计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(7): 1854−1868, 2026

面向多智能体强化学习中部分可观测性的主动通信掩码重建方法

贾月洋　刘文彬　王　恩
（吉林大学软件学院　长春　130012）
（符号计算与知识工程教育部重点实验室（吉林大学）　长春　130012）
（yyjia24@mails.jlu.edu.cn）

Active Communication Masked Reconstruction Method for Partial Observability in Multi-Agent Reinforcement Learning

Jia Yueyang, Liu Wenbin, and Wang En
（*College of Software*, *Jilin University*, *Changchun* 130012）
（*Key Laboratory of Symbol Computation and Knowledge Engineering* (*Jilin University*), *Ministry of Education*, *Changchun* 130012）

Abstract　In multi-agent reinforcement learning (MARL), partial observability poses a significant challenge to effective collaboration among agents. Although the centralized training with decentralized execution (CTDE) paradigm alleviates this challenge by leveraging global state information during training, agents must still rely solely on local observations during execution. To address this limitation, existing research has primarily followed two directions. The first direction introduces communication mechanisms to enable information sharing among agents; however, these methods often suffer from high communication overhead and redundant information. The second direction leverages mask reconstruction techniques to infer the missing global information; however, due to sample sparsity, the inferred information depends only on local observations, resulting in low inference accuracy and potential interference with decision-making. To overcome these limitations, we propose an active communication masked reconstruction method. During centralized training, the method dynamically groups agents based on global observation trajectories and allows each agent to select the most relevant communication targets from these groups. During decentralized execution, agents proactively communicate according to their selected target sets, enabling accurate reconstruction of global information and ultimately supporting their decision-making. Extensive experiments conducted on the StarCraft Ⅱ Multi-Agent Challenge (SMAC) demonstrate the effectiveness of the proposed approach in improving coordination and overall performance in partially observable multi-agent scenarios.

Key words　multi-agent reinforcement learning; partial observability; dynamic grouping; multi-agent communication; masked autoencoder; StarCraft Ⅱ Multi-Agent Challenge (SMAC)

摘　要　在多智能体强化学习（multi-agent reinforcement learning, MARL）中，部分可观测性问题严重制约了智能体间的协作。尽管集中训练和分散执行（centralized training and decentralized execution, CTDE）范式通过在训练阶段引入全局状态缓解了该问题，但在执行阶段，智能体仍只能依赖局部观测。为进一步解决该限制，目前有2种方案：一种通过引入通信机制实现智能体间的信息共享，但会面临通信开销

收稿日期：2025-09-21；**修回日期**：2026-03-04
基金项目：国家自然科学基金项目（92567204, 62272193, 62472194）；中央高校基本科研业务费专项资金；吉林省自然科学基金项目（20260101016JJ）
This work was supported by the National Natural Science Foundation of China (92567204, 62272193, 62472194), the Fundamental Research Funds for the Central Universities, and Jilin Science and Technology Research Project (20260101016JJ).
通信作者：王恩（wangen@jlu.edu.cn）

大、信息冗余等问题；另一种则通过掩码重建推理全局信息，但受限于样本稀疏性，重建信息仅来自于单一智能体的局部观测，导致重建精度低，可能会干扰决策。基于此，提出一种主动通信掩码重建方法。在集中训练阶段，该方法根据全局观测轨迹对智能体进行动态分组，并为每个智能体从各组中选取最相关的通信对象，进而在分散执行阶段，智能体依据通信对象集合进行主动通信，从而实现全局信息的准确重建，最终辅助智能体决策。通过在星际争霸多智能体挑战赛（StarCraft Ⅱ Multi-Agent Challenge，SMAC）上的一系列实验，证明了该方法在部分可观测场景下对提升多智能体系统协作与整体策略性能的有效性。

关键词 多智能体强化学习；部分可观测性；动态分组；多智能体通信；掩码自编码器；SMAC

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550701 **CSTR:** 32373.14.issn1000-1239.202550701

多智能体强化学习（multi-agent reinforcement learning，MARL）通过学习多智能体与环境之间的交互，以优化个体或整体的策略表现[1-3]。近年来，该算法在自动驾驶[4]、订单调度[5-6]、机器人系统[7-8]、游戏AI系统[9]等多个领域得到了广泛应用。然而，由于环境的部分可观测性，每个智能体只能获取局部观测信息，难以感知全局状态，这在很大程度上制约了智能体之间的协作与整体策略性能。因此，解决部分观测性问题是多智能体强化学习中提升智能体间协作与整体策略性能的关键挑战[3,10]。

为了应对该挑战，最直接有效的方式是采用中心化思想对所有智能体进行统一训练与控制，即采用集中训练和集中执行（centralized training and centralized execution，CTCE）框架[11]。然而，随着智能体数量的增加，状态-动作联合空间的维度将呈指数级增长，进而引发维度灾难，导致系统面临计算成本高、扩展性低等问题，难以适应复杂的多智能体场景[2]。集中训练和分散执行（centralized training and decentralized execution，CTDE）框架的提出缓解了上述问题[12]。虽然该框架允许在训练阶段访问全局状态，但在执行阶段，智能体仍只能依赖局部观测。因此，部分可观测性所带来的挑战仍未彻底解决。

现有2种针对缓解分散执行阶段缺乏全局信息的策略：一种通过引入通信机制，允许智能体之间在执行阶段共享观测信息，通过消息传递的方式直接获取其他智能体的局部观测，从而模拟全局状态以获取全局视角[13]。该方法虽然有利于智能体间的协作，但往往面临带宽有限、通信冗余等问题。另一种则通过掩码重建，根据智能体局部观测推理全局信息[14]。此方法虽然不依赖于通信，但受样本稀疏性的影响，推理过程仅依赖单一智能体的局部观测，导致重建精度存在偏差，使得智能体决策容易陷入局部最优。尽管可以利用融合思想，在现有掩码重建框架中引入通信机制以缓解样本稀疏性问题，但由于环境是动态变化的，因此难以满足智能体对于全局信息重建的动态通信需求。

为此，本文提出了一种主动通信掩码重建方法。在保持低通信成本的前提下，在训练阶段利用基于谱聚类（spectral clustering）[15]的动态分组机制对智能体进行自适应的多样化分组，实现对智能体间关系的动态建模，并通过掩码交叉注意力（masked cross-attention）机制[16]从各组中挑选最相关的通信对象，从而为每个智能体构建通信对象集合；在执行阶段，每个智能体依据通信对象集合与其他智能体建立通信，从而为智能体引入少量且高度相关的多样化观测信息。随后，结合掩码自编码器（masked autoencoder，MAE）[17]的推理能力，利用整合后的局部观测信息来重建全局信息，从而为智能体构建更准确的全局视角，进一步提升多智能体系统的协作与整体决策性能。本文的主要贡献有4个方面：

1）提出了一种基于谱聚类的动态分组机制。在集中训练阶段，能够根据全局观测轨迹自适应调整分组，并为智能体确定通信对象，从而为后续智能体在执行阶段进行主动通信提供有效指导。

2）提出了一种融合主动通信与掩码重建的重建思想。在保持低通信成本的前提下，突破传统仅依赖局部观测的限制，通过引入少量且相关的多样化外部信息辅助智能体构建全局信息，有效缓解传统掩码重建中因样本稀疏导致的重建误差问题，以提升多智能体系统的协作与整体决策性能。

3）设计了一种主动通信掩码重建框架。该框架利用基于动态分组的主动通信模块，为智能体级掩码自编码器提供多样化样本，以实现全局信息的精确重建，进而通过自注意力机制从重建的全局信息

中提取全局特征，从而帮助智能体有效决策。

4）在星际争霸多智能体挑战赛（StarCraft Ⅱ Multi-Agent Challenge, SMAC）[18]上进行了一系列实验评估。综合结果显示，所提方法在多个场景下取得了良好的性能提升，验证了该方法在部分可观测场景下的有效性。

1 相关工作

在多智能体强化学习中，如何在部分可观测环境下实现高效的协同决策一直是研究的主要问题。现有研究主要从缓解部分可观测性出发，围绕多智能体间通信机制与掩码重建2种方法进行展开。同时，分组机制作为一种辅助手段，可为实现高效协作提供支持。因此，本文将从以下4个方面进行详细介绍。

1.1 部分可观测性

在多智能体系统中，环境通常表现为部分可观测的，每个智能体只能获取自身观测范围内的有限信息。因此，在早期的分散训练和分散执行（decentralized training and decentralized execution, DTDE）框架中，每个智能体通过将其他智能体视为环境的一部分进行独立学习[19]。但由于其他智能体的策略是动态变化的，因此系统是不稳定的[20]。为了缓解非平稳问题，在这之后的许多研究开始采用中心化的思想，即集中训练和集中执行（centralized training and centralized execution, CTCE）框架对所有智能体进行统一训练和控制[11]。然而，智能体数量的增加会引发维度灾难，导致系统面临计算成本高、扩展性低等问题[2]。为了提高系统的适应性，CTDE框架被广泛应用，并已成为如今最流行的框架[12]。该框架下有2种主要的MARL算法：基于值函数的方法和基于策略的方法。基于值函数的方法以值函数分解为导向，将联合值函数因式分解为多个单独的局部值函数，经典方法有VDN[21]，QMIX[22]等；基于策略的方法以学习参数化策略为导向，其基于Actor-Critic结构，经典方法有A3C[23]，DDPG[24]等。尽管基于策略的方法能够处理连续动作空间，但与基于值函数的方法相比，它通常表现出更高的方差和更低的样本效率，并且可能收敛于局部最优值。因此，为了方便设计，本文基于值函数方法进行框架设计。

1.2 分组机制在强化学习中的相关工作

在多智能体强化学习研究过程中，分组机制作为提升智能体间信息交互效率以及捕获和建模智能体间关系的重要手段，近年来受到了广泛的关注。该类方法通过将智能体划分为若干子集，在保证组间多样性的同时，使组内智能体之间能够高效的信息共享。早期研究如Q-decomposition[25]，MACKRL[26]等方法通常依赖于先验知识实现智能体划分。随后，一部分研究如M^3RL[27]，ALMA[28]等方法尝试通过任务分配实现隐式动态分组。此外，以ROMA[29]和RODE[30]为代表的角色分解方法通过学习智能体在不同情境下的角色策略，实现上下文相关的功能划分，这类方法虽然能提高多智能体间协作的灵活性和适应性，但需要手动设定角色数量或动作空间划分，难以充分适应复杂多变的动态环境。近年来，以数据为驱动的方法如GoMARL[31]利用智能体之间的相似性度量实现自动分组，尽管该方法能够在动态环境下进行自适应调整，但通常缺乏可解释性。因此，为了兼顾自适应性与可解释性，受HYGMA[32]工作的启发，本文提出一种基于谱聚类的动态分组机制。通过使用具有严格数学基础的谱聚类方法，对智能体观测轨迹进行谱分析，可以发现智能体间复杂的潜在关系，从而实现有效的划分。同时，通过引入CH（Calinski-Harabasz）指数[33]，可以实现在动态环境下评估组内紧密度与组间多样性，从而实现对聚类数的自适应调整，能够为后续主动通信提供多样化且最优的分组参考。

1.3 通信机制在强化学习中的相关工作

为了缓解智能体在分散执行阶段的信息缺失问题，现有研究尝试利用通信机制学习有效的通信策略来实现多智能体之间信息的共享，从而辅助独立决策。早期方法如RIAL和DIAL[34]引入了可学习的通信信道，为通信建模奠定了基础。CommNet[35]则提出了基于消息聚合的共享机制，通过广播实现状态信息的融合。随后，ATOC[36]，IC3Net[37]等方法进一步引入门控机制和注意力结构，在减轻通信冗余的同时增强了通信选择的灵活性。近年来，部分工作开始关注有向通信结构与消息选择内容的建模，如G2ANet[38]利用图神经网络动态指定通信目标，TarMAC[39]，MASIA[40]等则试图提升消息的表达性与接收端的信息利用效率。虽然这些方法能够有效缓解部分可观测性，但由于训练过程中高度依赖稀疏的强化学习信号，导致通信内容与策略联合优化存在训练不稳定的问题，并且大规模通信会增加系统的复杂性与计算负担。与上述方法不同，本文采用了一种相对轻量的主动通信机制，该机制依据训练阶段所获得的分组结果，通过掩码交叉注意力机制为每个智能

体从各组中选取最相关的其他智能体建立通信，从而在执行阶段为智能体提供少量且高相关性的多样化观测信息。这样不仅避免了过多的通信开销，同时能为后续的掩码重建提供多样化的样本，从而有效地缓解样本稀疏性问题。

1.4 掩码重建在强化学习中的相关工作

受自然语言处理[41]、计算机视觉[17]等领域中自监督学习掩码重建思想的启发，近年来，许多研究将掩码重建思想引入到了强化学习任务中，使得智能体在分散执行阶段能够通过推理全局信息获取全局视野。在早期的单智能体强化学习中，MaskDP[42]，MTM[43]等方法通过对输入轨迹进行随机遮蔽并重建原始信息，有效提升了策略的泛化能力。而在多智能体场景下，部分研究也尝试将掩码重建用于缓解部分可观测性问题。例如，MA2CL[44]利用对比损失联合训练策略网络与重建模块，以增强智能体间协作；MA^{2}E [45]则采用独立训练的掩码自编码器，将局部观测重建为全局观测信息以辅助策略学习。尽管上述方法在全局信息重建方面取得了一定成效，但由于输入样本的稀疏性，即每个智能体仅依赖自身的局部观测进行推理，导致难以准确重建复杂环境中的全局信息。为此，本文在MA^{2}E工作的基础上，提出了结合主动通信机制的掩码重建方法，通过为模型引入少量且高度相关的多样化的外部观测信息，引导模型更准确地重建全局信息，从而更有效地辅助决策。

2 方　　法

在本节中，我们将系统性地介绍所提出的主动通信掩码重建方法。首先，我们对多智能体强化学习问题进行了形式化定义。随后，我们详细介绍了该方法的整体框架，重点阐述了框架的整体结构组成和处理流程。此外，我们对框架中关键的基于谱聚类的动态分组机制、主动通信模块和智能体级掩码自编码器进行了深入解析。最后，我们讲解了整个训练过程的具体步骤，并介绍了所采用的损失函数与优化目标。

2.1 问题定义

我们将多智能体强化学习问题建模为一个分布式部分可观测的马尔可夫决策过程（decentralized partially observable Markov decision process, Dec-POMDP）[46]，具体来讲，Dec-POMDP可以被定义为一个元组$\mathcal{G}=(\mathcal{N},\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\mathcal{Z},\mathcal{O},\gamma)$，其中$\mathcal{N}$表示$n$个智能体的集合，$\mathcal{S}$表示全局状态信息的集合，$\mathcal{A}$表示智能体动作的集合，$\mathcal{P}:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$是状态转移函数，其中$\Delta(\mathcal{S})$是$\mathcal{S}$上概率分布的集合，$\mathcal{R}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$表示奖励函数，$\mathcal{Z}$表示智能体的观测集合，$\mathcal{O}:\mathcal{S}\times\mathcal{N}\to\mathcal{Z}$表示观测函数，$\gamma\in[0,1)$表示折扣因子。在每个时间步$t$，每个智能体$i\in\mathcal{N}$从观测函数$\mathcal{O}(\boldsymbol{s}_t,i)$接收一个观测$\boldsymbol{o}_t^i\in\mathcal{Z}$，其中$\boldsymbol{s}_t\in\mathcal{S}$表示在时间步$t$下的全局状态，每个智能体有自己的策略$\pi^i(\boldsymbol{a}_t^i|\boldsymbol{\tau}_t^i)$，其中$\boldsymbol{a}_t^i$是每个智能体的动作，$\boldsymbol{\tau}_t^i=(\boldsymbol{o}_1^i,\boldsymbol{a}_1^i,\cdots,\boldsymbol{o}_{t-1}^i,\boldsymbol{a}_{t-1}^i,\boldsymbol{o}_t^i)$是每个智能体的观测-动作的历史轨迹，环境会执行联合动作$\boldsymbol{a}_t$以进入新的状态$\boldsymbol{s}_{t+1}\sim\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$，并且多智能体系统会收到一个整体的全局奖励$r_t=\mathcal{R}(\boldsymbol{s}_t,\boldsymbol{a}_t)$，最终的目标是找到一个联合策略$\pi=(\pi^1,\pi^2,\cdots,\pi^n)$以最大化期望折扣总奖励$\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r\right]$。在本文中，我们使用基于值函数的方法去更新策略，在分散执行阶段，每个智能体都会学习一个个体Q网络$Q^i(\boldsymbol{\tau}^i,\boldsymbol{a}^i)$，之后在集中训练阶段，会通过一个混合网络$\mathcal{M}_\theta$将每个智能体的个体$Q$值聚合为一个联合$Q$值，即$Q_{\text{tot}}(\boldsymbol{\tau},\boldsymbol{a})=\mathcal{M}_\theta(Q^1(\boldsymbol{\tau}^1,\boldsymbol{a}^1),Q^2(\boldsymbol{\tau}^2,\boldsymbol{a}^2),\cdots,Q^n(\boldsymbol{\tau}^n,\boldsymbol{a}^n);\boldsymbol{s})$，混合网络的参数$\theta$通过最小化时序差分（temporal difference, TD）损失函数进行学习：

$$\mathcal{L}(\theta)=\sum_{i=1}^{b}\left[(y_i^{\text{tot}}-Q_{\text{tot}}(\boldsymbol{\tau}_t,\boldsymbol{a}_t,\boldsymbol{s}_t;\theta))^2\right],\tag{1}$$

其中$y^{\text{tot}}=r+\gamma\max_{\boldsymbol{a}_{t+1}}Q_{\text{tot}}(\boldsymbol{\tau}_{t+1},\boldsymbol{a}_{t+1},\boldsymbol{s}_{t+1};\theta^-)$，$\theta^-$是目标网络的参数，$b$是重放缓冲区$\mathcal{D}$中采样的批量大小。

2.2 整体框架

本文方法在QMIX框架的基础上进行构建，在继承CTDE范式的同时，保证了值函数分解的单调性IGM（individual-global-max）条件[21]。方法的整体框架如图1所示，其中框架的整体结构包含n个智能体的个体Q网络和1个混合网络。在混合网络中，超网络会根据全局状态为混合网络生成权重和偏置，用于混合个体Q值，从而得到联合Q值。对于个体Q网络，该网络依据由主动通信模块所生成的通信对象集合C_t^i进行通信，从而为智能体级掩码自编码器提供指导，帮助智能体重建全局信息。

然而，由于重建的全局信息中并不全是有效信息，因此我们通过自注意（self-attention）机制[16]从重建得到的全局信息中筛选重要信息，以获取全局特征，从而帮助智能体学习一个跨智能体的全局协作模式。在这之前，为避免重建误差对注意力权重分配的干扰，还需要将当前智能体重建表示替换为真实值，以确保特征查询过程的准确性。首先，通过自注意力机制计算智能体之间的注意力权重：

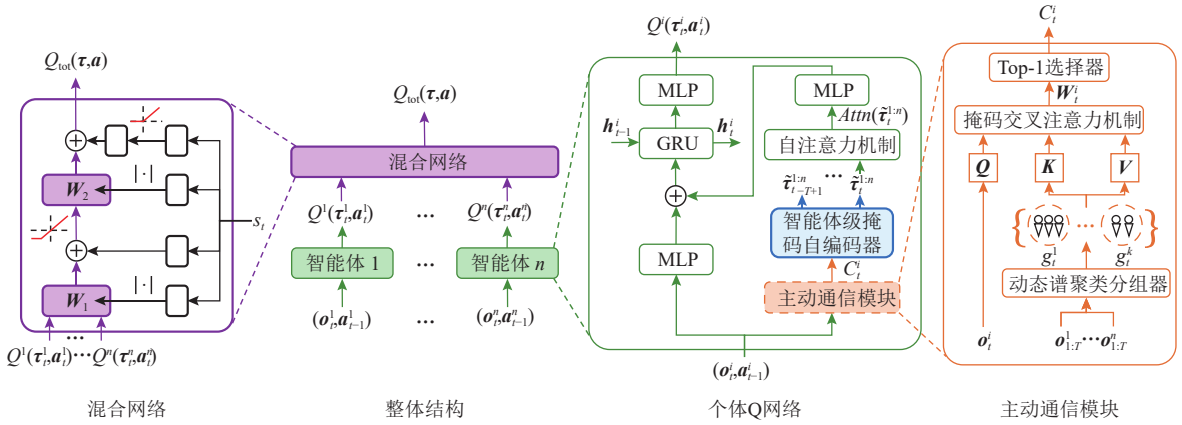


Fig. 1 Overall framework of our method

图 1 本文方法整体框架

$$\alpha_{i,j}=\frac{\exp(\tilde{\boldsymbol{\tau}}_i^{\mathrm{T}}\boldsymbol{W}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{q}}\tilde{\boldsymbol{\tau}}_j)}{\sum_{j=1}^{n}\exp(\tilde{\boldsymbol{\tau}}_i^{\mathrm{T}}\boldsymbol{W}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{q}}\tilde{\boldsymbol{\tau}}_j)}\,, \tag{2}$$

其中$\boldsymbol{W}_{\mathrm{q}}$和$\boldsymbol{W}_{\mathrm{k}}$分别是查询和键的可学习权重矩阵。然后再计算注意力权重和值的加权和得到加权值:

$$Attn(\tilde{\boldsymbol{\tau}}_t^{1:n})=\sum_{j=1}^{n}\alpha_{i,j}\boldsymbol{W}_{\mathrm{v}}\tilde{\boldsymbol{\tau}}_t^{j}\,, \tag{3}$$

其中$\boldsymbol{W}_{\mathrm{v}}$是值的可学习权重矩阵。最终,将该权重加权值与智能体自身观测相结合,用于最终的决策。

2.3 基于谱聚类的动态分组机制

在集中训练阶段,我们将n个智能体在T个时间步的观测轨迹$\{\boldsymbol{o}_{t-T+1:t}^{i}\}_{i=1}^{n}$作为输入,以此构成轨迹矩阵$\boldsymbol{O}_t\in\mathbb{R}^{b\times T\times d}$,其中$b$表示批量大小,$T$表示轨迹的时间步长,$d$表示轨迹维度。为了在保持组内紧密度的同时保持组间多样性,会通过最小化规范化切割(normalized cut, cut_{N})[47]来实现智能体分组,从而得到分组$G_t=\{g_t^1,g_t^2,\dots,g_t^k\}$:

$$cut_{\mathrm{N}}(g_t^1,g_t^2,\dots,g_t^k)=\sum_{i=1}^{k}\frac{cut(g_t^i,\bar{g}_t^i)}{vol(g_t^i)}\,, \tag{4}$$

其中k表示聚类数目,我们预定义初始聚类数目为$k=\lceil n/2\rceil$,而n表示智能体总数;g_t^i表示在第t个时间步下的第i个分组;$cut(g_t^i,\bar{g}_t^i)$表示分组g_t^i到其他群体的边的总权重,间接反映组间多样性;$vol(g_t^i)$则表示群体g_t^i内所有节点的度数之和,间接反映组内紧密度。但在离散划分空间里求最优解是NP难问题,因此需要采用谱松弛(spectral relaxation)[48]将问题转化为连续优化问题:

$$\arg\min_{\boldsymbol{H}\in\mathbb{R}^{n\times k}}\mathrm{tr}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}),\ \mathrm{s.t.}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}\,, \tag{5}$$

其中$\boldsymbol{H}\in\mathbb{R}^{n\times k}$表示特征矩阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$表示归一化后的拉普拉斯矩阵,$\boldsymbol{D}$是度矩阵,而$\boldsymbol{W}$是邻接矩阵,用于表示智能体观测轨迹之间的相似度。具体而言,$\boldsymbol{W}$是根据欧几里得距离计算K近邻所得到的,若智能体j是智能体i的K近邻,则$W_{ij}>0$;否则$W_{ij}=0$。最终,谱聚类得到的最优分组G_t^*满足:

$$cut_{\mathrm{N}}(G_t)\leqslant O(\log k)\cdot cut_{\mathrm{N}}(G_t^*)\,。 \tag{6}$$

为了能够适应环境的动态变化,找到最终的最佳分组,我们引入CH(Calinski-Harabasz)指数[33]用于评估聚类后的组内紧密度与组间多样性,从而不断调整聚类数目k以寻找最佳智能体分组:

$$CH_t^k=\frac{\sum_{i=1}^{k}|g_t^i|(\mu_t^i-\mu_t)^2}{\sum_{i=1}^{k}\sum_{x_t^j\in g_t^i}(x_t^j-\mu_t^i)^2}\cdot\frac{n-k}{k-1}\,, \tag{7}$$

其中x_t^j是分组g_t^i中的第j个样本,μ_t^i是分组g_t^i的质心,表示组内样本均值,$x_t^j-\mu_t^i$则表示样本与质心之间的偏差,而μ_t表示在第t个时间步下n个智能体的样本均值。至此,根据计算得到的CH值的大小可以评估聚类数目k的优劣,从而在集中训练阶段对聚类数目k以及分组G_t进行更新:

$$\begin{aligned}&k_{\mathrm{best}}=\arg\max_{\substack{k'\in\{k-1,k,k+1\}\\ k_{\min}\leqslant k'<k_{\max}}}CH_t^{k'}\,,\\ &\mathrm{s.t.}\ CH_t^{k'}-CH_t^{k_{\mathrm{current}}}>\delta\ \text{且}\ CH_t^{k'}>CH_t^{\mathrm{best}}\,,\end{aligned} \tag{8}$$

其中k有2种选择:$k-1$和$k+1$。分别计算2种选择下的分组CH值,以此寻找最佳的CH值,即CH_t^{best},从而得到对应的最佳分组G_t^*。

2.4 主动通信模块

通过动态分组机制获得分组结果G_t(最终固定为最佳分组G_t^*)以后,为后续智能体级掩码自编码器提供多样化的样本,在训练阶段,我们还将采用掩码交叉注意力(masked cross-attention)机制[16]为每个智

能体从各组中选取最相关的其他智能体对象，从而确立通信。对于掩码交叉注意力机制，令当前智能体观测作为查询，每个分组中的智能体观测作为键和值，从而计算注意力权重加权值：

$$CrossAttn(\boldsymbol{o}_t^i, g_t^j) = softmax\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}} + \boldsymbol{M}\right)\cdot \boldsymbol{V}, \quad (9)$$

其中$\boldsymbol{o}_t^i$表示第i个智能体的局部观测，g_t^j表示第j组的智能体观测，同时我们加入掩码矩阵$\boldsymbol{M}$用以防止智能体在确定通信对象时选择自身。在计算得到权重加权值以后，我们利用 Top-1 选择器选取每组中权重值最高的智能体作为通信对象，从而为每个智能体构建通信对象集合$C_t^i \subseteq \{1,2,\cdots,n\}\backslash\{i\}, \left|C_t^i\right| = c$，进而在执行阶段为智能体主动通信提供有效参考，缓解智能体级掩码自编码器所面临的样本稀疏性问题。

2.5 智能体级掩码自编码器

为实现对其他智能体信息的有效重建，本文引入了 MA^{2}E[45] 工作中的智能体级掩码自编码器结构。其中，掩码自编码器的输入为每个智能体在T个时间步内的观测-动作历史轨迹，其包含的编码器和解码器由多头注意力机制（multi-head attention, MHA）和前馈神经网络（feed-forward network, FFN）组成，从而能够有效捕捉轨迹中的智能体之间的依赖关系与相关性。需要注意的是，该掩码自编码器的解码器中并未使用传统 Transformer[16] 的掩码多头注意力机制（masked MHA），而是在输入阶段额外引入了独立的智能体级掩码层来实现遮蔽处理。该掩码层能够在输入轨迹中对智能体的整个信息进行选择性的遮蔽，从而能够更好地模拟部分可观测环境。考虑到集中训练阶段与分散执行阶段在可访问信息上的差异，本文针对这 2 个阶段分别设计了不同的掩码策略，以适应训练与执行的需求。

在集中训练阶段，系统可以访问所有智能体的观测-动作历史轨迹$\boldsymbol{\tau}_{t-T+1:t}^{1:n} = \left(\boldsymbol{\tau}_{t-T+1}^{1:n}, \boldsymbol{\tau}_{t-T}^{1:n}, \cdots, \boldsymbol{\tau}_t^{1:n}\right)$，其中$\boldsymbol{\tau}_t^i = (\boldsymbol{o}_t^i, \boldsymbol{a}_{t-1}^i)$表示单个智能体的观测-动作历史轨迹。我们采用随机掩码策略[43]，即随机选择m个智能体，并将其全部的历史轨迹进行掩蔽，迫使模型从未遮蔽的轨迹中恢复完整信息$\tilde{\boldsymbol{\tau}}_{t-T+1:t}^{1:n}$，以便在分散执行阶段为智能体推断出更准确的信息。

在分散执行阶段，智能体仅能获取自身的局部观测信息，无法直接访问其他智能体的观测轨迹。为了缓解样本稀疏性，我们通过基于动态分组的主动通信模块所生成的通信对象集合C_t^i进行通信，从而为智能体级掩码自编码器提供多样化的样本，帮助智能体级掩码自编码器更好地推断全局信息，智能体级掩码自编码器的具体设计如图 2 所示。在构建输入轨迹时，除了包含当前智能体的历史轨迹以外，还应该包含通信对象的历史轨迹$\tau_{t-T+1:t}^{C_t^i} = \{\boldsymbol{\tau}_{t-T+1:t}^j \mid j \in C_t^i\}$，其中，通信对象的历史轨迹会在每次通信后进行更新，以确保当前时间步为最新信息。最终，为了在掩码层能够模拟智能体分散执行时的部分可观测性，仅保留当前智能体及其通信智能体的历史轨迹，其余z个未观测智能体的历史轨迹信息均被掩蔽。

同时，为了能够区分多个智能体的历史轨迹，该掩码自编码器在构建位置编码时，同时引入了时间和智能体索引以更好地捕捉多智能体的时序交互特征。

2.6 训　练

本文方法中的智能体级掩码自编码器与智能体的训练是相互独立的，其中掩码自编码器模型的训练遵循预训练和微调[17]2 个阶段。在预训练阶段，我

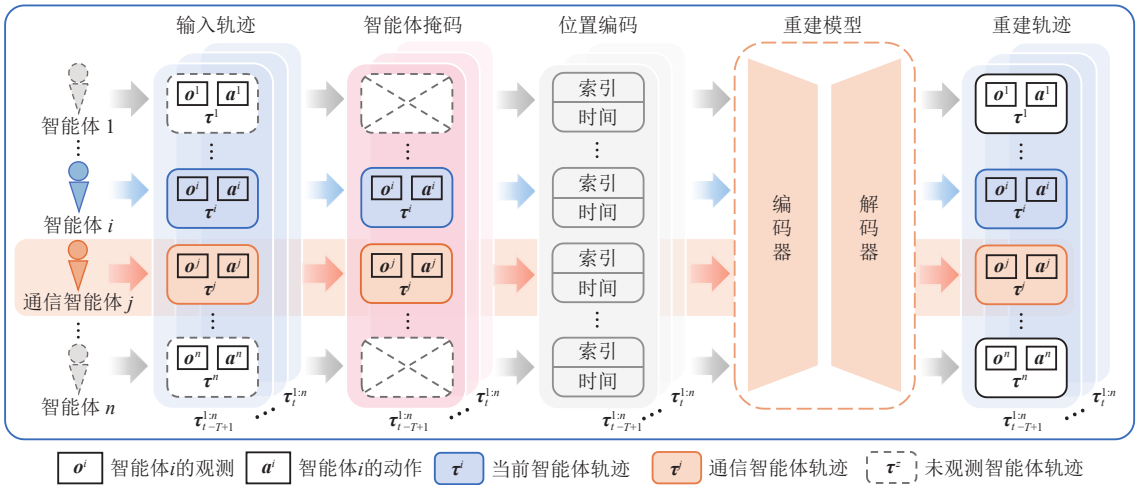


Fig. 2 Agent-level masked autoencoder combined with active communication

图 2 结合主动通信的智能体级掩码自编码器

们的目的是为每个智能体预先训练出一个可以直接使用的重建模型,因此在智能体策略训练开始之前,会先与环境交互收集初始样本,以进行随机掩码,从而确保自编码器能够正确推断全局信息。该模型的参数通过最小化真实历史轨迹与重建轨迹之间的均方误差(mean squared error, MSE)来进行学习:

$$\mathcal{L}_{\text{rec}}=\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(\boldsymbol{\tau}_t^i-\tilde{\boldsymbol{\tau}}_t^i)^2\text{。} \tag{10}$$

预训练结束以后,在智能体策略训练的过程中,模型会周期性地进行微调以实现进一步的优化,微调操作具体参考 2.5 节中的集中训练阶段。

2.7 复杂度分析

本节从时间复杂度和空间复杂度 2 个角度对方法进行了定量分析。为了方便表达,设n表示智能体的数量,T表示每个智能体的历史轨迹长度,$k<n-1$表示聚类数目,d表示自编码器中嵌入向量维度,f表示每个时间步的观测-动作向量维度,h表示自编码器中多头注意力头数,b表示批量大小。

时间复杂度主要包含基于谱聚类的动态分组机制、主动通信、智能体级掩码自编码器以及自注意力机制 4 部分。在训练阶段,由于引入了基于谱聚类的动态分组机制,需要对智能体轨迹构建 K 近邻相似图并进行谱松弛,其复杂度约为 $O(n^2Td+n^2k)$,之后,每个智能体需通过主动通信模块构建通信对象集合,其复杂度约为 $O(nd^2)$;同时,智能体级掩码自编码器在集中可见的条件下可处理的序列长度为nT,则单层复杂度为 $O(hnTd^2)$,又因为自编码器中包含多层多头注意力机制,因此,当自编码器中的编码器和解码器总共有L层时,其复杂度为 $O(LhnTd^2)$。而在执行阶段,每个智能体会依据通信对象集合与其他智能体建立通信,则最终输入到智能体级掩码自编码器中的序列长度为$(r+1)\cdot T$,其中$r\leqslant k$,则掩码自编码在执行阶段的复杂度为 $O(nLh(r+1)Td^2)$;此外,重建后的全局信息需经过自注意力机制筛选和混合 Q 网络更新,复杂度分别为 $O(nd^2)$和 $O(nd)$。综上所述,相较于传统全通信方法,本文方法在保证通信有效性的同时,将全局信息的重建规模从与n成正比下降至与r成正比,因此更适用于大多数中等规模的多智能体任务。

对于空间复杂度,本文方法的主要开销来自于智能体级掩码自编码器的模型参数存储 $O(Lhd^2)$、历史轨迹缓存 $O(bnTf)$、训练阶段谱聚类中矩阵与向量缓存 $O(nK+nk)$(其中 K 为近邻数),以及执行阶段通信样本缓存 $O(n(r+1)Tf)$。总体而言,相比于全通信方法,本文方法在保证全局信息重建精度的同时,有效降低了执行阶段的通信所需的存储开销。

3 实验与结果分析

3.1 实验设置

为了验证本文方法在部分可观测场景下的有效性,我们选择在 SMAC 平台上进行实证评估。SMAC 是基于暴雪公司旗下的经典即时战略游戏《星际争霸Ⅱ》开发的多智能体强化学习测试平台,它提供了多个专门的多智能体协作场景,能够有效评估方法在部分可观测性下的性能表现[18]。

我们分别在 Easy, Hard, Super Hard 这 3 个难度等级的场景地图上进行了测试,选择的地图及地图设置如表 1 所示。其中 Easy 难度选取了具有异构特性的地图 2s3z,该地图中的友军单位和敌军单位包含 2 名狂战士和 3 名潜行者 2 种异构单位,体现了不同单位属性的组合,用于验证方法在智能体间存在异构差异时的协同与微观操作能力;Hard 难度地图选取了具有代表性的地图 5m_vs_6m,该地图包含 5 名海军陆战队友军单位和 6 名海军陆战队敌军单位,具有一定的不对称性与协同难度,可用于评估方法在同质单位条件下的协作效果;而 Super Hard 难度选取的地图是 6h_vs_8z,该地图由 6 名水蛛友军远程单位和 8 名狂战士敌军近战单位组成,具备敌我异构与不对称 2 个特性。该场景的策略方向较为凸显,可用于评估方法在高难度条件下的策略调整能力,并以此侧面评估整体性能。

Table 1 The Map Settings

表 1 地图设置

难度等级	地图名称	友军单位类型及数量	敌军单位类型及数量	特点
Easy	2s3z	2 名狂战士,3 名潜行者	2 名狂战士,3 名潜行者	异构,对称
Hard	5m_vs_6m	5 名海军陆战队	6 名海军陆战队	同构,不对称
Super Hard	6h_vs_8z	6 名水蛛(远程)	8 名狂战士(近战)	异构,不对称,策略凸显

我们将本文方法与经典的 CTDE 基线算法 QMIX 和仅利用单个智能体的局部观测进行掩码重建的多智能体重建方法 MA^{2}E 进行比较。每组实验运行 3 次,时间步数设置为 2E6。为了全面评估各方法的性能表现,我们采用平均胜率和平均回报作为评估指标。平均胜率作为多智能体强化学习中被广泛使用,且具备较强可比性和通用性的指标,能够充分反

映方法的整体表现，在 SMAC 平台中，若在限定时间步内，友方单位能成功击败所有敌方单位，则判定为胜利，否则判定为失败，最终统计获胜回合数占总实验回合数的比值作为胜率。同时，我们还在胜率曲线上通过设置下 25% 和上 75% 的分位数区域用以表示每种方法的鲁棒性。而平均回报则是智能体在若干评估回合中获得的累计奖励的平均值，相比于平均胜率，该指标能够更好地衡量智能体在整个过程中的综合表现，展现策略的优劣。

需要注意的是，由于许多 MARL 方法都能从调整的超参数中表现出优异的性能[49]，因此，为了确保实验的可对比性和严谨性，我们采用了对齐超参数的实验策略来确保所有方法使用相同的超参数进行对比，超参数设置如表 2 所示。其中，考虑到 6h_vs_8z 的地图特性，为了能更加公平地评估各方法在该地图上的性能表现，我们对部分超参数进行了相应的调整，具体而言，我们将 $TD(\lambda)$ 设置为 0.3，同时将退火时间设置为 500 000，以确保各方法能够在该地图上发挥良好。

Table 2 The Hyperparameter Settings

表 2 超参数设置

超参数名称	变量符号	取值
缓冲区大小	*buffer_size*	1 000
批量大小	*batch_size*	128
批量运行大小	*batch_size_run*	8
运行器	*runner*	parallel
优化器	*optimizer*	Adam
学习率	*lr*	0.001
$TD(\lambda)$	*td_lambda*	0.6/0.3
隐藏状态维度	*rnn_hidden_dim*	128
混合嵌入维度	*mixing_embed_dim*	32
超网络嵌入维度	*hypernet_embed*	64
γ	*gamma*	0.99
动作选择器	*action_selector*	epsilon_greedy
退火时间	*epsilon_anneal_time*	100 000/500 000

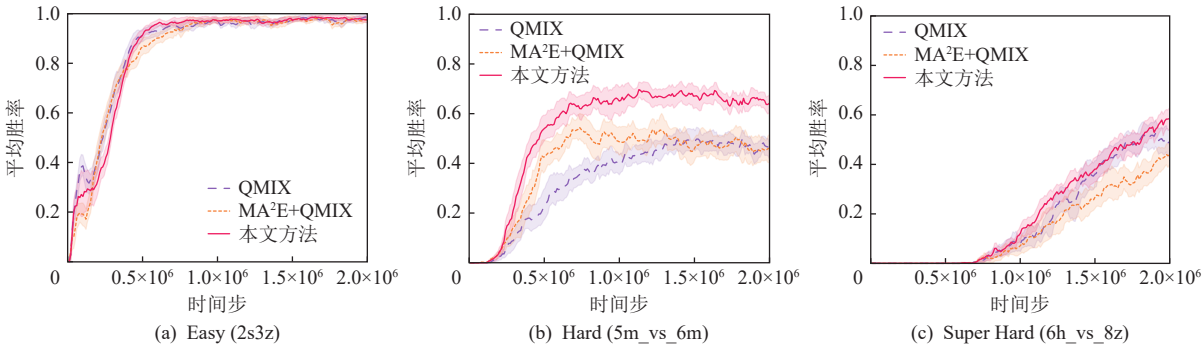
注：$TD(\lambda)$ 和退火时间中斜杠后的值专门用于 6h_vs_8z 地图。

3.2 平均胜率对比分析

平均胜率对比结果如图 3 所示。本文方法在整个训练过程中表现出较强的性能优势，在 3 种不同难度等级场景下的最终平均胜率均高于 MA^{2}E 和 QMIX 这 2 种方法。其中，在 5m_vs_6m 地图中，本文方法在结果上的提升最为明显，最终平均胜率能够稳定在约 67%，相较于 MA^{2}E 方法性能直接提升了 19%；而在 6h_vs_8z 地图中，本文方法依旧能保持稳定的性能，在整个训练过程中，其平均胜率整体上高于性能较优的 QMIX 方法。对于 MA^{2}E 方法，受样本稀疏性的影响，智能体仅能依赖自身局部观测重建全局信息，使得重建结果准确性偏低，这种对全局信息认知的偏差导致其在各个场景中的平均胜率均低于本文方法。然而，本文方法在此基础上引入了主动通信机制，为每个智能体的掩码自编码器提供了额外的多样化信息，从而缓解了样本稀疏问题，帮助重建更准确的全局信息，进而帮助智能体实现更有效的决策，最终提升作战胜率。对于鲁棒性，从胜率曲线的下 25% 及上 75% 的分位数可观察到本文方法在保证稳定性的前提下能够实现更高的胜率。

相比之下，传统的 QMIX 方法在执行阶段仅允许智能体依赖自身局部观测进行决策，因而缺乏对全局信息的重建能力，使得 QMIX 容易陷入局部最优，在 5m_vs_6m 上表现相对较弱。虽然在 6h_vs_8z 地图中因策略凸显使其性能表现较好，但由于缺乏对全局的感知能力，使得 QMIX 不能更有效地学习，导致综合性能弱于本文方法。

为了更加直观地展示各方法在不同场景下的性能差异，我们统计并比较了各方法在后 1E6 时间步内的平均胜率作为最终胜率，结果如表 3 所示。从



(a) Easy (2s3z) (b) Hard (5m_vs_6m) (c) Super Hard (6h_vs_8z)

Fig. 3 Average win rate comparison on SMAC map scenarios

图 3 SMAC 地图场景下的平均胜率对比

结果中可以看出，本文方法的最终胜率均优于 MA^{2}E 和 QMIX 这 2 种方法，说明了本文方法能够帮助智能体学到更有效的策略，从而提高作战胜率，进一步验证了本文方法在性能提升方面的有效性。

Table 3 Final Win Rate Comparison on SMAC Map Scenarios

表 3 SMAC 地图场景下的最终胜率对比

方法	2s3z	5m_vs_6m	6h_vs_8z
QMIX	0.97	0.47	0.39
MA^{2}E	0.97	0.48	0.31
本文	0.98	0.67	0.43

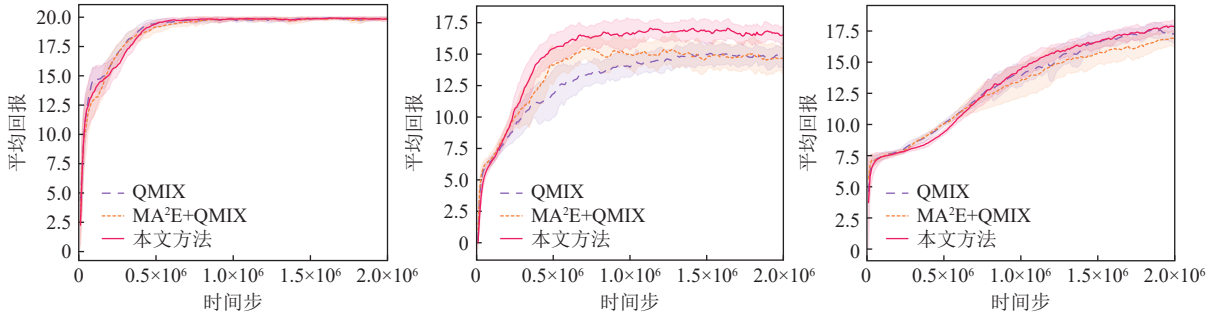
注：最终胜率并非理论意义上的最终收敛胜率。

3.3 平均回报对比分析

为了更好地评估智能体在整个过程中的综合表现，我们测试了各方法在不同难度等级场景下的平均回报表现，对比结果如图 4 所示。可以看出，本文方法在整个过程中始终保持较高的平均回报，在 3 种场景中综合优于 MA^{2}E 和 QMIX 这 2 种方法。

具体而言，在 Easy 场景中，3 种方法之间的差距相对较小，在 2s3z 地图中，由于 QMIX 结构相对简单，因此在前期平均回报的提升相对较快，而本文方法由于在前期需要广泛探索以寻找最佳分组，因此提升相对较慢，但最终表现与 2 种方法基本持平，并没有因此造成损失。

而在难度等级为 Hard 的 5m_vs_6m 地图中，由于 QMIX 方法仅依赖智能体自身的局部观测信息，缺乏对全局环境状态的充分把握，导致其策略调整能力受限，难以有效应对复杂的环境变化，从而影响最终的奖励回报。相较之下，本文方法与 MA^{2}E 表现更为稳定，主要原因在于二者均在智能体执行阶段对全局信息进行了重建，使智能体能够更全面地感知全局，进而根据当前局势动态调整和优化中间策略，提升决策质量，获得更高的奖励回报。同时，本文方法引入的主动通信模块为全局信息重建提供了多样化的样本，使得智能体能够感知更为准确的全局信息，从而做出更有效的决策，使奖励回报得到了进一步提升。



(a) Easy (2s3z) (b) Hard (5m_vs_6m) (c) Super Hard (6h_vs_8z)

Fig. 4 Mean return comparison on SMAC map scenarios

图 4 SMAC 地图场景下的平均回报对比

对于 6h_vs_8z 地图，由于该地图本身的策略特征较为凸显，QMIX 虽然仅依赖局部观测，但仍能够较快地捕捉关键战术，从而在训练前期展现出较优的性能。然而，受限于缺乏全局感知能力，QMIX 在复杂局势下的策略调整效率不足，难以持续优化，因而在训练中后期逐渐被本文方法反超。相比之下，MA^{2}E 虽然尝试通过掩码重建全局信息，但受样本稀疏性的影响，重建结果往往存在误差，从而干扰决策调整，因此整体表现未能优于 QMIX。由此可见，即便在具有较强策略信号的前提下，提升全局信息重建的精度仍然是优化策略并进一步增强整体性能的关键。

为了直观地反映各方法在不同场景下的策略调整能力，我们统计并比较了各方法在后 1E6 时间步内的平均回报作为最终回报，结果如表 4 所示。结果表明，本文方法在多个场景下均取得了更高的最终回报。这主要得益于本文方法在执行阶段对全局信息的有效重建与利用，从而使智能体能够在复杂环境下实现了更有效的决策。

Table 4 Final Return Comparison on SMAC Map Scenarios

表 4 SMAC 地图场景下的最终回报对比

方法	2s3z	5m_vs_6m	6h_vs_8z
QMIX	19.85	14.77	16.59
MA^{2}E	19.83	14.89	15.94
本文	19.87	16.77	16.90

注：最终回报并非理论意义上的最终收敛回报。

3.4 信息完整性对性能的影响分析

为深入探究信息完整性对智能体策略协作性能的影响，我们在对比实验基础上进一步扩展了实验。按照全局观测信息的不同完整程度，分别选取了具有代表性的 4 种情况：局部观测（QMIX）、单独局部观测信息推断全局信息（MA^{2}E）、随机信息推断全局信息（用随机通信代替主动通信，且随机通信的智能体数量设置为 3）和高相关的多样化信息推断全局信息（本文方法）。实验选择在地图 5m_vs_6m 上进行，这是因为该场景为同构单位且具有一定不对称性，能有效放大信息完整性对协作性能的影响；同时其难度适中，不会像地图 2s3z 那样差异不显著，也避免了地图 6h_vs_8z 中因策略特征过于凸显而掩盖方法差异的情况，因此更适合用于验证不同信息完整程度下性能的变化趋势。我们依旧采用平均胜率作为评估指标，用以对比信息完整性对智能体策略协作性能的影响，最终的实验结果如图 5 所示。结果表明，提高信息完整程度对于提升平均胜率具有积极作用。

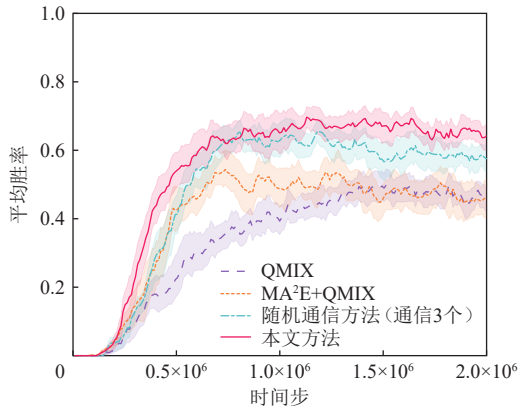


Fig. 5 Impact of information integrity on performance

图 5 信息完整性对性能的影响

为了更加清楚地反映信息完整性对智能体策略协作性能的影响，我们取后 1E6 步的平均胜率，以反映不同信息完整性下的性能表现，最终的性能表现柱状图如图 6 所示。结果表明，随着信息完整程度的提升，平均胜率呈现上升趋势，验证了增强信息完整性对提升智能体策略协作性能的重要性。

局部观测（QMIX）与本文方法的对比结果如图 7 所示，实验选择在地图 5m_vs_6m 上进行。QMIX 作为传统方法，仅使用智能体的局部观测信息进行独立决策，由于缺乏全局视野，导致智能体无法感知其他智能体的行为，使得不能对于智能体之间的协作做出针对性的策略调整。因此，验证了部分可观测场景下局部观测信息在复杂协作任务中的固有局限性。

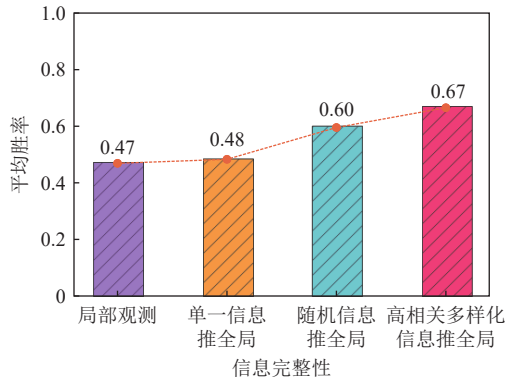


Fig. 6 Performance under different information integrity

图 6 不同信息完整性下的性能表现

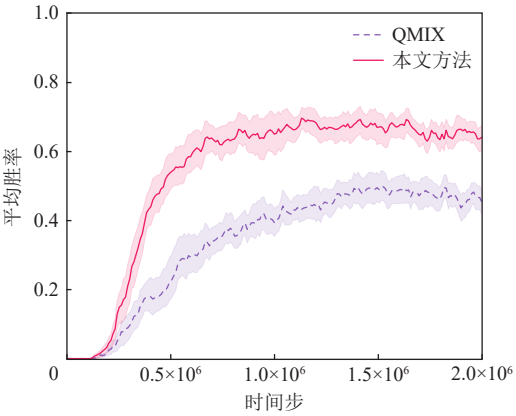


Fig. 7 Comparison between QMIX and the proposed method

图 7 QMIX 与本文方法的对比

单独局部观测信息推断全局信息（MA^{2}E）与本文方法的对比结果如图 8 所示，实验选择在地图 5m_vs_6m 上进行。MA^{2}E 作为多智能体掩码重建的基础方法，其利用掩码重建方法，根据智能体的局部观测信息来推断全局观测信息，从而辅助智能体的决策过程。相较于 QMIX 方法，虽然在一定程度缓解了缺乏全局视野的问题，但由于受样本稀疏性的影响，仅依靠单独的局部观测推断全局信息准确性仍然不高，相比于本文方法存在一定差距，因此仍具有局限性。

为了深入分析通信机制对全局信息推断以及性能的影响，我们还测试了随机通信方法在不同通信数量配置下的结果，并与本文方法进行了对比，实验选择在地图 5m_vs_6m 上进行，实验结果如图 9 所示。实验结果表明，随机通信下的智能体协作性能并不会随通信数量的增加而单调提升，这是因为智能体可能接收到关联度较低的信息，导致形成“虚假全局”，从而干扰智能体决策。相比之下，本文方法使

智能体能提供更加相关且多样化的观测信息,从而增强掩码自编码器的全局推断能力,提升决策效果。因此,该结果验证了主动通信机制在通信层面的有效性与先进性。

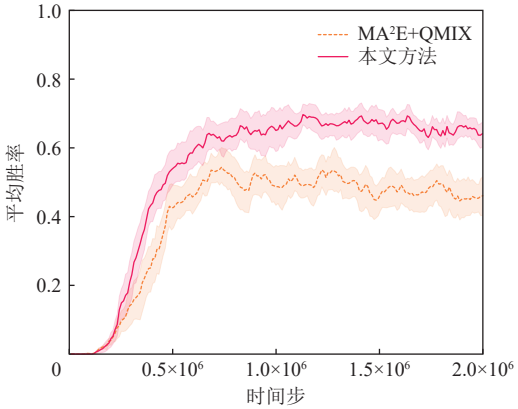


Fig. 8 Comparison between MA²E and our method

图 8 MA²E 与本文方法的对比

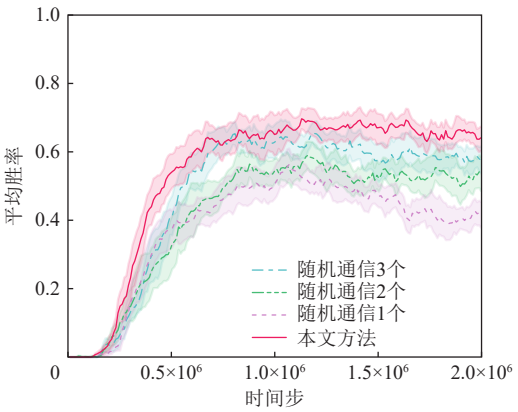


Fig. 9 Comparison between different communication mechanisms

图 9 不同通信机制之间的对比

综上所述,提升智能体间协作性能的关键在于增强信息完整性。因此,为了验证本文方法对信息完整性的影响,我们额外测试了不同方法下智能体级掩码自编码器在预训练前、预训练后和微调后对所有智能体信息的重建精度。具体而言,我们通过计算 3 次实验中所有智能体观测真实值与重建值之间的平均均方误差(MSE),来对比 MA²E、随机通信(通信 3 个)与本文方法在观测值重建精度上的差异。需要说明的是,为确保本文方法能够实现有效的通信调度,并进一步评估模型在动态复杂环境下的重建稳健性,表 5 所示的测试结果均在训练模式下获得。结果表明,相比于 MA²E 和随机通信方法(通信 3 个)这 2 种方法,本文方法在预训练阶段对于重建精度的提升最为明显,并在最终微调后取得了最高的重建精度,原因在于本文方法能够利用主动通信机制为智能体级掩码自编码器提供重要且多样化的样本,从而帮助智能体更准确地重建全局信息,验证了本文方法在提升信息重建精度方面的有效性。

Table 5 Comparison of Reconstruction Accuracy

表 5 重建精度对比

方法	预训练前	预训练后	微调后
MA²E	0.542 7	0.472 0	0.277 7
随机通信(通信 3 个)	0.458 9	0.392 9	0.273 2
本文	0.458 2	**0.352 6**	**0.264 5**

注:加黑数值表示不同阶段完成后的重建精度的最优值。

3.5 适用性及适用条件分析

本文方法虽然基于 QMIX 框架进行设计,但同样适用于所有基于值函数的方法。因此,为了测试本文方法的适用性,我们将本文方法移植到了 VDN[21] 和 QPLEX[50] 这 2 种基于值函数的方法中,测试了移植前后在地图 5m_vs_6m 上的平均胜率,实验设置与 3.1 节保持一致,最终 VDN 与本文方法的对比结果如图 10 所示,QPLEX 与本文方法的对比结果如图 11 所示。观察结果可知,2 种方法在移植本文方法以后,平均胜率均得到了一定程度的提升,验证了本文方法在基于值函数方法层面的适用性。

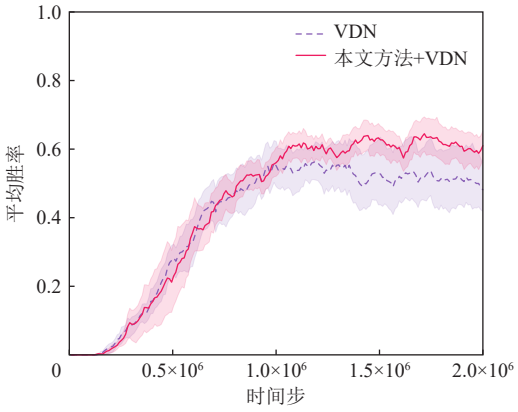


Fig. 10 Comparison between VDN and our method

图 10 VDN 与本文方法的对比

同时,为了保证方法的有效性和适用性,本文方法强调在集中训练和分散执行(CTDE)范式下进行实现。在集中训练阶段,智能体要求能够共享观测信息进行动态分组并确定通信对象集合;在分散执行阶段,每个智能体应具备通信能力,从而获取其他智能体的信息,帮助重建更准确的全局信息以辅助自身决策。保证以上条件才能确保本文方法能够在部分可观测场景下提升多智能体系统性能。

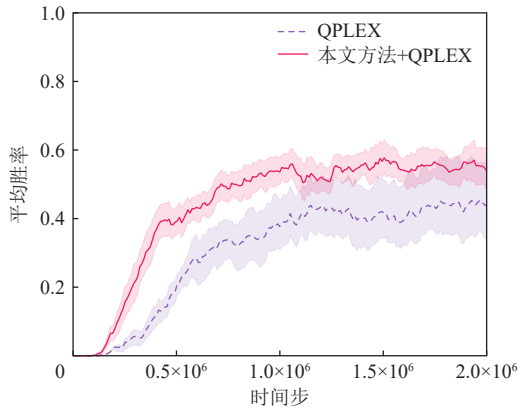


Fig. 11 Comparison between QPLEX and our method

图 11 QPLEX 与本文方法的对比

3.6 代价分析

由于模型在实际运行过程中受模型规模和通信开销等因素的影响，因此，我们进一步对不同方法进行了代价分析，主要从训练阶段的模型参数总量、收敛所需时间步以及执行阶段的单步通信开销 3 个方面进行了评估，所有数据均在 5m_vs_6m 地图上进行收集，最终的统计结果如表 6 所示。表中的结果表明，本文方法由于引入主动通信机制，确实导致模型参数量有所增加。然而，与随机通信方法相比，本文方法在通过基于谱聚类的动态分组机制寻找到最佳分组并确定通信对象以后，智能体能够在执行阶段通信少量且相关的智能体信息，从而在保证重建效果的同时降低通信开销。与此同时，机制所带来的高效信息利用能力也使得整体训练过程更易收敛，从而缩短了达到稳定性能所需的时间步。

Table 6 Cost Comparison of Different Methods on the Map of 5m_vs_6m

表 6 5m_vs_6m 地图下不同方法的代价对比

方法	单步通信开销/b	收敛时间步	模型参数量
本文	**30 368**	**0.767**×10⁶	699 342
随机通信（通信 3 个）	35 040	1.049×10⁶	683 574
MA²E	0（无通信）	1.332×10⁶	680 017
QMIX	0（无通信）	1.423×10⁶	**141 485**

注：黑体数值表示最优值。

综上所述，本文方法在增加少量模型代价的同时，有效降低了通信开销并加快了收敛速度，实现了更优的综合性能。

3.7 智能体间协作表现分析

为了更加直观地展现本文方法对于智能体间协作表现的影响，我们在 SMAC 环境中实现了可视化。具体来讲，我们截取了本文方法与 QMIX 在 SMAC 地图 5m_vs_6m 中的实机运行画面，详细观察并记录了这 2 种方法下盟军用来对抗敌军的协作策略，最终得到的可视化结果如图 12 所示。



(a) 本文方法



(b) QMIX

Fig. 12 Visualization of QMIX and our method

图 12 QMIX 与本文方法的可视化

图 12 中，QMIX 由于缺乏全局视角，使得每个智能体只能根据自身的观测信息进行决策，导致未能集火单一目标，而是呈现出各自为战的局面，且智能体后撤并返回时，由于局部观测只能观测到最近可见目标，导致无法与其他智能体协同攻击同一目标，最终以只有 2 个智能体存活的局面险胜。

而对于本文方法，智能体不仅能根据血量动态调整站位，以保证自身存活与整体阵型稳定，同时还得益于全局视角，智能体能在后撤并返回时优先攻击集火目标，而非最近可见目标，从而实现了有效的集火策略，最终以仅牺牲 1 个智能体的局面获胜。

综上所述，本文方法通过为智能体提供一种全

局视角，从而帮助它们学习得到一个跨智能体的全局协作策略，进而促进协作，体现了本文方法对于增强智能体间协作表现的有效性。

4 结 论

本文针对多智能体强化学习中的部分可观测性问题，提出了一种主动通信掩码重建方法。该方法在集中训练阶段利用谱聚类实现智能体的动态分组，并通过掩码交叉注意力机制为每个智能体选择少量且高度相关的智能体来确定通信关系，进而在分散执行阶段，每个智能体依据通信对象集合主动与对应智能体建立通信，从而为智能体级掩码自编码器提供多样化信息，实现全局信息的准确重建，之后再利用自注意力机制构建全局特征以辅助智能体决策。在 SMAC 平台上的实验结果表明，本文方法能够显著提高多个协作任务中的性能，验证了方法在提升协作多智能体系统性能上的可行性与有效性。

在未来的研究中，我们将优化通信策略，进一步考虑智能体间的距离以及通信拓扑结构等因素对全局状态推断与协作性能的影响；同时，还将继续探索如何降低计算复杂度的同时，提高掩码重建的精确度；最后，我们将进一步把该方法扩展至完全分布式的智能体系统，并应用于更加复杂的任务环境中。

作者贡献声明：贾月洋负责提出方法思路，完成实验并撰写论文；刘文彬负责论文审定和修改；王恩提出指导意见并修改论文。

参 考 文 献

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

[2] Canese L, Cardarilli G C, Nunzio L D, et al. Multi-agent reinforcement learning: A review of challenges and applications[J]. Applied Sciences, 2021, 11(11): 4948

[3] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 3826−3839

[4] Zhou Ming, Luo Jun, Villella J, et al. Smarts: An open-source scalable multi-agent RL training school for autonomous driving[C]//Proc of the 4th Conf on Robot Learning. New York: PMLR, 2021: 264−285

[5] Yang Yaodong, Luo Rui, Li Minne, et al. Mean field multi-agent reinforcement learning[C]//Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 5571−5580

[6] Di Yuanzhu, Deng Libao, Wang Ling, et al. A review of deep reinforcement learning methods for production scheduling[J/OL]. Tsinghua Science and Technology, 2025[2026-01-23]. https://doi.org/10.26599/TST.2024.9010247

[7] Yu Chao, Yang Xinyi, Gao Jiaxuan, et al. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration[C]//Proc of the 2023 Int Conf on Autonomous Agents and Multi-agent Systems. Berlin: Springer, 2023: 1107−1115

[8] Han Sheng, Zhang Hengrui, Wu Hao, et al. Multi-constraint reinforcement learning in complex robot environments[J]. Frontiers of Computer Science, 2025, 19(8) : 198353

[9] Peng Peng, Wen Ying, Yang Yaodong, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games[J]. arXiv preprint, arXiv: 1703.10069, 2017

[10] Lu Wenhao, Zhao Yong, Ji Yatai, et al. Task allocation method for collaborative crowdsensing of heterogeneous entities in emergency response[J]. Chinese Journal on Internet of Things, 2025, 9(4): 77−90 (in Chinese)
(路文浩，赵勇，季雅泰，等. 面向应急响应群智感知的异构群体协作任务分配方法[J]. 物联网学报，2025，9(4)：77−90)

[11] Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning[C]//Proc of the 2017 Int Conf on Autonomous Agents and Multi-agent Systems. Berlin: Springer, 2017: 66−83

[12] Kraemer L, Banerjee B. Multi-agent reinforcement learning as a rehearsal for decentralized planning[J]. Neurocomputing, 2016, 190(19): 82−94

[13] Zhu Changxi, Dastani M, Wang Shihan. A survey of multi-agent deep reinforcement learning with communication[J]. arXiv preprint, arXiv: 2203.08975, 2024

[14] Seo Y, Hafner D, Liu Hao, et al. Masked world models for visual control[C]//Proc of the 6th Conf on Robot Learning. New York: PMLR, 2023: 1332−1344

[15] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm[C]//Proc of the 15th Int Conf on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA: MIT, 2001: 849−856

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 6000−6010

[17] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//Proc of the 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16000−16009

[18] Samvelyan M, Rashid T, De Witt C S, et al. The StarCraft multi-agent challenge[J]. arXiv preprint, arXiv: 1902.04043, 2019

[19] Lauer M, Riedmiller M A. An algorithm for distributed reinforcement learning in cooperative multi-agent systems[C]//Proc of the 17th Int Conf on Machine Learning. New York: ACM, 2000: 535−542

[20] Foerster J, Nardelli N, Farquhar G, et al. Stabilising experience replay for deep multi-agent reinforcement learning[C]//Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 1146−1155

[21] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning[J]. arXiv preprint, arXiv: 1706.05296, 2017

[22] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(1): 1−51

[23] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]//Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1928−1937

[24] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint, arXiv: 1509.02971, 2015

[25] Russell S J, Zimdars A. Q-decomposition for reinforcement learning agents[C]//Proc of the 20th Int Conf on Machine Learning. New York: ACM, 2003: 656−663

[26] Schroeder de Witt C, Foerster J, Farquhar G, et al. Multi-agent common knowledge reinforcement learning[C]//Proc of the 33rd Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 9927−9939

[27] Shu Tianmin, Tian Yuandong. M^{3}RL: Mind-aware multi-agent management reinforcement learning[J]. arXiv preprint, arXiv: 1810.00147, 2018

[28] Iqbal S, Costales R, Sha F. ALMA: Hierarchical learning for composite multi-agent tasks[C]//Proc of the 36th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 7155−7166

[29] Wang Tonghan, Dong Heng, Lesser V, et al. ROMA: Multi-agent reinforcement learning with emergent roles[C]//Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 9876–9886

[30] Wang Tonghan, Gupta T, Mahajan A, et al. RODE: Learning roles to decompose multi-agent tasks[J]. arXiv preprint, arXiv: 2010.01523, 2020

[31] Zang Yifan, He Jinmin, Li Kai, et al. Automatic grouping for efficient cooperative multi-agent reinforcement learning[C]//Proc of the 37th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2023: 46105−46121

[32] Liu Chiqiang, Li Dazi. HYGMA: Hypergraph coordination networks with dynamic grouping for multi-agent reinforcement learning[J]. arXiv preprint, arXiv: 2505.07207, 2025

[33] Caliński T, Harabasz J. A dendrite method for cluster analysis[J]. Communications in Statistics-theory and Methods, 1974, 3(1): 1−27

[34] Foerster J, Assael I A, De Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning[C]//Proc of the 30th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2016: 2145−2153

[35] Sukhbaatar S, Fergus R. Learning multiagent communication with backpropagation[C]//Proc of the 30th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2016: 2252−2260

[36] Jiang Jiechuan, Lu Zongqing. Learning attentional communication for multi-agent cooperation[C]//Proc of the 32nd Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 7265−7275

[37] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multiagent cooperative and competitive tasks[J]. arXiv preprint, arXiv: 1812.09755, 2018

[38] Liu Yong, Wang Weixun, Hu Yujing, et al. Multi-agent game abstraction via graph attention neural network[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 7211−7218

[39] Das A, Gervet T, Romoff J, et al. TarMAC: Targeted multi-agent communication[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 1538−1546

[40] Guan Cong, Chen Feng, Yuan Lei, et al. Efficient multi-agent communication via self-supervised information aggregation[C]//Proc of the 36th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 1020−1033

[41] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[42] Liu Fangchen, Liu Hao, Grover A, et al. Masked autoencoding for scalable and generalizable decision making[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2022: 12608−12618

[43] Wu P, Majumdar A, Stone K, et al. Masked trajectory models for prediction, representation, and control[C]//Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 37607−37623

[44] Song Haolin, Feng Mingxiao, Zhou Wengang, et al. MA2CL: Masked attentive contrastive learning for multi-agent reinforcement learning[J]. arXiv preprint, arXiv: 2306.02006, 2023

[45] Kang S, Lee Y, Kim G, et al. MA^{2}E: Addressing partial observability in multi-agent reinforcement learning with masked auto-encoder[C]// Proc of the 13rd Int Conf Learning Representations. Mountain View, CA: OpenReview. net, 2025: 51818−51838

[46] Oliehoek F A, Amato C. A Concise Introduction to Decentralized POMDPs[M]. Berlin: Springer, 2016

[47] Shi Jianbo, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888−905

[48] Zha Hongyuan, He Xiaofeng, Ding C, et al. Spectral relaxation for k-means clustering[C]//Proc of the 15th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2001: 1057−1064

[49] Hu Jian, Jiang Siyang, Harding S A, et al. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning[J]. arXiv preprint, arXiv: 2102.03479, 2021

[50] Wang Jianhao, Ren Zhizhou, Liu T, et al. QPLEX: Duplex dueling multi-agent Q-learning[C]//Proc of the 9th Int Conf on Learning Representations. Mountain View, CA: OpenReview.net, 2021: 11609−11635



Jia Yueyang, born in 2001. Master. His main research interest includes multi-agent reinforcement learning.

贾月洋, 2001 年生。硕士。主要研究方向为多智能体强化学习。



Liu Wenbin, born in 1990. PhD, associate professor. His main research interests include multi-agent reinforcement learning, mobile computing, and mobile crowdsensing.

刘文彬, 1990 年生。博士，副教授。主要研究方向为多智能体强化学习、移动计算、移动群智感知。



Wang En, born in 1987. PhD, professor. His main research interests include multi-agent reinforcement learning, mobile computing, and mobile crowdsensing.

王　恩, 1987 年生。博士，教授。主要研究方向为多智能体强化学习、移动计算、移动群智感知。