GRDC: A Unified Graph-Driven Framework for Role Discovery and Communication in Multi-Agent Reinforcement Learning

Zihong Gao[1], Hongjian Liang[1], Yuanhui Hao[1], Lei Hao[1], Liangjun Ke[1]*

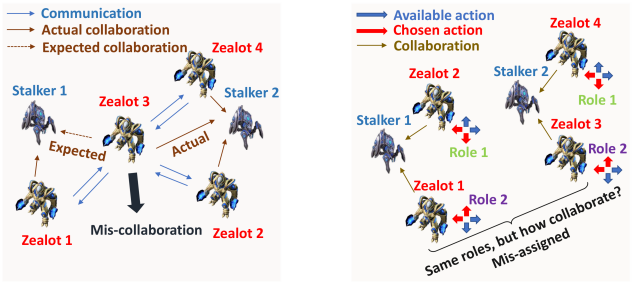
[1]The State Key Laboratory for Manufacturing Systems Engineering,
School of Automation Science and Engineering, Xi'an Jiaotong University g1246830895@stu.xjtu.edu.cn,
lianghj@stu.xjtu.edu.cn, haoyuanhui@stu.xjtu.edu.cn, lei_hao@stu.xjtu.edu.cn, keljxjtu@xjtu.edu.cn

Abstract

Effective coordination in Multi-Agent Reinforcement Learning (MARL) is particularly challenging under partial observability, where agents must identify and coordinate with task-relevant collaborators based solely on local information. Existing methods can be categorised into communication-based approaches, which allow message exchange but either rigidly predefine or misidentify collaborators, and role-based approaches, which promote functional specialization based on observed behavioural similarity. However, both paradigms overlook the dynamic and context-specific cooperative dependencies induced by the task, which determine which agents should collaborate, thereby leading to miscommunication or role misassignment under partial observability. We introduce GRDC (Graph-driven Role Discovery and Communication), a unified framework that approximates these dependencies by dynamically constructing local interaction graphs from trajectory embeddings, then uses these graphs to infer roles via prototype matching and to restrict communication to intra-role agents with attention-based aggregation. In addition to role inference and communication, GRDC promotes a structured and compact role space by maximising role entropy, decorrelating prototypes, and dynamically pruning redundant prototypes. Experimental results on Predator Prey, Cooperative Navigation, and SMACv2 demonstrate that GRDC consistently outperforms state-of-the-art communication- and role-based baselines, improving coordination efficiency and training stability across tasks.

Introduction

Multi-Agent Reinforcement Learning (MARL) has advanced numerous complex sequential decision-making tasks that require multi-agent coordination, such as game AI (Vinyals et al. 2019), autonomous driving (Kiran et al. 2021), and traffic signal control (Wang et al. 2021). A core challenge of MARL is enabling effective coordination under partial observability, where each agent relies exclusively on local observations (Nguyen, Nguyen, and Nahavandi 2020). Limited observability prevents agents from identifying potential collaborators, namely those whose behaviour is functionally coupled with theirs under the given task dynamics.



(a) Miscommunication. (b) Misassignment.

Figure 1: Two coordination failures in MARL.

The absence of collaborator information impedes the emergence of coordinated behaviour and degrades overall system performance.

Existing efforts address this challenge through two main research lines. Communication-based approaches, exemplified by CommNet (Sukhbaatar, Fergus et al. 2016), DIAL (Foerster et al. 2016) and ATOC (Jiang and Lu 2018), learn differentiable message-passing mechanisms that enable agents to share internal states or local observations, thereby improving situational awareness and reducing decision-making uncertainty (Das et al. 2019). In contrast, role-based approaches, such as ROMA (Wang et al. 2020a) and RODE (Wang et al. 2020b), learn functional roles during training, thereby promoting task specialization and structured cooperation. Although the two paradigms differ in implementation, both aim to model and exploit the latent cooperative structure among agents. Yet neither paradigm offers a unified perspective on how cooperation emerges or how it influences message flow and action consistency, resulting in two critical limitations.

First, many communication methods implicitly equate the set of agents that exchange messages with the set of agents that should collaborate (Liu et al. 2023). They ignore that cooperative relations are dynamic and heterogeneous, varying over time and across agent pairs, thus leading to misrouted messages and collaboration conflicts. Fig. 1a illustrates such a conflict: Zealot 1 communicates only with Zealot 3 and thus expects support, whereas Zealot 3 also messages Zealot 2 and Zealot 4 and actually cooperates with them, leaving Zealot 1 unsupported and causing its attack

*Corresponding author.

on Stalker 1 to fail. Second, role-based methods typically cluster agents by action-space similarity, implicitly assuming that similar observable behaviour implies similar function (Wang et al. 2020b). This assumption inverts causality: cooperation induces behavioural similarity, not the reverse (Sievers et al. 2024). Fig. 1b shows that Zealot 1 and Zealot 3 behave similarly yet do not collaborate; a similar mismatch occurs for Zealot 2 and Zealot 4, leading to incorrect role assignments and functional misalignment.

To address these limitations, we introduce GRDC (Graph-driven Role Discovery and Communication), a unified framework that dynamically constructs interaction graphs to approximate local cooperative dependencies, and subsequently uses these graphs to infer functional roles and regulate intra-role communication. GRDC comprises three components. First, each agent dynamically constructs a local interaction graph over observable teammates, capturing cooperative dependencies despite partial observability. Second, guided by the graph topology, each agent selects a role via prototype matching, which promotes consistent and discriminative role prototypes while avoiding misclassification due to behavioural similarity. Third, GRDC employs intra-role communication: message exchange is confined to intra-role agents, and attention-based aggregation ensures consistent communication while suppressing cross-role interference. Beyond role inference and communication, GRDC encourages diversity and compactness through role-entropy maximisation and prototype decorrelation, and further eliminates under-utilised roles via dynamic prototype pruning.

In contrast to prior methods that address coordination through either communication or role discovery in isolation, GRDC introduces a unified framework that grounds both paradigms structurally and integrates them functionally, yielding more expressive representations and more reliable communication pathways. Our main contributions are summarized as follows:

- We propose GRDC, a unified graph-driven framework that jointly infers agent roles and models inter-agent cooperative patterns using interaction graphs. This enables role assignment based on relational semantics rather than action similarity, reducing misclassification under partial observability.
- We design an intra-role communication mechanism in which information exchange is restricted to agents sharing the same role. Combined with attention-based message aggregation, this mechanism mitigates collaboration conflicts arising from misaligned communication.
- We evaluate GRDC on three representative partially observable benchmarks: Predator Prey, Cooperative Navigation, and SMACv2 (Ellis et al. 2023). Results show that GRDC consistently outperforms state-of-the-art communication- and role-based baselines, achieving higher coordination efficiency and training stability.

Problem Formulation

Fully cooperative tasks under partial observability are formally modelled as a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) (Bernstein et al. 2002), represented by the tuple $(\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{R}, \mathcal{P}, \gamma, \mu)$. In this definition, $\mathcal{N}$ is the set of agents; $\mathcal{S}$ the state space; $\mathcal{O} = \prod_{i \in \mathcal{N}} \mathcal{O}_i$ the joint observation space; $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$ the joint action space; $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the shared reward; $\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ the transition function; $\gamma \in [0, 1)$ the discount factor; and μ the initial state distribution. At each timestep t, each agent $i \in \mathcal{N}$ receives a private observation $o_i^t \in \mathcal{O}_i$ and selects an action $a_i^t \sim \pi_i(\cdot \mid o_i^t)$ under its stochastic policy π_i. The joint action $\boldsymbol{a^t} = (a_1^t, \ldots, a_{|\mathcal{N}|}^t)$ induces a transition to the next state $s^{t+1} \sim \mathcal{P}(s^t, \boldsymbol{a^t})$, and all agents obtain the shared reward $r^t = \mathcal{R}(s^t, \boldsymbol{a^t})$. The objective is to learn a joint policy $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{|\mathcal{N}|}\}$ that maximizes the expected discounted return $J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}, \mathcal{P}}[\sum_{t=0}^{T} \gamma^t r^t]$, where T denotes the finite time horizon.

Related Work

Existing efforts to enhance coordination under partial observability fall into two broad categories: communication-based methods and role-based methods.

Communication-based MARL. Inter-agent communication is critical for coordination under partial observability. Early studies such as CommNet (Sukhbaatar, Fergus et al. 2016), DIAL (Foerster et al. 2016), and IC3Net (Singh, Jain, and Sukhbaatar 2018) adopt centralized protocols where agents broadcast messages through fully connected topologies. Building on these foundations, FCMNet (Wang and Sartoretti 2022) and MASIA (Guan et al. 2022) incorporate recurrent encoders and self-supervised objectives to improve temporal reasoning and message aggregation. To mitigate bandwidth constraints, subsequent work proposes selective or structured communication strategies. Methods such as SchedNet (Kim et al. 2018), NeurComm (Chu, Chinchali, and Katti 2020), and ETCNet (Hu et al. 2021) learn scheduling mechanisms, whereas I2C (Ding, Huang, and Lu 2020), CMVC (Gao et al. 2025), VBC (Zhang, Zhang, and Lin 2019), and NDQ (Wang et al. 2019) explicitly minimise communication overhead by value or mutual-information criteria. More recently, attention-based and graph-structured communication has emerged, exemplified by ATOC (Jiang and Lu 2018), DGN (Jiang et al. 2019), G2ANet (Liu et al. 2020), T2MAC (Sun et al. 2024), SeqComm (Ding et al. 2024), and TGCNet (Zhang et al. 2025).

Role-based MARL. Role-based approaches aim to improve scalability and coordination by partitioning agents into roles, each associated with distinct sub-tasks or specialised policies. ROMA (Wang et al. 2020a) introduces a stochastic role embedding space in which agents learn identifiable and dynamic roles conditioned on local observations. RODE (Wang et al. 2020b) further advances role modeling by decomposing the action space based on learned effect-based action representations. SR-MARL (Zeng, Peng, and Li 2023) proposes structural information principles to stabilize and enhance role learning. LDSA (Yang et al. 2022) formulates role learning as a dynamic sub-task assignment problem based on structural latent variables, while Nguyen et al. (Nguyen et al. 2022) study the generalization and trans-

fer of role assignments across teams of varying sizes. GoMARL (Zang et al. 2023) clusters agents into functional groups to facilitate structured and efficient cooperation.

Method

We introduce Graph-driven Role Discovery and Communication (GRDC), a unified framework that combines graph-structured local-interaction modelling, prototype-based role discovery, and intra-role communication to enable coordination under partial observability in multi-agent environments. Fig. 2 illustrates the overall framework of GRDC and the corresponding pseudocode is shown in Algorithm 1.

Graph-based Interaction Modeling

GRDC builds on the insight that coordination in multi-agent systems is more likely to emerge among spatially adjacent agents (Ohtsuki et al. 2006). To approximate local inter-agent dependencies, each agent dynamically constructs a visibility-constrained interaction graph; two agents are neighbours if either can partially observe the other.

For agent i, let o_i^t be its current observation and $\tau_i^{t-1} = \{o_i^1, a_i^1, \ldots, o_i^{t-1}, a_i^{t-1}\}$ its trajectory history. A trajectory encoder f_{en} maps (o_i^t, τ_i^{t-1}) to an embedding $h_{\tau_i}^t = f_{\text{en}}(o_i^t, \tau_i^{t-1})$. For notational clarity, the superscript t is omitted in the remainder of this section. Let $\mathcal{G}_i$ be the visible neighbourhood graph of agent i. The interaction weights are computed by a soft-attention mechanism (Xu et al. 2015):

$$\omega_{ij}^g = \frac{\exp((W_Q^i h_{\tau_i})^\top W_K^i h_{\tau_j})}{\sum_{m \in \mathcal{G}_i} \exp((W_Q^i h_{\tau_i})^\top W_K^i h_{\tau_m})} \tag{1}$$

where W_Q^i, W_K^i, and W_V^i are learnable projection matrices. Self-connections are excluded by enforcing $\omega_{ii}^g = 0$. The neighbourhood feature is aggregated as $x_i = \sum_{j \in \mathcal{G}_i} \omega_{ij}^g W_V^i h_{\tau_j}$, and the domain representation is defined as the concatenation $z_i = [h_{\tau_i} \| x_i]$.

Role Discovery via Prototype Matching

Given the domain representation z_i, which combines individual trajectory features with local relational context, we formulate a structured role-inference mechanism based on prototype matching.

The aim is to capture functional heterogeneity by inferring agents' latent roles instead of clustering solely by observable similarity. Such functional diversity manifests as divergent policies or decision patterns. Directly inferring roles from observed actions can be unreliable under partial observability because limited context obscures the agents' true functions. Fig. 1b illustrates that this limitation yields functionally inconsistent role assignments.

To remedy this, we infer discrete role assignments from z_i, embedding both the agent's private trajectory and its local interaction context. This integration endows z_i with a locality-aware inductive bias that facilitates role discovery grounded in functional and relational semantics.

We define a shared set of K role prototypes, denoted as $\rho = \{\rho_1, \ldots, \rho_K\}$, where each prototype ρ_k is parameterized by a learnable vector $P_{\rho_k} \in \mathbb{R}^D$. Each agent i matches

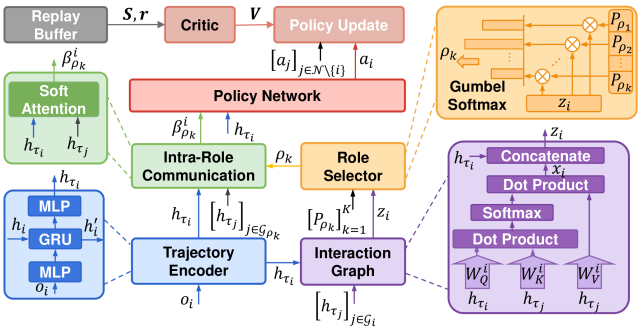


Figure 2: Framework of GRDC.

z_i to the role prototypes via a Gumbel-Softmax distribution (Jang, Gu, and Poole 2017):

$$f_{\text{rs}}(\rho_k \mid z_i) = \text{GumbelSoftmax}\left(\left[z_i^\top P_{\rho_1}, \ldots, z_i^\top P_{\rho_K}\right]\right) \tag{2}$$

This formulation yields discrete yet differentiable role assignments, enabling end-to-end optimisation.

Intra-role Communication

Building on the inferred roles, we introduce an intra-role communication mechanism that enables message exchange exclusively among agents sharing the same role, thereby enforcing structured coordination. The interaction graph captures spatial dependencies and prototype matching identifies functional groups, while intra-role communication reinforces behavioural consistency within each group.

Communication within a role group is beneficial because agents perform similar functions and therefore share relevant information. Let $\mathcal{G}_{\rho_k}$ denote the agents assigned to role ρ_k. For agent $i \in \mathcal{G}_{\rho_k}$, its role-level aggregated message is computed as:

$$\beta_{\rho_k}^i = \sum_{j \in \mathcal{G}_{\rho_k} \setminus \{i\}} \omega_{ij}^\rho h_{\tau_j} \tag{3}$$

where ω_{ij}^ρ is obtained with an attention mechanism analogous to Eq. (1). The policy of agent i is conditioned on h_{τ_i} and the aggregated intra-role message $\beta_{\rho_k}^i$, producing the action $a_i = \pi_i(h_{\tau_i}, \beta_{\rho_k}^i)$.

Structural Regularization on Role Representations

To preserve both diversity and utility of role prototypes, we employ three structural regularisation strategies: role entropy maximisation, prototype decorrelation, and dynamic prototype pruning.

Role Entropy Maximization. To prevent role collapse, we maximise the entropy of the empirical role distribution across agents. Let $q \in \mathbb{R}^{B \times N \times K}$ be the soft role-assignment tensor, where B is the batch size. The empirical marginal assignment probability for each role ρ_k is defined as $\bar{q}_k = \frac{1}{BN} \sum_{b=1}^{B} \sum_{n=1}^{N} q_{b,n,k}$. The entropy regularization loss is defined as:

$$\mathcal{L}_{\text{re}} = -\sum_{k=1}^{K} \bar{q}_k \cdot \log \bar{q}_k \tag{4}$$

Prototype Decorrelation. To reduce redundancy and promote diversity, we explicitly encourage the prototypes to be mutually dissimilar. Let $P_\rho \in \mathbb{R}^{K\times D}$ be the prototype matrix. Each prototype is normalised as $\tilde{P}_{\rho_k} = \frac{P_{\rho_k}}{\|P_{\rho_k}\|_2}$. The cosine similarity matrix among prototypes is then computed as $C = \tilde{P}_\rho \cdot \tilde{P}_\rho^\top \in \mathbb{R}^{K\times K}$, and the decorrelation loss (Zhu et al. 2022) is defined as:

$$\mathcal{L}_{\text{pd}} = \|C - I_K\|_F^2 \tag{5}$$

where I_K is the $K \times K$ identity matrix and $\|\cdot\|_F$ denotes the Frobenius norm. This loss penalizes off-diagonal similarity, thereby encouraging orthogonality among prototypes.

Dynamic Prototype Pruning. To promote compactness, under-utilised prototypes are pruned dynamically during training. The utilisation of prototype ρ_k is $u_k = \sum_{b=1}^{B}\sum_{n=1}^{N} q_{b,n,k}$. A threshold η_{min} is defined as $\eta_{\text{min}} = \kappa \cdot (B \cdot N)$, where $\kappa \in (0, 1)$ is a pruning sensitivity hyperparameter. The set of active prototypes is then given by $\mathcal{G}_{\rho_{\text{act}}} = \{k \in \{1, \dots, K\} \mid u_k > \eta_{\text{min}}\}$. The prototype matrix is updated by retaining only active prototypes:

$$P \leftarrow P[\mathcal{G}_{\rho_{\text{act}}}, :] \tag{6}$$

To ensure stable optimization, the optimizer state associated with the pruned parameters is reinitialized after each pruning operation.

Optimization with MAPPO

GRDC can be instantiated with either value-based or policy-gradient paradigms. In this study, we adopt the centralized training and decentralized execution (CTDE) paradigm (Lowe et al. 2017) and employ Multi-Agent Proximal Policy Optimization (MAPPO) (Yu et al. 2022) as the learning backbone.

Let θ_{ϕ_i} denote the parameters of agent i's policy, encompassing the trajectory encoder, interaction graph module, role assignment, intra-role communication, and the policy network itself. Let θ_{ψ_i} denote the parameters of its critic network.

The policy is updated by the PPO-clip objective:

$$\mathcal{L}_{\text{clip}} = \mathbb{E}_t\left[\min\left(\psi_i^t(\theta_{\phi_i}), \text{clip}(\psi_i^t(\theta_{\phi_i}), 1 \pm \epsilon)\right)\hat{A}^t\right] \tag{7}$$

where $\psi_i^t(\theta_{\phi_i}) = \frac{\pi_i(a_i^t|h_{\tau_i}^t, \beta_{\rho_k}^{i,t}; \theta_{\phi_i})}{\pi_i(a_i^t|h_{\tau_i}^t, \beta_{\rho_k}^{i,t}; \theta_{\phi_i}^{\text{old}})}$ denotes the probability ratio, and the generalized advantage estimate is computed as $\hat{A}^t = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^V$ with $\delta_{t+l}^V = r^{t+l} + \gamma V(s^{t+l+1}) - V(s^{t+l})$.

The actor loss additionally contains two regularisation terms that encourage diverse role usage and reduce prototype redundancy:

$$\mathcal{L}(\theta_{\phi_i}) = \mathcal{L}_{\text{clip}} + \lambda_{\text{re}}\mathcal{L}_{\text{re}} + \lambda_{\text{pd}}\mathcal{L}_{\text{pd}} \tag{8}$$

where λ_{re} and λ_{pd} are the coefficients for the role entropy and prototype decorrelation regularisations, respectively.

The critic is optimized by minimizing the temporal difference (TD) error:

$$\mathcal{L}(\theta_{\psi_i}) = \mathbb{E}_t\left[r^t + \gamma V'(s^{t+1}; \theta'_{\psi_i}) - V(s^t; \theta_{\psi_i})\right]^2 \tag{9}$$

where θ'_{ψ_i} denotes the parameters of the target critic network.

Algorithm 1: GRDC (Graph-driven Role Discovery and Communication)

Input: number of prototypes K, pruning sensitivity κ, $\lambda_{\text{re}}, \lambda_{\text{rd}}$, PPO clip ϵ, GAE λ, horizon T, episode budget E
Initialisation: prototype matrix $P_\rho \in \mathbb{R}^{K\times D}$, policy params $\{\theta_{\phi_i}\}_{i=1}^{|\mathcal{N}|}$, critic params $\{\theta_{\psi_i}\}$, replay buffer $\mathcal{B}$
Output: trained policies $\{\pi_i\}_{i=1}^{|\mathcal{N}|}$

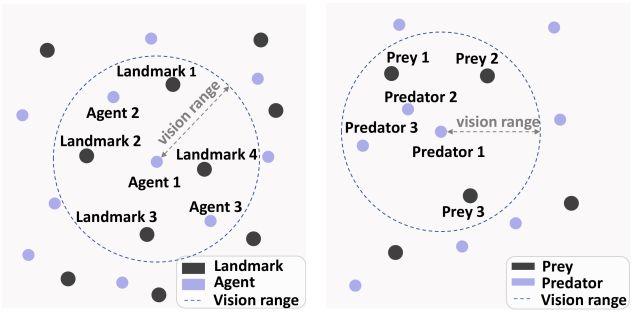
Initialise $P_\rho, \theta_\phi, \theta_\psi$, replay buffer $\mathcal{B}$
for episode $= 1$ to E **do**
  Reset env; get $\{o_i^0, s^0\}$; $t \leftarrow 0$
  while $t < T$ and not terminal **do**
    Encode trajectory: $h_{\tau_i} = f_{\text{en}}(o_i^t, \tau_i^{t-1})$
    Build visibility-constrained $\mathcal{G}_i$, compute $z_i = [h_{\tau_i}||x_i]$
    Infer role $\rho_k \sim \text{GumbelSoftmax}(z_i^\top P_\rho)$
    Aggregate intra-role message $\beta_{\rho_k}^i$ via attention
    Act: $a_i^t \sim \pi_i(h_{\tau_i}, \beta_{\rho_k}^i)$; step env; store transition in $\mathcal{B}$; $t \leftarrow t + 1$
  end while
  Sample minibatch $\mathcal{D}$; compute $\mathcal{L}_{\text{clip}}, \mathcal{L}_{\text{re}}, \mathcal{L}_{\text{pd}}$
  Update actor: $\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{clip}} + \lambda_{\text{re}}\mathcal{L}_{\text{re}} + \lambda_{\text{pd}}\mathcal{L}_{\text{pd}}$; update critic by TD error
  Prune prototypes with usage $u_k < \kappa|\mathcal{D}|$; reset optimiser states if pruned
end for
return $\{\pi_i\}$

Time Complexity Analysis

The overall time complexity of GRDC is given by $O(b_g n + Kn + b_\rho n)$, where n is the number of agents. Here, b_g denotes the maximum number of neighbours per agent, and b_ρ represents the maximum number of agents that share the same role. The term $O(b_g n)$ accounts for constructing local interaction graphs. Once the graphs are constructed, GRDC performs role assignment for each agent. Since each agent matches against K prototypes, this step results in a complexity of $O(Kn)$. Each agent then engages in intra-role communication with at most b_ρ peers, resulting in an additional cost of $O(b_\rho n)$. In our experimental settings, both b_g and b_ρ are constants significantly smaller than n, which ensures that the overall complexity scales linearly with the number of agents. In contrast, prior methods such as SR-MARL, MASIA, T2MAC, and TGCNet incur $O(n^2)$ time complexity due to fully connected or dense communication topologies among agents.

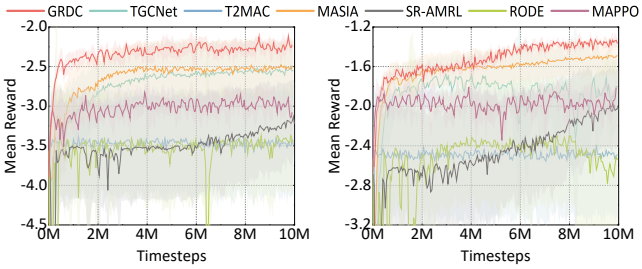
Experiments

Our experiments aim to answer the following questions: (1) Performance: Does GRDC outperform strong MARL baselines across diverse benchmarks? (2) Role quality: Do roles induced from local interaction graphs display clear functional differentiation and semantic coherence while reducing misassignment under partial observability? (3) Communication benefit: Does restricting messages to intra-role peers improve within-team coordination? (4) Structural regularisation: Do the role-entropy maximisation, prototype decor-



(a) Cooperative Navigation (b) Predator Prey

Figure 3: Illustration of modified Cooperative Navigation and Predator Prey.



(a) Cooperative Navigation (b) Predator Prey

Figure 4: Learning curves of GRDC and baselines on Cooperative Navigation and Predator Prey.

relation, and dynamic prototype pruning terms jointly yield a compact yet diverse role space?

Experimental Setting

We evaluate GRDC on three standard multi-agent benchmarks: Cooperative Navigation (CN), Predator Prey (PP), and SMACv2. To strengthen partial observability in CN and PP, each agent's field of view is limited to a fixed local region. As shown in Fig. 3, the vision radius is set to 0.3. Consequently, an agent perceives only the relative position and velocity of teammates or targets located within this radius. We compare GRDC with communication-based methods (TGCNet (Zhang et al. 2025), T2MAC (Sun et al. 2024), MASIA (Guan et al. 2022)), role-based methods (SR-MARL (Zeng, Peng, and Li 2023), RODE (Wang et al. 2020b)), and the standard MAPPO (Yu et al. 2022). All methods share the same hyperparameter settings to ensure a fair comparison, and complete configurations are reported in Appendix A. All metrics are averaged over four independent runs with distinct random seeds. The reported learning curves show the mean and standard deviation with 95% confidence intervals.

Overall Performance

Table 1, together with the learning curves in Fig. 4 and Fig. 5, demonstrates that GRDC achieves state-of-the-art performance across all three benchmark suites. The improvements are evident in both steady-state performance and

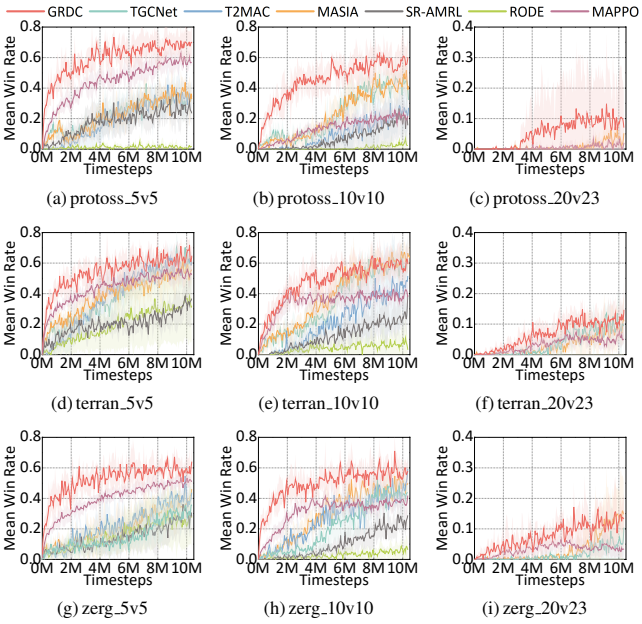


Figure 5: Learning curves of GRDC and baselines on nine SMACv2 maps.

sample efficiency.

On CN and PP, GRDC obtains the highest mean step rewards of -2.29 and -1.33, improving upon the next-best method MASIA by 8.4% and 10.1%, respectively, and outperforming the weakest baselines (T2MAC and RODE) by up to 47.6%. Furthermore, as shown in Fig. 4, GRDC converges with nearly half the sample budget required by MASIA on CN, while maintaining narrower confidence intervals that indicate more stable learning dynamics.

On the nine SMACv2 combat scenarios, GRDC achieves the highest win rate on eight of them, with the only exception being terran_10v10. On average, GRDC improves win rates by 5.95% over all baselines. On the Protoss maps (5v5, 10v10, 20v23), GRDC achieves win rates of 71.88%, 60.16%, and 22.81%, outperforming the strongest baselines by 23.2%, 24.2%, and 212.5%, respectively. For the Terran maps, GRDC achieves 61.72% in 5v5 and 14.70% in 20v23, outperforming the best baselines by 5.3% and 17.6%, respectively. On terran_10v10, the marginal gap between GRDC and MASIA falls within MASIA's standard deviation, suggesting no statistically significant difference. On the Zerg maps, GRDC achieves the highest win rates across all three tasks, surpassing the strongest baselines by 18.4%, 2.6%, and 5.3%, respectively. As shown in Fig. 5, GRDC consistently achieves higher win rates during training and converges faster than all baselines. Its performance curves maintain a consistent statistical advantage throughout most scenarios.

Across all eleven tasks, GRDC's confidence intervals do not overlap with those of the weakest baselines, even when mean performance is similar, indicating statistically significant improvements at the 95% confidence level. These results confirm that graph-driven role discovery combined

Method	Cooperative Navigation	Predator Prey	SMACv2 (%)								
			protoss_5v5	protoss_10v10	protoss_20v23	terran_5v5	terran_10v10	terran_20v23	zerg_5v5	zerg_10v10	zerg_20v23
GRDC	**-2.29(0.12)**	**-1.33(0.06)**	**71.88(10.82)**	**60.16(13.35)**	**22.81(18.13)**	**61.72(10.33)**	59.37(8.46)	**14.70(9.03)**	**57.81(7.44)**	**60.93(8.26)**	**15.62(5.10)**
TGCNet	-2.55(0.38)	-1.80(0.32)	41.67(4.78)	42.19(2.21)	0.23(0.39)	58.59(11.79)	61.72(4.68)	7.30(6.61)	30.47(8.97)	35.93(9.02)	5.21(4.77)
T2MAC	-3.49(0.65)	-2.53(0.64)	41.41(7.81)	27.08(10.97)	0.00(0.00)	60.15(8.61)	51.56(24.30)	0.00(0.00)	35.94(11.05)	59.37(4.41)	0.00(0.00)
MASIA	-2.50(0.34)	-1.48(0.18)	43.75(7.66)	48.43(7.86)	7.30(4.78)	52.34(5.33)	**63.28(14.06)**	12.50(8.07)	35.15(3.93)	49.21(11.23)	14.84(7.81)
SR-MARL	-3.22(0.51)	-2.03(0.39)	29.69(16.63)	24.22(16.21)	0.00(0.00)	36.71(13.59)	32.03(7.38)	0.00(0.00)	28.12(10.52)	28.12(6.25)	0.00(0.00)
RODE	-3.43(0.61)	-2.54(0.92)	1.56(2.21)	3.91(2.99)	0.00(0.00)	32.03(22.73)	6.25(7.21)	0.00(0.00)	24.22(15.59)	4.68(5.41)	0.00(0.00)
MAPPO	-2.94(0.23)	-1.96(0.12)	58.32(3.74)	21.37(5.27)	3.12(5.41)	55.03(2.29)	46.05(2.88)	5.71(1.93)	48.83(4.14)	35.18(2.28)	3.87(3.79)

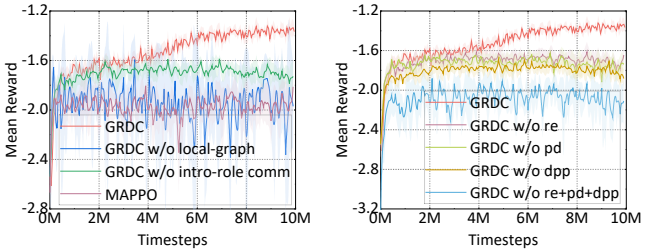
Table 1: Performance comparison across Cooperative Navigation (CN), Predator Prey (PP), and nine SMACv2 maps. CN and PP results are reported as the mean step reward, whereas SMACv2 results are reported as the mean win rate (%). Each algorithm is evaluated with four random seeds, each seed being run for 32 evaluation episodes. Bold numbers mark the best-performing algorithm in each environment.

with intra-role communication yields efficient coordination under partial observability.

Ablation Studies

Role Quality. A key contribution of GRDC is its use of local interaction graphs to structurally guide role discovery under partial observability. To assess the effectiveness of this graph-based inductive bias, we ablate the graph construction module from GRDC. In this variant, each agent infers its role solely from its own trajectory embedding. This variant is evaluated in PP, and the learning curves are shown in Fig. 6a. Without the graph structure, this variant underperforms MAPPO and exhibits significantly less stable learning dynamics. This performance degradation arises because trajectory embeddings encode only limited self-observation history, which hinders the identification of agents' functional characteristics under partial observability. This limitation mirrors the failure modes seen in methods such as RODE, where behaviour-based role discovery is unreliable due to the absence of local contextual information. In contrast, GRDC constructs visibility-constrained local graphs that allow each agent to access information from neighbouring agents. These structured interactions enable more accurate modelling of role-function dependencies, leading to enhanced performance and training stability.

To further examine the quality of the learned roles, we analyse the role assignments and prototype embeddings in CN and PP. We apply 3D Principal Component Analysis (PCA) to the learned role prototypes and visualise the resulting distributions in Fig.7. In CN, GRDC consistently assigns the same role to spatially adjacent agents that tend to occupy nearby landmarks, indicating semantic coherence and spatial coordination. The active roles are $\{0, 1, 3, 4, 9\}$, whose prototypes are well separated in the PCA space. In contrast, the pruned (inactive) prototypes are tightly clustered, reflecting their semantic redundancy. In PP, GRDC assigns the same role to predators operating within similar spatial regions, who often pursue the same prey. The active roles are $\{1, 3, 4, 5, 8, 9\}$, which also exhibit dispersed and discriminative prototype representations. However, when multiple predators perform highly similar functions, such as pursuing prey in overlapping areas, their role prototypes may become



(a) Ablation for local interaction graph and intro-role communication.

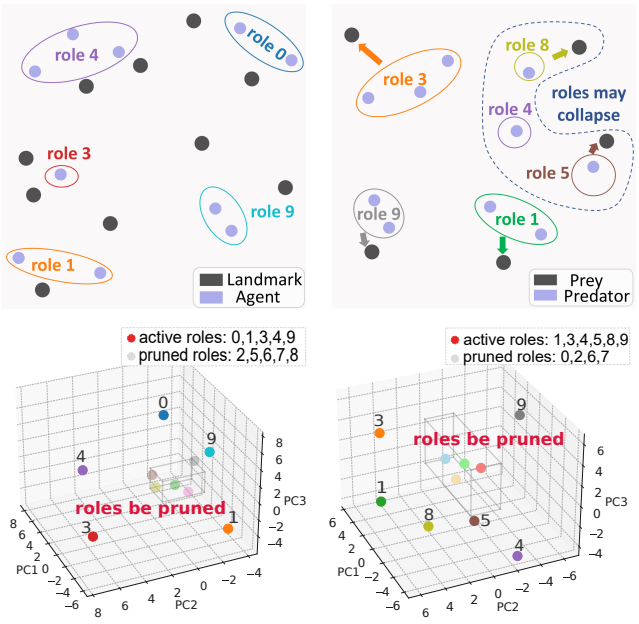
(b) Ablation for three structural regularisation methods.

Figure 6: Ablation studies: framework components (left) and regularisation terms (right).

closer in embedding space, as seen with prototypes $\{4, 5, 8\}$.

Overall, GRDC induces roles with clear functional differentiation and semantic coherence, validating the effectiveness of local interaction graphs under partial observability.

Communication Benefit. Unlike existing role-based methods such as RODE and SR-MARL, which coordinate agents through learned features of role-specific action spaces, GRDC introduces explicit communication among agents sharing the same role, enabling more direct and consistent intra-role coordination. To isolate the impact of intra-role communication, we remove this module from GRDC and let each agent i make decisions based solely on its trajectory encoding and the matched role prototype, i.e., $a_i = \pi_i(h_{\tau_i}, P_{\rho_k})$. As shown in Fig. 6a, although this variant outperforms MAPPO, it suffers a 29.3% performance drop compared to the full GRDC. This gap suggests that the learned role prototypes, obtained from trajectory and graph-based features $[h_{\tau_i} \| x_i]$, form semantic clusters of agents that share similar functional contexts. Although such role representations offer semantic interpretability, the lack of communication limits the model's capacity to dynamically align behaviours across agents within the same role. In contrast, GRDC's attention-based intra-role communication allows agents to aggregate task-relevant messages from same-role collaborators, thereby enhancing behavioural consistency and enabling context-aware adaptation. This



(a) Role dynamics and the corresponding 3D PCA projection of role prototypes on Cooperative Navigation

(b) Role dynamics and the corresponding 3D PCA projection of role prototypes on Predator Prey

Figure 7: Role dynamics and 3D Principal Component Analysis (PCA) projection analysis of role prototypes learned by GRDC on Cooperative Navigation and Predator Prey.

mechanism improves coordination precision and reduces policy variance within role groups.

These results affirm that intra-role communication is critical for aligning behaviours and enabling effective coordination under partial observability.

Structural Regularisation. To ensure that the learned role space is both compact and diverse, GRDC integrates three structural regularisation techniques: role entropy maximisation, prototype decorrelation, and dynamic prototype pruning. To evaluate their effectiveness, we construct four ablated variants by removing each regulariser individually, as well as all three simultaneously. Experimental results on PP are shown in Fig. 6b. All ablations result in notable performance drops compared to the full model: 30.1% (without entropy), 34.58% (without decorrelation), 38.35% (without pruning), and 68.42% (without any regularisation). Moreover, the model becomes unstable when no regularisers are applied. Among the three, dynamic prototype pruning has the most significant impact, while entropy maximisation and prototype decorrelation yield comparable gains. Each regularisation term plays a distinct role. Pruning suppresses redundancy by eliminating inactive prototypes, thereby refining the semantic space and improving compactness. Entropy maximisation prevents role collapse and encourages balanced usage across role types. Prototype decorrelation promotes semantic orthogonality and spatial separation in embedding space, enhancing role discriminability and structural diversity. Collectively, these regularisers guide the formation of well-structured role representations that are semantically meaningful, functionally diverse, and stable to train, effectively answering question about regarding the necessity of structured constraints in role discovery.

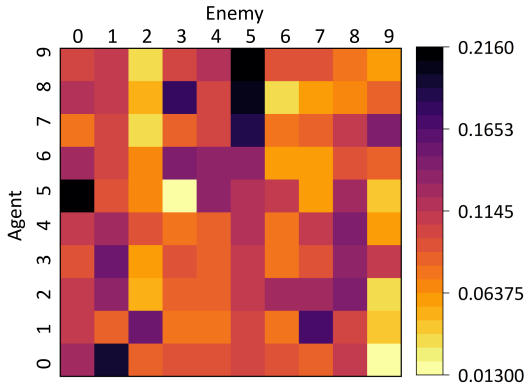


Figure 8: Episode-averaged action distribution of agents with different roles learned by GRDC in protoss_10v10.

Action Space Similarity

To further examine the causal relationship between cooperation and action similarity, we analyse whether similar action patterns emerge as a consequence of learned cooperation. We visualise the target selection behaviours of agents in the protoss_10v10 scenario, as shown in Fig. 8. Basic movement commands are excluded from this analysis, and only attack-related decisions towards enemy targets are considered. The set of active roles learned by GRDC in this task is $\{0, 2, 5, 8\}$. Agents assigned to different roles exhibit clearly distinct action distributions, while agents sharing the same role tend to display highly similar target selection patterns. For example, agents $\{7, 8, 9\}$ assigned to role 2 primarily focus on enemy targets $\{3, 4, 5\}$, whereas agents in role 0 concentrate on enemies $\{1, 8\}$. These observations support the view that functional coordination in GRDC drives the emergence of action similarity within roles, rather than clustering agents based on pre-existing behavioural similarity.

Conclusions

This paper presents GRDC as a unified framework for multi-agent coordination under partial observability. It integrates local interaction graph construction, discrete role assignment via prototype matching, and intra-role communication, all guided by structural regularisation to ensure a compact and semantically coherent role space. By explicitly modelling task-induced cooperative dependencies, GRDC enables scalable and robust cooperation. Experimental results across Cooperative Navigation, Predator Prey, and SMACv2 demonstrate consistent improvements over strong baselines, while ablation studies validate the interpretability of learned roles, the utility of intra-role communication, and the effectiveness of structural regularisation. GRDC provides a principled and effective solution for structured coordination in multi-agent systems.

References

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research*, 27(4): 819–840.

Chu, T.; Chinchali, S.; and Katti, S. 2020. Multi-agent Reinforcement Learning for Networked System Control. In *International Conference on Learning Representations*.

Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; and Pineau, J. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on machine learning*, 1538–1546. PMLR.

Ding, G.; Liu, Z.; Fang, Z.; Su, K.; Zhu, L.; and Lu, Z. 2024. Multi-agent coordination via multi-level communication. *Advances in Neural Information Processing Systems*, 37: 118513–118539.

Ding, Z.; Huang, T.; and Lu, Z. 2020. Learning individually inferred communication for multi-agent cooperation. *Advances in Neural Information Processing Systems*, 33: 22069–22079.

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

Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.

Gao, Z.; Qu, C.; Hao, Y.; and Ke, L. 2025. Communication in Multiagent Reinforcement Learning via Counterfactual Message Value. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(11): 8152–8165.

Guan, C.; Chen, F.; Yuan, L.; Wang, C.; Yin, H.; Zhang, Z.; and Yu, Y. 2022. Efficient multi-agent communication via self-supervised information aggregation. *Advances in Neural Information Processing Systems*, 35: 1020–1033.

Hu, G.; Zhu, Y.; Zhao, D.; Zhao, M.; and Hao, J. 2021. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 3966–3978.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Jiang, J.; Dun, C.; Huang, T.; and Lu, Z. 2019. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*.

Jiang, J.; and Lu, Z. 2018. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31.

Kim, D.; Moon, S.; Hostallero, D.; Kang, W. J.; Lee, T.; Son, K.; and Yi, Y. 2018. Learning to Schedule Communication in Multi-agent Reinforcement Learning. In *International Conference on Learning Representations*.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6): 4909–4926.

Liu, Y.; Wang, W.; Hu, Y.; Hao, J.; Chen, X.; and Gao, Y. 2020. Multi-agent game abstraction via graph attention neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7211–7218.

Liu, Z.; Wan, L.; Sui, X.; Chen, Z.; Sun, K.; and Lan, X. 2023. Deep Hierarchical Communication Graph in Multi-Agent Reinforcement Learning. In *IJCAI*, 208–216.

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

Nguyen, D.; Nguyen, P.; Venkatesh, S.; and Tran, T. 2022. Learning to Transfer Role Assignment Across Team Sizes. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 963–971.

Nguyen, T. T.; Nguyen, N. D.; and Nahavandi, S. 2020. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9): 3826–3839.

Ohtsuki, H.; Hauert, C.; Lieberman, E.; and Nowak, M. A. 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092): 502–505.

Sievers, B.; Welker, C.; Hasson, U.; Kleinbaum, A. M.; and Wheatley, T. 2024. Consensus-building conversation leads to neural alignment. *Nature communications*, 15(1): 3936.

Singh, A.; Jain, T.; and Sukhbaatar, S. 2018. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks. In *International Conference on Learning Representations*.

Sukhbaatar, S.; Fergus, R.; et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29.

Sun, C.; Zang, Z.; Li, J.; Li, J.; Xu, X.; Wang, R.; and Zheng, C. 2024. 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.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020a. ROMA: Multi-Agent Reinforcement Learning with Emergent Roles. In *International Conference on Machine Learning*, 9876–9886. PMLR.

Wang, T.; Gupta, T.; Mahajan, A.; Peng, B.; Whiteson, S.; and Zhang, C. 2020b. RODE: Learning Roles to Decompose Multi-Agent Tasks. In *International Conference on Learning Representations*.

Wang, T.; Wang, J.; Zheng, C.; and Zhang, C. 2019. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *International Conference on Learning Representations*.

Wang, X.; Ke, L.; Qiao, Z.; and Chai, X. 2021. Large-Scale Traffic Signal Control Using a Novel Multiagent Reinforcement Learning. *IEEE Transactions on Cybernetics*, 51(1): 174–187.

Wang, Y.; and Sartoretti, G. 2022. FCMNet: Full Communication Memory Net for Team-Level Cooperation in Multi-Agent Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1355–1363.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yang, M.; Zhao, J.; Hu, X.; Zhou, W.; Zhu, J.; and Li, H. 2022. Ldsa: Learning dynamic subtask assignment in cooperative multi-agent reinforcement learning. *Advances in neural information processing systems*, 35: 1698–1710.

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, Y.; He, J.; Li, K.; Fu, H.; Fu, Q.; Xing, J.; and Cheng, J. 2023. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in neural information processing systems*, 36: 46105–46121.

Zeng, X.; Peng, H.; and Li, A. 2023. Effective and stable role-based multi-agent collaboration by structural information principles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11772–11780.

Zhang, S. Q.; Zhang, Q.; and Lin, J. 2019. Efficient communication in multi-agent reinforcement learning via variance based control. *Advances in Neural Information Processing Systems*, 32.

Zhang, Z.; He, B.; Cheng, B.; and Li, G. 2025. Bridging training and execution via dynamic directed graph-based communication in cooperative multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23395–23403.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6908–6917.