Role Discovery with Dynamic Role Cardinality for Zero-Shot Scalable Cooperation

Wanjun Jing[1, 2], Zhen Liu[*, 1, 2], Zhiming Zhou[1], Keyan Guo[3] Jiebo Chen[3], and Zixin 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:{jingwanjun2024, liuzhen, zhiming.zhou}@ia.ac.cn, 380593465@qq.com
jiebo_chen@foxmail.com, liuzixin2024@ia.ac.cn

***Abstract*—Inspired by the natural emergence of roles in biological systems for efficient coordination, role-based multi-agent reinforcement learning (MARL) methods have been proposed to enhance policy heterogeneity and cooperative capability in complex tasks. However, existing approaches typically rely on a fixed number of roles, whereas in practical multi-agent systems, the required role cardinality often varies across task stages and agent numbers, causing learned cooperative strategies to transfer poorly under varying agent numbers. In this paper, we propose SRD-DRC, a scalable role discovery framework with dynamic role cardinality. The core idea is to dynamically aggregate a global context from the local interaction histories of all agents via an attention mechanism, and to adaptively estimate the number of roles required at each cooperation stage, thereby eliminating reliance on a predefined role cardinality. Based on this estimation, SRD-DRC employs contrastive learning to acquire discriminative role representations that capture behavior patterns evolving across cooperation stages. Moreover, role representations are integrated at both the policy and value levels by parameterizing role-conditioned heterogeneous layers and guiding centralized value decomposition to enable structured credit assignment. Experimental results on the StarCraft Multi-Agent Challenge (SMAC) benchmark demonstrate that SRD-DRC achieves superior performance across diverse cooperative scenarios and exhibits effective zero-shot scalability.**

***Index Terms*—Multi-Agent Reinforcement Learning, Dynamic Role Cardinality, Zero-Shot Scalability, Centralized Training Decentralized Execution.**

I. INTRODUCTION

Multi-agent reinforcement learning (MARL) has been widely applied in complex decision-making domains, including board games [1], financial trading [2], autonomous driving [3], and robotic cooperation [4]. However, learning stable and efficient cooperative strategies remains challenging. Centralized training with decentralized execution (CTDE) is the dominant paradigm for cooperative MARL, leading to representative methods such as COMA [5], VDN [6], QMIX [7], and MAPPO [8], which typically rely on parameter sharing to enhance training efficiency and scalability.

Although parameter sharing reduces training complexity, it often leads to policy homogenization, which constrains exploration and limits performance in complex cooperative tasks. Several approaches have been proposed to address this issue. Methods that learn fully independent policies introduce heterogeneity but suffer from rapidly growing parameter complexity and lack effective mechanisms to train newly introduced agents [9], [10]. Role-based approaches learn emergent roles from historical observations [11]–[14], but typically rely on a predefined and fixed role cardinality, limiting scalability under varying task stages or agent numbers. Partial parameter sharing methods improve diversity by adjusting the degree of sharing [15]–[17], yet still depend on static schemes or incur substantial implementation complexity. Therefore, existing approaches often fail to generalize under varying agent numbers, resulting in limited zero-shot scalability.

In human team collaboration, efficient cooperation relies on a global understanding of the collaborative context, in which roles can be dynamically adjusted based on current states and historical evolution. Motivated by this observation, we propose a scalable role discovery framework with dynamic role cardinality (SRD-DRC), which constructs an informative global context and adaptively infers role cardinality to guide scalable role representation learning. Specifically, a Dynamic Role Cardinality Module first aggregates local observations from all agents via multi-head attention to build a global context and estimate the required role cardinality. Based on the estimated cardinality, a periodic clustering process determines the number of clusters, enabling contrastive learning to form positive and negative samples for role representation learning. The learned role representations are then used to parameterize heterogeneous layers in the decision network. Finally, SRD-DRC incorporates role representations into centralized value decomposition to improve structured credit assignment.

Our main contributions are summarized as follows:

- A novel scalable role discovery framework with dynamic role cardinality is proposed. By adaptively inferring the required role cardinality, it removes reliance on predefined roles and enables zero-shot scalable cooperation under varying agent numbers.
- The learned role representations are integrated into both policy learning and centralized value decomposition, promoting role-conditioned policy heterogeneity and effi-

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

cient coordination among agents.
- Experimental results on the SMAC benchmark using QMIX and MAPPO across both original and zero-shot scalability tasks demonstrate significant performance and scalability advantages.

II. PRELIMINARIES

A fully cooperative multi-agent task can be formulated as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [18], defined by $G = \langle I, S, A, P, R, \Omega, O, n, \gamma \rangle$. Here, $I = \{1, 2, \ldots, n\}$ denotes the set of agents, S represents the global state space, and $\gamma \in [0, 1]$ is the discount factor. Due to partial observability, each agent i receives a local observation $o_i^t \in O_i$ according to $\Omega(s^t, i) : S \times I \rightarrow O_i$, forming the joint observation space $O = O_1 \times O_2 \times \cdots \times O_n$. At each time step, agent i selects an action $a_i^t \in A_i$ from its discrete action space based on a stochastic policy $\pi_i : \tau_i^t \times A_i \rightarrow [0, 1]$, where $\tau_i^t \in (O_i \times A_i)^t$ denotes the action–observation history accessible to agent i. Given the joint action a^t, the system transitions to the next state s^{t+1} according to the transition dynamics $P(s^{t+1} \mid s^t, a^t)$, and a shared team reward $r^t = R(s^t, a^t)$ is received by all agents.

Building on this formulation, we adopt a role-based perspective to model heterogeneous cooperation under parameter sharing. Each agent i is associated with a time-varying latent role $m_i^t \in \mathcal{M}$, which is encoded into a continuous representation z_i^t. Existing methods typically assume a predefined role set with a fixed cardinality $|\mathcal{M}|$ [12]–[14], which simplifies learning but is often violated in practical cooperative scenarios. Both temporal evolution within a task and variations in agent numbers may lead to changes in the optimal role cardinality. In such cases, a fixed role cardinality limits the expressiveness of role representations and reduces policy adaptability to dynamic cooperation requirements. Therefore, enabling dynamic role cardinality is essential for scalable role representation learning in multi-agent systems.

III. METHOD

In this section, the proposed SRD-DRC framework is presented through a QMIX-based instantiation (Fig. 1), which can be readily extended to MAPPO. The framework consists of four core modules: (1) The **Dynamic Role Cardinality Module** aggregates global context from historical interactions among all agents and estimates the required role cardinality at each cooperation stage. (2) The **Contrastive Role Representation Module** learns discriminative role representations. Guided by the estimated role cardinality, this module periodically applies clustering to construct positive and negative sample pairs. (3) The **Heterogeneous Layer** is incorporated into the decision network, with layer parameters generated conditioned on the learned role representations. (4) The **Role Attention Fusion Module** integrates role representations into value decomposition via an attention mechanism to capture inter-role interactions.

A. Dynamic Role Cardinality Module

The proposed approach aims to adaptively infer role cardinality from real-time global context, providing a foundation for role representation learning. This module consists of two components: an Attention-Enhanced Dynamic State Estimator (AE-DSE) and a Dynamic Role Cardinality Estimator (DRCE), as illustrated in Fig. 1(b).

1) Attention-Enhanced Dynamic State Estimator (AE-DSE): Accurate role cardinality estimation relies on an informative global context, which is not directly accessible under decentralized execution. To address this limitation, AE-DSE aggregates local information from all agents into a compact yet information-rich global context g_t. Specifically, AE-DSE comprises a local trajectory encoder that summarizes individual interaction histories and an attention-enhanced global aggregator for information fusion.

Local Trajectory Encoder. A Gated Recurrent Unit (GRU) is employed to capture temporal dynamics in agent trajectories. At each time step t, the encoder takes the current observation o_i^t, previous action a_i^{t-1}, and prior embedding e_i^{t-1} as input, producing the agent embedding $e_i^t = f_\phi(o_i^t, a_i^{t-1}, e_i^{t-1})$, which serves as the basis for global state estimation.

Attention-Enhanced Global Aggregator. To generate a global context g_t from agent embeddings $E^t = [e_1^t, \ldots, e_n^t]$, we employ an attention-based aggregator. Here, E^t serves as the keys and values, while a learnable global query Q_{global} acts as the query.

For each attention head, the Q, K and V representations are obtained via linear projection:

$$Q = Q_{\text{global}} W_q, \quad K = E^t W_k, \quad V = E^t W_v \tag{1}$$

where W_q, W_k, and W_v are learnable projection matrices. The output of each attention head τ_{atten}^h is computed using scaled dot-product attention:

$$\tau_{\text{atten}}^h = \text{Softmax}\left(\frac{QK^\top}{\sqrt{l_k}}\right) V \tag{2}$$

where l_k denotes the key dimension.

Considering that multi-agent systems often involve diverse latent roles and interaction patterns, a single attention head may not capture complex cooperative structures adequately. To enhance representational capacity, we adopt a multi-head attention mechanism with H parallel heads. The final global context g_t is obtained by concatenating the outputs of all heads:

$$g_t = \text{Concat}\left(\tau_{\text{atten}}^1, \cdots, \tau_{\text{atten}}^H\right) W_O \tag{3}$$

where W_O is a learnable projection matrix.

2) Dynamic Role Cardinality Estimator (DRCE): The DRCE module models role cardinality as a function of the current system state, rather than relying on predefined hyperparameters. Specifically, it is implemented as a fully-connected MLP that takes the global context g_t as input and outputs a vector of dimension $k_{\max}$, where $k_{\min}$ and $k_{\max}$ denote the predefined minimum and maximum numbers of roles. A

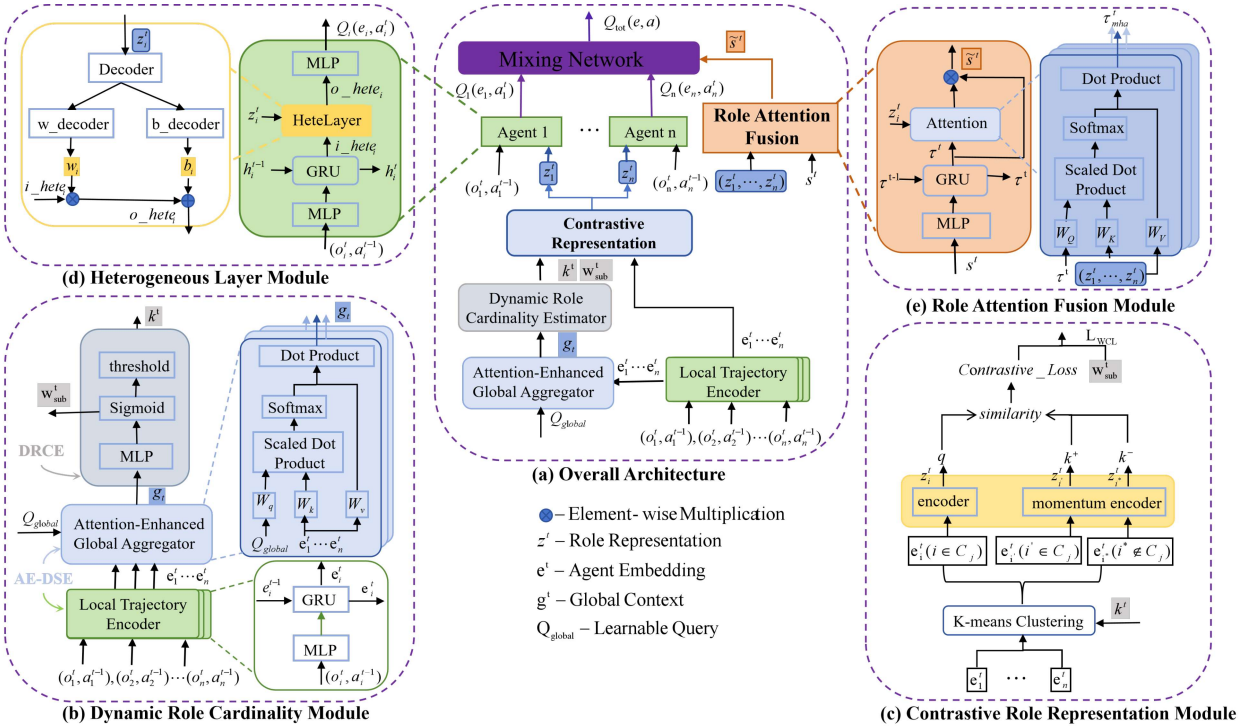

Fig. 1. The SRD-DRC framework based on QMIX. (a) The overall architecture. (b) **Dynamic Role Cardinality Module**: inferring the number of roles required for cooperation by aggregating global context. (c) **Contrastive Role Representation Module**: learning discriminative role representations via contrastive learning guided by the inferred role cardinality. (d) **Heterogeneous Layer Module**: generating role-conditioned heterogeneous parameters. (e) **Role Attention Fusion Module**: incorporating role representations into centralized value decomposition.

sigmoid activation is then applied to obtain the activation weight vector $\mathbf{w}_{sub}^t \in \mathbb{R}^{k_{\max}}$:

$$\mathbf{w}_{sub}^t = \sigma(\text{MLP}(g_t)) \tag{4}$$

Each element in $\mathbf{w}_{sub}^t$ represents the confidence associated with the corresponding role. The role cardinality k^t is determined by applying a predefined activation threshold τ to count the number of elements in $\mathbf{w}_{sub}^t$ that exceed this threshold. For stability, the resulting estimate is further clamped to a predefined valid range $[k_{\min}, k_{\max}]$:

$$k^t = \text{clamp}\left(\sum_{k=1}^{k_{\max}} \mathbb{I}(\mathbf{w}_{sub,k}^t > \tau), k_{\min}, k_{\max}\right) \tag{5}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

The predicted role cardinality k^t is subsequently used to parameterize K-means clustering for agent grouping, while $\mathbf{w}_{sub}^t$ is further used to weight the contrastive loss, effectively incorporating confidence information into role representation learning.

B. Contrastive Role Representation Module

In this work, a contrastive role representation learning approach is adopted to learn discriminative role representations from multi-agent interaction data in an unsupervised manner (Fig. 1(c)).

Role Encoder. The agent embeddings extracted by AE-DSE inherently capture behavioral differences, with similar agents exhibiting closer representations in the embedding space. Building upon these embeddings, a role encoder is introduced to generate role representations. Specifically, the encoder is formulated as a probabilistic mapping from agent embeddings to a role representation space:

$$z^t \sim f_\theta(z^t \mid e^t) \tag{6}$$

where e^t denotes the agent embedding at time step t, and z^t represents the corresponding role representation. This design enables role representations to capture long-term behavioral patterns at a higher semantic level.

Contrastive Learning. To learn discriminative role representations without ground-truth labels, we optimize an unsupervised contrastive objective. During training, the DRCE module predicts the role cardinality k^t, which is then used to periodically perform K-means clustering over agent embeddings, producing adaptive agent groups. At each time step t, the role representation z_i^t of agent i is treated as a query. Representations from the same cluster form the positive set K^+, while those from other clusters form the negative set K^-. This encourages intra-cluster compactness and inter-cluster separation in the representation space. The contrastive loss $\mathcal{L}$

is defined as

$$\mathcal{L} = -\log \frac{\sum_{k\in K^+} \exp((z_i^t)^\top W k)}{\sum_{k\in K^+} \exp((z_i^t)^\top W k) + \sum_{k\in K^-} \exp((z_i^t)^\top W k)} \tag{7}$$

where W is a learnable parameter matrix.

To improve training stability, we adopt the momentum update mechanism from MoCo [19] to update the parameters of the key encoder θ_k:

$$\theta_k = (1-\zeta)\theta_q + \zeta\theta_k \tag{8}$$

where $\zeta \in [0, 1)$ denotes the momentum coefficient. Only the query encoder θ_q is updated via backpropagation, while θ_k is updated using the momentum mechanism.

Dynamic Weighted Contrastive Learning. The quality of pseudo-labels generated by K-means depends on the accuracy of the predicted role cardinality k^t. In complex multi-agent systems, inaccurate estimates of k^t may lead to erroneous clustering and noisy gradient signals. To improve robustness, we introduce a dynamic weighted contrastive learning mechanism. The key idea is that the internal activation states leveraged by DRCE to predict k^t can be interpreted as a measure of confidence in the current estimation. Specifically, the mean activation value $\overline{\mathbf{w}}_{sub}^t$ is used to weight the contrastive loss at each time step, adaptively controlling the learning signal strength. The weighted contrastive loss $\mathcal{L}_{\text{WCL}}$ is defined as

$$\mathcal{L}_{\text{WCL}} = \overline{\mathbf{w}}_{sub}^t \cdot \mathcal{L} \tag{9}$$

When the role cardinality prediction is confident, the contrastive loss is assigned a higher weight; otherwise, the corresponding gradient contribution is suppressed. This mechanism enables contrastive learning to adapt to uncertainty in role cardinality estimation and mitigates the effects of noisy pseudo-labels.

C. Heterogeneous Layer Module

A heterogeneous layer is incorporated into the Q-network to inject role information into policy learning (Fig. 1(d)). Instead of using agent-specific network parameters, the heterogeneous layer maps role representations to layer-wise modulation parameters via role decoders. Specifically, given the role representation z_i^t and the heterogeneous layer input i_{hete}^t, the personalized output o_{hete}^t is obtained via a linear modulation:

$$o_{\text{hete}}^t = w_i^t \odot i_{\text{hete}}^t + b_i^t, \tag{10}$$

where $\odot$ denotes element-wise multiplication, and w_i^t and b_i^t are generated from z_i^t via linear decoders w_{decoder} and b_{decoder}.

This design enables role-conditioned modulation of shared network parameters, promoting policy heterogeneity while preserving parameter sharing.

D. Role Attention Fusion Module

To capture inter-role interactions, we introduce a role attention fusion module that integrates the learned role representations into the mixing network (Fig. 1(e)). The core idea is to adaptively weight different roles from a global perspective via an attention mechanism.

A GRU encodes the global state sequence $(s^0, s^1, \ldots, s^t)$ into a global embedding τ_t. This embedding serves as the query, while role representations $z = [z_1, \ldots, z_n]^\top$ act as keys and values. The attention weights are computed using scaled dot-product attention:

$$\alpha_i = \text{softmax}\left(\frac{(\tau_t W_Q)(z_i W_K)^\top}{\sqrt{d_k}}\right), \tag{11}$$

where W_Q and W_K are learnable projection matrices, and d_k is the key dimension. The resulting role-aware context τ_{atten} is given by

$$\tau_{\text{atten}} = \sum_{i=1}^{n} \alpha_i \, z_i W_V, \tag{12}$$

where W_V is the value projection matrix.

Furthermore, a multi-head attention is adopted to enhance the aggregation of role representations. Finally, the role-aware context is fused with the global state embedding and fed into the mixing network to guide the generation of its weights.

IV. Experiments

In this section, we design experiments to investigate the following research questions: (1) Evaluation of the learning performance of SRD-DRC compared with state-of-the-art baseline methods in complex multi-agent environments (Section IV-B). (2) Investigation of the contributions of key components, including DRCE, contrastive role representation learning, the heterogeneous layer, and the role attention fusion module, to training stability and cooperative performance (Section IV-C). (3) Analysis of the learned role representations in capturing agent behavior patterns and supporting zero-shot scalability under varying agent numbers (Section IV-D).

A. Experimental Setup

Environment. Experiments are conducted on the StarCraft II micromanagement (SMAC) [20] benchmark to evaluate the proposed method. Additional scalability tests are performed to assess generalization to unseen scenarios. In the original SMAC settings, each agent observes all allies and enemies within its field of view, which limits scalability under varying agent numbers. To enable zero-shot transfer across tasks with different agent numbers without additional observation mapping, the scalability tests (see Section IV-D) adopt a modified setting in which the number of observable allies and enemies is fixed for each agent. All other experiments follow the original SMAC configuration.

Evaluation Settings. All experiments are conducted with five random seeds. The mean test win rate is reported as the solid line, with the 95% bootstrap confidence interval shown as a shaded region. For scalability tests, trained models are directly evaluated on the corresponding scalability tasks under the modified environment. For each random seed, the trained model is evaluated over 40 episodes, and the mean and standard deviation of the win rate are reported to measure zero-shot transfer under varying agent numbers.

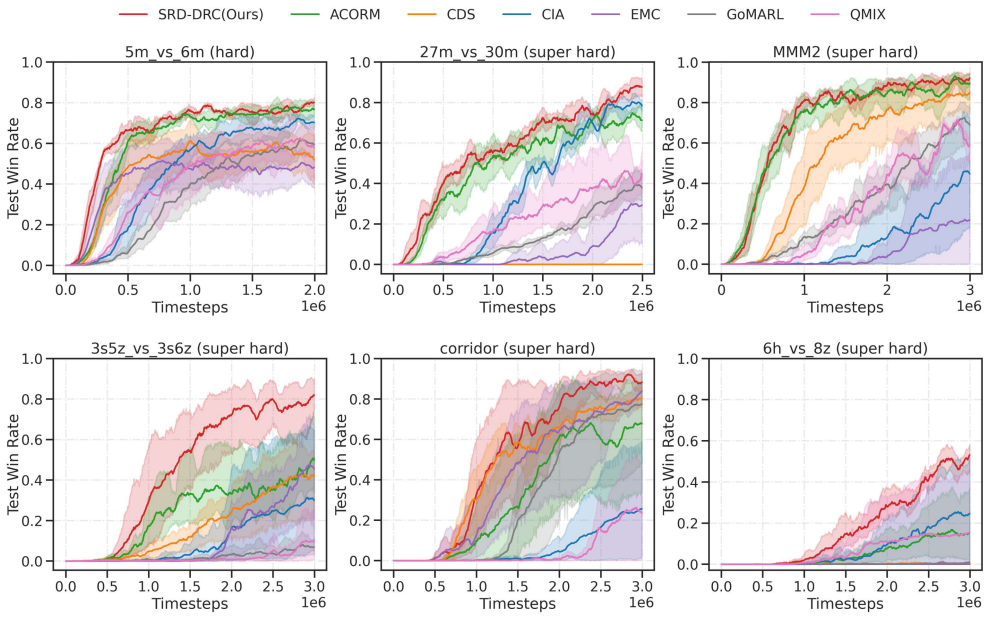

Fig. 2. Performance comparison of the QMIX-based SRD-DRC with baselines on six hard and super hard SMAC maps.

TABLE I
WIN RATE OF SCALABILITY TEST ON SMAC.

Task/Algorithms	SDR-DRC	SDR-DRC_w/o_hete	SHPPO	MAPPO	HAPPO	HAPPO (share)
MMM2(721_831)	**0.995±0.010**	0.990±0.012	0.712±0.065	0.627±0.101	0.763±0.051	0.312±0.204
621_631	**0.660±0.070**	0.370±0.099	0.625±0.025	0.625±0.159	0.351±0.224	0.264±0.141
621_821	**0.795±0.060**	0.730±0.062	0.775±0.078	0.752±0.125	0.568±0.112	0.251±0.125
711_731	**0.065±0.010**	0.050±0.010	0.056±0.025	0.051±0.025	0.025±0.018	0.000±0.000
722_831	**0.995±0.010**	0.985±0.020	0.963±0.015	0.961±0.038	0.875±0.101	0.956±0.025
731_831	**1.000±0.000**	1.000±0.000	0.956±0.045	0.975±0.025	0.318±0.302	0.964±0.035
821_831	0.990±0.010	**1.000±0.000**	0.825±0.075	0.955±0.045	0.125±0.049	0.700±0.192
8m_vs_9m	**0.900±0.085**	0.855±0.133	0.855±0.058	0.652±0.128	0.810±0.105	0.705±0.093
6m_vs_7m	**0.350±0.102**	0.220±0.109	0.175±0.104	0.157±0.119	0.025±0.012	0.127±0.098
7m_vs_8m	0.581±0.243	**0.595±0.237**	0.425±0.107	0.371±0.235	0.025±0.025	0.405±0.102
10m_vs_11m	**0.794±0.107**	0.760±0.208	0.702±0.201	0.425±0.183	0.000±0.000	0.513±0.238

Baselines. SRD-DRC instantiated with QMIX and MAPPO is evaluated against a diverse set of baselines. Under the QMIX-based setting, SRD-DRC is compared with QMIX [7], ACORM [13], CDS [10], CIA [21], EMC [22] and GoMARL [23]. Similarly, the MAPPO-based SRD-DRC is compared with MAPPO [8], HAPPO and HATRPO [9]. Furthermore, for scalability evaluations of the MAPPO-based SRD-DRC, additional baselines are included, such as SHPPO [24], MAPPO, HAPPO and a parameter-sharing variant of HAPPO.

B. Performance on SMAC

The SMAC benchmark consists of maps with varying difficulty levels (easy, hard, and super hard), where super hard tasks require more sophisticated coordination and provide a stringent test of SRD-DRC.

Fig. 2 reports the performance of the QMIX-based SRD-DRC on six hard and super hard maps. The results show that SRD-DRC converges faster than all baselines and achieves higher final win rates across all tasks. Moreover, SRD-DRC exhibits more stable training dynamics, as evidenced by consistently narrower confidence intervals across different runs, particularly on super hard maps. Fig. 3 illustrates the performance of MAPPO-based SRD-DRC on six representative tasks spanning different difficulty levels. For easy tasks, SRD-DRC achieves faster convergence with performance comparable to strong baselines. As task difficulty increases, it consistently attains higher win rates and maintains more stable training. Overall, these results demonstrate that SRD-DRC is particularly effective in scenarios requiring complex multi-agent coordination.

C. Ablation Studies

To better understand the contribution of each core component to the performance of SRD-DRC, ablation studies are conducted by comparing SRD-DRC with four ablated variants: (1) SRD-DRC_w/o_DRCE, where the Dynamic Role

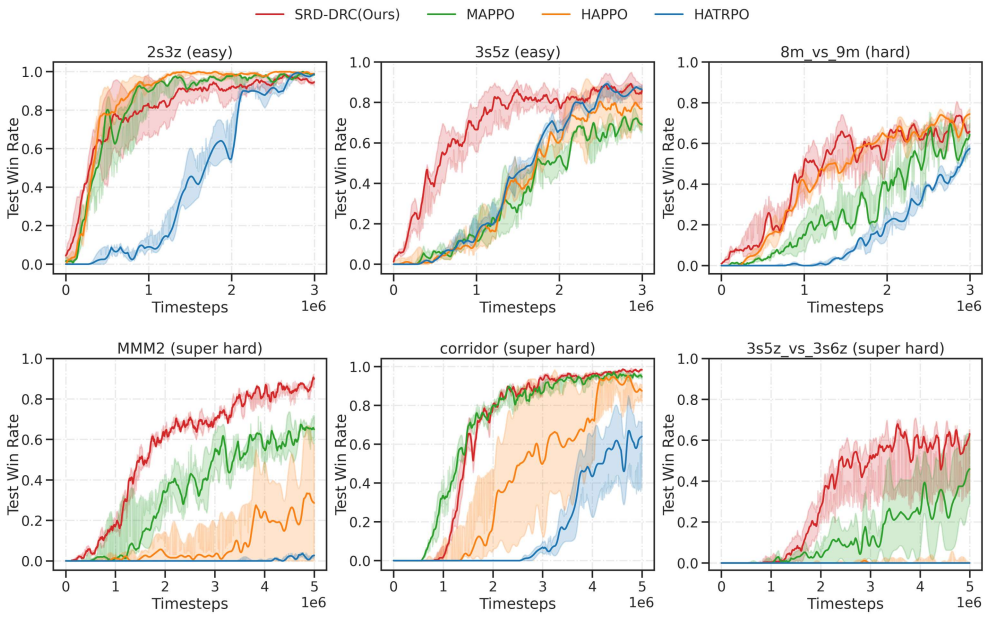

Fig. 3. Performance comparison between the MAPPO-based SRD-DRC and baselines on six representative maps.

Cardinality Estimator is disabled; (2) SRD-DRC_w/o_CL, where the contrastive role representation module is removed; (3) SRD-DRC_w/o_hete, where the heterogeneous layer is removed and role representations are directly concatenated with observations and actions to form $(o_i^t, a_i^{t-1}, z_i^t)$; (4) SRD-DRC_w/o_MHA, where the role attention fusion module is removed.

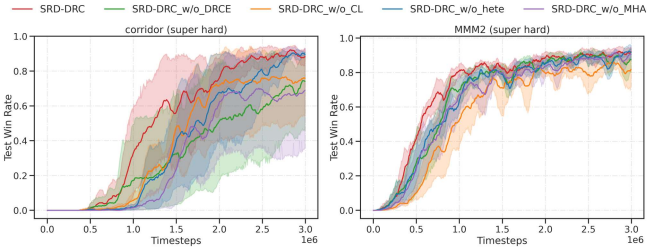

Fig. 4. Ablation results of the QMIX-based SRD-DRC on representative SMAC maps.

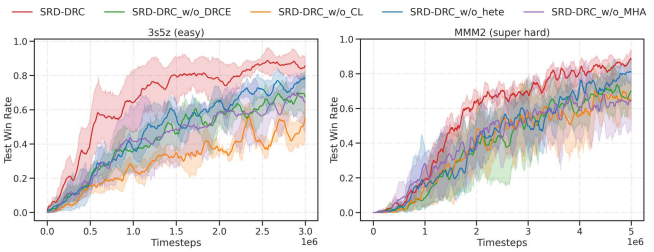

Fig. 5. Ablation results of the MAPPO-based SRD-DRC on representative SMAC maps.

As shown in Fig. 4, ablation studies are conducted on the QMIX-based SRD-DRC on the corridor and MMM2 maps. Similarly, Fig. 5 presents the ablation results of the MAPPO-based SRD-DRC. The results indicate that removing any individual component leads to consistent performance degradation, often accompanied by slower convergence and increased variance on hard and super hard maps, demonstrating that each module plays a necessary role in both performance and training stability. Moreover, the observed performance drops differ across ablated variants, suggesting that these components exhibit clear functional complementarity in supporting effective multi-agent coordination.

D. Scalability Test

To evaluate scalability under varying agent numbers, we select the homogeneous 8m_vs_9m map and the heterogeneous MMM2 map as original tasks. All models are trained for 20 million timesteps before zero-shot scalability evaluation. For MMM2 and 8m_vs_9m, each agent observes a fixed number of nearest enemies and allies (10/8 and 6/5, respectively) to ensure consistent observation space across different agent configurations. Based on this setting, we construct a set of extended tasks with varying agent numbers and enemy-ally configurations, and directly evaluate zero-shot transfer performance without fine-tuning. Specifically, six extended tasks are derived from MMM2 and three from 8m_vs_9m. We use the notation ABC_DEF to denote team compositions, where A, B, and C represent the numbers of Marines, Marauders, and Medivacs on the allied side, and D, E, and F denote those on the enemy side. For example, the original MMM2 task is denoted as 721_831.

For HAPPO, which follows an agent-independent modeling paradigm, changes in agent numbers are handled by reusing

or removing per-agent policies: policies of the same type are reused for newly introduced agents, while redundant policies are removed when the agent number decreases. As shown in Table I, SRD-DRC not only achieves strong performance on the original SMAC tasks, but also attains the highest win rates on almost all unseen zero-shot extended tasks. Notably, the performance advantages are particularly pronounced on heterogeneous MMM2 variants, where SRD-DRC remains highly robust under changes in both agent numbers and unit composition. Further comparison shows that SRD-DRC_w/o_hete, which removes the heterogeneous layer, performs consistently worse than the full model on most extended tasks, highlighting the critical role of the heterogeneous layer in supporting policy transfer under varying agent numbers.

V. Conclusion

In this paper, we propose a scalable role discovery framework with dynamic role cardinality (SRD-DRC), which aims to overcome the reliance of existing role-based methods on a fixed role cardinality and thereby support zero-shot scalable cooperation under varying agent numbers. The framework adaptively infers the number of roles required during cooperation, guides the generation of discriminative role representations, and leverages the learned role representations at both the policy and value-function levels to enable role-conditioned policy heterogeneity and structured credit assignment. Experimental results and ablation studies on the SMAC benchmark validate the effectiveness of the proposed method in terms of both performance and scalability. Future work will further explore incorporating historical exploratory trajectories into the role discovery process and combining them with agent-number-invariant modeling mechanisms to continuously improve generalization in large-scale and dynamic multi-agent systems.

References

[1] J. Perolat, B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony *et al.*, "Mastering the game of stratego with model-free multiagent reinforcement learning," *Science*, vol. 378, no. 6623, pp. 990–996, 2022.

[2] Y. Fang, Z. Tang, K. Ren, W. Liu, L. Zhao, J. Bian, D. Li, W. Zhang, Y. Yu, and T.-Y. Liu, "Learning multi-agent intention-aware communication for optimal multi-order execution in finance," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4003–4012.

[3] J. Dinneweth, A. Boubezoul, R. Mandiau, and S. Espié, "Multi-agent reinforcement learning for autonomous vehicles: A survey," *Autonomous Intelligent Systems*, vol. 2, no. 1, p. 27, 2022.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "Rode: Learning roles to decompose multi-agent tasks," in *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[12] 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.

[13] Z. Hu, Z. Zhang, H. Li, C. Chen, H. Ding, and Z. Wang, "Attention-guided contrastive role representations for multi-agent reinforcement learning," in *The Twelfth International Conference on Learning Representations*.

[14] H. Goel, M. Omama, B. Chalaki, V. Tadiparthi, E. M. Pari, and S. Chinchali, "R3dm: Enabling role discovery and diversity through dynamics models in multi-agent reinforcement learning," *arXiv preprint arXiv:2505.24265*, 2025.

[15] F. Christianos, G. Papoudakis, M. A. Rahman, and S. V. Albrecht, "Scaling multi-agent reinforcement learning with selective parameter sharing," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1989–1998.

[16] D. Li, N. Lou, B. Zhang, Z. Xu, and G. Fan, "Adaptive parameter sharing for multi-agent reinforcement learning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6035–6039.

[17] X. Li, L. Pan, and J. Zhang, "Kaleidoscope: Learnable masks for heterogeneous multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 22 081–22 106, 2024.

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

[19] M. Laskin, A. Srinivas, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *International conference on machine learning*. PMLR, 2020, pp. 5639–5650.

[20] S. Whiteson, M. Samvelyan, T. Rashid, C. De Witt, G. Farquhar, N. Nardelli, T. Rudner, C. Hung, P. Torr, and J. Foerster, "The starcraft multi-agent challenge," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2019, pp. 2186–2188.

[21] S. Liu, Y. Zhou, J. Song, T. Zheng, K. Chen, T. Zhu, Z. Feng, and M. Song, "Contrastive identity-aware learning for multi-agent value decomposition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, 2023, pp. 11 595–11 603.

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

[23] Y. Zang, J. He, K. Li, H. Fu, Q. Fu, J. Xing, and J. Cheng, "Automatic grouping for efficient cooperative multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 36, pp. 46 105–46 121, 2023.

[24] X. Guo, D. Shi, J. Yu, and W. Fan, "Heterogeneous multi-agent reinforcement learning for zero-shot scalable collaboration," *Neurocomputing*, p. 130716, 2025.