RPMAT: Role-Guided Prioritized Multi-Agent Transformer for Dynamic Action Ordering

Zixin Liu[1, 2], Zhiming Zhou[1, *], Xiaohui Sun[3], Jiebo Chen[3] Wanjun Jing[1, 2], and Zhen Liu[1, 2]
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Beijing Institute of Electrical Engineering, Beijing 100854, China
Email: {liuzixin2024, zhiming.zhou}@ia.ac.cn, sxhbuaa@163.com, jiebo_chen@foxmail.com
{jingwanjun2024, liuzhen}@ia.ac.cn

***Abstract*—Sequential decision making paradigms have recently emerged in Multi-Agent Reinforcement Learning (MARL) to address the coordination limitations of simultaneous execution. However, existing approaches typically employ fixed or observation-driven heuristic orderings, overlooking the intrinsic dependency of decision priority on agent roles. In this paper, we propose the Role-Guided Prioritized Multi-Agent Transformer (RPMAT), a novel framework that optimizes action generation order by leveraging explicit role semantics. Specifically, we design a Role-Guided Cross-Attention (RGCA) mechanism that queries global observation features using learned latent role representations to dynamically quantify the decision urgency of each agent. This mechanism allows high-contribution roles to preemptively broadcast their intentions, thereby guiding the team's sequential execution. Theoretically, we reinterpret this order optimization as anterior structural credit assignment, where agents are ranked based on their predicted potential contribution to the global value function. Extensive experiments on SMAC, Google Research Football, and Multi-Agent MuJoCo benchmarks demonstrate that RPMAT achieves consistent superiority over state-of-the-art baselines in both heterogeneous and homogeneous scenarios, validating the effectiveness of role-driven coordination.**

***Index Terms*—Multi-Agent Reinforcement Learning, Decision Order Optimization, Role-Guided Cross Attention, Credit Assignment**

I. Introduction

Multi-Agent Reinforcement Learning (MARL) has achieved remarkable success in complex cooperative tasks [1]–[3]. However, effectively coordinating multiple agents to achieve a global optimal policy remains a core challenge. Traditional Centralized Training with Decentralized Execution (CTDE) [4] algorithms, such as QMIX [5], MADDPG [6] and MAPPO [7], typically adopt a simultaneous decision-making paradigm. As illustrated in Figure 1 (Top Left), this approach forces agents to act blindly regarding others' current intentions, overlooking complex action-level dependencies and often causing execution conflicts [8].

To address these limitations, recent works have introduced sequence modeling into MARL. The Multi-Agent Transformer (MAT) [9] recasts multi-agent cooperation as a sequence modeling problem, where agents generate actions sequentially. This paradigm allows subsequent agents to observe the intentions of their predecessors. However, MAT typically employs a fixed or random decision order. As depicted in Figure 1 (Bottom Left), such sequencing might lead to a priority mismatch: **a high-value agent may be forced to act last, rendering their potential guidance useless to the rest of the team**. While subsequent works like PMAT [10] attempt to dynamically optimize this order, they rely solely on local observations, neglecting the intrinsic influence of agent **Role** attributes on decision priority.

In human collaboration, decision making is an organic combination of role and timing (e.g., a Striker with a scoring opportunity takes precedence over a Midfielder). To explicitly leverage role semantics, we propose the **Role-Guided Prioritized Multi-Agent Transformer (RPMAT)**, which optimizes action generation order via a novel Role-Aware Ranking Mechanism. RPMAT utilizes explicit role semantics to determine decision priorities (see Figure 1, Right). Specifically, we design a **Role-Guided Cross-Attention (RGCA)** module that queries local observations using learned latent role representations to assess each agent's potential contribution. This generates a ranking score to guide a Plackett-Luce (P-L) sampling [11], [12], ensuring high-value agents are prioritized in the decision sequence to maximize information broadcasting.

Theoretically, this ranking functions as an *anterior structural credit assignment* [13], proactively granting early-decision rights to agents predicted to contribute most to the global advantage, contrasting with traditional posterior reward-based methods [13]–[15].

Our main contributions are summarized as follows:

- **RPMAT**: A novel framework that dynamically optimizes the action generation order via explicit role semantics.
- **Theoretical Insight**: Reinterpreting action order optimization as an anterior structured credit assignment.
- **Empirical Superiority**: RPMAT consistently outperforms state-of-the-art baselines on SMAC, GRF, and MA-MuJoCo benchmarks.

*Corresponding authors:Zhiming Zhou(email:zhiming.zhou@ia.ac.cn)
This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102402.

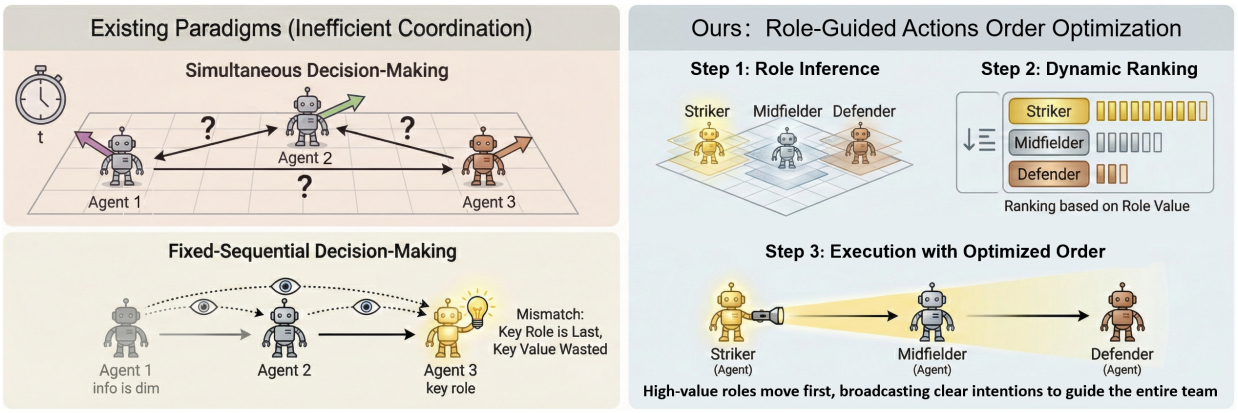

Fig. 1. Comparison of Decision-Making Paradigms. **Left:** Existing paradigms suffer from either blind execution conflicts due to isolation (Top) or inefficient information flow caused by fixed, mismatched decision priorities (Bottom). **Right:** Our proposed **RPMAT** leverages explicit role semantics to dynamically optimize the execution order. By prioritizing high-value roles, it ensures critical intentions are broadcasted first to effectively guide the team's coordination.

II. Preliminaries

A. Problem Formulation

We model cooperative MARL as a Dec-POMDP: $\mathcal{G} = \langle \mathcal{N}, \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{O}}, \boldsymbol{\mathcal{A}}, R, P, \gamma \rangle$ [16]. Here, $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents. At each timestep t, the system is in state $\boldsymbol{s}_t \in \boldsymbol{\mathcal{S}}$, each agent i observes $o_t^i \in \mathcal{O}^i$ and executes $a_t^i \in \mathcal{A}^i$ according to a dynamically determined order. The environment transitions via $P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a})$ and agents share a joint reward $R(\boldsymbol{s}_t, \boldsymbol{a})$ to maximize the discounted return $R^\gamma = \sum_{t=0}^{\infty} \gamma^t R_t$.

To handle agent heterogeneity, we categorize agents into distinct *Roles*: **Definition 1 (Role and Subtask).** A Role $\varphi_j \in \Phi$ is defined as a tuple $\langle Z_{\varphi_j}, \mathcal{G}_{\varphi_j} \rangle$, where Z_{φ_j} is the latent embedding of the role, and $\mathcal{G}_{\varphi_j}$ represents a subtask context for the subset of agents $\mathcal{N}_j \subseteq \mathcal{N}$ identified with this role.

B. Multi-Agent Advantage Decomposition

To justify the sequential decision-making paradigm, we leverage the Multi-Agent Advantage Decomposition Theorem [8]. Let $i_{1:n}$ be an ordered permutation of agents. The multi-agent observation-action value function $Q_{\boldsymbol{\pi}}^{i_{1:m}}$ is the expected return given actions $\boldsymbol{a}^{i_{1:m}}$. The local advantage of agent i_m, conditioned on predecessors $i_{1:m-1}$, is defined as:

$$A_{\boldsymbol{\pi}}^{i_m}(\boldsymbol{o}, \boldsymbol{a}^{i_{1:m-1}}, a^{i_m}) \triangleq Q_{\boldsymbol{\pi}}^{i_{1:m}}(\boldsymbol{o}, \boldsymbol{a}^{i_{1:m}}) - Q_{\boldsymbol{\pi}}^{i_{1:m-1}}(\boldsymbol{o}, \boldsymbol{a}^{i_{1:m-1}}). \tag{1}$$

Crucially, the global joint advantage decomposes into the sum of these local advantages: **Theorem 1 (Multi-Agent Advantage Decomposition [8]).**

$$A_{\boldsymbol{\pi}}(\boldsymbol{o}, \boldsymbol{a}) = \sum_{m=1}^{n} A_{\boldsymbol{\pi}}^{i_m}(\boldsymbol{o}, \boldsymbol{a}^{i_{1:m-1}}, a^{i_m}). \tag{2}$$

This theorem guarantees that maximizing each agent's local advantage sequentially leads to monotonic improvement in the joint policy, providing the theoretical foundation for our auto-regressive architecture.

C. Multi-Agent Transformer (MAT)

MAT [9] recasts MARL as a sequence modeling task using an **encoder-decoder** [17] architecture. The **Encoder** processes the joint observation sequence $(\mathbf{o}^{i_1}, \ldots, \mathbf{o}^{i_n})$ via unmasked self-attention to generate representations $(\hat{\mathbf{o}}^{i_1}, \ldots, \hat{\mathbf{o}}^{i_n})$. It functions as a centralized critic, approximating value functions by minimizing the empirical Bellman error:

$$\mathcal{L}_{\text{Enc}}(\phi) = \frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} \left[R_t + \gamma V_{\bar{\phi}}(\hat{\mathrm{o}}_{t+1}^{i_m}) - V_{\phi}(\hat{\mathrm{o}}_t^{i_m}) \right]^2, \tag{3}$$

where ϕ and $\bar{\phi}$ denote the online and target network parameters, respectively.

The **Decoder** generates actions auto-regressively. To maintain sequential dependency where agent i_m depends on $\boldsymbol{a}^{i_{1:m-1}}$, it employs masked self-attention with triangular masks. Parameterized by ρ, the decoder maximizes the expected return via a clipped PPO [18] objective:

$$\mathcal{L}_{\text{Dec}}(\rho) = -\frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} \min \left(r_t^{i_m} \hat{A}_t, \text{clip}(r_t^{i_m}, 1 \pm \epsilon) \hat{A}_t \right), \tag{4}$$

where $r_t^{i_m}(\rho) = \frac{\pi_\rho^{i_m}(\mathrm{a}_t^{i_m} | \hat{\mathbf{o}}_t, \hat{\mathbf{a}}_t^{i_{1:m-1}})}{\pi_{\rho_{\text{old}}}^{i_m}(\mathrm{a}_t^{i_m} | \hat{\mathbf{o}}_t, \hat{\mathbf{a}}_t^{i_{1:m-1}})}$ is the probability ratio, ϵ is the clipping range hyperparameter, and $\hat{A}_t$ is the GAE estimate derived from the encoder's value function.

III. Method

As illustrated in Figure 2, the proposed RPMAT framework is elaborated upon in this section. The overall pipeline begins with a standard Multi-Agent Transformer Encoder that extracts latent representations from local observations. Building upon this backbone, our core contributions consist of three specialized modules: (1) a **Role Encoder** that disentangles latent tactical intents from historical trajectories; (2) a **Role-Guided Cross-Attention (RGCA)** mechanism that assesses decision priority by querying global observations with role semantics; and (3) a **Sequential Action Generation** scheme via P-L sampling, which dynamically optimizes the execution

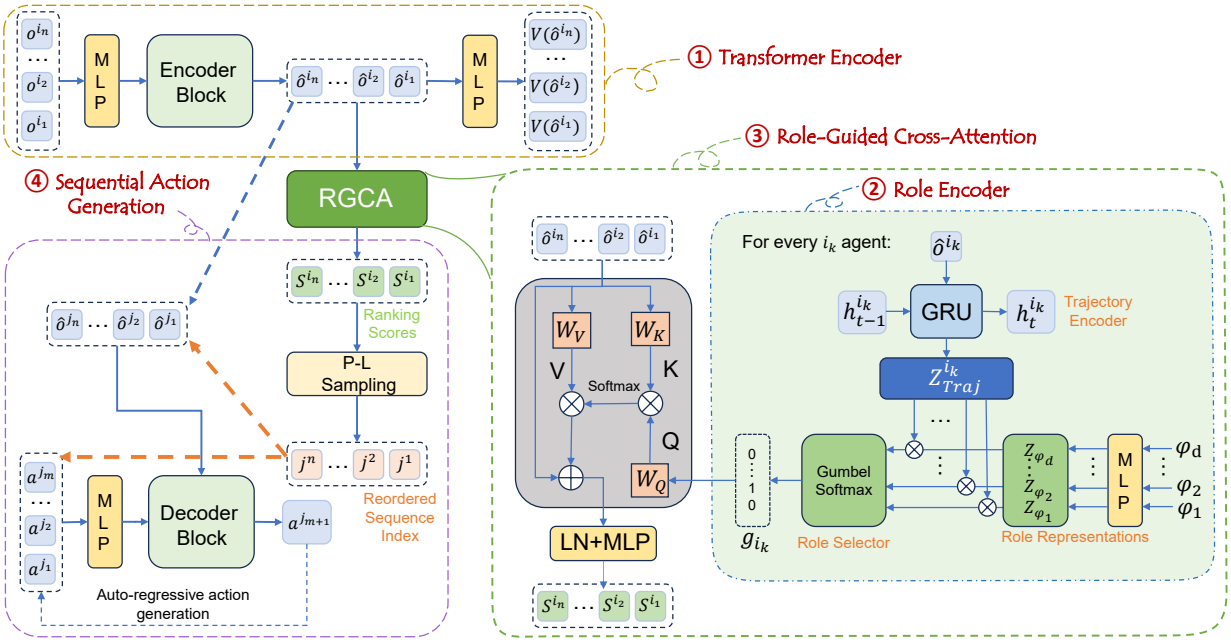

Fig. 2. The overall framework of *RPMAT*. The model optimizes action order via role semantics through: (1) **Transformer Encoder**: extracting latent representations from local observations; (2) **Role Encoder**: inferring latent role features from historical trajectories; (3) **Role-Guided Cross-Attention (RGCA)**: Priority Evaluation by calculating ranking scores; and (4) **Sequential Action Generation**: generating actions auto-regressively via the Transformer Decoder based on the sampled action order.

order for the auto-regressive decoder. Finally, we detail the joint training objective.

A. Role Encoder

In cooperative multi-agent tasks, agents often exhibit distinct behavioral patterns that correspond to specific tactical roles. We utilize the historical trajectories to capture dynamic intentions and infer its current roles, ensuring that the learned roles are both semantically distinct and temporally stable.

We employ a GRU-based trajectory encoder to map agent i_k's observation history into an embedding $Z^{i_k}_{Traj}$. Simultaneously, a set of learnable role prototypes $\mathcal{Z}_\Phi = \{Z_{\varphi_1}, \ldots, Z_{\varphi_d}\}$ is generated via an MLP. Role assignment is determined by the dot-product similarity $L_{i_k} = Z^{i_k}_{Traj} \cdot (\mathcal{Z}_\Phi)^T$. To enable differentiable discrete selection, we apply the Gumbel-Softmax [19] to sample a **role assignment vector**:

$$g_{i_k} = \text{Gumbel-Softmax}(L_{i_k}, \tau). \tag{5}$$

This vector g_{i_k} encapsulates dynamic and long-term behavioral patterns, serving as the contextual basis for subsequent priority evaluation.

To ensure the interpretability and robustness of learned roles, we introduce two auxiliary losses. First, to prevent role collapse and ensure that different roles capture distinct tactical semantics, we use the **Distinct Role Loss** that maximizes the separability of role prototypes by minimizing the negative pairwise squared Euclidean distance:

$$\mathcal{L}_{\text{distinct}} = -\frac{1}{d(d-1)} \sum_{j \neq l} \|Z_{\varphi_j} - Z_{\varphi_l}\|^2. \tag{6}$$

Second, the **Temporal Consistency Loss** enforces roles to maintain stability over short time horizons to prevent erratic behavior switching by minimizing the KL divergence between role distributions of consecutive timesteps ($P^{i_k}_t$ and $P^{i_k}_{t-1}$):

$$\mathcal{L}_{\text{consist}} = \frac{1}{N} \sum_{k=1}^{N} D_{\text{KL}}(P^{i_k}_{t-1} \| P^{i_k}_t). \tag{7}$$

This regularization term penalizes drastic changes in role selection between consecutive timesteps, promoting smooth role evolution throughout the episode.

B. Role-Guided Cross-Attention (RGCA)

Unlike conventional approaches that use roles merely for parameter sharing [20]–[22], RPMAT exploits explicit role semantics to guide the action generation sequence, emulating human team coordination where role attributes dictate decision timing.

Existing ranking methods typically rely solely on local observations, overlooking that identical observations can imply vastly different urgencies for different roles. To address this, we propose the **Role-Guided Cross-Attention (RGCA)** module, which evaluate the agent's potential contribution to the global objective conditioned on its role and real time observations.

Specifically, RGCA aims to leverage explicit role representations to retrieve critical information within observations across the entire team. To capture the interplay between roles and observations efficiently, we perform calculations using attention module.

We designate the role representations $\hat{\mathbf{G}}^t = \{g^t_{i_1}, \ldots, g^t_{i_n}\}$ extracted in the previous section as the Query, while the observation representations $\hat{\mathbf{O}}^t = \{\hat{o}^t_{i_1}, \ldots, \hat{o}^t_{i_n}\}$ at the current timestep serve as both the Key and Value. This design compels the network to query salient environmental features conditioned on the specific tactical roles. The calculation is formulated as :

$$\mathbf{Q} = \mathbf{G}^t \mathbf{W}_Q, \quad \mathbf{K} = \hat{\mathbf{O}}^t \mathbf{W}_K, \quad \mathbf{V} = \hat{\mathbf{O}}^t \mathbf{W}_V, \tag{8}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices. The attention output is then computed as:

$$\mathbf{Att} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{9}$$

where d_k is the dimension of the key vectors.

We apply a residual connection followed by a MLP, and normalized to a specific range to obtain decision priority scores $S^t = \{s^t_1, \ldots, s^t_n\}$:

$$\mathbf{S}^t = Sigmoid\left(\text{MLP}(\mathbf{Att} + \hat{\mathbf{O}}^t)\right). \tag{10}$$

Through this mechanism, the S^t output by the RGCA integrates the tactical priority associated with the agent's current role. By assigning higher scores to high-value roles, RGCA implements **Anterior Structural Credit Assignment**, ensuring critical agents act early to broadcast high-value intentions.

C. Sequential Action Generation Scheme via P-L Sampling

To transform the latent decision priority scores S^t into a concrete execution sequence σ, we employ Plackett-Luce (P-L) sampling. For an arbitrary permutation of agents $\sigma = (j_1, j_2, \ldots, j_n)$, the probability of sampling this specific sequence given the ranking scores S^t is:

$$P(\sigma|S^t) = \prod_{k=1}^{n} \frac{\exp(s^t_{j_k})}{\sum_{m=k}^{n} \exp(s^t_{j_m})}. \tag{11}$$

This formulation models the decision process as sampling without replacement. The P-L sampling offers two advantages. (1) **gradient smoothing**: The Softmax-based structure prevents gradient vanishing when parameter updates are insufficient to alter the hard relative ranking. (2) **exploration-exploitation trade-off**. It allows stochastic sampling during training for diverse coordination patterns and greedy execution during evaluation.

Upon determining the decision order σ, the MAT decoder rearranges agents according to σ for auto-regressive generation. Following the sequential decision-making paradigm, the policy for agent j_m is:

$$\pi_{\text{joint}}(\mathbf{a}|o) = \prod_{k=1}^{m} \pi(a_{j_k}|o, a_{j_1}, \ldots, a_{j_{k-1}}). \tag{12}$$

To ensure the RGCA module correctly identifies high-contribution agents, we formulate an optimization objective that align ranking scores with global contribution. We treat the generated order σ as a ranking policy $\pi_\theta(\sigma|S^t)$ and optimize it to maximize the expected Multi-Agent Advantage Function via PPO-clipped [18] objective:

$$\mathcal{L}_{\text{Rank}}(\theta) = -\frac{1}{T}\sum_{t=0}^{T-1} \min\left(r_t(\theta)\hat{A}_\sigma, \text{clip}(r_t(\theta), 1 \pm \epsilon)\hat{A}_\sigma\right), \tag{13}$$

where $r_t(\theta) = \frac{P_\theta(\sigma|S^t)}{P_{\theta_{\text{old}}}(\sigma|S^t)}$, θ denotes the parameters of RGCA, $\hat{A}_\sigma$ represents the estimated global advantage under the execution order σ, and ϵ is the clipping range hyperparameter.

Intuitively, this loss creates a feedback loop: if a sampled sequence σ yields a positive advantage ($A_\sigma > 0$), the gradient update increases the generation probability of this sequence, thereby elevating the scores of agents positioned earlier in the chain. Through this, RGCA learns to align ranking scores with the agents' marginal contributions to the global value, effectively achieving anterior structural credit assignment.

D. Overall Training Objective

The entire RPMAT framework is trained end-to-end. The objective function is a composite of the MAT policy optimization, the sequence ranking optimization, and the role representation regularization.

The MAT component is optimized using the standard PPO objective to maximize the expected return, denoted as $\mathcal{L}_{\text{MAT}}$ (comprising both $\mathcal{L}_{\text{Enc}}$ and $\mathcal{L}_{\text{Dec}}$). The Role Encoder is regularized by the weighted sum of diversity and consistency losses:

$$\mathcal{L}_{\text{Role}} = \lambda_{\text{dist}}\mathcal{L}_{\text{distinct}} + \lambda_{\text{consist}}\mathcal{L}_{\text{consist}}. \tag{14}$$

Combining these components, the total training objective is formulated as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MAT}} + \lambda_{\text{Rank}}\mathcal{L}_{\text{Rank}} + \mathcal{L}_{\text{Role}}, \tag{15}$$

where λ_{Rank} controls the weight of the order optimization. This joint optimization ensures that the learned roles, the decision order, and the execution policy are mutually enhanced to maximize the global cooperative performance.

IV. Experiments

We evaluate RPMAT against SOTA baselines (MAT [9], PMAT [10], and MAPPO [7]) across three MARL benchmarks: StarCraft II Multi-Agent Challenge (SMAC) [23], Google Research Football (GRF) [24], and Multi-Agent MuJoCo (MA MuJoCo) [25].

A. Experimental Environments

SMAC is a challenging cooperative benchmark environment for MARL based on the game StarCraft II. We conduct comparison experiments on two challenging maps, *3s5z vs 3s6z* and *MMM2*, both maps are Super Hard difficulty, heterogeneous and asymmetric.

GRF is a physics-based football simulation environment designed to evaluate the strategic learning capabilities of multi-agent systems in complex adversarial and cooperative scenarios. We evaluate the proposed method using the following three scenarios: *academy counterattack easy*, *academy pass and shoot with keeper*, and *academy 3 vs 1 with keeper*.

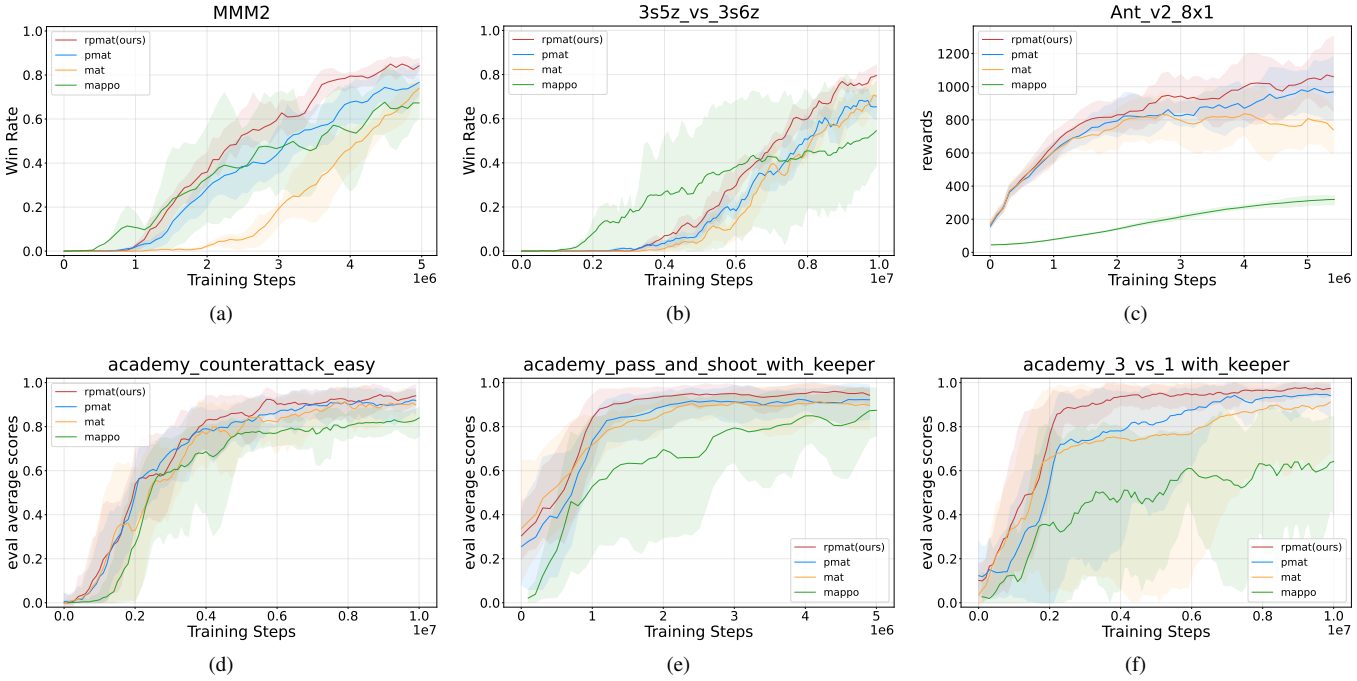

Fig. 3. Experimental results in StarCraft II, Google Research Football, and Multi-Agent MuJoCo.

MA MuJoCo is a robotic control benchmark platform built upon the MuJoCo [26]. By decomposing a single robot into multiple independently controlled joint segments, it establishes a suite of fine-grained cooperative control tasks. we utilize the *Ant-v2-8×1 agent* scenario for our algorithm evaluation.

B. Experimental Result

Performance on SMAC. The comparative results of different algorithms in the SMAC environment are illustrated in Figure 3a, Figure 3b and Table I. Performance is reported as the average win rate over 32 evaluation episodes across five different random seeds. To evaluate the final performance of each algorithm, we calculate the mean of the average win rates obtained from the last five evaluations. Specifically, in the *MMM2* scenario, RPMAT achieves a final win rate of 82.81% after 5×10^6 training steps, significantly outperforming the three baseline methods: MAT (71.71%), PMAT (73.75%), and MAPPO (67.29%). Similarly, in the *3s5z vs 3s6z* scenario, RPMAT reaches a win rate of 79.58%, surpassing MAT (69.93%), PMAT (68.81%), and MAPPO (51.56%). These results demonstrate that in these complex scenarios characterized by role heterogeneity, our method exhibits superior performance and faster convergence compared to the baselines.

Performance on GRF. The comparative results across all methods in the GRF benchmark are presented in Figure 3d, Figure 3e, Figure 3f and Table II. In the *academy pass and shoot with keeper* scenario, RPMAT achieves an average episode reward of 0.948, outperforming MAT (0.923), PMAT (0.940), and MAPPO (0.813). In the *academy counterattack easy* scenario, RPMAT attains an average episode reward of 0.946, consistently surpassing MAT (0.899), PMAT (0.912), and MAPPO (0.822). In the *academy 3 vs 1 with keeper* scenario, RPMAT achieves an average episode reward of 0.973, again outperforming MAT (0.912), PMAT (0.941), and MAPPO (0.642). The experimental results in the GRF domain indicate that RPMAT delivers superior performance in collaborative tasks exhibiting significant role heterogeneity.

TABLE I
PERFORMANCE ON SMAC (WIN RATE)

Scenario	Algorithms			
	RPMAT	MAT	PMAT	MAPPO
MMM2 (5M)	**0.828 ±0.063**	0.711 ±0.106	0.738 ±0.123	0.673 ±0.128
3s5z vs 3s6z (10M)	**0.796 ±0.073**	0.699 ±0.080	0.688 ±0.138	0.516 ±0.239

TABLE II
PERFORMANCE ON GRF (AVERAGE SCORES)

Scenario	Algorithms			
	RPMAT	MAT	PMAT	MAPPO
pass and shoot with keeper	**0.948 ±0.014**	0.923 ±0.051	0.940 ±0.084	0.813 ±0.097
counterattack easy	**0.946 ±0.031**	0.899 ±0.044	0.912 ±0.064	0.822 ±0.038
3 vs 1 with keeper	**0.973 ± 0.019**	0.912 ±0.107	0.941 ±0.053	0.642 ±0.182

Performance on MA MuJoCo. The comparison of all algorithms in the MA MuJoCo environment is shown in Figure 3c. Observing the training curves, it is evident that

RPMAT consistently outperforms baseline methods in the *Ant-v2-8×1 agent* scenario.

The aforementioned results reveal that sequential decision-making methods, such as MAT and PMAT, exhibit a significant performance advantage over simultaneous decision-making method like MAPPO. We attribute this phenomenon to the sequential decision paradigm, which provides critical information regarding the decisions of preceding agents. In most tasks, PMAT outperforms MAT, validating that an appropriate action sequence enables subsequent agents to better discern the intentions of preceding agents, thereby facilitating policy alignment. Furthermore, RPMAT, incorporating with the RGCA mechanism, achieves the best performance across all scenarios. This demonstrates the effectiveness of utilizing role information to optimize the sequence of agent action generation.

C. Structured Credit Assignment Study

In this section, we aim to empirically elucidate the intrinsic connection between our designed action order optimization and agent marginal contribution, thereby reinterpreting the optimization of action generation sequences through the lens of multi-agent credit assignment. In MARL, marginal contribution is a pivotal concept in credit assignment, typically defined as the variation in global team return should a specific agent deviate from its current action or be absent. This metric is widely used to quantify individual contributions to the total team reward and serves as the foundation for numerous algorithms [13], [14].

We evaluated two trained RPMAT models (differing only in the number of roles) within the *MMM2* scenario of the SMAC. At each timestep of an episode, we recorded the ranking scores output by the RGCA Scoring block. Simultaneously, we computed the marginal contribution for all agents. Specifically, we held the inputs of other agents constant while replacing the local observation of the target agent i with random noise. The resulting change in the value function served as a proxy for the agent's contribution to the team at that moment. We performed a correlation analysis on this data and visualized the results in Figure 4.

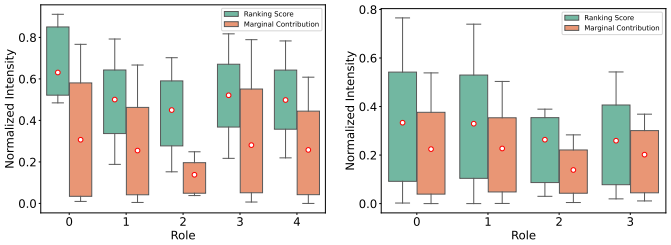

Fig. 4. Distribution of Scores & Marginal Contribution by Role, the x-axis represents different roles, and the y-axis denotes the normalized ranking scores and marginal contributions.

The statistical results reveal the following phenomena, supporting our hypotheses:

1) ***Effectiveness of Structural Credit Assignment*** *:* We observe a distinct positive correlation trend between the scores output by the RGCA network and the agents' marginal contributions. This phenomena indicates that, although we did not utilize marginal contributions as direct supervision signals during training, the RGCA module—guided by role attention—learned to assign decision priority to agents with greater marginal contributions to the team return. This demonstrates that our order optimization mechanism essentially functions as an effective form of structural credit assignment.

2) ***Role Heterogeneity in Decision Priority*** *:* The results show that roles with higher average marginal contributions consistently receive higher average ranking scores; conversely, roles with lower contributions receive lower scores. RPMAT avoids treating agents homogeneously; instead, it leverages the Role Encoder to capture agent heterogeneity. The scoring network demonstrates the adaptive capability to identify high-value roles and grant them higher decision priority, thereby preventing low-contribution agents from preempting critical decision channels.

Our algorithm aligns ranking scores with marginal contributions, adaptively prioritizing high-value roles to achieve structural credit assignment in action sequence generation.

D. Ablation Study

To verify the effectiveness of the RGCA mechanism, specifically its capability to dynamically optimize the decision order, we compared RPMAT against a variant of MAT that employs a random decision order at each timestep (referred to as MAT-Random). The results, as illustrated in Figure 5, demonstrate that a rational and optimized decision order is crucial for achieving superior performance.

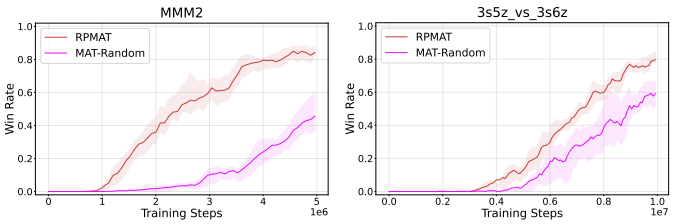

Fig. 5. Ablation studies investigating the effectiveness of the RGCA mechanism.

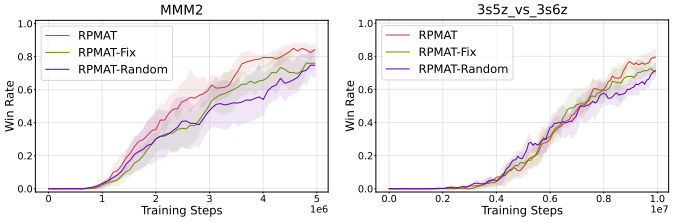

Fig. 6. Ablation studies investigating the effectiveness of the Role Encoder Module.

To validate the effectiveness of the proposed Role Information Representation Learning Module (Role Encoder), we

conducted two ablation studies. Both variants retain the cross-attention-based scoring network but exclude the Role Encoder. The first variant utilizes static, inherent agent IDs as the role representation g_i^k input to the RGCA (referred to as RPMAT-Fix). The second employs dynamic, randomly generated role representations (referred to as RPMAT-Random).

As shown in the Figure 6, RPMAT consistently outperforms both variants in terms of win rate and stability, validating the hypothesis that accurate role representation is essential for optimizing the sequential decision-making process.

V. Conclusion

In this paper, we introduced the **Role-Guided Prioritized Multi-Agent Transformer (RPMAT)**, a framework that rethinks the sequential decision-making paradigm in MARL by integrating explicit role semantics. To address the limitations of static orderings and purely observation-driven heuristics, RPMAT leverages a novel Role-Guided Cross-Attention (RGCA) mechanism to dynamically optimize the action generation sequence.

Experiments across diverse benchmarks—SMAC, GRF, and MA-MuJoCo—demonstrate that RPMAT consistently outperforms state-of-the-art baselines, exhibiting superior coordination efficiency and robustness, particularly in scenarios characterized by significant agent heterogeneity. Moreover, our analysis reveals that the optimized decision order is not merely a heuristic improvement but functions as an *anterior structural credit assignment*. By granting high-contribution roles the priority to broadcast their intentions, RPMAT effectively aligns individual action sequences with the global objective.

Our findings suggest that optimizing the *order* of decision is as critical as optimizing the *policy* itself in cooperative settings. Future work will explore extending this role-driven paradigm to large-scale, open-ended multi-agent systems and investigating the potential of such explicit role modeling in enhancing the interpretability of agent interactions.

References

[1] W. Jin, H. Du, B. Zhao, X. Tian, B. Shi, and G. Yang, "A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives," *arXiv preprint arXiv:2503.13415*, 2025.

[2] M. Hua, X. Qi, D. Chen, K. Jiang, Z. E. Liu, H. Sun, Q. Zhou, and H. Xu, "Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects," *IEEE Transactions on Automation Science and Engineering*, 2025.

[3] A. Krnjaic, R. D. Steleac, J. D. Thomas, G. Papoudakis, L. Schäfer, A. W. K. To, K.-H. Lao, M. Cubuktepe, M. Haley, P. Börsting *et al.*, "Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 677–684.

[4] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.

[5] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.

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

[7] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in neural information processing systems*, vol. 35, pp. 24 611–24 624, 2022.

[8] J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang, "Trust region policy optimisation in multi-agent reinforcement learning," *arXiv preprint arXiv:2109.11251*, 2021.

[9] M. Wen, J. Kuba, R. Lin, W. Zhang, Y. Wen, J. Wang, and Y. Yang, "Multi-agent reinforcement learning is a sequence modeling problem," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 509–16 521, 2022.

[10] K. Hu, M. Wen, X. Wang, S. Zhang, Y. Shi, M. Li, M. Li, and Y. Wen, "Pmat: Optimizing action generation order in multi-agent reinforcement learning," *arXiv preprint arXiv:2502.16496*, 2025.

[11] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 24, no. 2, pp. 193–202, 1975.

[12] J. I. Yellott Jr, "The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution," *Journal of Mathematical Psychology*, vol. 15, no. 2, pp. 109–144, 1977.

[13] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] J. Li, K. Kuang, B. Wang, F. Liu, L. Chen, F. Wu, and J. Xiao, "Shapley counterfactual credits for multi-agent reinforcement learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 934–942.

[15] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2085–2087.

[16] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[19] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[20] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "Rode: Learning roles to decompose multi-agent tasks," *arXiv preprint arXiv:2010.01523*, 2020.

[21] T. Wang, H. Dong, V. Lesser, and C. Zhang, "Roma: Multi-agent reinforcement learning with emergent roles," *arXiv preprint arXiv:2003.08039*, 2020.

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

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

[24] K. Kurach, A. Raichuk, P. Stańczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet *et al.*, "Google research football: A novel reinforcement learning environment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 4501–4510.

[25] C. S. de Witt, B. Peng, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson, "Deep multi-agent reinforcement learning for decentralized continuous cooperative control," *arXiv preprint arXiv:2003.06709*, vol. 19, 2020.

[26] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.