

A Related Works

In this section, we review related works in the multi-agent reinforcement learning (MARL) domain concerning heterogeneity. These works can be broadly categorized into three types: one focuses on traditional heterogeneous MARL, another on policy heterogeneity in multi-agent systems (MAS), and the last on the identification and quantification of heterogeneity¹.

Related Work on Traditional Heterogeneous MARL. This group of works focuses on reinforcement learning for traditional heterogeneous MAS. [Marques and do Patrocínio Júnior, 2023] discuss the advantages of heterogeneous MAS over homogeneous MAS in several toy environments, where agent heterogeneity is reflected in differences in environment observations and the impact of actions on the environment. [Meneghetti and Bianchi, 2020, Seraj et al., 2021] focus on establishing learnable communication mechanisms for heterogeneous MARL to enhance multi-agent collaboration, with agent heterogeneity characterized by observation spaces, action spaces, and local state transitions. [Jiang et al., 2023, Yu et al., 2024] concentrate on the credit assignment problem in heterogeneous MARL, where agent heterogeneity is defined by observations, local state transitions, and rewards. [Guo et al., 2024] propose a scalable and heterogeneous PPO algorithm to address zero-shot generalization in heterogeneous MARL, describing heterogeneity as the diversity of agents' functions and abilities. [Hu et al., 2024a] provide a detailed analysis of observation heterogeneity in MARL and design a graph neural network to help agents aggregate heterogeneous observation information. [Hu et al., 2025] develop a hybrid actor-critic method to tackle the challenges of joint policy training instability in physically heterogeneous MARL [Bettini et al., 2023a].

However, the aforementioned works do not provide a rigorous definition of heterogeneity in MARL. They typically address only one or a few aspects of heterogeneity discussed in this paper, with their perspective largely limited to traditional functional heterogeneity.

Related Work on Policy Heterogeneity in MARL. This group of works focus on the heterogeneity of agents' policies in MARL. [Bettini et al., 2023a] categorizes heterogeneity in MAS into physical heterogeneity and behavioral heterogeneity, and develops the HetGPPO algorithm to ensure training stability of the multi-agent PPO under heterogeneous parameter settings. [Zhong et al., 2024] discusses the advantages and disadvantages of homogeneous and heterogeneous policies in MARL, and designs an asynchronous update method that theoretically guarantees monotonically increasing learning. Similarly, some methods identify the limitations of homogeneous policies and excessive parameter sharing, resulting a series of algorithms based on policy diversity. [Wang et al., 2020, 2021, Liu et al., 2022, Nguyen et al., 2022] introduce the concept of *roles* to help MAS learn diverse policies, but their definitions of roles vary. This is reflected in the different *prior constraints* they impose on policy learning, which may include learning distinguishable policy encodings for agents, grouping agent policies based on environmental learning, or varying discount factors for agents.

Other works focus on the paradigm of parameter sharing. Some propose static or dynamic pruning on a large shared-parameter network to achieve policy heterogeneity [KIM and Sung, 2023, Li et al., 2024b], while others divide MAS policies into a combination of shared and non-shared parameters [Li et al., 2021]. Some methods directly group agent policies statically or dynamically [Christianos et al., 2021, Li et al., 2024a, Hu et al., 2024b]. These works apply different prior constraints to agent policy learning. For example, some aim for policies to be as heterogeneous as possible, some group policies based on static or dynamic quantification of agent heterogeneity, and others impose no constraints on policy learning. This essentially reflects their different approaches to utilizing agent heterogeneity.

It is evident that works on policy heterogeneity and diversity not only fall within the scope of agent heterogeneity, but also (directly or indirectly) identify and utilize heterogeneity in MARL. Our work can effectively categorize and integrate these approaches into a unified framework, and the proposed methods will contribute to advancing this field.

Related Work on Heterogeneity Identification and Quantification in MARL. This group of works further advances the understanding of heterogeneity in the MARL field, focusing primarily on the identification and quantification of heterogeneity. As mentioned above, some works on policy heterogeneity have implicitly identified agent heterogeneity. For instance, RODE [Wang et al., 2021] and SePS [Christianos et al., 2021] group agents by learning a latent variable related to the agents.

¹This classification also indirectly reflects the gradual expansion of understanding agent heterogeneity within the MARL domain, aligning with the development of heterogeneity in the broader MAS field [Bennett, 2024].

In this latent variable learning, the input consists of the agents’ observations and actions, while the output, reconstructed from the latent variable, includes the next-time observation and corresponding reward. Under our methodology, these works essentially identify partial meta-transition heterogeneity of agents. However, these methods directly map agents to a single latent variable, which leads to significant information loss, whereas our approach also accounts for the implicit distribution differences across the entire variable space.

Compared to works identifying environment-related heterogeneity, more research focuses on identifying and quantifying policy heterogeneity among agents. [McKee et al., 2022] introduces a "policy distance" by sampling states and recording the ratio of different actions selected by the population. [Hu et al., 2022] employs KL divergence to measure distances between action-based roles, while [Masood and Doshi-Velez, 2019] uses maximum mean discrepancy to compare distributions over trajectories under different policies. [Liu et al., 2021] formulates behavioral diversity as discrepancies in occupancy measures. Some studies explore policy representations. In [Grover et al., 2018], episodes generated by agent policies are mapped to real-valued vectors, and the Euclidean distances between these vectors are used. [Jiang et al., 2021] maps agent policies to a distribution and employs KL divergence and one-dimensional Wasserstein distance to handle discrete and continuous action spaces, respectively. These approaches, which map overall policies to low-dimensional vectors or distributions, inevitably result in significant information loss. [Bettini et al., 2023b] calculates the "behavioral distance" between pairs of agents and uses the average of these distances as an indicator of system diversity. Furthermore, [Hu et al., 2024b] builds on this method by learning representations of policy distributions to quantify "policy distance," establishing the current state-of-the-art approach. Under our methodology, the aforementioned policy distance work is just a special case of our proposed quantification method, under model-based and policy heterogeneity settings. Our approach not only accommodates both model-based and model-free cases, but also can quantify a broader range of agent heterogeneity.

B Experimental Details

In this section, we first provide the detailed setup of the experiments, followed by a more detailed discussion on the interpretability and adaptability of our proposed method.

Table 1: Hyperparameters used for actor-critic-based algorithms.

	SMAC	Multi-agent Spreading
learning rate	0.001	0.0005
discount factor	0.99	0.99
GAE lambda	0.95	0.95
coefficient of entropy	0.01	0.01
coefficient for value loss	0.5	0.5
optimization epochs	16	4
clip parameter	0.2	0.2
clip parameter of value function	0.2	0.2
maximum gradient norm	0.5	0.5

B.1 Experimental Settings

Common Settings. For the overall hyperparameters of the experiment, all experiments in SMAC include 50 timesteps per update (5 timesteps in Multi-agent Spreading). Each experiment has 32 parallel environments. Therefore, one update corresponds to 50×32 or 5×32 timesteps. Team rewards, win rates, and other information are recorded every 1,000 updates for plotting in SMAC, and 100 updates for plotting in Multi-agent Spreading.

Common Algorithm Settings. Since we compare the performance of different parameter-sharing methods, we use the same network architecture and MARL algorithm to ensure fairness. Specifically, we adopt the PPO algorithm [Schulman et al., 2017, Yu et al., 2022] as the base algorithm for agent training, with detailed parameters shown in Table 1. For the network architecture, we ensure that the actor and critic networks are identical and follow the same parameter-sharing paradigm. All networks use multilayer perceptrons with a hidden layer dimension of 64.

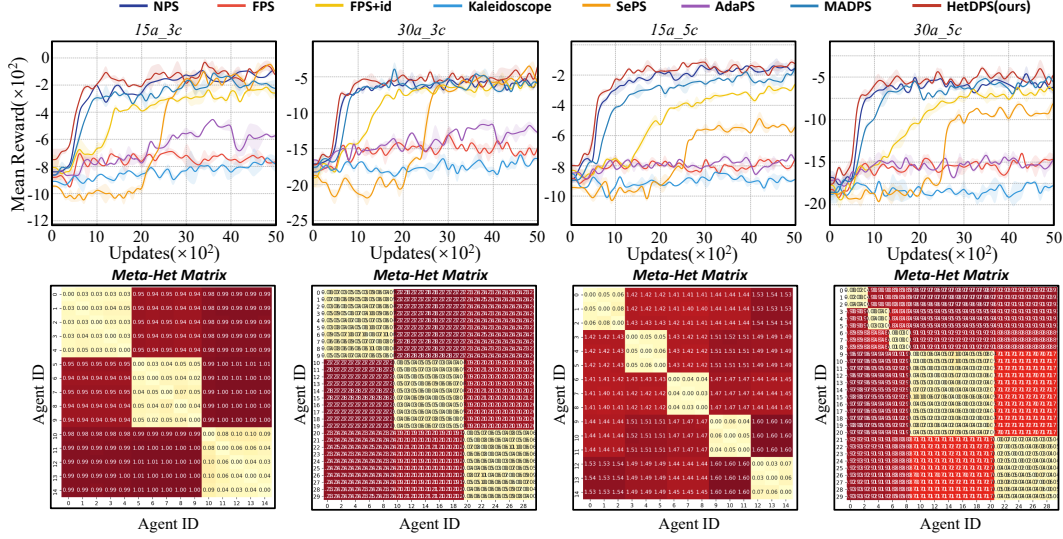


Figure 1: Results on Partial-based Multi-agent Spreading.

Algorithm-specific Settings. In HetDPS, the hidden dimension of the CVAE is 40, the learning rate is 0.0005, the optimization epochs are 10, and the KL weight is 0.0001. For other algorithms, the hyperparameters are kept at their standard default configurations as described in the original papers. In particle-based tasks, the update interval for all intervalic parameter-sharing paradigm algorithms is 2,000 updates, with quantification based on the last 200 updates. In SMAC tasks, the update interval is 5,000 updates, with quantification based on the last 500 updates.

B.2 Strong Interpretability

Figure 1 presents the experimental results of HetDPS and other baseline methods on the multi-agent spreading task, while Figure 2 and Figure 3 show the results of the tested algorithms on 8 hard-level tasks in SMAC. The experimental results for each task include the reward curves of the tested algorithms, as well as the multi-agent meta-transition heterogeneity distance matrices calculated using our proposed quantification method.

From the results on the multi-agent spreading task, our HetDPS outperforms others in both convergence speed and final convergence performance. Moreover, the results of the heterogeneous distance matrices are fully consistent with the agent distribution in each task. This demonstrates that our method can not only identify an appropriate parameter-sharing paradigm but also explain why agents are grouped. In contrast, the SePS method cannot dynamically adjust the parameter-sharing of agents during training, leading to poor performance. MADPS groups agents based on policy distances, relying on the assumption that agent policy learning can capture environmental heterogeneity, which results in relatively low learning efficiency. AdaPS struggles with handling transitions when agents switch parameter-sharing modes, leading to unstable training. Kaleidoscope fails to find a suitable parameter-sharing paradigm, and its prior assumption of overly pursuing diversity hinders the policy training of multi-agents.

From the results on SMAC, our HetDPS achieves the best performance across all tasks. Compared to its performance on the multi-agent spreading task, the NPS method performs well in particle environments but struggles in SMAC, where stronger collaboration is required. In contrast, Kaleidoscope performs significantly better in SMAC than in multi-agent spreading tasks, as its loss function promoting policy diversity aids exploration but hinders precise agent grouping. These results highlight the significant differences between the two selected environments. SMAC, as a more complex environment, exhibits agent heterogeneity that depends not only on the agents' types but also on their interactions with the environment. Figure 2 displays the meta-transition heterogeneity matrices of agents in two groups of SMAC tasks. In less challenging tasks (*3s5z* and *MMM*), agents can complete tasks without clear role division, leading to policy distance matrices that deviate significantly from

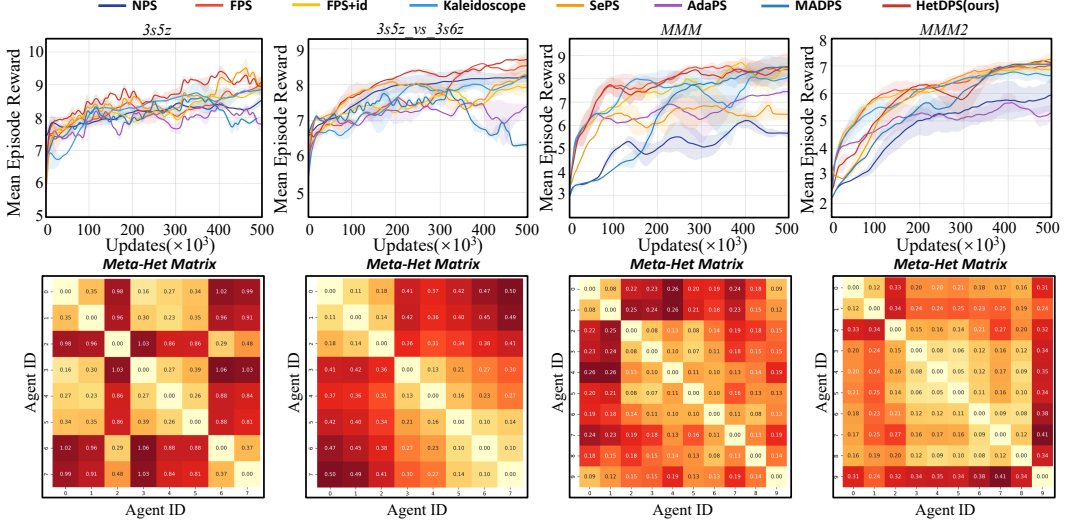


Figure 2: Results on StarCraft Multi-Agent Challenge (Part I).

the agents’ original types. In such cases, agents can quickly learn to complete tasks under parameter sharing. However, in more difficult tasks (*3s5z vs 3s6z* and *MMM2*), agents require distinct roles to collaborate effectively. Here, the heterogeneity distance matrices closely aligns with the agents’ original type distribution. These results indicate that agent heterogeneity is influenced not only by the agents’ inherent properties but also by their interactions with the environment.

Similarly, Figure 3 shows the meta-transition heterogeneity matrices for agents in four "homogeneous" agent tasks in SMAC. The results reveal that even in so-called homogeneous agent settings, agents still exhibit heterogeneity. This heterogeneity emerges from the agents’ interactions with the environment. Although agents share identical physical attributes, their interactions with the environment lead to heterogeneity, resulting in further role division. The performance difference between HetDPS and FPS also reflects the impact of role division versus non-division. These experiments demonstrate that our method not only outperforms other baselines across tasks but also provides strong interpretability. This facilitates further exploration of the deeper relationships between heterogeneity in various MARL tasks.

B.3 Strong Adaptability

As demonstrated by the results above, our approach achieves optimal performance across all tested tasks, significantly outperforming all baselines. This is attributed to our method’s ability to identify the optimal parameter-sharing paradigm for all tasks. Moreover, we emphasize that for all tasks within the same environment, our method uses identical hyperparameters. Our approach does not rely on task-specific hyperparameters, which refer to algorithm hyperparameters related to the characteristics of the environment’s tasks, beyond common hyperparameters. For other baselines, typical task-specific hyperparameters include the reset interval, reset rate, and diversity loss coefficient for Kaleidoscope; the number of clusters and update interval for SePS and AdaPS; and the fusion/division threshold and quantization interval for MADPS.

Compared to these methods, HetDPS employs a distance-based clustering approach, eliminating the need for hyperparameters such as the number of clusters or fusion threshold. Furthermore, by fully accounting for the smooth evolution of the policy-sharing paradigm across two periods, and utilizing a quantization method which is independent of the amount of data per quantization (as CVAE supports continuous learning), HetDPS is insensitive to the quantization interval hyperparameter. We test the training performance of the algorithm under different quantization intervals in all tasks of multi-agent spreading, with results shown in Table 2. The results indicate that even when the quantization interval varies from 20 to 2000, the training performance remains unaffected.

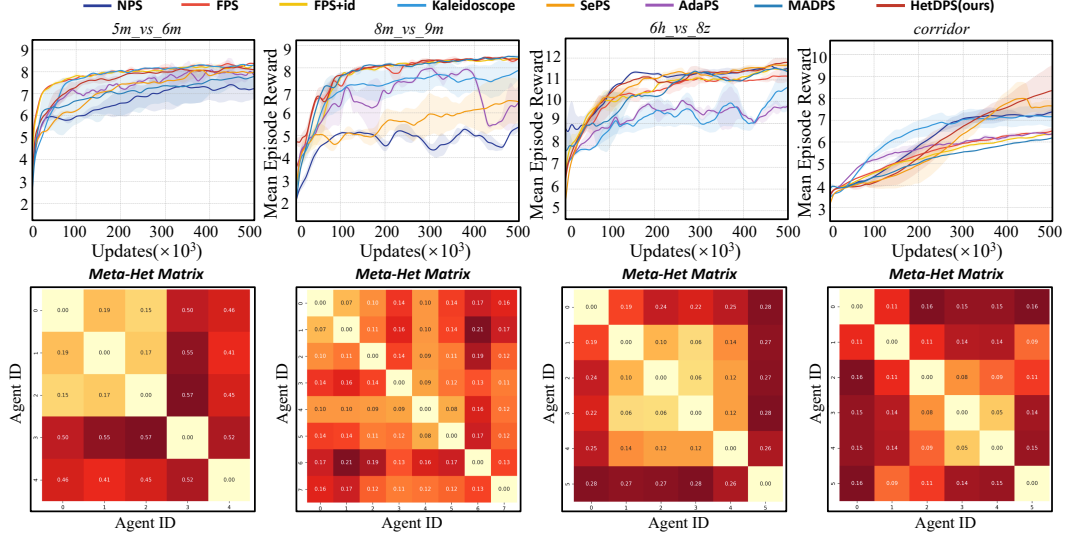


Figure 3: Results on StarCraft Multi-Agent Challenge (Part II).

In summary, HetDPS not only offers greater interpretability but also does not rely on task-specific hyperparameters, enabling easy deployment across various tasks.

Table 2: Results of varying quantization intervals in the Multi-agent Spreading environment, showing the average rewards of agents.

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

References

- Chris Bennett. *Heterogeneity in multi-agent systems*. PhD thesis, University of Bristol, 2024.
- Matteo Bettini, Ajay Shankar, and Amanda Prorok. Heterogeneous multi-robot reinforcement learning. In *AAMAS*, 2023a.
- Matteo Bettini, Ajay Shankar, and Amanda Prorok. System neural diversity: Measuring behavioral heterogeneity in multi-agent learning. *arXiv preprint arXiv:2305.02128*, 2023b.
- Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*, pages 1989–1998. PMLR, 2021.
- Aditya Grover, Maruan Al-Shedivat, Jayesh Gupta, Yuri Burda, and Harrison Edwards. Learning policy representations in multiagent systems. In *International conference on machine learning*, pages 1802–1811. PMLR, 2018.
- Xudong Guo, Daming Shi, Junjie Yu, and Wenhui Fan. Heterogeneous multi-agent reinforcement learning for zero-shot scalable collaboration. *CoRR*, 2024.
- Siyi Hu, Chuanlong Xie, Xiaodan Liang, and Xiaojun Chang. Policy diagnosis via measuring role diversity in cooperative multi-agent rl. In *International Conference on Machine Learning*, pages 9041–9071. PMLR, 2022.
- Tianyi Hu, Xiaolin Ai, Zhiqiang Pu, Tenghai Qiu, and Jianqiang Yi. Heterogeneous observation aggregation network for multi-agent reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2024a.
- Tianyi Hu, Zhiqiang Pu, Xiaolin Ai, Tenghai Qiu, and Jianqiang Yi. Measuring policy distance for multi-agent reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 834–842, 2024b.
- Tianyi Hu, Zhiqiang Pu, Xiaolin Ai, Tenghai Qiu, Yanyan Liang, and Jianqiang Yi. Hybrid actor-critic for physically heterogeneous multi-agent reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.
- Haobin Jiang, Yifan Yu, and Zongqing Lu. Metric policy representations for opponent modeling. *arXiv preprint arXiv:2106.05802*, 2021.
- Kun Jiang, Wenzhang Liu, Yuanda Wang, Lu Dong, and Changyin Sun. Credit assignment in heterogeneous multi-agent reinforcement learning for fully cooperative tasks. *Applied Intelligence*, 53(23):29205–29222, 2023.
- WOOJUN KIM and Youngchul Sung. 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, 2023.
- Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3991–4002, 2021.
- Dapeng Li, Na Lou, Bin Zhang, Zhiwei Xu, and Guoliang Fan. Adaptive parameter sharing for multi-agent reinforcement learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6035–6039. IEEE, 2024a.
- Xinran Li, Ling Pan, and Jun Zhang. Kaleidoscope: Learnable masks for heterogeneous multi-agent reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.
- Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, Zhipeng Hu, and Yaodong Yang. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. *Advances in Neural Information Processing Systems*, 34:941–952, 2021.

- Yuntao Liu, Yuan Li, Xinhai Xu, Donghong Liu, and Yong Dou. Rogc: Role-oriented graph convolution based multi-agent reinforcement learning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.
- Rodrigo Fonseca Marques and Zenilton Kleber Gonçalves do Patrocínio Júnior. Impact of heterogeneity on multi-agent reinforcement learning. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1048–1062. SBC, 2023.
- Muhammad A Masood and Finale Doshi-Velez. Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies. *arXiv preprint arXiv:1906.00088*, 2019.
- Kevin R McKee, Joel Z Leibo, Charlie Beattie, and Richard Everett. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(1):21, 2022.
- Douglas Meneghetti and Reinaldo Bianchi. Towards heterogeneous multi-agent reinforcement learning with graph neural networks. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2020)*, 2020.
- Dung Nguyen, Phuoc Nguyen, Svetha Venkatesh, and Truyen Tran. Learning to transfer role assignment across team sizes. *International Foundation for Autonomous Agents and Multiagent Systems*, 2022.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Esmail Seraj, Zheyuan Wang, Rohan Paleja, Matthew Sklar, Anirudh Patel, and Matthew Gombolay. Heterogeneous graph attention networks for learning diverse communication. *arXiv preprint arXiv:2108.09568*, 2021.
- Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. Roma: Multi-agent reinforcement learning with emergent roles. *International Conference on Machine Learning*, pages 9876–9886, 2020.
- Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. Rode: Learning roles to decompose multi-agent tasks. *International Conference on Representation Learning*, 2021.
- Chao Yu, Akash Velu, Eugene Vinitzky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.
- Xiaoyang Yu, Youfang Lin, Xiangsen Wang, Sheng Han, and Kai Lv. Ghq: grouped hybrid q-learning for cooperative heterogeneous multi-agent reinforcement learning. *Complex & Intelligent Systems*, 10(4):5261–5280, 2024.
- Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.