Heterogeneity in Multi-Agent Reinforcement Learning

Tianyi Hu
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
hutianyi2021@ia.ac.cn

Zhiqiang Pu
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
zhiqiang.pu@ia.ac.cn

Yuan Wang
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
wangyuan2025@ia.ac.cn

Tenghai Qiu
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
tenghai.qiu@ia.ac.cn

Min Chen
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
min.chen@ia.ac.cn

Xin Yu
National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, CASIA
Beijing, China
xin.yu@ia.ac.cn

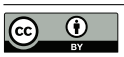
ABSTRACT

Heterogeneity is a fundamental property in multi-agent reinforcement learning (MARL), which is closely related not only to the functional differences of agents, but also to policy diversity and environmental interactions. However, the MARL field currently lacks a rigorous definition and deeper understanding of heterogeneity. This paper systematically discusses heterogeneity in MARL from the perspectives of *definition*, *quantification*, and *utilization*. First, based on an agent-level modeling of MARL, we categorize heterogeneity into five types and provide mathematical definitions. Second, we define the concept of heterogeneity distance and propose a practical quantification method. Third, we design a heterogeneity-based multi-agent dynamic parameter sharing algorithm as an example of the application of our methodology. Case studies demonstrate that our method can effectively identify and quantify various types of agent heterogeneity. Experimental results show that the proposed algorithm, compared to other parameter sharing baselines, has better interpretability and stronger adaptability. The proposed methodology will help the MARL community gain a more comprehensive and profound understanding of heterogeneity, and further promote the development of practical algorithms.[1]

KEYWORDS

Multi-Agent Reinforcement Learning; Heterogeneity

ACM Reference Format:
Tianyi Hu, Zhiqiang Pu, Yuan Wang, Tenghai Qiu, Min Chen, and Xin Yu. 2026. Heterogeneity in Multi-Agent Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/10.65109/HFKR5027



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/10.65109/HFKR5027

1 INTRODUCTION

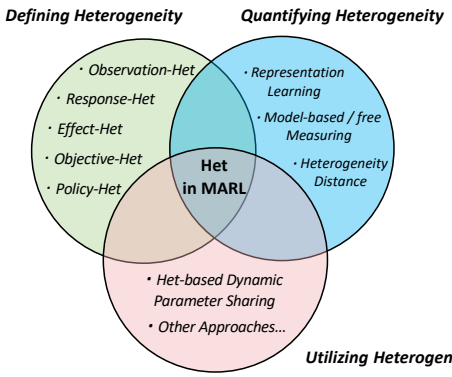


Figure 1: Our Philosophy. We aim to systematically discuss heterogeneity in MARL, establishing methodologies for defining, quantifying and utilizing heterogeneity.

Multi-agent reinforcement learning (MARL) has achieved success in various real-world applications, such as swarm robotic control [13], autonomous driving [37], and large language model fine-tuning [23]. However, most MARL studies focus on policy learning for homogeneous multi-agent systems (MAS), overlooking in-depth discussions of heterogeneous multi-agent scenarios [24]. *Heterogeneity* is a common phenomenon in multi-agent systems. For example, in nature, different species of fish collaborate to find food [5]; in human society, diverse teams demonstrate higher intelligence and resilience [7, 34]; and in artificial systems, aerial drones and ground vehicles cooperate to monitor forest fires [21]. Heterogeneity can enhance system functionality, reduce costs, and improve robustness, but effectively leveraging heterogeneity remains a key challenge in multi-agent system [1]. As an approach of learning through environmental interactions, MARL can effectively enable multi-agent systems to learn collaborative policies. Hence,

[1]Our code is available at https://github.com/Harry67Hu/HetDPS. Supplementary material (Appendix) is available at https://arxiv.org/pdf/2512.22941.

exploring heterogeneity from a reinforcement learning perspective would significantly broaden the applicability of MARL.

In the current MARL field, although some works explicitly or implicitly mention agent heterogeneity, only a few focus on its definition and identification. Regarding explicit discussion of heterogeneity, studies have explored communication issues [29], credit assignment [35], and zero-shot generalization [10] in heterogeneous MARL. However, these works limit their focus to agents with clear functional differences and lack definitions of agent heterogeneity. On the other hand, many studies explore policy diversity in MARL. Some encourage agents to learn distinguishable behaviors based on identity or trajectory information [12, 16], some works group agents using specific metrics [6, 33], and some quantify policy differences [4, 11] and design algorithms to control policy diversity [2].

However, these works do not adequately address where policy diversity originates or how it fundamentally relates to agent differences. In terms of defining and classifying heterogeneity in MARL, [3] divides heterogeneity into physical and behavioral types but lacks a mathematical definition. [29] provides extended POMDP for heterogeneous MARL settings, but do not classify or define heterogeneity. Others introduce the concept of local transition heterogeneity [35], but does not cover all elements of MARL. Currently, there is still a lack of *systematic analysis of agent heterogeneity from the MARL perspective*. To fill the aforementioned gaps, we conduct a series of studies on defining, quantifying, and utilizing heterogeneity in the MARL domain, the philosophy of our study can be found in Figure 1. And more details of related work can be found in Appendix A. Our contributions are summarized as follows:

- **Defining Heterogeneity:** Based on an agent-level model of MARL, we categorize heterogeneity into observation heterogeneity, response transition heterogeneity, effect transition heterogeneity, objective heterogeneity, and policy heterogeneity, and provide corresponding definitions.
- **Quantifying Heterogeneity:** We define the heterogeneity distance, and propose a quantification method based on representation learning, applicable to both model-free and model-based settings. Additionally, we give the concept of meta-transition heterogeneity to quantify agents' comprehensive heterogeneity.
- **Utilizing Heterogeneity:** We develop a multi-agent dynamic parameter-sharing algorithm based on heterogeneity quantification, which offers better interpretability and fewer task-specific hyperparameters compared to other related parameter-sharing methods.

2 PRELIMINARIES

Primal Problem of MARL. In this paper, we use Partially Observable Markov Game (POMG) [15, 19] as the general model for the primal problem of MARL.[1] To better study agent heterogeneity, we adopt an agent-level modeling approach similar to that in [9, 29]. A POMG is defined as an 8-tuple, represented as follows:

$$\langle N, \{S^i\}_{i\in N}, \{O^i\}_{i\in N}, \{A^i\}_{i\in N}, \{\Omega^i\}_{i\in N}, \{\mathcal{T}^i\}_{i\in N}, \{r_i\}_{i\in N}, \gamma\rangle, \tag{1}$$

Among all elements in Expression 1, N is the set of all agents, $\{S^i\}_{i\in N}$ is the global state space which can be factored as $\{S^i\}_{i\in N} = \times_{i\in N} S^i \times S^E$, where S^i is the state space of an agent i, and S^E is the environmental state space, corresponding to all the non-agent components. $\{O^i\}_{i\in N} = \times_{i\in N} O^i$ is the joint observation space and $\{A^i\}_{i\in N} = \times_{i\in N} A^i$ is the joint action space of all agents. $\{\Omega^i\}_{i\in N}$ is the set of observation functions. $\{\mathcal{T}^i\}_{i\in N} = (\mathcal{T}^1, \cdots, \mathcal{T}^{|N|}, \mathcal{T}^E)$ is the collection of all agents' transitions and the environmental transition. Finally, $\{r_i\}_{i\in N}$ is the set of reward functions of all agents and γ is the discount factor.

Here, we give the independent and dependent variables for each function and their notation. At each time step t, an agent i receives an observation $o_t^i \sim \Omega^i(\cdot|\hat{s}_t)$, where $\hat{s}_t \in \{S^i\}_{i\in N}$ is the global state at time t. Then, agent i makes a decision based on its observation, resulting in an action $a_t^i \sim \pi_i(\cdot|o_t^i)$. The environment then collects actions from all agents to form the global action $\hat{a}_t = (a_t^1, \ldots, a_t^{|N|})$. We assume that the local state transition of agent i is influenced by the global state and global action, so its local state transitions to a new state $s_{t+1}^i \sim \mathcal{T}^i(\cdot|\hat{s}_t, \hat{a}_t)$. Similarly, the states of other agents and the environment also transition, yielding the next global state $\hat{s}_{t+1} = (s_{t+1}^1, \ldots, s_{t+1}^{|N|}, s_{t+1}^E) \sim (\mathcal{T}^1(\cdot|\hat{s}_t, \hat{a}_t), \ldots, \mathcal{T}^{|N|}(\cdot|\hat{s}_t, \hat{a}_t), \mathcal{T}^E(\cdot|\hat{s}_t, \hat{a}_t)) = \{\mathcal{T}^i\}_{i\in N}(\cdot|\hat{s}_t, \hat{a}_t)$. At the same time, all agents receive rewards, with the reward for a specific agent i given by $r_t^i \sim r^i(\cdot|\hat{s}_t, \hat{a}_t)$.

The objective of MARL is to solve POMG by finding an optimal joint policy that maximizes the cumulative reward for all agents. We denote the individual optimal policy for agent i as π_i^* and the optimal joint policy as $\hat{\pi}^*$, which can be expressed as $\hat{\pi}^* = (\pi_1^*, \ldots, \pi_{|N|}^*)$. The optimal joint policy for a POMG can be obtained through the following equation:

$$\pi_i^* = \arg\max_{\hat{\pi}} \mathbb{E}_{\hat{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k \sum_{i\in N} r_{t+k}^i \middle| \hat{s}_t = \hat{s}_0\right], \tag{2}$$

where γ is the discount factor, and the expectation is taken over the trajectories via joint policy $\hat{\pi}$ starting from the initial state $\hat{s}_0$.

3 TAXONOMY AND DEFINITION OF HETEROGENEITY IN MARL

Heterogeneity in MAS. Our goal is to define agent heterogeneity from the perspective of MARL. Before achieving this, we discuss heterogeneity in MAS across various disciplines. Early studies [8, 27] define heterogeneity as differences in *physical structure* or *functionality* of agents, which aligns with common understanding. Later work [26] describes heterogeneity as differences in agent *behavior*, further expanding its meaning. Recently, [1] points out that heterogeneity may be a complex phenomenon, related not only to the *inherent properties* of agents, but also to their *interactions with environment*. Thus, heterogeneity in MARL should not be limited to inherent functional differences of agents, but should also fully consider various coupling effects of agents within the environment.

Heterogeneity in MARL. The fundamental modeling of MARL primal problem provides convenience for defining heterogeneity.

[1] POMG is an extension of POMDP for multi-agent settings, with the basic extension path being MDP → POMDP → POMG [31]. Please refer to Appendix D to see a more detailed explanation of POMG.

Table 1: Five Types of Heterogeneity in MARL

Heterogeneity Type	Heterogeneity Description	Related POMG Elements	Mathematical Definition
Observation Heterogeneity	Describes the differences of agents in observing global information	Agent's observation space and observation function	Agents i and j are observation heterogeneous if: ① $O^i \neq O^j$; or ② $\exists \hat{s} \in \{S^i\}_{i\in N}$, $\Omega^i(\cdot\|\hat{s}) \neq \Omega^j(\cdot\|\hat{s})$
Response Transition Heterogeneity	Describes the differences of agents in how their state transitions are affected by global environmental components (*environment-to-self*)	Agent's state space and local state transition function	Agents i and j are response transition heterogeneous if: ① $S^i \neq S^j$; or ② $\exists \hat{s} \in \{S^i\}_{i\in N}$, $\hat{a} \in \{A^i\}_{i\in N}$, $\mathcal{T}^i(\cdot\|\hat{s},\hat{a}) \neq \mathcal{T}^j(\cdot\|\hat{s},\hat{a})$
Effect Transition Heterogeneity	Describes the differences of agents in how their states and actions impact global state transitions (*self-to-environment*)	Agent's action space, state space, and global state transition function	Agents i and j are effect transition heterogeneous if: ① $S^i \neq S^j$; or ② $A^i \neq A^j$; or ③ $\exists s' \in S^{-i}, a' \in A^{-i}, s \in S^i, a \in A^i$, $\mathcal{T}^{-i}(\cdot\|s',s,a',a) \neq \mathcal{T}^{-j}(\cdot\|s',s,a',a)$
Objective Heterogeneity	Describes the differences of agents in the objective they aim to achieve	Agent's reward function	Agents i and j are objective heterogeneous if: ① $\exists \hat{s} \in \{S^i\}_{i\in N}, \hat{a} \in \{A^i\}_{i\in N}, r^i(\cdot\|\hat{s},\hat{a}) \neq r^j(\cdot\|\hat{s},\hat{a})$
Policy Heterogeneity	Describes the differences of agents in their decision-making based on observations	Agent's observation space, action space, and policy	Agents i and j are policy heterogeneous if: ① $O^i \neq O^j$; or ② $A^i \neq A^j$; or ③ $\exists o \in O^i$, $\pi_i(\cdot\|o) \neq \pi_j(\cdot\|o)$

This modeling specifies all MARL elements, delineating the boundaries of the problem discussion [2] and ensuring the completeness of the discussion.

We focus on the heterogeneity *among agents* within a same POMG. As mentioned in the previous sections, functions in POMG can serve as bridges linking other elements. Therefore, we focus on the functions and classify heterogeneity into five types. This approach can avoid redundant classification, and ensure coverage of each agent-level element. Specifically, these five types of heterogeneity are: *Observation heterogeneity*, *Response transition heterogeneity*, *Effect transition heterogeneity*, *Objective heterogeneity*, and *Policy heterogeneity*. Their specific descriptions and definitions are given in Table 1. In this table, $S^{-i} = \times_{k\in N, k\neq i} S^k \times S^E$ represents the joint state space of all agents except agent i, reflecting the influence of the agent on other states. Similarly, A^{-i} denotes the joint action space excluding agent i, and $\mathcal{T}^{-i}$ is the collection of state transitions excluding agent i.

The definitions in this section are relatively straightforward: if there are any differences in the associated elements, the agents are considered heterogeneous. We need to emphasize that our work goes beyond this. The quantification methods provided in the next section will be able to characterize the degree of agent heterogeneity related to certain attributes in practical scenarios, which far exceeds the level of definition.

4 QUANTIFYING HETEROGENEITY IN MARL

4.1 Heterogeneity Distance

According to the definition, each type of heterogeneity corresponds to a core function which connects relevant elements in the heterogeneity type. Therefore, we quantify the differences in these core functions to characterize the degree of heterogeneity.[3] To make the quantification results simpler and more practical, we draw upon the ideas of policy distance from the works [4] and [11], and present the concept of heterogeneity distance.

Let the core function corresponding to a certain heterogeneity type F be denoted as $y \sim F(\cdot|x)$. The formula for calculating the F-heterogeneous distance between two agents i and j is given by:

$$d_{ij}^F = \int_{x\in X} D[F_i(\cdot|x) \parallel F_j(\cdot|x)] \cdot p(x)\, dx, \tag{3}$$

where X is the space of independent variables, $p(x)$ is the probability density function, and $D[\cdot \parallel \cdot]$ is a measure that quantifies the difference between distributions. Unlike the works in [4] and [11], we add probability density terms to ensure accuracy and consider the case of multivariate variables. When the independent variables x consist of multiple factors, the above integral becomes a multivariate integral. Based on Equation 3, we provide the specific expressions for quantifying all heterogeneous distances in Appendix G and discuss the properties of heterogeneous distance in Appendix F.

4.2 Practical Method

To compute Equation 3 in practice, we need to address several core issues: 1) Full space traversal. In practice, it is impossible to traverse the entire space X. 2) Measure D is difficult to compute. Even assuming we can obtain model F for each agent, the distribution types of F may vary, making it difficult to calculate measures between different distributions. 3) Handling cases when model F is not available. More commonly, it is hard to obtain environment-based agent models, especially in practical MARL tasks.

For issue 1, our approach is sampling based on the interaction between agents and the environment. Instead of simply traversing

[2] In this paper, we focus on the heterogeneity of MARL under the conventional POMG problem. Additional discussions on unconventional heterogeneity types are provided in Appendix E.

[3] Quantifying space elements is feasible and even easier to implement. But a space element may appear across multiple heterogeneity types, making it unsuitable as unique identifiers for specific heterogeneity types.

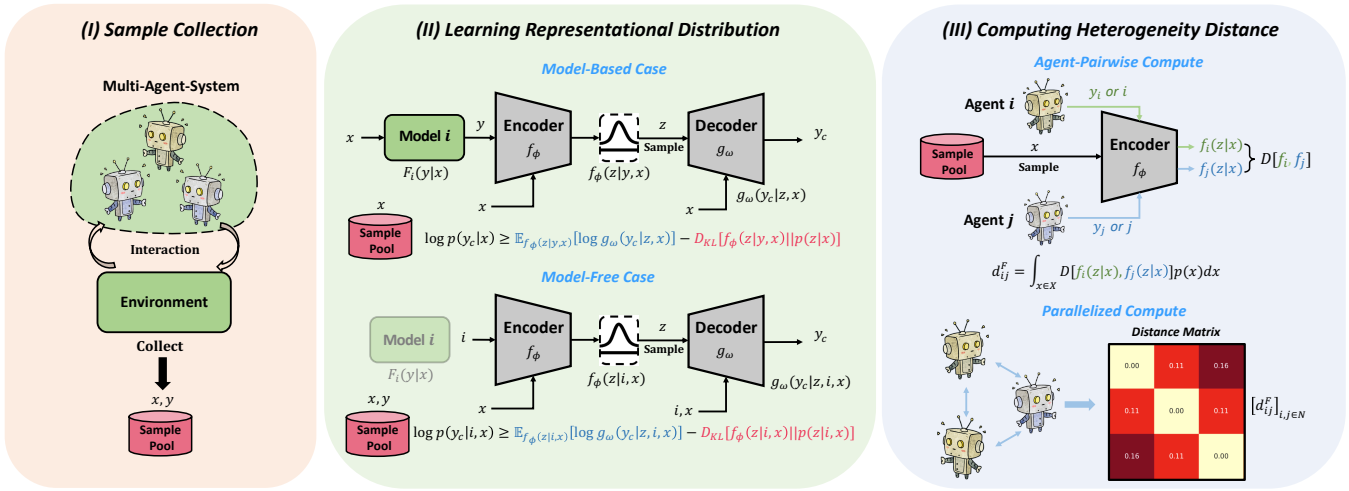


Figure 2: The method of measuring heterogeneity distance based on representation learning.

the space or using random policy exploration, we construct a sample pool using trajectories from the training phase of MARL. This significantly reduces computational load and filters out excessive marginal spaces, benefiting the use of heterogeneity distance in subsequent MARL tasks (Section 5). For issue 2, this has been solved in paper [11] (computing policy distance). We follow their approach, which performs representation learning on F and maps it to a standardized distribution. For issue 3, we extend the representation learning method to model-free cases. This helps apply the method in real-world settings and enables us to propose the concept of *Meta-Heterogeneity Distance*. By freely combining different attributes to construct *Meta-Transitions*, the proposed method can quantify the "comprehensive heterogeneity" of agents.

Combining these ideas, we propose a practical method as shown in Figure 2. **In the first step**, the agents interact with the environment during MARL training to build a sample pool. Notably, the sample pool data is shuffled to ensure that the learned function follows the *Markov* property (independent of historical information).

In the second step, the representational distributions are learned. We discuss this in both model-based and model-free settings, corresponding to cases function F is known and unknown. We adopt the conditional variational autoencoder (CVAE) [30] for representation learning. In the model-based case, CVAE performs a reconstruction task [20]. The optimization goal is to maximize the likelihood of the reconstructed variable $\log p(y|x)$. Through derivation, we obtain the evidence lower bound (ELBO) as:

$$\begin{aligned} ELBO_{\text{model-based}} =& \mathbb{E}_{f_\phi(z|y,x)}\left[\log g_\omega(y|z,x)\right] \\ & - D_{KL}\left[f_\phi(z|y,x) \parallel p(z|x)\right], \end{aligned} \tag{4}$$

where f_ϕ and g_ω represent the encoder and decoder, respectively, and $p(z|x)$ is the prior conditional latent distribution. The relevant losses are designed based on ELBO, including a reconstruction loss and a prior-matching loss.

In the model-free case, CVAE essentially performs a prediction task [36], capturing the model characteristics of each agent. The network takes the independent variable x and agent ID i as inputs, using both as conditions to predict y. The optimization goal is to maximize the likelihood of the predicted y given conditions. Similarly, the corresponding ELBO can be derived as (the derivation for this part can be found in Appendix I):

$$\begin{aligned} ELBO_{\text{model-free}} =& \mathbb{E}_{f_\phi(z|i,x)}\left[\log g_\omega(y|z,i,x)\right] \\ & - D_{KL}\left[f_\phi(z|i,x) \parallel p(z|i,x)\right]. \end{aligned} \tag{5}$$

In the third step, the heterogeneity distances for MAS are computed. For each x, we obtain the distribution representation using the encoder in either the model-based or model-free manner. The distance under a specific x is computed using the *Wasserstein distance* [32] of the prior distribution (*standard Gaussian*). The heterogeneity distance is then calculated via multi-rollout Monte Carlo sampling. In practice, we parallelize this operation, enabling simultaneous computation of distances between all agents on GPUs, significantly improving computational efficiency.

Meta-Transition. The aforementioned method can quantify the heterogeneity of agents for specific types. In practical applications, researchers may also want to quantify the **comprehensive** heterogeneity of agents to enable operations such as grouping. To this end, we give the *Meta-Transition* model (see Appendix H for details). By measuring the differences between meta-transitions, the comprehensive heterogeneity related to environment can be quantified. We refer to this as the meta-transition heterogeneity distance (Hereafter referred to as *Meta-Het*).

4.3 Case Study

We design a multi-agent spread scenario for case study. In the basic scenario, there are two groups, each with two agents, and their goal is to move to randomly generated landmarks. We create 6 versions of the scenario to show the quantitative results of different types of heterogeneity and *Meta-Het*. As shown in Figure 3, the first 4 versions correspond to the 4 environment-related types of heterogeneity, while the last 2 versions represent cases where multiple types of heterogeneity exist. We use the model-based manner to

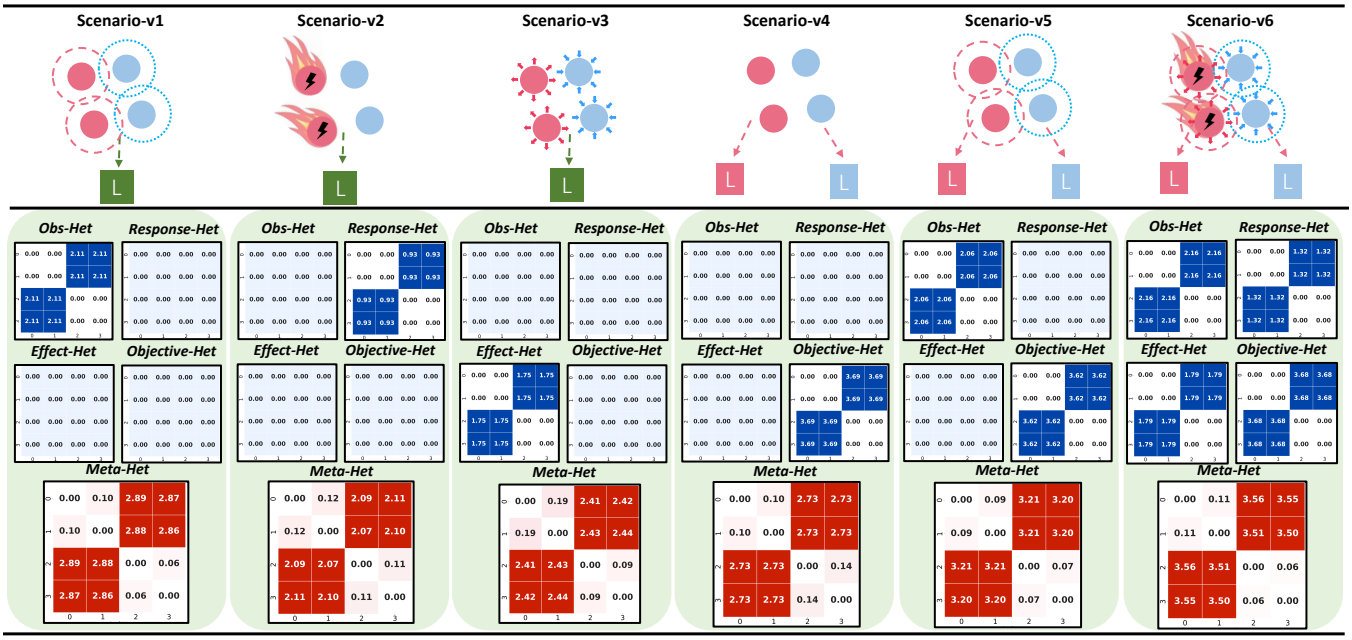


Figure 3: The scenario illustration and heterogeneity distance matrices. In *v1*, the observations of agents from different groups are shuffled in different orders. In *v2*, the max speeds of agents are different. In *v3*, one group of agents applies repulsive force to surrounding entities, while the other attractive force. In *v4*, agents need to move to different landmarks. In *v5*, both the observations and objectives of agents are heterogeneous. In *v6*, all the above properties are heterogeneous. Below each scenario illustration, the corresponding heterogeneity distance matrices are shown. Specifically, *Obs-Het*, *Response-Het*, *Effect-Het*, and *Objective-Het* correspond to observation / response transition / effect transition / objective heterogeneity, respectively.

compute the first four distance matrices, and the model-free manner to compute the *Meta-Het* distance matrix.

The results show that for each type of heterogeneity, our method can accurately capture and identify the differences. And the *Meta-Het* distance between agents in the same group is much smaller than that in different groups. Moreover, as the number of heterogeneity types increases, the *Meta-Het* distance between different groups also increases. These results demonstrate the effectiveness of our method for various environment-related heterogeneities.

We further quantify the policy heterogeneity distance (*Policy-Het*) and *Meta-Het* distance of agents during the training process. We select two algorithms at the extreme cases of parameter sharing: fully parameter sharing (FPS) and no parameter sharing (NPS) for training in the above scenarios. Figure 4 shows the measurement results at 500 and 1500 updates. From the *Policy-Het* results, the policy distance can effectively reveal the evolution of agent policy differences in MARL. From the *Meta-Het* results, the comprehensive agent heterogeneity measurement remains consistent across different learning algorithms, and can identify environmental heterogeneous characteristics in scenarios more rapidly compared to policy evolution.

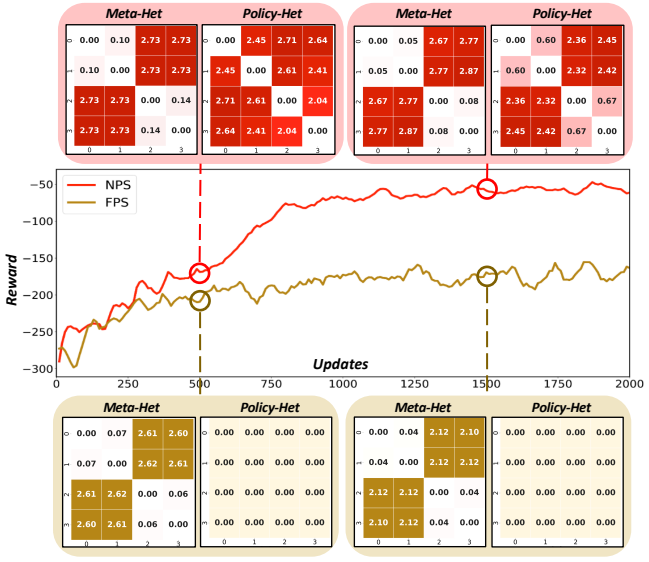


Figure 4: Meta-transition heterogeneity and policy heterogeneity distance matrices during training in our case study.

5 UTILIZING HETEROGENEITY IN MARL

Based on the case study in Section 4.3, the proposed method can not only accurately quantify all types of heterogeneity, but also the "comprehensive heterogeneity" among agents. Additionally, the method is independent of the parameter-sharing type used in MARL and can be deployed online, thereby further enhancing its

Table 2: Comparison of different methods and their properties.

Method	Paradigm	Adaptive	Relation to Heterogeneity Utilization
NPS	No Sharing	No	None
FPS	Full Sharing	No	None
FPS+id	Full Sharing	No	None
Kaleidoscope [18]	Partial Sharing	Yes	No utilization, increases agent policy heterogeneity as the bias
SePS [6]	Group Sharing	No	Implicitly utilizes objective heterogeneity and response transition heterogeneity
AdaPS [17]	Group Sharing	Yes	Implicitly utilizes objective heterogeneity and response transition heterogeneity
MADPS [11]	Group Sharing	Yes	Explicitly utilizes policy heterogeneity only
HetDPS (ours)	Group Sharing	Yes	Explicitly utilizes heterogeneity, leveraging heterogeneous distance

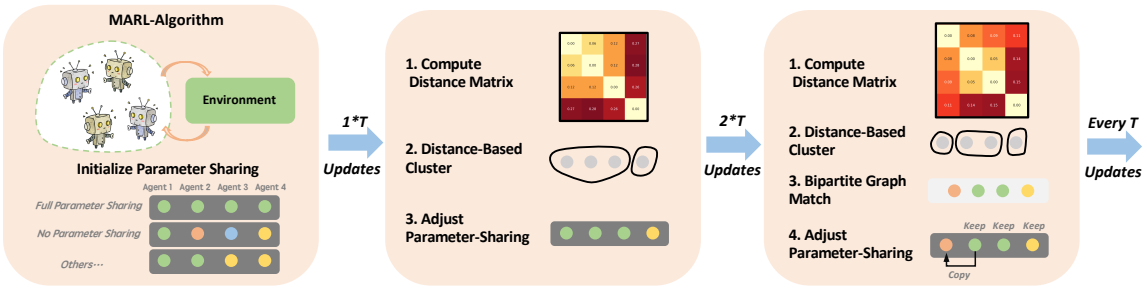


Figure 5: The method of multi-agent dynamic parameter sharing algorithm based on heterogeneity quantification.

practicality. In this section, we provide a practical application of our methodology to demonstrate its potential in empowering MARL.

We select parameter sharing in MARL as our application context. As a common technique in MARL, parameter sharing can improve sample utilization efficiency [14], but its excessive use may inhibit agents' policy heterogeneity expression [11]. Many works have attempted to find a balance between parameter sharing and policy heterogeneity [18]. However, existing approaches suffer from two main problems: *poor interpretability*, unable to explain why policy heterogeneity is necessary and to what extent; and *poor adaptability*, manifested by numerous task-specific hyperparameters

To address these issues, we propose a **Het**erogeneity-based multi-agent **D**ynamic **P**arameter **S**haring algorithm (HetDPS) with two core ideas (More details can be found in Appendix J):

♠ **Grouping agents for parameter sharing through heterogeneity distances**. We utilize distance-based clustering methods to group agents, thus avoiding the introduction of task-specific hyperparameters like group number [6, 17] or fusion thresholds [11]. The heterogeneity distance matrices also enhance the algorithm's interpretability.

♣ **Periodically quantifying heterogeneity and modifying agents' parameter sharing paradigm**. This approach can help policies escape local optima [22], the effectiveness of such a mechanism has been verified in the MARL domain [18], and even in broader RL areas such as large model fine-tuning [23, 25].

Combining the above ideas, we present the method of HetDPS as illustrated in Figure 5. This approach can be combined with common MARL algorithms and supports various parameter-sharing initialization (e.g., FPS and NPS). After every T updates, the algorithm computes the distance matrix of agents and groups them via distance-based clustering. If clustering exists from the previous cycle, bipartite graph matching is performed between the two clustering results to help agents determine policy inheritance relationships. This *dual-clustering mechanism* effectively enhances the algorithm's adaptability.

We emphasize that utilization of MARL heterogeneity extend beyond this scope. Through our method, researchers can quantify specific types of heterogeneity or composite heterogeneity, which can be integrated with cutting-edge MARL research directions, as detailed in Appendix C.

Table 3: Task information for PMS.

Task	Agent Type Distribution
15a_3c	5 − 5 − 5
30a_3c	10 − 10 − 10
15a_5c	3 − 3 − 3 − 3 − 3
30a_5c	3 − 3 − 3 − 12 − 9

Table 4: Agent distribution in four heterogeneous SMAC tasks.

Task	Agent Type Distribution
3s5z	3 Stalkers (0–2) − 5 Zealots (3–7)
3s5z_vs_3s6z	3 Stalkers (0–2) − 5 Zealots (3–7)
MMM	2 Marauders (0–1) − 7 Marines (2–8) − 1 Medivac (9)
MMM2	2 Marauders (0–1) − 7 Marines (2–8) − 1 Medivac (9)

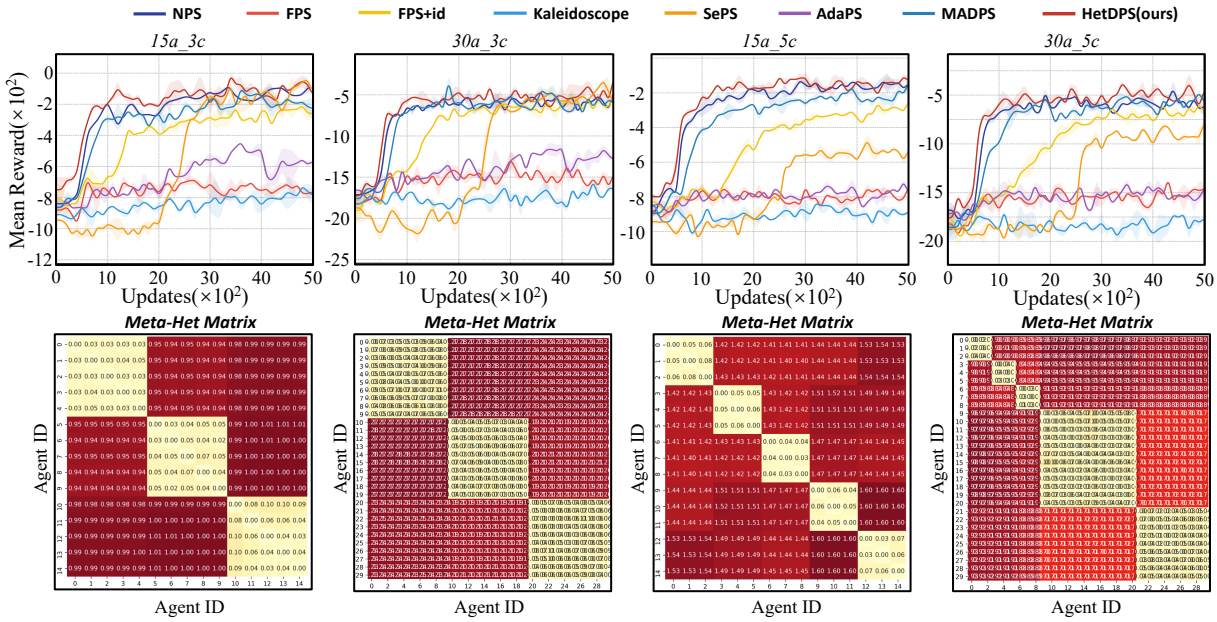


Figure 6: Results on Partical-based Multi-agent Spreading.

6 EXPERIMENTS

In this section, we conduct comprehensive comparisons between HetDPS and other parameter sharing methods. Beyond performance comparisons, we also analyze the heterogeneity characteristics of each MARL task with our methodology, to demonstrate the algorithm's interpretability. Additionally, we conduct hyperparameter experiments and efficiency and resource consumption experiments, to show the adaptability and practicality of HetDPS.

6.1 Experimental Setups

Environments. Partical-based Multi-agent Spreading (PMS) [11] is a typical environment in the policy diversity domain. In this environment, multiple agents are randomly generated in the center of the map, while multiple landmarks are generated near the periphery. Both agents and landmarks have various colors, and agents need to move to landmarks with matching colors. Additionally, agents need to form tight formations when they reach the vicinity of landmarks. We employ 4 typical tasks, corresponding to different numbers and color distributions, as detailed in Table 3. **The StarCraft Multi-Agent Challenge (SMAC)** [28] is a popular MARL benchmark, where multiple ally units controlled by MARL algorithms aim to defeat enemy units controlled by built-in AI.

Baselines and training. We compare HetDPS with other parameter sharing baselines, as listed in Table 2. As seen from the table, current methods can not effectively utilize heterogeneity. Although some methods implicitly use certain heterogeneity quantification results, the elements they involve are not comprehensive. MADPS, as the only method that explicitly uses policy distance for dynamic grouping, relies on the assumption that policy learning can effectively capture heterogeneity, which lacks practicality. All parameter-sharing methods are integrated with MAPPO, and we use official implementations of the baselines wherever available. For more details of the experiments, see the Appendix K.

6.2 Results

Performance and interpretability. The reward curves and corresponding heterogeneity distance matrices are shown in Figure 6, Figure 7 (More in Appendix). From the reward curve results, we can see that HetDPS achieves either optimal or comparable results across all tasks.

The *Meta-Het* distances in Multi-agent Spreading scenario closely match the type distributions in Table 3, validating demonstrating the effectiveness of our method in identifying agent heterogeneity. In SMAC, we observe that in simpler tasks like *3s5z* and *MMM*, the heterogeneity distances often do not closely match the original agent types. In *MMM*, agents even tend toward homogeneous policies to improve training efficiency. However, in more difficult tasks such as *3s5z_vs_3s6z* and *MMM2*, agents' quantification results closely match their original types for better coordination. This confirms that agent heterogeneity depends on both functional attributes and environment interactions.

Similarly, results in "homogeneous" tasks (details in Appendix) reveals that even "homogeneous" agents exhibit emergent heterogeneity from environment interactions, leading to role division. The performance difference between HetDPS and FPS also reflects the impact of role division versus non-division. Our method thus provides both superior performance and strong interpretability for exploring heterogeneity in MARL tasks.

Adaptability. Our approach achieves comparable performance across all tested tasks. Moreover, we emphasize that **for all tested tasks, our method uses identical hyperparameters**, without requiring task-specific tuning. Other baselines require task-specific hyperparameters: e.g., reset interval, reset rate, and diversity loss

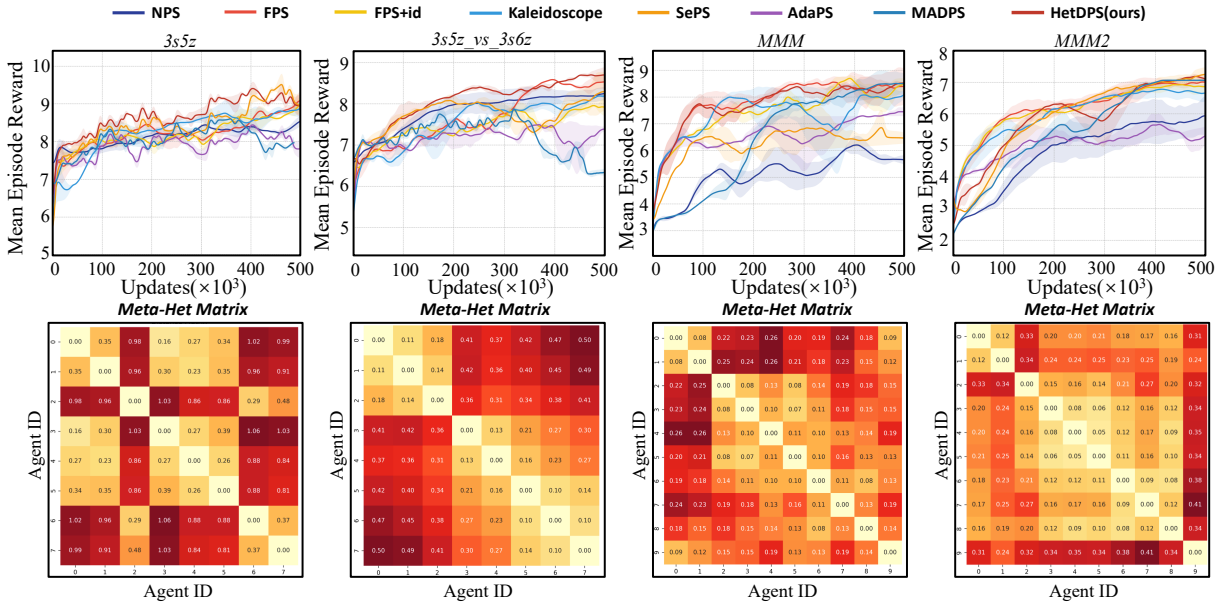


Figure 7: Results on four "heterogeneous" tasks in SMAC.

Table 5: Training efficiency metrics across different methods. Results are normalized with respect to the FPS method.

	NPS	FPS	FPS+id	Kaleidoscope	SePS	AdaPS	MADPS	HetDPS (ours)
Training Speed	0.952x	1.000x	0.992x	0.974x	0.986x	0.614x	0.539x	0.712x

Table 6: Results of varying quantization intervals in PMS, showing the average rewards of agents.

Quantization Interval	*15a_3c*	*30a_3c*	*15a_5c*	*30a_5c*
20 *Updates*	-10.12	-300.56	-50.89	-350.23
100 *Updates*	-9.45	-298.91	-49.32	-349.67
200 *Updates*	-10.78	-301.34	-51.15	-351.45
1000 *Updates*	-11.23	-299.67	-50.44	-350.89
2000 *Updates*	-9.87	-300.12	-49.78	-349.12

coefficient for Kaleidoscope; number of clusters and update interval for SePS and AdaPS; fusion/division threshold and quantization interval for MADPS.

HetDPS employs distance-based clustering, eliminating hyperparameters such as cluster number or fusion threshold. Furthermore, by fully accounting for dual-clustering mechanism, HetDPS is insensitive to the quantization interval. Table 6 shows that performance remains stable across quantization intervals ranging from 20 to 2000 in all multi-agent spreading tasks.

Cost Analysis. We conduct an experiment to investigate training efficiency. The experimental results are shown in Table 5. The results indicate that although our method introduces periodic heterogeneity quantification, it does not significantly reduce algorithm efficiency.

7 CONCLUSION

Heterogeneity manifests in various aspects of MARL. It is not only related to the inherent properties of agents but also to the coupling factors arising from agent-environment interactions. Consequently, agents that appear homogeneous may develop heterogeneity under environmental influences. In this paper, we categorize heterogeneity in MARL into five types and provide definitions. Meanwhile, we propose methods for quantifying these heterogeneities and conduct case studies. Under our theoretical framework, policy diversity is merely a manifestation of policy heterogeneity, fundamentally originating from the division of labor necessitated by agents' environmental heterogeneity (*cause*), serving as an inductive bias (*result*) for solving optimal joint policies. Thus, we introduce the quantification of heterogeneity as prior knowledge into multi-agent parameter-sharing learning, resulting in HetDPS, an algorithm with strong interpretability and adaptability. HetDPS is not the endpoint of our research, but rather a starting point for heterogeneity applications. We believe that by systematically studying the definition, quantification, and application of heterogeneity, future MARL research will more profoundly understand the complex collaboration mechanisms between agents, and pave the way for more intelligent and adaptive collective decision-making systems.

REFERENCES

[1] Chris Bennett. 2024. *Heterogeneity in multi-agent systems.* Ph.D. Dissertation. University of Bristol.

[2] Matteo Bettini, Ryan Kortvelesy, and Amanda Prorok. 2024. Controlling Behavioral Diversity in Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*. PMLR, 3611–3636.

[3] Matteo Bettini, Ajay Shankar, and Amanda Prorok. 2023. Heterogeneous Multi-Robot Reinforcement Learning. In *AAMAS*.

[4] Matteo Bettini, Ajay Shankar, and Amanda Prorok. 2023. System neural diversity: Measuring behavioral heterogeneity in multi-agent learning. *arXiv preprint arXiv:2305.02128* (2023).

[5] Alicia L Burns, Alexander DM Wilson, and Ashley JW Ward. 2019. Behavioural interdependence in a shrimp-goby mutualism. *Journal of Zoology* 308, 4 (2019), 274–279.

[6] Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. 2021. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*. PMLR, 1989–1998.

[7] Emiliano Dall'Anese, Hao Zhu, and Georgios B Giannakis. 2013. Distributed optimal power flow for smart microgrids. *IEEE Transactions on Smart Grid* 4, 3 (2013), 1464–1475.

[8] Gregory Dudek, Michael RM Jenkin, Evangelos Milios, and David Wilkes. 1996. A taxonomy for multi-agent robotics. *Autonomous Robots* 3 (1996), 375–397.

[9] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review* 55, 2 (2022), 895–943.

[10] Xudong Guo, Daming Shi, Junjie Yu, and Wenhui Fan. 2024. Heterogeneous Multi-Agent Reinforcement Learning for Zero-Shot Scalable Collaboration. *arXiv preprint arXiv:2404.03869* (2024).

[11] Tianyi Hu, Zhiqiang Pu, Xiaolin Ai, Tenghai Qiu, and Jianqiang Yi. 2024. Measuring Policy Distance for Multi-Agent Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 834–842.

[12] Jiechuan Jiang and Zongqing Lu. 2021. The emergence of individuality. In *International Conference on Machine Learning*. PMLR, 4992–5001.

[13] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*. PMLR, 651–673.

[14] WOOJUN KIM and Youngchul Sung. 2023. Parameter Sharing with Network Pruning for Scalable Multi-Agent Deep Reinforcement Learning. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. AAMAS.

[15] Mykel J Kochenderfer, Tim A Wheeler, and Kyle H Wray. 2022. *Algorithms for decision making*. MIT press.

[16] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 3991–4002.

[17] Dapeng Li, Na Lou, Bin Zhang, Zhiwei Xu, and Guoliang Fan. 2024. Adaptive parameter sharing for multi-agent reinforcement learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6035–6039.

[18] Xinran Li, Ling Pan, and Jun Zhang. 2024. Kaleidoscope: Learnable Masks for Heterogeneous Multi-agent Reinforcement Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[19] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[20] Manuel Lopez-Martin, Belen Carro, Antonio Sanchez-Esguevillas, and Jaime Lloret. 2017. Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot. *Sensors* 17, 9 (2017), 1967.

[21] Jonathan Lwowski, Patrick Benavidez, John J Prevost, and Mo Jamshidi. 2017. Task allocation using parallelized clustering and auctioning algorithms for heterogeneous robotic swarms operating on a cloud network. *Autonomy and artificial intelligence: A threat or savior?* (2017), 47–69.

[22] Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado P van Hasselt, Razvan Pascanu, and Will Dabney. 2024. Normalization and effective learning rates in reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 106440–106473.

[23] Hao Ma, Tianyi Hu, Zhiqiang Pu, Liu Boyin, Xiaolin Ai, Yanyan Liang, and Min Chen. 2024. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 15497–15525.

[24] Zepeng Ning and Lihua Xie. 2024. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence* 3, 2 (2024), 73–91.

[25] Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C Courville. 2023. Language model alignment with elastic reset. *Advances in Neural Information Processing Systems* 36 (2023), 3439–3461.

[26] Liviu Panait and Sean Luke. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems* 11 (2005), 387–434.

[27] Lynne E Parker. 2000. Lifelong adaptation in heterogeneous multi-robot teams: Response to continual variation in individual robot performance. *Autonomous Robots* 8 (2000), 239–267.

[28] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR* abs/1902.04043 (2019).

[29] Esmaeil Seraj, Zheyuan Wang, Rohan Paleja, Matthew Sklar, Anirudh Patel, and Matthew Gombolay. 2021. Heterogeneous graph attention networks for learning diverse communication. *arXiv preprint arXiv:2108.09568* (2021).

[30] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[31] Lijun Sun, Yu-Cheng Chang, Chao Lyu, Ye Shi, Yuhui Shi, and Chin-Teng Lin. 2023. Toward multi-target self-organizing pursuit in a partially observable Markov game. *Information Sciences* 648 (2023), 119475.

[32] Leonid Nisonovich Vaserstein. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii* 5, 3 (1969), 64–72.

[33] T Wang, T Gupta, B Peng, A Mahajan, S Whiteson, and C Zhang. 2021. RODE: learning roles to decompose multi- agent tasks. In *Proceedings of the International Conference on Learning Representations*. OpenReview.

[34] H Peyton Young. 1993. The evolution of conventions. *Econometrica: Journal of the Econometric Society* (1993), 57–84.

[35] Xiaoyang Yu, Youfang Lin, Xiangsen Wang, Sheng Han, and Kai Lv. 2024. GHQ: grouped hybrid Q-learning for cooperative heterogeneous multi-agent reinforcement learning. *Complex & Intelligent Systems* 10, 4 (2024), 5261–5280.

[36] Chen Zhang, Riccardo Barbano, and Bangti Jin. 2021. Conditional variational autoencoder for learned image reconstruction. *Computation* 9, 11 (2021), 114.

[37] Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. 2021. Smarts: An open-source scalable multi-agent rl training school for autonomous driving. In *Conference on robot learning*. PMLR, 264–285.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62322316, the National Natural Science Foundation of China under Grant No. 62503472, the Open Fund of National Key Laboratory of Information Systems Engineering (No. 6142101240203), and the Young Scientists Foundation of CSAA (Guidance Navigation and Control, GNC) under Grant CSAA-YSF2025-GNC-08. (Corresponding author: Zhiqiang Pu.)