

Concept Learning for Cooperative Multi-Agent Reinforcement Learning

Zhonghan Ge

School of Management Engineering
Nanjing University
Nanjing, China
1476360984@qq.com

Yuanyang Zhu*

School of Information Management
Nanjing University
Nanjing, China
yuanyangzhu@nju.edu.cn

Chunlin Chen

School of Management Engineering
Nanjing University
Nanjing, China
clchen@nju.edu.cn

Abstract—Despite substantial progress in applying neural networks (NN) to multi-agent reinforcement learning (MARL) areas, they still largely suffer from a lack of transparency and interpretability. However, its implicit cooperative mechanism is not yet fully understood due to black-box networks. In this work, we study an interpretable value decomposition framework via concept bottleneck models, which promote trustworthiness by conditioning credit assignment on an intermediate level of human-like cooperation concepts. To address this problem, we propose a novel value-based method, named Concepts learning for Multi-agent Q-learning (CMQ), that goes beyond the current performance-vs-interpretability trade-off by learning interpretable cooperation concepts. CMQ represents each cooperation concept as a supervised vector, as opposed to existing models where the information flowing through their end-to-end mechanism is concept-agnostic. Intuitively, using individual action value conditioning on global state embeddings to represent each concept allows for extra cooperation representation capacity. Empirical evaluations on the StarCraft II micromanagement challenge and level-based foraging (LBF) show that CMQ achieves superior performance compared with the state-of-the-art counterparts. The results also demonstrate that CMQ provides more cooperation concept representation capturing meaningful cooperation modes, and supports test-time concept interventions for detecting potential biases of cooperation mode and identifying spurious artifacts that impact cooperation.

Index Terms—Multi-agent reinforcement learning, value decomposition, concept learning, interpretability

I. INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) has achieved notable success in complex domains such as autonomous driving [1], sensor networks [2], and robotics control [3]. A core driver behind this progress is the development of value decomposition methods [4], [5], [6], [7], [8], which effectively factorize the joint action-value function to enable decentralized execution with centralized training. However, these methods often operate as black boxes, providing limited visibility into individual agent behaviors or their contributions to team performance, hindering transparency, trust, and practical deployment in safety-critical domains.

Efforts to enhance interpretability in RL typically fall into post-hoc analysis or global approximation [9], [10]. Instance-based methods (e.g., Shapley values [11], clustering [12]) offer local explanations but suffer from instability and computational cost [13], [14]. Global imitation-based techniques [15], [16], [17] distill agent behaviors but often fail to retain fidelity in complex tasks. Recently, concept-based learning [18], [19], [20] has emerged as a promising paradigm, offering high-level, human-understandable representations of a model’s reasoning process. Concept bottleneck models, in particular, separate learning into two stages: predicting intermediate concepts, then making final decisions based on these concepts, enabling both interpretability and test-time human intervention [21].

Inspired by this, we explore whether concept learning can enhance interpretability and coordination in MARL. We propose CMQ (Concept learning for Multi-agent Q-learning), a novel value decomposition method that integrates rich cooperation concepts as intermediate representations. Each agent’s local action-value is combined with a global concept embedding to compute the joint action-value and agent-wise credit. The concept bottleneck not only improves interpretability but also retains expressiveness through concept-conditioned temporal Q-values, mitigating the linearity limitations in existing decomposition methods [6]. Furthermore, CMQ supports test-time concept interventions, allowing practitioners to diagnose and adjust specific cooperation failures.

Our contributions are summarized as follows: 1) We propose a novel value decomposition method, called Concept learning for Multi-agent Q-learning (CMQ), which moves a step toward modeling explicit collaboration among agents; 2) We provide the test-time concept intervention to diagnose which cooperation concepts are incorrect or do not align with human experts; 3) Through extensive experiments on challenging MARL benchmarks, CMQ not only consistently achieves superior performance compared to different state-of-the-art baselines but also allows for an easy-to-understand interpretation of credit assignment.

II. PRELIMINARIES AND RELATED WORK

Dec-POMDP. Cooperative multi-agent tasks are generally formulated as decentralized partially observable Markov decision processes (Dec-POMDPs), defined by the tuple $G =$

*Corresponding Author.

This work was supported in part by the China Postdoctoral Science Foundation under Grant Number 2025T180877, the National Key Research and Development Program of China under Grant 2023YFD2001000 and the Nanjing University Integrated Research Platform of the Ministry of Education-Top Talents Program.

$\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \Omega, O, r, \gamma \rangle$, where \mathcal{N} is the agent set, $s \in \mathcal{S}$ is the global state, γ is the discount factor, and $a_i \in \mathcal{A}$ is the action of agent i . At each timestep, agents take actions \mathbf{a} and receive a shared reward $r(s, \mathbf{a})$ with state transitions governed by \mathcal{P} . Under partial observability, each agent receives local observation o_i from $O(o_i|s, a_i)$ and forms a history τ_i . The goal is to learn a joint policy $\pi = \langle \pi_1, \dots, \pi_n \rangle$ that maximizes the expected cumulative return $R_t = \sum_{t=0}^{\infty} \gamma^t r_t$.

Centralized Training with Decentralized Execution.

CTDE is a prevalent paradigm in the MARL, where each agent learns its policy solely from its own action–observation history, while a centralized critic exploits the full joint state–action information to compute gradient updates. A natural extension of this framework is value decomposition, in which agents learn individual utility functions whose combination reconstructs the global action-value function. This decomposition not only facilitates efficient credit assignment, but—if properly constrained—also preserves optimality guarantees. In particular, any value decomposition scheme must satisfy the *Individual–Global–Max* (IGM) principle [7], which demands that the joint action maximizing the total value coincides with the tuple of per-agent maximizers:

$$\operatorname{argmax}_{\mathbf{a}} Q_{tot}(\boldsymbol{\tau}, \mathbf{a}) = \left(\begin{array}{c} \operatorname{argmax}_{a_1} Q_1(\tau_1, a_1) \\ \vdots \\ \operatorname{argmax}_{a_n} Q_n(\tau_n, a_n) \end{array} \right), \quad (1)$$

where \mathbf{a} is the joint action-observation history.

Concept Bottleneck Models. Concept Bottleneck Models (CBMs) structure the learning process into two stages: concept learning and decision-making. Given an input x , a CBM first maps x to an intermediate representation c , which encodes a set of predefined concepts. These concepts are chosen for their interpretability and relevance to the task. The model then uses c to make a final prediction y . Formally, a CBM can be represented as a composite function $y = g(f(x))$, where $f : \mathcal{X} \rightarrow \mathcal{C}$ is the concept learning function and $g : \mathcal{C} \rightarrow \mathcal{Y}$ is the decision-making function. The overall objective in training a CBM is to optimize both f and g to improve the interpretability of the model while maintaining or even enhancing its predictive performance.

Related work. Effective credit assignment is at the core of value-decomposition methods for cooperative MARL, as it allocates the global action-value among agents to guide decentralized policy updates [4], [5]. VDN [4] achieves this via a simple linear sum of individual Q-values, while QMIX [5] enforces a monotonic mixing network to guarantee consistency with the IGM principle. Building on this foundation, QTRAN [7] relaxes the monotonicity constraint through specialized transformation objectives, QPLEX [6] introduces a dueling-style non-monotonic mixer, and both WQMIX [22] and Qatten [23] enrich the mixer with distributional and attention-based mechanisms for finer credit estimation. Despite these advances, the mixing networks in such approaches remain largely opaque and ignore the underlying semantic factors that drive effective coordination.

Alternatives, QPD [8] computes credit weights directly from each agent’s action–observation history, but their black-box mixers still lack transparency. In contrast, we propose to embed concept learning into the decomposition process, thereby structuring the joint value function around human-interpretable semantic units. This not only preserves strong cooperative performance but also yields transparent credit assignments that align with interpretable concepts, closing the gap between high-performance and interpretability.

III. CONCEPT LEARNING FOR COOPERATION MARL

To overcome the limitations of existing value decomposition methods, we introduce Concept-based Multi-agent Q-learning (CMQ)—a novel framework that integrates concept bottleneck learning into value decomposition, aiming to make multi-agent cooperation both effective and interpretable. CMQ factorizes the global action-value function as a weighted sum over concept-conditioned temporal Q-values:

$$Q_{tot}^{\phi}(\boldsymbol{\tau}, \mathbf{a}) = \sum_k \alpha_k \hat{Q}_k(\boldsymbol{\tau}, \mathbf{a}) + f(s), \quad (2)$$

where $\hat{Q}_k(\boldsymbol{\tau}, \mathbf{a})$ denotes the concept-specific temporal Q-value generated via agent activation probabilities p_i and individual value components, while α_k is a state-dependent weight reflecting global semantics. This decomposition not only enhances transparency through interpretable cooperation concepts but also retains policy learning effectiveness. Although linear decomposition may be limited in expressiveness [6], we show that CMQ maintains competitive performance without sacrificing interpretability.

As illustrated in Figure 1, CMQ consists of local agent value functions $Q_i(\tau_i, u_i)$, which are aggregated through a concept-driven mixing network. Local Q-values $[Q_i]_{i=1}^n$ are mapped into cooperation concept bottlenecks \hat{Q}_i , and global state embeddings are used to compute corresponding credit assignments. The final joint Q-value is computed from these concept-conditioned values and their associated credits. Importantly, CMQ introduces explicit cooperation semantics, enabling test-time concept interventions to simulate how specific modes of coordination impact performance.

In the following, we elaborate on the structure of the mixing network, how it integrates dual semantic embeddings per concept, and how CMQ enables meaningful concept-level interventions, making it responsive and reliable in real-world cooperative scenarios.

A. Architecture of CMQ

CMQ comprises a concept predictor and a joint value function predictor. The concept predictor generates a mixture of two global state semantics for each cooperation concept to explicitly predict the probability of the concept’s activity. The joint value function predictor estimates the joint value from the predicted concept embeddings together with the global state s and the cooperation concept bottleneck \hat{Q}_i . This architecture endows our model with the capability to assemble arguments both supporting and contesting the activation of a

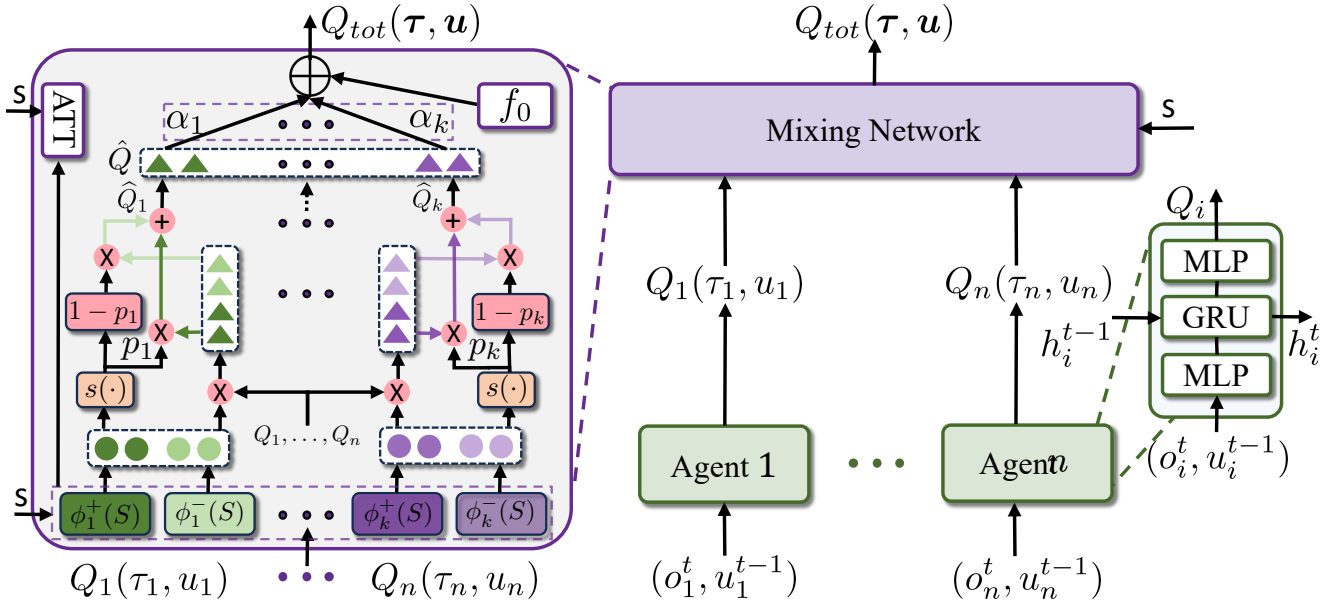


Fig. 1: The framework of our method CMQ. First, each agent models a value function $Q_i(\tau_i, a_i)$ conditioned on its individual action-observation history τ_i . To model the activation of these concepts, we parameterize each one with a pair of global-state-dependent embeddings: \hat{s}_i^+ when the concept is active, and \hat{s}_i^- when inactive. These embeddings are aligned with agent-level Q-values via dot-product interactions, serving as semantic interpreters of cooperation patterns. A learnable scoring function $s(\cdot)$ estimates the activation probability \hat{p}_i for each concept based on its dual embeddings. These probabilities are used to interpolate between the active and inactive representations, yielding a fused embedding for each concept: $\tilde{Q}_i = \hat{p}_i \cdot \tilde{Q}_i^+ + (1 - \hat{p}_i) \cdot \tilde{Q}_i^-$. This formulation enables concept-aware credit assignment while preserving semantic interpretability and enforcing structural disentanglement between latent cooperation modes.

concept, thereby facilitating straightforward interventions in cooperation concepts. At the point of intervention, one can deftly alternate between the global state semantics, thereby tailoring concept activation dynamics.

In practice, we use the concept predictor to produce cooperation concept $c_i(s)$ with two global state semantics $\hat{c}_i^+(s), \hat{c}_i^-(s) \in \mathbb{S}^m$. We define the specific semantics: $\hat{c}_i^+(s)$ and $\hat{c}_i^-(s)$, which represent the existence and non-existence of concept $c_i(s)$, respectively. To achieve this, the global state s is fed into two concept-specific fully connected layers, yielding the following embeddings $\hat{c}_i^+(s) = \phi_i^+(s) = \text{Relu}(W_i^+ s + \mathbf{b}_i^+)$ and $\hat{c}_i^-(s) = \phi_i^-(s) = \text{Relu}(W_i^- s + \mathbf{b}_i^-)$, where W_i^+, W_i^- and $\mathbf{b}_i^+, \mathbf{b}_i^-$ are learnable parameters. These dual embeddings provide a semantic grounding that distinguishes whether the corresponding concept is active or inactive under a given global state. To align the embeddings with the underlying ground-truth semantics, we employ a learnable and differentiable scoring function $s: \mathbb{S}^{2m} \rightarrow [0, 1]$, which is adeptly crafted to estimate the probability $\hat{p}_i \triangleq s([\hat{c}_i^+(s), \hat{c}_i^-(s)]^T) = \sigma(W_s [\hat{c}_i^+(s), \hat{c}_i^-(s)]^T + \mathbf{b}_s)$ of concept $c_i(s)$ is in existence, based on the conjoint embedding space, where σ is a sigmoid function. To promote parameter efficiency and generalization, the scoring function parameters W_s and \mathbf{b}_s are shared across all concepts $c_i(s)$, enabling compact modeling of semantic decision boundaries across multiple concept dimensions.

To relieve the limited expressivity of linear decomposition for the joint value function, we factor the centralized critic as a weighted summation of individual critics across agents. We

first build the temporal Q-values \tilde{Q}_i that comprise the positive $\tilde{Q}_i^+ = W_i^+(s)[Q_1(\tau, \mu), \dots, Q_n(\tau, \mu)]$ and negative $\tilde{Q}_i^- = W_i^-(s)[Q_1(\tau, \mu), \dots, Q_n(\tau, \mu)]$. Given the concept activation probability \hat{p}_i , we compute the final concept-level Q-value via a convex combination of the positive and negative projections:

$$\hat{Q}_i \triangleq (\hat{p}_i \tilde{Q}_i^+ + (1 - \hat{p}_i) \tilde{Q}_i^-). \quad (3)$$

Intuitively, this serves a two-fold purpose: (i) it enforces selective dependence on \hat{s}_i^+ when the concept i is active (i.e., $\hat{p}_i \approx 1$), and on \tilde{Q}_i^- otherwise, thereby inducing two semantically meaningful latent spaces, and (ii) it enables a clear intervention strategy where one switches the embedding states when mispredicted concepts can be corrected by swapping their embedding states. All k mixed concept Q-values $\{\hat{Q}_k\}_{k=1}^k$ are then concatenated to form a cooperation bottleneck. To realize the full joint value function Q_{tot} in Eq. 2, we compose it as a weighted sum over the concept-conditioned Q-values:

$$Q_{tot}^\phi(\tau, \mathbf{a}) = \sum_k \alpha_k \hat{Q}_k(\tau, a_i) + f(s), \quad (4)$$

where $f_0(s)$ is a bias term, $\hat{Q}_k(\tau, a_i)$ is the cooperation bottleneck and α_i is credits. The credit α_k is computed using a dot-product attention mechanism between the concept embedding $\hat{c}_i(s)$ and the global state s :

$$\alpha_k = \frac{\exp\left((\mathbf{w}_i \hat{c}_i(s))^\top \text{ReLU}(\mathbf{w}_s s)\right)}{\sum_{k=1}^m \exp\left((\mathbf{w}_i \hat{c}_i(s))^\top \text{ReLU}(\mathbf{w}_s s)\right)}, \quad (5)$$

where w_i and s are the learnable parameters, and ReLU serves as the activation function to preserve non-negativity. To ensure the monotonicity constraint of Eq. (1), the credits α_k are enforced to be non-negative, e.g., via absolute-value operations if necessary. Finally, similar to vanilla CBMs, our framework CMQ provides interpretable reasoning via its concept probability vector $\hat{\mathbf{p}}(\mathbf{x}) \triangleq [\hat{p}_1, \dots, \hat{p}_k]$, which explicitly indicates the predicted activation of each cooperation concept involved in shaping Q_{tot} .

Under the CTDE paradigm, CMQ is optimized by minimizing the standard squared temporal-difference (TD) error. Specifically, during training, a batch of transitions is sampled from a replay buffer b , and the loss is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\tau, \mathbf{a}, r, \tau') \in b} \left[(y - Q_{tot}(\tau, \mathbf{a}; \theta))^2 \right], \quad (6)$$

where θ denotes the current network parameters, and τ represent the joint action-observation trajectory. The TD target value is computed as $y = r + \gamma \max_{\mathbf{a}'} Q_{tot}(\tau', \mathbf{a}'; \theta^-)$, where θ^- denotes the parameters of a target network that are periodically updated from θ and held fixed for a number of training iterations to stabilize learning. It allows CMQ to learn semantically grounded joint value estimates by integrating concept-based representations with standard RL optimization.

B. Intervening with Cooperation Concept

In our CMQ framework, concept intervention is straightforward: it entails substituting the predicted cooperation concept with the one that aligns semantically with the ground truth. For example, at time step t within the context of a given global state s , individual action values and concept $\hat{c}_i(s)$, CMQ predicted $\hat{p}_i = 0.1$ while a human expert knows that concept $\hat{c}_i(s)$ is activate ($\hat{c}_i(s) = 1$), an intervention can be enacted to adjust $\hat{p}_i = 1$. This revision recalibrates the cooperation concept representation within the CMQ bottleneck from a weighted mixture ($0.1\hat{c}_i^+(s) + 0.9\hat{c}_i^-(s)$) to the pure concept \hat{c}_i^+ . Such refinement directly informs the joint value function predictor with the accurate concept context.

Furthermore, a regularization technique is employed during training to acclimate the CMQ model to concept interventions, enhancing its responsiveness during deployment. This involves executing random concept interventions with a probability \tilde{p} , which can be set to $\tilde{p}_i := c_i(s)$ for the concept $\hat{c}_i(s)$ with probability \tilde{p} . The computation of the concept $\hat{c}_i(s)$ during training process is as follows:

$$\hat{c}_i(s) = \begin{cases} c_i(s)\hat{c}_i^+(s) + (1 - c_i(s))\hat{c}_i^-(s), & \text{with } \tilde{p} \\ \hat{p}_i\hat{c}_i^+(s) + (1 - \hat{p}_i)\hat{c}_i^-(s), & \text{with } 1 - \tilde{p} \end{cases} \quad (7)$$

while the predicted probability mixture is systematically employed at inference time. During the backpropagation phase, the updating process is selective: only the relevant concept embedding (i.e., $\hat{c}_i^+(s)$ if $\hat{c}_i(s) = 1$) is updated based on task-specific feedback, whereas concept prediction feedback impacts both embeddings $\hat{c}_i^+(s)$ and $\hat{c}_i^-(s)$. Conceptually, this approach encapsulates the learning across a vast array of CMQ configurations, where individual concept representations

are either honed by isolated concept feedback or by an amalgamation of concept and task-derived feedback.

IV. EXPERIMENTS

In this section, we evaluate the performance of our proposed approach, CMQ, against several strong baselines. Specifically, we compare with three representative value decomposition methods, VDN [4], QMIX [5], and QTRAN [7], and four state-of-the-art MARL algorithms, including WQMIX [22](both CW-QMIX and OW-QMIX), QPLEX [6], CDS [24], and SHAQ [25]. We conduct experiments on two widely-used cooperative MARL benchmarks: the Level-Based Foraging (LBF) environment [26], and the StarCraft II micromanagement benchmark (SMAC) [27], using SC2 version 2.4.10. All experiments are repeated across five random seeds to account for stochasticity, and we report the mean performance along with the 75% confidence intervals. Our results demonstrate that CMQ consistently outperforms all baselines in both final performance and learning efficiency. We also conduct ablation studies to analyze the effect of varying the number of semantic concepts, providing insight into how concept granularity influences coordination. Finally, we evaluate the interpretability of CMQ by visualizing semantic contributions and demonstrating how concept interventions affect decision-making behavior.

To ensure fair comparisons, all methods are trained under consistent hyperparameter settings: a batch size of 32, discount factor $\gamma = 0.99$, and an exploration rate linearly decaying from 1.0 to 0.05. The replay buffer size is set to 5000 transitions, and the target networks are updated every 200 episodes. Optimization is performed using RMSprop with a learning rate of 0.0005, and gradients are clipped to a maximum norm of 10. Our CMQ model is implemented within the PyMARL framework. We use 16 semantic concepts to construct the concept space, and the bias term is modeled via a single-layer MLP with 32 hidden units. All attention modules use a hidden dimension of 64. Training is conducted on a single machine equipped with an NVIDIA RTX 4090 GPU and an Intel i9-13900K CPU. Depending on the scenario complexity, training requires between 2M and 5M timesteps, which corresponds to approximately 0.5 to 5 hours of wall-clock time.

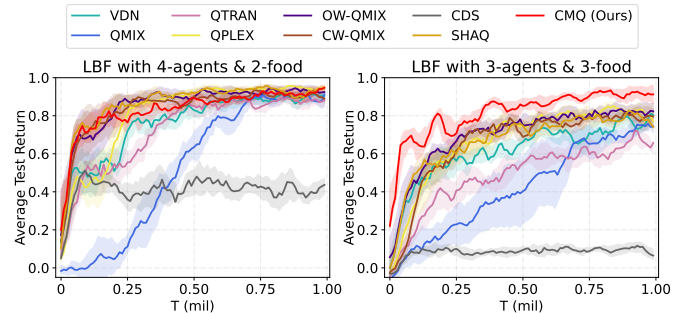


Fig. 2: Average test return on two constructed tasks of LBF.

A. Level Based Foraging

We conducted experiments on two self-created LBF tasks within a 10×10 grid world where agents gather food, requiring

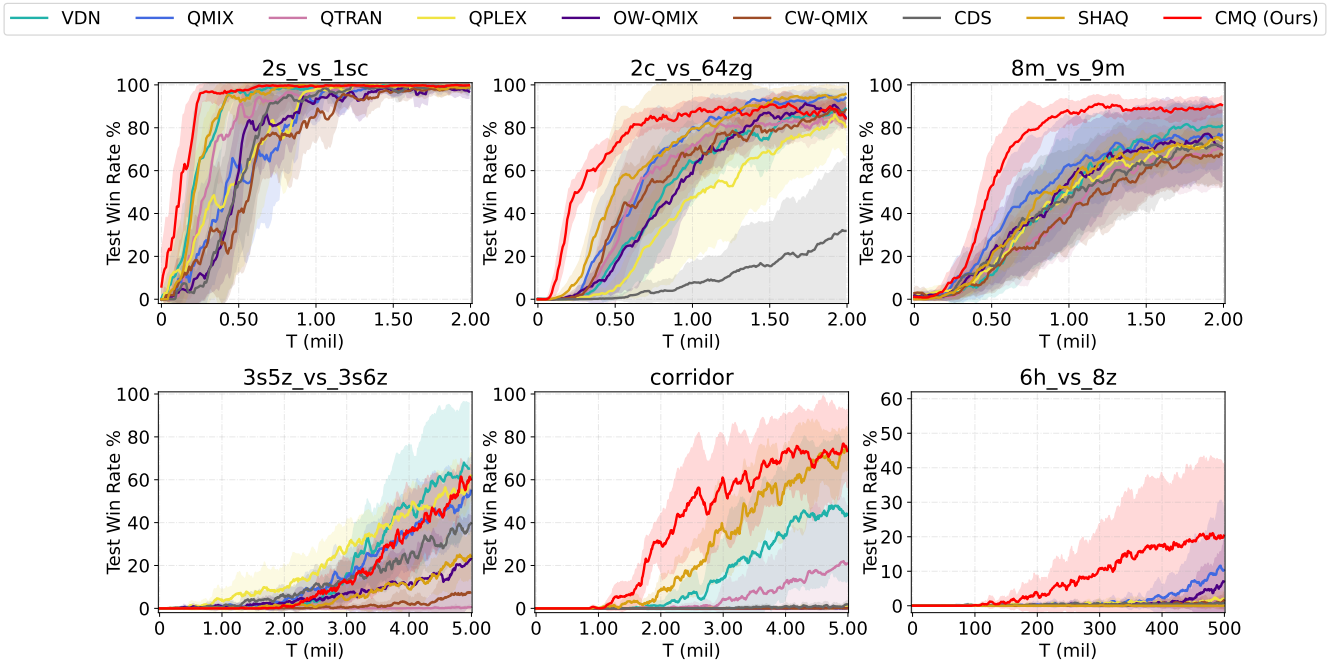


Fig. 3: Performance comparison with baselines on easy and hard (first line), super hard (second line) scenarios.

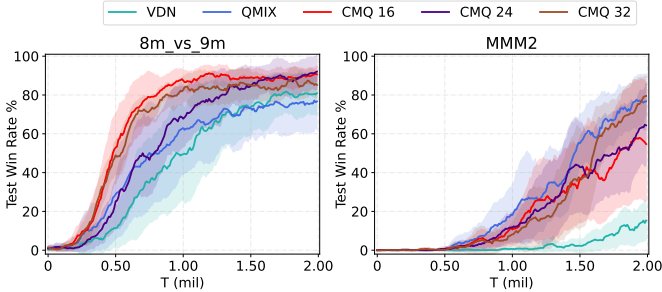


Fig. 4: Performance with different number of concepts on 8m_vs_9m and MMM2 scenarios.

cooperation at times. Agents have a limited perception, seeing only a 5×5 area around them. Cooperative actions are eating lower-level food, resulting in positive rewards, while non-cooperative moves result in a small penalty of -0.002 . The agents’ actions include moving in four directions, eating, or doing nothing. We assessed different algorithms by altering the number of agents and food amounts.

Figure 2 shows the average test return curves of CMQ and baselines. As the agent decreases, each agent needs more cooperative skills to find the food, and it takes more timesteps to obtain a high return. CDS obtains a lower win rate, which may lack of diverse agents to explore collaborative strategies. We find that CMQ obtains higher returns in 3 agents with the 3 food task. It indicates that CMQ can better solve complex tasks with a more cooperative policy capacity, benefiting from the rich representation of cooperation concepts.

B. SMAC Benchmark.

Figure 3 presents the performance of CMQ and baseline methods across various SMAC scenarios. On easy maps like

2s_vs_1sc, CMQ performs on par with other algorithms, all of which are able to learn optimal policies. As the task difficulty increases, especially in maps requiring fine-grained coordination (e.g., 2s_vs_64zg and 8m_vs_9m), CMQ consistently outperforms existing methods. Notably, CMQ surpasses SHAQ and VDN by a small but consistent margin and significantly outperforms QMIX, QTRAN, QPLEX, and CDS. While VDN performs reasonably well in tasks with limited coordination demand, it fails to scale under complex cooperative settings. SHAQ benefits from its enhanced credit assignment but still lags behind CMQ, which leverages cooperation concepts to achieve more accurate decomposition and faster convergence. On 8m_vs_9m, CMQ learns to execute focused fire and tactical retreats, outperforming all baselines by a large margin.

In super-hard scenarios, CMQ improves the average win rate of state-of-the-art methods by nearly 20%. For instance, in the heterogeneous 6h_vs_8z map, agents must alternate kiting and attacking to succeed—a behavior that CMQ successfully learns, while others like VDN, QPLEX, and CDS nearly fail. On corridor, CMQ achieves near-perfect results, while other algorithms struggle to even explore viable strategies. Although performance on 3s5z_vs_3s6z is not superior to all baselines, this may be due to the reduced need for complex credit assignment. Still, CMQ shows faster convergence (within 1M steps) and maintains over 90% win rate on maps that demand consistent coordination. These results suggest that embedding cooperation concepts enhances both interpretability and policy quality, enabling CMQ to balance expressiveness and performance without added model complexity.

Further, we study the learning performance as the number of concepts increases. As shown in Figure 4, as the number

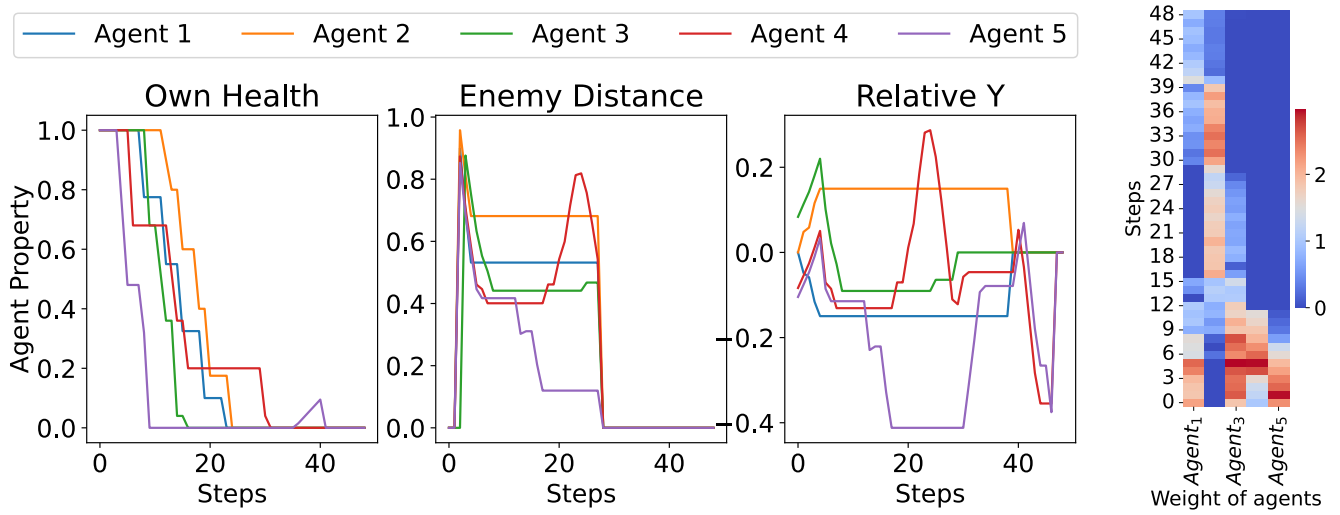


Fig. 5: Visualisation of property semantics and agent contributions on the 2s3z scenario.

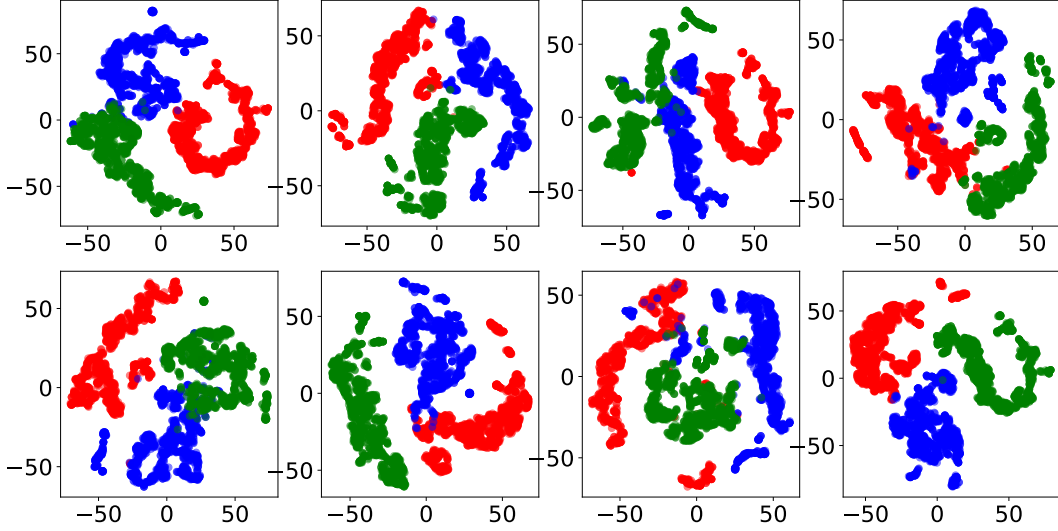


Fig. 6: The t-SNE visualization of cooperative concepts learned of CMQ sample points for 30 episodes, where three of all concepts are randomly selected to be combined into a group.

of concepts increases, the learning performance is improved slightly, i.e., on the 8m_vs_9m map. On MMM2 maps, the extra cooperation concept representation can improve the performance, indicating that a larger number of concepts is required. It is consistent with our motivation that linear decomposition would not harm the expression of the model when given a providential concept representation space. On the other hand, a larger number of concepts will lead to large computational costs. Generally, a moderate number of concepts is enough for an appropriate trade-off between performance improvement and computation costs.

C. Interpretability

To further demonstrate the interpretability of CMQ, we conduct an analysis on the 2s3z scenario using three agent-related features: own health, enemy distance, and relative Y-position. We visualize one randomly selected active concept and the corresponding agent contributions over an episode

(Figure 5). Notably, agent 5 shows high contribution early in the episode, as reflected by its strong activation in the heatmap—likely due to being attacked and experiencing a sharp drop in health. By timestep 10, agent 5’s health reaches zero, and its contribution becomes strongly negative, indicating it has been eliminated. In contrast, agent 4 maintains a low contribution in early stages due to its distance from enemies, reflected by minimal activation. These findings illustrate how CMQ captures meaningful patterns in agent behavior, enabling attribution of action distributions to semantically grounded feature importance.

Beyond attribution, CMQ’s cooperation concepts serve as rich and structured policy representations. To explore this qualitatively, we apply 2D t-SNE visualization to the embeddings of three randomly selected concepts on the 2s3z map (Figure 6). The results show clear clustering in the latent space: samples are grouped both by concept activation and by distinct cooperation modes within each concept. This suggests

that CMQ learns a latent hierarchy over cooperation semantics, enabling interpretable structure without supervision. Such structured embeddings facilitate more informed exploration and contribute to CMQ’s superior learning efficiency and generalization in complex tasks.

V. CONCLUSION

In this work, we propose CMQ to facilitate more efficient linear weighted value decomposition by incorporating concept learning into the mixer, which can maintain representational capability while enjoying interpretability. Technically, we build the concept bottleneck via embedding the global state, which is used to deduce the contributions of each agent in value decomposition and predict the joint value function together with the individual value function. The concept of bottleneck can incentivize our model to react positively to interventions at test time. Empirical evaluations demonstrate that CMQ can achieve stable and efficient learning and help humans understand the cooperation behavior of MARL. For future work, we aim to introduce concept learning into the entire pipeline of the MARL framework to yield further improvement and explore further theoretical properties of concept representation capacity in the MARL community. Our results indicate that CMQ advances the state-of-the-art for the performance-vs-interpretability trade-off, which can make progress on a crucial concern in explainable AI.

REFERENCES

- [1] Y. Cao, W. Yu, W. Ren, and G. Chen, “An overview of recent progress in the study of distributed multi-agent coordination,” *IEEE Trans. Industr. Inform.*, pp. 427–438, 2012.
- [2] C. Zhang and V. Lesser, “Coordinated multi-agent reinforcement learning in networked distributed pomdps,” in *AAAI*, 2011.
- [3] M. Hüttenrauch, A. Šošić, and G. Neumann, “Guided deep reinforcement learning for swarm systems,” *arXiv:1709.06011*, 2017.
- [4] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, “Value-decomposition networks for cooperative multi-agent learning based on team reward,” in *AAMAS*, 2018, pp. 2085–2087.
- [5] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning,” in *ICML*, 2018, pp. 4295–4304.
- [6] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, “QPLEX: Duplex dueling multi-agent Q-learning,” in *ICLR*, 2020, pp. 1–27.
- [7] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, “QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning,” in *ICML*, 2019, pp. 5887–5896.
- [8] Y. Yang, J. Hao, G. Chen, H. Tang, Y. Chen, Y. Hu, C. Fan, and Z. Wei, “Q-value path decomposition for deep multiagent reinforcement learning,” in *ICML*, 2020, pp. 10 706–10 715.
- [9] Z. Liu, Y. Zhu, Z. Wang, Y. Gao, and C. Chen, “Mixrts: Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 4090–4107, 2025.
- [10] Z. Liu, Y. Zhu, and C. Chen, “Na²q: Neural attention additive model for interpretable multi-agent q-learning,” in *International Conference on Machine Learning*, vol. 202, 2023, pp. 22 539–22 558.
- [11] J. Wang, Y. Zhang, T.-K. Kim, and Y. Gu, “Shapley Q-value: A local reward approach to solve global reward games,” in *AAAI*, 2020, pp. 7285–7292.
- [12] T. Zahavy, N. Ben-Zrihem, and S. Mannor, “Graying the black box: Understanding DQNs,” in *ICML*, 2016, pp. 1899–1908.
- [13] A. Ghorbani, A. Abid, and J. Zou, “Interpretation of neural networks is fragile,” in *AAAI*, 2019, pp. 3681–3688.
- [14] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju, “Reliable post hoc explanations: Modeling uncertainty in explainability,” in *NeurIPS*, 2021, pp. 9391–9404.
- [15] O. Bastani, Y. Pu, and A. Solar-Lezama, “Verifiable reinforcement learning via policy extraction,” in *NeurIPS*, 2018, p. 2494–2504.
- [16] A. Silva, M. Gombolay, T. Killian, I. Jimenez, and S.-H. Son, “Optimization methods for interpretable differentiable decision trees applied to reinforcement learning,” in *AISTATS*, 2020, pp. 1855–1865.
- [17] Y. Zhu, X. Yin, and C. Chen, “Extracting decision tree from trained deep reinforcement learning in traffic signal control,” *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1997–2007, 2023.
- [18] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *ICML*. PMLR, 2020, pp. 5338–5348.
- [19] W. Stammer, M. Memmel, P. Schramowski, and K. Kersting, “Interactive disentanglement: Learning concepts by interacting with their prototype representations,” in *CVPR*, 2022, pp. 10 317–10 328.
- [20] M. Havasi, S. Parbhoo, and F. Doshi-Velez, “Addressing leakage in concept bottleneck models,” *NeurIPS*, vol. 35, pp. 23 386–23 397, 2022.
- [21] A. Bai, C.-K. Yeh, P. Ravikumar, N. Y. Lin, and C.-J. Hsieh, “Concept gradient: Concept-based interpretation without linear assumption,” *ICLR*, 2022.
- [22] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, “Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning,” in *NeurIPS*, 2020.
- [23] Y. Yang, J. Hao, B. Liao, K. Shao, G. Chen, W. Liu, and H. Tang, “Qatten: A general framework for cooperative multiagent reinforcement learning,” *arXiv:2002.03939*, 2020.
- [24] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, “Celebrating diversity in shared multi-agent reinforcement learning,” in *NeurIPS*, 2021, pp. 3991–4002.
- [25] J. Wang, J. Wang, Y. Zhang, Y. Gu, and T.-K. Kim, “SHAQ: Incorporating shapley value theory into multi-agent Q-learning,” in *NeurIPS*, 2022.
- [26] F. Christianos, L. Schäfer, and S. Albrecht, “Shared experience actor-critic for multi-agent reinforcement learning,” in *NeurIPS*, 2020, pp. 10 707–10 717.
- [27] M. Samvelyan, T. Rashid, C. Schroeder de Witt, G. Farquhar, N. Nardelli, T. G. Rudner, C.-M. Hung, P. H. Torr, J. Foerster, and S. Whiteson, “The StarCraft Multi-Agent Challenge,” in *AAMAS*, 2019, pp. 2186–2188.