控制与决策
Control and Decision

面向智能空中博弈的大语言模型-强化学习分层决策算法

骞晨旭, 张雪波, 李论, 赵铭慧, 黄魁华

引用本文:
骞晨旭, 张雪波, 李论, 等. 面向智能空中博弈的大语言模型-强化学习分层决策算法[J]. 控制与决策, 2026, 41(3): 855-864.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.1215

面向智能空中博弈的大语言模型-强化学习分层决策算法

骞晨旭[1,2]，张雪波[1,2†]，李 论[3]，赵铭慧[1,2]，黄魁华[4]

(1. 南开大学 机器人与信息自动化研究所，天津 300350；2. 南开大学 人工智能学院，天津 300350；
3. 广州大学 网络空间安全学院，广州 510006；4. 国防科技大学 系统工程学院，长沙 410072)

摘 要: 在多机智能空中博弈等复杂且高对抗性的场景下, 同时具备精准微操决策能力与高效战术推理能力, 是实现多机紧密协同并夺取制胜优势的关键. 针对现有强化学习方法在多机智能空中博弈过程中面临的策略泛化性差且缺乏高层推理能力的挑战, 提出一种融合大语言模型与深度强化学习的分层决策算法 (LRHDF). 首先, 借鉴人类飞行员的决策机制, 构建“大语言模型-强化学习”(大脑-躯干) 分层决策架构, 有效提高算法的底层微操决策性能与上层认知推理能力; 其次, 基于大语言模型反思的提示迭代机制, 利用环境反馈作为优化信号, 驱动提示指令的持续自主进化; 最后, 受人类团队协同决策机理启发, 设计序贯协同决策机制, 显式建模多智能体协作模式, 提高多智体间协同效率. 在高保真空中博弈平台下的仿真结果与消融结果表明, 相较于传统强化学习类算法, 所提出算法在多类博弈场景下表现出更强的博弈性能与泛化能力, 为多机空中博弈问题的求解提供了一条可行的技术路径.

关键词: 智能空中博弈；超视距；多智能体；大语言模型；强化学习；分层决策框架

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1215

引用格式: 骞晨旭, 张雪波, 李论, 等. 面向智能空中博弈的大语言模型-强化学习分层决策算法 [J]. 控制与决策, 2026, 41(3): 855-864.

LLM-RL hierarchical decision-making algorithm for intelligent aerial combat

QIAN Chen-xu[1,2]，*ZHANG Xue-bo*[1,2†]，*LI Lun*[3]，*ZHAO Ming-hui*[1,2]，*HUANG Kui-hua*[4]

(1. Institute of Robotics and Automatic Information Systems， Nankai University， Tianjin 300350， China；
2. College of Artificial Intelligence， Nankai University， Tianjin 300350， China； 3. School of Cyber Security，
Guangzhou University， Guangzhou 510006， China； 4. School of Systems Engineering， National University of
Defense Technology，Changsha 410072，China)

Abstract: In complex and highly adversarial scenarios such as multi-UAV intelligent aerial combat, simultaneous mastering precise micro-operation decision-making and efficient tactical reasoning is essential for achieving close coordination and gaining dominant advantages. To address the limitations of poor policy generalization and inadequate high-level reasoning capabilities in existing reinforcement learning (RL) methods for such scenarios, this paper introduces a hierarchical decision-making framework integrating large language models (LLMs) with RL (LRHDF). First, inspired by human pilots' decision-making processes, a “LLM-RL” hierarchical framework is constructed, which effectively enhances both low-level micro-operation performance and high-level cognitive reasoning ability. Then, a reflection-based prompt iteration mechanism is implemented, which uses environmental feedback as an optimization signal to optimize prompt instructions continuously. Finally, drawing from human team collaboration, a sequential cooperative decision-making module is developed to explicitly model multi-agent collaboration patterns, thereby improving coordination efficiency. Simulation results and ablation studies on a high-fidelity aerial combat platform demonstrate that the proposed algorithm outperforms traditional RL methods in adversarial performance and generalization across diverse combat scenarios, providing a viable solution for multi-UAV aerial combat challenges.

Keywords: intelligent aerial combat； beyond-visual-range； multi-agent； large language model； reinforcement learning；hierarchical decision-making framework

收稿日期: 2025-11-24；**录用日期:** 2025-12-18.

基金项目: 国家自然科学基金项目 (62293510/62293513)；天津市自然科学基金项目 (22JCZDJC00810).

†通信作者. E-mail: zhangxuebo@nankai.edu.cn.

0 引 言

随着无人系统智能的持续发展, 固定翼无人机凭借其卓越的机动性能、快速响应能力以及高负载动作执行特性, 展现出重要的应用价值[1-2]. 自主空中博弈通常以固定翼无人机为博弈主体, 其在复杂对抗环境下争取制空权方面所表现出的巨大潜力, 已引发学术界的广泛关注[3-5]. 然而, 智能空中博弈策略的设计面临诸多难题, 一方面, 其决策空间高维且复杂, 传统基于规则或博弈论的方法难以同时兼顾无人机机动控制与战术决策; 另一方面, 空中博弈问题回报稀疏且交互序列冗长, 信用分配困难[6]. 尤其在多机博弈场景下, 智能体不仅需要具备人类飞行员般的精准微操决策能力, 还必须拥有支持多机协同的高级战术推理能力, 以应对不同场景和任务需求的变化[7], 这为其设计带来了更加严峻的挑战.

近年, 深度强化学习 (deep reinforcement learning, DRL) 在智能空中博弈领域取得了显著成果, 已在高敏捷固定翼无人机机动控制[8-9]、多无人机视距内缠斗[10-11]、多无人机超视距博弈[12-14] 等多种任务与场景中展现出良好性能. 然而, 基于强化学习的方法通常更擅长在特定任务下进行微观操作决策, 而在需要较强推理与规划能力的新场景中则面临较大挑战. 具体表现为: 首先, 受限于泛化能力[15-16], 该类方法严重依赖训练数据分布. 在面对训练分布外 (out of distribution, OOD) 的新颖或突发状态时, 其性能可能急剧下降, 甚至产生不可控的决策行为; 其次, 其适应速度较慢, 面对新态势或对手策略变化时, 常需在大量数据上进行二次训练, 难以实现快速在线适应; 此外, 基于强化学习的博弈策略通常缺乏高层、可解释的推理能力, 难以显示形成如“诱敌深入”或“协同包抄”等抽象战术概念. 其决策过程类似黑盒, 限制了在真实场景中的应用.

与此同时, 大语言模型 (large language model, LLM) 在自然语言理解、常识推理和情境适应等方面展现出的巨大潜力, 推动了基于大语言模型的智能体研究的发展. 一方面, 部分研究尝试将 LLM 作为规划器 (planner), 利用其推理能力将复杂任务分解为基本指令或技能, 如 Dasgupta 等[17]、Yuan 等[18]、Wang 等[19] 的工作; 为进一步提高推理与反思能力, Zhu 等[20] 为 LLM 智能体设计了记忆模块, 并与已有文本知识融合, 在“我的世界 (minecraft)”困难任务中相较传统强化学习方法取得了更高成功率与泛化性能. 然而, 上述研究多将 LLM 定位于规划器, 且主要聚焦于单智能体任务规划, 其在多智体协同中的潜力尚未充分挖掘. 另一方面, 为促进 LLM 在现实物理世界中的应用, 部分研究探索了 LLM 与 RL 的融合方式. 例如, Ahn 等[21] 通过将 LLM 与强化学习生成的值函数进行决策融合, 增强了智能体对环境的感知能力, 实现了 LLM 与现实世界的有效衔接, 在多项真实机器人任务中表现出更优的控制性能. 另有研究将 RL 作为优化器, 基于环境反馈对 LLM 主体[22]、文本编码[23]或提示 (prompt)[24] 等进行微调优化, 在机器人导航、机械臂控制与抓取等多种任务上展现出更强的规划与迁移能力. 然而, 上述研究多选用机器人控制等动态变化相对缓慢、环境可控程度较高的任务场景, 针对多机博弈这类高对抗、高动态且多智能体协同决策关系复杂场景下的应用, 相关探索仍较为缺乏.

为此, 为系统融合大语言模型的推理泛化能力与强化学习的精确控制优势, 本文面向多机空中博弈, 提出一种面向多机协同的大语言模型-强化学习分层决策框架 (LLM-RL hierarchical decision-making framework, LRHDF), 主要贡献如下:

1) 提出一种大语言模型-强化学习分层决策框架, 通过大语言模型实现高层策略规划, 强化学习智能体负责底层动作执行, 实现策略生成与动作执行的有效解耦, 有效提升算法的底层决策性能与上层策略推理能力.

2) 设计一种动态提示反思与循环优化机制, 通过大语言模型对决策结果进行自动评估与迭代修正, 实现提示文本的自主进化, 降低对领域专家知识的依赖的同时, 赋予系统自主进化的能力.

3) 引入基于思维链 (chain-of-thought, COT) 的多智能体协同决策机制, 通过序列化推理过程将隐式协作转化为显式可追溯的决策链, 提升多智能体协同效率与决策可解释性.

4) 通过在多种典型空中博弈环境下的仿真与消融分析, 表明所提出方法在新颖场景下的有效性.

1 问题建模

1.1 多无人机空中博弈

多无人机超视距空中博弈是典型的信息化、智能化对抗场景, 其核心问题在于实现多机协同决策与动态任务规划. 为便于表述, 定义博弈双方为红方与蓝方, 每架固定翼无人机携带m枚超视距导弹, 红方无人机编号为$R_i(i \in \{1,2,\dots,N_r\})$, 蓝方无人机编号为$B_j(j \in \{1,2,\dots,N_b\})$, 已发射导弹的编号为$M_k(k \in \{1,2,\dots,N_m\})$, 其中$N_r$、$N_b$、$N_m$分别为红蓝方无人机数量与当前导弹的数量.

为提高系统泛化能力, 采用相对位姿描述无人机之间的空间关系, 以替代全局坐标系下的表示方式. 如图 1 所示, h^{ij}、$D^{ij} \in \mathbb{R}$分别表征R_i与B_j间的高度差距与距离, D_M^{ik}、D_M^{jk}分别为R_i、B_j与M_k的距离. 定义两机质心连线为视距线 (line-of-sight, LOS), φ_{ij}、$q_{i,j}$分别表征R_i的提前角与B_j的进入角.

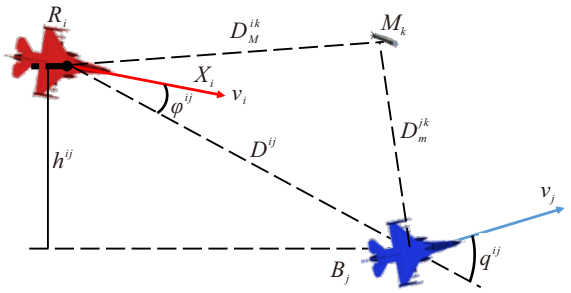


图1　空中博弈态势建模

在该问题设定下, 固定翼无人机与导弹均遵循六自由度动力学模型. 智能体可控制的动作包括无人机油门$c_T \in [0,1]$、副翼偏转$c_a \in [-1,1]$、升降舵偏转$c_e \in [-1,1]$、方向舵偏转$c_r \in [-1,1]$以及导弹发射指令$a^m \in \{0,1\}$. 红蓝双方的博弈目标均为在保证自身生存的前提下, 尽可能击落更多敌方无人机. 胜负判定以双方剩余无人机数量多者为胜, 后续研究将基于此多机智能空中博弈建模方式展开.

1.2　马尔可夫决策建模

在多机智能空中博弈场景中, 雷达有限的探测范围常导致“战争迷雾”, 使得探测信息不完整. 鉴于此, 将多无人机智能空战问题建模为部分可观测马尔可夫博弈 (partially observable Markov game, POMG). 通常, 一个 POMG 可表示为元组$\langle N, \mathcal{S}, \mathcal{O}, \mathcal{A}, R, T, \gamma \rangle$. 其中: N为智能体数量, $\mathcal{S}$为环境全局状态空间. 每个智能体i首先从环境中获得观测值$o_t^i \in \mathcal{O}$, 随后根据自身策略π^i, 基于该观测o_t^i生成动作a_t^i, 环境执行联合策略$\boldsymbol{\pi}(s_t) = \prod \pi^i(o_t^i)$产生的联合动作$\boldsymbol{a_t} = [a_t^1, \ldots, a_t^N]$, 并根据状态转移函数$T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]^N$生成下一状态$s_{t+1}$, 同时根据奖励函数$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$为各智能体生成奖励$\boldsymbol{r_t} = [r_t^1, \ldots, r_t^N]$. 每个智能体的目标是最大化其折扣累积回报$G_t^i = \sum_{k=0}^{H} \gamma^k r_{t+k+1}^i$. 其中: t为当前时间步; H为步数上限; $\gamma \in [0,1]$为折扣因子, 用于衡量对未来回报的重视程度.

2　方法框架概述

图 2 展示了面向多机空中博弈的大语言模型-深度强化学习分层决策方法的基本框架, 其核心包含 3 项关键设计.

首先, 构建“大语言模型-强化学习”(大脑-躯干)分层决策架构, 利用大语言模型的推理与规划能力作为“大脑”进行高层策略生成, 利用强化学习在特定任务上的性能优势设计智能体作为“躯干”负责底层战术执行. 该设计有效融合了大语言模型的推理泛化能力与强化学习的实时精准控制能力, 使系统能够理解复杂任务指令, 并对突发与新颖态势作出合理快速的战略响应.

其次, 针对大语言模型应用中依赖人工提示设计、成本高且泛化性受限的问题, 设计一种提示指令迭代优化机制. 以博弈过程反馈为优化信号, 通过大语言模型对自身决策结果进行反思, 迭代修正并优化任务提示 (prompt), 降低对领域先验知识依赖的

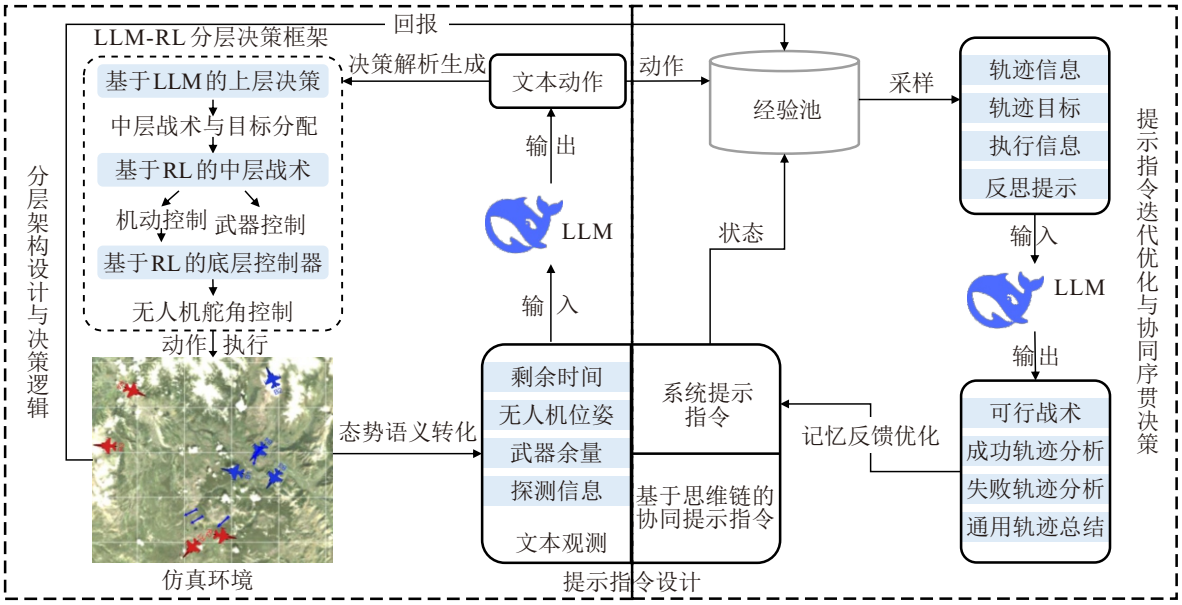


图2　面向多机空中博弈的大语言模型-深度强化学习分层决策方法

同时, 赋予智能体不断演进的能力.

最后, 设计一种基于思维链的多智能体协同序贯决策机制. 该机制通过引入序贯推理过程, 要求每个智能体基于前置智能体的已执行动作进行显式推理, 继而生成自身动作, 从而将传统方法中隐式的协作关系转化为可追溯的显式推理链. 这一设计在提升多机策略协同效率的同时, 也增强了多智能体决策过程的透明性与可解释性.

3 面向多机空中博弈的大语言模型-深度强化学习分层决策算法

3.1 分层架构设计

在多机智能博弈场景中, 由于交互序列长、奖励稀疏等特点, 智能体需具备对长远目标的规划能力. 尽管最终执行动作体现为无人机的油门、舵面偏转及武器控制指令, 但这些动作必须经过策略性选择, 以服务于在更长时间尺度上定义的全局目标, 例如实现无人机的平稳精准控制、执行典型攻防战术、乃至完成多机之间的高效协同. 为此, 借鉴人类飞行员在空战中的分层决策机制, 将复杂的多机博弈问题分解为“底层机动控制-中层战术生成-上层战术调配”的 3 层架构. 在该分层架构中, 底层与中层均采用强化学习方法进行独立训练, 以分别获得鲁棒的机动控制能力与灵活的战术执行能力.

1) 底层控制器负责实现高敏捷、高精度的无人机机动控制, 其核心功能是将上层下达的语义化指令 (如目标航点) 实时转化为平滑可行的底层控制信号. 本文采用 Li 等[9] 提出的高敏捷固定翼控制器, 该控制器基于强化学习与课程学习框架训练, 能够根据当前飞行状态实时解算油门与舵面控制量, 为中层战术决策提供可靠的动作执行基础, 具体实现细节见上述文献.

2) 中层战术生成模块则基于底层控制器, 负责针对不同博弈态势生成相应的战术策略, 如探测、攻击、防御、导弹规避等基本博弈战术. 该模块通过构建“规则-模仿-强化”的训练范式, 将不同战术需求与意图转化为具体的战术动作序列, 并交由底层控制器执行. 本文采用 Qian 等[25] 提出的中层战术训练框架, 设计 4 种典型博弈策略: 探测、攻击、防御、导弹躲避, 涵盖超视距空中博弈的基本博弈战术, 具体实现细节见上述文献.

3) 上层战术调配智能体作为系统的指挥中心, 采用大语言模型实现全局战场态势理解与协同攻防决策. 其通过语义化战场感知、结构化提示引导和指令解析映射, 动态分配目标并分派中层博弈战术, 以实现多无人机协同决策任务, 共包含态势语义化、提示指令设计、决策解析与记忆反馈 4 部分, 具体阐述如下:

① 态势语义转化: 将数值表征的状态信息 (如位姿、弹药存量等) 转化为结构化自然语言描述, 构建包含空间关系、战术上下文与任务目标的态势文本, 为模型提供可理解的博弈认知基础.

② 提示指令设计: 构建具有角色定义、先验知识与输出约束的系统提示模板, 结合动态优化的任务提示与实时态势文本, 组成完整的模型输入, 引导大语言模型进行战术级推理.

③ 决策解析生成: 通过定义结构化输出模板与解析规则, 将模型生成的文本推理链自动映射为可执行的中层策略编号及目标索引, 实现从语义决策到具体指令的转化.

④ 记忆反馈优化: 构建决策经验库对历史交互序列进行存储与表征分析, 将历史决策及其产生的战场效能作为提示词优化的经验证据, 为提示词的迭代优化提供理论依据与数据支撑, 形成具有自我改进能力的决策闭环.

3.2 提示指令迭代优化机制

在基于大语言模型的决策生成时, 提示指令的质量对决策逻辑的合理性与系统性能起着关键作用. 本节详细阐述所提出的动态提示优化机制, 该机制旨在解决 LLM 应用中过度依赖人工提示设计、成本高昂的关键问题, 同时赋予算法自主演进的能力. 如算法 1 所示, 通过构建“决策-存储-反思-进化”的闭环, 实现对提示指令的自主与持续优化.

算法 1 提示指令迭代优化算法.

输入: 预设提示P^0, 更新阈值N_s, 更新轮数N_u;

输出: 迭代优化后的提示P^*.

1: for $k=0$ to N_u do

2: for $i=0$ to N_s do

3: 基于当前提示P^k采样轨迹τ;

4: 将元组$(\tau, P^k, R(\tau))$存入经验池D中;

5: end for

6: for $j=0$ to $N_s//b$ do

7: 从D中随机抽取b个样本;

8: 拼接反思提示P^r, 样本与当前提示;

9: 输入LLM进行反思, 生成提示$\hat{P}^j$;

10: end for

11: 赋值$P^{k+1}=\hat{P}^j$;

12: end for

13: $P^*=\arg\max_P\{R(P^1), R(P^2), \ldots, R(P^{N_u})\}$.

定义当前任务提示指令为P^k, 优化目标是寻找一个最优提示P^*, 以最大化在目标任务上的期望回报, 即$P^* = \arg\max_P \mathbb{E}_{\tau\sim\pi(P)}[R(\tau)]$. 其中: $\pi(P)$为由提示P参数化的 LLM 决策策略; τ为决策轨迹; R为回报函数, 用于评估轨迹τ的优劣.

1) 决策与存储: 在每个数据采集周期中, LLM 基于当前提示P^k与所有无人机观测o_t生成决策a_t, 并与环境交互获得瞬时回报r_t与下一时刻观测o_{t+1}, 重复该过程至回合结束, 得到完整决策轨迹

$$\tau = \{(o_t, a_t, r_t, o_{t+1})\}_{t=0}^{T-1}, \tag{1}$$

其中T为轨迹长度. 将决策轨迹、提示指令与轨迹评估值组成经验元组$(\tau, P^k, R(\tau))$存入记忆池D中.

2) 反思与迭代: 当记忆池D中积累经验达到更新阈值后, 启动反思过程. 首先, 使用洗牌算法 (shuffle algorithm) 将记忆池D中所有样本随机排列; 然后, 从D中依次不重复地采样b个样本, 构建反思提示P^r, 要求 LLM 基于当前决策轨迹与性能评估结果, 分析提示P^k的缺陷, 生成提示$\hat{P}^k$; 接着, 重复上述反思过程, 将记忆池中所有样本迭代完成后, 生成改进版本P^{k+1}; 最后, 用P^{k+1}替换P^k, 并重置记忆池, 完成一次优化迭代.

通过“决策-存储-反思-进化”的循环过程, 系统在完成预设迭代轮次后, 基于验证集性能从所有生成的提示序列$\{P^1, P^2, \ldots, P^K\}$中选取在验证环境中获得最高累积回报的提示作为最终输出P^*, 即

$$P^* = \arg\max_P \{R(P^1), R(P^2), \ldots, R(P^k)\}. \tag{2}$$

该动态优化机制实现了提示指令的自主演进与持续改进, 不仅显著降低了对人工提示工程的依赖, 更通过数据驱动的迭代优化过程, 确保了系统能够自主适应复杂多变的博弈环境.

3.3　基于思维链的多智能体协同序贯决策机制

为提升多智能体系统在复杂动态环境中的协同决策性能, 本文设计一种基于提示工程的序贯决策机制. 其核心思路是通过提示指令明确规定智能体的决策顺序与依赖关系, 强制后续智能体在生成动作时参考前置智能体已生成的动作与分析结果, 从而将隐式协作转化为显式推理, 增强团队决策的一致性与互补性.

图 3 展示了序贯决策机制与标准决策指令的差异. 具体而言, 该机制通过结构化提示模板将多智能体决策过程建模为一个序贯生成问题, 其中每个智能体的输出不仅包含动作指令, 还涵盖对战场态势的语义理解与战术意图推断, 形成可追溯的推理链条. 考虑一个由N个智能体组成的系统, 智能体索引为$i \in 1, 2, \ldots, N$. 在每个决策时刻t, 环境状态为s_t, 智能体按预定义顺序依次生成动作. 设a_i表示智能体i的动作, c_i表示其生成的思维链分析 (包括态势理解、意图推断和动作理由). 序贯决策过程可表示为

$$a_i, c_i = \pi_i^{\mathrm{LLM}}(s_t, h_{<i}). \tag{3}$$

其中: $h_{<i} = (a_1, c_1), (a_2, c_2), \ldots, (a_{i-1}, c_{i-1})$为前序智能体的动作与分析序列; π_i^{LLM}为智能体i的策略函数, 由大语言模型实例化.

4　仿真与消融

本节旨在对所提出算法 (LRHDF) 进行全面评估. 通过在高保真空中博弈仿真环境中, 设计与基线方法的系统性对比及关键组件的消融实验, 验证该方法在泛化性、协同效率与应对突发态势方面的综合性能.

4.1　仿真设置

本文基于“智空仿真推演与训练平台”开展算

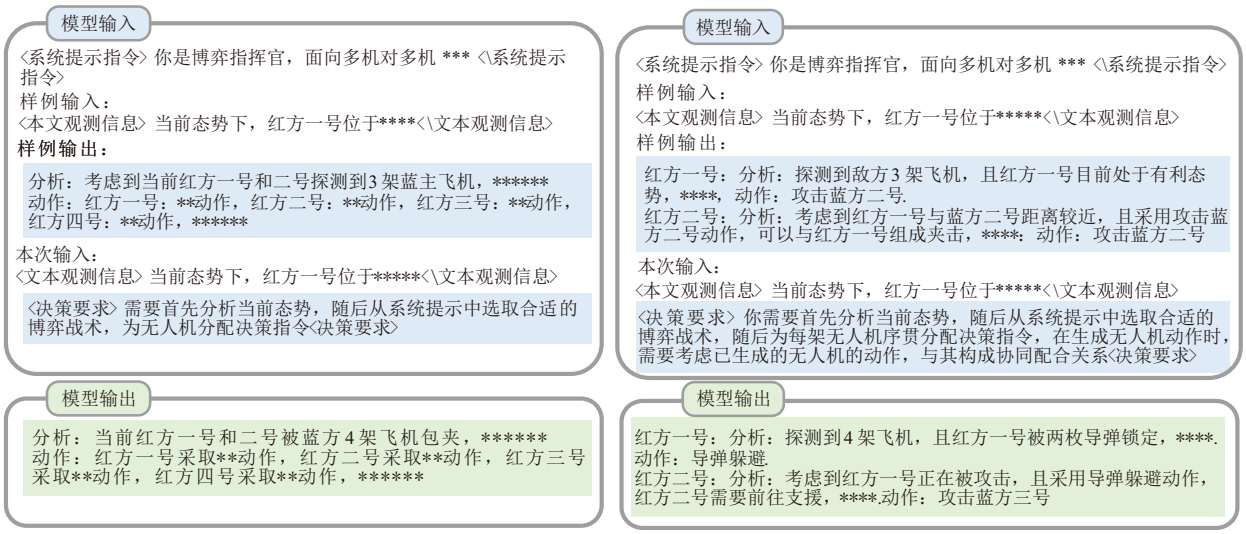
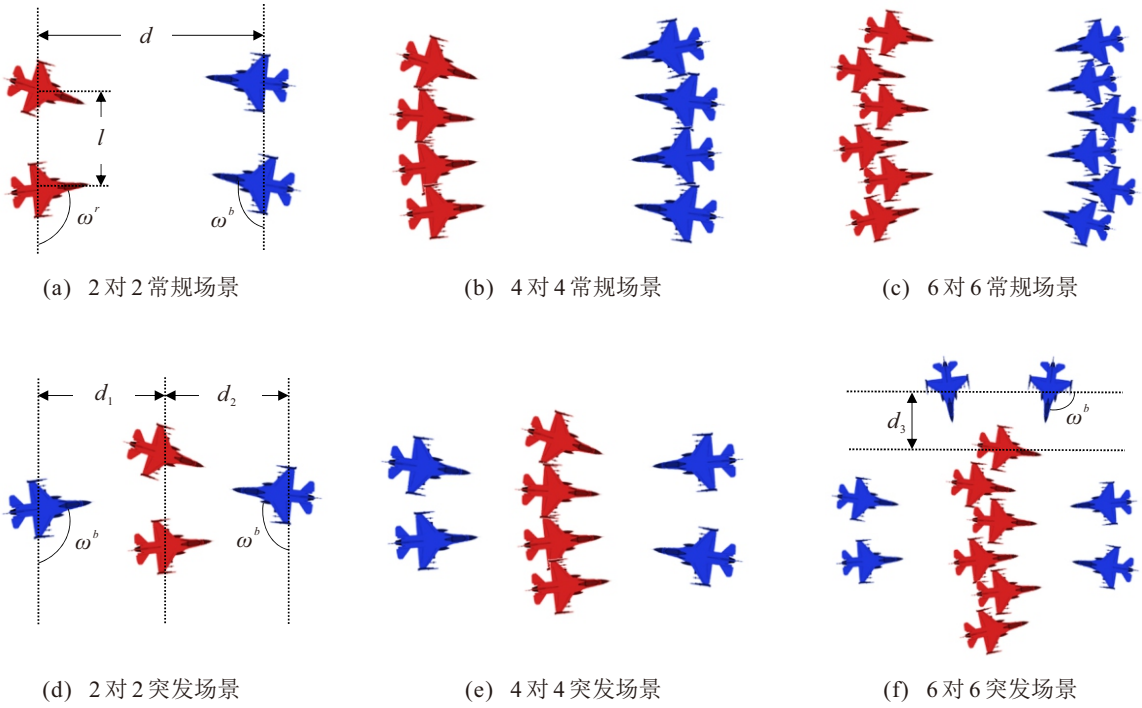


图3　基于思维链的协同提示指令设计

法验证. 该平台集成六自由度无人机与导弹动力学模型, 配备高精度机载雷达与武器系统, 能够为智能空中博弈任务提供高保真的仿真环境.

为全面评估算法性能, 本文针对 2 对 2、4 对 4 和 6 对 6 三种典型博弈规模, 分别设计了“常规场景”与“突发场景”两类初始化条件. 每种规模下均实施一系列红蓝双方的仿真对抗实验, 以系统评估无人机在自主决策与协同作战中的表现.

具体场景设计如图 4 所示, 所有场景均设计为红蓝对抗模式, 其中红方代表待评估的智能体算法, 蓝方则选自平台内置策略库或采用自对抗策略 (根据不同算法需求), 需要说明的是, 为充分检验算法的泛化能力, 所有场景下的初始变量 (无人机位姿, 速度等) 将均在合理范围内随机生成.



(a) 2 对 2 常规场景 (b) 4 对 4 常规场景 (c) 6 对 6 常规场景

(d) 2 对 2 突发场景 (e) 4 对 4 突发场景 (f) 6 对 6 突发场景

图4 仿真场景示意图

常规场景下, 红蓝双方初始呈相向对峙态势, 处于相对均势. 为保证双方均势且具备足够的调整空间, 设导弹的最远攻击距离为d_m, $l \in [0.1d_m, 0.15d_m]$表征友方相邻无人机间的距离, $d \in [2d_m, 2.5d_m]$表征红蓝双方无人机所在经线的距离, 红蓝双方的初始航向角ω^r、ω^b均在 $[0,2\pi]$ 中随机生成, 无人机的高度h在 [20 000 ft, 28 000 ft]中随机初始, 速度v在 [180 m/s, 280 m/s] 中随机初始.

突发场景下, 红方初始即处于蓝方包围圈内且尚未完成目标探测, 处于明显劣势. 如图 4(d) 与图 4(f) 所示, d_2、$d_1 \in [1.2d_m, 1.5d_m]$分别表征红方无人机所在经线与前方、后方蓝队无人机所在经线的距离. 随后, 标注侧方蓝队无人机与距离其最近的红方无人机, 两者所在纬线间的距离被表征为$d_3 \in [1.2d_m, 1.5d_m]$. 最后, 其余变量的初始化策略与常规场景相同.

4.2 对比结果与分析

为深入探究 LRHDF 算法的有效性, 分别选取最优水平 (state of the art, SOTA) 的智能空中博弈算法与多智体强化学习方法算法作为基线算法, 对比算法具体信息阐述如下:

1) LRHDF: 本文所提出算法, 采用大语言模型-强化学习分层决策架构. 本次实验中, 选用 DeepSeek-V3.2-Exp 的非思考模式作为大语言模型底座.

2) PJOH-TED2: PJOH-TED2[25] 是面向超视距空中博弈场景的智能博弈算法, 其基于分层强化学习, 通过构建分层局部联合优化机制与时间事件双驱动算法, 在超视距空中博弈场景下达到 SOTA 水平.

3) MAPPO: 多智能体近端策略优化算法 (multi-agent proximal policy optimization, MAPPO[26]) 是一个广泛应用于多智能体环境中的强化学习方法, 其通过集中式训练、分布式执行的框架, 使各个智能体可以共享信息, 促进多智体间的策略协同. 考虑到空战探索空间巨大的挑战, 本文使用 MAPPO 训练上层战术调配智能体, 中层与底层的设计与 LRHDF 相同.

3 种算法在中层和底层设计上保持一致, 仅在上

表1　与 SOTA 算法对比结果 (胜率/平率/负率)

博弈规模	场景类型	LRHDF	PJOH-TED2	MAPPO
2对2	常规场景	84% / 10% / 6%	**96%** / 2% / 2%	84% / 12% / 4%
	突发场景	**58%** / 24% / 18%	22% / 74% / 4%	34% / 50% / 16%
	综合结果	**71%** / 17% / 12%	59% / 38% / 3%	59% / 31% / 10%
4对4	常规场景	76% / 12% / 12%	**96%** / 0% / 4%	72% / 22% / 6%
	突发场景	**72%** / 20% / 8%	36% / 48% / 16%	16% / 70% / 14%
	综合结果	**74%** / 16% / 10%	66% / 24% / 10%	44% / 46% / 10%
6对6	常规场景	78% / 14% / 8%	**92%** / 2% / 6%	86% / 8% / 6%
	突发场景	**50%** / 36% / 14%	32% / 50% / 18%	22% / 68% / 10%
	综合结果	**64%** / 25% / 11%	62% / 26% / 12%	54% / 38% / 8%

层设计上存在差异. 其中: PJOH-TED2 与 MAPPO 为基于强化学习的决策算法, 二者均针对 2 对 2、4 对 4 及 6 对 6 三种博弈规模独立训练智能体为模拟训练分布内与分布外的仿真条件; 训练阶段仅采用常规场景, 测试阶段则同时涵盖常规场景与突发场景, 且所有场景下的初始变量 (无人机位姿, 速度等) 均在合理范围内随机生成. 上述两个算法均基于分布式训练框架完成 1 000 轮训练, 每轮训练采集 50 幕交互数据并执行 200 次参数更新, 最终选取训练过程中性能最优的模型用于测试. LRHDF 则为基于大语言模型的决策算法, 其提示指令迭代优化模块共设计 100 轮迭代过程, 每轮迭代采集 50 幕样本. 针对 2 对 2、4 对 4 与 6 对 6 三种不同博弈规模, 采用统一的大语言模型基座与提示指令范式, 无需针对不同规模单独训练.

训练完成后, 以各算法生成的最优模型为红方策略, 平台内置策略库内策略为蓝方策略, 每个博弈规模与初始化场景下测试 50 局, 得到的博弈胜率结果展示在表 1 中, 为清晰展示性能对比, 每一组中胜率最高的数值均以加粗形式突出显示. 仿真数据表明以下现象:

1) 在常规场景下, 基于强化学习的 PJOH-TED2 与 MAPPO 算法的胜率略优于 LRHDF 算法;

2) 在突发场景 (即训练分布外场景) 下, PJOH-TED2 与 MAPPO 的胜率出现显著下降, LRHDF 方法则保持了相对稳定的胜率;

3) 在突发场景及最终综合结果中, LRHDF 算法的胜率优于 PJOH-TED2 与 MAPPO 算法.

深入分析上述现象, 可得到以下结论:

1) PJOH-TED2 与 MAPPO 作为针对性训练的强化学习算法, 在与训练数据分布一致的常规场景中, 通过训练能够充分拟合场景特征, 因此表现出更高的胜率, 展现了强化学习在特定训练任务与场景下的优越性能.

2) 在作为训练分布外测试的突发场景中, 两类强化学习算法因缺乏对未知态势的针对性训练, 且其决策机制缺少对战场态势的语义级理解与抽象推理能力, 导致胜率显著下降. 相比之下, LRHDF 依托大语言模型基座所具备的通用知识与逻辑推理能力, 无需额外训练即可迁移适配新场景特征, 展示了更好的泛化性能.

3) 综合常规场景与突发场景下的结果分析, LRHDF 在维持常规场景可接受性能的同时, 在突发场景中仍具备稳定决策能力; 而 PJOH-TED2 与 MAPPO 的性能过度依赖训练分布, 其在分布外场景的性能崩塌严重制约了整体表现.

此外, 从学习机制角度看, 传统强化学习依赖大量交互数据以迭代优化策略参数, 而 LRHDF 通过提示工程的动态优化与思维链的显式推理, 实现了知识的高效迁移与快速适应. 这一差异在突发场景测试中尤为显著: 传统方法需重新采样并训练模型, 而 LRHDF 仅通过语义层面的提示调整即可应对新型战场态势, 为空中博弈等复杂任务的智能体设计提供了具有前景的技术路径.

综上所述, 仿真结果表明, LRHDF 方法在不同场景下的泛化能力方面具有明显优势, 验证了其在高动态不确定环境中的有效性与鲁棒性.

4.3　消融结果与分析

为深入探究 LRHDF 框架中各核心组件的贡献, 本节设计 3 组消融实验并进行具体分析. 具体而言, 通过修改提示指令优化方法与多智体协同决策框架, 生成 3 种变体, 具体阐述如下:

1) LRHDF: 本文所提出算法, 同时采用提示指令迭代优化算法与采用基于思维链的多智能体协同序贯决策机制.

2) w/o reflection: 在 LRHDF 基础上, 移除提示

表2 消融结果 (胜率/平率/负率)

博弈规模	场景类型	完整LRHDF	w/o reflection	w/o COT
2对2	常规场景	**84%** / 10% / 6%	70% / 22% / 8%	6% / 14% / 80%
	突发场景	**58%** / 24% / 18%	28% / 50% / 22%	16% / 54% / 30%
4对4	常规场景	**76%** / 12% / 12%	56% / 32% / 12%	14% / 56% / 30%
	突发场景	**72%** / 20% / 8%	28% / 54% / 18%	10% / 72% / 18%
6对6	常规场景	**78%** / 14% / 8%	42% / 42% / 16%	8% / 78% / 14%
	突发场景	**50%** / 36% / 14%	40% / 46% / 14%	20% / 56% / 24%

指令迭代优化机制, 使用基于先验知识的初始提示 P^0.

3) w/o COT: 在 LRHDF 基础上, 移除基于思维链的多智能体协同序贯决策机制, 即多智能体将并行生成决策并执行.

本节消融实验的设置与对比实验部分保持一致, 采用相同的博弈场景与对手策略库. 为评估算法的泛化性能, 每个场景与博弈规模组合下的初始变量均在合理范围内随机生成, 每组实验进行 50 次独立博弈, 消融结果记录于表 2 中. 为清晰展示性能对比, 每一组中胜率最高的数值均以加粗形式突出显示.

表 2 所示的消融实验结果指出, 完整的 LRHDF 算法在所有博弈规模与场景下均表现出最优的博弈性能. 当移除提示迭代优化机制或基于思维链的多智能体协同序贯决策机制时, 智能体的性能均出现显著衰退. 其中, 协同序贯决策机制的移除导致的影响更为严重: 在 300 局博弈中, 智能体仅取得不足 15% 的平均胜率.

通过可视化分析, 性能差异的主要成因可归纳为以下两方面: 1) 去除提示迭代优化机制后, 智能体依赖于人为设计的专家提示指令. 尽管这些指令包含基本博弈战术知识, 使智能体能够执行基础战术, 但由于缺乏与仿真环境的交互优化, 智能体在关键决策时机 (如躲避导弹和攻防切换) 上出现偏差, 导致性能下降. 2) 去除基于思维链的多智能体协同序贯决策机制后, 各无人机倾向于采取局部最优的个体策略, 难以实现“诱骗弹药”“协同包夹”等高效协同战术. 推理能力与协同能力的缺失严重制约了整体性能. 综上所述, 动态提示优化机制通过持续的策略反思与演进赋予系统自适应能力, 而协同序贯决策机制则通过显式的推理链共享确保多智能体间的高效协作. 消融实验结果证实了所提组件的有效性与必要性.

4.4 可解释性博弈策略分析

值得注意的是, 根据可视化分析, 所提出算法不仅在基本攻防博弈战术中表现出优异的控制性能, 还展现出更具前瞻性的策略规划能力, 演化出多种兼具高控制精度、强逻辑推理能力与良好可解释性的博弈策略. 为直观分析智能体的博弈行为, 本节以 6 对 6 突发场景为例, 对 LRHDF 算法所涌现的可解释性战术进行分析. 图 5 展示了一幕典型博弈对局中的 8 个关键帧及相应战术目标, 具体分析如下.

首先, 在 6 对 6 突发场景下, 初始化时红方六架无人机被蓝方六架无人机包夹, 且蓝方单位处于红方的探测范围之外. 随着博弈的推进, 红方探测到来自 3 个方向的蓝方编队 (图 5 的 A), 在识别自身处于被包夹的不利态势后, 首先采取向两侧出击以扩展战场空间, 增强攻防转换的灵活性, 从而缓解敌方

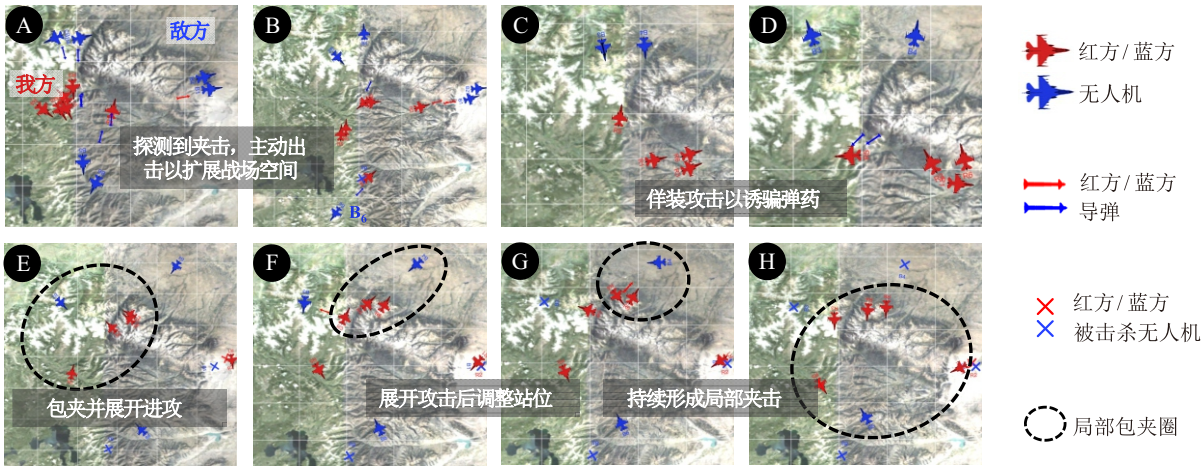


图5 突发场景博弈战术分析

夹击所带来的威胁 (图 5 的 B). 在此过程中, 依托基于 RL 的中层导弹规避战术优势, 红方无人机成功化解第一波导弹攻击. 待红方与蓝方拉开一定安全距离后, 智能体借助优越的防御性能, 持续采用佯装攻击战术诱骗对手弹药 (图 5 的 C 和 D); 当对手弹药与油量出现显著损耗、智能体获得导弹射程优势后, LRHDF 指导多架无人机以编队形式分批次展开进攻, 分别在地图的左上方 (图 5 的 E), 右上方 (图 5 的 F) 及下方 (图 5 的 H) 形成局部包夹. 完成对单个蓝方目标的编队围剿后, 智能体可迅速调整编队构成, 转向攻击蓝方下一目标无人机 (图 5 的 E ∼ H), 通过局部多打一的战术实现对蓝方无人机的逐个击破. 最后, 凭借优异的攻防博弈性能与多智能体协同决策战术, LRHDF 仅以一架红方无人机为代价, 成功击落了 6 架蓝方无人机, 取得博弈胜利.

综上所述, 可视化分析结果表明, 所提出大语言模型-强化学习分层决策算法不仅演化出高性能底层攻防博弈战术, 还涌现出具备长远规划性与协同能力的高层博弈策略, 如佯装攻击、诱骗弹药消耗、编队动态调整、局部包夹与逐个击破等. 即使在多机包夹的突发场景中, 该算法仍能依托其上层高效的逻辑推理能力与底层的精准控制性能, 展现出优秀的博弈水平与泛化能力, 为多机空中博弈等复杂问题的求解及实际应用提供了可行路径.

5　结　论

本文面向多机智能空中博弈场景, 提出了一种融合大语言模型与强化学习的分层决策框架. 设计提示指令迭代优化机制, 通过环境反馈驱动提示文本的自主演进, 以有效提升智能体对动态环境的适应能力; 构建基于思维链的协同序贯决策机制, 显式建模多智能体协作模式, 以提高多智体间协同效率. 高保真空中博弈平台下的仿真结果表明, 所提出框架在多种博弈规模与场景下, 均展现出更高的博弈胜率与优异的泛化性能, 消融结果进一步证实了算法各核心组件的有效性与必要性.

未来研究将着力于优化强化学习与大语言模型的深度融合架构. 首先, 在决策架构方面, 需在保留大语言模型跨场景泛化能力的基础上, 深度融合强化学习在特定任务中的专家级性能, 并系统优化分层架构的计算效能与响应机制, 着力降低大模型决策延迟, 以满足高动态博弈的实时性要求; 其次, 在应用落地方面, 深入探究算法在通信延迟、感知误差等复杂扰动下的稳定性, 构建贴近实际应用的扰动补偿机制, 从而在动态对抗环境中实现兼具广泛适应性与领域最优性的鲁棒智能决策系统.

参考文献 (References)

[1] Jordan J. The future of unmanned combat aerial vehicles: An analysis using the Three Horizons framework[J]. Futures, 2021, 134: 102848.

[2] Ernest N, Carroll D. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions[J]. Journal of Defense Management, 2016, 6(1): 1000144.

[3] 马钧文, 毕文豪, 张安, 等. 基于模糊动态权重的近距空战态势评估方法[J]. 控制与决策, 2024, 39(9): 2995-3005.
(Ma J W, Bi W H, Zhang A, et al. Close-range air combat situation assessment based on fuzzy dynamic weight[J]. Control and Decision, 2024, 39(9): 2995-3005.)

[4] 王国岩, 赵旭华, 解宇轩, 等. 基于态势感知的无人机空战协同决策方法[J]. 控制与决策, 2025, 40(6): 1847-1854.
(Wang G Y, Zhao X H, Xie Y X, et al. A collaborative decision-making method for unmanned aerial vehicles in aerial combat based on situational awareness[J]. Control and Decision, 2025, 40(6): 1847-1854.)

[5] 倪浩, 章胜, 刘福炜, 等. 基于域随机化增强 EfficientZero 的无人机空战智能决策[J]. 控制与决策, 2025, 40(11): 3273-3286.
(Ni H, Zhang S, Liu F W, et al. UAV air combat intelligent decision-making based on domain randomization enhanced EfficientZero[J]. Control and Decision, 2025, 40(11): 3273-3286.)

[6] 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.

[7] 江碧涛, 温广辉, 周佳玲, 等. 智能无人集群系统跨域协同技术研究现状与展望[J]. 中国工程科学, 2024, 26(1): 117-126.
(Jiang B T, Wen G H, Zhou J L, et al. Cross-domain cooperative technology of intelligent unmanned swarm systems: Current status and prospects[J]. Strategic Study of CAE, 2024, 26(1): 117-126.)

[8] Li L, Zhang X B, Qian C X, et al. Basic flight maneuver generation of fixed-wing plane based on proximal policy optimization[J]. Neural Computing and Applications, 2023, 35(14): 10239-10255.

[9] Li L, Zhang X B, Qian C X, et al. Autopilot controller of fixed-wing planes based on curriculum reinforcement learning scheduled by adaptive learning curve[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(3): 2182-2196.

[10] Pope A P, Ide J S, Mićović D, et al. Hierarchical reinforcement learning for air combat at DARPA's AlphaDogfight trials[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(6): 1371-1385.

[11] Chen C, Song T, Mo L, et al. Autonomous dogfight

decision-making for air combat based on reinforcement learning with automatic opponent sampling[J]. Aerospace, 2025, 12(3): 265.

[12] Sun Z X, Piao H Y, Yang Z, et al. Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

[13] Piao H Y, Han Y, Chen H C, et al. Complex relationship graph abstraction for autonomous air combat collaboration: A learning and expert knowledge hybrid approach[J]. Expert Systems with Applications, 2023, 215: 119285.

[14] Qian C X, Zhang X B, Li L, et al. H3E: Learning air combat with a three-level hierarchical framework embedding expert knowledge[J]. Expert Systems with Applications, 2024, 245: 123084.

[15] Wu J L, Wu H X, Qiu Z H, et al. Supported policy optimization for offline reinforcement learning[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022: 31278-31291.

[16] 王雪松, 杨露, 程玉虎. 基于温和泛化的不确定性离线强化学习[J]. 控制与决策, 2025, 40(11): 3329-3339.
(Wang X S, Yang L, Cheng Y H. Uncertainty-aware offline reinforcement learning with mild generalization[J]. Control and Decision, 2025, 40(11): 3329-3339.)

[17] Dasgupta I, Kaeser-Chen C, Marino K, et al. Collaborating with language models for embodied reasoning[J/OL]. 2023, arXiv: 2302.00763.

[18] Yuan H Q, Zhang C, Wang H C, et al. Skill reinforcement learning and planning for open-world long-horizon tasks[J/OL]. 2023, arXiv: 2303.16563.

[19] Wang Z H, Cai S F, Chen G Z, et al. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents[C]. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, 2023: 34153-34189.

[20] Zhu X Z, Chen Y T, Tian H, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory[J/OL]. 2023, arXiv: 2305.17144.

[21] Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: Grounding language in robotic affordances[J/OL]. 2022: arXiv: 2204.01691.

[22] Carta T, Romac C, Wolf T, et al. Grounding large language models in interactive environments with online reinforcement learning[J/OL]. 2023, arXiv: 2302.02662.

[23] Driess D, Xia F, Sajjadi M S M, et al. PaLM-E: An embodied multimodal language model[C]. Proceedings of the 40th International Conference on Machine Learning. Hawaii: PMLR, 2023: 8469-8488.

[24] Yao W R, Heinecke S, Niebles J C, et al. Retroformer: Retrospective large language agents with policy gradient optimization[J/OL]. 2023, arXiv: 2308.02151.

[25] Qian C X, Zhang X B, Li L, et al. A partial joint optimization algorithm for autonomous air combat based on hierarchical reinforcement learning[J]. IEEE Transactions on Cybernetics, 2025, 55(9): 4145-4157.

[26] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]. Neural Information Processing Systems. New Orleans, 2022, 35: 24611-24624.

作者简介

骞晨旭 (1999−), 男, 博士生, 主要研究方向为强化学习与智能博弈, E-mail: qianchenxu@mail.nankai.edu.cn;

张雪波 (1984−), 男, 教授, 博士, 主要研究方向为机器人与人工智能, E-mail: zhangxuebo@nankai.edu.cn;

李论 (1993−), 男, 讲师, 博士, 主要研究方向为强化学习, 多智能体博弈, E-mail: theory@gzhu.edu.cn;

赵铭慧 (1995−), 女, 实验师, 硕士, 主要研究方向为强化学习与智能博弈, E-mail: zhaomh@nankai.edu.cn;

黄魁华 (1986−), 男, 副教授, 博士, 主要研究方向为智能任务规划、智能辅助决策, E-mail: khhuang@nudt.edu.cn.