

北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于强化学习的有限燃料下航天器轨道追逃博弈
作者：赵庭睿，袁洋，霍梦真
DOI：10.13700/j.bh.1001-5965.2026.0339
收稿日期：2026-05-31
网络首发日期：2026-09-05
引用格式：赵庭睿，袁洋，霍梦真．基于强化学习的有限燃料下航天器轨道追逃博弈[J/OL]．北京航空航天大学学报.
https://doi.org/10.13700/j.bh.1001-5965.2026.0339



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

基于强化学习的有限燃料下航天器轨道追逃博弈

赵庭睿，袁洋，霍梦真✉

（北京航空航天大学 自动化科学与电气工程学院，北京 100191）

*通信作者 E-mail： mzhuo@buaa.edu.cn

摘要 针对近圆参考轨道下非合作目标机动引发的空间安全威胁，研究有限燃料约束下的航天器轨道追逃自主决策问题。基于 Clohessy-Wiltshire（CW）相对运动模型，构建了包含短时脉冲机动、长时滑行和燃料约束的多智能体并行追逃仿真环境。在此基础上，提出一种基于多智能体近端策略优化算法（MAPPO）的非对称课程自博弈训练方法。该方法采用集中式训练、分布式执行架构，并在 Actor 与 Critic 网络中引入长短期记忆网络（LSTM），以增强非完备观测下的时序状态表征能力；同时，设计由相位—高度解耦引导、速度漏斗、静默滑行增益和燃料消耗惩罚组成的复合奖励函数，引导追击者在有限燃料条件下形成脉冲—滑行式追踪策略。进一步地，引入基于 CW 哈密顿量变化率的能量观测特征，并设计目标动作辅助预测头，以提升短时目标信息缺失情况下的对手机动估计能力。仿真结果表明，所提方法能够在 CW 相对运动场景和有限燃料约束下实现稳定追逃决策，并改善追击者的燃料利用效率和短时观测缺失时的目标状态预测能力。

关键词 航天器；轨道力学；强化学习；追逃博弈；多智能体系统

中图分类号 V221⁺.3；TB553 文献标志码 A DOI： DOI：10.13700/j.bh.1001-5965.2026.0339

Orbital pursuit-evasion game for spacecraft with limited fuel based on reinforcement learning

ZHAO Tingrui, YUAN Yang, HUO Mengzhen✉

（School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China）

*E-mail: mzhuo@buaa.edu.cn

Abstract This paper investigates an autonomous pursuit-evasion decision-making method for spacecraft under limited Delta-v constraints in a Clohessy-Wiltshire (CW) relative orbital dynamics scenario. A 2-vs-2 parallel pursuit-evasion environment is established to describe impulsive maneuvers, gravitational coasting and fuel limitation near a geostationary reference orbit. On this basis, a self-play decision-making method based on multi-agent proximal policy optimization (MAPPO) is developed. The method adopts a centralized training and decentralized execution architecture, and introduces long short-term memory (LSTM) networks after the MLP encoders of both actor and critic to aggregate historical observations under incomplete information. A composite reward function combining phase-altitude guidance, velocity funnel guidance, coasting reward and fuel penalty is designed to induce fuel-efficient pulse-coast pursuit behavior. In addition, a Hamiltonian variation-rate-based energy observation feature and an auxiliary target-action prediction head are introduced to enhance opponent maneuver estimation during short-term observation loss. Simulation results show that the proposed method improves decision stability and fuel efficiency in the CW relative-motion scenario while providing reliable supplementary target-state prediction under incomplete information.

Key words spacecraft; orbital mechanics; reinforcement learning; pursuit-evasion game; multi-agent systems

收稿日期：2026-05-31

基金项目：国家自然科学基金（U2541218，T2121003）;北京市自然科学基金（L252094）;空间智能控制技术全国重点实验室开放基金课题

Fund：National Natural Science Foundation of China (U2541218，T2121003); Beijing Municipal Natural Science Foundation（L252094）; Open Fund Project of the National Key Laboratory of Space Intelligent Control Technology

网络首发时间： 2026-09-05 10:07:03 网络首发地址： https://link.cnki.net/urlid/11.2625.V.20260904.1707.001

随着近地空间环境日益拥挤以及非合作目标的增加，航天器在执行任务时面临的轨道威胁呈现高动态、对抗性特征。如何赋予航天器在复杂空间环境下自主规避、抵近与对抗决策能力，已成为空间资产安全保障的重要问题[1-2]。航天器间攻防交互通常可抽象为“轨道追逃博弈”，其理论基础可追溯至 Isaacs 提出的微分对策理论[3]。然而，轨道动力学约束、燃料受限以及探测信息滞后等因素，使传统依赖地面测控的开环决策难以满足高动态对抗场景对实时性的要求。

针对轨道追逃博弈，早期研究侧重于利用数值优化或启发式搜索寻找纳什均衡解。有学者通过改进蒙特卡洛树搜索算法解决了回合制博弈中的奖励稀疏问题[4]；针对多对一协同场景，基于计算博弈理论的一步前瞻搜索策略也被提出以寻找均衡解[5]。然而，此类方法在处理连续动作空间或大尺度动态过程时，往往面临计算维度灾难。近年来，深度强化学习凭借“离线训练、在线执行”的毫秒级决策优势被引入轨道控制领域，基于双延迟深度确定性策略梯度算法（twin delayed deep deterministic policy gradient，TD3）以及深度 Q 网络（deep Q-network，DQN）等单智能体算法实现了特定约束下的快速规避规划[6-7]。尽管如此，单智能体方法通常将对手视为环境的一部分，难以捕捉博弈双方策略动态耦合、相互演化的本质特性[8]。

在对抗场景中，有效对手建模是提升策略鲁棒性的关键[9]。为突破单向策略训练的局限，相关研究逐渐引入多智能体强化学习与自博弈机制，使追击者与逃逸者在交替训练中演化出更复杂的对抗策略[10-11]。此外，针对信息非完备及多航天器场景，融合长短期记忆网络（long short-term memory，LSTM）的多智能体强化学习（multi-agent reinforcement learning，MARL）架构能够增强历史信息利用能力[12]；分布式强化学习也在混合动作空间与多目标约束博弈中展现出应用潜力[13]。

值得注意的是，轨道追逃策略的有效性高度依赖动力学模型与燃料约束。在近圆参考轨道邻域内，CW 相对运动模型能够以较低计算代价描述局部相对运动特征，适合用于强化学习策略训练与对比验证。同时，航天器受燃料预算限制，追逃过程通常表现出脉冲滑行特征[14]。经典最优控制理论指出，有限燃料轨迹需要在主动机动与自然轨道演化之间取得平衡[15]。因此，如何在强化学习框架下诱导智能体学习节能机动、相位覆盖和末端稳定交会，是本文关注的核心问题。

针对上述问题，本文提出一种基于多智能体近端策略优化算法（multi-agent proximal policy optimization，MAPPO）的有限燃料轨道追逃自博弈决策方法。本文采用集中式训练、分布式执行架构，以缓解多智能体交互中的环境非平稳性；基于燃料控制设计复合奖励函数，引导追击者利用相位—高度关系和滑行阶段提高燃料利用效率；并引入 CW 哈密顿量变化率观测和辅助预测头，以增强对逃逸者机动状态的估计能力。

1 问题建模

本文对 CW 相对运动模型下的多智能体轨道追逃问题进行建模。基于经典轨道力学理论，建立追击者与逃逸者的相对运动模型，并给出宏观几何约束与末端视线走廊约束，为后续强化学习策略设计提供物理基础。

1.1 坐标系定义

为描述航天器间的相对运动，本研究主要采用以下两个坐标系：

1) 地心惯性坐标系：以地心为原点，x 轴指向春分点，z 轴指向地球北极，y 轴位于赤道面内并与 x、z 轴构成右手系。该坐标系用于描述航天器的绝对运动状态。

2) 局部轨道坐标系：原点设在参考轨道上某一个做匀速圆周运动的参考点的质心上。x 轴沿地心指向航天器的径向，z 轴沿轨道角动量方向，y 轴在轨道面内与 x、z 轴构成右手系。

1.2 航天器相对动力学建模与线性化

假设追击者与逃逸者均主要受地心引力作用。在地心惯性系下，航天器的二体运动方程为

$$\ddot{\boldsymbol{r}}_i = -\frac{\mu}{r_i^3}\boldsymbol{r}_i + \frac{\boldsymbol{f}_i}{m_i} \tag{1}$$

式中：i 为航天器索引（$i\in\{\mathrm{p},\mathrm{e}\}$），p 表示追击者，e 表示逃逸者；$\boldsymbol{r}_i$ 为航天器的地心距离矢量；μ 为地球引力常数；$\boldsymbol{f}_i$ 为航天器受到的控制推力矢量；m_i 为航天器质量。

当追击者与逃逸者之间的相对距离远小于参考轨道半径时，可在近圆参考轨道附近进行线性化，得到描述局部相对运动的 Clohessy-Wiltshire（CW）方程：

$$\begin{cases}\ddot{x}-3n^2x-2n\dot{y}=a_x\\ \ddot{y}+2n\dot{x}=a_y\\ \ddot{z}+n^2z=a_z\end{cases}\tag{2}$$

式中：x，y，z 分别为相对位置矢量在局部轨道坐标系下的径向、切向和面外分量；n 为参考轨道的平均角速度；a_x，a_y，a_z 分别为单位质量受到的控制加速度分量。

1.3 宏观轨道力学几何约束与特性解耦

在远距离交会阶段，追击者需利用轨道相位与高度关系进行节能抵近，而非单纯沿视线方向推进。本文采用如下约束逻辑：

1) 面外偏差优先消除原则。在轨道力学中，改变轨道倾角通常需要较大的法向速度增量。定义追击者与逃逸者单位角动量矢量 $\hat{\boldsymbol{h}}_\mathrm{p}$ 与 $\hat{\boldsymbol{h}}_\mathrm{e}$，夹角的余弦值为 $\cos i_{\mathrm{pe}}=\hat{\boldsymbol{h}}_\mathrm{p}\cdot\hat{\boldsymbol{h}}_\mathrm{e}$。由于末端拦截阶段燃料裕度有限，宏观阶段应优先减小面外偏差，以避免末端因面外误差过大导致捕获失败。

2) 相位与高度的动力学解耦。在中心引力场中，航天器的引力加速度 $\boldsymbol{a}$ 满足：

$$\boldsymbol{a}=-\frac{\mu}{r^3}\boldsymbol{r}\tag{3}$$

由此推导得到开普勒角速度 n 与轨道半长轴 a 的关系：

$$n=\sqrt{\frac{\mu}{a^3}}\tag{4}$$

该关系揭示了低轨快、高轨慢的相位演化特性。若直接以欧氏距离作为唯一引导量，追击者可能因加速升轨导致角速度下降，从而产生不利的相位漂移。因此，本文将距离约束解耦为高度差与相位差，引导追击者通过适当调整轨道高度形成角速度差，实现更有效的相位覆盖。轨道追逃过程如图 1 所示。

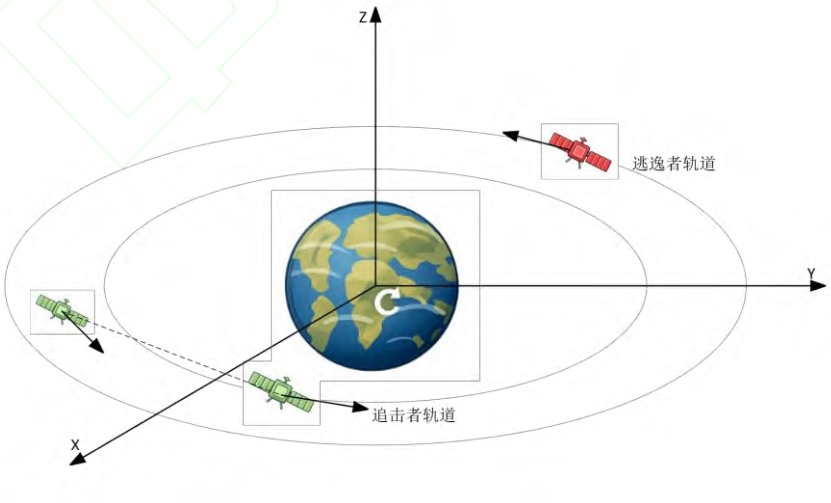

图 1　轨道追逃过程示意图
Fig.1　Schematic diagram of the orbital pursuit-evasion process

1.4 末端视线几何与锥形安全走廊

进入战术交接区后，博弈重点转向相对矢量对齐。定义视线单位矢量与相对接近方向单位矢量，接近角定义为

$$\cos\alpha = \boldsymbol{e}_{\mathrm{r}} \cdot \boldsymbol{e}_{\mathrm{v}} \tag{5}$$

本文设置半锥角为 30° 的安全接近走廊，鼓励末端抵近过程中尽量满足 $\alpha \le 30^{\circ}$，以降低侧向速度过大导致的交叉飞越风险。

1.5 任务判定准则

1) 有效压制：追击者与目标相对距离不超过 40 km 并连续保持 200 个仿真步。压制时间定义为首次满足该连续区间的起始时刻，平均值仅统计压制成功样本。

2) 完美捕获：追击者与目标相对距离不超过 10 km、相对速度不超过 150 m/s，并连续保持 50 个仿真步。捕获时间同样定义为首次满足该连续区间的起始时刻。

2 基于自博弈强化学习的轨道追逃算法设计

针对本文建立的 CW 相对运动追逃模型，传统预规划方法难以在多智能体对抗环境下实现实时闭环决策。为此，本文提出非对称课程自博弈训练架构，并结合 MAPPO 算法完成状态空间、动作空间、奖励函数和辅助预测模块设计。

2.1 非对称课程自博弈训练架构

在高动态对抗环境中，若追击者与逃逸者同时从随机策略开始探索，容易出现策略坍塌或长期低效探索。为缓解该问题，本文设计分阶段非对称课程自博弈训练架构，其流程如图 2 所示。第一阶段为课程预训练，追击者在弱机动或静态目标环境下学习基础抵近规律；第二阶段为自博弈强化训练，追击者加载预训练权重，逃逸者逐步提高机动能力，双方在闭环交互中提升对抗策略强度。阶段切换时仅将预训练追击者 Actor 参数载入第二阶段追击者动作网络，逃逸者动作网络随机初始化，两阶段不共享采样缓存。

2.2 MAPPO 算法机理与运行流程

2.2.1 目标函数与策略裁剪机制

强化学习是一类基于环境交互和奖励反馈的序贯决策学习方法。为限制策略更新幅度并提高训练稳定性，MAPPO 采用 PPO 裁剪目标函数。本文定义的总目标函数由策略裁剪项、价值函数项和熵正则项组成：

$$L_t^{\mathrm{Total}}(\theta,\phi) = \mathrm{E}_t\left[L_t^{\mathrm{CLIP}}(\theta) - c_1 L_t^{\mathrm{VF}}(\phi) + c_2 S[\pi_\theta](s_t)\right] \tag{6}$$

式中：$L_t^{\mathrm{CLIP}}(\theta)$ 为策略裁剪损失函数；$L_t^{\mathrm{VF}}(\phi)$ 为价值函数损失项；$S[\pi_\theta](s_t)$ 为策略熵正则项；c_1 和 c_2 为常数系数。训练目标为最大化该代理目标。

同时由于轨道博弈环境的高度不稳定性，为了防止策略在更新过程中产生过大偏移，导致训练不稳定，$L_t^{\mathrm{CLIP}}(\theta)$ 采用了裁剪机制：

$$\begin{aligned} &L_t^{\mathrm{CLIP}}(\theta) = \\ &\mathrm{E}_t\left[\min\{r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-蝌1+\ )\hat{A}_t\}\right] \end{aligned} \tag{7}$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|\ s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|\ s_t)} \tag{8}$$

式中：$r_t(\theta)$ 为新旧策略的概率比，$\pi_\theta(a_t|\ s_t)$ 为当前网络参数下在状态 s_t 条件下选择动作 a_t 的概率，ò 为裁剪超参数，$\hat{A}_t$ 为优势函数。

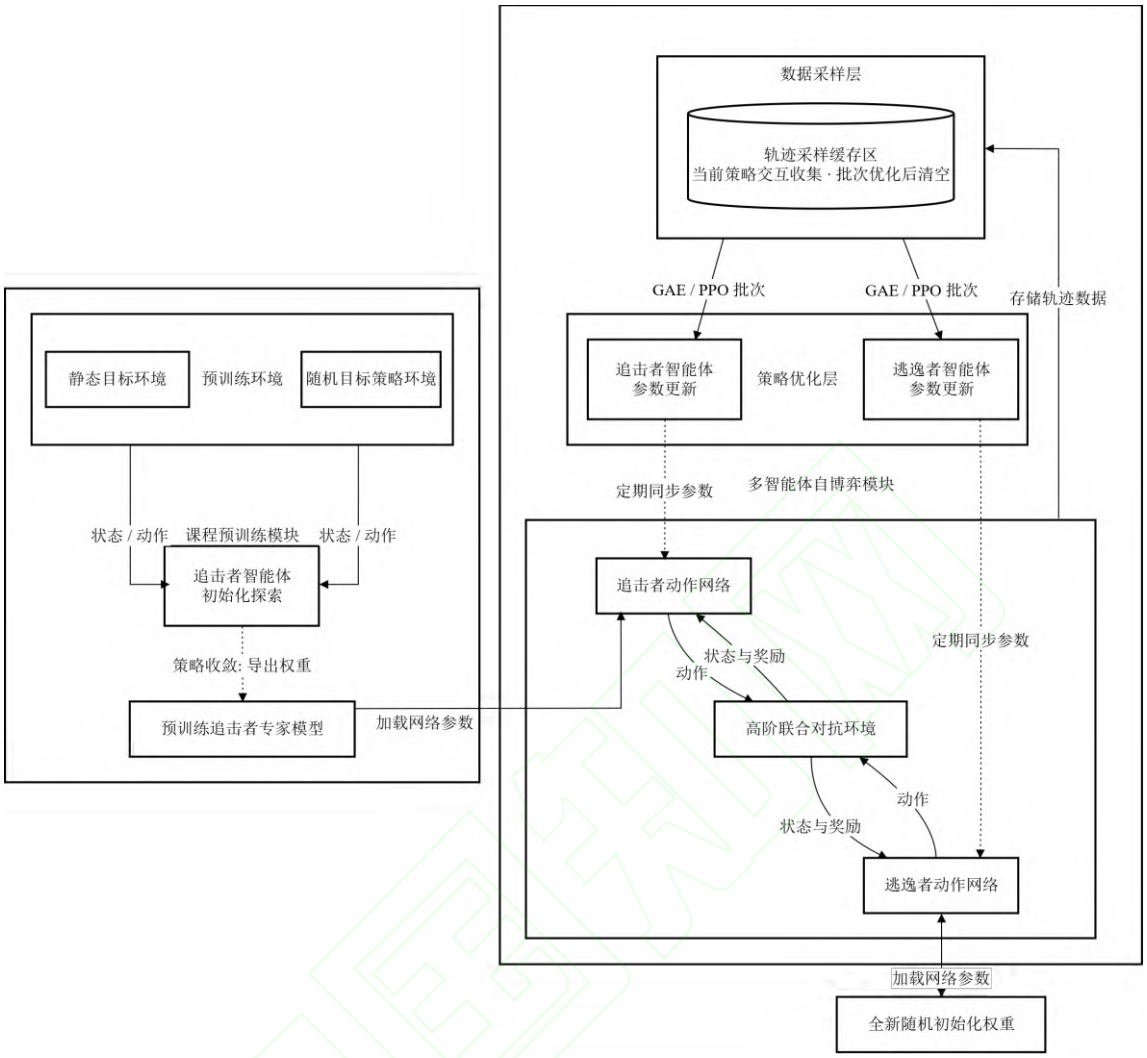

图 2　分阶段训练架构示意图
Fig.2　Schematic diagram of the staged training framework

2.2.2 优势函数与网络更新

优势函数 $\hat{A}_t$ 为当前的动作价值函数与状态价值函数的差值，表示当前策略相较于平均策略的优势：

$$\hat{A}_t = q_{\pi_\theta}(s_t, a_t) - V_\phi(s_t) \tag{9}$$

式中：$V_\phi(s_t)$ 为参数 ϕ 估计的状态价值。由于真实的动作价值函数未知，一般通过计算累积奖励与价值估计的均方误差来更新 Critic 网络，其价值损失函数可表示为：

$$L_t^{\mathrm{VF}}(\phi) = \left(\sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} - V_\phi(s_t) \right)^2 \tag{10}$$

此外，熵正则项用于保持策略探索性。对于本文 33 维离散动作空间，其熵项表示为

$$S[\pi_\theta](s_t) = -\sum_{i=1}^{33} \pi_\theta(a_i \mid s_t) \ln \pi_\theta(a_i \mid s_t) \tag{11}$$

当各动作选择概率接近均匀分布时，熵达到较大值，有助于避免训练早期过早收敛到局部策略。本文利用下式的梯度方法对策略网络与价值网络参数进行更新：

$$\Theta \leftarrow \Theta + \alpha \frac{\partial L^{\text{Total}}(\Theta)}{\partial \Theta} \tag{12}$$

式中：Θ 表示神经网络参数集合，α 为学习率。由于本文将 $L^{\text{Total}}(\Theta)$ 定义为待最大化的代理目标，因此采用梯度上升形式进行更新；若在程序实现中采用损失函数最小化形式，则等价于最小化其相反数。

动作空间方面，本文将 LVLH 坐标系下三轴推力指令分别离散为{-1.0，-0.5，0，0.5，1.0}五级取值，三轴笛卡尔积得到 125 个候选推力组合。为避免离散组合超过单位推力约束，仅保留模长不大于 1 的动作向量，最终形成 33 个合法离散动作。该动作集合覆盖静默、单轴、双轴及部分三轴组合机动，兼顾推力表达能力与离散策略学习稳定性。

2.2.3 算法运行流程

1) 环境初始化：在每个回合开始时，对追击者和逃逸者的相对位置、速度及轨道要素进行随机初始化，以提升策略对不同初始态势的适应性。

2) 状态观测与动作输出：智能体根据局部观测和 LSTM 隐状态输出 33 维离散动作概率，并采样得到三轴推力控制指令。

3) 动力学转移：将智能体输出的推力输入到环境中，经过 CW 方程积分，得到下一时刻的新状态，并计算当前步的奖励。

4) 数据存储与网络更新：重复上述过程并存储轨迹数据，当样本量达到训练需求后，采用 MAPPO 目标函数批量更新网络参数。在自博弈架构下，追击者与逃逸者交替提升策略能力。

2.3 复合奖励函数设计

为了将物理约束转化为强化学习引导信号。本文设计了一种覆盖全任务周期的复合奖励函数 R_{total}：

$$\begin{aligned} R_{\text{total}} &= R_{\text{dist}} + w_{\text{macro}} R_{\text{macro}} + w_{\text{micro}} R_{\text{micro}} + \\ & R_{\text{appr}} + R_{\text{glide}} + R_{\text{ctrl}} + R_{\text{term}} \end{aligned} \tag{13}$$

式中：R_{dist} 为基础距离惩罚；R_{macro} 为宏观轨道要素奖励；R_{micro} 为微观视线引导奖励；w_{macro} 与 w_{micro} 为随距离动态调整的权重系数；R_{appr} 为连续抵近奖励；R_{glide} 为静默滑行奖励；R_{ctrl} 为控制能耗惩罚；R_{term} 为终端任务结算奖励。追击者奖励针对其分配目标计算，逃逸者奖励针对其最近追击者计算。

2.3.1 动态尺度权重分配机制

为实现从宏观变轨到末端拦截的平滑交接，权重定义为

$$\begin{cases} w_{\text{macro}} = \text{clip}\left(\dfrac{d - d_{\min}}{d_{\max} - d_{\min}}, 0, 1 \right) \\ w_{\text{micro}} = 1 - w_{\text{macro}} \end{cases} \tag{14}$$

式中：d 为相对距离；$d_{\max} = 80\ \text{km}$ 为宏观引导上限；$d_{\min} = 40\ \text{km}$ 为末端引导下限。基础惩罚 $R_{\text{dist}} = -(d / 200000) \times 0.5$ 始终激活。动态权重仅用于追击者，逃逸者以捕获区边界的 10km 为阈值区分奖励。

2.3.2 宏观轨道博弈奖励与速度漏斗

宏观奖励 R_{macro} 包含轨道面夹角惩罚、相位对齐奖励及高度死区约束：

$$R_{\text{macro}} = \lambda_1 (\cos i_{\text{pe}} - 1) + \lambda_2 \exp(-\theta_{\text{pe}} / \sigma) - \zeta_{\text{alt}} \tag{15}$$

式中：i_{pe} 为两星轨道面夹角；θ_{pe} 为轨道相位差；ζ_{alt} 为高度死区惩罚项。本文取相位衰减尺度 σ =0.1rad，$\xi_{\text{alt}} = (|\Delta h| - 20000) / 50000$，仅当 $\theta_{pe} < 0.05\ \text{rad}$ 且 $|\Delta h| > 20000$ 时启用，d≤40km 时 λ_1 / λ_2 =0/1，d≥80km 时为 1/0，40～80km 内线性过渡。

其中，轨道面夹角的调整通常需要较大的法向速度增量，若在远距离阶段不及时校正，末端抵近时将产生较大的面外误差并显著增加燃料消耗。因此，本文在宏观奖励中对轨道面夹角设置较高权重。

相位对齐奖励用于引导追击者在远距离阶段逐渐缩小与逃逸者之间的轨道相位差。考虑到低轨角速度较高的特性，本文不在远距离阶段直接强制两者保持相同轨道高度，从而允许追击者通过适当升轨或降轨形成角速度差，以利用自然轨道演化完成相位追赶。

高度死区约束仅在相位差较小时激活。当双方相位基本对齐后，奖励函数开始约束轨道高度差，引导追击者进入与目标相近的轨道邻域。高度死区可避免智能体因微小高度误差频繁机动，从而兼顾远距离相位覆盖自由度与近距离交会稳定性。

同时为了协调不同距离下的相对速度大小，引导智能体在远端快速接近，近端缓慢抵近，引入基于距离映射的速度漏斗奖励：

$$R_{\text{funnel}}=\begin{cases}1.0\times(1-e_v/e_{\text{allow}}), & \text{if } e_v<e_{\text{allow}}\\ -(e_v/e_{\text{allow}})\times 0.2, & \text{otherwise}\end{cases} \tag{16}$$

式中：$e_v=|v_{\text{rel}}-v_{\text{ideal}}|$ 为速率偏差；$v_{\text{ideal}}=\text{clip}(d/200,50,500)$，为理想接近速率；$e_{\text{allow}}=\text{clip}\left(0.4v_{\text{ideal}},20,100\right)$ 为允许误差带。

2.3.3 末端视线对齐与安全走廊奖励

为了构建 30° 安全接近走廊，设计微观奖励 R_{micro} 如下：

$$R_{\text{micro}}=2\cos\alpha+\begin{cases}2\left(1-\dfrac{\alpha}{30^\circ}\right)+1.5, & \alpha\le 30^\circ,\ v_{\text{rel}}<200\ \text{m/s}\\ 2\left(1-\dfrac{\alpha}{30^\circ}\right)-1.0, & \alpha\le 30^\circ,\ v_{\text{rel}}\ge 200\ \text{m/s}\\ -1.5\dfrac{\alpha-30^\circ}{30^\circ}, & \alpha>30^\circ\end{cases} \tag{17}$$

式中：α 为视线方向与相对接近速度方向之间的夹角。若满足走廊条件，则给予递减走廊奖励；否则施加偏离惩罚。该项并非严格约束，而是用于降低初始抵近时因角度偏差和侧向速度过大造成的掠过风险。

2.3.4 抵近奖励的燃料灵敏度调节

在连续抵近奖励 R_{appr} 中引入了燃料灵敏度系数 β：

$$R_{\text{appr}}=\text{clip}(\Delta d/20,-1.5,1.5)\cdot\beta \tag{18}$$

式中：Δd 为单步距离变化量；$\beta=\text{clip}(m_{\text{ratio}}/0.3,0,1)$ 为燃料敏感系数；m_{ratio} 为剩余燃料比例。该机制引导智能体在燃料受限时减少激进动作。

2.3.5 静默滑行奖励与动能耗散惩罚

为抑制意义不大的速度波动从而节省燃料，设计了静默滑行奖励 R_{glide}：

$$R_{\text{glide}}=\begin{cases}1.5, & \text{if } \|\boldsymbol{u}\|<0.05 \text{ and } \Delta d>50\\ 0, & \text{otherwise}\end{cases} \tag{19}$$

式中：$\boldsymbol{u}$ 为单步燃料消耗量。

当 $\boldsymbol{u}\ge 0.05$ 时，施加燃耗惩罚 $R_{\text{fuel}}=-\boldsymbol{u}^2\eta(f_t)\zeta(d_t)$，其中 $f_t=F_t/F_{\max}$ 为归一化剩余燃料比例，$\zeta(d_t)=\text{clip}(d_t/80000,0.2,2.5)$ 为距离调节系数，$\eta(f_t)=0.05/(f_t+0.05)$。该惩罚在远距离阶段具有较大权重，在近距离阶段平滑降低至 0.2，但不会完全取消，以兼顾燃料节约与末端精细机动。

此处逃逸者奖励与追击者一致，仅去除距离调节系数。

2.3.6 终端结算奖励与防漂移机制

当 $d\le 40$ km 时激活终端结算：

$$R_{\text{term}} = R_{\text{base}} + R_{\text{align}} + R_{\text{drift}} + R_{\text{success}} \tag{20}$$

式中：$R_{\text{base}} = \text{clip}((40000 - d)/6000, 0, 5)$ 为基础奖励；$R_{\text{align}} = 5.0 \times (\boldsymbol{v}_{\text{p}} \cdot \boldsymbol{v}_{\text{e}} / \| \boldsymbol{v}_{\text{p}} \| \, \| \boldsymbol{v}_{\text{e}} \|)$ 为速度矢量对齐奖励，用于抑制末端相对速度漂移；$R_{\text{drift}} = -5$ 仅在 $d_t - d_{t-1} > 2$ m 即脱离锁定条件下触发。若进入捕获区并保持伴飞状态，则给予任务完成奖励和燃料保留奖励，R_{success} 采用分段奖励：40km 内且速度达标奖励 5，10km 内且速度达标额外奖励 15，捕获时给予 10 倍 f_t 的燃料保留奖励。

2.3.7 逃逸者奖励函数

逃逸者奖励由基础距离奖励、分阶段规避引导奖励、连续逃离奖励、燃料奖励和终端状态奖励组成：

$$R_e = R_{\text{dis}}^e + R_{\text{guide}}^e + R_{\text{esc}}^e + R_{\text{fuel}}^e + R_{\text{term}}^e \tag{21}$$

式中：R_{dis}^e 为基础距离奖励，R_{guide}^e 为分阶段规避引导奖励，R_{esc}^e 为连续逃离奖励，R_{fuel}^e 为燃料奖励，R_{term}^e 为终端状态奖励。

基础距离奖励为

$$R_{\text{dis}}^e = 0.5 d_t / 100000 \tag{22}$$

式中：d_t 为逃逸者与当前主要威胁追击者之间的距离，该项使逃逸者在扩大安全距离时获得更高奖励。

当 $d_t > 10$ km 时，采用远距离规避引导奖励。相位奖励取 $R_{\text{phase}}^e = -2\exp(-\Delta\phi_t / 0.1)$。当 $\Delta\phi_t < 0.05$ rad 且 $|\Delta h_t| > 15$ km 时，增加高度差奖励 $R_h^e = (|\Delta h_t| - 15000)/50000$。速度偏离奖励取参考相对速度 $v_t = \text{clip}(d_t / 1000, 2, 200)$，允许误差为 $\varepsilon_{v,t} = \text{clip}(0.5 v_t, 5, 50)$。当 $|v_{\text{rel},t} - v_t^*| > \varepsilon_{v,t}$ 时，给予速度偏离奖励 $R_v^e = 1$。远距离引导奖励为

$$R_{\text{far}}^e = R_{\text{phase}}^e + R_h^e + R_v^e \tag{23}$$

当 $d_t \le 10$ km 时，采用近距离反视线引导奖励。令 a_t 表示视线方向与相对接近速度方向的对齐程度，基础奖励取 $R_{\text{near}}^e = -2a_t$；当 $a_t > 0.5$ 时，在此基础上额外施加 1.5 的锁定惩罚。

$$R_{\text{guide}}^e = \begin{cases} R_{\text{far}}^e, & d_t > 10 \text{ km} \\ R_{\text{near}}^e, & d_t \le 10 \text{ km} \end{cases} \tag{24}$$

为直接鼓励逃逸者拉开距离，设置连续逃离奖励：

$$R_{\text{esc}}^e = \text{clip}(\Delta d_t / 20, -2, 3) \tag{25}$$

式中：$\Delta d_t > 0$ 表示逃逸者与追击者之间的距离增大，$\Delta d_t < 0$ 表示追击者正在缩短追逃距离。

逃逸者的燃料奖励同时考虑静默滑行和主动机动。令 $q_t = |\mathbf{u}_t|$ 为单步控制量，f_t 为归一化剩余燃料比例。当 $q_t < 0.05$ 且 $\Delta d_t > -50$ m 时，给予 1.5 的静默滑行奖励；当 $q_t \ge 0.05$ 时，施加与控制量平方及剩余燃料比例相关的燃耗惩罚：

$$R_{\text{fuel}}^e = \begin{cases} 1.5, & q_t < 0.05,\ \Delta d_t > -50 \text{ m} \\ 0, & q_t < 0.05,\ \Delta d_t \le -50 \text{ m} \\ -q_t^2 [0.1/(f_t + 0.1)], & q_t \ge 0.05 \end{cases} \tag{26}$$

若 $q_t < 0.05$，但逃逸者在单步内被追击者拉近超过 50 m，则不提供额外奖励。

最后，根据逃逸者所处的近距离状态设置终端奖励。进入 30 km 范围后施加危险区域惩罚；进入 10 km 范围且相对速度不超过 10 m/s 时，施加低速锁定惩罚；若进入 10 km 范围但仍保持较大的相对速度，则给予飞掠奖励和燃料保留奖励。终端奖励为

$$R_{\text{term}}^e = \begin{cases} 0, & d_t \ge 30 \text{ km} \\ -2, & 10 \text{ km} \le d_t < 30 \text{ km} \\ -22, & d_t < 10 \text{ km},\ v_{\text{rel},t} \le 10 \text{ m/s} \\ 13 + 0.2F_t, & d_t < 10 \text{ km},\ v_{\text{rel},t} > 10 \text{ m/s} \end{cases} \tag{27}$$

式中：F_t 为逃逸者当前剩余燃料。上述 10 m/s 仅为逃逸者奖励函数中的低速锁定阈值，与任务评

价中的捕获判定阈值相互独立。

2.4 基于辅助预测头的策略网络设计

在传统 MAPPO 中，Actor 主要负责策略输出。为增强轨道动力学表征和对手建模能力，本文在 Actor 侧引入物理状态预测头与对手意图预测头。辅助预测头仅作为多任务学习约束和短时观测缺失时的补充估计，不改变主策略动作空间。图 3 为追击者网络结构示意图。

网络结构上，Actor 与 Critic 均采用多层感知机（multi-layer perceptron, MLP）编码器—LSTM 循环层—输出头的形式。其中，MLP 用于对当前时刻的观测输入进行非线性特征提取，LSTM 用于聚合历史观测信息并刻画非完备信息条件下的时序依赖。Actor 以各智能体局部观测为输入，经 MLP 提取单步特征后输入 LSTM，以聚合历史相对运动和对手机动信息，随后由动作头输出 33 维离散动作概率；Critic 以集中式状态信息为输入，经 MLP 和 LSTM 后输出状态价值估计。

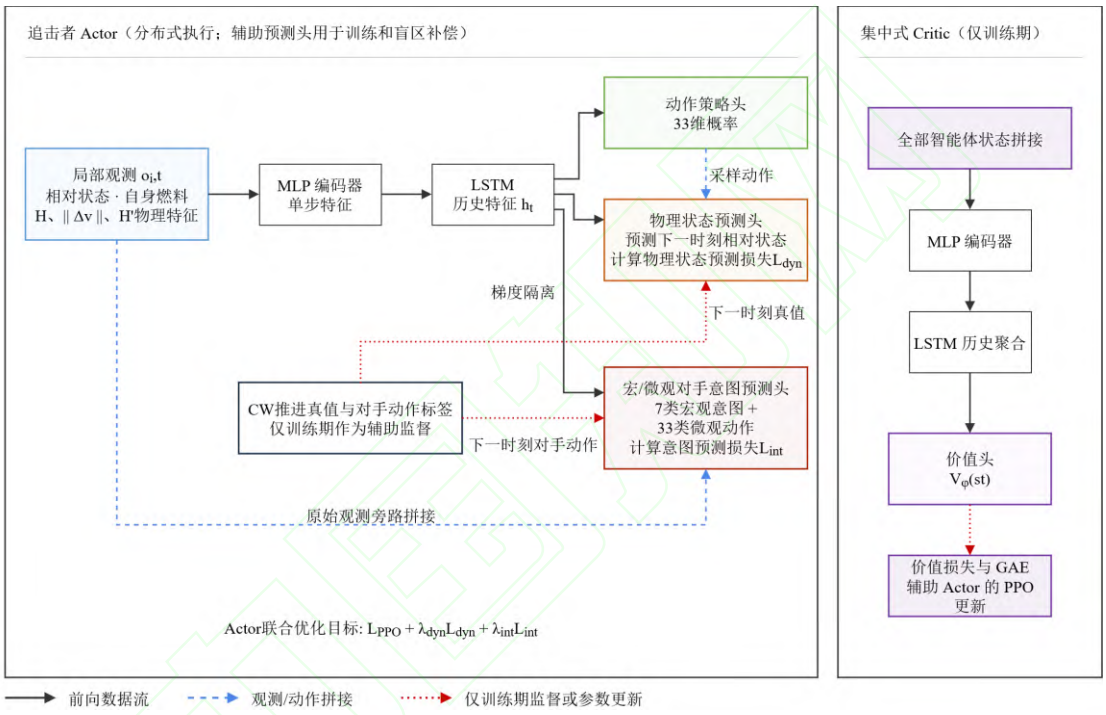

图 3　追击者网络结构示意图

Fig.3　Schematic diagram of the pursuer network architecture

2.4.1 物理状态预测头

物理状态预测头用于约束网络隐层特征符合 CW 相对运动规律。该分支基于当前状态与采样动作预测下一时刻相对状态，从而为 Actor 特征提供物理一致性约束。

1) 将 Actor 骨干网络输出的隐层特征 h_t 与当前采样的动作特征 a_t 进行拼接，输入到一个包含 ReLU 激活函数的 MLP 中，输出预测的下一时刻相对状态 $\hat{s}_{t+1}$。

$$\hat{s}_{t+1}=g_{\mathrm{dyn}}([(h_t),\boldsymbol{e}(a_t)]) \tag{28}$$

式中：$\boldsymbol{h}_t$ 为 Actor 骨干网络输出的隐层特征；$\boldsymbol{e}(a_t)$ 为当前采样动作 a_t 的独热编码；$g_{\mathrm{dyn}}(\cdot)$ 为物理状态预测网络；$\hat{s}_{t+1}$ 为预测的下一时刻六维相对状态。

2) 损失函数：训练阶段，将预测状态与基于 CW 方程推进得到的下一时刻参考状态进行比较。相对位置和相对速度分别按照 150000m 和 500m/s 归一化，物理状态预测损失定义为：

$$L_{dyn}=1/(6B)\sum_{b=1}^{B}\|\hat{s}_{t+1}^{(b)}-s_{t+1}^{CW,(b)}\|_2^2 \tag{29}$$

式中：B 为批量样本数；s_{t+1}^{CW} 为根据当前相对状态、采样动作和 CW 动力学模型计算得到的下

一时刻参考状态。当前实现中取 $\lambda_{\mathrm{dyn}}=0.1$ 从而与策略损失尺度匹配。

2.4.2 宏微观双尺度对手意图预测头

博弈过程中，对手动作预测能够为非完备观测下的状态估计提供辅助。本文在 Actor 网络中并联意图预测头，用于推断逃逸者下一步可能的机动方向。

该模块的设计主要包含两部分。

1) 梯度隔离与特征增强。在网络前向传播时，意图预测头不直接采用 Actor 主干特征，而采用梯度截断后的主干特征并与原始观测重新拼接。该梯度隔离机制可减弱辅助分类任务对 Actor 主策略优化方向的干扰；同时结合 LayerNorm 层，提高深层 MLP 对微弱观测变化的敏感度。

2) 7D/33D 双尺度分类架构。逃逸者实际动作空间为 33 维离散推力矢量。直接进行 33 分类在训练初期较难收敛，因此本文将预测任务解耦为 7 维宏观意图分类和 33 维微观动作分类。宏观意图按推力象限划分，微观动作对应具体离散推力索引。

损失函数根据预测意图概率分布与实际意图交叉熵计算，分为宏观意图预测损失与微观推力预测损失混合构成，设置总预测损失 $L_{int}=0.3L_7+0.7L_{33}$，由于此处采用梯度隔离，所以无需严格匹配策略损失尺度，设置 $\lambda_{\mathrm{int}}=1$。

通过双尺度预测结构，意图预测分支能够从 Actor 隐层表征和原始观测中提取宏观机动趋势与细粒度推力模式，从而为追击者非完备信息决策提供对手行为先验。

2.4.3 基于哈密顿量变化率的燃料状态估计观测

在有限燃料追逃博弈中，对手剩余燃料是影响策略选择的重要隐变量，但追击者无法直接获得逃逸者燃料状态。为此，本文在追击者观测空间中引入基于 CW 相对动力学的哈密顿量及其变化率，将目标相对参考轨道的能量变化作为对手主动机动强度的代理特征。

为刻画轨道追逃过程中由主动机动引起的相对轨道能量变化，本文在观测向量中引入基于 CW 相对动力学的哈密顿量及其变化率。根据三维 CW 相对动力学，可构造如下能量型哈密顿量：

$$H_{CW}=\frac{1}{2}\left(\dot{x}^2+\dot{y}^2+\dot{z}^2\right)-\frac{3}{2}n^2x^2+\frac{1}{2}n^2z^2 \tag{30}$$

式中：第一项表示三维相对运动的动能项；第二项表示径向相对偏移引起的等效势能项；第三项表示面外方向相对运动对应的回复势能项。该哈密顿量不表示航天器绝对轨道能量，其构造基于 CW 线性化相对运动模型，可用于描述局部轨道坐标系下相对运动状态的能量变化特征。

下面说明该量在无控制输入条件下保持恒定。对其关于时间求导，可得

$$\dot{H}_{CW}=\dot{x}\ddot{x}+\dot{y}\ddot{y}+\dot{z}\ddot{z}-3n^2x\dot{x}+n^2z\dot{z} \tag{31}$$

由无控 CW 方程可得

$$\begin{cases}\ddot{x}=2n\dot{y}+3n^2x\\ \ddot{y}=-2n\dot{x}\\ \ddot{z}=-n^2z\end{cases} \tag{32}$$

代入上式后，科氏耦合项、径向势能项以及面外势能项相互抵消，因此有

$$\dot{H}_{CW}=0 \tag{33}$$

这说明在无控制输入、无外部扰动且满足 CW 线性化假设的情况下，三维 CW 哈密顿量保持恒定。因此，当航天器仅依靠自然轨道动力学滑行时，其哈密顿量理论上不会发生变化。

当考虑控制加速度输入时，此时对 H_{CW} 求导并代入受控 CW 方程，可得

$$\dot{H}_{CW}=\dot{x}a_x+\dot{y}a_y+\dot{z}a_z \tag{34}$$

记三维相对速度向量和控制加速度向量分别为 $\boldsymbol{v}_{rel}=\begin{bmatrix}\dot{x} & \dot{y} & \dot{z}\end{bmatrix}^T$，$\boldsymbol{a}=\begin{bmatrix}a_x & a_y & a_z\end{bmatrix}^T$，则哈密顿量变化率可写为

$$\dot{H}_{CW}=\boldsymbol{v}_{rel}^T\boldsymbol{a} \tag{35}$$

在无控制滑行阶段，目标状态主要由自然轨道动力学推进，在短时间内变化缓慢，仅会由于计算误差影响出现少量变化。当逃逸者实施脉冲机动时，相对速度和径向能量项会出现突变，使哈密顿量及其变化率出现显著响应。因此，可以通过哈密顿量变化率表征对手当前是否正在向轨道系统注入控制能量。

实际输入网络时，对哈密顿量和哈密顿量变化率分别采用能量基准 1.0e5 和变化率基准 5.0e3 进行截断归一化，并与目标相对速度模共同组成能量观测子向量。

在本文 2v2 并行追逃场景中，追击者与逃逸者的观测结构并不完全一致。逃逸者需要同时观测两个追击者的相对位置、相对速度、距离关系、轨道高度/相位关系、最近追击者标识以及自身燃料状态，其观测维度为 29 维；而追击者采用按索引绑定目标的局部追逃结构，只需观测自身相对参考速度、另一追击者相对位置、目标逃逸者相对位置与速度、距离、高度/相位关系、光照方向优势及自身燃料比例，原始观测维度为 17 维。为适配统一的多智能体训练接口和批量张量存储形式，本文将追击者观测补齐至与逃逸者一致的 29 维输入。补齐的 12 维并非单纯无效填充，其中前 3 维用于引入物理辅助观测，分别为归一化 CW 哈密顿量、目标相对速度模长和归一化哈密顿量变化率；其余 9 维置零，用于保持输入维度一致。该设计既保证了异构智能体观测维度兼容，又为追击者提供了与目标主动机动强度相关的物理先验信息。

图 4 给出了 CW 动力学演化过程中哈密顿量变化率的响应情况。约第 100 个仿真步（约 2000 s）处哈密顿量变化率快速衰减至零附近，主要对应目标由主动脉冲机动转入无控滑行阶段；根据式(33)和式(35)，控制输入停止后相邻时刻哈密顿量差值减小，变化率自然趋于稳定。该现象表明所构造特征能够反映目标主动机动强度的变化。

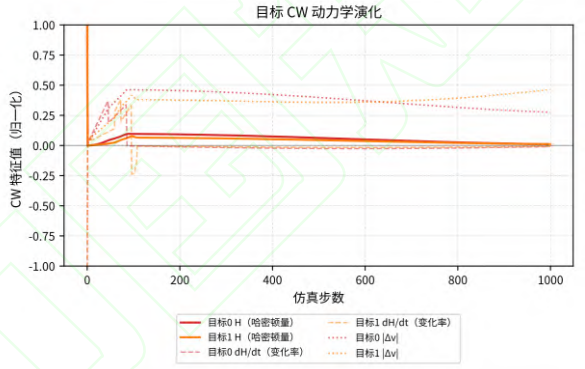

图 4　CW 方程动力学演化
Fig.4　Dynamical evolution curve of the CW equations

该估计量不直接测量逃逸者真实燃料，其本质是由可观测状态构造的物理一致代理特征。其优势在于不破坏非完备信息假设，同时为策略输出和意图预测头提供燃料相关的时序线索。

3 仿真分析

为验证本文提出的非对称课程自博弈 MAPPO 策略的有效性，本文基于 Python 和 MPE 框架构建 GEO 附近 CW 相对运动仿真环境，并引入航天器推力与燃料预算约束。

3.1 仿真环境与物理参数设置

实验设定在地球同步轨道附近，采用 2v2 并行追逃场景。追击者按照索引与逃逸者形成局部追逃关系，该设置主要用于验证共享策略、集中式价值函数和分布式执行机制在多目标并行交互环境下的稳定性。当前奖励函数主要围绕局部追逃关系构建，双方均具备有限燃料预算和三维离散推力机动能力，具体参数如表 1 所示。

表 1 轨道博弈环境与航天器物理仿真参数

Table 1 Physical simulation parameters of the orbital game environment and spacecraft

参数类别	参数名称	参数数值/设定
轨道动力学环境	地球引力常量	3.986e14m³/s²
	虚拟参考圆轨道半径	42164km
	参考轨道平均速度	3074.6m/s
	单步仿真时间步长	20s
	单回合最大仿真步数	1000
追击者设置	智能体数量	2
	速度增量上限	300m/s
	推力机动放大系数	3.0
	初始相对逃逸者距离分布	80km~120km
逃逸者设置	智能体数量	2
	速度增量上限	150m/s
	推力机动放大系数	2.0
	初始相对参考点距离分布	10km~60km
控制约束	控制动作空间	33 维离散推力矢量

为保证深度强化学习网络对不同量纲物理状态的泛化提取能力，必须对输入观测空间进行严格的边界归一化，归一化参数如表 2 所示。

表 2 状态空间归一化基准系数

Table 2 Normalization constants for the state space

参数名称	变量符号	归一化上限/系数
最大相对位置基准	MAX_POS_NORM	150,000 m
最大相对速度基准	MAX_VEL_NORM	500 m/s

3.2 逃逸者策略生成与追击者课程训练

3.2.1 自博弈获得强逃逸者策略

为获得具有足够对抗强度的固定测试对象，本文首先通过自博弈提升逃逸者策略能力。初始阶段利用已具备基础抵近能力的追击者与逃逸者进行闭环对抗，使逃逸者在有限燃料和推力约束下逐渐学习利用轨道相位差、高度差以及脉冲机动进行规避。该阶段追击者成功率可能出现波动，但能够产生更具对抗强度和多样性的逃逸者策略集合。

图 5 和图 6 展示了自博弈前后追逃模型的表现变化。可以看出，自博弈后的逃逸者与最近追击者间的距离发散概率大幅提高，且即使被成功拦截的样本也有一部分相对距离会在后期发散，这表示在自博弈中逃逸者占据优势，形成了更复杂的规避行为。在此基础上，本文冻结自博弈获得的强逃逸者策略，并将其作为后续实验的固定对手。

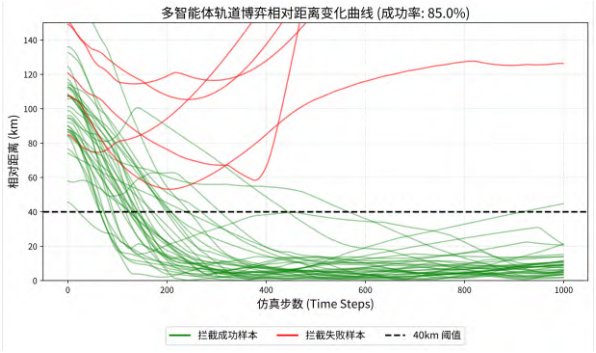

图 5　自博弈前的预训练追击逃逸模型表现
Fig.5　Performance of the pre-trained pursuit-evasion models prior to self-play

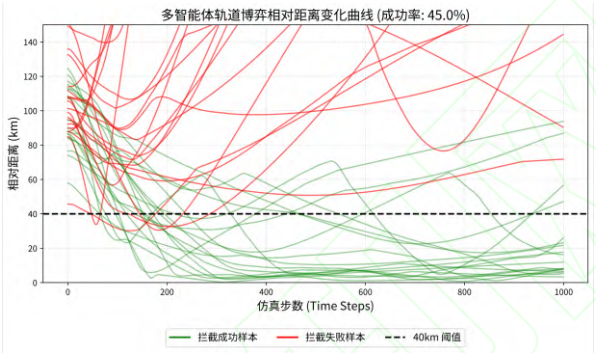

图 6　自博弈后的预训练追击逃逸模型表现
Fig.6　Performance of pre-trained pursuit-evasion models after self-play

3.2.2 基于强逃逸者的追击者课程训练

在上述强逃逸者策略基础上，本文进一步训练追击者。首先设置无课程学习对照实验，使随机初始化追击者直接面对固定强逃逸者进行训练。结果表明，直接训练时，复合奖励与常规距离奖励对应策略的压制率和捕获率均为 0%，追击者平均燃料消耗均达到 300m/s，说明策略主要表现为高燃料消耗下的无效机动，未形成稳定追捕能力。

随后采用分阶段课程训练：先在无限燃料条件下针对弱机动或静态目标学习基础轨道抵近规律，获得无限燃料基础追击者；再逐步引入有限燃料约束和强逃逸者对抗环境。该流程使追击者在面对高强度逃逸者前已具备基本抵近和末端速度控制能力。在此基础上，本文进一步比较常规距离-终端奖励与所提复合奖励在有限燃料条件下的追逃性能差异。

3.3 不同奖励函数下的追逃性能对比

为验证本文复合奖励函数在有限燃料轨道追逃任务中的有效性，本文以课程训练得到的无限燃料基础追击者为初始策略，并在固定强逃逸者条件下设置常规距离-终端奖励作为基线进行对比。该基线奖励由距离缩短奖励、距离惩罚、压制区奖励、终端捕获奖励和动作平方惩罚组成，不包含相位—高度解耦、速度漏斗、视线走廊、静默滑行奖励以及终端燃料保留奖励等物理先验项。两组实验采用相同 MAPPO 网络结构、训练轮数、学习率衰减策略、燃料约束和评估随机种子。

表 3 汇总了无课程训练与课程训练后的主要评估结果。可以看出，随机初始化追击者直接训练至 300 代后，无论采用复合奖励还是常规距离奖励，压制率和捕获率均为 0%，且平均燃料消耗达到 300m/s。相比之下，采用课程基础模型继续训练后，本文复合奖励下的压制率和捕获率分别达到 90.50%和 63.50%，高于常规距离奖励的 59.50%和 31.00%；平均燃料消耗由 277.21m/s 降低至 207.79m/s，最终相对速度由 23.42m/s 降低至 4.39m/s。结果表明，课程学习能够显著降低初始策略搜索难度，而复合奖励进一步提高了有限燃料条件下的捕获稳定性和末端速度控制能力。

表 3 不同训练方式与奖励函数下的轨道追逃性能对比

Table 3 Performance comparison under different training settings and reward functions

指标	无课程 复合	无课程 常规	课程 复合	课程 常规
压制率/%	0.00	0.00	90.50	59.50
捕获率/%	0.00	0.00	63.50	31.00
压制时间/s	—	—	5224	5265
捕获时间/s	—	—	10669	8786
燃料消耗/(m/s)	300.00	300.00	207.79	277.21
末端速度/(m/s)	211.63	246.04	4.39	23.42

注：压制/捕获时间均指首次连续满足相应判据区间的起始时刻；平均值仅统计成功样本。

选取两种方法在相同初始条件下均成功捕获的 50 组样本，比较累计燃料消耗和相对速度演化，如图 7 和图 8 所示。细线表示单个样本曲线，粗线表示 50 组样本平均曲线。可以看出，常规距离奖励在成功样本中仍表现出更高的燃料消耗，且后期相对速度存在回升和波动；本文复合奖励能够在保持低相对速度的同时减少燃料消耗，体现了相位—高度解耦、速度漏斗和滑行奖励对有限燃料追逃任务的约束作用，常规距离奖励在成功样本中的平均捕获时间较短，主要是因为该奖励更直接地鼓励欧氏距离快速缩短，容易形成较激进的直接追赶策略，但从整体评估结果看，其捕获率较低，燃料消耗和末端相对速度明显偏高，说明其快速捕获主要出现在少量较易成功样本中，而本文复合奖励更侧重于提高整体捕获稳定性和燃料利用效率。

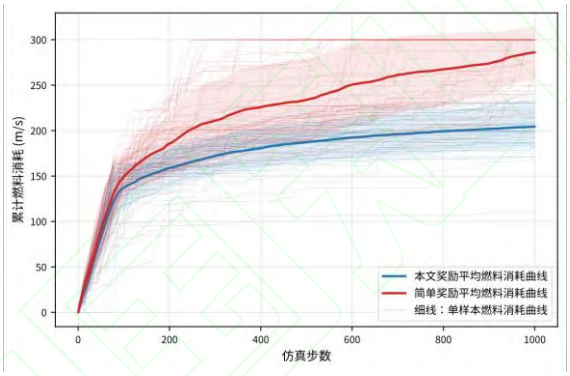

图 7 不同奖励函数下的累计燃料消耗对比

Fig.7 Cumulative fuel consumption under different reward functions

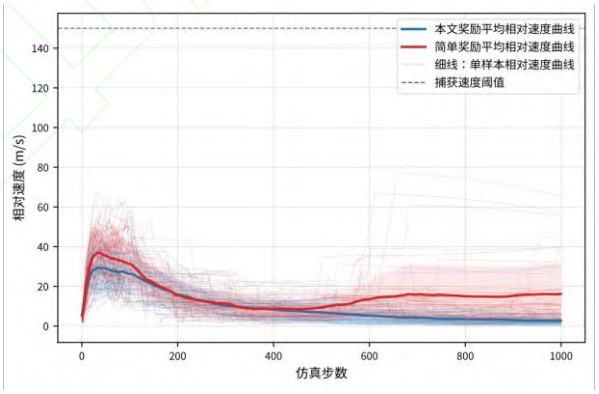

图 8 不同奖励函数下的相对速度演化对比

Fig.8 Relative-speed evolution under different reward functions

为进一步分析相位—高度解耦奖励是否诱导出符合轨道相位演化规律的控制行为，本文对两种奖励函数下均成功捕获的 90 组配对样本进行了统计。设沿轨相位差为 $rel_y = y_{evader} - y_{pursuer}$，其中 $rel_y > 0$ 表示逃逸者位于追击者前方；定义高度差为 $\Delta h = \| \boldsymbol{p}_{pursuer} \| - \| \boldsymbol{p}_{evader} \|$，其中 $\Delta h > 0$ 表示追击者处于更高轨道。根据低轨快、高轨慢的相位演化特性，当 $rel_y \cdot \Delta h < 0$ 时，可认为追击者的高度调节方向与相位追

赶规律一致。统计结果表明，本文复合奖励策略的相位—高度调节一致性比例为 90.75%，高于常规距离奖励的 86.25%；在满足该调节关系的状态中，本文奖励策略使沿轨相位误差继续缩小的比例为 77.61%，高于常规距离奖励的 70.74%，说明相位—高度解耦项能够在策略行为层面体现低轨快、高轨慢的轨道动力学先验。

3.4 辅助预测头与短时盲区预测分析

为进一步验证辅助预测头在对手建模和短时观测缺失条件下的作用，本文对有无哈密顿量观测两种设置下的预测头输出进行了统计分析，结果如图 9、图 10 所示。引入哈密顿量变化率观测后，预测结果更集中于高置信度区间，不同置信度阈值下的预测准确率整体提高，说明 CW 哈密顿量变化率能够为网络提供目标主动机动强度的辅助信息。由于逃逸者动作具有脉冲机动与自然滑行交替的非连续特征，预测置信度难以长期保持在极高水平，但整体结果表明，辅助预测头能够较好识别目标宏观机动倾向。

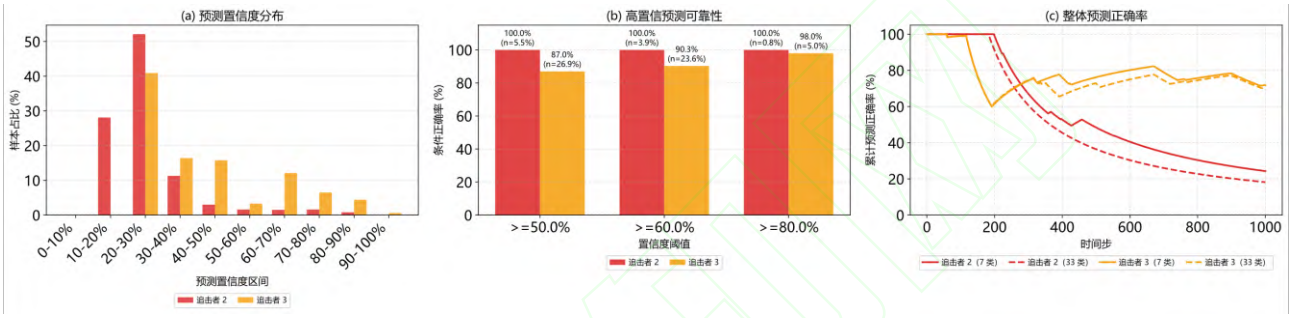

图 9　未引入哈密顿量观测的辅助预测头结果
Fig.9 Auxiliary prediction head results without Hamiltonian observation

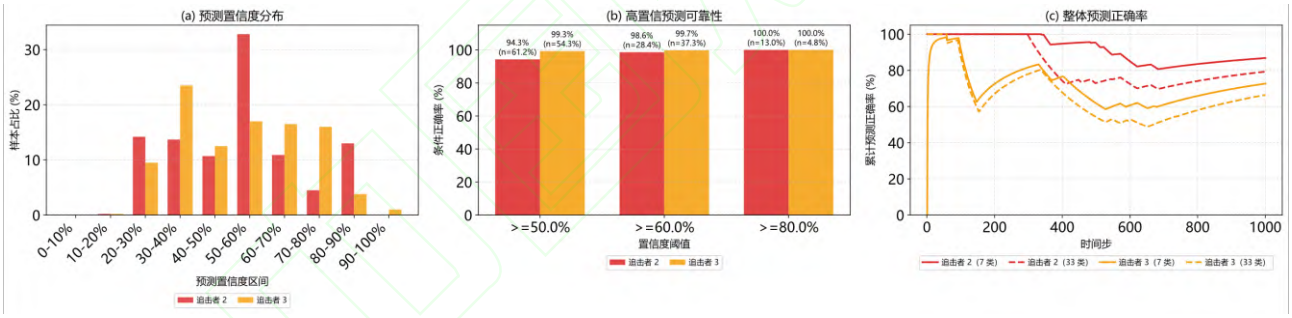

图 10　引入哈密顿量观测后的辅助预测头结果
Fig.10 Auxiliary prediction head results with Hamiltonian observation

进一步地，本文设置 50~100 步和 150~200 步两个短时盲区区间，将神经网络预测、多项式外推和置信度门控混合预测进行对比，结果如图 11 和图 12 所示。多项式外推仅利用目标进入盲区前的历史位置序列进行二阶拟合，不引入轨道动力学模型和对手意图信息。实验结果表明，在预测头置信度较高的区间，神经网络预测误差低于多项式外推；而在置信度较低的区间，纯神经网络预测容易产生误差累积。置信度门控方法根据预测头最大输出概率判断是否采用神经网络预测，从而在高置信度区间利用神经网络的预测能力，在低置信度区间避免误差快速扩散。

图 11 对应高置信度盲区，神经网络预测能够较好地保持目标状态估计精度。

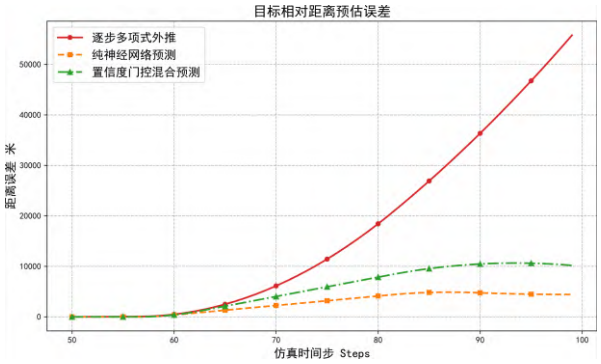

图 11　第 50–100 步盲区轨迹预测误差
Fig.11 Trajectory prediction error during the blind interval from step 50 to 100

图 12 对应低置信度盲区，门控混合预测能够避免纯神经网络误差快速累积，体现了利用预测置信度选择状态补偿方式的必要性。

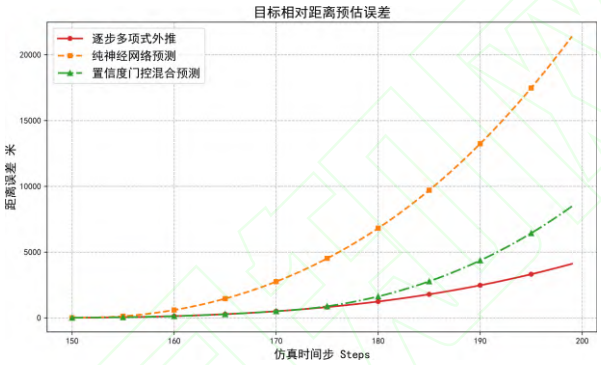

图 12 第 150–200 步盲区轨迹预测误差
Fig.12 Trajectory prediction error during the blind interval from step 150 to 200

4 结 论

1) 自博弈训练能够获得具有较强规避能力的逃逸者策略集合，为追击者训练提供稳定的高强度对抗对象。在该强逃逸者条件下，随机初始化追击者直接训练至 50 代和 300 代时压制率与捕获率均为 0%，而分阶段课程学习能够使追击者先获得基础轨道抵近能力，再逐步适应有限燃料与强机动逃逸者约束，从而提高训练稳定性。

2) 复合奖励函数能够诱导追击者形成脉冲—滑行式控制行为，有效降低追击者的整体燃料使用量，并提升有限燃料条件下的压制与捕获稳定性、燃料利用效率及末端速度控制能力。

3) 基于 CW 哈密顿量变化率的物理先验观测和辅助预测头可为短时观测缺失提供补充状态估计。置信度门控混合预测能够在高置信度区间利用神经网络预测优势，并在低置信度区间降低误差累积风险。

此外，本文验证主要基于 GEO 附近 CW 线性化相对运动模型，尚未考虑 J2 摄动、非圆参考轨道及高保真非线性相对动力学影响，相关扩展将在后续工作中进一步研究。

参考文献（References）

[1] 袁利，姜甜甜．航天器威胁规避智能自主控制技术研究综述[J]．自动化学报，2023，49（2）：229-245.
YUAN L,JIANG T T.Review on intelligent autonomous control for spacecraft confronting orbital threats[J].Acta Automatica Sinica,2023,49(2):229-245(in Chinese).

[2] 高婉莹，吴健发，魏春岭．航天器威胁规避自主决策规划方法研究综述[J]．中国空间科学技术（中英文），2024，44（4）：71-89.
GAO W Y,WU J F,WEI C L.Review on spacecraft autonomous decision-making and planning for orbital threat avoidance[J].Chinese Space Science and Technology,2024,44(4):71-89(in Chinese).

[3] ISAACS R.Differential games[M].New York:Wiley,1965.

[4] 郑鑫宇，张轶，周杰，等．回合制轨道博弈中 MCTS 算法的改进与应用[J]．中国空间科学技术（中英文），2024，44（5）：75-82．
ZHENG X Y,ZHANG Y,ZHOU J,et al.Improvement and application of MCTS in turn-based orbital games[J].Chinese Space Science and Technology,2024,44(5):75-82(in Chinese).
[5] LIU Y,ZHANG Y,JIANG J,et al.Multiple-to-one orbital pursuit:a computational game strategy[J].IEEE Transactions on Aerospace and Electronic Systems,2025,61(2):2213-2225.
[6] 吴健发，魏春岭，张海博，等．基于深度强化学习的航天器多约束规避动作快速规划[J]．空间控制技术与应用，2023，49（2）：1-9．
WU J F,WEI C L,ZHANG H B,et al.Spacecraft multi-constraint rapid avoidance motion planning based on deep reinforcement learning[J].Aerospace Control and Application,2023,49(2):1-9(in Chinese).
[7] 高婉莹，吴健发，魏春岭．不完全信息下的航天器事件触发威胁规避控制[J]．空间控制技术与应用，2025，51（3）：46-59．
GAO W Y,WU J F,WEI C L.Event-triggered impulsive evasion control for spacecraft under incomplete information[J].Aerospace Control and Application,2025,51(3):46-59(in Chinese).
[8] PONTANI M,CONWAY B A.Numerical solution of the three-dimensional orbital pursuit-evasion game[J].Journal of Guidance,Control,and Dynamics,2009,32(2):474-487.
[9] HE H,BOYD-GRABER J,KWOK K,et al.Opponent modeling in deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning.New York:PMLR,2016,48:1804-1813.
[10] 耿远卓，袁利，黄煌，等．基于终端诱导强化学习的航天器轨道追逃博弈[J]．自动化学报，2023，49（5）：974-984．
GENG Y Z,YUAN L,HUANG H,et al.Terminal guidance based reinforcement-learning for orbital pursuit-evasion game of the spacecraft[J].Acta Automatica Sinica,2023,49(5):974-984(in Chinese).
[11] 袁利，耿远卓，汤亮，等．航天器轨道追逃博弈多阶段强化学习训练方法[J]．上海航天（中英文），2022，39（4）：33-41．
YUAN L,GENG Y Z,TANG L,et al.Multi-stage reinforcement learning method for orbital pursuit-evasion game of spacecrafts[J].Aerospace Shanghai(Chinese & English),2022,39(4):33-41(in Chinese).
[12] 王英杰，袁利，汤亮，等．信息非完备下多航天器轨道博弈强化学习方法[J]．宇航学报，2023，44（10）：1522-1533．
WANG Y J,YUAN L,TANG L,et al.Reinforcement learning method for multi-spacecraft orbital game with incomplete information[J].Journal of Astronautics,2023,44(10):1522-1533(in Chinese).
[13] YU T,WU H,YIN Q,et al.Diverse strategy generation for orbital pursuit-evasion games using distributed reinforcement learning[J].Aerospace Science and Technology,2026,168:111052.
[14] VENIGALLA C,SCHEERES D J.Delta-V-based analysis of spacecraft pursuit-evasion games[J].Journal of Guidance,Control,and Dynamics,2021,44(11):1961-1971.
[15] LI Z,LUO Y.Orbital impulsive pursuit-evasion game formulation and Stackelberg equilibrium solutions[J].Journal of Spacecraft and Rockets,2025,62(2):349-364.