基于互信息引导信息聚合的多智能体强化学习方法

魏 巍[1),2)] 宋慧忠[1)] 李 琳[1)] 张利军[1)]
马 亿[1),2)] 郝建业[3)] 梁吉业[1),2)]

[1)](山西大学计算机与信息技术学院 太原 030006)
[2)](山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)
[3)](天津大学智能与计算学部 天津 300050)

摘 要 多智能体强化学习(MARL)通过将强化学习理论延伸至多智能体领域,构建一种适用于多种现实场景交互的通用建模求解范式,使智能体能够在与环境及其他智能体的交互中动态优化策略。然而,如何让智能体在高维复杂场景中从观测中提取有效信息并做出正确决策仍是当前 MARL 方法面临的一大挑战。现有方法通常将盟友观测视为确定性输入,在采用统一或者分组编码策略时,未能有效刻画盟友协作意图背后的不确定性,导致智能体难以准确推断并响应盟友的多样化协作行为,使协作效率提升受到限制。为此,本文提出变分信息聚合(Variational Information Aggregation,VIA)模块以提高智能体对盟友行为意图的感知与利用能力。该模块将盟友观测与自身状态的联合表征建模为高斯分布并进行采样,通过显式刻画盟友协作意图的不确定性增强智能体对盟友行为意图的推断能力。多个 MARL 基准上的实验结果表明,VIA 模块可嵌入任意基于值函数分解的 MARL 算法。在性能方面,嵌入 VIA 的 MARL 算法平均胜率提升约 30%;在样本效率方面,达到相同胜率所需的训练时间至多可减少 60%。

关键词 多智能体强化学习;信息聚合;互信息;变分推断;智能体建模
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2026.01514

Enhancing Multi-Agent Cooperation via Mutual Information-Guided Information Aggregation

WEI Wei[1),2)] SONG Hui-Zhong[1)] LI Lin[1)] ZHANG Li-Jun[1)]
MA Yi[1),2)] HAO Jian-Ye[3)] LIANG Ji-Ye[1),2)]

[1)](School of Computer and Information Technology, Shanxi University, Taiyuan 030006)
[2)](Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006)
[3)](College of Intelligence and Computing, Tianjin University, Tianjin 300050)

Abstract Multi-agent reinforcement learning (MARL) extends reinforcement learning to interactive multi-agent systems (MAS), providing a common modeling and solving paradigm for many real-world situations, enabling agents to learn and update their strategies in the process of interacting with the environment and other agents. Through the continuous interaction and policy iteration between agents, MARL gradually adapts to the influence of allies' policy evolution. MARL can

收稿日期:2025-10-16;在线发布日期:2026-03-17。本课题得到国家自然科学基金项目(Nos. 62276160,62506221)、山西省自然科学基金面上项目(No. 202203021211294)、山西省留学基金项目(No. 20240002)、山西省高等学校科技创新项目资助。**魏 巍**,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为强化学习、多智能体系统、大语言模型。E-mail:weiwei@sxu.edu.cn。**宋慧忠**,博士研究生,中国计算机学会(CCF)学生会员,主要研究领域为多智能体强化学习。**李 琳(通信作者)**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为强化学习、无人系统。E-mail:lilynn1116@sxu.edu.cn。**张利军**,博士研究生,中国计算机学会(CCF)学生会员,主要研究领域为强化学习、大语言模型。**马 亿**,博士,讲师,中国计算机学会(CCF)会员,主要研究领域为强化学习、具身智能。**郝建业**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为强化学习、具身智能。**梁吉业**,博士,教授,中国计算机学会(CCF)会士,主要研究领域为机器学习、人工智能。

model high-dimensional, non-stationary and partially observable environments. In recent years, MARL make significant progress in the fields of game AI, autonomous driving, UAVs, power grid control and embodied AI. However, the increasing number of agents leads to an exponential state space expansion and complexity of agents' environmental sensory information in MAS. Every agent needs to decode allies' communication and fuse multi-source observations while keeping its own state. Thus, it is still a key challenge to extract useful information and make decisions in complex environments. Existing methods often consider allies' observations deterministic inputs and use unified or grouped encoding methods, which can not describe the uncertainty of allies' behavior intentions. This makes it difficult for agents to reason and respond to allies' behavior intentions, limiting the agents' policy updating efficiency. In this paper, we propose a Variational Information Aggregation (VIA) module, which improves the agents' ability to perceive and utilize allies' behavior intentions. VIA divides the state space into three subsets according to the information source: the own-information subset, representing the agent's own state; the ally-information subset, representing the state of the surrounding allies; and the other-information subset, representing the state of enemies or other entities. Each information subset is processed by independent encoding to obtain the corresponding embedding. VIA integrates the own-information subset and other-information subset into the agent's own state. Subsequently, VIA models these states together with the ally-information subset as a Gaussian distribution. By using variational inference and the reparameterization trick, VIA explicitly models the uncertainty of allies' behavior intentions. We also propose a collaborative intention-action semantic regularization which include agent modeling regularization and mutual information regularization. Extensive experiments on benchmarks with different difficulty levels verify the effectiveness of VIA. With high compatibility and low training cost, VIA provides an efficient solution for promoting multi-agent collaboration. Compared with the baseline, the integrated VIA method not only increases the average test win rate by 30%, but also reduces the training time required to achieve the same win rate by up to 60%.

Keywords　multi-agent reinforcement learning; information aggregation; mutual information; variational inference; agent modeling

1　引　言

多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)主要研究多个智能体的协同控制,以实现协作或者竞争性目标[1]。近年来,MARL在许多领域都取得了重要进展,例如游戏AI[2]、智能网联汽车[3]、无人机编队控制[4]以及具身智能[5-6]。在多智能体系统中,智能体观测信息通常更为多样和复杂,这对智能体的信息处理和决策能力提出更高要求。以多无人机协同搜索任务[7]为例,每架无人机不仅要处理自身状态数据(如位置、速度、电池电量等),还需要解析来自协作无人机的通信消息(如位置、速度、已探测区域等),以及自身传感器获取的目标观测信息(如追踪目标、障碍区域等)。因此,如何有效聚合智能体自身与来自盟友的信息,并提取关键特征,是多智能体强化学习研究急需解决的问题。

为此,研究人员开展了大量相关工作,现有方法主要可分为两类。一类方法是采用单一网络架构编码智能体接收到的观测信息[8-9];另一类方法则考虑观测空间中存在不同类型的观测空间子集(例如反应自身状态的信息,来自盟友的通信信息以及其他信息等),设计独立的编码网络对不同的观测空间子集进行独立编码[10-11]。由于不同观测空间子集对决策的贡献程度不同,对其分别处理是自然有效的方法。然而,现有方法通常将盟友观测作为确定性特征进行编码与聚合,这种处理方式忽略了盟友行为背后意图的动态性与多样性,导致智能体在决策时难以准确推断盟友的潜在协作意图,限制协作行为的产生。

针对现有方法的局限，本文提出了变分信息聚合(Variational Information Aggregation，简称 VIA)模块，该模块基于盟友观测与自身状态的联合表征生成一个高斯分布，并对其采样获得刻画盟友协作意图的特征，以提升智能体对盟友协作意图的感知与利用能力。

本文的主要贡献可归纳为以下三点：

(1)提出一种建模盟友协作意图的模块(VIA)，该模块可以嵌入任意基于值函数分解的多智能体强化学习方法。

(2)构建一种协作意图正则化机制，通过联合优化智能体自身决策动作与盟友观测的相关性，引导策略网络编码盟友的协作意图。

(3)在多种多智能体测试平台上的实验分析表明，VIA 增强了智能体对盟友动作预测的准确性，且相比于更复杂的信息融合方案，能在取得相近效果的情况下显著降低运行时间。

2 相关工作

2.1 智能体观测处理

在多智能体系统中，每个智能体通常需要处理多样化的观测信息并提取准确的特征表示。现有方法大致可分为两类。第一类方法采用单一网络对观测信息进行编码。例如，QMIX[8]和 QPLEX[9]利用多层感知机(Multi-Layer Perceptron，MLP)编码当前观测信息，并通过循环神经网络[12](Recurrent Neural Networks，RNN)生成动作。UpDeT[13]和 ATM[14]引入 Transformer[15]替代 RNN 模块，提升智能体建模历史轨迹的能力。TransfQMix[16]采用图神经网络(Graph Neural Network，GNN)对智能体观测进行结构化建模。LINDA[17]通过在局部网络中构建对队友的行为意图，以缓解部分可观测问题。LIA[18]提出一种基于邻近关系的固定加权机制用于聚合观测信息。然而，智能体的观测空间内容通常存在不同类型的观测空间子集，这些子集包含的信息语义存在显著差异。每个智能体的观测空间通常由反映自身状态的自身观测信息子集，反映盟友状态的盟友信息子集以及来自对手或者其他目标信息构成的其他实体信息子集构成，使用单一网络结构难以有效捕捉不同类型观测空间子集的差异。因此，第二类方法关注不同观测空间子集的相关性，通过设计专用编码器分别处理各类特征。ASN[10]和 UNMAS[11]构建语义网络建模智能体状态和动作的关系，DyMA-CL[19]和 InforMARL[20]利用 GNN 处理维度可变的观测信息。EPC[21]和 CollaQ[22]针对盟友数量可变问题引入自注意力机制处理盟友信息。MASIA[23]结合自监督学习与多步动作预测的方法提升智能体的信息聚合能力。然而，现有 MARL 方法的智能体将盟友信息编码为确定性特征，忽略盟友协作意图固有的不确定性，制约协作效率提升。

2.2 智能体建模

在多智能体系统中，每个智能体都应具备感知其他智能体行为意图的能力，这种能力通常通过智能体建模[24](Agent Modeling，AM)实现。SOM[25]通过观察其他智能体的动作来推断其目标。MAIC[26]构建盟友行为模型，增强智能体间观测信息交互能力。心智理论[27](Theory of Mind，ToM)是心理学中的重要概念，核心在于通过递归推理盟友状态与意图实现协作。ToM2C[28]利用 ToM 指导智能体实现去中心化通信，优化多智能体系统协作性能。然而，现有 MARL 方法的智能体在预测其他智能体动作时，多依赖完整局部观测。实际场景中，智能体观测常混杂环境状态、自身状态及盟友信息，其中大量与盟友无关的冗余噪声制约了行为建模准确性。针对这一问题，本文方法使用观测中仅与盟友相关的观测空间子集来预测其动作，通过剔除无关信息干扰，提升盟友行为意图建模质量。

2.3 互信息在多智能体强化学习中的应用

互信息[29](Mutual Information，MI)是衡量变量间统计依赖关系的常用指标，在 MARL 中应用广泛。现有方法多利用 MI 最大化优化交互策略。NDQ[30]通过最大化动作与消息互信息并最小化消息的熵提升通信效率；MAIC[26]则通过最大化观测信息与盟友动作的 MI 优化协作行为。为增强学习稳定性，PMIC[31]引入双 MI 神经网络估计器建模全局状态与联合动作依赖，VM3-AC[32]将动作间的 MI 作为回报正则项以促进协作。MIPI[33]针对智能体数目动态变化场景，通过最小化策略间的 MI 降低个体依赖，从而提升策略泛化性。ROMA[34]和 CDS[35]利用 MI 鼓励智能体探索，引导智能体产生多样化的策略。BVM-CMARL[36]通过在局部视角下最大化局部变分与自身轨迹的 MI，激励智能体策略的多样性，同时在全局视角下最大化全局变分与盟友动作的 MI，强化智能体之间的协作。然而，现有基于 MI 的 MARL 方法大多依赖盟友的完整观测或者完整通信，缺乏对观测信息中语义

成分的区分与利用。为此,本文提出一种增强智能体动作与盟友行为意图相关性的互信息正则项,引导智能体关注盟友的状态,从而增强智能体的协作能力。

3　背景知识

完全协作的多智能体强化学习问题通常被建模为去中心化的部分可观测马尔可夫决策过程[37](Decentralized Partially Observable Markov Decision Process,Dec-POMDP),其可形式化为一个九元组 $G=\langle I,S,A,P,R,\Omega,O,n,\gamma\rangle$,其中 $I\equiv\{1,2,\cdots,n\}$,表示由 n 个智能体组成的有限集合,S 为状态集合,A 为动作集合,$\gamma\in[0,1)$为折扣因子,用于衡量未来奖励的重要性。在该框架下的每个时间步 t,每个智能体通过观测函数 $O(s,i)$接收得到局部观测 $o_i\in\Omega$,并维护历史轨迹信息嵌入 $\tau_i\in T\equiv(\Omega\times A)^*$ 来选择动作 $a_i\in A$,所有智能体的动作将共同构成联合动作 $a\in A^n$。随后,系统根据状态转移函数 $P(s'|s,a)$转移到下一个时刻的状态 s',并获得所有智能体共享的奖励 $r=R(s,a)$。联合策略 π 通过联合动作价值函数进行表示:

$$Q_{tot}(\tau,a)=\mathbb{E}_{s_{0:\infty},a_{0:\infty}}\left[\sum_{t=0}^{\infty}\gamma_t r_t \mid s^0=s,a^0=a\right] \tag{1}$$

其遵循个体-全局-最大(Individual-Global-Max,简称 IGM)条件[38]:

定义 1(IGM 条件)对于联合动作值函数 Q_{tot},存在个体动作值函数 Q_i 满足:

$$\operatorname{argmax}_{\mathbf{a}}Q_{tot}(\tau,a)=\begin{pmatrix}\operatorname{argmax}_{a_1}Q_1(\tau_1,a_1)\\ \cdots\\ \operatorname{argmax}_{a_n}Q_n(\tau_n,a_n)\end{pmatrix} \tag{2}$$

则称 Q_i 满足 Q_{tot} 的 IGM 条件。

基于该原则,Q_{tot} 对应联合状态动作空间的贪婪动作可通过每个智能体 Q_i 的贪婪动作得到。

4　方　法

本节提出一种用于增强智能体对盟友协作意图感知与利用的变分信息聚合(Variational Information Aggregation,简称 VIA)模块,其整体结构如图 1 所示。4.1 节阐述 VIA 的整体网络架构;4.2 节介绍协作意图正则化机制;4.3 节介绍将 VIA 嵌入到基于值函数分解的 MARL 方法后的联合优化目标。

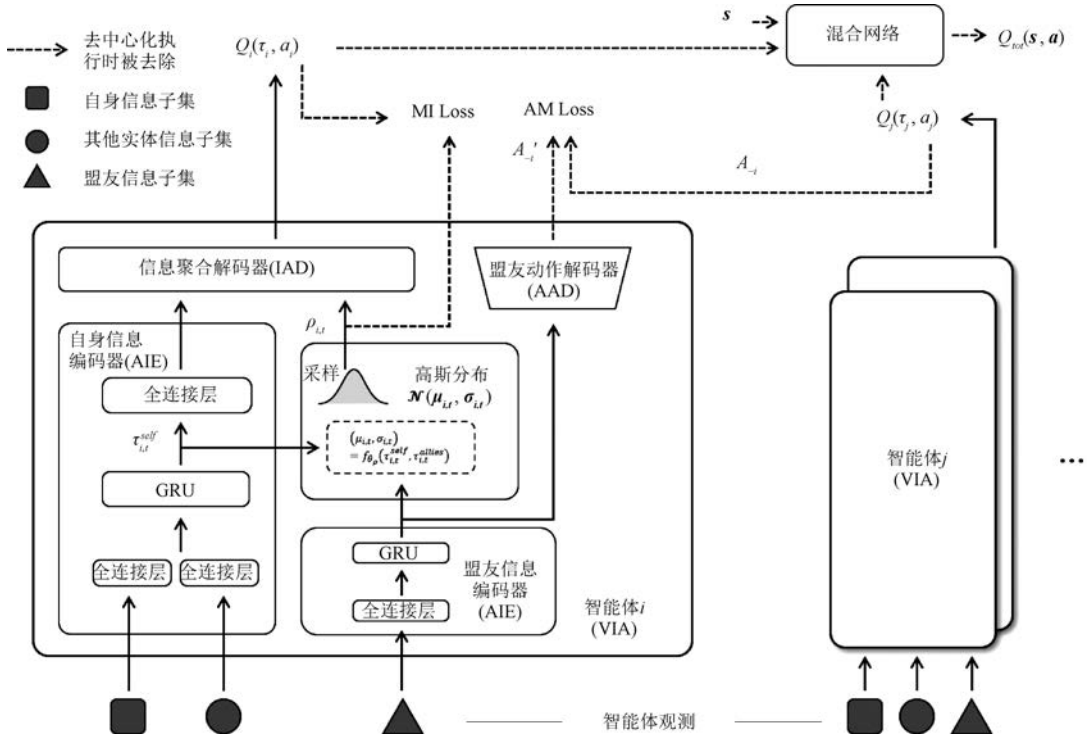


图 1　算法整体结构

4.1　VIA 结构

VIA 主要由两个观测编码器和两个信息解码

器组成。其中,编码器用于对不同类型的观测空间子集进行编码;两个解码器则分别用于聚合各观测空间子集的编码表示以及预测盟友的动作。

观测编码器(Observation Encoder)。在多智能体系统中,智能体的观测空间可根据信息来源划分为三类观测空间子集:自身信息子集 $o_{i,t}^{\text{self}}$,表示智能体自身的状态;盟友信息子集 $o_{i,t}^{\text{allies}}$,表示周围盟友智能体的状态;其他实体信息子集 $o_{i,t}^{\text{others}}$,表示观测范围内敌方或其他实体信息的状态。每个观测空间子集均通过独立编码处理,得到相应的嵌入表示,如公式(3)所示:

$$\begin{cases} e_{i,t}^{\text{self}} = MLP(o_{i,t}^{\text{self}}) \\ e_{i,t}^{\text{others}} = MLP(o_{i,t}^{\text{others}}) \\ e_{i,t}^{\text{allies}} = MLP(o_{i,t}^{\text{allies}}) \end{cases} \tag{3}$$

将这些嵌入表示分别输入自身信息编码器(Self-Information Encoder,SIE)和盟友信息编码器(Ally Information Encoder,AIE)中,分别产生智能体独立决策的状态-动作值函数 $Q_{i,t}^{\text{self}}$ 以及盟友的历史轨迹信息嵌入 $\tau_{i,t}^{\text{allies}}$:

$$\begin{cases} Q_{i,t}^{\text{self}} = SIE(e_{i,t}^{\text{self}}, e_{i,t}^{\text{others}}) \\ \tau_{i,t}^{\text{allies}} = AIE(e_{i,t}^{\text{allies}}) \end{cases} \tag{4}$$

其中,SIE 将自身信息 $e_{i,t}^{\text{self}}$ 与其他信息 $e_{i,t}^{\text{others}}$ 进行拼接后,利用 GRU[39] 对拼接信息进行编码,得到智能体自身的历史轨迹信息嵌入 $\tau_{i,t}^{\text{self}}$,通过 MLP 将该嵌入映射为状态-动作值函数 $Q_{i,t}^{\text{self}}$。

SIE 与 AIE 的主要差异在于:AIE 使用 GRU 对盟友信息 $e_{i,t}^{\text{allies}}$ 进行独立编码,生成盟友历史轨迹信息嵌入 $\tau_{i,t}^{\text{allies}}$;而 SIE 则在输出端额外引入一个 MLP,将编码后的嵌入直接映射为智能体决策的状态-动作值函数 $Q_{i,t}^{\text{self}}$。将盟友信息单独处理可避免不同类型信息在编码过程中相互干扰,有助于保留各部分的语义特征。

信息解码器(Observation Decoder)。VIA 包含两个解码器:信息聚合解码器(Information Aggregation Decoder,IAD)和盟友动作解码器(Allies Action Decoder,AAD)。IAD 用于将智能体接收到的信息进行聚合,生成智能体状态-动作值函数 $Q_{i,t}$,其形式如下:

$$Q_{i,t}(\tau_{i,t}^{\text{self}}, \tau_{i,t}^{\text{allies}}, \cdot) = Q_{i,t}^{\text{self}}(\tau_{i,t}^{\text{self}}, \cdot) + \rho_{i,t} \tag{5}$$

其中,$Q_{i,t}^{\text{self}}$ 表示智能体不依赖盟友时进行独立决策的状态-动作值函数。偏置项 $\rho_{i,t}$ 为协同价值修正项,通过联合建模智能体自身状态与盟友历史轨迹信息嵌入,实现对智能体独立决策价值的动态修正。$\rho_{i,t}$ 服从多元高斯分布 $\mathcal{N}(\mu_{i,t}, \sigma_{i,t})$,该分布可由 MLP($f_{\theta_\rho}$)计算得到:

$$(\mu_{i,t}, \sigma_{i,t}) = f_{\theta_\rho}(\tau_{i,t}^{\text{self}}, \tau_{i,t}^{\text{allies}}),$$
$$\rho_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma_{i,t}) \tag{6}$$

为支持梯度反向传播,$\rho_{i,t}$ 的采样过程引入了重参数化技巧。

AAD 通过心智理论[30]建模盟友信息,增强了智能体对盟友状态的理解。VIA 利用盟友的历史轨迹信息嵌入推断对应盟友的动作:

$$A'_{-i} = AAD(\tau_{i,t}^{\text{allies}}; \theta_{am}) \tag{7}$$

如公式(7)所示,智能体 i 中的盟友历史轨迹信息嵌入 $\tau_{i,t}^{\text{allies}}$ 被输入至 AAD,生成盟友的动作预测 A'_{-i},AAD 的参数为 θ_{am}。与采用统一编码方式处理智能体完整观测的方法不同,VIA 仅利用智能体观测空间中的盟友信息子集 $o_{i,t}^{\text{allies}}$ 预测盟友动作,而非依赖智能体的完整观测 $o_{i,t}$,这样能够有效地去除冗余信息,提高对盟友行为意图建模的质量。

算法 1. VIA-QMIX 算法

1. 输入:每个智能体的观测 o_i 以及全局状态信息 s_t
2. 输出:联合状态-动作值函数 $Q_{tot}(\tau, a)$
3. 参数初始化
4. FOR 每个回合 DO
5. FOR 每个时间步 t DO
6. FOR 每个智能体 i DO
7. 通过公式(3)(4)对智能体观测进行编码
8. 通过公式(5)聚合嵌入并计算 Q_i
9. 通过 argmax Q_i 选择动作 a_i
10. END FOR
11. 所有智能体与环境交互,获得奖励 r_t 和下一时刻状态 s_{t+1}
12. 收集经验$<s_t, a, r_t, s_{t+1}>$
13. END FOR
14. 将$\{Q_1, \cdots, Q_n\}$输入混合网络,计算 $Q_{tot}(\tau, a)$
15. 利用混合网络计算目标 $Q_{tot}(\tau, a)$
16. 根据公式(19)更新 $Q_{tot}(\tau, a)$与目标 $Q_{tot}(\tau, a)$
17. END FOR

4.2 协作意图正则化机制

本节主要介绍协作意图正则化机制,增强多智能体间的协作。该机制由两个正则项构成:一是智能体建模正则项,提高对盟友行为意图建模的质量;二是互信息正则项,通过最大化智能体观测信息与自身动作之间的互信息强化两者之间的语义关联,引导策略网络编码盟友协作意图。

智能体建模正则项(AM Loss)。为增强智能体对盟友状态的推断能力,构建一个用于准确预测盟

友动作的智能体建模正则项:

$$\mathcal{L}_{am}(\theta,\theta_{am})=\mathbb{E}_{o_{i,t},s_{i,t}}\|AAD(\tau_{i,t}^{\text{allies}};\theta_{am})-A_{-i}\|_2^2 \tag{8}$$

其中,AAD 表示盟友动作解码器,其输出为对盟友动作的预测;监督标签为盟友实际执行的动作,通过均方误差损失度量预测动作与真实动作之间的差异。该损失以监督学习方式更新 AAD 的参数 θ_{am},并作为正则项引入 MARL 的优化目标,以促进策略网络显式建模盟友行为。

互信息正则项(MI Loss)。盟友信息在智能体决策过程中发挥着重要作用。为了鼓励智能体在决策过程中关注盟友状态,促进智能体之间的协作,通过最大化偏置项 $\rho_{i,t}$ 与智能体动作 $a_{i,t}$ 之间的 MI,提出一个互信息正则项:

$$I(\rho_{i,t};a_{i,t})=H(\rho_{i,t}\mid\tau_{i,t}^{\text{self}},\tau_{i,t}^{\text{allies}})-H(\rho_{i,t}\mid\tau_{i,t}^{\text{self}},\tau_{i,t}^{\text{allies}},a_{i,t}) \tag{9}$$

公式(9)表示在给定智能体自身和盟友历史信息轨迹$(\tau_{i,t}^{\text{self}},\tau_{i,t}^{\text{allies}})$的前提下,动作 $a_{i,t}$ 对智能体对推断盟友协作意图 $\rho_{i,t}$ 的信息增益。令 $\tau_{i,t}=(\tau_{i,t}^{\text{self}},\tau_{i,t}^{\text{allies}})$:

$$I(\rho_{i,t};a_{i,t})=\underbrace{\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log p(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho\mathrm{d}a}_{\text{(a)}}-\underbrace{\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log p(\rho_{i,t}\mid\tau_{i,t})\mathrm{d}\rho\mathrm{d}a}_{\text{(b)}} \tag{10}$$

由于公式(10)的(a)项表达式的真实后验分布难以计算,本文借鉴信息瓶颈[34](Information Bottleneck,IB)的思想,利用变分近似分布 $g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t})$近似求解,建立一个互信息正则项的下界。根据 KL 散度(Kullback-Leibler Divergence,KL 散度)的非负性:

$$D_{KL}(p(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\|g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t}))\geqslant 0 \tag{11}$$

可得

$$\int p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log p(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho\geqslant\int p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho \tag{12}$$

将上述不等式左侧的表达式用公式(10)的(a)项表达式进行替换:

$$I(\rho_{i,t};a_{i,t}\mid\tau_{i,t})=\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log p(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho\mathrm{d}a\geqslant\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho\mathrm{d}a \tag{13}$$

公式(10)的(b)项表达式 $\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log p(\rho_{i,t}\mid\tau_{i,t})\mathrm{d}\rho\mathrm{d}a$ 是边缘熵,记作 $H(\rho\mid\tau)$。融合公式(10)的(a)项表达式和(b)项表达式可以得到互信息正则项的下界:

$$I(\rho_{i,t};a_{i,t}\mid\tau_{i,t})\geqslant\iint p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})\log g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t})\mathrm{d}\rho\mathrm{d}a-H(\rho\mid\tau) \tag{14}$$

令 $P_\rho=p(\rho_{i,t},a_{i,t}\mid\tau_{i,t})$,$G_\xi=g_\xi(\rho_{i,t}\mid\tau_{i,t},a_{i,t})$,公式(14)可简记为

$$I(\rho_{i,t};a_{i,t}\mid\tau_{i,t})\geqslant\iint P_\rho\log G_\xi\mathrm{d}\rho\mathrm{d}a-H(\rho\mid\tau) \tag{15}$$

即

$$I(\rho_{i,t};a_{i,t})\geqslant\mathbb{E}_{\mathcal{D}}[D_{KL}(P_\rho\parallel G_\xi)] \tag{16}$$

其中,D_{KL} 表示 KL 散度,分布 P_ρ 与 G_ξ 中涉及的变量均从经验回放缓冲区$\mathcal{D}$中采样得到。进一步可得到互信息正则项:

$$\mathcal{L}_\rho(\theta_\rho,\xi)=\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}[-D_{KL}(P_\rho\parallel G_\xi)] \tag{17}$$

4.3　优化目标

VIA 可以嵌入到任意基于值函数分解的 MARL 方法。以 QMIX[8] 为例,VIA-QMIX 以端到端的方式训练,方法中的所有参数均通过强化学习中的标准 TD 误差损失以及两个正则项(智能体建模正则项和互信息正则项)的梯度进行联合更新。在计算 TD 误差损失的过程中,每个智能体的状态-动作值函数 Q_i 被输入到混合网络中,输出联合状态-动作值函数 Q_{tot}。混合网络的参数由超网络[40] 基于全局状态 S_t 提供。

VIA 的最终学习目标为

$$\mathcal{L}(\theta,\theta_{am},\theta_\rho,\xi)=\mathcal{L}_{TD}(\theta)+\lambda_{am}\mathcal{L}_{am}(\theta,\theta_{am})+\lambda_\rho\mathcal{L}_\rho(\theta_\rho,\xi) \tag{18}$$

其中,$\mathcal{L}_{TD}(\theta)$表示强化学习的 TD 误差损失。$\mathcal{L}_{TD}(\theta)$定义如下:

$$\mathcal{L}_{TD}(\theta)=[r_t+\gamma\max_{a_t'}Q_{tot}(s_t',a_t';\theta^-)-Q_{tot}(s_t,a_t;\theta)]^2 \tag{19}$$

其中,θ 表示 Q_{tot} 网络的参数,θ^- 表示目标 Q_{tot} 网络的参数,θ 会周期性地赋值给 θ^-。λ_{am} 和 λ_ρ 为超参数,分别用于平衡 TD 误差损失与两个正则项之间

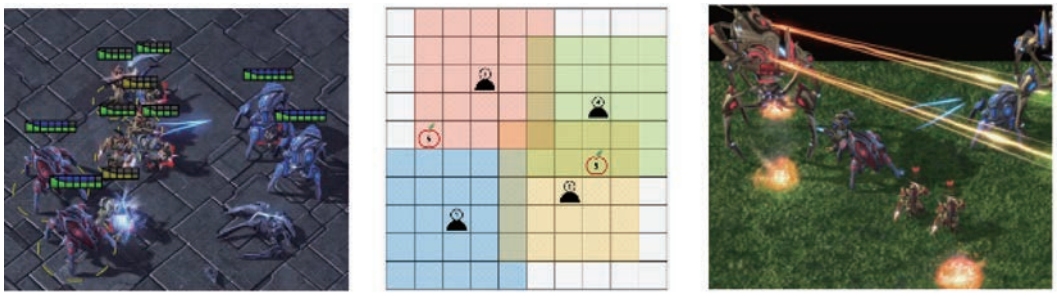
的权重。在基于值函数分解的 MARL 方法中，混合网络与 AAD 仅在集中式训练阶段使用，执行阶段将被移除。

5 实 验

为评估所提出方法的有效性，实验部分选取了三个常见的 MARL 基准测试环境：星际争霸多智能体挑战[41]（StarCraft Ⅱ Multi-Agent Challenges，SMAC）、基于层级的觅食环境[42]（Level-Based Foraging，LBF）以及星际争霸多智能体挑战 v2[43]（StarCraft Ⅱ Multi-Agent Challenges v2，SMACv2）。实验分析旨在回答以下四个问题：（1）VIA 相比现有基线方法是否具有更优性能？（2）将 VIA 嵌入到不同的基于值函数分解的 MARL 基线方法中，表现如何？（3）本文所提出方法中各组件对最终性能有何影响？（4）与具备通信条件的算法相比，VIA 表现如何？

5.1 实验设定

（1）SMAC。SMAC 环境包含一系列地图，其中一方由 MARL 算法控制的多个智能体组成，目标是击败敌方队伍，如图 2（a）所示。每个智能体包含了移动，攻击，无操作与停止 4 种类型的动作。在每个时间步中，智能体会根据敌方受到的累计伤害获得奖励，每杀死一个敌方智能体会额外获得 10 点奖励，每赢得一场战斗则会获得 200 点奖励。在不同难度的地图上进行测试（困难：5mvs6m、10mvs11m；困难＋：MMM2、3s5zvs3s6z），通过比较测试胜率来评估算法性能。



(a) SMAC (b) LBF (c) SMACv2

图 2 算法基准测试环境

（2）LBF。在 LBF 环境中，智能体需在网格地图中协作采集随机分布的食物资源，如图 2（b）所示。智能体与食物均具有等级属性，当且仅当参与协作的智能体等级总和不低于目标食物等级时才能成功拾取。智能体可在四个方向上移动，并对相邻食物执行“拾取”动作，动作是否成功取决于其自身等级与目标食物等级的匹配关系。在 10x10_3p4f 和 10x10_4p5f 两个具有不同智能体数量的环境中进行实验。所有实验均基于 Epymarl 框架实现。

（3）SMACv2。相较于 SMAC，SMACv2 环境更为复杂，主要进行了三项关键改进：初始位置随机化、智能体类型随机化以及视野与攻击范围的锥形限制。游戏环境如图 2（c）所示。前两项改进旨在提升环境的随机性，从而更严格地检验 MARL 算法的泛化与协作能力。其中，初始位置随机化包含两种模式：其一为“包围模式”，友方智能体在地图中心生成，敌方智能体环绕其四周，迫使智能体应对来自多个方向的威胁；其二为“镜像模式”，友方智能体随机生成后，敌方智能体位置通过地图中心对称镜像确定，从而形成对称对抗形态。在智能体类型随机化方面，SMACv2 不再采用固定兵种组合，而是依据预设概率随机生成作战智能体，例如神族（Protoss）可包含三类不同兵种。SMACv2 以星际争霸 II 游戏内原始数值替代原版固定值，可在战术上凸显近战与远程智能体的差异性，更贴近真实游戏机制。对于攻击范围和视野，SMACv2 将智能体的视野范围和攻击范围限制为 30°的扇形区域，并将动作空间扩展至 12 维，使智能体能够选择不同的观察方向。在 Protoss_5vs5 和 Protoss_10vs10 两个具有不同智能体数量的地图上进行相关实验。

5.2 基线对比结果

VIA 将与以下基线方法进行对比。为确保实验结果的公平性，所有方法均采用与 QMIX 相同的混合网络结构。（1）门控循环单元[8]（Gated Recurrent Unit，GRU）：智能体使用单一 GRU 对整体观测信息进行编码，并生成对应动作。（2）动作语义网络[15]（Action-Semantic Network，ASN）：根据语义信息对智能体观测进行拆分，并将其映射到相应的动作上。（3）图神经网络[17]（Graph Neural Network，GNN）：利用 GNN 对观测信息进行编码，分

别处理维度可变与维度不变的信息。(4)Transformer[10](UpDeT):UpDeT 使用 Transformer 对智能体信息进行全面编码。VIA 的相关参数设置可查看表 1。图 3 给出了 VIA 与各基线方法在 SMAC 环境中的性能对比。VIA 的整体性能要优于 GRU、ASN 和 GNN,在多个任务上和 UpDeT 表现相近,部分任务中甚至更优。在 5m vs 6m、10m vs 11m 和 3s5z vs 3s6z 地图上,VIA 的表现优于相关基线方法,表明有效聚合多样化观测信息能够促进多智能体协作。在 Hard+地图中,VIA 测试胜率早期增长较慢,但训练后期提升速度加快。这可能是因为 Hard+环境的状态-动作空间相对较大,需要进行更充分的探索,而 VIA 引入的两个正则项在训练初期对策略产生了一定的约束。当智能体逐渐适应这些优化约束,胜率在后期快速提升。这一现象在 3s5z vs 3s6z 地图中更为明显,具体的消融实验分析见 5.3 节。与基于 Transformer 的方法相比,VIA 在性能相近的情况下效率更高。从图 3(c)和表 2 可以看出,在相同实验条件下,VIA 在 MMM2 任务上达到与 UpDeT 相近的胜率时,仅需后者训练时间的 60%。

表 1　VIA 超参数设置

超参数名称	数值
批大小	32
经验池大小	5000
学习率	0.0005
折扣系数	0.99
互信息损失权重	0.01
智能体建模损失权重	0.001
RNN 隐藏层	64

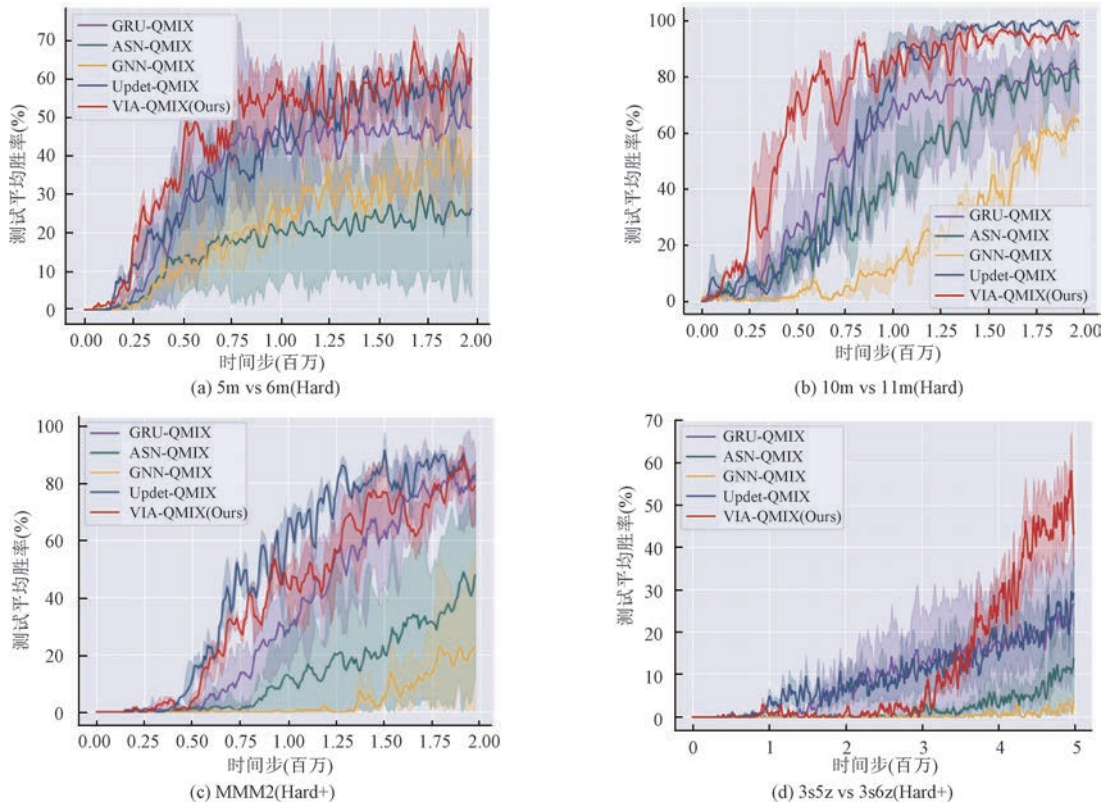


图 3　SMAC 环境中的性能对比

表 2　SMAC 各算法需要花费的运行时间(单位:分钟)

算法名称	5 m vs 6 m	10 m vs 11 m	MMM2	3s5z vs 3s6z
GRU-QMIX	319±2	510±3	755±1	1237±0
ASN-QPLEX	356±1	655±2	785±3	2274±0
GNN-QMIX	684±1	929±2	975±2	2271±2
Updet-QMIX	984±2	1875±0	1432±2	2896±1
VIA-QMIX	450±1	955±0	860±2	1756±0

5.3　嵌入到不同基于值函数分解的 MARL 方法下的性能表现

图 4 和图 5 分别展示了 VIA 嵌入到不同的基于值函数分解的 MARL 方法后的实验结果。在 LBF 环境中,与 VIA 结合的方法均优于对应基线方法,表明变分信息聚合机制能够有效建模盟友的不确定性协作行为,从而加速协作行为产生。为验证 VIA 嵌入到不同的基于值函数分解的 MARL 方法(如 QMIX、QPLEX 及 DuelMix[44])的潜力,进一步在 SMACv2 上开展了实验。SMACv2 环境比 LBF 更具挑战性。训练初期,各方法与基线之间的性能差异不显著,但到训练后期,VIA-QPLEX 和 VIA-

DuelMix 的性能明显提升。不同方法所需运行时间的详细数据可查看表 3 和表 4。

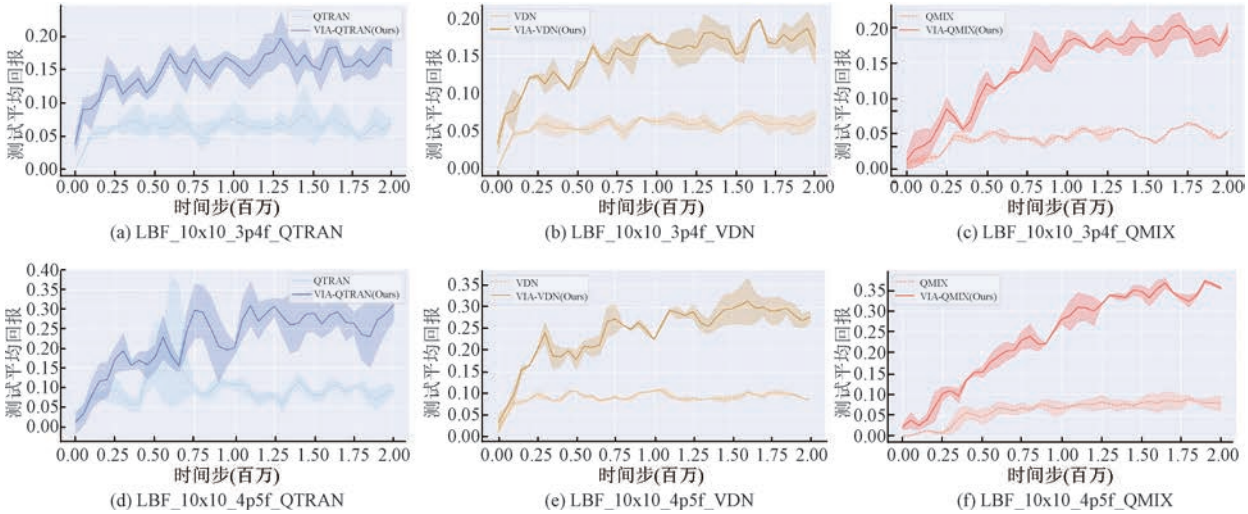


图 4 在 LBF 上与不同算法结合的实验结果(图(a)～(c) 为在 10x10_3p4f 上与不同算法结合的实验结果。图(d)～(f)为在 10x10_4p5f 上与不同算法结合的实验结果)

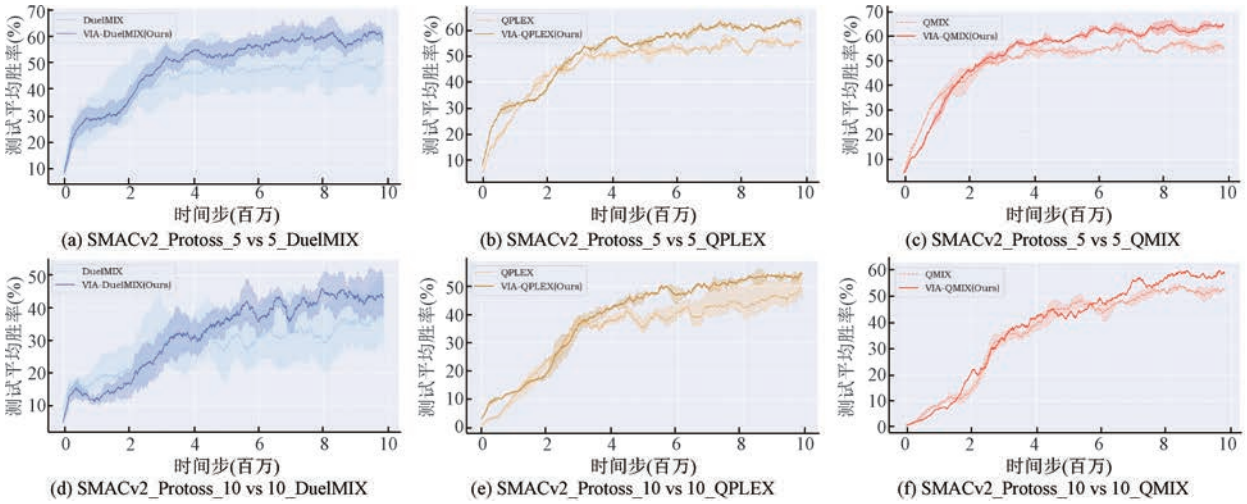


图 5 SMACv2 上与不同算法结合的实验结果(图(a)～(c)为在 Protoss_5 vs 5 上与不同算法结合的实验结果。图(d)～(f)为在 Protoss_10 vs 10 上与不同算法结合的实验结果)

表 3 在 LBF 环境中各算法的运行时间(单位:分钟)

算法名称	10x10_3p4f	10x10_4p5f
GRU-VDN	218±1	224±0
GRU-QMIX	216±3	227±2
GRU-QTRAN	223±0	244±1
VIA-VDN	918±2	920±2
VIA-QMIX	916±2	922±1
VIA-QTRAN	926±0	928±0

表 4 在 SMACv2 环境中各算法的运行时间(单位:分钟)

算法名称	Protoss_5vs5	Protoss_10vs10
GRU-QMIX	2420±5	3736±0
GRU-QPLEX	2036±2	3533±6
GRU-DuelMix	2122±0	3638±5
VIA-QMIX	3382±2	4569±3
VIA-QPLEX	3562±3	4545±4
VIA-DuelMix	3710±2	4642±2

5.4 消融实验分析

为验证 VIA 内部不同结构对最终性能的影响,实验设计如下变体:(1)VIA-QMIX w/o gs and mi and am:移除高斯采样模块,互信息正则项和智能体建模正则项;(2)VIA-QMIX w/o mi and am:移除互信息正则项和智能体建模正则项;(3)VIA-QMIX w/o mi:仅移除互信息正则项;(4)VIA-QMIX w/o am:仅移除智能体建模正则项。实验结果如图 6 所示:(1)与 VIA-QMIX w/o am 相比,VIA-QMIX 呈现出显著性能优势,验证了智能体建模正则项对整体性能的贡献,特别是在 5m vs 6m 和 MMM2 两个环境中,不采用正则化的 VIA-

QMIX w/o am 的性能甚至劣于原始 QMIX。其原因可能在于:在智能体数量少、环境简单的任务中,仅需有效处理智能体观测信息,就可以显著提升算法性能。而在 MMM2 和 3s5z vs 3s6z 等复杂环境中,智能体需要准确理解盟友的行为意图,才能实现有效协作。(2)VIA-QMIX 和 VIA-QMIX w/o mi 对比表明,引入互信息正则项可以提升算法性能,在 3s5z vs 3s6z 环境中尤为明显。实验结果表明,互信息正则项可提高智能体自身决策动作与盟友观测信息间的关联性,增强其对盟友状态的感知能力,从而促进协作产生。(3)训练初期,嵌入 VIA 的 MARL 方法胜率增速较慢,随着训练步数增加,性能提升速度逐渐加快。然而,移除两个正则项的 VIA-QMIX w/o mi and am 与 QMIX 相比,VIA-QMIX w/o mi and am 在训练初期稳定性有所下降;引入互信息正则项(VIA-QMIX w/o am)后,胜率明显提升,表明互信息正则项增强了智能体的行为策略与盟友协作意图之间的相关性,提升智能体对盟友行为意图的感知能力,但在训练后期胜率仍然有所波动,表明智能体之间的协作仍不够稳定。而借助智能体建模正则项,智能体能持续调整对盟友的特征表示,提高对盟友行为意图建模的质量以及策略的稳定性。(4)VIA-QMIX w/o gs and mi and am 与 VIA-QMIX w/o mi and am 的对比表明,高斯采样对提升模型性能有重要影响。带有高斯采样的 VIA-QMIX w/o mi and am 方法会牺牲一定的稳定性,但能提升智能体对盟友协作意图的感知与利用能力以及更好的协作性能。

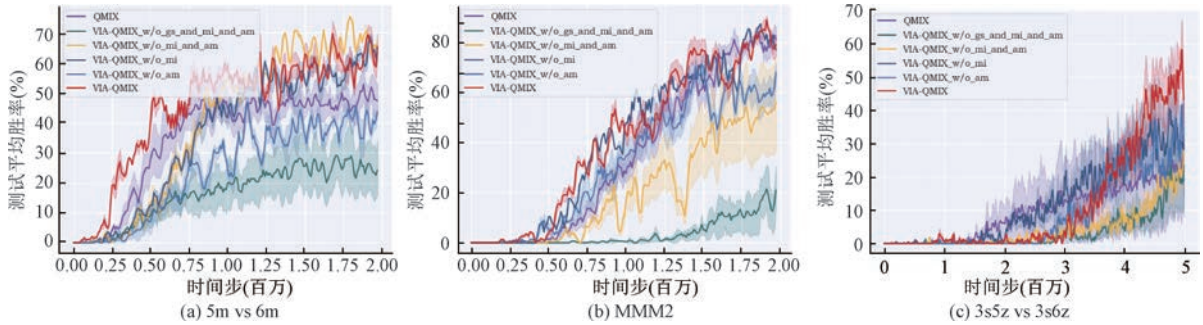


图 6　VIA 消融实验

5.5　与基于通信的方法进行比较

为进一步说明嵌入 VIA 的 MARL 方法在信息聚合层面的优势,与基于通信的 MARL 方法 MASIA[18] 进行对比实验,平均胜率如图 7 所示。MASIA 通过自监督学习与多步预测聚合盟友的信息。为确保实验的公平性,将 VIA 中盟友的观测信息作为通信信息,输入至 AIE 中进行编码。由于 MASIA 忽略重构信息与智能体动作之间的关联,而 VIA 引入协作意图正则化机制,强化观测信息与动作之间的关联,促进智能体之间有效协作。

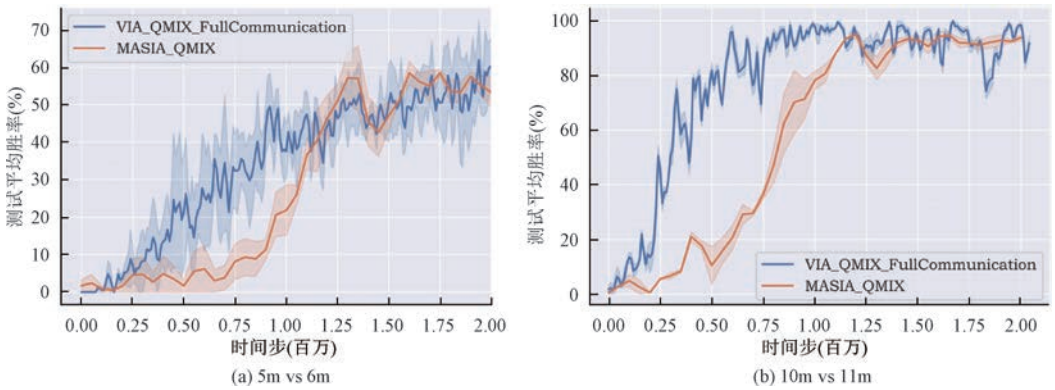


图 7　与 MASIA 的比较结果

6　结　论

本文提出一种用于增强智能体对盟友协作意图感知与利用的变分信息聚合模块,解决现有 MARL 研究忽略盟友行为意图动态性与多样性的问题。VIA 通过将盟友观测与自身状态的联合表征建模为高斯分布并进行采样,同时引入一种协作意图正

则化机制提升智能体协作能力。在多个 MARL 任务上的实验结果表明，VIA 不仅可以提升多智能体算法协作性能，还能与不同 MARL 算法有效结合，并在实现相似性能的条件下显著降低训练时间。

未来工作将探索该方法在开放场景下与现有多智能体协作方法结合。VIA 有望随着大语言模型的发展应用于多智能体大语言模型的决策过程，以提升其协同推理与行为生成能力。同时，VIA 具备与现有视觉-语言-行动模型结合的潜力，有望在多智能体具身协作等领域拓展应用前景。

参 考 文 献

[1] Ding Shi-Fei, Du Wei, Zhang Jian, et al. Research progress of multi-agent deep reinforcement learning. Chinese Journal of Computers, 2024,47(7):1547-1567 (in Chinese)
(丁世飞,杜威,张健等.多智能体深度强化学习研究进展.计算机学报,2024,47(7):1547-1567)

[2] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in starcraft II using multi-agent reinforcement learning. Nature, 2019,575:350-354

[3] Li Q, Peng Z, Feng L, et al. Metadrive: composing diverse driving scenarios for generalizable reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022,45(3):3461-3475

[4] Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, et al. Survey on multi-agent reinforcement learning for control and decision-making. Acta Automatica Sinica, 2025,51(3):510-539 (in Chinese)
(罗彪,胡天萌,周育豪等.多智能体强化学习控制与决策研究综述.自动化学报,2025,51(3):510-539)

[5] Yuan Lei, Zhang Zi-qian, Li Li-he, et al. Progress on cooperative multi-agent reinforcement learning in open environment. Scientia Sinica Informationis, 2025,55(2):217-298 (in Chinese)
(袁雷,张子谦,李立和等.开放环境下的协作多智能体强化学习进展.中国科学:信息科学,2025,55(2):217-268)

[6] Feng Z, Xue R, Yuan L, et al. Multi-agent embodied ai: advances and future directions. arXiv preprint arXiv:2505.05108, 2025

[7] Fei B, Bao W, Zhu X, et al. Autonomous cooperative search model for multi-uav with limited communication network. IEEE Internet of Things Journal 2022,9(19):19346-19361

[8] Rashid T, Samvelyan M, Schroeder C, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning. The Journal of Machine Learning Research, 2020,21(1):7234-7284

[9] Wang J, Ren Z, Liu T, et al. Qplex: duplex dueling multi-agent q-learning//Proceedings of the International Conference on Learning Representations. Virtual, 2020

[10] Yang T, Wang W, Hao J, et al. Asn: action semantics network for multiagent reinforcement learning. Autonomous Agents and Multi-Agent Systems, 2023,37(2)

[11] Chai J, Li W, Zhu Y, et al. Unmas: multiagent reinforcement learning for unshaped cooperative scenarios. IEEE Transactions on Neural Networks and Learning Systems, 2021,34(4):2093-2104

[12] Sherstinsky A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena, 2020,404:132306

[13] Hu S, Zhu F, Chang X, et al. Updet: universal multi-agent reinforcement learning via policy decoupling with transformers//Proceedings of the International Conference on Learning Representations. Virtual, 2021

[14] Yang Y, Chen G, Wang W, et al. Transformer-based working memory for multiagent reinforcement learning with action parsing//Proceedings of the Neural Information Processing Systems. New Orleans, USA, 2022:34874-34886

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Neural Information Processing Systems. Long Beach, USA, 2017,30:6000-6010

[16] Gallici M, Martin M, Masmitja I. Transfqmix: transformers for leveraging the graph structure of multi-agent reinforcement learning problems//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. London, UK, 2023:1679-1687

[17] Cao J, Yuan L, Wang J, et al. Linda: multi-agent local information decomposition for awareness of teammates. Science China Information Sciences, 2023,66(8):182101

[18] Hu T, Ai X, Pu Z, et al. Heterogeneous observation aggregation network for multi-agent reinforcement learning//Proceedings of the International Joint Conference on Neural Networks. Yokohama, Japan, 2024,1-9

[19] Wang W, Yang T, Liu Y, et al. From few to more: large-scale dynamic multiagent curriculum learning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA., 2020:7293—7300

[20] Nayak I, Choi K, Ding W, et al. Scalable multi-agent reinforcement learning through intelligent information aggregation//Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023,202:25817-25833

[21] Long Q, Zhou Z, Gupta A, et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning//Proceedings of the International Conference on Learning Representations. Virtual, 2020

[22] Zhang T, Xu H, Wang X, et al. Multi-agent collaboration via reward attribution decomposition//Proceedings of the International Conference on Learning Representations. Virtual, 2021

[23] Guan C, Chen F, Yuan L, et al. Efficient communication via self-supervised information aggregation for online and offline multiagent reinforcement learning. IEEE Transactions on Neural

Networks and Learning Systems, 2025,36(5):9044-9056

[24] Huang Kai-qi, Xing Jun-liang, Zhang Junge, et al. Intelligent technologies of human-computer gaming. Scientia Sinica Informationis, 2020,50(4):540-550 (in Chinese)
(黄凯奇,兴军亮,张俊格等.人机对抗智能技术.中国科学:信息科学,2020,50(4):540-550)

[25] Raileanu R, Denton E, Szlam A, et al. Modeling others using oneself in multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018.4257-4266

[26] Yuan L, Wang J, Zhang F, et al. Multi-agent incentive communication via decentralized teammate modeling//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022,36(9):9466-9474

[27] Etel E and Slaughter V. Theory of mind and peer cooperation in two play contexts. Journal of Applied Developmental Psychology, 2019,60:87-95

[28] Wang Y, Zhong F, Xu J, et al. Tom2c: target-oriented multi-agent communication and cooperation with theory of mind//Proceedings of the International Conference on Learning Representations. Virtual, 2021

[29] Li S, Xu R, Xiu J, et al. Robust multi-agent reinforcement learning by mutual information regularization. IEEE Transactions on Neural Networks and Learning Systems, 2025,36(10):18118-18132

[30] Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization//Proceedings of the International Conference on Learning Representations. Virtual, 2020

[31] Li L, Tang H, Yang T, et al. Pmic: improving multi-agent reinforcement learning with progressive mutual information collaboration//Proceedings of the International Conference on Machine Learning. Baltimore, USA., 2022,162:12979-12997

[32] Kim W, Jung W, Cho M, et al. A variational approach to mutual information-based coordination for multi-agent reinforcement learning//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. London, UK, 2023:40-48

[33] Ye D, Lu Z. Mutual-information regularized multi-agent policy iteration//Proceedings of the Neural Information Processing Systems. New Orleans, USA., 2023,36:2617-2635

[34] Wang T, Dong H, Lesser V, et al. Roma: multi-agent reinforcement learning with emergent roles//Proceedings of the International Conference on Machine Learning. Virtual, 2020, 119:9876-9886

[35] Li C, Wang T, Wu C, et al. Celebrating diversity in shared multi-agent reinforcement learning//Proceedings of the Neural Information Processing Systems. Virtual, 2021,34:3991-4002

[36] Liu Quan, Shi Mei-long, Huang Zhi-gang, et al. Multi-agent collaborative reinforcement learning method based on bi-view modeling. Chinese Journal of Computers, 2024,47(7):1582-1594 (in Chinese)
(刘全,施眉龙,黄志刚等.基于双视角建模的多智能体协作强化学习方法.计算机学报,2024,47(7):1582-1594)

[37] Oliehoek F, Amato C. A concise introduction to decentralized pomdps, volume 1. Springer, 2016

[38] Son K, Kim D, Kang W, et al. Qtran: learning to factorize with transformation for cooperative multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019,97:5887-5896

[39] Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015,45:141-149

[40] Ha D, Dai A, Le Q. Hypernetworks//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017

[41] Samvelyan M, Rashid T, Schroeder C, et al. The starcraft multi-agent challenge//Proceedings of the International Conference on Autonomous Agents and Multi Agent Systems. Montreal, Canada, 2019,2186-2188

[42] Papoudakis G, Christianos F, Schafer L, et al. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks//Proceedings of the Neural Information Processing Systems. Virtual, 2021

[43] Ellis B, Cook J, Moalla S, et al. Smacv2: an improved benchmark for cooperative multi-agent reinforcement learning//Proceedings of the Neural Information Processing Systems. New Orleans, USA , 2023,36:37567-37593

[44] Enrico M, Andrea B, Rupali B, et al. On stateful value factorization in multi-agent reinforcement learning//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Michigan, USA, 2025:1445-1453

WEI Wei, Ph. D., professor. His research interests include reinforcement learning, multi-agent system, and large language model.

SONG Hui-Zhong, Ph. D. candidate. His research interests include multi-agent reinforcement learning.

LI Lin, Ph. D., associate professor. Her main research interests include reinforcement learning and unmanned system.

ZHANG Li-Jun, Ph. D. candidate. His research interests include reinforcement learning and large language model.

MA Yi, Ph. D., lecture. His research interests include reinforcement learning and embodied AI.

HAO Jian-Ye, Ph. D., professor. His research interests include reinforcement learning and embodied AI.

LIANG Ji-Ye, Ph. D., professor. His research interests include machine learning and artificial intelligence.

Background

Multi-Agent Reinforcement Learning (MARL) enables multiple agents to independently learn policies from a common environment by interacting with it for a particular goal. To tackle this problem, each agent must not merely optimize its own decisions based on local observations, but also take into account the non-stationarity caused by the policies of allies. Significant traction has been gained for MARL in complex fields. Nevertheless, correctly reasoning about allies' behavioral intentions presents a major obstacle for agents. Existing methods are either using one neural network that encodes the complete observation or using unified encoding modules for different information subsets, so that distinct modeling is enabled. However, these methods generally capture and aggregate allies' observations as deterministic inputs, failing to capture uncertainty and diversity of their behaviors.

In this paper, we propose a module, Variational Information Aggregation (VIA), that enables agents to better perceive and utilize the collaborative intents of allies. VIA learns a probabilistic model of the agent's state and its allies' observations, and samples from the model to explicitly represent the uncertainty in the allies' intentions. Then, we propose a collaboration-intent-action semantic regularization mechanism to improve the accuracy of predictions of allies' behavior. The experimental results indicate that VIA improves prediction accuracy by modeling uncertainty in allies' behavioral intentions explicitly and achieves performance comparable to baselines while significantly reducing training time.

This work was supported by the National Natural Science Foundation of China (Nos. 62276160, 62506221), the Shanxi Provincial Natural Science Foundation General Project (No. 202203021211294), and the Shanxi Provincial Overseas Study Fund Project (No. 20240002).