GSLHD: Global Skill Learning with Heterogeneous Decoding for Multi-Agent Reinforcement Learning

Tenghai Qiu[12⋆] Ruoyang Gao[4⋆] Jinyuan Feng[23†] Zhiqiang Pu[23] Yuqian Zhao[1] Biao Luo[1]
[1]School of Automation, Central South University
[2]National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]College of Civil Engineering, Tongji University

***Abstract*—In heterogeneous multi-agent reinforcement learning, each type of agent can naturally emerge as different roles, meaning that diversity of agents is conducive to cooperation within the team. However, adopting a fully heterogeneous policy, where each agent learns an independent policy without any shared components, suffers from sample inefficiency. To achieve a better trade-off of diversity and efficiency, we propose a novel framework named Global Skill Learning with Heterogeneous Decoding (GSLHD), which fosters sample efficiency by collaboratively sharing generalizable skill semantics and promotes diversity by heterogeneously decoding according to agent type. Specifically, we introduce a concept of skill, which is extracted from the Global Skill Assignment (GSA) module, to adaptively encode skill based on the inference of global states. Then, generalizable skill semantics are learned by an attention mechanism and shared across types. Finally, a heterogeneous decoder promotes policy diversity by decoding cross-type shared skills into specific actions tailored for each agent type. Extensive experiments are conducted on SMAC and SMACv2 to verify the remarkable performance of GSLHD.**

***Index Terms*—Multi-agent reinforcement learning, Heterogeneity, Skill learning**

I. Introduction

Recently, Multi-Agent reinforcement learning (MARL) has achieved remarkable success across a variety of challenging domains - from AlphaStar in StarCraft II [1], to real-world applications such as cooperative search task in unknown environment [2] or optimizing metropolitan traffic signals in real time [3]. These breakthroughs underscore MARL's potential for scalable coordination, autonomous decision-making, and adaptive control. Yet existing studies assume homogeneous agents that share identical observation and action spaces, an assumption rarely satisfied in real-world multi-agent systems (MAS). This translates to an urgent demand for heterogeneous policies that simultaneously exhibit diverse and flexible behaviours. [4]–[7]

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62503472 and Grant 62322316, the Young Scientists Foundation of CSAA (Guidance Navigation and Control, GNC) under Grant CSAA-YSF2025-GNC-08, and the CAS Project for Young Scientists in Basic Research (No. YSBR-123).

⋆ Equal contributions.

† Corresponding Author.

The performance of MARL benefits from the diversity of agent behaviors, which is vital for complex tasks but difficult to achieve without losing the efficiency of parameter sharing. No Parameter Sharing (NoPS) [8] equips each agent with an independent network, enhancing behavioural diversity at the cost of severe inefficiency, while Full Parameter Sharing (FuPS) trains a single policy, improving data usage but failing to capture heterogeneity [9]–[11]. To overcome this limitation, recent research has explicitly focused on heterogeneous-agent reinforcement learning. Zhong et al. [12] propose approaches that directly model agent heterogeneity beyond parameter sharing, offering an alternative but still facing scalability challenges. To reconcile this trade-off, prior works also exploit role-based learning [13] or encode agent types into observations, yet suffer from gradient interference [14], [15]. We instead introduce a concept of skill, which is learned jointly across agents [16], [17]. A type-conditioned heterogeneous decoder then maps each skill into agent-specific actions, preserving heterogeneity without duplicating networks. Our GSLHD framework thus combines FuPS-level efficiency with NoPS-level diversity, offering a principled solution for heterogeneous MARL.

Motivated by the aforementioned discussions and the limitations observed in existing multi-agent reinforcement learning approaches, we propose a novel framework named Global Skill Learning with Heterogeneous Decoding (GSLHD). The central idea of GSLHD is to endow agents with the ability to acquire shared high-level skill semantics through a global assignment mechanism, while simultaneously enabling type-specific execution via a heterogeneous decoding strategy. In this way, the framework not only enhances the generalization and sample efficiency of skill learning across diverse agents, but also preserves the capacity for specialization that is essential in heterogeneous multi-agent systems. The main contributions of our method can be summarized as follows:

1). **Global State Inference and Skill Allocation** We employ an assignment module that infers the global observation and then allocates a skill for each agent.
2). **Generalizable Skill Semantics Learning** The learned Skill Embeddings are seamlessly injected into the Query

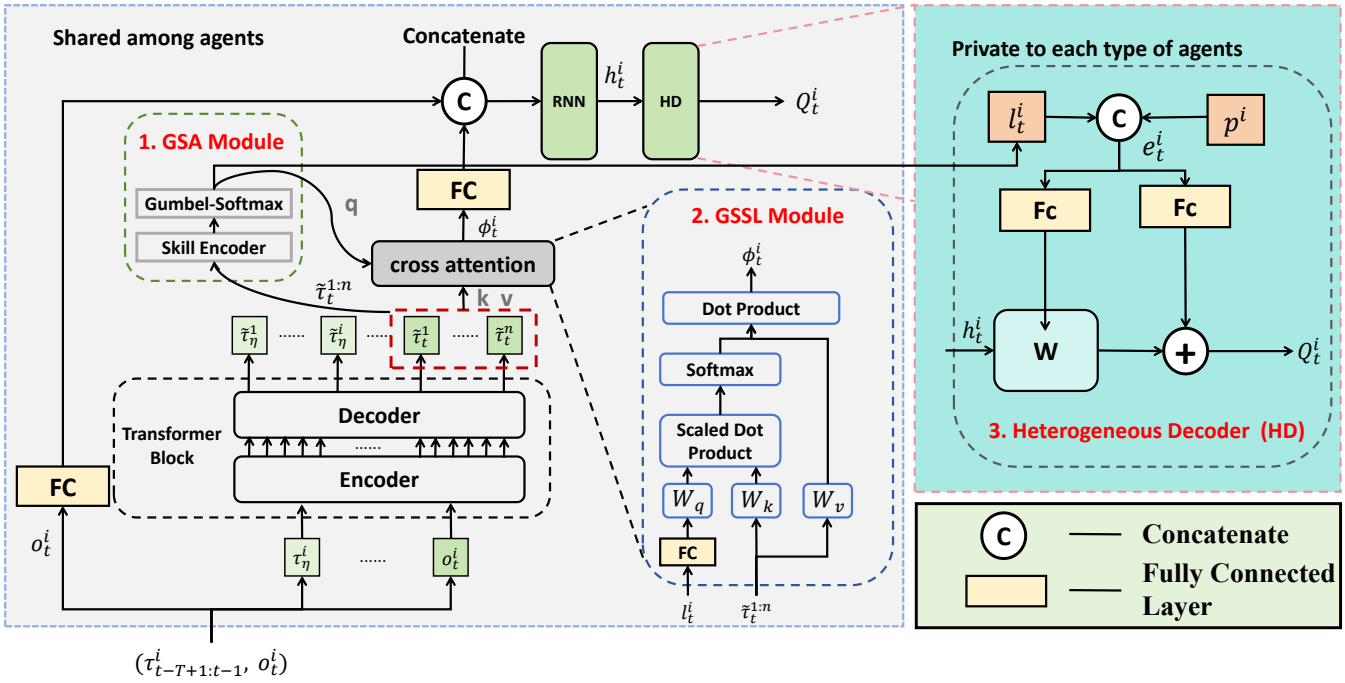

Fig. 1: The overall structure of the GSLHD. The GSLHD is composed of three modules: 1. Global Skill Assignment (GSA) module; 2. Generalizable Skill Semantics Learning (GSSL) module and 3. Heterogeneous Decoder (HD) module. The GSLHD adopts the centralized training and decentralized execution paradigm. During training, with a centralized critic, agent i can access global states. During decentralized execution, each agent executes actions based solely on its local observations.

and Key interaction of a multi-head attention mechanism, which leverages the inductive bias that skill semantics refers to the effects of other agents.

3). **Heterogeneous Decoder** The heterogeneous decoder takes the concatenation of skill and agent type as input and dynamically generates the parameters of each agent's policy network, thus promoting policy diversity.

II. Related Work

A. Parameter Sharing and Heterogeneity

The centralized training with decentralized execution (CTDE) paradigm [8], [18], [19] has enabled remarkable progress in multi-agent reinforcement learning (MARL), giving rise to numerous algorithms that exploit parameter sharing to improve sample efficiency and scalability. However, full parameter sharing (FuPS) often leads to homogeneous behaviors [9]–[11], limiting behavioral diversity among agents. To alleviate this issue, variants such as SePS [10] group agents by type and share parameters within groups. Recent studies have further examined the role of policy diversity as an indicator of cooperative performance [20], proposing diagnostic metrics for evaluating the degree of agent specialization. Other works improve individuality through decomposition or auxiliary learning. For example, Li et al. [4] introduces mutual information regularization between agent identities and trajectories to promote behavioral diversity. Similarly, Liu et al. [21] employs contrastive objectives to enhance agent distinguishability within value decomposition frameworks. Despite these advances, most approaches emphasize diversity at the cost of coordination efficiency or rely on fixed grouping schemes, whereas our framework aims to reconcile both via global skill learning and type-conditioned decoding that preserve heterogeneity without sacrificing scalability.

B. Role and Skill Learning

A parallel line of research introduces roles or skills to encode structured heterogeneity within cooperative MARL. Early role-based methods such as ROMA [14]and RODE [13]condition individual policies on latent role embeddings to reduce learning complexity. Later works such as [22] extend this idea to hierarchical action-space clustering., while LDSA [23] formulates dynamic subtask assignment to improve adaptability. Transformer-based extensions such as ROMAT [24] and attention-guided role models [25] leverage sequence modeling and attention to infer generalizable cooperation patterns, resonating with recent views that "MARL is a sequence modeling problem" [26]. Beyond discrete roles, policy transfer approaches like TRMARL [27]highlight task-relationship modeling to facilitate knowledge reuse across agents, reflecting the growing interest in transferable skill semantics. Compared with these studies, our GSLHD framework integrates global skill assignment with heterogeneous decoding, enabling both generalizable skill sharing and type-specific specialization under the CTDE paradigm.

III. METHODS

A. Problem Formulation

We consider a partially observable multi-agent MDP (PO-MAMDP) $\langle \mathcal{N}, \mathcal{P}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \{\mathcal{O}^i\}_{i\in\mathcal{N}}, T, R, \Omega, \gamma\rangle$. Here $\mathcal{N} = \{1, \ldots, n\}$ indexes agents and $\mathcal{P}$ is the set of unit *types* (e.g., Marine, Marauder, Medivac). At time t, the environment is in state $s_t \in \mathcal{S}$, each agent i receives a private observation $o_t^i \sim \Omega(\cdot \mid s_t, i)$ from its observation space $\mathcal{O}^i$, chooses an action $a_t^i \in \mathcal{A}^i$, and the joint action $\mathbf{a}_t = (a_t^1, \ldots, a_t^n)$ drives the transition $s_{t+1} \sim T(\cdot \mid s_t, \mathbf{a}_t)$ and shared reward $r_t = R(s_t, \mathbf{a}_t)$. Each agent maintains a local history (trajectory prefix) $\tau_t^i = (o_{1:t}^i, a_{1:t-1}^i)$; we write a truncated window $(\tau_{t-T+1:t-1}^i, o_t^i)$ when using a fixed length T. Under the centralized-training–decentralized-execution (CTDE) paradigm, a centralized critic may access global state s_t for training targets, while *actors* at test time must condition only on local information. In our setting, agent i holds a type label $p^i \in \mathcal{P}$ and learns an action-value function Q_t^i that estimates the expected return of its actions given its internal representation. Our operational goal is standard: to maximize the team return $J(\pi) = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t R(s_t, \mathbf{a}_t)\big]$ under a joint policy $\pi = (\pi^1, \ldots, \pi^n)$, with value-based learning providing temporal-difference (TD) targets during training. To mitigate partial observability without violating CTDE, each actor forms a belief-like global *context estimate* $\tilde{\tau}_t^{1:n}$ from its own history via MA^{2}E [28]: $\tilde{\tau}_t^{1:n} = \mathrm{MA}^2\mathrm{E}(\tau_{t-T+1:t-1}^i, o_t^i)$. This quantity is used only as an internal representation; no inter-agent communication or centralized signals are required at execution. Fig. 1 summarizes this setting and the information flow under CTDE.

B. Method Pipeline

GSLHD maps each agent's local history to action-values under CTDE. At each step, agent i applies MA^{2}E to $(\tau_{t-T+1:t-1}^i, o_t^i)$ to form an inferred global context $\tilde{\tau}_t^{1:n}$, while a recurrent backbone maintains temporal consistency. From this context, GSA samples a discrete skill l_t^i (Gumbel–Softmax) as a shared intent code. GSSL then uses the skill as a query over the inferred context to produce a semantic vector ϕ_t^i, which is fused with a local stream and fed to the recurrent state. Finally, the Heterogeneous Decoder concatenates l_t^i with the type p^i and, via a small hypernetwork, generates a linear head mapping the hidden state to Q_t^i, preserving type-specific execution without duplicating policies. Training uses the same value-based objective as our baselines; at test time, each agent independently computes $\tilde{\tau}_t^{1:n}$, selects a skill, derives ϕ_t^i, and acts from Q_t^i.

C. Global Skill Assignment

Global Skill Assignment (GSA) module fetches the inherent skill latent variable based on inference of global observation by MA^{2}E [28], flattens and feeds it into the GSA module to generate the latent skill variable as is shown in the left part of Fig. 1 Specifically, for agent i, MA^{2}E firstly infers the observations of other agents $\tilde{\tau}_t^{-i}, -i = 0, ..., i-1, i+1, ..., n-1$ and concatenate it to the local observation $\boldsymbol{o}_t^i$ to form the inference on global observation which is denoted as $\tilde{\tau}_t^{1:n}$. The next step is to generate a latent skill variable by the Gumbel-Softmax operator, a powerful differentiable sampling technique enabling discrete skill selection, allowing our process of skill assignment trainable by the way of gradient backpropagation.

$$l_i^t \sim \text{Gumbel-Softmax}(f_\phi(\tilde{\tau}_t^{1:n})) \tag{1}$$

where f_ϕ is a skill encoder in GSA, and l_i^t is the latent skill one-hot variable generated by the GSA module, serving as a compact yet expressive representation that will be thereafter leveraged by the attention-based component of our method to enhance coordination and decision-making efficiency.

D. Generalizable Skill Semantics Learning

Attention mechanism behaves as a cornerstone in the current field of AI that exhibits superiority when dealing with extracting dependencies between sequences. Building on this foundation, researchers have proposed a cross-attention-based approach to enable adaptive subtask semantics learning [16]. Yet these approaches still rely on local observation, our methods' general skill semantics learning mechanism leverages embedding skill variable as query, which is combined with inference on global observation as keys' input and values' input to promote team coordination while maintaining generalizability. Initially, we utilize the generated skill variable to produce the query

$$Q_i^t = W_Q l_i^t \tag{2}$$

where $W_Q \in \mathbb{R}^{n_s \times (nd)}$ is a learnable parameter and n_s and d are hyperparameters denoting the total number of skills and embedding dimension of the query, respectively. To align the key with the query, we map l_i^t into the d-dimensional query vector Q_i^t, thereby enabling precise semantic matching between the latent skill space and the query space.

$$K_i^t = W_K \tilde{\tau}_t^{1:n}, V_i^t = V_Q \tilde{\tau}_t^{1:n} \tag{3}$$

Then, the adaptively general skill semantics are easily obtained by the following equation

$$\boldsymbol{\phi_t^i} = softmax(\frac{Q_i^t {K_i^t}^T}{\sqrt{d}}) V_i^t \tag{4}$$

where d is the feature dimension of K_i^t. The garnered skill semantics, enhanced through a carefully designed MLP for semantic refinement, are subsequently aggregated with the output of another branch that takes o_t^i as input, achieving a balanced fusion of complementary information from distinct representation streams. The concatenation of the above-mentioned was then passed through a RNN neural network, thus generating the input of the heterogeneous decoder, a critical step that empowers the decoder to generate type- and skill-specific actions with extensibility.

E. Heterogeneous Decoder (HD)

To achieve heterogeneity while conserving the flexibility of inputs, the heterogeneous decoder employs a hypernetwork that adapts each agent's low-level actions to both its skill intention and physical embodiment. HD is implemented to dynamically generate the parameters of a type-specific linear projection, thereby converting the recurrent backbone's hidden state into action logits. In practice, we design the decoder to accommodate heterogeneous input e_t^i containing the agent's unit type p^i and the learned latent skill variable l_t^i. This is sent into a hypernetwork, which outputs the weights and biases of a linear mapping tailored to the specific combination of skill and embodiment. In detail, the weights W and bias b are generated by the following equation:

$$W = f(e_t^i; \theta^{(1)}) \tag{5}$$

$$b = f(e_t^i; \theta^{(2)}) \tag{6}$$

Here, $\theta^{(1)}$ and $\theta^{(2)}$ correspond to two neural networks' parameters. Lastly, the HD conducts a batch matrix multiplication

$$Q_t^i = BMM(h_t^i, W) + b \tag{7}$$

to gain the final action value Q_t^i. To sum up, this design ensures that, even when agents share a common high-level skill representation, the final motor commands are specialized to their unit types. Consequently, the framework enables fine-grained behavioral diversity without the need to train an entirely separate policy for each agent type.

IV. Experiments

A. Experimental Settings

We evaluate our method GSLHD on the StarCraft Multi-agent Challenge (SMAC) [29] and SMACv2 [30] to prove its effectiveness. SMAC is a benchmark based on StarCraft II micromanagement tasks, designed to test team coordination among multiple cooperative agents, while SMACv2 introduces more diverse unit compositions (reflect, surround) to better reflect real-world complexity. For both benchmarks, we adopt the centralized training with decentralized execution (CTDE) paradigm and all models are trained for 2 million timesteps using the PyMARL framework [31].

- 5m_vs_6m:A benchmark SMAC task where 5 Marines must defeat 6 opponents, highlighting coordination under numerical disadvantage and requiring precise micromanagement.
- MMM2:A mixed-unit scenario with Marines, Marauders, and Medivacs, testing heterogeneous team coordination, including ranged attacks, healing, and strategic positioning.
- 10gen_terran:A large-scale scenario with 10 diverse Terran units (e.g., Marines, Marauders, Medivacs, Siege Tanks), emphasizing combined-arms tactics, formation control, and synergy across roles.
- 10gen_zerg:A large-scale scenario with 10 varied Zerg units (e.g., Zerglings, Hydralisks, Banelings, Mutalisks), requiring swarm tactics, synchronized assaults, and mobility-based coordination.

We carried out a series of experiments to validate the effectiveness of our methods compared with other MARL algorithms. We selected three algorithms as baseline methods: MA^{2}E [28], VDN [32], OW_QMIX [33], RIIT [34]. MA^{2}E leverages masked auto-encoder techniques to address partial observability in multi-agent settings. VDN employs value decomposition to factorize the joint action-value function into a sum of individual value functions, enabling decentralized execution. OW_QMIX extends the QMIX framework with an optimistically-weighted mixing network, improving the approximation of the optimal joint action-value function.

B. Results and Discussion

As shown in Figure 3, across both homogeneous (e.g., MMM2, 5m_vs_6m) and heterogeneous (e.g., 10gen_terran, 10gen_zerg) MARL scenarios, the mean episode reward curves of GSLHD consistently converge to higher levels than those of other baselines, demonstrating its robust generalization capability. Notably, in heterogeneous settings, GSLHD significantly outperforms methods employing a homogeneous decoder, highlighting its advantage in handling diverse agent types. This performance gain can be attributed to the synergy between the Global Skill Assignment (GSA) module and the Heterogeneous Decoder (HD): GSA adaptively assigns skills from the global context, while HD maps them into type-specific actions, enabling more precise coordination and strategy adaptation. Such a design not only facilitates better information utilization but also ensures stable policy learning across tasks of varying interaction complexity.

To further examine the contribution of each component in our framework, we conduct ablation studies on both the 5m_vs_6m and MMM2 scenarios. The 5m_vs_6m task represents a homogeneous combat setting with only slight unit imbalance, whereas MMM2 is a heterogeneous mixed-arms battle involving Marines, Marauders, and Medivacs with healing–support dynamics. Fig. 4 presents the win rates, where the full GSLHD framework consistently achieves the highest performance across both scenarios, demonstrating the effectiveness of integrating global skill assignment, general skill semantics learning, and the heterogeneous decoder. In contrast, w/o GSA leads to a notable drop in performance, indicating that the absence of global skill assignment weakens the coordination among agents. w/o GSSL further reduces the win rate, showing that general skill semantics learning is crucial for capturing transferable knowledge and enhancing sample efficiency. Finally, w/o HD also suffers from degraded performance, confirming that the heterogeneous decoder is indispensable for adapting shared skills into agent-specific actions. These results collectively highlight that each component contributes significantly to the overall success of GSLHD: in the homogeneous 5m_vs_6m scenario, the gains primarily stem from improved cooperation, while in the heterogeneous MMM2 task, the benefits are more pronounced in enabling specialization across distinct unit types.

Fig. 2: Illustration of heterogeneous behaviors in the MMM2 task. *Left:* Medivac and one Marauder choose to escape while another Marauder and a Marine continue attacking, showing both inter-type and intra-type action divergence. *Right:* The same attack skill is executed differently across types: Marines sustain direct fire, whereas Medivacs perform continuous move-attack to maintain mobility.

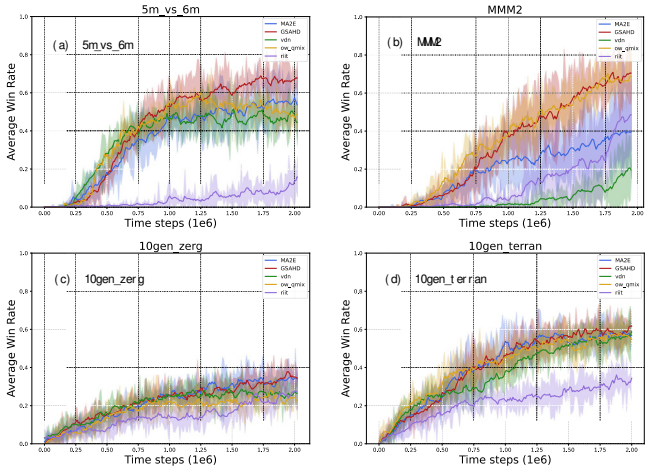

Fig. 3: The mean episode reward curves. (a) 5m_vs_6m, a homogeneous combat scenario with slight unit imbalance. (b) MMM2, a heterogeneous mixed-arms scenario with Marines–Marauders–Medivacs and support-healing dynamics under asymmetric compositions. (c) 10gen_zerg, a heterogeneous multi-agent task with various Zerg unit types. (d) 10gen_terran, a heterogeneous multi-agent task involving diverse Terran unit types.

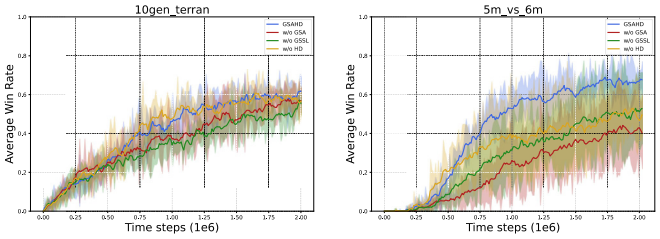

Fig. 4: Win rates on the 5m_vs_6m map. The comparison includes the full GSLHD framework and its three ablated variants: w/o GSA, w/o GSSL, and w/o HD, demonstrating the contribution of each component to the overall performance.

In Table I, on **MMM2** both IQL and IPPO remain near zero (0.1% and 0.0%), reflecting the difficulty of coordinating asymmetric unit types under fully independent learning. By contrast, GSLHD attains a much higher final win rate of 71.5% and is the only method that reaches the 80% threshold during training (at 1.64×10^6 steps), indicating substantially improved sample efficiency and stability in this heterogeneous setting.

On **1o_2r_vs_4r**, the steps-to-80% column shows that IQL crosses the threshold earliest (at 7×10^4 steps), while GSLHD reaches it later (at 1.1×10^5 steps). However, neither sustains that level by the end of training: IQL finishes at 41.9% and GSLHD at 46.5%. This suggests that early threshold crossing alone does not guarantee high final performance, and that our approach retains a higher terminal win rate even when late-training nonstationarity reduces absolute returns.

On **10gen_zerg**, no method reaches 80%, yet GSLHD still more than doubles IQL (35.3% vs. 16.1%) and vastly surpasses IPPO (1.6%), highlighting its advantage when many unit types and interactions increase coordination complexity. Overall, these results show that sharing skill semantics while decoding by type enables GSLHD to combine the efficiency of shared representation with the behavioral diversity needed for heterogeneous cooperation.

C. Skill Visualization

As illustrated in Fig. 2, the left sub-figure presents a snapshot from the MMM2 task where heterogeneous behaviors emerge both across and within unit types. Specifically, the

TABLE I: Comparison of final win rate (%) and convergence speed (steps to reach 80%) across three scenarios: **MMM2** (medium multi-agent movement), **10gen_zerg** (heterogeneous 10v10 combat), and **1o2r** (one-offensive-two-reconnaissance cooperation). Win rate is reported as the mean ± std over the last 20 evaluations averaged across 3 random seeds. Convergence speed is defined as the number of environment steps required to reach a smoothed win rate of 80%; lower is better.

Scenario	Steps	IQL		IPPO		Ours	
		Win rate (%)	Steps to 80%	Win rate (%)	Steps to 80%	Win rate (%)	Steps to 80%
MMM2	2M	0.1 ± 0.1	N/A	0.0 ± 0.0	N/A	**71.5 ± 9.0**	**1.64e6**
1o_2r_vs_4r	2M	41.9 ± 0.0	**7e4**	18.6 ± 0.0	N/A	**46.5 ± 1.9**	1.1e5
10gen_zerg	2M	16.1 ± 4.7	N/A	1.6 ± 0.4	N/A	35.3 ± 1.2	N/A

Medivac and one Marauder choose to retreat for survival, while another Marauder and a Marine continue to launch attacks. This divergence reflects the role of the Global Skill Assignment (GSA) module, which infers a shared skill representation (e.g., attack vs. escape) while still allowing agents of the same type to specialize under different contexts. In contrast, the right sub-figure shows differentiated executions of the same attack skill: Marines remain stationary and sustain fire, whereas Medivacs perform continuous move-attack operations to maximize mobility. Such type-specific realizations are enabled by the Heterogeneous Decoder (HD) that translates the shared skill semantics into distinct action patterns tailored to each unit type.

D. Overall Discussion

The experimental results across all benchmarks consistently verify the effectiveness and generality of the proposed GSLHD framework. From homogeneous to highly heterogeneous tasks, GSLHD maintains both stable convergence and high asymptotic performance, demonstrating its ability to reconcile coordination and specialization. The reward curves in Fig. 3 and the quantitative results in Table I reveal that the key advantage stems from the cooperative interaction between the *Global Skill Assignment* and *Heterogeneous Decoder*. GSA enables agents to infer globally consistent yet compact intent codes from partial observations, which mitigates the non-stationarity typical of MARL. At the same time, the HD ensures that these shared intentions are realized differently according to agent type, allowing specialization to emerge naturally without sacrificing training efficiency. This balance explains why GSLHD achieves both higher win rates and faster convergence across multiple environments.

Moreover, the ablation results in Fig. 4 further emphasize the complementary nature of each module: removing any of them leads to clear degradation, confirming that GSA, GSSL, and HD jointly contribute to coordination, generalization, and diversity. Visualizations in Fig. 2 also reveal interpretable behaviors—shared skills translate into distinct execution strategies depending on agent type—illustrating that GSLHD not only improves quantitative metrics but also leads to qualitatively meaningful cooperation patterns. In summary, the proposed framework establishes a scalable paradigm for heterogeneous MARL by combining shared skill semantics with type-conditioned decoding, bridging the gap between efficient centralized training and diverse decentralized execution.

V. Conclusion

We propose a novel framework, GSLHD, that addresses the long-standing tension between specialization and sample efficiency in heterogeneous multi-agent reinforcement learning. The method incorporates a global skill assignment framework equipped with a heterogeneous decoder that simultaneously distills general skill semantics from global observation through an attention-based extractor and transforms those skills into agent-specific behaviors via a hypernetwork conditioned on both the skill variable and unit type. By sharing skill semantics while maintaining a role-aware decoder, GSLHD enables effective knowledge transfer across agents yet preserves the diversity needed for fine-grained coordination. Extensive evaluations on SMAC and the more challenging SMACv2 benchmarks show that GSLHD consistently surpasses state-of-the-art baselines in win rate, convergence speed, and sample efficiency, confirming the advantages of its hybrid design.

References

[1] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. P. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, pp. 350 – 354, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204972004

[2] Y. Li, M. Ma, J. Cao, G. Luo, D. Wang, and W. Chen, "A method for multi-auv cooperative area search in unknown environment based on reinforcement learning," *Journal of Marine Science and Engineering*, vol. 12, no. 7, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/12/7/1194

[3] D. K. Kwesiga, A. Guin, and M. Hunter, "Adaptive traffic signal control based on multi-agent reinforcement learning. case study on a simulated real-world corridor," 2025. [Online]. Available: https://arxiv.org/abs/2503.02189

[4] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, "Celebrating diversity in shared multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 3991–4002. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/20aee3a5f4643755a79ee5f6a73050ac-Paper.pdf

[5] J. Feng, M. Chen, J. Li, J. Chen, Z. Pu, and M. Chen, "Knowledge-based and data-driven integrating design methodology for air combat strategy in multi-opponent adversarial game," *Acta Electronica Sinica*, vol. 52, no. 11, pp. 3809–3822, 2024.

[6] J. Feng, M. Chen, Z. Pu, T. Qiu, J. Yi, and J. Zhang, "Efficient multi-task reinforcement learning via task-specific action correction," *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

[7] J. Feng, M. Chen, Z. Pu, Y. Xu, and Y. Liang, "Ma2rl: Masked autoencoders for generalizable multi-agent reinforcement learning," *IEEE Transactions on Artificial Intelligence*, 2026.

[8] R. Lowe, Y. WU, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf

[9] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, "Dealing with Non-Stationarity in Multi-Agent Deep Reinforcement Learning," *arXiv e-prints*, p. arXiv:1906.04737, Jun. 2019.

[10] F. Christianos, G. Papoudakis, A. Rahman, and S. V. Albrecht, "Scaling multi-agent reinforcement learning with selective parameter sharing," 2021. [Online]. Available: https://arxiv.org/abs/2102.07475

[11] W. Kim and Y. Sung, "Parameter sharing with network pruning for scalable multi-agent deep reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2303.00912

[12] Y. Zhong, J. G. Kuba, X. Feng, S. Hu, J. Ji, and Y. Yang, "Heterogeneous-agent reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2304.09870

[13] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "{RODE}: Learning roles to decompose multi-agent tasks," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=TTUVg6vkNjK

[14] T. Wang, H. Dong, V. Lesser, and C. Zhang, "ROMA: Multi-agent reinforcement learning with emergent roles," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9876–9886. [Online]. Available: https://proceedings.mlr.press/v119/wang20f.html

[15] K. ab Abebe Tessera, A. Rahman, A. Storkey, and S. V. Albrecht, "Hypermarl: Adaptive hypernetworks for multi-agent rl," 2025. [Online]. Available: https://arxiv.org/abs/2412.04233

[16] Z. Tian, R. Chen, X. Hu, L. Li, R. Zhang, F. Wu, S. Peng, J. Guo, Z. Du, Q. Guo, and Y. Chen, "Decompose a task into generalizable subtasks in multi-agent reinforcement learning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=aky0dKv9ip

[17] J. Yang, I. Borovikov, and H. Zha, "Hierarchical cooperative multi-agent reinforcement learning with skill discovery," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2020, p. 1566–1574.

[18] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," 2016. [Online]. Available: https://arxiv.org/abs/1605.06676

[19] Y. Chen, J. Feng, W. Yang, M. Zhong, Z. Shi, R. Li, X. Wei, Y. Gao, Y. Wu, Y. Hu *et al.*, "Self-compression of chain-of-thought via multi-agent reinforcement learning," *arXiv preprint arXiv:2601.21919*, 2026.

[20] S. Hu, C. Xie, X. Liang, and X. Chang, "Policy diagnosis via measuring role diversity in cooperative multi-agent rl," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9041–9071.

[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] X. Zeng, H. Peng, and A. Li, "Effective and stable role-based multi-agent collaboration by structural information principles," 2023. [Online]. Available: https://arxiv.org/abs/2304.00755

[23] 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.

[24] D. Wang, F. Zhong, M. Li, M. Wen, Y. Peng, T. Li, and A. Yang, "Romat: Role-based multi-agent transformer for generalizable heterogeneous cooperation," *Neural Networks*, vol. 174, p. 106129, 2024.

[25] Z. Hu, Z. Zhang, H. Li, C. Chen, H. Ding, and Z. Wang, "Attention-guided contrastive role representations for multi-agent reinforcement learning," *arXiv preprint arXiv:2312.04819*, 2023.

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

[27] R. Qin, F. Chen, T. Wang, L. Yuan, X. Wu, Y. Kang, Z. Zhang, C. Zhang, and Y. Yu, "Multi-agent policy transfer via task relationship modeling," *Science China Information Sciences*, vol. 67, no. 8, p. 182101, 2024.

[28] S. Kang, Y. Lee, G. Kim, S. Chong, and S.-Y. Yun, "MA^{2}e: Addressing partial observability in multi-agent reinforcement learning with masked auto-encoder," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=klpdEThT8q

[29] M. Samvelyan, T. Rashid, C. S. de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C. Hung, P. H. S. Torr, J. N. Foerster, and S. Whiteson, "The starcraft multi-agent challenge," *CoRR*, vol. abs/1902.04043, 2019. [Online]. Available: http://arxiv.org/abs/1902.04043

[30] B. Ellis, J. Cook, S. Moalla, M. Samvelyan, M. Sun, A. Mahajan, J. N. Foerster, and S. Whiteson, "Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2212.07489

[31] M. Samvelyan, T. Rashid, C. S. de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C.-M. Hung, P. H. S. Torr, J. Foerster, and S. Whiteson, "The StarCraft Multi-Agent Challenge," *CoRR*, vol. abs/1902.04043, 2019.

[32] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning," 2017. [Online]. Available: https://arxiv.org/abs/1706.05296

[33] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," 2020. [Online]. Available: https://arxiv.org/abs/2006.10800

[34] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. N. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," *CoRR*, vol. abs/1803.11485, 2018. [Online]. Available: http://arxiv.org/abs/1803.11485