Offline Multi-Agent Reinforcement Learning with Global Moderate Generalization

Yuanrui Duan
University of Science and Technology of China
Hefei, China
dyr_1221@mail.ustc.edu.cn

Wengang Zhou*
University of Science and Technology of China
Hefei, China
zhwg@ustc.edu.cn

Yufeng Shi
University of Science and Technology of China
Hefei, China
shiyufeng@mail.ustc.edu.cn

Xiancheng Gao
University of Science and Technology of China
Hefei, China
gxcnb@mail.ustc.edu.cn

Lin Liu
University of Science and Technology of China
Hefei, China
linliu2021@mail.ustc.edu.cn

Houqiang Li*
University of Science and Technology of China
Hefei, China
lihq@ustc.edu.cn

ABSTRACT

Offline multi-agent reinforcement learning (MARL) suffers from severe value overestimation and extrapolation errors. These challenges lead to over-generalization when value functions or policies encounter out-of-distribution (OOD) actions. The considerable body of work on generalization issues has yielded successful in-sample learning methods, which avoid OOD actions altogether. However, we argue that the conservatism inherent in this approach could impose unnecessary limitations. This study demonstrates that moderate generalization can be both reliable and beneficial for improving performance. Building on this insight, this paper introduces the offline multi-agent reinforcement learning algorithm with global moderate generalization (OMGMG). OMGMG enforces moderate generalization at the global level and dynamically distributes the generalization effects to individual agents through value decomposition techniques, thereby achieving macro-level control over the generalization process. OMGMG comprises two core components: global moderate action generalization and global moderate generalization propagation. The former approach improves value function estimation by selecting joint actions within the vicinity of the dataset. The latter approach ensures the effective propagation of reinforcement learning signals while mitigating the issue of erroneous generalization propagation during the bootstrapping process. Extensive experimental evaluations on the multi-agent Mujoco and StarCraft II benchmark demonstrate that our OMGMG surpasses the current state-of-the-art offline MARL methods across the majority of tasks.

KEYWORDS

Offline Multi-Agent Reinforcement Learning; Global Moderate Generalization

*Corresponding author

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CVXE7640

ACM Reference Format:
Yuanrui Duan, Wengang Zhou, Yufeng Shi, Xiancheng Gao, Lin Liu, and Houqiang Li. 2026. Offline Multi-Agent Reinforcement Learning with Global Moderate Generalization. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/CVXE7640

1 INTRODUCTION

In recent years, online reinforcement learning (RL) has achieved significant progress in various fields including autonomous driving [11, 40], intelligent control [28], and natural language processing [38]. However, practical deployment remains challenging as online RL typically requires millions of trial-and-error interactions. This process consumes substantial resources [12] and introduces safety risks in real-world scenarios [8]. To address these limitations, offline RL has emerged as a promising alternative [17], enabling policy learning solely from pre-collected datasets without the costs and risks of real-time interaction [3, 16, 33, 43].

Although offline RL shows advantages, it still faces several critical challenges [7, 15]. Among these, distributional shift is particularly prominent [15]: Since the learned policy can generate actions that fall outside the distribution of the dataset, these out-of-distribution (OOD) actions are often systematically overestimated, which can lead to instability or failure in the training process [6]. From a generalization perspective, this phenomenon essentially reflects the lack of robust generalization capability of the value function or policy when faced with samples outside the data-supported region [17], a problem widely known as over-generalization [24]. In Multi-agent Reinforcement Learning (MARL), this problem is significantly exacerbated. Complex agent interactions cause the joint state-action space to expand exponentially, aggravating data sparsity and insufficient coverage [41]. This increases the probability of OOD actions and amplifies the value estimation bias [30].

To address the challenges posed by OOD estimation, researchers have drawn inspiration from the *centralized training with decentralized execution* (CTDE) framework [21, 31], extending single-agent conservative algorithms to the multi-agent setting [30, 32, 40, 44]. Specifically, existing methods primarily reduce the overestimation of OOD actions by imposing penalties on unseen actions [32] or regularizing based on behavior policy [20]. Notably, recent studies have

proposed strategies that rely entirely on in-sample data [30, 40] to completely avoid dependence on OOD actions. However, these methods often neglect the generalization potential of deep neural networks, preventing convergence to the optimal solution [22, 27, 45]. Therefore, effectively leveraging model generalization while reducing OOD risks remains a key problem in offline MARL.

In this paper, we theoretically demonstrate that moderate generalization beyond dataset is both reliable and beneficial in continuous multi-agent environments, and its rational utilization can significantly enhance the collaborative performance of multiple agents. Therefore, to reasonably leverage the generalization capability of the network, we propose an innovative offline MARL algorithm with global moderate generalization (OMGMG). The proposed algorithm introduces moderate generalization at the global level and dynamically distributes it to the single-agent level through value decomposition, thereby achieving control over the overall generalization from a global perspective.

OMGMG consists of two core components: global moderate action generalization and global moderate generalization propagation. The former expands the range of joint action selection to the dataset's neighborhood, enabling agents to explore optimal strategies in a broader action space. However, relying exclusively on action generalization can result in the accumulation of erroneous generalization, especially during the multi-agent bootstrapping process. In such scenarios, minor generalization errors can be amplified through TD bootstrapping, ultimately compromising the stability of the system. To mitigate this, we propose the moderate generalization propagation mechanism. This mechanism effectively reduces the generalization propagation speed while preserving the integrity of multi-agent reinforcement learning signal propagation.

In terms of implementation, OMGMG integrates two expectations into the Bellman objective. Specifically, the first is the *moderate generalization expectation*, achieved through a regularized actor-critic objective that biases towards high-advantage actions; the second is the *in-sample expectation*, realized through in-sample learning methods such as implicit local value regularization [41]. This dual-expectation mechanism ensures that the algorithm maintains generalization capabilities while effectively controlling value overestimation and error propagation.

In our experiments, we further extended the proposed approach to discrete offline multi-agent environments, such as SMACv2, and were pleasantly surprised to observe that it achieved strong performance. To comprehensively assess its effectiveness, this study conducts systematic evaluations on both the continuous Multi-agent MuJoCo tasks [34] and the discrete StarCraft II Multi-Agent Challenge (SMACv2) [4] benchmarks using multiple offline datasets. The experimental results demonstrate that the proposed method exhibits significant advantages in key performance metrics compared to existing state-of-the-art baselines.

2 RELATED WORK

Offline Reinforcement Learning. Offline RL, also known as batch RL, aims to learn policies solely using pre-collected datasets without relying on online interactions [19, 39]. Compared to online RL, the main challenges faced in the offline setting are distributional shift and amplification of estimation errors: the trained value function tends to produce overly optimistic estimates when the state-action distribution deviates from the target policy [15]. Recent research has primarily addressed this issue by incorporating various forms of "pessimism" [35]. One class of methods explicitly or implicitly constrains the learned policy to remain close to the behavior policy in the dataset to mitigate the risks posed by distributional shift [3, 5, 7, 23, 29]. This includes direct behavioral cloning, which mimics the actions in the dataset [5, 23], and behavioral constraints, such as Kullback-Leibler (KL) divergence regularization, which penalize deviations from the behavior policy during policy optimization [29]. Another class of methods focuses on conservative value estimation, proposing objectives based on lower-bound optimization or conservative Q-function learning [10, 16, 26, 43]. These methods ensure that the learned Q-function remains conservative for unobserved actions, thus avoiding excessive optimism during policy improvement [10, 16]. A third approach is uncertainty-based methods, which incorporate uncertainty quantification into the reward [47, 48] or the value function [1, 42]. By modeling uncertainty, these methods achieve pessimistic learning and reduce the risk of overestimation in out-of-distribution regions.

Recently, a new learning paradigm has emerged: during the learning of policies and value functions, OOD actions are not queried, and the training process is entirely confined to the dataset, thus completely avoiding distributional extrapolation [13, 25, 43]. However, existing methods tend to be conservative, often resulting in pessimistic value function estimates and an excessive dependence on the quality of the dataset, which limits their generalization capabilities [22, 24, 45]. To mitigate this issue, researchers have proposed the use of *softer constraints*, which relax the conservatism of the value function estimation and provide better generalization [22, 45]. Although these methods have demonstrated excellent performance in the context of single-agent RL, they have not been thoroughly investigated in the realm of offline MARL.

Offline Multi-Agent Reinforcement Learning. Offline MARL is an emerging research domain that integrates principles from MARL and offline RL [9]. It aims to derive optimal collaborative strategies from pre-collected datasets of multi-agent interactions. This challenging field is confronted with two primary challenges: (1) addressing the inherent non-stationarity and credit assignment issues characteristic of MARL [36]; (2) mitigating the multi-agent extrapolation error, which arises from distributional shifts in offline RL due to the mismatch between the dataset and the learned policy [32]. The complexity of these challenges is exacerbated by the exponential growth of the joint action space as the number of agents increases. In multi-agent settings, even minor OOD behaviors by a single agent can be significantly amplified through inter-agent interactions, potentially leading to severe value overestimation and catastrophic outcomes [41].

Early research in offline MARL focused primarily on mitigating OOD risks through explicit constraints or value penalties. For instance, MABCQ extended the concept of Batch-Constrained Q-learning (BCQ) [7] to the multi-agent setting by explicitly constraining the policy outputs of individual agents to the behavior distribution. Similarly, conservatism-based methods like OMAR[32] adapted the Conservative Q-Learning (CQL) [16] objective, which suppresses value overestimation by introducing a regularization term. While these approaches effectively discourage deviations

from the dataset, they still rely on querying or estimating values for unseen actions, which can introduce instability or excessive conservatism in high-dimensional joint action spaces.

To avoid the challenges associated with OOD estimation, recent research has shifted towards in-sample learning paradigms within the CTDE framework. For example, ICQ [44] avoids querying unseen actions by utilizing an implicit constraint mechanism, effectively converting policy updates into a weighted supervised regression problem. OMIGA [41] achieves the transformation from global to local regularization by ensuring global joint optimality through implicit global-local value regularization. Furthermore, ComaDICE [30] combines the Distribution Correction Estimation (DICE) method [18] with a value decomposition strategy. It casts the offline MARL problem as a stationary distribution correction task solvable strictly within the data support, demonstrating superior performance on various benchmarks.

3 PRELIMINARIES

This study addresses the problem of Cooperative Multi-Agent Reinforcement Learning (Cooperative MARL), which can be formally modeled as a multi-agent Partially Observable Markov Decision Process (POMDP) [2].

The framework is defined by the tuple $G = \langle S, A, P, r, Z, O, n, N, \gamma \rangle$, where S represents the set of true states of the environment, and $\boldsymbol{s} \in S$ denotes the true state of the multi-agent system. The system comprises n agents, indexed by the set $N = \{1, \ldots, n\}$. The joint action space is defined as $A = \prod_{i \in N} A_i$, where each agent $i \in N$ has its own set of executable actions A_i. At each time step, each agent i selects an action $a_i \in A_i$, and the actions of all agents collectively form the joint action $\boldsymbol{a} = (a_1, \ldots, a_n) \in A$. The state transition function $P : S \times A \times S \to [0, 1]$ specifies the probability of transitioning to state $\boldsymbol{s'}$ after the agents execute the joint action $\boldsymbol{a}$ in state $\boldsymbol{s}$. Due to the partial observability of the environment, each agent i receives a local observation o_i through the observation function $Z_i(\boldsymbol{s}) : S \to O_i$. The observations of all agents collectively form the joint observation $\boldsymbol{o} = (o_1, \ldots, o_n)$. The system employs a shared reward function $r(\boldsymbol{o}, \boldsymbol{a}) : O \times A^n \to \mathbb{R}$, and $\gamma \in [0, 1)$ represents the discount factor.

In cooperative MARL, the agents aim to learn a joint policy $\boldsymbol{\pi}_{tot} = \{\pi_1, \ldots, \pi_n\}$, which consists of local policies $\{\pi_i\}_{i=1}^n$. The objective is to find the joint policy $\boldsymbol{\pi}_{tot}$ that maximizes the expected discounted return: $\mathbb{E}_{\boldsymbol{o} \in O, \boldsymbol{a} \sim \boldsymbol{\pi}_{tot}} \left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{o}_t, \boldsymbol{a}_t) \right]$. During the training of the global Q-function Q_{tot}, which evaluates joint observation-action pairs, the Bellman expectation equation is typically utilized for updates, with the loss function defined as follows:

$$\min_{Q_{tot}} \mathbb{E}_{(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o'}) \sim \mathcal{D}} \Big[\big(r(\boldsymbol{o}, \boldsymbol{a}) + \gamma \mathbb{E}_{\boldsymbol{a'} \sim \boldsymbol{\pi}_{tot}(\cdot | \boldsymbol{o'})} \bar{Q}_{tot}(\boldsymbol{o'}, \boldsymbol{a'}) - Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) \big)^2 \Big]. \tag{1}$$

where $\mathcal{D}$ represents the offline dataset, $\bar{Q}_{tot}$ denotes the global target Q-function.

In the offline setting, the training data is collected through interactions between the behavior policy $\hat{\boldsymbol{\beta}}_{tot} = \{\hat{\beta}_1, \ldots, \hat{\beta}_n\}$ and the environment. During training, only the data sampled from the dataset are used, without further interaction with the actual multi-agent environment.

4 METHOD

This section formally introduces the application of moderate generalization to offline MARL. Section 4.1 provides a formal analysis of the impact of generalization on the performance of MARL, covering both over-generalization and non-generalization. Section 4.2 presents a theoretical analysis demonstrating that moderate generalization beyond the dataset is effective in continuous multi-agent settings. In addition, it introduces the core concept of OMGMG, which is made up of two essential components: global moderate action generalization and global moderate generalization propagation. Finally, we elaborate on the implementation details of OMGMG, specifically how global moderate generalization is propagated to each agent through value decomposition methods.

4.1 Exploration of Generalization in Offline MARL

Offline MARL algorithms typically update the value function using the Bellman expectation operator. While the specific instantiation of this operator varies across different methods, we formalize the standard representation of the Bellman expectation operator $\mathcal{T}_{\boldsymbol{u}_{tot}}$:

$$\mathcal{T}_{\boldsymbol{u}_{tot}} Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) := r(\boldsymbol{o}, \boldsymbol{a}) + \gamma \mathbb{E}_{\substack{\boldsymbol{o'} \sim P(\cdot | \boldsymbol{o}, \boldsymbol{a}), \\ \boldsymbol{a'} \sim \boldsymbol{u}_{tot}}} \left[Q_{tot}(\boldsymbol{o'}, \boldsymbol{a'}) \right], \tag{2}$$

where $\boldsymbol{a'}$ represents the joint action in the subsequent time step generated by an arbitrary joint policy $\boldsymbol{u}_{tot}$.

During offline training, this operator is only performed for pairs of joint observation-action $(\boldsymbol{o}, \boldsymbol{a})$ within the dataset $\mathcal{D}$, and the value estimates for pairs outside the dataset $(\boldsymbol{o}, \boldsymbol{a})$ are influenced by generalization. Since the target distribution depends on $\boldsymbol{u}_{tot}$, a bidirectional interaction arises between the Bellman expectation operator and the generalization mechanism [24], which is manifested as follows: first, when $(\boldsymbol{o}, \boldsymbol{a}) \in \mathcal{D}$, the operator update indirectly affects the value function estimation of $(\boldsymbol{o}, \boldsymbol{a}) \notin \mathcal{D}$ via generalization; second, when the target value involves $(\boldsymbol{o}, \boldsymbol{a}) \notin \mathcal{D}$, this update process in turn affects the value estimation of $(\boldsymbol{o}, \boldsymbol{a}) \in \mathcal{D}$.

However, this interaction in offline MARL often leads to the prevalent issue of value overestimation. Specifically, the complex generalization mechanism inherent in such systems tends to overestimate the global Q-values associated with OOD joint actions. During the update process, the algorithm is inclined to select these overestimated Q-values, thereby intensifying the overestimation phenomenon. This overestimation effect distorts the value estimation for $(\boldsymbol{o}, \boldsymbol{a}) \in \mathcal{D}$ and propagates to the values of $(\boldsymbol{o}, \boldsymbol{a}) \notin \mathcal{D}$ through the generalization capability of the network. As bootstrapping continues iteratively throughout the training process, this issue may ultimately lead to the divergence of the value function. The underlying mechanism driving this detrimental process can be primarily attributed to the problem of over-generalization.

To address the issue of over-generalization in RL, researchers have proposed various conservative algorithms [32, 44]. In recent years, scholars have further advanced the field by introducing in-sample learning algorithms tailored for offline MARL [30, 41]. These algorithms fundamentally circumvent extrapolation errors by completely avoiding generalization behaviors during the training process. Specifically, they construct the Bellman target exclusively using the data available within the dataset, which is equivalent

to setting $\boldsymbol{u}_{tot} = \boldsymbol{\beta}_{tot}$ in the operator $\mathcal{T}_{\boldsymbol{u}_{tot}}$. The policy is subsequently extracted by updating the obtained value function. From the perspective of network generalization, such algorithms can be characterized as non-generalization. However, generalization is one of the core strengths that makes neural networks highly versatile and widely applicable. Consequently, in-sample learning algorithms may appear overly conservative due to their limited utilization of the network's generalization capabilities. This limitation becomes particularly evident in scenarios where the offline dataset fails to adequately cover the optimal actions in large-scale or continuous action spaces, thereby restricting the algorithm's effectiveness in more complex environments.

4.2 Global Moderate Generalization

This section aims to demonstrate the effectiveness of moderate generalization and introduces the core concept of this study: global moderate generalization (GMG). This concept emphasizes the reasonable and effective utilization of generalization capabilities while maintaining model performance.

Previous discussions have qualitatively suggested that leveraging the generalization capability of neural networks can enhance performance. To theoretically validate the acceptability of moderate generalization in continuous multi-agent settings, we begin with a minimal update setting, defined as a single gradient step with a small learning rate α, and analyze the generalization behavior induced by updating the global Q-function Q_{tot}. The parameters θ are updated to θ' via a gradient step based on a single in-sample transition $(\boldsymbol{o}, \boldsymbol{a}) \in \mathcal{D}$. Specifically, for an OOD joint action $\tilde{\boldsymbol{a}}$ where $(\boldsymbol{o}, \tilde{\boldsymbol{a}}) \notin \mathcal{D}$, we derive the following theorem:

Theorem 4.1. *(Informal) Under specific continuity conditions, when the learning rate α is sufficiently small and $\tilde{\boldsymbol{a}}$ is sufficiently close to $\boldsymbol{a}$, the following equality holds:*

$$Q_{tot}(\boldsymbol{o}, \tilde{\boldsymbol{a}}; \theta') = Q_{tot}(\boldsymbol{o}, \tilde{\boldsymbol{a}}; \theta) + C_1(\mathcal{T}_u Q_{tot}(\boldsymbol{o}, \tilde{\boldsymbol{a}}; \theta) - Q_{tot}(\boldsymbol{o}, \tilde{\boldsymbol{a}}; \theta) + C_2\|\tilde{\boldsymbol{a}} - \boldsymbol{a}\|) + O(\|\theta' - \theta\|^2), \quad (3)$$

where $C_1 \in [0, 1]$ and C_2 is a finite constant. The last term represents the higher-order error of the update.

The formal statement of the theorem and its complete proof are provided in Appendix A.1. The theorem indicates that when the learning rate is sufficiently small, for $(\boldsymbol{o}, \tilde{\boldsymbol{a}}) \notin \mathcal{D}$, the update increment of the global Q-function is primarily governed by the target difference and the interaction bias term. Specifically, when $\tilde{\boldsymbol{a}}$ approaches $\boldsymbol{a}$, this bias tends to vanish, and the coefficient C_1 further modulates the magnitude of the update, thereby causing the overall update behavior to approximate a target-based gradient update explicitly executed at $(\boldsymbol{o}, \tilde{\boldsymbol{a}})$, and this update is entirely induced by the generalization effect. Therefore, Theorem 4.1 states that, under certain continuity conditions, if the perturbation of joint actions is controlled within a local neighborhood, the update of the global Q-function can formally approximate the true global target update. This result suggests that moderately allowing generalization beyond the dataset support is expected to yield a better global Q-function, achieving superior policy performance. Based on this, we define a novel global moderate generalization policy $\tilde{\boldsymbol{\beta}}_{tot}$ as follows:

Definition 4.2. $\tilde{\boldsymbol{\beta}}_{tot}$ is referred to as a global moderate generalization policy if it satisfies the following conditions:

$$\begin{aligned} &\mathrm{supp}(\hat{\boldsymbol{\beta}}_{\mathrm{tot}}(\cdot \mid \boldsymbol{o})) \subseteq \mathrm{supp}(\tilde{\boldsymbol{\beta}}_{\mathrm{tot}}(\cdot \mid \boldsymbol{o})), \\ &\text{and} \max_{\boldsymbol{a}_1 \sim \hat{\boldsymbol{\beta}}_{\mathrm{tot}}(\cdot|\boldsymbol{o})} \min_{\boldsymbol{a}_2 \sim \hat{\boldsymbol{\beta}}_{\mathrm{tot}}(\cdot|\boldsymbol{o})} \|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \leq \epsilon_{\boldsymbol{a}}, \end{aligned} \quad (4)$$

where $\hat{\boldsymbol{\beta}}_{tot}$ is the behavior policy observed in the offline dataset.

This indicates that under the same observation, the joint action space covered by the global moderate generalization policy is larger, but it must satisfy that for any joint action $\boldsymbol{a}_1$ obtained from the global moderate generalization policy, there exists a joint action $\boldsymbol{a}_2$ obtained from the behavior policy such that $\|\boldsymbol{a}_1 - \boldsymbol{a}_2\| \leq \epsilon_{\boldsymbol{a}}$. Therefore, the support set of $\tilde{\boldsymbol{\beta}}_{tot}$ is broader, meaning that it can explore actions not present in the dataset, yet it ensures the reliability of generalization through the "moderate" constraint. Consequently, according to Theorem 4.1, within this moderate generalization range, the global Q-function has a high probability of achieving a good generalization.

However, the generalization capability of neural networks presents a dual-edged challenge. Even when confined to a credible moderate generalization region, the value function still exhibits a non-negligible generalization error. This form of bias is systematically propagated and amplified during the update of temporal differences through the bootstrapping process, which is weighted by the discount factor γ, which ultimately leads to value overestimation. To mitigate this issue, we introduce the concept of global moderate generalization propagation and propose a novel GMG operator, defined as follows:

Definition 4.3. The GMG operator is defined as follows:

$$\begin{aligned} \mathcal{T}_{GMG} Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) := r(\boldsymbol{o}, \boldsymbol{a}) \\ + \gamma \mathbb{E}_{\boldsymbol{o}' \sim P(\cdot|\boldsymbol{o}, \boldsymbol{a})} \Big[\lambda \mathbb{E}_{\boldsymbol{a}' \sim \hat{\boldsymbol{\beta}}_{tot}(\cdot|\boldsymbol{o}')} Q_{tot}(\boldsymbol{o}', \boldsymbol{a}') \\ + (1 - \lambda) \mathbb{E}_{\boldsymbol{a}' \sim \tilde{\boldsymbol{\beta}}_{tot}(\cdot|\boldsymbol{o}')} Q_{tot}(\boldsymbol{o}', \boldsymbol{a}') \Big], \end{aligned} \quad (5)$$

where $\hat{\boldsymbol{\beta}}_{tot}$ is the behavior policy observed in the offline dataset and $\tilde{\boldsymbol{\beta}}_{tot}$ is global moderate generalization policy.

The GMG operator differs from the traditional Bellman expectation operator in two key aspects. First, by introducing a generalization discount factor λ, the effective discount factor is reduced to $\lambda\gamma$. Unlike explicit value penalties used in standard pessimism, this mechanism serves as a structural constraint that dampens the propagation of extrapolation errors. This adjustment mitigates the accumulation of overestimation bias during the Bellman iteration, thereby enhancing the stability of value estimation. Second, the operator does not rely exclusively on in-sample learning but also incorporates OOD actions. This approach leverages the generalization capability of neural networks to achieve global moderate generalization. Furthermore, Theorem 4.1 demonstrates that the value function has a high probability of achieving robust generalization within the moderate generalization region.

In general, the generalization of GMG exhibits moderation in two aspects: 1. *global moderate actions generalization*: Agents select actions within the joint action space by employing a global moderate

generalization policy $\tilde{\boldsymbol{\beta}}_{tot}$. This policy operates with a generalization scope broader than $\hat{\boldsymbol{\beta}}_{tot}$, thereby endowing the moderate generalization operator with favorable generalization properties. 2. *global moderate generalization propagation*: The GMG integrates globally generalized Q-values, derived from moderate generalization, with in-sample global Q-values using a generalization discount factor. This approach enables the effective propagation of reinforcement learning signals across multiple agents while regulating the degree of generalization. By reducing the sensitivity of generalization propagation to the discount factor, our method significantly alleviates the issue of global value estimation amplification caused by the bootstrapping mechanism in multi-agent collaborative environments. These techniques collectively contribute to a more robust and scalable framework for multi-agent reinforcement learning.

4.3 Practical Algorithm

Building upon the theoretical foundations laid out in the previous sections, this section provides a comprehensive exposition of the proposed OMGMG algorithm. In the implementation process, we not only validated its effectiveness in continuous multi-agent environments but also successfully extended it to discrete multi-agent scenarios, where it also achieved promising empirical results. This algorithm is structured around several key components, including the global policy $\boldsymbol{\pi}_{tot}$, local policy π_i, global target policy $\bar{\boldsymbol{\pi}}_{tot}$, local target policy $\bar{\pi}_i$, global Q-function Q_{tot}, local Q-function Q_i, global target Q-function $\bar{Q}_{tot}$, local target Q-function $\bar{Q}_i$, global V-function V_{tot}, and local V-function V_i. Here, each local function corresponds to each agent i.

Policy Learning. In practical applications, the quality of the dataset often exhibits significant variability. To improve performance, we aim for OMGMG to achieve moderate generalization in the vicinity of high-quality joint actions within the dataset. This approach increases the likelihood of obtaining superior results. To this end, we reconstruct the behavior policy $\hat{\boldsymbol{\beta}}_{tot}$ in the dataset, biasing it towards actions associated with higher advantage values $\hat{\boldsymbol{\beta}}^*_{tot}(\boldsymbol{a} \mid \boldsymbol{o}) \propto \hat{\boldsymbol{\beta}}_{tot}(\boldsymbol{a} \mid \boldsymbol{o}) \exp(A_{tot}(\boldsymbol{o}, \boldsymbol{a}))$. Additionally, we introduce a proximity constraint between the training policy and the reconstructed behavior policy to effectively regulate the extent of generalization. Specifically, the generalization set is defined as follows:

$$\Pi_G = \{\boldsymbol{\pi}_{tot} \mid \mathrm{KL}(\hat{\boldsymbol{\beta}}^*_{tot}(\cdot \mid o) \parallel \boldsymbol{\pi}_{\boldsymbol{tot}}(\cdot \mid o)) \leq \epsilon\}. \tag{6}$$

Incorporating the KL divergence constraint into the formulation enables the policy $\boldsymbol{\pi}_{tot}$ to explore joint actions beyond those present in the dataset with a controlled probability. This ensures that the generalization set encompasses joint actions that are not explicitly included in the dataset. Once the generalization set is defined, the primary objective is to train the policy within this set to maximize the global Q-function. To accomplish this, the study adopts the actor-critic framework for policy optimization.

$$\max_{\boldsymbol{\pi}_{tot}} \mathbb{E}_{\boldsymbol{o}\sim\mathcal{D}, \boldsymbol{a}\sim\boldsymbol{\pi}_{tot}(\cdot|\boldsymbol{o})} Q_{tot}(\boldsymbol{o}, \boldsymbol{a}), \quad \text{s.t.} \quad \boldsymbol{\pi}_{tot} \in \Pi_G. \tag{7}$$

Using the Lagrange multiplier method, the constraint term is integrated into the objective function. Here, ν serves as the Lagrange multiplier (penalty coefficient) that regulates the KL-divergence constraint. This leads to the maximization of the following modified objective function:

$$\begin{aligned} \max_{\boldsymbol{\pi}_{tot}} \quad & \mathbb{E}_{\boldsymbol{o}\sim\mathcal{D}, \boldsymbol{a}\sim\boldsymbol{\pi}_{tot}(\cdot|\boldsymbol{o})} Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) \\ & - \nu\, \mathbb{E}_{\boldsymbol{o}\sim\mathcal{D}} \mathrm{KL}\big(\hat{\boldsymbol{\beta}}^*_{tot}(\cdot|\boldsymbol{o}) \| \boldsymbol{\pi}_{tot}(\cdot|\boldsymbol{o})\big). \end{aligned} \tag{8}$$

At this stage, we have derived the global policy optimization framework in terms of the global state-action value function Q_{tot} and the global state value function V_{tot}. However, in multi-agent systems, the joint state-action space expands exponentially with the number of agents, rendering the precise estimation of global value functions highly challenging. To mitigate this issue, we propose a value decomposition method, which is formulated as follows:

$$\begin{aligned} Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) &= \sum_i w_i(\boldsymbol{o}) Q_i(o_i, a_i) + b(\boldsymbol{o}), \\ V_{tot}(\boldsymbol{o}) &= \sum_i w_i(\boldsymbol{o}) V_i(o_i) + b(\boldsymbol{o}), \\ w_i &\geq 0, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{9}$$

Here, Q_i and V_i denote the value functions of the i-th agent, where the weights w and the bias b are generated by a learnable hyperparameter network that takes the joint observation $\boldsymbol{o}$ as input. To ensure consistency in the allocation of values between Q_{tot} and V_{tot}, this study employs a unified decomposition coefficient to link the two. Furthermore, to establish a mapping between local policies and local value functions, we assume that the global policy can be decomposed as: $\boldsymbol{\pi}_{tot}(\boldsymbol{a} \mid \boldsymbol{o}) = \prod_{i=1}^n \pi_i(a_i \mid o_i)$. Based on this assumption, we derive the training objective function for the local policy as follows:

$$\begin{aligned} \max_{\pi_i} \; & \mathbb{E}_{o_i\sim\mathcal{D}, a_i\sim\pi_i(\cdot|o_i)} \left[Q_i(o_i, a_i)\right] \\ & + \nu\, \mathbb{E}_{(o_i,a_i)\sim\mathcal{D}} \Bigg[\exp\left(\frac{w_i(\mathbf{o})}{\alpha} \left(\bar{Q}_i(o_i, a_i) - V_i(o_i)\right)\right) \\ & \cdot \log \pi_i(a_i \mid o_i) \Bigg]. \end{aligned} \tag{10}$$

This approach facilitates effective generalization and rational allocation of global policy. The detailed mathematical derivation is provided in Appendix A.2.

Value Learning. We now apply the proposed moderate generalization operator to the training of the value function. Through the policy optimization process described above, the expectation $E_{\boldsymbol{a}'\sim\tilde{\boldsymbol{\beta}}_{tot}}$ can be effectively approximated by $E_{\boldsymbol{a}'\sim\bar{\boldsymbol{\pi}}_{tot}}$. To enhance the stability of the training process, we represent the expected value $E_{\boldsymbol{a}'\sim\hat{\boldsymbol{\beta}}(\cdot|\boldsymbol{o}')} Q_{tot}(\boldsymbol{o}', \boldsymbol{a}')$ within the moderate generalization operator using the state value function V_{tot}. This state value function can be derived using any in-sample learning algorithm. In this work, we adopt the formula value decomposition method and integrate it with the implicit local value regularization technique within the OMIGA algorithm [41] for training.

$$\begin{aligned} \min_{V_i} \quad & \mathbb{E}_{(o_i,a_i)\sim\mathcal{D}} \Bigg[\exp\big(\frac{w_i(\boldsymbol{o})}{\alpha} \left(\bar{Q}_i(o_i, a_i) - V_i(o_i)\right)\big) \\ & + \frac{w_i(\boldsymbol{o})}{\alpha} V_i(o_i) \Bigg]. \end{aligned} \tag{11}$$

Table 1: Average returns of OMGMG and baselines with the mean and standard on MaMujoco tasks.

Task	Dataset	MABCQ [7]	MACQL [16]	ICQ [44]	OMIGA [41]	ComaDICE [30]	OMGMG(ours)
HalfCheetah	expert	2992.7±629.7	1189.5±1034.5	2955.9±459.2	3383.6±552.7	4082.9±45.7	**4707.2±140.1**
	medium	2590.5±1110.4	1011.3±1016.9	2549.3±96.3	3608.1±237.4	2664.7±54.2	**3976.7±271.3**
	m-replay	-333.6±152.1	1998.7±693.9	1922.4±612.9	2504.7±83.5	2855.0±242.2	**4422.8±199.3**
	m-expert	3543.7±780.9	1194.2±1081.0	2834.0±420.3	2948.5±518.9	3889.7±81.6	**4656.6±145.1**
Hopper	expert	77.9±58.0	159.1±313.8	754.7±806.3	859.6±709.5	**2827.7±62.9**	2683.9±470.2
	medium	44.6±20.6	401.3±199.9	501.8±14.0	1189.3±544.3	822.6±66.2	**2054.9±134.9**
	m-replay	26.5±24.0	31.4±15.2	195.4±103.6	774.2±494.3	906.3±242.1	**1678.0±218.4**
	m-expert	54.3±23.7	64.8±123.3	355.4±373.9	709.0±595.7	1362.4±522.9	**1990.4±288.6**
Ant	expert	1317.7±286.3	1042.4±2021.6	2050.0±11.9	2055.5±1.6	2056.9±5.9	**2062.3±3.8**
	medium	1059.6±91.2	533.9±1766.4	1412.4±10.9	1418.4±5.4	1425.0±2.9	**1428.1±3.2**
	m-replay	950.8±48.8	234.6±1618.3	1016.7±53.5	1105.1±88.9	1122.9±61.0	**1151.1±25.0**
	m-expert	1020.9±242.7	800.2±1621.5	1590.2±85.6	1720.3±110.6	**1813.9±68.4**	1801.1±22.8

Finally, by combining Equation (1) and Equation (5), we derive the following training objective for the global Q-function:

$$\min_{\substack{Q_i,w_i,b\\ i=1,\dots,n}} \mathbb{E}_{(\boldsymbol{o},\boldsymbol{a},\boldsymbol{o}')\sim\mathcal{D}}\Big[\big(r(\boldsymbol{o},\boldsymbol{a}) + \lambda\gamma\mathbb{E}_{\boldsymbol{a}'\sim\bar{\pi}_{tot}(\cdot|\boldsymbol{o}')}\bar{Q}_{tot}(\boldsymbol{o}',\boldsymbol{a}') + (1-\lambda)\gamma V_{tot}(\boldsymbol{o}') - Q_{tot}(\boldsymbol{o},\boldsymbol{a})\big)^2\Big]. \quad (12)$$

Overall Algorithm Building upon the aforementioned learning process, we implement global moderate generalization within offline multi-agent systems. By leveraging a value decomposition mechanism, this global capability is dynamically decomposed to each agent. The detailed implementation of this process is presented in Algorithm 1, which provides the complete pseudocode of OMGMG. Key hyperparameters include the generalization discount factor λ, designed to dampen generalization error propagation, and the penalty coefficient ν, which regulates the trade-off between policy optimization and behavior regularization.

Algorithm 1 OMGMG

Require: offline dataset $\mathcal{D}$, generalization discount factor λ, and penalty coefficient ν.

1: Initialize w, b, and π_i, $\bar{\pi}_i$, Q_i, $\bar{Q}_i$, V_i for each agent $i = 1, 2, \dots, n$;
2: **for** each gradient step **do**
3: Sample a mini-batch of transitions $(\boldsymbol{o}, \boldsymbol{a}, r, \boldsymbol{o}')$ from $\mathcal{D}$;
4: Update $V_i(o)$ for each agent i by minimizing Eq. (11);
5: Update $Q_i(o_i, a_i)$, $w(o)$ and $b(o)$ using Eq. (9) by minimizing Eq. (12);
6: Update π_i for each agent i by maximizing Eq. (10);
7: Update $\bar{Q}_i(o_i, a_i)$ by $Q_i(o_i, a_i)$ and $\bar{\pi}_i$ by π_i for each agent i;
8: **end for**

5 EXPERIMENTS

5.1 Datasets

This study employs two widely recognized MARL benchmark environments for algorithm evaluation: Multi-agent MuJoCo [34] (MaMujoco), and SMACv2 [4]. Among these, MaMujoco represents a continuous action space environment, while SMACv2 are discrete action space environments. This selection of environments ensures the rigor and comprehensiveness of the experimental evaluation.

MaMujoco is a multi-agent continuous control task built on the MuJoCo physics simulator. Unlike single-agent MuJoCo tasks, MaMujoco distributes the degrees of freedom for robot or system control across multiple agents. This setup is designed to study the collaboration and division of labor in high-dimensional continuous action spaces under realistic physical constraints. The dataset used in this study was collected by Wang et al. [41] using the HAPPO method [14]. It includes three tasks: Hopper-v2, Ant-v2, and HalfCheetah-v2. Each task comprises four quality levels: expert, medium, medium-replay, and medium-expert.

SMACv2 builds upon SMACv1 [37], which is a multi-agent micromanagement benchmark based on StarCraft II. It is primarily used to evaluate the collaborative capabilities of CTDE strategies in partially observable discrete action spaces. SMACv2 introduces two key enhancements: (1) randomly generated unit types and initial positions, and (2) restricted field of view and attack range. Additionally, the action space is extended to 12 dimensions, allowing agents to select the field of view area, significantly increasing the complexity of the task. SMACv2 includes multiple tasks involving the three StarCraft races: Protoss, Terran, and Zerg. Since there is no publicly available offline dataset for SMACv2, this study generated four datasets of varying quality using the MAPPO algorithm [46] for three tasks (Protoss_5_vs_5, Terran_5_vs_5, and Zerg_5_vs_5): random, medium, medium_expert, and expert. More details are provided in Appendix B.1.

5.2 Baselines

This study employs a diverse and comprehensive selection of classical and state-of-the-art (SOTA) algorithms in the domain of offline MARL as experimental baselines, including BC, MABCQ [7], MACQL [16], ICQ [44], OMAR [32], OMIGA [41], and ComaDICE [30]. Specifically, MABCQ extends the batch-constrained policy of BCQ to multi-agent scenarios. Both MACQL and OMAR are grounded in the CQL framework, incorporating conservative terms into the value function through distinct methodologies and adapting them to multi-agent settings. ICQ, OMIGA, and ComaDICE leverage advanced in-sample learning techniques tailored for multi-agent environments. Among these, ICQ adapts to multi-agent tasks by decomposing joint policies under implicit constraints. OMIGA ensures the global joint optimality of local policies by transforming global regularization into local regularization. ComaDICE, on the

Table 2: Average winrate of OMGMG and baselines with the mean and standard on SMACv2 tasks.

Method	expert	medium	m-expert	random
Protoss_5_vs_5				
BC	35.0±4.6	26.9±5.4	35.6±6.7	3.8±3.1
ICQ [44]	7.5±3.7	5.6±6.1	6.2±4.0	4.4±4.2
OMAR [32]	11.9±1.2	7.5±2.5	8.8±2.3	3.8±1.3
OMIGA [41]	37.5±10.8	26.2±4.2	24.4±9.1	5.0±2.5
ComaDICE [30]	22.5±3.6	19.4±1.2	20.0±9.2	4.4±4.7
OMGMG(ours)	**49.4±4.1**	**31.9±3.6**	**39.4±4.2**	**10.0±3.1**
Terran_5_vs_5				
BC	26.9±4.2	21.2±3.1	28.7±6.1	6.2±2.0
ICQ [44]	20.6±6.4	10.0±7.2	16.9±5.4	4.5±3.2
OMAR [32]	15.0±1.1	9.4±5.7	10.0±6.5	3.8±1.8
OMIGA [41]	32.5±5.3	23.8±1.5	31.9±7.0	5.0±1.5
ComaDICE [30]	33.1±3.2	25.0±2.8	29.4±4.7	**13.8±3.2**
OMGMG(ours)	**37.5±4.8**	**26.2±4.7**	**36.9±5.0**	12.5±2.8
Zerg_5_vs_5				
BC	15.6±6.2	13.1±5.4	15.6±5.9	4.4±3.8
ICQ [44]	7.5±2.5	5.9±2.5	6.2±6.6	0.6±1.3
OMAR [32]	12.5±6.8	10.6±2.7	11.3±5.3	3.1±0.4
OMIGA [41]	22.5±9.4	18.8±6.6	19.4±5.4	3.8±3.6
ComaDICE [30]	18.1±1.2	17.5±1.5	18.8±5.2	6.9±1.3
OMGMG(ours)	**28.7±7.8**	**20.0±1.5**	**26.9±7.6**	**7.5±1.5**

other hand, integrates stationary distribution shift regularization into the objective function and utilizes a hybrid network to decompose global value and advantage functions, facilitating efficient and stable learning. In particular, ComaDICE is currently regarded as the SOTA method in this field.

5.3 Comparison with Baselines

In this work, performance evaluation metrics were carefully selected based on the specific characteristics of the tasks. For reward-oriented MaMujoco tasks, return was used as the primary performance metric. In adversarial-based SMACv2 game environments, the winrate was utilized as the core evaluation criterion. Consistent with standard offline RL protocols, all algorithms were trained strictly on static datasets, and the learned policies were evaluated via interaction with the environment. To ensure the reliability and reproducibility of the experimental results, each experiment was conducted independently using five different random seeds. The mean and standard deviation of the corresponding evaluation metrics were calculated to provide a comprehensive assessment of the algorithm's performance.

The experimental results are summarized in Tables 1 and Tables 2. These results demonstrate that the algorithm proposed in this paper surpasses the baseline methods in the majority of task scenarios, both in the continuous-action MaMujoco environment and the discrete-action SMACv2 environment, achieving the highest overall performance score. Specifically, methods such as MABCQ and MACQL, which simply apply conservative constraints to each agent, result in overly strong constraints and generally lower performance. Although methods like OMIGA and ComaDICE have shown some improvement by introducing single-agent in-sample learning combined with multi-agent value decomposition, they still exhibit significant performance bottlenecks due to insufficient utilization of network generalization capabilities. In contrast, OMGMG introduces moderate generalization at the global level and dynamically allocates it to each agent through value decomposition methods, allowing the exploration of better strategies within a broader feasible domain and thus achieving significant performance improvements. In particular, on the Medium and Expert datasets of the Ant environment, OMGMG and ComaDICE exhibit comparable performance. This is mainly because the average trajectory returns of these datasets already establish a practical upper bound (Medium: 1418.7 ± 37.0; Expert: 2055.1 ± 22.1). Consequently, both methods successfully recover policies close to these high-quality behaviors. In contrast, for tasks with greater optimization potential, such as HalfCheetah, OMGMG achieves comprehensive leading performance. The comparative experimental results strongly confirm the significant superiority of the proposed method in the MaMujoco and SMACv2 environments. More details of the experimental settings are provided in Appendix B.2.

5.4 Ablation Study

Generalization Discount Factor λ: This study investigates the influence of the generalization discount factor on the extent of global generalization propagation by fixing the action generalization parameter ν and conducting experiments with $\lambda \in [0, 1]$. Figure 1 presents trends in performance metrics and value estimation results for various tasks as λ varies. The experimental findings reveal that as λ increases, the bootstrapping effect in OMGMG intensifies, thus increasing the extent of generalization propagation. This process is accompanied by a corresponding rise in the learned value Q_{tot}, albeit at the risk of inducing value divergence should λ become excessive. Consequently, selecting an appropriate λ to ensure moderate generalization propagation has a significant impact on algorithmic performance across diverse tasks. By maintaining the same degree of action generalization, moderate generalization propagation effectively mitigates value overestimation, enhances policy learning stability, and ultimately improves overall performance.

Penalty Coefficient ν: The penalty coefficient, as a crucial hyperparameter for governing the global action generalization degree, significantly impacts algorithmic performance. To systematically assess the influence of ν, this study employs controlled experiments by keeping λ constant and varying $\nu \in \{0.001, 0.01, 0.1, 1, 10\}$ for ablation experiments. Figure 2 presents trends in performance metrics and value estimation results for various tasks as ν varies. As observed, a decrease in ν significantly expands the action generalization range in OMGMG, leading to a general increase in the learned values of Q_{tot}. However, this expansion may trigger value overestimation. Specifically, an excessively large action generalization range causes the agent to produce biased value estimates for OOD actions, resulting in abnormally high estimated values and significantly increased variance. Consequently, selecting an appropriate ν not only enhances the overall performance of the model, but also effectively improves algorithmic stability.

Comparing Figure 1 and Figure 2, it is evident that λ exerts a more pronounced influence on the training of Q_{tot}. This is mainly because λ directly modulates the update process of Q_{tot} by controlling the scale of generalization propagation. In contrast, while ν has

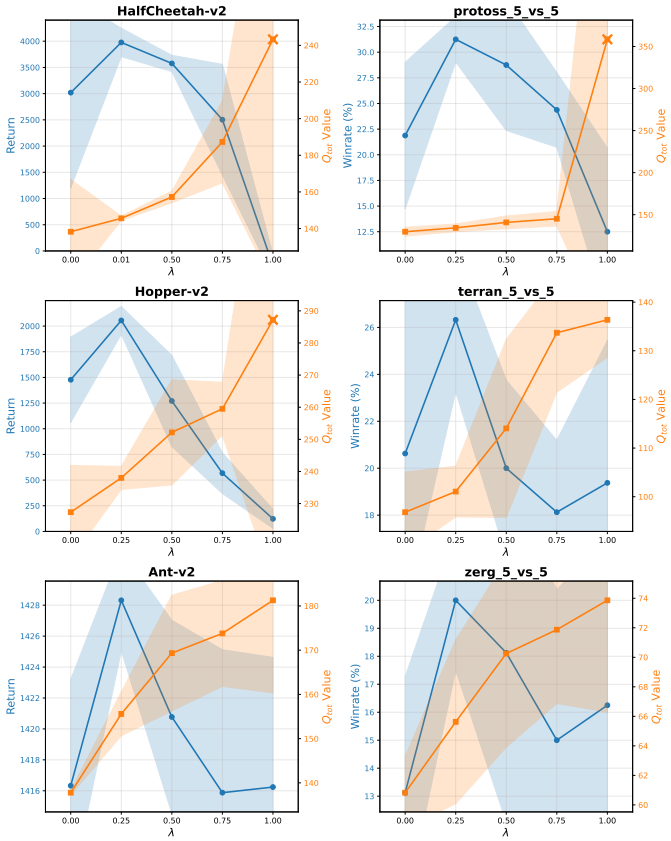

Figure 1: Impact of generalization discount factor λ on performance and Q_{tot} values in different tasks over 5 random seeds. The crosses $\times$ indicate value function divergence under corresponding settings.

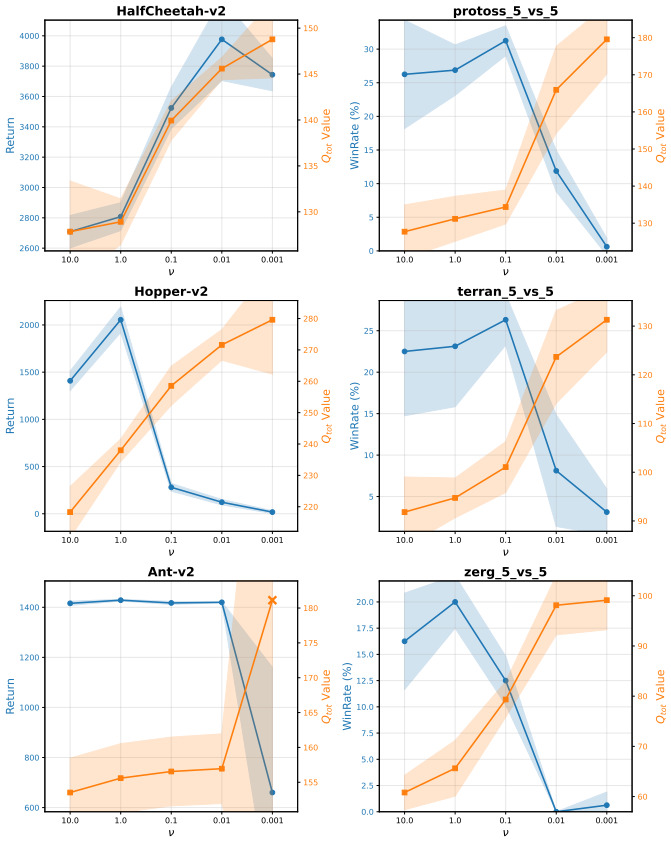

Figure 2: Impact of penalty coefficient ν on performance and Q_{tot} values in different tasks over 5 random seeds. The crosses $\times$ indicate value function divergence under corresponding settings.

a relatively indirect effect on Q_{tot}, it contributes significantly to the stability of the learning process. Consequently, both hyperparameters are indispensable for effective training, as they collectively ensure the accuracy and robustness of the value estimates.

6 CONCLUSION

In this paper, we systematically analyze the limitations of current offline MARL algorithms and their generalization requirements from theoretical and practical perspectives. To address these challenges, we propose OMGMG, which comprises two core components: global moderate action generalization and global moderate generalization propagation, thereby effectively leveraging network generalization. To avoid imposing generalization on each agent individually, we adopt a value decomposition approach from a global perspective, dynamically allocating global generalization capabilities to each agent, thus regulating the intensity of generalization at a macro level. Theoretical analysis demonstrates that OMGMG is mathematically sound, and the moderate use of generalization mechanisms can lead to significant performance improvements. Empirical evaluations further confirm that OMGMG outperforms existing methods in various tasks, achieving state-of-the-art results.

Despite these contributions, this study has some potential limitations. Although OMGMG exhibits robust performance in most environments, its application in discrete action spaces presents challenges due to the inherent discontinuity of the action space, which complicates the fine-grained regulation of global moderate action generalization scales. Future research could focus on developing adaptive generalization scale adjustment mechanisms capable of automatically tuning generalization parameters based on specific task characteristics. Furthermore, bridging the gap between standard benchmarks and real-world deployment remains a critical objective. We plan to investigate the application of OMGMG to high-stakes domains such as autonomous driving, leveraging its moderate generalization framework to enhance safety and robustness in real-world settings.

ACKNOWLEDGMENTS

This work was supported by National Key RD Program of China under Contract 2022ZD0119802 and the Youth Innovation Promotion Association CAS. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution and the Supercomputing Center of USTC.

REFERENCES

[1] Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. 2022. Pessimistic Bootstrapping for Uncertainty-Driven Offline Reinforcement Learning. In *International Conference on Learning Representations*.

[2] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK*, Vol. 96. 195–210.

[3] Tianyuan Chen, Ronglong Cai, Faguo Wu, and Xiao Zhang. 2025. ACTIVE: Offline Reinforcement Learning via Adaptive Imitation and In-sample V-Ensemble. In *International Conference on Learning Representations*.

[4] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. 2023. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 37567–37593.

[5] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 20132–20145.

[6] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*. PMLR, 1587–1596.

[7] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*. PMLR, 2052–2062.

[8] Javier García and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[9] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 750–797.

[10] Longyang Huang, Botao Dong, Wei Xie, and Weidong Zhang. 2024. Offline reinforcement learning with behavior value regularization. *IEEE Transactions on Cybernetics* 54, 6 (2024), 3692–3704.

[11] Bangalore Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on Intelligent Transportation Systems* 23, 6 (2021), 4909–4926.

[12] Jens Kober, James Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.

[13] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

[14] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. 2021. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251* (2021).

[15] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems* 32 (2019).

[16] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

[17] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*. Springer, 45–73.

[18] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. 2021. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*. PMLR, 6120–6130.

[19] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[20] Qian Lin, Chao Yu, Zongkai Liu, and Zifan Wu. 2024. Policy-regularized offline multi-objective reinforcement learning. *arXiv preprint arXiv:2401.02244* (2024).

[21] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems* 30 (2017).

[22] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. 2022. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 1711–1724.

[23] Yi Ma, Jianye Hao, Xiaohan Hu, Yan Zheng, and Chenjun Xiao. 2024. Iteratively refined behavior regularization for offline reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 56215–56243.

[24] Yi Ma, Hongyao Tang, Dong Li, and Zhaopeng Meng. 2023. Reining generalization in offline reinforcement learning via representation distinction. *Advances in Neural Information Processing Systems* 36 (2023), 40773–40785.

[25] Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. 2024. Diffusion-dice: In-sample diffusion guidance for offline reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 98806–98834.

[26] Yixiu Mao, Qi Wang, Chen Chen, Yun Qu, and Xiangyang Ji. 2024. Offline reinforcement learning with ood state correction and ood action suppression. *Advances in Neural Information Processing Systems* 37 (2024), 93568–93601.

[27] Yixiu Mao, Qi Wang, Yun Qu, Yuhang Jiang, and Xiangyang Ji. 2024. Doubly mild generalization for offline reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 51436–51473.

[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[29] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359* (2020).

[30] Thanh Hong Nguyen, Tien Anh Mai, et al. 2024. Comadice: Offline cooperative multi-agent reinforcement learning with stationary distribution shift regularization. In *International Conference on Learning Representations*.

[31] Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[32] Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. 2022. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International Conference on Machine Learning*. PMLR, 17221–17237.

[33] Seohong Park, Qiyang Li, and Sergey Levine. 2025. Flow Q-Learning. In *International Conference on Machine Learning*.

[34] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. 2021. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems* 34 (2021), 12208–12221.

[35] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. 2023. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (2023), 10237–10257.

[36] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51.

[37] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2186–2188.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[39] Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. 2023. Dual rl: Unification and new methods for reinforcement and imitation learning. *arXiv preprint arXiv:2302.08560* (2023).

[40] Letian Wang, Jie Liu, Hao Shao, Wenshuo Wang, Ruobing Chen, Yu Liu, and Steven L Waslander. 2023. Efficient reinforcement learning for autonomous driving with parameterized skills and priors. *arXiv preprint arXiv:2305.04412* (2023).

[41] Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. 2023. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. *Advances in Neural Information Processing Systems* 36 (2023), 52413–52429.

[42] Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. 2021. Uncertainty weighted actor-critic for offline reinforcement learning. *arXiv preprint arXiv:2105.08140* (2021).

[43] Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Victor Wai Kin Chan, and Xianyuan Zhan. 2023. Offline RL with No OOD Actions: In-Sample Learning via Implicit Value Regularization. arXiv:2303.15810

[44] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. 2021. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 10299–10312.

[45] Qingmao Yao, Zhichao Lei, Tianyuan Chen, Ziyue Yuan, Xuefan Chen, Jianxiang Liu, Faguo Wu, and Xiao Zhang. 2025. Offline rl with smooth ood generalization in convex hull and its neighborhood. *arXiv preprint arXiv:2506.08417* (2025).

[46] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[47] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems* 33 (2020), 14129–14142.

[48] Xianyuan Zhan, Xiangyu Zhu, and Haoran Xu. 2022. Model-Based Offline Planning with Trajectory Pruning. In *International Joint Conference on Artificial Intelligence*. 3716–3722.