2 Related Work

2.1 Continual Learning in RL and MARL

Continual Reinforcement Learning (CRL) aims to enable agents to learn sequential tasks while avoiding catastrophic

forgetting. Early approaches [Rolnick *et al.*, 2019; Chaudhry *et al.*, 2019] alleviate forgetting by reusing past experiences through replay buffers. Regularization-based methods [Kirkpatrick *et al.*, 2017; Aljundi *et al.*, 2018] mitigate forgetting by constraining parameter updates to ensure stable learning. Another class of methods tackles the stability-plasticity trade-off by dynamically adjusting model architectures, such as employing new networks [Kim *et al.*, 2023] or multi-head structures [Kessler *et al.*, 2022]. However, while these methods have shown progress in single-agent tasks, achieving effective continual learning in multi-agent tasks remains a significant challenge.

In Co-MARL, the complex interactions between agents make continual learning significantly more challenging. Agents must coordinate in diverse teams, requiring them to acquire new knowledge while retaining the ability to address previous tasks. To address these challenges, MALT [Shi *et al.*, 2021] employs progressive neural networks to facilitate knowledge transfer across tasks, while MACPro [Yuan *et al.*, 2024] leverages a shared feature extractor and dynamically expands policy heads to effectively prevent catastrophic forgetting. LL-Hanabi [Nekoei *et al.*, 2021] provides a benchmark for testing continual learning in multi-agent settings. MARR [Yang *et al.*, 2024] adopts a reset strategy to avoid the loss of network plasticity caused by high replay rates. However, these methods fail to thoroughly analyze the fundamental bottlenecks of Co-MARL. In this paper, our approach tackles the core challenges of Co-MARL, providing a stable and efficient solution to address catastrophic forgetting.

2.2 Hypernetworks in MARL

Hypernetworks have demonstrated strong performance in meta-reinforcement learning and continual learning [Beck *et al.*, 2023; Chandra *et al.*, 2023; Ehret *et al.*, 2021]. In MARL, QMIX [Rashid *et al.*, 2020b] introduces hypernetworks into the mixing network to enable effective credit assignment, while MAVEN [Mahajan *et al.*, 2019] is the first to leverage hypernetworks to generate decision-layer parameters, facilitating efficient exploration for agents. HPN [Jianye *et al.*, 2022] employs hypernetworks to achieve permutation equivariance and invariance, enabling it to handle observations with varying dimensions. Concord [Guan *et al.*, 2024] pioneers the use of hypernetworks for continual AI-human collaboration by generating agent parameters to adapt to diverse teammate strategies, though it incurs significant computational overhead and overlooks the challenge of varying agent numbers in continual tasks. In this work, we utilize hypernetworks to generate decision parameters for agents, providing the capability to learn new tasks and achieve stable continual coordination.

3 Problem Formalization

In a sequence of continual multi-agent coordination tasks, all agents can only access partial local observations. We model this setting as a Continual Decentralized Partially Observable Markov Decision Process (Co-Dec-POMDP) [Oliehoek *et al.*, 2016]. The Co-Dec-POMDP for each task is formal-

ized as a tuple:

$$\mathcal{Z}_c = \langle c, \mathcal{N}_c, \mathcal{M}_c, \mathcal{S}_c, \mathcal{O}_c, \mathcal{A}_c, \mathcal{R}_c, \mathcal{P}_c, \gamma \rangle,$$

where, in task c , \mathcal{M}_c denotes the set of all entities in the environment, and $\mathcal{N}_c \subseteq \mathcal{M}_c$ represents the set of agents. At each time step t , the global state $s_t \in \mathcal{S}_c$ represents the complete state of the environment. Each agent $i \in \mathcal{N}_c$ receives a local observation $o_t^i \in \mathcal{O}_c$ and selects an action $a_t^i \in \mathcal{A}_c$. The joint action $\mathbf{a}_t = \{a_t^i\}_{i=1}^{|\mathcal{N}_c|}$ is then executed, and the environment transitions to the next state s_{t+1} according to the state transition function $\mathcal{P}_c(s_{t+1} | s_t, \mathbf{a}_t)$, while generating a shared reward $r_t \in \mathcal{R}_c$. The objective of the multi-agent system is to maximize the global value function for the current task: $Q_{tot}(s, \mathbf{a}) = \mathbb{E}_{s_t, \mathbf{a}_t} [\sum_{t=0}^{\infty} \gamma^t r(s_t, \mathbf{a}_t)]$, where $\gamma \in [0, 1]$ is the discount factor, and π represents the joint policy.

In a Co-Dec-POMDP, when a new task $c + 1$ arrives, the entity composition changes from \mathcal{M}_c to \mathcal{M}_{c+1} , resulting in corresponding dimensional changes in \mathcal{S} , \mathcal{O} , and \mathcal{A} . Within our continual learning framework, agents are designed to handle these dynamically changing spaces, with the objective of maximizing the global reward in the current task while maintaining the performance of the current agent network across all previously encountered tasks.

4 Method

We believe the core of Co-MARL lies in the relation patterns that manifest distinct properties across different spaces, i.e., observation and action spaces. On the one hand, general relation patterns emerge across tasks with varying numbers of entities, facilitated by the decomposition and combination of the observation space in an entity-wise manner. On the other hand, even similar relation patterns may yield different values when mapped to distinct action spaces. To address this, we propose a General Relation Patterns-guided Task-specific Decision-Maker (RPG) framework. First, to capture and preserve the general agent-to-entity relations from the dynamic observation space, we design a scalable relation capturer with stability. Next, we inject plasticity into the decision-maker using a hypernetwork, enabling it to map the general relation patterns to task-specific action spaces.

4.1 Scalable Relation Capturer with Stability

In sequential Co-MARL tasks, extracting general relation patterns and maintaining their stability is crucial. To this end, we develop a scalable relation capturer to dynamically capture the interaction relationships between entities. The scalable relation capturer has two components: attention-based relation patterns and return-aware anti-forgetting.

Attention-Based Relation Patterns. To extract relation patterns at the entity level, we model the observation with non-fixed dimensions as a combination of multiple entity features. Consider a multi-agent system with m entities. Given the agent’s observation $o_{t,i}$ at time step t , we decompose $o_{t,i}$ into the agent’s own features $o_{t,i}^{\text{se}}$, $n - 1$ teammate entity features $o_{t,i}^{\text{te}}$, and $m - n$ other entity features $o_{t,i}^{\text{oe}}$.

Since the relation patterns are independent of the position of entities in the sequence, a permutation-invariant cross-attention mechanism can be used to capture general relation

patterns. Specifically, as shown in Figure 2(a), we treat $o_{t,i}^{\text{se}}$ as the query, while $o_{t,i}^{\text{te}}$ or $o_{t,i}^{\text{oe}}$ serve as the keys and values, respectively. We then calculate the attention scores for teammate entities and other entities, corresponding to collaborative partners and potential threats or targets, respectively.

$$\alpha_{t,i}^{\text{te}} = \text{Softmax}\left(\frac{qK_{\text{te}}^\top}{\sqrt{d_x}}\right), \alpha_{t,i}^{\text{oe}} = \text{Softmax}\left(\frac{qK_{\text{oe}}^\top}{\sqrt{d_x}}\right), \quad (1)$$

where $\alpha_{t,i}$ represents the attention level of agent i towards teammates and other entities, and $\sqrt{d_x}$ is the scaling coefficient. Additionally, we use $\alpha_{t,i}$ to perform a weighted sum of values, thereby obtaining the relation pattern from the agent’s first-person perspective to teammates and others, as shown below:

$$z_{t,i}^{\text{te}} = \alpha_{t,i}^{\text{te}} V_{\text{te}}, z_{t,i}^{\text{oe}} = \alpha_{t,i}^{\text{oe}} V_{\text{oe}}. \quad (2)$$

The relation pattern $z_{t,i}$ is formed by concatenating the outputs from two cross-attention modules. Furthermore, to establish closer relation patterns with more relevant entities, inspired by MAIC [Yuan *et al.*, 2022] and CoDe [Song *et al.*, 2025], we introduce a regularization term to sparsify the attention scores:

$$\mathcal{L}_{\text{att}} = \frac{1}{T} \sum_{t=1}^T \mathcal{H}(\alpha_t) = -\frac{1}{T} \sum_{t=1}^T \sum_i (\alpha_{t,i} \log \alpha_{t,i}). \quad (3)$$

By minimizing the entropy of the attention scores α_i , we encourage the agent to focus on several key entities.

Finally, to capture the agent’s information, we encode its own features $o_{t,i}^{\text{se}}$ as $z_{t,i}^{\text{se}}$, which is then concatenated with the relation pattern to form $z_{t,i} = [z_{t,i}^{\text{se}}, z_{t,i}^{\text{te}}, z_{t,i}^{\text{oe}}]$. It is then fed into the GRU along with the previous hidden state to generate the historical relation pattern $h_{t,i} = \text{GRU}(z_{t,i}, h_{t-1,i})$, which serves as the output of the scalable relation capturer.

Return-Aware Anti-Forgetting. Although we can extract common relation patterns across different tasks, excessive parameter updates of the capturer can lead to the forgetting of past relation patterns. Therefore, applying varying degrees of constraints to the different parameters ensures the stability of the capturer. Inspired by Taylor Pruning [Molchanov *et al.*, 2019], under the i.i.d. assumption of parameters, we measure the importance of a parameter θ_i by the squared difference in the TD loss when θ_i is set to zero:

$$\mathcal{L}_{TD} = \mathbb{E} \left[(Q_{tot}(s, \mathbf{a}) - (r + \gamma \max_{\mathbf{a}'} Q_{tot}(s', \mathbf{a}'))^2 \right], \quad (4)$$

$$I(\theta_j) = (\mathcal{L}_{TD} - \mathcal{L}_{TD}(\theta_j = 0))^2. \quad (5)$$

Directly computing the parameter importance using Equation (5) is computationally expensive. Therefore, we approximate the importance of θ_j^c for the c -th task using a Taylor expansion. The second-order Taylor expansion is as follows:

$$I(\theta_j^c) = \left(\left(\frac{\partial \mathcal{L}_{TD}}{\partial \theta_j^c} \right) \theta_j^c + \frac{1}{2} \left(\frac{\partial^2 \mathcal{L}_{TD}}{\partial (\theta_j^c)^2} \right) (\theta_j^c)^2 \right)^2. \quad (6)$$

We typically retain only the quadratic term in Equation (6). The final average importance of θ^c is then given by:

$$I(\theta_j^c) = \frac{1}{N} \sum_{n'=1}^N \left(\left(\frac{\partial \mathcal{L}_{TD}^c}{\partial \theta_j^c} \right) \theta_j^c \right)^2 / I(\theta^c)_{\text{max}}, \quad (7)$$

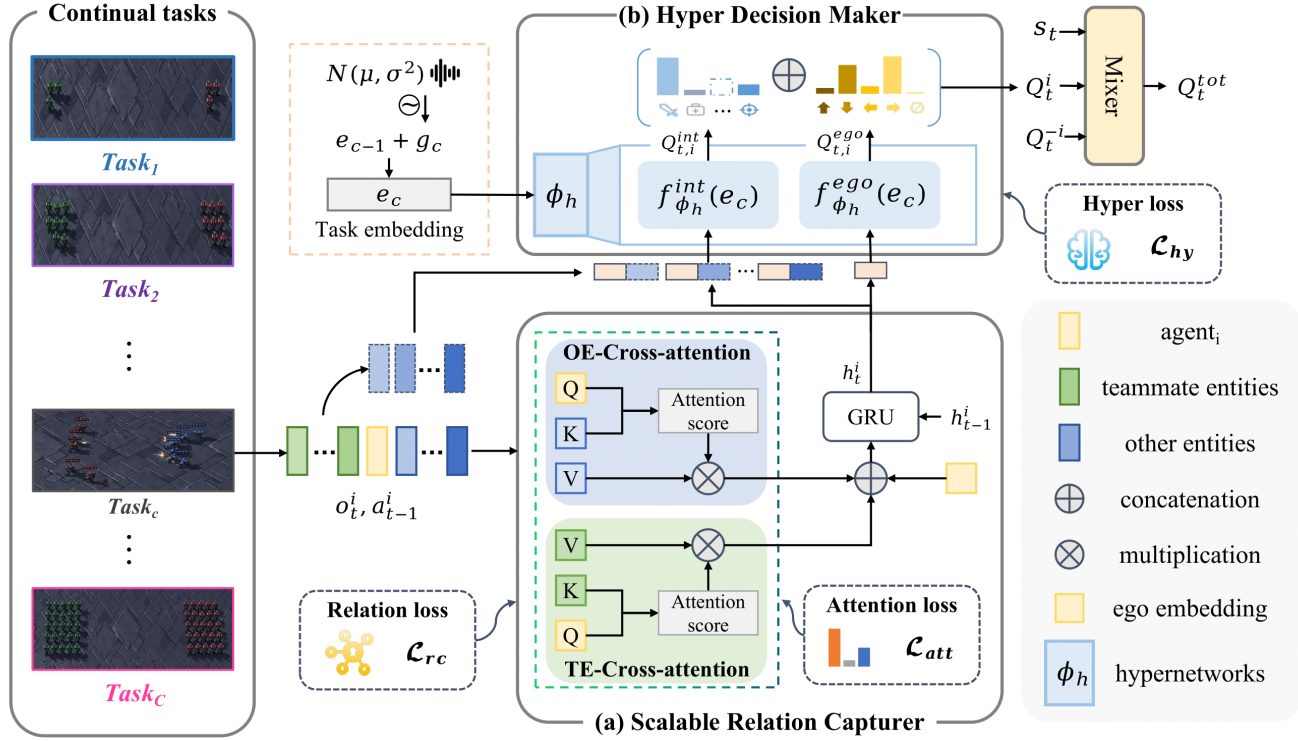


Figure 2: An overview of the RPG framework. In the c -th task, the observations are input into the (a) **Scalable Relation Capturer**, which captures the historical relational patterns. These captured patterns are then fed into the (b) **Hyper Decision-Maker**, where the hypernetworks generate the decision-makers that map the relation patterns to local Q-values.

where N denotes the number of samples from the previous buffer, and all parameter importance values of ϕ_{rc} are normalized by dividing by the maximum $I(\theta^c)_{\max}$. Specifically, for the importance matrices $I(\theta^1), I(\theta^2), \dots, I(\theta^{c-1})$ of the past $c-1$ tasks, the parameter regularization item can be expressed as the sum of the regularization losses for the relation capturers of each task:

$$\mathcal{L}_{rc} = \frac{\lambda_{rc}}{2} \sum_{c'=1}^{c-1} \sum_{j=1}^{|\phi_{rc}|} I(\theta_j^{c'}) (\theta_j^* - \theta_j^{c'})^2, \quad (8)$$

where λ_{rc} represents the regularization strength for the relation capturer parameters ϕ_{rc} . The learning process can be stabilized by penalizing inconsistencies in the parameters, which helps the agent retain key relation patterns and avoid catastrophic forgetting.

In addition, overly conservative regularization can prevent the agent from extracting all possible relation patterns in more complex tasks. During the early stages of learning a new task, the team's return reflects its ability to solve the current problem. Therefore, we use the team's normalized cumulative return over the first N episodes as the discount factor for \mathcal{L}_{rc} :

$$\gamma_p = \frac{1}{N} \sum_{n'=1}^N G_{\text{norm}}^{n'}, \quad (9)$$

where $G_{\text{norm}}^{n'}$ denotes the raw cumulative return of the n' -th episode. A larger γ_p indicates that the current policy is well-suited to the current task, while a smaller γ_p suggests relaxing

the constraints in the new observation space to learn more potential coordination forms.

4.2 Hyper-Decision-Maker with Plasticity

Another key challenge in Co-MARL is that general relation patterns exhibit varying values across continual tasks. Agents must learn a new decision-making policy to adapt to changing action spaces. Therefore, we introduce a task-conditioned hypernetwork, which aims to generate task-specific decision-makers that can make rapid behavioral adjustments based on the captured relation patterns.

Using a traditional MLP as the policy layer may lead to a loss of plasticity in the agent. Instead, we utilize a hypernetwork to generate the decision-maker, which further maps the historical relation pattern to the action space. When learning the first task, we use a randomly initialized, Gaussian-distributed, learnable task embedding e_1 as the input to the hypernetwork ϕ_h .

The hypernetwork generates an ego-decision-maker (right) and an interaction-decision-maker (left), as shown in Figure 2(b), corresponding to fixed ego actions and variable interaction actions, respectively. For ego actions, we use the hypernetwork ϕ_h^{ego} to generate the ego-decision-maker $f_{\phi_h}^{\text{ego}}(e_1)$, and then map the historical relation pattern $h_{t,i}$ to a fixed-dimension $Q_{t,i}(a_{\text{ego}}|o_{t,i})$. For interactive actions, directly mapping $h_{t,i}$ to variable-dimension $Q_{t,i}(a_{\text{int}}|o_{t,i})$ is not feasible. Therefore, we design an entity-wise interaction decision-maker that focuses on entity interactions. By concatenating

the historical relation pattern $h_{t,i}$ with the features of $m - n$ other entities $e_{t,i}^{\text{oe}}$, we map them to each interaction entity's $Q_{t,i}(a_{\text{int}}(\cdot)|o_{t,i})$. Finally, we concatenate the state-action values from both parts to obtain the complete local state-action $Q_{t,i}(a|o_{t,i})$.

In addition, the hypernetwork itself faces the issue of catastrophic forgetting across tasks. We introduce an additional regularization term for the hypernetworks:

$$\mathcal{L}_{\text{hy}} = \frac{\lambda_{\text{hy}}}{c-1} \sum_{c'=1}^{c-1} \left\| f_{\phi_h}(e_{c'}) - f_{\tilde{\phi}_h}(e_{c'}) \right\|_2^2, \quad (10)$$

where $\tilde{\phi}_h$ refers to the hypernetworks obtained after the $c-1$ -th task. By minimizing the above loss term, we encourage the hyper-decision-maker to learn new behaviors while preserving the decision-maker specific to previous tasks. Furthermore, to ensure a smooth task transition, we initialize the task embedding e_c for the new task as follows:

$$e_c = \alpha_{\text{init}} e_{c-1} + (1 - \alpha_{\text{init}}) g_c, \quad (11)$$

where α_{init} controls the degree of retention of the previous task's embedding, and g_c is the embedding initialized from a Gaussian distribution. This soft initialization method allows the agent to inherit knowledge from previous tasks and smoothly transition to the new task.

4.3 Overall Algorithm

In this section, we provide an overall description of RPG and present the pseudocode of our algorithm in the Appendix.

In the training phase of each task, we train the overall agent network and mixing network using the loss terms in Equation (3) and (4). Additionally, We integrate a multi-head attention module into the mixing network to accommodate the varying state spaces across tasks. This design facilitates the agent's ability to learn collaborative behaviors from a global perspective. A detailed explanation of the mixing network is provided in the Appendix.

Furthermore, during the training phase of the c -th ($c > 1$) task, to prevent catastrophic forgetting, we introduce regularization loss terms for the relation capturer and decision-maker. The final overall loss function is as follows:

$$\mathcal{L}_{\text{tot}}^c = \mathcal{L}_{TD} + \alpha_{\text{att}} \mathcal{L}_{\text{att}} + \mathbf{1}_{c>1} \gamma_p \mathcal{L}_{\text{rc}} + \mathbf{1}_{c>1} \mathcal{L}_{\text{hy}}, \quad (12)$$

where α_{att} is a coefficient used to balance the various loss terms. Notably, $\mathbf{1}_{c>1}$ is an indicator function that takes the value 0 in the first task ($c = 1$) and 1 otherwise. Additionally, as tasks are trained sequentially, we evaluate the current agent network on the prior $c-1$ tasks to assess the anti-forgetting effect of RPG.

5 Experiments

To validate the effectiveness of RPG, we design comprehensive experiments to address the following questions: 1) Does our method outperform other baselines in learning new knowledge while preventing catastrophic forgetting? 2) How does each component of our approach impact continual learning performance? 3) Does the relation pattern extracted by RPG facilitate better multi-agent coordination compared to other baselines? 4) Does the general relation pattern captured by RPG contribute to zero-shot generalization?

5.1 Experimental Setups

Benchmarks. The experiments in this study are conducted using the Level-Based Foraging (LBF) [Papoudakis *et al.*, 2020] and the StarCraft Multi-Agent Challenge (SMAC) [Samvelyan *et al.*, 2019] benchmarks. The former involves multiple agents collaborating in a 2D grid to collect scattered food items, where the sum of the agents' levels must exceed the food's level to successfully collect it and receive a reward. The latter is a complex real-time strategy game that requires multiple agents to cooperate to achieve objectives, such as destroying enemy units or protecting allied ones. SMAC provides a dynamic state space, heterogeneous unit types, and diverse map settings, effectively evaluating the agents' abilities in both cooperation and competition.

Baselines. To evaluate whether RPG performs well on these benchmark tasks as they continuously appear, we apply it to a popular value-based method, QMIX [Rashid *et al.*, 2020b]. Additionally, we use MACPro [Yuan *et al.*, 2024], EWC [Kirkpatrick *et al.*, 2017], MAS [Aljundi *et al.*, 2018], Replay, L2, and Finetuning as baselines, where MACPro is the latest method in Co-MARL.

Furthermore, to validate the role of relation patterns in multi-agent cooperation, we combine the relation capturer with QMIX and compare it with common value-based and policy-based methods. Specifically, in SMAC, we use methods such as VDN [Sunehag *et al.*, 2017], QMIX [Rashid *et al.*, 2020b], WQMIX [Rashid *et al.*, 2020a], QTRAN [Son *et al.*, 2019], ResQ [Shen *et al.*, 2022], HPN [Jianye *et al.*, 2022], and FtQMIX [Hu *et al.*, 2023] as baselines. Additionally, we explore policy-based methods like MAA2C, COMA [Foerster *et al.*, 2018], MADDPG [Lowe *et al.*, 2017], MAPPO [Yu *et al.*, 2022], and IPPO [De Witt *et al.*, 2020] in LBF.

5.2 Continual-tasks Performance and Ablation

CL Performance Comparison. In Figure 3, we present a comprehensive result of the continual learning experiments conducted in SMAC, highlighting the superior performance of RPG. RPG outperforms other methods in both mitigating catastrophic forgetting and learning new knowledge. Specifically, the gray background indicates the current task and the training procedure synchronously tests all previous tasks using the current agent parameters. The results show that methods such as EWC, Replay, and MAS experience a performance drop after learning the first task, and fail to achieve optimal performance when learning new tasks, exhibiting weak capabilities in mitigating catastrophic forgetting and acquiring new knowledge. This indicates that directly applying CL methods to MARL settings is not feasible. Finetuning and L2 represent the upper bounds for plasticity and stability, respectively. However, Finetuning exhibits the worst catastrophic forgetting performance as it rapidly forgets knowledge after each task is learned, while L2 regularization, due to its excessive constraints, prevents the agent from adapting to task changes.

The Co-MARL method MACPro, while performing well among the baselines, is inferior to RPG in both single-task performance and plasticity. RPG adapts to 8m.vs.9m training

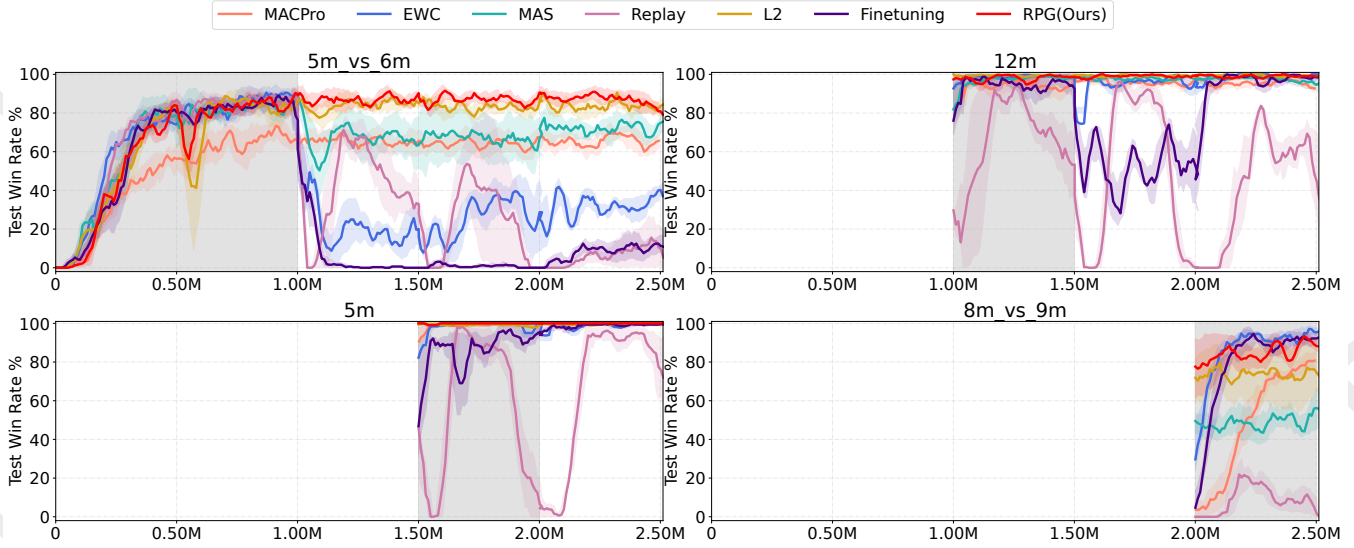


Figure 3: The complete continual learning results in StarCraft II. The tasks appear in the order of {5m_vs_6m, 12m, 5m, 8m_vs_9m}. The gray background indicates that the task is in the training phase, and the blank areas represent tasks that have not yet been encountered. The curves after the gray background represent the performance of the agent trained later when tested on that task. Each curve represents the average win rate across 3 random seeds.

Methods	5m_vs_6m	12m	5m	8m_vs_9m
RPG	0.875	1.00	1.00	0.91
w/o \mathcal{L}_{rc}	0.86	1.00	1.00	0.95
w/o \mathcal{L}_{hy}	0.84	1.00	1.00	0.875
w/o $\mathcal{L}_{rc}, \mathcal{L}_{hy}$	0.865	1.00	1.00	0.86
RPG(EWC)	0.875	1.00	1.00	0.93
RPG(MLP)	0.84	1.00	1.00	0.59

Table 1: Ablation analysis of plasticity performance. “w/o \mathcal{L} ” indicates the removal of \mathcal{L} from the RPG, and “(·)” denotes the use of · as a replacement.

faster than MACPro. Moreover, RPG’s agent parameter size (112K) is smaller than MACPro’s (377K), and MACPro’s dynamic policy head further increases the parameter size. Overall, RPG demonstrates comparable performance to L2 in anti-forgetting and is more adaptable to new tasks. In addition, we compare RPG with several mentioned baselines on the continual tasks of LBF, as detailed in the Appendix.

Ablation Study. In this section, we investigate the impact of different components in RPG through an ablation study conducted on SMAC. First, we remove the loss term \mathcal{L}_{rc} from the scalable relation capturer to obtain w/o \mathcal{L}_{rc} , in order to study its effect on preventing catastrophic forgetting. Next, we replace \mathcal{L}_{rc} with EWC in RPG (denoted as RPG(EWC)) and apply it to the relation capturer to validate the effectiveness of our parameter importance estimation method. Then, in w/o \mathcal{L}_{hy} , we eliminate the hypernetwork regularization loss term \mathcal{L}_{hy} to investigate whether a free hypernetwork could still facilitate continual learning. Additionally, we introduce RPG(MLP), which replaces the decision-maker generated by the hypernetwork with an MLP for decision-making. Finally,

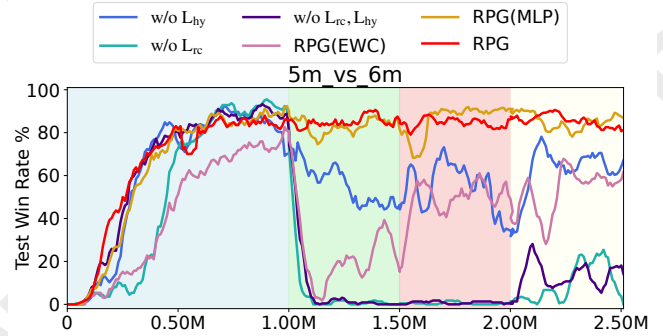


Figure 4: Ablation analysis of anti-forgetting performance. The 4 different colored backgrounds correspond to the training processes of {5m_vs_6m, 12m, 5m, and 8m_vs_9m}. Each curve represents the agent’s average win rate on 5m_vs_6m at distinct testing stages.

we remove both \mathcal{L}_{rc} and \mathcal{L}_{hy} from RPG, and we refer to this setup as w/o $\mathcal{L}_{rc}, \mathcal{L}_{hy}$.

Continual learning aims to balance stability and plasticity, requiring joint analysis of Table 1 and Figure 4 for a comprehensive assessment. In Figure 4, we test the algorithm on 5m_vs_6m at four stages during training to evaluate its forgetting resistance. Table 1 shows the algorithm’s performance at different stages of task completion. We find that w/o \mathcal{L}_{rc} , lacking regularization, improves plasticity, but its performance rapidly degrades when the second task begins. This occurs because relaxing the relation capture constraints enhances the agent’s learning capacity but compromises its memory retention. Similarly, RPG(EWC) shows a similar performance drop, confirming the reliability of our parameter importance estimation. Removing \mathcal{L}_{hy} allows w/o \mathcal{L}_{hy} to continue learning new tasks, but its forgetting resistance

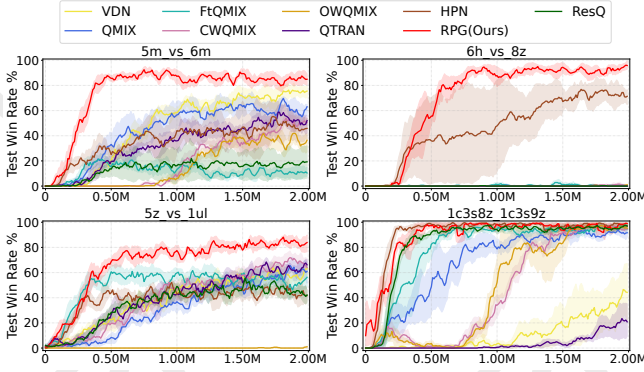


Figure 5: Algorithms Performance in 4 super-hard SMAC tasks.

Tasks	RPG	EWC	MAS
3m	86.5 ± 8.3	47.7 ± 13.9	91.6 ± 8.5
8m	100	95.3 ± 5.4	99.6 ± 1.5
10m	100	87.3 ± 7.9	99.6 ± 1.5
10m_vs_11m	90.3 ± 7.9	89.8 ± 6.6	63.1 ± 15.3
15m	96.3 ± 4.9	73.2 ± 10.6	76.9 ± 16.6

Table 2: Zero-shot results. Bold indicates the best performance.

declines, highlighting the need for a constrained hypernetwork. Notably, while RPG(MLP) maintains coordination on 5m_vs_6m, it loses some plasticity, emphasizing the necessity of task-specific decision-makers. This is because the MLP cannot isolate the decision-makers of different tasks. Finally, w/o \mathcal{L}_{rc} , \mathcal{L}_{hy} loses the ability to resist forgetting.

5.3 Single-task Performance

In this section, we conduct comparative experiments to validate the coordination performance of our method in single tasks. Figure 5 and 6 show the comparison of RPG with baselines in SMAC and LBF.

In several super-hard maps of SMAC, RPG achieves the best performance, demonstrating that agents dynamically selecting interaction relationships with different teammates or opponents based on their own states is more conducive to achieving better coordination. In the LBF, particularly in maps with more entities and larger areas, RPG outperforms other baselines. This advantage is attributed to the ability of the relation pattern capturer to quickly identify key interactive entities in large-scale tasks, enabling efficient collaboration. These results indicate that the relation pattern is not only effective in complex, continual multi-agent cooperation tasks but also exhibits superior performance in more challenging single-task scenarios.

As shown in Figure 7, we conduct an ablation study on the sparse attention mechanism using w/o \mathcal{L}_{att} to analyze its impact on the generation of relational patterns by agents. The results indicate that removing sparse attention slightly reduces the performance of RPG, confirming that interacting with a few key entities facilitates better collaboration among agents.

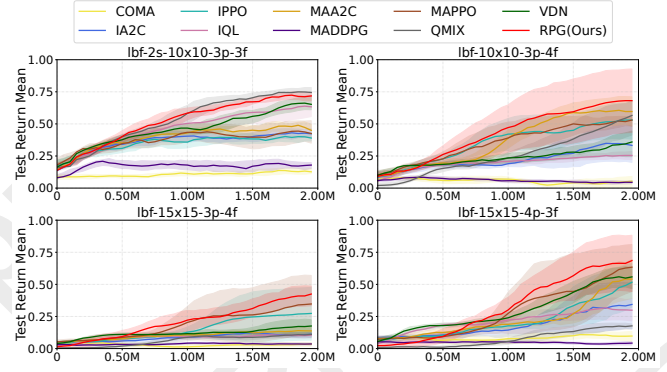


Figure 6: Algorithms Performance in 4 LBF tasks.

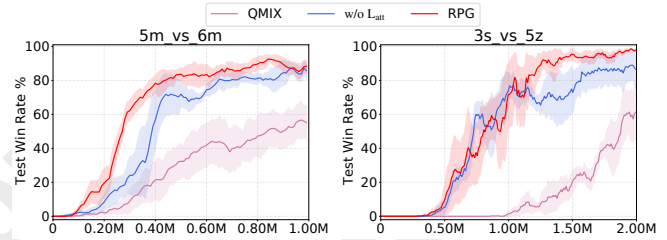


Figure 7: Ablation results of the sparse attention loss term.

5.4 Zero-shot Performance

In addition to learning sequential tasks, agents should also learn to leverage prior coordination experience to handle unseen tasks. To end this, we compare RPG with EWC and MAS to assess its zero-shot generalization capability in tackling novel tasks. Specifically, in the SMAC scenarios, agents trained on continual tasks {5m_vs_6m, 12m, 5m, 8m_vs_9m} are tested on 5 additional unseen tasks to validate the generalization performance. As shown in Table 2, compared to other baselines, RPG demonstrates superior generalization performance in 4 out of the 5 tasks. However, RPG’s generalization ability on 3m does not match that of MAS. This indicates that in simple tasks with few agents, complex relation patterns may instead hinder simple cooperation. This highlights the ability of RPG to extract and retain generalizable relation patterns, enabling robust generalization to new tasks.

6 Conclusion

This paper focuses on the problem of continual coordination in MARL. To address this issue, we analyze the essence of continual multi-agent coordination—relation patterns—and propose a novel framework, RPG. RPG extracts general relation patterns from dynamic observation spaces and maps them to varying action spaces. Results demonstrate that RPG consistently outperforms baselines across various scenarios. However, RPG focuses solely on homogeneous multi-agent collaborative tasks and currently lacks the capability for perpetual knowledge acquisition. In future work, we aim to address these challenges. We hope RPG inspires further interest in continual multi-agent coordination research.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 2024XKRC080) and the National Key Laboratory of Air-based Information Perception and Fusion (Grant No. 202300010M5001).

References

- [Aljundi *et al.*, 2018] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.
- [Beck *et al.*, 2023] Jacob Beck, Matthew Thomas Jackson, Risto Vuorio, and Shimon Whiteson. Hypernetworks in meta-reinforcement learning. In *Conference on Robot Learning*, pages 1478–1487. PMLR, 2023.
- [Chandra *et al.*, 2023] Dupati Srikanth Chandra, Sakshi Varshney, PK Srijith, and Sunil Gupta. Continual learning with dependency preserving hypernetworks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2339–2348, 2023.
- [Chaudhry *et al.*, 2019] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.
- [De Witt *et al.*, 2020] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makovychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.
- [Ehret *et al.*, 2021] Benjamin Ehret, Christian Henning, Maria R. Cervera, Alexander Meulemans, Johannes von Oswald, and Benjamin F. Grewe. Continual learning in recurrent neural networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.
- [Foerster *et al.*, 2018] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [Guan *et al.*, 2024] Cong Guan, Feng Chen, Ke Xue, Chunpeng Fan, Lichao Zhang, Ziqian Zhang, Pengyao Zhao, Zongzhang Zhang, Chao Qian, Lei Yuan, et al. One by one, continual coordinating with humans via hyper- teammate identification. 2024.
- [Hu *et al.*, 2023] Jian Hu, Siying Wang, Siyang Jiang, and Weixun Wang. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. In *The Second Blogpost Track at ICLR 2023*, 2023.
- [Ji *et al.*, 2023] Zhong Ji, Zhishen Hou, Xiyao Liu, Yanwei Pang, and Xuelong Li. Memorizing complementation network for few-shot class-incremental learning. *IEEE Transactions on Image Processing*, 32:937–948, 2023.
- [Jianye *et al.*, 2022] HAO Jianye, Xiaotian Hao, Hangyu Mao, Weixun Wang, Yaodong Yang, Dong Li, Yan Zheng, and Zhen Wang. Boosting multiagent reinforcement learning via permutation invariant and permutation equivariant networks. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Kessler *et al.*, 2022] Samuel Kessler, Jack Parker-Holder, Philip Ball, Stefan Zohren, and Stephen J Roberts. Same state, different task: Continual reinforcement learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7143–7151, 2022.
- [Kim *et al.*, 2023] Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11930–11939, 2023.
- [Kiran *et al.*, 2021] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.
- [Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.
- [Mahajan *et al.*, 2019] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. Maven: Multi-agent variational exploration. *Advances in neural information processing systems*, 32, 2019.
- [Molchanov *et al.*, 2019] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272, 2019.
- [Nekoei *et al.*, 2021] Hadi Nekoei, Akilesh Badri-naaraayanan, Aaron Courville, and Sarath Chandar. Continuous coordination as a realistic scenario for life-long learning. In *International Conference on Machine Learning*, pages 8016–8024. PMLR, 2021.
- [Oliehoek *et al.*, 2016] Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

- [Oroojlooy and Hajinezhad, 2023] Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.
- [Papoudakis *et al.*, 2020] Georgios Papoudakis, Filippas Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.
- [Rashid *et al.*, 2020a] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020.
- [Rashid *et al.*, 2020b] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.
- [Robins, 1995] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.
- [Rolnick *et al.*, 2019] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.
- [Samvelyan *et al.*, 2019] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.
- [Shen *et al.*, 2022] Siqi Shen, Mengwei Qiu, Jun Liu, Weiquan Liu, Yongquan Fu, Xinwang Liu, and Cheng Wang. Resq: A residual q function-based approach for multi-agent reinforcement learning value factorization. *Advances in Neural Information Processing Systems*, 35:5471–5483, 2022.
- [Shi *et al.*, 2021] Haobin Shi, Jingchen Li, Jiahui Mao, and Kao-Shing Hwang. Lateral transfer learning for multi-agent reinforcement learning. *IEEE Transactions on Cybernetics*, 53(3):1699–1711, 2021.
- [Son *et al.*, 2019] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*, pages 5887–5896. PMLR, 2019.
- [Song *et al.*, 2025] Shoucheng Song, Youfang Lin, Sheng Han, Chang Yao, Hao Wu, Shuo Wang, and Kai Lv. Code: Communication delay-tolerant multi-agent collaboration via dual alignment of intent and timeliness. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pages 23304–23312, 2025.
- [Suneag *et al.*, 2017] Peter Suneag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
- [Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [Wang *et al.*, 2024a] Shuo Wang, Zhihao Wu, Xiaobo Hu, Jinwen Wang, Youfang Lin, and Kai Lv. What effects the generalization in visual reinforcement learning: policy consistency with truncated return prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5590–5598, 2024.
- [Wang *et al.*, 2024b] Shuo Wang, Zhihao Wu, Jinwen Wang, Xiaobo Hu, Youfang Lin, and Kai Lv. How to learn domain-invariant representations for visual reinforcement learning: an information-theoretical perspective. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1389–1397, 2024.
- [Yang *et al.*, 2024] Yaodong Yang, Guangyong Chen, HAO Jianye, and Pheng-Ann Heng. Sample-efficient multiagent reinforcement learning with reset replay. In *Forty-first International Conference on Machine Learning*, 2024.
- [Yu *et al.*, 2021] Tian Yu, Jing Huang, and Qing Chang. Optimizing task scheduling in human-robot collaboration with deep multi-agent reinforcement learning. *Journal of Manufacturing Systems*, 60:487–499, 2021.
- [Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitzky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.
- [Yuan *et al.*, 2022] Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, Zongzhang Zhang, Yang Yu, and Chongjie Zhang. Multi-agent incentive communication via decentralized teammate modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9466–9474, 2022.
- [Yuan *et al.*, 2024] Lei Yuan, Lihe Li, Ziqian Zhang, Fuxiang Zhang, Cong Guan, and Yang Yu. Multiagent continual coordination via progressive task contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [Zhang *et al.*, 2023] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023.
- [Zhang *et al.*, 2024] Hengrui Zhang, Youfang Lin, Shuo Shen, Sheng Han, and Kai Lv. Enhancing off-policy constrained reinforcement learning through adaptive ensemble c estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21770–21778, 2024.