控制与决策
Control and Decision

# 具备可解释性的决策依据自编码多智能体强化学习方法

李佩璋, 费庆, 陈振, 张言军, 王博

引用本文:
李佩璋, 费庆, 陈振, 等. 具备可解释性的决策依据自编码多智能体强化学习方法[J]. 控制与决策, 2025, 40(9): 2748-2758.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0026

---

# 具备可解释性的决策依据自编码多智能体强化学习方法

李佩璋[1]，费 庆[1†]，陈 振[1]，张言军[1]，王 博[2]

(1. 北京理工大学 自动化学院，北京 100081；2. 中船智海创新研究院，北京 100036)

**摘 要:** 深度强化学习已成为无人集群在复杂未知环境中实现自主决策的关键技术方案, 但是, 内部不可解释的“黑盒”结构使得人类难以理解、信任和监督智能体的自主决策, 严重阻碍其在高安全需求领域中的应用. 鉴于此, 提出一种具备可解释性的多智能体强化学习方法. 首先, 设计具备可解释性的狄利克雷变分自编码器, 从隐空间中编码匹配物理语义信息的决策依据概率分布; 然后, 使用门控网络线性混合决策依据生成动作决策; 最后, 在多智能体近端策略优化强化学习网络框架下完成可解释自编码器的集成训练. 所提出方法将智能体的决策表征为若干具备物理含义依据的混合概率分布, 使得人类可通过概率密度直观地理解智能体行为, 并可通过调整门控权重直接干预智能体决策. 仿真对比实验验证了所提出方法的决策性能, 所设计的可视化方法展示了智能体决策的可解释性以及人类干预决策的效果.

**关键词:** 强化学习；可解释性；多智能体系统；变分自编码器；决策依据；门控加权

**中图分类号:** TP391.9 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2025.0026

**引用格式:** 李佩璋, 费庆, 陈振, 等. 具备可解释性的决策依据自编码多智能体强化学习方法 [J]. 控制与决策, 2025, 40(9): 2748-2758.

# Interpretable decision-basis autoencoder for multi-agent reinforcement learning

*LI Pei-zhang*[1]，*FEI Qing*[1†]，*CHEN Zhen*[1]，*ZHANG Yan-jun*[1]，*WANG Bo*[2]

(1. School of Automation， Beijing Institute of Technology， Beijing 100081， China； 2. CSSC Intelligent Innovation Research Institute，Beijing 100036，China)

**Abstract:** Deep reinforcement learning (DRL) has emerged as a pivotal technology for enabling autonomous decision-making in unmanned swarms operating within complex and unstructured environments. Nevertheless, the inherent “black-box” nature and lack of interpretability of DRL models impede human understanding, trust, and oversight of agents’ autonomous behaviors, thereby significantly restricting their deployment in safety-critical applications. To address this challenge, this paper proposes an interpretable multi-agent reinforcement learning approach. Firstly, an interpretable Dirichlet variational autoencoder is designed to encode decision rationales as probability distributions in a latent space aligned with physically meaningful semantics. Secondly, a gating network is employed to generate action decisions by linearly combining the encoded rationales. Lastly, the interpretable autoencoder is integrated into and jointly trained within a multi-agent proximal policy optimization framework. This method represents the agent’s decision-making process as a mixture of probability distributions, each corresponding to interpretable physical semantics, thereby enabling intuitive human understanding of agent behavior through probability densities and allowing direct human intervention by adjusting the gating weights. Comparative simulation experiments validate the decision-making performance of the proposed approach, while the developed visualization techniques demonstrate both the interpretability of agent decisions and the efficacy of human intervention.

**Keywords:** reinforcement learning； interpretability； multi-agent systems； variational autoencoders； decision basis；gated weighting

## 0 引 言

深度强化学习 (deep reinforcement learning, DRL) 技术因其在决策领域展现出的巨大优势和潜力[1-3], 已被视为求解智能无人系统自主决策问题的关键解决方案[4]. 但是, 由于深度学习“黑盒”结构的不可解释性, 人们对于这一算法部署于现实场景仍然抱有

**收稿日期:** 2025-01-07；**录用日期:** 2025-04-28.

**基金项目:** 国家自然科学基金项目 (62495090).

†通信作者. E-mail: feiqing@bit.edu.cn.

本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

警惕和质疑的态度[5], 尤其是敏感性和安全性要求高的军事以及医疗领域. 此外, DRL 的不可解释性使得人类难以理解机器决策, 无法对机器决策进行有效的干预控制和有人/无人协同控制[6-7].

面对这一阻碍, 可解释人工智能技术被逐渐重视并发展, 目前, 已有许多基于视觉解释的方法被广泛应用, 如 LIME[8]、Grad-CAM[9] 和 Kernel SHAP[10] 等. 但是, 这些方法大多用于求解深度学习中的有监督分类问题, 很难直接应用于序列决策问题. 作为序列决策的关键方法, 可解释的 DRL 网络成为获得可信的人工智能决策网络的关键途径[11]. 区别于有监督的深度学习, 强化学习 (reinforcement learning, RL) 通过与环境交互学习任务奖励最大化的策略, 这种独有的特性促使研究人员将对其可解释的研究更多地聚焦于分析环境交互和任务奖励[12], 即通过建立决策模型与环境态势数据间的关系, 或建立决策模型与任务拆解目标或奖励间的关系来理解智能体的行为[13-14]. 还有许多通过构建可解释模型近似智能体策略的方法, 如基于模仿学习的方法[15-16] 和逆强化学习方法[17] 等. 显然, 这些方法本质上只实现了间接解释, 并未实现对决策网络模型结构的直接解释, 即获得透明的内部结构和决策路径[18], 因此, 无法彻底消除人们对 "黑盒" 模型的担忧. 近年来, 为了对 RL 决策进行直接诠释, 研究人员借助反事实、显著性图以及注意力等工具提出了 SAFE-RL[19]、FLS[20] 以及 i-DQN[21] 等方法, 旨在找到对智能体决策过程有突出影响的语义. 然而, 此类方法只能实现解释决策模型本身而无法生成匹配人类理解的因果决策路径. 只有在物理因果模型中组合人类可理解决策依据生成的决策才能与人类的决策理解对齐, 从而容许人类对机器决策进行干预控制, 实现人机协同决策.

此外, 现有对可解释 RL 的研究还较少地考虑在工程应用中的问题, 尤其是在应用部署和多智能体扩展方面. 具体而言, 现有的直接解释方法[19-21] 通常需要对网络架构重新设计, 在实际应用部署时需要承担额外的验证成本和测试开销; 且这些方法[22-23] 也暂未考虑在集群设置下解释 RL 决策, 团队策略的复杂交互为设计可解释的 RL 方法带来了更大的挑战.

综上所述, 为了在多智能体设置中形成以物理语义为决策依据的因果解释策略网络, 并能够以 "插件" 形式直接替换现有多智能体 RL 框架中不可解释的策略网络完成训练, 本文提出一种基于门控狄利克雷变分自编码器的新型多智能体强化学习方法. 具体而言, 使用狄利克雷变分自编码器从匹配不同物理语义的状态隐空间中编码出不同决策依据对于离散动作的影响, 从而实现自编码器由数据驱动学习向物理语义驱动学习的转变; 并使用门控网络计算权重线性组合不同类型的决策依据为离散动作概率. 该变分自编码器能够以 "插件" 形式直接替换多智能体近端策略优化网络中不可解释的动作网络来完成训练, 通过仿真环境中的对比实验验证所提出方法的有效性. 为了直观地显示所提出方法的可解释性, 本文设计可视化方法展示各决策依据影响决策的过程. 此外, 通过设计对照实验验证所提出的可解释性对于引导奖励重塑和人机协同决策的作用. 实验结果表明, 编码可解释语义能够在满足决策任务基本要求的同时诠释决策依据, 为在安全敏感领域中部署智能决策方法提供可行路径.

## 1　预备知识

### 1.1　多智能体马尔可夫决策过程

多智能体强化学习 (multi-agent reinforcement learning, MARL) 是一类算法的统称, 其核心在于多个智能体通过与同伴以及环境的交互学习进行序列决策. 具体而言, 在每个时间步$t$, 智能体基于其观测$o^t$选择一个动作$a^t$, 并从环境中获得一个标量奖励$r^t$, 同时, 环境会更新其观测至$o^{t+1}$. 更为规范地, 一个 MARL 任务可表示为$N$个智能体的马尔可夫决策过程 (Markov decision process, MDP), 记为$(S_1, \ldots, S_N, A_1, \ldots, A_N, \gamma, P, R_1, \ldots, R_N)$元组. 该 MDP 元组包括$t$时刻$N$个智能体的状态$s_n^t \in S_n^t$, 动作$a_n^t \in A_n^t$, 折扣因子$\gamma \in [0,1]$, 状态转移概率$p(s_n^{t+1}|s_n^t, a_n^t)$以及奖励函数$r_n^t \in R_n(s_n^t, a_n^t)$. 每个智能体的行为通过其策略$\pi_n$决定, 策略$\pi_n$将每个状态映射为一个在所有可能动作上的分布. 价值函数$v_n^\pi(s_n^t)$表示从$t$时刻状态$s_n^t$出发, 遵循策略$\pi_n$行动产生的期望折扣累计奖励, 即

$$v_n^\pi(s_n^t) = \mathbb{E}_{\pi_n}\left[\sum_{k=0}^{\infty}\gamma^k r_n^{t+k}\right].$$

类似地, 从状态$s_n^t$出发, 基于策略$\pi_n$执行动作$a_n^t$能够得到状态-动作价值函数$Q^{\pi_n}(s_n^t, a_n^t) = \mathbb{E}_{\pi_n}[R_n^t | s_n^t = s_n, a_n^t = a_n]$, 其中$R_n^t = \sum_{k=t}^{T-1} r_n^{k+1}$为从$t$时刻起的累计奖励.

### 1.2　变分自编码器

自编码器是一种无监督神经网络模型, 常被用于重构输入数据. 与简单的复制输入数据不同, 自编码器在重构过程中能够学习并表示关键特征[24]. 自编码器的结构由编码器和解码器两部分组成, 编码器将输入$x$映射至一个隐变量$z \in \mathbb{R}^{\mathcal{L}}$, 此过程表示为$z = f_\theta(x)$, $\theta$为编码器网络参数; 解码器则从$z$中

重构输入$r$, 该过程表达为$r=g_\omega(z)$, $\omega$为解码器网络参数, 自编码器网络通过最小化输入数据与其重构间的差异惩罚项来实现对网络的训练. 变分自编码器 (variational autoencoder, VAE) 是在自编码器的基础上进行扩展, 将隐变量$z$建模为条件分布$p(z|x)$以增强特征表示能力. 然而, 真实的条件分布通常难以求得, 因此, VAE 使用变分分布$q_\phi(z|x)$近似真实分布$p(z|x)$, 并通过损失函数实现优化, 有

$$\mathcal{L}_{\rm VAE}(x)=\mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_2}(x|z)]-D_{\rm KL}(q_\phi(z|x)\|p_{\theta_1}(z)). \tag{1}$$

其中: $\phi$为编码器网络参数, $\theta_1$为先验概率参数, $\theta_2$为解码器网络参数. 该损失函数由作为正则项的 Kullback-Leibler (KL) 散度$D_{\rm KL}(q_\phi(z|x)\|p_{\theta_1}(z))$和作为期望项的重构误差$\mathbb{E}_{q_\phi(z|x)}[\log p_{\theta_2}(x|z)]$两部分组成. 最后, 在反向传播过程中, 为了解决期望项中存在的随机节点导致的梯度信息丢失问题, 引入重参数化技巧. 通过可微分的方式重新参数化隐变量, 实现编码器参数和解码器参数的联合优化训练.

## 2 决策可解释性问题描述

本文旨在设计具备可解释性的多智能体策略网络, 顺序地接收输入状态$s$决策输出动作$a$, 并通过隐变量$z$, 即决策依据, 诠释决策过程. 该网络基于 VAE 架构设计, 使用$S=\{s_1,s_2,\dots,s_{|S|}\}$表示智能体状态空间, $A=\{a_1,a_2,\dots,a_{|A|}\}$表示智能体动作空间; 对$S$根据物理语义分割后的子集提取隐变量, 生成直接匹配不同物理语义的隐变量集合$Z=\{z_1,z_2,\dots,z_{|Z|}\}$, 表示不同决策依据对于离散动作的影响. 具体而言, 第$i$个物理语义单元由大小为$k$的状态空间子集$\{s_{i:1},s_{i:2},\dots,s_{i:k}\}$的多元概率分布$f_S$描述, 即

$$z_i=f_S^i(s_{i:1},s_{i:2},\dots,s_{i:k};\alpha_i),\ i=1,2,\dots,|Z|, \tag{2}$$

其中$\alpha_i$为条件概率参数群. 智能体的离散动作概率由网络$g_Z$混合各决策依据计算产生, 有

$$a_j=g_Z^j(z_1^j,z_2^j,\dots,z_{|Z|}^j),\ j=1,2,\dots,|A|. \tag{3}$$

具体而言, 本文考虑多智能体自主路径规划任务中多架无人机自主规划路线完成任务目标点全覆盖问题, 任务目标点位置随机刷新, 智能体间的碰撞会导致任务失败. 基于此, 根据人类理解设置智能体决策依据为任务目标吸引作用和同伴避碰排斥作用, 智能体的动作为不同方向的离散概率. 在这一设置下, 本文旨在设计一种全新的动作网络来实现可解释的智能体决策, 其可解释性体现于: 1) 能够训练出任务目标吸引和同伴避碰排斥两种匹配物理语义决策依据的多元概率分布$f_S$; 2) 能够训练出线性组合两种决策依据的权重网络$g_Z$为智能体输出动作决策. 特别地, 该动作网络可在 MARL 框架下完成参数训练.

## 3 可解释狄利克雷变分自编码器动作网络

基于上述决策可解释问题描述, 需要设计可解释的 VAE 网络从匹配不同物理语义的状态隐空间中提取不同决策依据对离散动作影响的概率分布, 并组合输出可解释的动作决策. VAE 提取的潜在特征本质上并不具备可解释性, 因此, 需要进行概念调整使其能够编码有意义的语义, 即将数据驱动学习的编码器-解码器架构转变为以专家知识 (物理概念) 为基础的编码器-解码器架构. 这种设置迫使编码器提取具有语义意义的潜在参数, 隐空间将与确定性映射中的参数具有相应的语义, 只要隐空间具有语义解释, 就不能任意地选择潜在先验[25].

传统 VAE 使用高斯分布建模隐变量分布, 这种连续分布描述难以用于描述具有离散概率且语义抽象的决策依据, 需要使用概率范围在$0\sim1$之间的离散随机分布代替连续高斯分布来完成建模. 同时, 所选择随机分布需要满足可解释决策依据的“多元总和归一”性质, 即决策依据对于不同动作的影响为一组概率和为 1 的随机变量. 狄利克雷分布能够通过单纯形 (simplex) 表示概率密度而拥有这一特性, 因此, 本文使用狄利克雷分布来构造 VAE, 并进一步设计可解释狄利克雷变分自编码器 (interpretable Dirichlet-based variational autoencoder, iDVAE) 动作网络. iDVAE 构造与决策依据数量相同的编码器, 编码任务目标吸引作用相关状态$S_{\rm Att.}$和同伴避碰排斥作用相关状态$S_{\rm Rep.}$生成决策依据的概率分布$f_S$. 对概率分布重参数化采样得到决策依据$Z_{\rm Att.}$和$Z_{\rm Rep.}$, 并在混合网络$g_Z$中混合各依据来计算智能体动作, 最后, 通过解码器计算输入状态重构$\hat{S}_{\rm Att.}$和$\hat{S}_{\rm Rep.}$. 网络通过对概率分布参数计算 KL 散度, 对智能体动作计算价值以及对重构状态计算误差完成训练, iDVAE 结构如图 1 所示.

iDVAE 的训练包括无监督的自编码器训练和有监督的强化学习训练两部分. 其中: 强化学习训练通过动作价值完成, 将在后文第 4 节说明; 自编码器训练通过 KL 散度和重构误差完成, 可描述为编码器学习$q_{\phi_*}(Z_*|S_*)$逼近真实后验分布$P(Z_*|S_*)$(使用$*$表示 Att.和 Rep.两部分), 解码器学习$p_{\theta_*}(S_*|Z_*)$生成数据. 状态的边际似然表示为$p_{\theta_*}(S_*)=\sum_{Z_*}[p_{\theta_*}(S_*|Z_*)p(Z_*)]$, 为了最大化似然, 由 Jensen 不等式可得到$\log p_{\theta_*}(S_*|Z_*)\geqslant\mathcal{L}(\theta_*,\phi_*;S_*)$, 这里$\mathcal{L}(\theta_*,\phi_*;S_*)$为变分下界 (evidence lower bound,

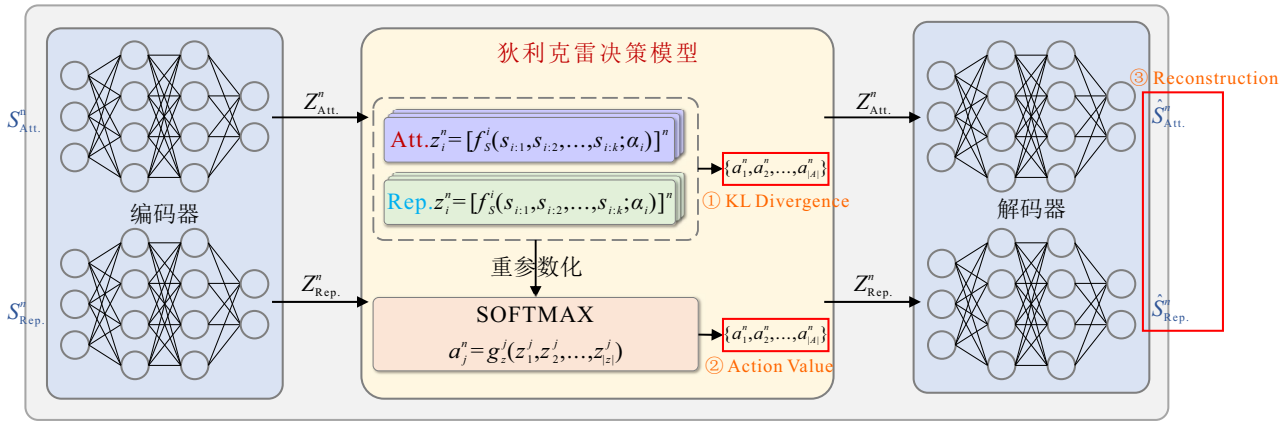



**图1　可解释狄利克雷变分自编码器网络架构**

ELBO).

在离散动作空间问题中, 智能体的动作表示为0 ~ 1之间的离散概率, 因此, 考虑使用狄利克雷分布描述这一离散概率分布, 即

$$q_{\phi_*}(Z_*|S_*)=\frac{1}{B(f(S_*))}\prod_{k=1}^{K}(Z_*)_k^{f(S_*)_k-1}. \tag{4}$$

其中: $\{(Z_*)_k\}_{k=1}^{k=K}$属于标准$K$-1阶单纯形, 即$\sum_{i=1}^{K}(Z_*)_i=1$且对于全体$i\in\{1,2,\ldots,K\}$有$(Z_*)_i\geqslant 0$; $f(\cdot)$表示将输入状态变换为狄利克雷参数, 可通过神经网络实现; 分母上用于归一化的常数为多元Beta函数, 即

$$B(f(S_*))=\frac{\prod_{k=1}^{K}\Gamma(f(S_*)_k)}{\Gamma\left(\sum_{k=1}^{K}f(S_*)_k\right)}. \tag{5}$$

与传统基于高斯分布设计的VAE不同, 使用狄利克雷分布难以使用“重参数化”技巧, 因此需要设计替代先验分布, 一种简单有效的方法是对狄利克雷构建拉普拉斯近似[26], 这种方法首先选用Softmax函数$\sigma(\cdot)$替代单纯形, 因此, 概率分布函数能够使用新的变量$u$表示为

$$q_{\phi}(\sigma(u)_*|S_*)=\frac{\prod_{k=1}^{K}(\sigma(u)_*)_k^{f(S_*)_k}g(1^{\mathrm{T}}u)}{B(f(S_*))}, \tag{6}$$

其中$g(\cdot)$为一个任意密度函数, 用于通过限制冗余自由度来确保可积性.

进一步地, Dirichlet先验$p(\sigma(u)_*|\alpha)$能够通过Laplace近似表示为变量$u$的多元Logistic-Normal分布$\hat{p}(\sigma(u)_*|\mu_0,\Sigma_0)=\mathcal{LN}(\sigma(u)_*|\mu_0,\Sigma_0)$. 在本文中, 潜在先验被设定为决策依据对于不同离散动作的影响, 因此, 近似多元Logistic-Normal分布的维度与动作空间维度同为$|A|$, 先验分布均值$\mu_0$和协方差$\Sigma_0$计算如下所示:

$$\mu_{1;k}=\log\alpha_k-\frac{1}{|A|}\sum_i\log\alpha_i, \tag{7}$$

$$\Sigma_{1;kk}=\frac{1}{\alpha_k}\left(1-\frac{2}{|A|}\right)+\frac{1}{|A|^2}\sum_i\frac{1}{\alpha_i}. \tag{8}$$

ELBO能够展开为重构误差项和正则化(KL散度)项, 即

$$\mathcal{L}(\theta_*,\phi_*;S_*)=\mathbb{E}_{Z_*\sim q_{\phi}(Z_*|S_*)}[\log p_{\theta_*}(S_*|Z_*)]-D_{\mathrm{KL}}(q_{\phi_*}(Z_*|S_*)\|p(Z_*)). \tag{9}$$

狄利克雷分布下的KL散度计算[26]表示为

$$D_{\mathrm{KL}}=\frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0)+\log\frac{|\Sigma_1|}{|\Sigma_0|}-|A|\right)+\frac{1}{2}(\mu_1-\mu_0)^{\mathrm{T}}\Sigma_1^{-1}(\mu_1-\mu_0). \tag{10}$$

其中: $\mu_1$和$\Sigma_1$分别为近似多元Logistic-Normal条件分布的均值和方差, 能够通过给定超参数$\alpha$后计算求得.

可解释狄利克雷变分自编码器算法伪代码归纳如算法1所示.

算法1　可解释狄利克雷变分自编码器.

输入: 智能体观测状态$S$;

输出: 智能体动作概率$Z$, 变分下界ELBO.

基于物理先验分割观测: $S=\{S_*\}=\bigcup(\{S_{\mathrm{Att.}}\},\{S_{\mathrm{Rep.}}\})$

for all $S_*$ do

$\mu_\phi(S_*),\Sigma_\phi(S_*)\leftarrow\mathrm{Encoder}(S_*;\phi)$

$\epsilon\sim N(0,I)$

重参数化采样: $Z_*\leftarrow\mu_\phi(S_*)+\sigma_\phi(S_*)\odot\epsilon$

$Z'_*\leftarrow\mathrm{softmax}(Z_*)$

end for

计算重构状态: $\hat{S}\leftarrow\mathrm{Decoder}\left(\sum_* Z'_*;\theta\right)$

计算重构损失: $L_{\rm RE} \leftarrow {\rm MSE}(S, \hat{S})$

计算 Dirichlet 先验均值: $\mu_0 \leftarrow \log \alpha_0 - \dfrac{1}{|A|}\sum\limits_i \log(\alpha_0)_i$

计算 Dirichlet 先验方差: $\Sigma_0 \leftarrow \dfrac{|A|-2}{\alpha_0 \cdot |A|} + \dfrac{1}{|A|^2}\sum\limits_i \dfrac{1}{(\alpha_0)_i}$

由式 (10) 计算KL损失$L_{\rm KL}$

计算ELBO: ELBO $\leftarrow L_{\rm RE} - L_{\rm KL}$

计算采样隐变量集合: $Z' = \{Z'_*\} = \bigcup(\{Z'_{\rm Att.}\}, \{Z'_{\rm Rep.}\}$

返回 $Z'$, ELBO

## 4 门控狄利克雷变分自编码器多智能体近端策略优化强化学习网络

通过 iDVAE, 实现了从观测状态中编码可解释的决策依据. 在此基础上, 需要基于物理因果混合决策依据为动作决策, 同时, 为了实现对决策过程中各决策依据的干预和控制, 需要在混合模型中引入可控权重. 基于上述原因, 本文设计了门控决策依据混合网络计算动作决策. 网络接收状态输入, 输出不同决策依据的重要度权重系数, 并以符合物理因果的概率混合策略加权混合决策依据为动作决策. 最后, 将 iDVAE 和门控决策依据混合网络集成于 MARL 框架中来完成联合训练. 本文旨在构造可解释的策略网络以理解 RL 智能体的决策, 因此, 需要集成在基于策略优化的 MARL 框架中完成训练, 多智能体近端策略优化 (multi-agent proximal policy optimization, MAPPO) 网络作为典型的基于策略优化的 MARL 网络, 以其在离散动作空间卓越的决策性能和泛化性, 被广泛应用于各类工程场景. 因此, 本文以 MAPPO 用于基准网络为例来说明门控狄利克雷变分自编码器多智能体近端策略优化强化学习网络的训练过程, 该算法可直接推广至其他基于策略优化的离散 MARL 网络中.

基于第 3 节所述 iDVAE 的拉普拉斯近似, 对编码器产生的概率分布进行重参数化采样, 有

$$Z_* = r_\phi(S_*, \epsilon) = \mu_\phi(S_*) + \Sigma_\phi^{1/2}(S_*) \odot \epsilon, \quad (11)$$

其中$\epsilon \sim \mathcal{N}(0, I)$. 对重参数化采样得到的决策依据$Z_*$使用 softmax 变换后组合为动作概率, 引入 Hadamard 乘积组合集合内独立的决策依据为动作概率$A$. 基于 Hadamard 乘积的概率混合能够在随机概率背景下近似实现具备物理模型概念的混合模型, 即使用大概率表示有较大数值的力, 小概率表示有较小数值的力, 混合概率表示合力计算. 同时, 通过门控网络$G(S;\psi)$计算的权重对各决策依据加权, 即

$$A = \mathbb{N}(Z_{\rm Att.}^{G_{\rm Att.}(S;\psi)} \times Z_{\rm Rep.}^{G_{\rm Rep.}(S;\psi)}). \quad (12)$$

这里: $\mathbb{N}$表示向量归一化; 门控网络基于 MLP 构造, 使用$\psi$表示其参数, 在接收全部状态数据后计算输出不同决策依据的归一化权值.

使用 iDVAE 和门控网络构造 Actor 网络生成第$n$个智能体动作策略$\pi_{\theta_n,\phi_n}(A_n^t|S_n^t)$, 接收状态$S_n^t$计算动作概率$A_n^t$. 使用 MLP 构造 Critic 网络和 Target-Critic 网络, 利用全部智能体的状态动作计算$t$时刻动作价值$q_n^t = C(s^t, a^t; \omega_n)$和下一时刻动作价值$q_n^{t+1} = C(s^{t+1}, a^{t+1}; \hat{\omega}_n)$. 根据 TD($\lambda$)误差$\delta_t$近似计算优势函数$\hat{A}_n^t = \sum\limits_{l=0}^{T-t}(\gamma\lambda)^l \delta_n^{t+l}$, 其中$\lambda$为 GAE 中的权重参数. 随后, 定义策略更新比例$\rho^t(\theta_n, \phi_n, \psi_n) = \pi_{\theta_n,\phi_n,\psi_n}(a_n^t|s_n^t)/\pi_{[\theta_n,\phi_n,\psi_n]_{\rm old}}(a_n^t|s_n^t)$. Critic 网络通过 TD 目标计算损失完成训练, 损失函数定义为

$${\rm Loss}_C(\omega^n) = \mathbb{E}[(q_n^t - (r_n^t + \gamma\cdot(1-d_n^t)\cdot q_n^{t+1}))^2]. \quad (13)$$

Actor 网络通过 clip 函数计算损失完成训练, 同时, 需要考虑最大化全部 iDVAE 的联合 ELBO, 因此, 定义 Actor 网络损失函数如下所示:

$$\begin{aligned}&{\rm Loss}_A(\theta_n, \phi_n, \psi_n) = \\ &\eta \cdot \mathcal{L}(\theta^n, \phi^n; x) + \mathbb{E}_t[\min(\rho_t \cdot \hat{A}_n^t, \\ &{\rm clip}(\rho_t, 1-\epsilon, 1+\epsilon)\cdot \hat{A}_n^t)], \end{aligned} \quad (14)$$

这里$\eta$为缩放系数. Target Critic 网络通过软更新完成训练, 即

$$\hat{\omega}_n^{\rm new} = \tau \cdot \omega_n^{\rm new} + (1-\tau)\cdot \hat{\omega}_n. \quad (15)$$

iDVAE-MAPPO 网络训练框架如图 2 所示.

## 5 仿真实验分析

本节主要对所提出可解释多智能体强化学习网络进行性能评估, 并在此基础上分析和验证该网络的可解释性和可干预性.

### 5.1 仿真场景设置

本文基于 OpenAI 公开的基准强化学习测试环境 MPE[27] 开发仿真验证场景, 以实现第 2 节描述的多智能体路径规划问题. 在该场景中, 3 架无人机在避免相互碰撞的前提下合作占领 3 处目标点, 状态空间为 14 维向量: 自身位置 (2 维), 自身速度 (2 维), 相对目标点的位置 (6 维) 以及相对同伴的位置 (4 维); 离散动作空间为 5 维向量: 保持禁止, 向右加速, 向左加速, 向上加速以及向下加速; 决策依据隐空间为$Z_{\rm Att.}$和$Z_{\rm Rep.}$对 5 维动作空间不同分量的影响, 即$Z \in \mathbb{R}^{2\times 5} = \{Z_{\rm Att.} \bigcup Z_{\rm Rep.}\}$. 无人机使用集中式训练分布式执行决策, 第$i$架无人机$a_i$的奖励函数$r_i$为

$$\begin{aligned} r_i = &-c_1 \cdot \sum_j \min_i \|a_i - o_j\| + \\ &- c_2 \cdot \sum_{j\neq i} \mathbf{1}(\|a_i - a_j\| \leqslant \delta). \end{aligned} \quad (16)$$

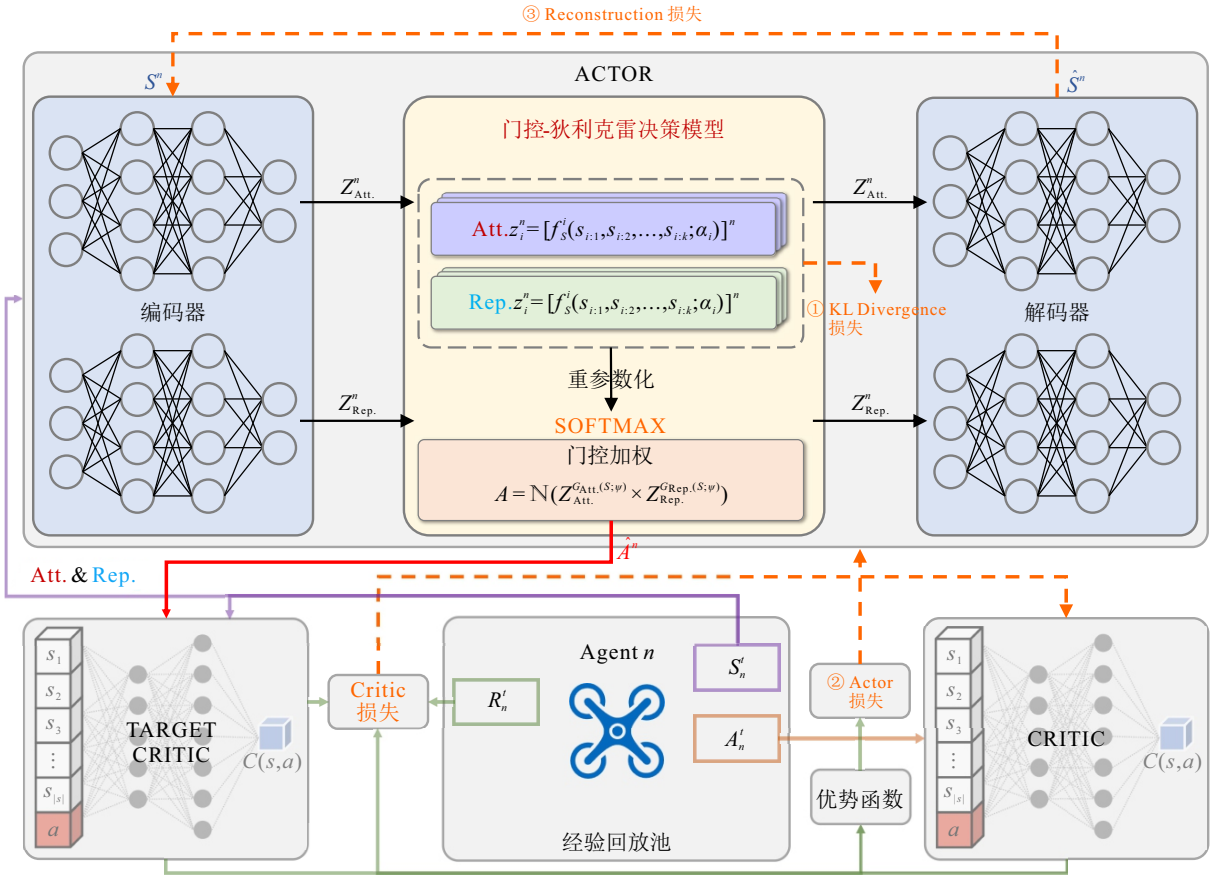



图2　门控狄利克雷变分自编码器多智能体近端策略优化强化学习网络训练框架

其中: $c_1$和$c_2$为大于 0 的超参数, $o_j$为第$j$个目标点, $\mathbf{1}(\cdot)$表示是否发生碰撞的指示函数.

仿真实验使用的超参数如表 1 所示.

表1　仿真实验使用超参数

| 参数 | 取值 | 参数 | 取值 |
|---|---|---|---|
| 学习率$l_r$ | 5e-4 | 奖励参数$c_1$ | 1.2 |
| 折扣因子$\gamma$ | 0.99 | 奖励参数$c_2$ | 1.0 |
| GAE加权参数$\lambda$ | 0.95 | 奖励参数$\delta$ | 0.02 |
| Dirichlet先验$\alpha_0$ | 0.3 | 软更新参数$\tau$ | 0.01 |
| 策略剪切$\epsilon$ | 0.2 | 缩放系数$\eta$ | 1e-4 |

## 5.2　路径规划性能分析

所设计 iDVAE 能够以“插件”形式便捷地集成至基于 Actor-Critic 架构的多智能体强化学习网络中, 即在不改变原有 Actor 网络的输入和输出条件下, 直接将不可解释的 MLP-Actor 网络替换为可解释的 iDVAE-Actor 网络. 本文使用 MLP-MAPPO[28] 作为基础网络, 通过对比该基准网络与替换后的 iDVAE-MAPPO 网络在训练和推理阶段的表现来验证 iDVAE 的决策性能. 此外, 额外增加 Baseline 对照网络 MADDPG[27]; 增加同样基于构造可解释语义实现直接解释的对照网络 CPM (concept policy model)[21]. 同时, 为确保可比性, 本文对 CPM 编码的可解释概念进行了设计, 将其设置为最近目标点吸引和最近同伴避障, 从而与所使用的决策依据在可解释语义上对齐. 与 CPM 的对比包括硬概念 (完全基于可解释概念决策) 和软概念 (基于可解释概念和不可解释残差共同决策) 两部分. 图 3 为 iDVAE-MAPPO、MLP-MAPPO、MLP-MADDPG、CPM-HARD 和 CPM-SOFT 在仿真场景中的训练奖励表现.

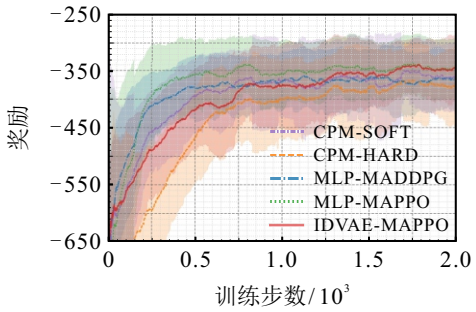



图3　各方法在仿真场景中的训练奖励表现

由图 3 可见: 所提出 iDVAE-MAPPO 网络在稳态性能方面与基准网络 MLP-MAPPO 相当, 优于对照网络 MLP-MADDPG. 相较于不可解释的基准网络, iDVAE-MAPPO 因需要额外从状态空间的可解释子空间提取决策依据, 在收敛速度上表现出一定的下降. 这种收敛时延能够被用于定量描述可解释性与决策效率间的权衡关系, 但是, iDVAE 的可解

释性并不影响稳态性能. 与可解释方法 CPM 的对比结果表明, 在设置本文使用的可解释概念的条件下, 完全基于可解释概念进行决策的硬概念方法 (CPM-HARD) 的决策性能欠佳, 弱于所提出方法与基准方法. 引入残差后的软概念方法 (CPM-SOFT) 能够在一定程度上提升性能并在最后接近所提出方法和基准方法, 但是, 其可解释性已经遭到了破坏, 其决策的产生并不是完全来自于可解释的概念. 上述对比实验结果有效验证了所提出方法在具备可解释性的同时, 仍然能够保持良好的决策性能.

进一步地, 为验证所提出方法在更大规模问题场景中的可扩展性, 在 5 智能体和 10 智能体环境中与基准网络进行路径规划性能对比. 表 2 为不同智能体规模场景中 iDVAE-MAPPO 与基准网络在推理阶段的平均奖励对比. 由表 2 对比结果可见, iDVAE-MAPPO 网络在不同规模的环境中均展现出与基准方法相当甚至更优的路径规划性能, 且在规模扩大时表现愈加突出, 有效表明其良好的可扩展性.

**表2 不同智能体规模场景中各方法的推理奖励对比**

| 推理算法 | Env: Agent-3 | Env: Agent-5 | Env: Agent-10 |
|---|---|---|---|
| IDVAE-MAPPO | −348.05 | −1 471.95 | −10 794.94 |
| MLP-MAPPO | −347.64 | −1 469.02 | −10 918.52 |
| MLP-MADDPG | −364.67 | −1 739.57 | −11 777.69 |

对所提出方法进行敏感度分析以及消融实验来验证其有效性, 本节表明对路径规划性能的影响, 对可解释性的影响将在第 5.3 节表明. iDVAE 通过训练编码器实现产生逼近真实的后验分布, 因此, 不同先验参数的选择会对网络决策性能以及功能实现造成影响. 基于此, 对 iDVAE 的先验分布进行了参数敏感度分析, 以验证不同先验分布参数的选择对于网络性能的影响. 在 iDVAE 中, 由于狄利克雷分布的共轭先验特性, 其分布与后验分布有着相同的形式, 不同的狄利克雷先验分布参数$\alpha_0$会直接影响分布形状, 并导致其可解释性语义出现明显区别: $\alpha_0$越接近 1, 先验分布越均匀; 反之, $\alpha_0$越接近 0, 则先验分布越稀疏. 图 4 为对狄利克雷先验进行敏感性分析的实验结果.

实验中, 将狄利克雷先验分布参数$\alpha_0$分别设置为 0.1、0.3、0.5、0.8 和 1.0 进行训练, 对应的奖励曲线如图 4 所示. 由图 4 可见, 本文选择的 0.3 先验参数具有最优的性能. 进一步地, 当$\alpha_0 = 1.0$时, 狄利克雷分布将退化为均匀分布. 此时, 编码的决策依据将不具备狄利克雷分布特性, 能够直接完成对所使

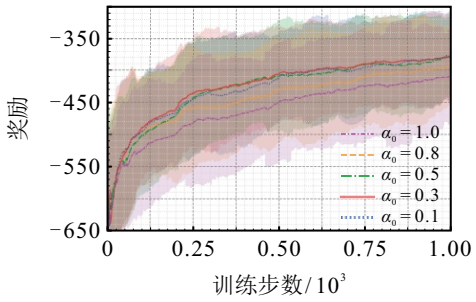



**图4 狄利克雷先验参数敏感度分析对比**

用狄利克雷分布自编码器的消融验证. 消融实验结果表明, 各可解释决策依据有着稀疏的特征, 当对狄利克雷分布消融为均匀分布时, 无法有效完成决策任务.

### 5.3 可解释性分析

本节通过分析多智能体系统使用 iDVAE-MAPPO 网络推理时的运动过程表明其可解释性, 图 5 为 3 架无人机在仿真场景中的运动过程, 使用$T_5 \sim T_{25}$表示推理过程中间时刻智能体位置. 智能体的决策由两种依据的门控混合概率分布决定, 通过门控权重可直观地得到智能体决策过程中各决策依据的重要程度, 基于此, 构造决策依据混合系数 (decision-basis mixing factor, DMF) 变化曲线. 图 6 为 3 架无人机推理阶段的 DMF 曲线. 其中: DMF-1 为任务目标决策依据, DMF-2 为同伴避碰决策依据. 结合图 5 对图 6 进行如下可解释分析.

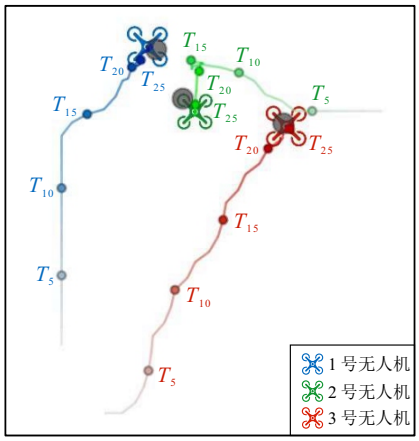



**图5 多智能体在仿真场景中的运动过程**

1) 1 号无人机在初始时由于与 3 号无人机距离较近, 同伴避碰依据有着较大的权重. 出发后至第$T_{13}$步, 任务目标依据权重不断增大, 同伴避碰依据权重不断降低, 第$T_{13}$步后, 考虑接近任务目标以及与其他无人机的碰撞风险, 分别减小和增大权重, 如图 5 所示轨迹中出现拐点;

2) 2 号无人机初始时的位置几乎没有碰撞风险, 主要受目标吸引影响, 随着推理进行靠近目标点, 目

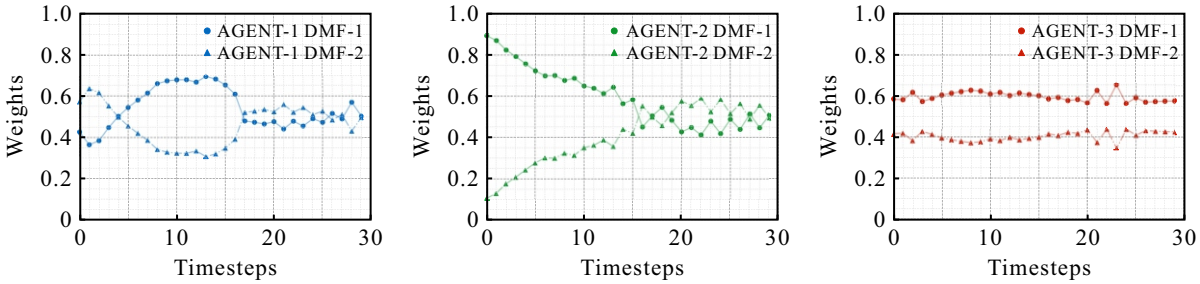



**图6　推理阶段决策混合系数变化曲线**

标吸引依据权重不断减小, 同伴避碰依据权重不断上升, 并在$T_{15}$步显著上升, 对照图 5 中的轨迹拐点可以看出, 因防止与 1 号无人机对任务目标的争抢碰撞切换了目标;

3) 3 号无人机的运动过程中没有碰撞风险, 两类依据权重变化平稳, 总体保持目标依据权重下降, 同伴避碰依据权重上升趋势.

为了在理解智能体决策依据权重变化的基础上进一步可视化决策依据对于智能体动作的影响, 设计决策依据概率分布 (secision-basis probability density, DPD) 变化曲线, 如图 7 所示. 图 7 以推理阶段$T_{14}$切换至$T_{15}$步过程为例展示 DPD 变化. 图 7 中: 每行的 5 幅子图表示一架无人机的 5 维动作空间, 红色、蓝色以及灰色曲线分别为任务目标依据、同伴避碰依据以及综合依据. DPD 基于对狄利克雷分布拉普拉斯近似后的均值和方差绘制, 并对极小值方差进行归一化处理以实现更好的可视化效果, 具体而言, 对于任意动作维度的 DPD 曲线, 其形状表示决策依据在该动作维度上的概率分布, 即峰值所在的横坐标为决策依据对该动作分量的最大影响概率, 综合依据表示该动作分类概率. 此外, 由于训练后的方差接近, 峰值直接等价于决策依据门控权重. 因此, DPD 曲线峰值越高综合依据越靠近该决策依据曲线, 均值越大该决策依据影响越大, 人类能够通过曲线分布直观地理解智能体决策依据. 以 2 号无人机 (第 2 行) 为例: 在向右加速子图中, 任务目标 DPD

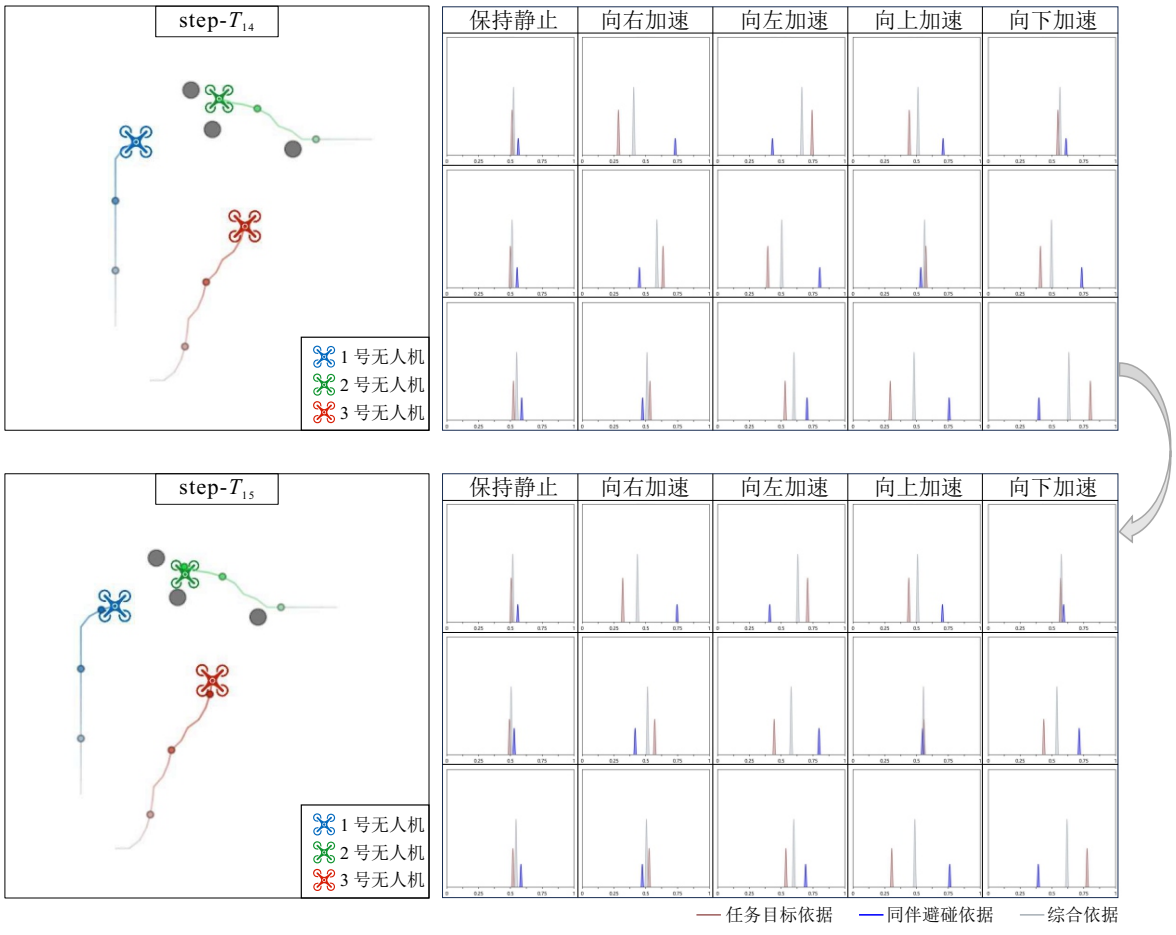



**图7　推理阶段决策依据概率密度变化曲线变化**

左移且峰值降低、同伴避碰 DPD 峰值增加, 导致向右加速动作概率下降; 相反, 在向左加速子图中, 同伴避碰 DPD 峰值增加、任务目标 DPD 右移, 使得向左加速动作概率上升. 同样地, 因同伴避碰 DPD 峰值的增加和任务目标 DPD 的右移, 2 号无人机向下加速的动作概率进一步增加. 无人机的 DPD 曲线变化规律与左侧仿真场景中的动作模式直接匹配, 类似地, 可对 1 号无人机和 3 号无人机进行分析.

### 5.4 可解释性对奖励函数重塑的引导分析

在本文所述问题中, 智能体同时拥有同伴和目标点状态信息, 使得集群中同时存在合作和竞争关系: 追求任务目标全覆盖时表现为合作关系, 追求优先同伴到达目标点时表现为竞争关系. 在这种混合关系作用下, 解释多智能体的决策对于改进强化学习网络奖励函数的设计、提升网络决策性能有着显著的意义.

通过第 5.3 节对于 DMF 和 DPD 曲线的分析可知, 3 架无人机在推理过程中持续维持一定比例的同伴避碰依据权重和概率密度, 然而, 以人类的决策理解对智能体运动过程分析, 绝大多数情况下智能体间不存在碰撞风险, 这表明智能体的决策模型存在过于保守的倾向. 基于该可解释性分析结果, 减小奖励函数中的碰撞损失系数后对模型进行重新训练. 此外, 为强化对比验证效果, 实验环境设置为高密度任务场景来增加潜在碰撞概率. 图 8 为奖励函数优化前后模型在相同仿真环境中的运动过程对比: 原始模型因过于考虑同伴避碰依据导致目标抵达率下降; 而重塑后的模型能够有效完成任务, 累计奖励显著提升.

### 5.5 可解释性对人机协同决策的指导分析

iDVAE-MAPPO 的门控机制提供给人类指挥员直接干预智能体决策的通道, 指挥员可实现全局调节决策偏好以及局部干预决策行为. 全局调节是指人类对智能体策略模式进行宏观的预设, 使得无人机集群在运动过程中展现出对某一类决策依据的偏好. 图 9 为人为修改决策依据权重对于无人机全局运动过程的影响. 其中: 图 9(a) 为任务目标依据与同伴避碰各 50% 权重下的无人机集群运动过程, 图 9(b) 为同伴避碰依据 70% 权重、任务目标 30% 权重下的无人机运动过程, 图 9(c) 为任务目标依据 70% 权重、同伴避碰 30% 权重下的无人机集群运动过程.

由图 9 仿真结果可见, 不同决策依据权重对无人机集群的决策偏好有着显著的影响, 进一步导致整体的运动策略与结果的不同. 在同伴避碰决策依据有较大权重的设置下, 无人机会以尽可能减少碰撞可能的方式规划任务目标; 而在任务目标决策依据有较大权重设置下, 无人机会优先选择更近的目标, 只有发生较大碰撞威胁的情况下才会切换任务目标.

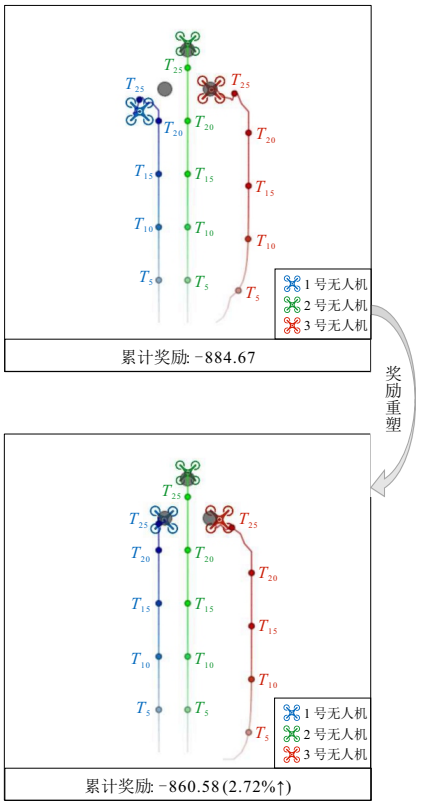


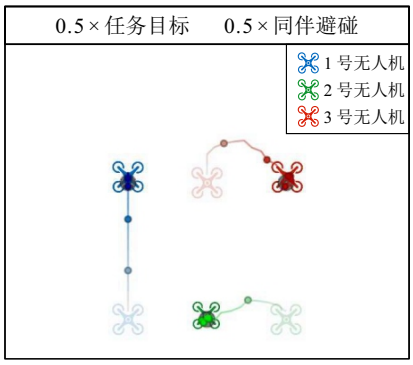

**图8 修改奖励重新训练后的智能体运动过程**

局部干预是指人类实时地对智能体策略进行调整以实现对决策行为的控制, 图 10 为$T_{15}$时刻突加干预信号, 将智能体策略从任务目标依据 70% 权重降低为 30% 权重, 并将同伴避碰依据权重提升后的运动过程. 由图 10 仿真结果可见, 智能体在$T_{15}$时刻前主要受到任务目标决策依据影响, 此时, 人类指挥员在观测到蓝色无人机与绿色无人机有碰撞风险, 实时增加同伴避碰权重控制智能体决策行为, 使得绿色无人机切换任务目标的选择.

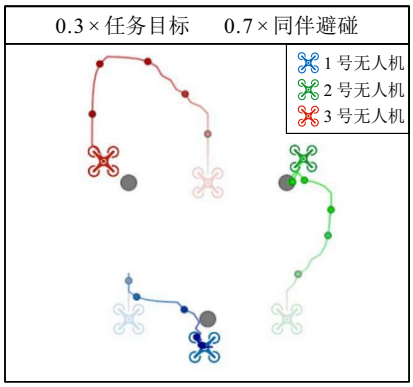
综上所述, 全局调节和局部干预的仿真实验结果有效验证了 iDVAE 的可解释性以及该可解释性对人机协同决策的指导作用, 人类对决策依据权重的调节能够在推理阶段直接影响智能体自主决策的偏好, 或进行控制干预, 实现有人/无人协同控制.

## 6 结 论

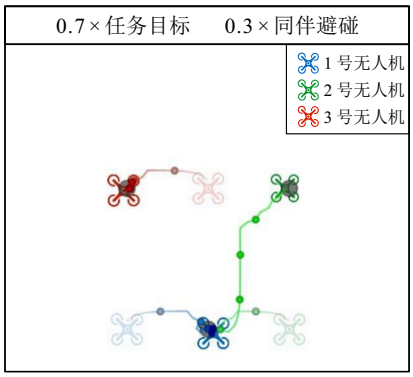
本文提出了一种新型可解释多智能体强化学习方法, 旨在通过变分自编码器网络和门控网络来实现匹配物理语义信息的决策依据概率分布编码和线性组合, 进而实现了多智能体自主决策的主动诠释. 具体而言, 首先, 将智能体的决策建模为若干狄利克雷分布的线性混合, 每种狄利克雷分布表示一种决策依据对于动作的影响; 然后, 决策依据从状态隐空




(a) 任务目标依据与同伴避碰各50%权重下无人机集群运动过程




(b) 同伴避碰依据70%权重、任务目标30%权重下无人机运动过程




(c) 任务目标依据70%权重、同伴避碰30%权重下无人机集群运动过程

**图9　全局调节决策偏好后的多智能体运动过程**

间中编码产生并由门控网络计算权重混合; 最后, 在多智能体近端策略优化网络框架下完成训练. 决策依据的概率分布描述和线性组合赋予了决策网络可解释性, 为人类重构奖励以提升网络性能和干预机器决策提供了支持. 仿真实验验证了所提出方法的有效性, 并直观地展示了网络的可解释性和可控性.

所提出可解释强化学习方法能够以插件的形式直接替换传统 MARL 中不可解释的动作网络, 在不影响决策性能的前提下实现了决策意图诠释, 对于实现多智能体可信自主决策具有重要意义. 未来希望将其进一步推广至语义复杂场景中并研究其对复杂语义信息和决策行为的编码能力以及表征能力,

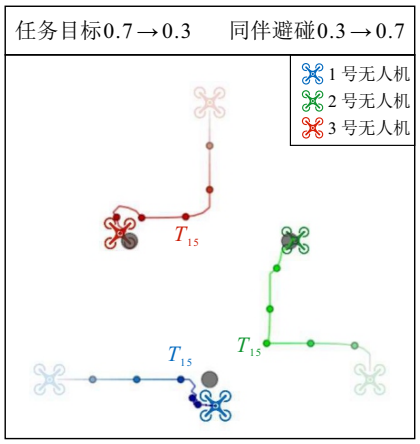



**图10　局部干预决策行为后的多智能体运动过程**

从而进一步增强人类对无人系统自主决策的理解和信任.

## 参考文献 (References)

[1] 伍国华, 李冰洁, 袁于斐, 等. 基于任务分解与强化学习的多平台协同火力分配方法[J]. 控制与决策, 2024, 39(5): 1727-1735.
(Wu G H, Li B J, Yuan Y F, et al. Multi-platform collaborative firepower allocation method based on task decomposition and reinforcement learning[J]. Control and Decision, 2024, 39(5): 1727-1735.)

[2] 夏家伟, 朱旭芳, 张建强, 等. 基于多智能体强化学习的无人艇协同围捕方法[J]. 控制与决策, 2023, 38(5): 1438-1447.
(Xia J W, Zhu X F, Zhang J Q, et al. Research on cooperative hunting method of unmanned surface vehicle based on multi-agent reinforcement learning[J]. Control and Decision, 2023, 38(5): 1438-1447.)

[3] 隋丽蓉, 高曙, 何伟. 基于多智能体深度强化学习的船舶协同避碰策略[J]. 控制与决策, 2023, 38(5): 1395-1402.
(Sui L R, Gao S, He W. Ship cooperative collision avoidance strategy based on multi-agent deep reinforcement learning[J]. Control and Decision, 2023, 38(5): 1395-1402.)

[4] Chen J, Sun J, Wang G. From unmanned systems to autonomous intelligent systems[J]. Engineering, 2022, 12: 16-19.

[5] Zhou Z Y, Liu G J, Tang Y. Multi-agent reinforcement learning: Methods, applications, visionary prospects, and challenges[J/OL]. 2023, arXiv: 2305.10091.

[6] 陈杰, 辛斌. 有人/无人系统自主协同的关键科学问题[J]. 中国科学: 信息科学, 2018, 48(9): 1270-1274.
(Chen J, Xin B. Key scientific problems in the autonomous cooperation of manned-unmanned systems[J]. SCIENTIA SINICA Informations, 2018, 48(9): 1270-1274.)

[7] 张杰勇, 钟赟, 孙鹏, 等. 有人/无人机协同作战指挥控制系统技术[J]. 指挥与控制学报, 2021, 7(2): 203-214.
(Zhang J Y, Zhong Y, Sun P, et al. Command and control system and technology for manned/unmanned

aerial vehicle cooperative operation[J]. Journal of Command and Control, 2021, 7(2): 203-214.)

[8] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016: 1135-1144.

[9] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision. Venice, 2017: 618-626.

[10] Lundberg S, Lee S I. A unified approach to interpreting model predictions[J/OL]. 2017, arXiv: 1705.07874.

[11] 杨书恒, 张栋, 熊威, 等. 基于可解释性强化学习的空战机动决策方法[J]. 航空学报, 2024, 45(18): 329922.
(Yang S H, Zhang D, Xiong W, et al. Decision-making method for air combat maneuver based on explainable reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(18): 329922.)

[12] 刘潇, 刘书洋, 庄韫恺, 等. 强化学习可解释性基础问题探索和方法综述[J]. 软件学报, 2023, 34(5): 2300-2316.
(Liu X, Liu S Y, Zhuang Y K, et al. Explainable reinforcement learning: Basic problems exploration and method survey[J]. Journal of Software, 2023, 34(5): 2300-2316.)

[13] Sequeira P, Gervasio M. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations[J]. Artificial Intelligence, 2020, 288: 103367.

[14] Jenner E, Gleave A. Preprocessing reward functions for interpretability[J/OL]. 2022, arXiv: 2203.13553.

[15] Bewley T, Lawry J. TripleTree: A versatile interpretable representation of black box agents and their environments[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(13): 11415-11422.

[16] Bastani O, Inala J P, Solar-Lezama A. Interpretable, verifiable, and robust reinforcement learning via program synthesis[C]. International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. Vienna, 2020: 207-228.

[17] Kaiser M, Otte C, Runkler T, et al. Interpretable dynamics models for data-efficient reinforcement learning[J/OL]. 2019, arXiv: 1907.04902.

[18] Milani S, Topin N, Veloso M, et al. Explainable reinforcement learning: A survey and comparative review[J]. ACM Computing Surveys, 2024, 56(7): 1-36.

[19] Samadi A, Koufos K, Debattista K, et al. SAFE-RL: Saliency-aware counterfactual explainer for deep reinforcement learning policies[J]. IEEE Robotics and Automation Letters, 2024, 9(11): 9994-10001.

[20] Nikulin D, Ianina A, Aliev V, et al. Free-lunch saliency via attention in atari agents[C]. IEEE/CVF International Conference on Computer Vision Workshop. Seoul, 2019: 4240-4249.

[21] Zabounidis R, Campbell J, Stepputtis S, et al. Concept learning for interpretable multi-agent reinforcement learning[C]. Conference on Robot Learning. Atlanta, 2023: 1828-1837.

[22] Annasamy R M, Sycara K. Towards better interpretability in deep $Q$-networks[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, 2019: 4561-4569.

[23] Shi W J, Huang G, Song S J, et al. Self-supervised discovering of interpretable features for reinforcement learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2712-2724.

[24] Fang H L, Xiao J W, Wang Y W. Self-training convolutional autoencoder for consumer characteristics identification with imbalance datasets[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106605.

[25] Neumeier M, Botsch M, Tollkühn A, et al. Variational autoencoder-based vehicle trajectory prediction with an interpretable latent space[C]. IEEE International Intelligent Transportation Systems Conference. Indianapolis, 2021: 2712-2720.

[26] Srivastava A, Sutton C. Autoencoding variational inference for topic models[J/OL]. 2017, arXiv: 1703.01488.

[27] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, 2017: 6379-6390.

[28] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[J]. Advances in Neural Information Processing Systems. New Orleans, 2022: 24611-24624.

## 作者简介

李佩璋 (1996−), 男, 博士生, 主要研究方向为可解释强化学习、多智能体系统智能决策, E-mail: lpz0805@qq.com;

费庆 (1970−), 男, 高级工程师, 博士, 主要研究方向为无人运动平台智能控制、工程系统的综合控制与优化, E-mail: feiqing@bit.edu.cn;

陈振 (1976−), 男, 教授, 博士, 主要研究方向为机器人驱动与控制、航天器姿态机动控制, E-mail: chenzhen76@bit.edu.cn;

张言军 (1987−), 男, 教授, 博士, 主要研究方向为多智能体系统分布式协同控制理论、不确定系统自适应与智能控制理论, E-mail: yanjun@bit.edu.cn;

王博 (1992−), 男, 高级工程师, 博士, 主要研究方向为海上人工智能与智能决策, E-mail: wangbobit@163.com.