航天器控制自主推理：基于组相对策略优化的大语言模型框架

Amit Jain麻省理工学院航空航天系
马萨诸塞州剑桥市 02139
amitjain@mit.edu

Richard Linares麻省理工学院航空航天系马萨诸塞州剑桥市 02139
linaresr@mit.edu

arXiv:2601.04334v1 [cs.RO] 7 Jan 2026

摘要——**本文提出一种基于学习的制导与控制方法，将具备推理能力的大语言模型（LLM）与组相对策略优化（GRPO）相结合。该方法采用两阶段流程训练针对各类环境的控制器：首先通过监督微调（SFT）学习格式规范和控制原语，随后利用 GRPO 根据交互结果改进策略。本文在四种控制问题上展示了该框架，涵盖不同程度的动力学复杂性，从典型线性系统、非线性振荡动力学，到具有陀螺耦合和推力约束的三维航天器姿态控制。结果表明，在一致的训练设置下，通过 GRPO 优化且具备显式推理能力的大语言模型，能够合成可行的稳定策略，适用于线性和非线性系统。两阶段训练方法使模型能够生成控制序列，同时提供对其决策过程的可读解释。本研究为将基于 GRPO 的推理应用于自主控制系统奠定了基础，并有望用于航空航天及其他安全关键领域。**

I. 引言

随着任务不断远离地球、复杂程度日益提高，太空探索对自主能力的需求也越来越大。传统的航天器控制依赖地面操作和预编程序列，这限制了适应能力，也因通信延迟而带来脆弱性。人工智能领域的最新进展，尤其是大语言模型（LLM），为提升航天器自主能力提供了颇具前景的途径。这些模型已展现出在多种领域中进行推理、模式识别和策略生成的强大能力 [11]。

Zucchelli 等人 [20] 近期的突破性研究表明，经过微调的大语言模型能够有效控制空间系统，而且所需训练数据比传统神经网络更少。研究人员已证明，像 Llama-2 这样的微调大语言模型能够为航天器控制生成高精度推力指令，适用于轨道转移和月球着陆；另有研究则在坎巴拉太空计划等模拟环境中取得了令人鼓舞的成果 [3]。基于 Transformer 的强化学习在多阶段航天器轨迹优化方面也展现出应用潜力 [10]。先前的研究已成功将深度强化学习应用于航天器对接机动 [14]以及结合元学习的自适应制导 [6]。然而，这些方法通常在单一领域环境中运行，无法实现跨复杂度迁移，并且更倾向于将控制视为模式识别，而非由推理驱动的策略生成。

尽管基于大语言模型的控制取得了这些令人鼓舞的进展，当前研究仍存在两项关键局限。首先，缺乏能够将控制策略从简单的学术系统迁移到复杂航天器动力学系统，同时保留推理能力的多环境泛化框架。目前尚无系统的方法，能够在维持已学习的控制策略及其相关推理过程的同时，逐步将基础控制问题扩展至航空航天应用。其次，尽管组相对策略优化（GRPO）在数学推理和代码生成任务中已取得显著成功 [5, 4],，但其在控制理论中的应用几乎尚未得到探索，这对于实现样本高效的自主控制而言，是一个重大的错失机会。

关键的是，现有航空航天控制系统缺乏自主推理这一根本能力：分析复杂情景、提供决策依据说明、适应非标称条件以及实现分层决策的能力。这是当前文献中的一个关键空白，因为此前没有研究专门为控制应用开发具备推理能力的模型。不同于依赖预编程响应的反应式方法，基于推理的控制器能够动态评估系统状态、比较多种控制策略，并透明地解释所选动作；对于在不确定环境中运行的下一代航空航天系统而言，这些能力至关重要。近期研究已表明，迭代推理方法能够实现高效的最优控制综合 [8]。

组相对策略优化通过免去单独的价值函数估计，同时激励结构化推理 [1, 17]，解决了传统强化学习的局限性。GRPO 不再训练评论家网络，而是采用基于组的归一化，通过以下方式计算优势：

979-8-3315-7360-7/26/$31.00 ©2026 IEEE

将每个样本与同组的其他样本进行比较，从而通过相对性能比较，以较高的样本效率改进策略。这种方法天然适用于大语言模型推理架构，并显著降低内存需求；对于资源受限的航空航天应用而言，这是一项至关重要的优势。

本文提出了一种具备推理能力的框架，系统地将 GRPO 应用于自主控制问题。该方法实现了从反应式控制向推理驱动的策略生成的范式转变：控制器会明确分析系统动力学、评估不同控制方案，并逐步说明其决策依据。通过监督微调（SFT）建立基础控制知识，使大语言模型在生成控制策略之前学会推理动力学约束。在此阶段，模型会生成结构化推理轨迹，为其控制决策提供依据。随后，GRPO 通过基于组的比较实现高效的策略优化，使推理原则能够指导控制生成。

该框架在四个控制问题上进行了验证，这些问题的动力学复杂度逐步增加，涵盖从线性质点动力学、非线性振荡系统，到涉及连续推力优化和多轴姿态稳定的三维航天器操作。这一递进过程验证了该框架能够从基础控制理论基准扩展到真实的航空航天应用，同时保持推理保真度和控制性能。

本文其余部分安排如下：第二节阐述使用大语言模型和 GRPO 进行控制的问题；第三节详细介绍方法，包括模型架构、训练过程和实现考量；第四节展示四种控制场景下的仿真结果；第五节总结本文的贡献并展望未来研究方向。

II. 问题表述

为形式化自主控制挑战，控制被表述为一个序列决策过程，其中基于大语言模型的智能体必须在满足物理约束的同时生成精确的数值输出，并提供可解释的说明。

A. 控制作为序列决策过程

动力系统的控制被表述为由元组 $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ [16] 定义的马尔可夫决策过程（MDP）。状态空间 $\mathcal{S}$ 是连续的：双积分器和范德波尔振荡器为二维（$\mathcal{S} \subset \mathbb{R}^2$），轨道抬升和航天器消旋为三维（$\mathcal{S} \subset \mathbb{R}^3$）。动作空间 $\mathcal{A}$ 由受执行器限制约束的控制输入构成：较简单系统采用标量控制（$\mathcal{A} \subset \mathbb{R}$），轨道机动采用推力角（$\mathcal{A} \subset [0, 2\pi]$），姿态控制采用三轴力矩向量（$\mathcal{A} \subset \mathbb{R}^3$）。

对于所考虑的确定性系统，状态转移函数 P: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ 可简化为 $s_{t+1} = f(s_t, a_t)$ 其中，f 表示控制系统的微分方程。奖励函数 R: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 根据状态误差和控制 effort 量化控制器性能，而折扣因子 $\gamma \in [0, 1]$ 则权衡即时奖励与未来奖励。控制目标是找到一个策略 $\pi : \mathcal{S} \to \mathcal{A}$，使系统从任意初始条件到达目标状态，同时降低成本并满足约束。

B. 系统动力学概述

评估系统的选择遵循审慎的递进安排，旨在验证该框架在日益严峻的控制体制下的表现。该递进过程从解析解已充分明确的系统开始，以便直接与最优基准进行比较；随后逐步转向非线性、耦合和多维控制带来复杂性的问题，这些问题具有真实航空航天应用的代表性。

该方法在四种具有代表性的动力系统上进行评估：

双积分器作为基础基准，是一种表示平动的基本线性系统：

$$\ddot{x} = u \tag{1}$$

状态：$s = [x, \dot{x}]^T$。该系统的动力学较为简单，但需要谨慎控制以避免超调。

超越线性动力学，**范德波尔振荡器**引入了非线性极限环行为：

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = u \tag{2}$$

状态：$s = [x, \dot{x}]^T$。在不受控制时，非线性阻尼项会产生自持振荡，使稳定控制颇具挑战。

转向航空航天领域的特定挑战，**轨道提升问题**描述了一种具有非线性引力和推力动力学的航天器轨道转移场景：

$$\begin{aligned} \dot{r} &= u \\ \dot{u} &= \frac{v^2}{r} - \frac{\mu}{r^2} + \frac{T \sin(\phi)}{m(t)} \\ \dot{v} &= -\frac{uv}{r} + \frac{T \cos(\phi)}{m(t)} \end{aligned} \tag{3}$$

状态：$s = [r, u, v]^T$ 表示轨道半径、径向速度和切向速度。控制量：$\phi \in [0, 2\pi]$ 为推力角。随时间变化的质量 $m(t) = m_0 + m_1 t$（因燃料消耗而变化）以及耦合的引力-离心力动力学，构成了一个具有挑战性的最优控制问题，需要精确的推力矢量控制，才能在更高的高度 [18, 13]实现圆轨道捕获。

航天器消旋体现了复杂性逐步提升的最终阶段，它将非线性动力学与通过三轴旋转动力学实现的多维控制结合起来：

$$J\dot{\omega} = -\omega \times (J\omega) + u \tag{4}$$

其中，$\omega = [\omega_1, \omega_2, \omega_3]^T \in \mathbb{R}^3$ 表示角速度，$J = \text{diag}([J_1, J_2, J_3])$ 是惯性矩阵，而 $u =$

$[u_1, u_2, u_3]^T \in \mathbb{R}^3$ 是控制力矩向量。叉积项会在各轴之间引入强烈的非线性耦合 [19]。

这些系统涵盖了一系列控制挑战：线性与非线性动力学、标量与多维控制、解耦与强耦合状态变量，以及时不变与时变参数。从双积分器到轨道抬升，再到去翻滚，问题复杂度逐步提高，同时保持物理真实性。详细的系统参数、约束和边界见第四节。

C. 基于大语言模型的控制范式

该控制问题被重新表述为适用于大语言模型的条件文本生成任务。其关键洞见在于，控制问题可以自然地用结构化自然语言表达，其中包含物理背景、约束和目标；大语言模型能够凭借预训练获得的物理与数学理解来处理这些信息。

给定当前系统状态 $s_t \in \mathcal{S}$，大语言模型会接收描述控制场景的文本提示，并生成包含以下两个组成部分的结构化响应：

1) 推理： 以自然语言分析控制问题和系统动力学，并说明所选策略的依据 **2)控制序列：** 以结构化文本格式呈现的数值控制动作$\{a_t, a_{t+1}, \ldots, a_{t+N-1}\}$

该公式使策略 π_θ 能够由参数为 θ的大语言模型进行参数化，其中：

$$\pi_\theta : \text{State Description} \rightarrow (\text{Reasoning}, \text{Control Sequence}) \tag{5}$$

文本状态描述提供了全面的上下文信息，包括系统类型、带单位的当前状态数值、目标状态、时间约束和控制界限。具体的编码方案和提示模板详见第三节-C。

这一范式为自主控制带来了若干显著优势。自然语言推理使人们能够理解和验证控制决策；在安全关键型航空航天应用中，这一能力至关重要，因为操作人员必须信任并审查自主行为。无需重新设计特定于系统的控制器，同一大语言模型架构即可针对不同控制问题进行微调，从而实现跨问题领域的迁移。此外，大语言模型能够利用其对物理学和数学的预训练理解来生成天然符合物理约束的控制策略；与从零开始训练神经网络控制器相比，预训练表示还可能使模型通过更少的环境交互完成学习。

然而，这种方法带来了独特的挑战：确保文本生成模型输出精确的数值，在强化学习过程中保持输出格式一致性，并防止生成物理上不可行的控制序列。两阶段训练方法通过先进行监督微调，再进行 GRPO优化，来应对这些挑战。

D. 基于*GRPO*的强化学习目标

传统强化学习旨在寻找由 θ 参数化的策略$\pi_\theta(a|s)$，以最大化期望累积折扣奖励：

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right] \tag{6}$$

其中，$\tau = (s_0,\ a_0,\ s_1,\ a_1,\ \ldots\ldots,\ s_T,\ a_T)$ 表示按照策略 π_θ 采样得到的轨迹。

对于基于大语言模型的控制，本文采用组相对策略优化（GRPO）。该方法无需单独的价值函数网络，因此在微调大型语言模型时可显著降低内存需求。GRPO目标函数通过基于组的优势估计来优化策略：

$$L_{\text{GRPO}}(\theta) = -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=1}^{N}\min\Big(r_{ij}(\theta)A_{ij},\ \text{clip}\big(r_{ij}(\theta), 1-\epsilon, 1+\epsilon\big)A_{ij}\Big) + c \cdot L_{\text{KL}}(\theta \mid \theta_{\text{old}}) \tag{7}$$

其中：

- $r_{ij}(\theta) = \frac{\pi_\theta(a_{ij}|s_i)}{\pi_{\theta_{\text{old}}}(a_{ij}|s_i)}$ 是当前策略与用于生成响应的策略之间的重要性采样比率。

- $A_{ij} = R(s_i,\ a_{ij}) - R_i$ 是组相对优势，其中 $\bar{R}_i = \frac{1}{N}\sum_{k=1}^{N} R(s_i,\ a_{ik})$ 是状态 s_i下 N 个候选控制序列的平均奖励。

- $\text{clip}(r_{ij}(\theta)\ 1-\epsilon,\ 1+\epsilon)$ 对策略比率进行约束，以防止更新幅度过大而造成破坏（通常为 $\epsilon = 0.2$）。
- $L_{\text{KL}}(\theta|\theta_{\text{旧}})$ 是一种 KL 散度惩罚项，用于正则化策略更新（系数为 $c = 0.01$）。

GRPO 的关键创新是组相对优势A_{ij}，它将每个候选控制序列与针对同一状态生成的多个候选序列的平均表现进行比较。这样既无需单独的价值函数近似器，又能提供稳定的学习信号。负号表示，最小化此损失即可最大化策略梯度目标函数。

受近端策略优化 [15], 启发的裁剪机制通过防止策略发生剧烈变化来确保训练稳定性，而 KL 惩罚项则使策略保持接近先前策略。这种形式尤其适用于大语言模型微调，因为内存约束使得采用独立价值网络的 Actor-Critic 方法并不实用。

E. 控制目标与性能指标 *cs*

对于每个控制问题，目标是在 s_0 内将系统从任意初始状态驱动至目标状态

（通常为原点），同时在固定时间范围 T内最小化二次代价函数：

$$J = \|s_T\|_{Q_f}^2 + \sum_{t=0}^{T-1} \left(\|s_t\|_Q^2 + \|a_t\|_R^2 \right) \tag{8}$$

其中， Q_f， $Q \succeq 0$为状态代价矩阵，用于惩罚偏离目标的情况； $R \succ 0$ 为控制代价矩阵，用于惩罚过大的控制 effort。此外，必须满足以下约束：

- **控制界限：** $a_t \in \mathcal{A}$ 对于所有 t均成立，以限制执行器使用量
- **状态约束：** $s_t \in \mathcal{S}$ 对于所有 t均成立，以确保安全运行

控制器性能通过一系列指标进行评估，这些指标同时反映终端精度和轨迹质量。最终状态误差 $\|s_T\|$ 衡量控制器达到目标状态的精确程度，而式 8 中的总轨迹代价 J 则量化机动过程中产生的累计惩罚。约束满足率表示在未发生违规的情况下完成的轨迹百分比，这对于安全关键型应用至关重要。控制 effort $\sum_{t=0}^{T-1} \|a_t\|^2$ 量化执行器使用量，与燃料受限的航天器运行相关；收敛质量则评估整个轨迹中的状态误差降低速率，以区分平稳收敛的控制器与表现出振荡行为的控制器。

这些指标可全面评估大语言模型控制器相对于最优控制基线的表现 [9], ，同时考察控制性能和约束满足情况。各系统的具体代价矩阵取值和约束界限见第四节。

III. 方法

在建立了基于大语言模型并采用 GRPO 优化的控制理论框架后，本节将详细介绍其实际实现方式。下文将说明系统架构、模型选择、训练流程以及奖励函数设计，这些内容将第二节中的形式化描述转化为可运行的控制系统。

A. 系统架构

如图1所示，所提出的系统将大语言模型与两阶段训练方法相结合，用于动态系统的自主控制。其核心是一个大语言模型控制器，接收以文本形式输入的系统状态，并生成附带推理过程的控制动作序列。在输入大语言模型之前，状态编码器会将数值状态向量转换为格式化的文本提示，同时保留其物理含义。生成完成后，动作解码器使用正则表达式从大语言模型的文本响应中提取数值控制序列，并通过可靠的错误处理机制应对格式错误的输出。

控制器与模拟各动力系统的仿真环境交互，通过数值积分提供状态转移，并根据轨迹质量计算奖励。训练分为两个阶段：首先通过专家演示进行监督微调，以建立恰当的输出格式和基本控制直觉；然后通过环境交互进行 GRPO 强化学习，进一步优化策略。这种架构实现了从状态观测到控制输出的端到端学习，同时借助自然语言推理保持可解释性。

B. 基础模型选择与参数高效微调

选择 Qwen3-4B-Base [2] 作为基础模型，因为它兼顾了 40 亿参数规模与计算效率。通过结合监督微调和 GRPO 的两阶段训练方法，训练模型生成结构化控制输出，其中包含对控制策略的自然语言解释。虽然模型会在 <推理>部分生成类推理解释，但这些内容是通过训练过程习得的，而非涌现出的推理能力。

1) 基于 *LoRA* 的微调：本研究没有对模型的所有参数进行微调，而是采用低秩适配（LoRA） [7], 这一种参数高效微调技术：在冻结预训练权重的同时，将可训练的低秩分解矩阵引入 Transformer 层：$W' = W_0 + \Delta W = W_0 + BA$, 其中 $W_0 \in \mathbb{R}^{d \times k}$ 表示冻结的预训练权重，而 $B \in \mathbb{R}^{d \times r}$、$A \in \mathbb{R}^{r \times k}$ 是秩为 $r \ll \min(d,\ k)$ 的可训练低秩矩阵。

这种参数高效方法为控制应用带来了关键优势。LoRA 可将可训练参数减少 99% 以上（秩为 $r = 8$ 时，从 4B 个参数降至 40M 个参数），使模型能够在标准 GPU 硬件上进行微调，仅需 12GB GPU 显存，而全量微调则需要 80GB+ 。训练收敛速度显著加快，每个系统的监督微调（SFT）仅需 2 至 3 小时即可完成，而全量微调则需要数天。LoRA 适配器（通常为 100 至 200MB）可与基础模型分开保存，从而能够针对不同系统高效切换多个专用控制适配器。通过冻结基础模型权重，LoRA 在添加特定任务控制知识的同时，保留了大语言模型的通用推理能力，防止灾难性遗忘。

所有实验均采用秩为 $r = 8$ 的 LoRA，并将其应用于所有注意力模块中的查询和值投影层；LoRA alpha 参数设为 16，丢弃率设为 0.05。此配置使基础模型中可训练参数的比例低于 1%，大幅降低了显存需求，同时仍能学习特定任务的控制策略。对于结构化控制任务而言，较低的秩已足够，因为适配过程主要涉及学习数值输出模式和约束处理，而非获取全新的语言能力。

C. 状态与动作表示

1) 状态编码方案：数值状态向量 $s_t \in \mathcal{S}$被转换为结构化的自然语言提示，为控制决策提供全面的上下文信息。每个提示包含五个关键组成部分：系统规范，用于确定控制问题类型；当前状态信息，呈现带有适当单位的数值；

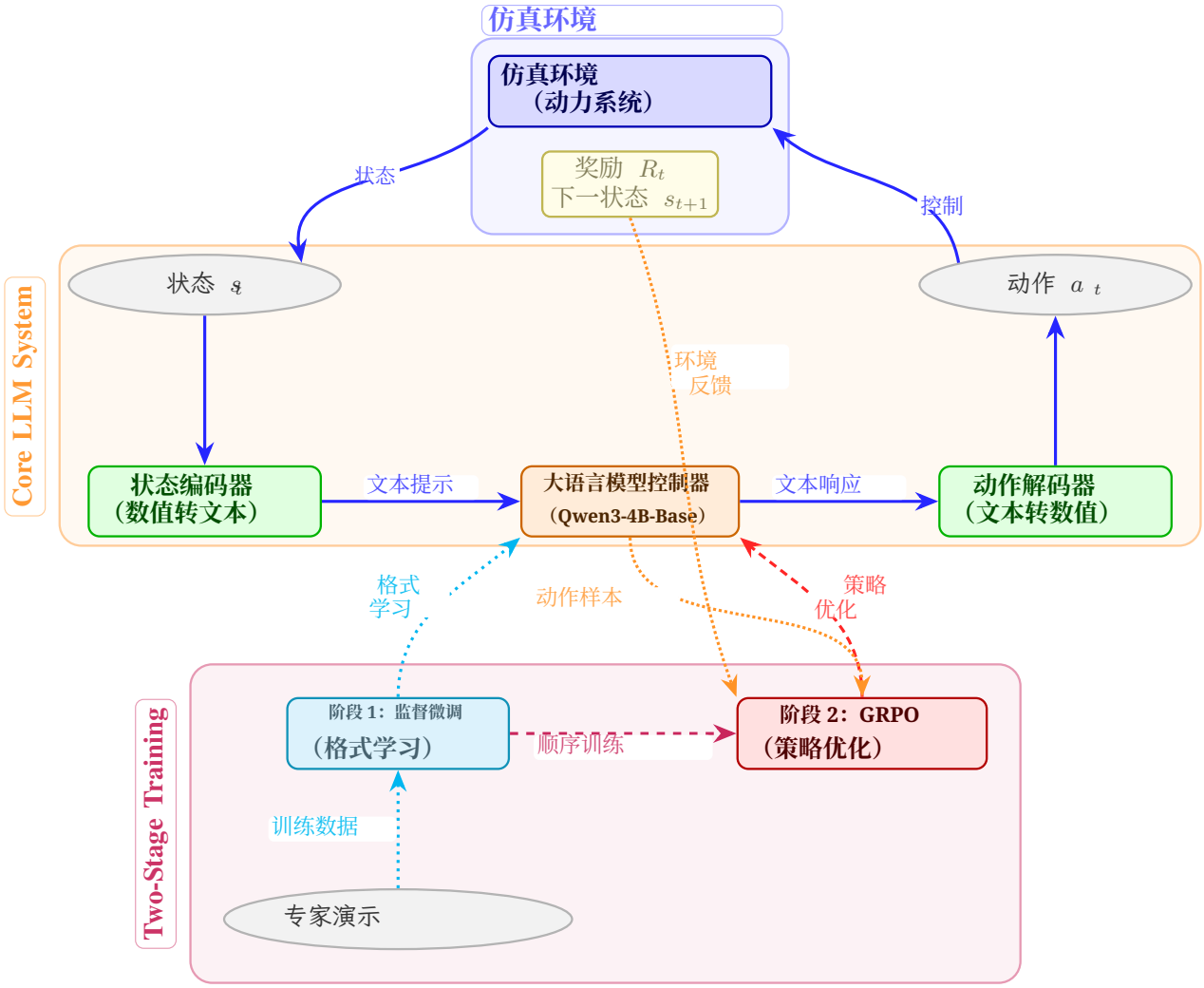


图1. 采用两阶段训练的基于大语言模型的控制增强系统架构。 **仿真环境** 提供动态系统上下文和反馈。 **核心大语言模型系统** 通过状态编码、大语言模型处理和动作解码来执行运行时推理。 **两阶段训练系统** 依次进行监督微调（SFT）以学习格式，随后进行组相对策略优化（GRPO）以优化策略。

目标状态规范，用于定义期望的最终状态；时间约束，用于规定时间范围和控制步数；以及控制规范，用于列举执行器界限和状态约束。这种结构化编码确保大语言模型能够接收全面的上下文信息，同时保持标准化格式，从而促进其在多种控制任务中的学习。重要的是，该编码通过使用“角速度”等描述性术语而非抽象标签来保留物理含义，使模型能够利用其预训练获得的物理概念理解。

2）动作解码方案：大语言模型生成的响应包含两个主要部分： <推理> ，其中以自然语言解释控制策略；以及<控制> ，其中包含数值控制序列，标量控制采用逗号分隔值格式，多维控制则采用方括号括起的向量格式。系统通过正则表达式进行稳健解析，提取数值并自动施加约束，将数值裁剪至执行器上下限范围内，验证序列长度是否正确，并通过回退机制处理解析失败：推理时应用安全的默认控制量，训练时则拒绝带有负奖励惩罚的样本。

D. 提示与响应格式

大语言模型以结构化自然语言格式接收控制问题，并生成同时包含推理内容和数值控制序列的响应。为说明这种格式，以下以航天器姿态去旋场景为例：

系统提示：

提示模板

你是一位航天器控制系统专家。

给定一个航天器消旋机动，其初始角速度为 [omega_1、omega_2、omega_3]，生成一组由 10 个三维力矩向量组成的序列，使

航天器恰好在 5.00 秒内静止（omega = [0,0,0]）。动力学：omega_dot = -J^(-1)(omega x J*omega)+J^(-1)*u，其中 J =diag([J1, J2, J3])。约束：|omega_i| <= omega_max rad/s，|u_i| <= u_max N*m。在 <REASONING> 与</REASONING>之间说明你的方法。然后在 <CONTROLS> 与</CONTROLS>之间提供恰好 10 个以逗号分隔的力矩向量。每个力矩向量的格式为 [u1, u2, u3]，保留 3 位小数。

用户提示（问题实例）：

提示模板

使用初始角速度 [omega_1=0.350、omega_2=-0.520、omega_3=0.180]rad/s，控制航天器消旋机动，使航天器在 5.00 秒内静止（omega =[0,0,0]），共使用 10 个步骤。惯性矩阵：J = diag([14.0, 10.0, 8.0]) kg*m^2。角速度保持在 [-1, 1] rad/s 范围内，转矩保持在 [-4, 4] N*m 范围内。

预期的大语言模型响应：

大语言模型响应

<推理>针对这一初始角速度为 [omega_1=0.350、omega_2=-0.520、omega_3=0.180] rad /s 的航天器消旋机动，我将采用边值问题最优控制，使航天器在 5.00 秒内静止。

航天器动力学遵循欧拉转动方程：omega_dot = -J^(-1)(omega x J*omega) + J^(-1)*u，其中惯性矩阵 J = diag([14.0, 10.0, 8.0]) kg*m^2。

分析：
- 初始角动量大小：0.649 rad /s - 主要翻滚轴：Y (omega_2) - 耦合常数：K_1=-0.143, K_2=0.600, K_3=-0.250_

策略：最优
- 在 10 个步骤中施加 3D 力矩序列 - 每步时长：0.5s - 目标：零角速度（消旋状态）- 约束：|omega_i| <= 1 rad/s，|u_i| <=4.0 N*m

该方法利用各轴之间的非线性耦合，同时尽量降低控制 effort 并遵守物理约束。</推理>

<控制>
[-1.245, 2.187, -0.658]-1.089, 1.923, -0.542]-0.876, 1.534, -0.398]-0.654, 1.142, -0.267]-0.445, 0.782, -0.154]-0.267, 0.478, -0.068]-0.134, 0.245, -0.012]-0.045, 0.089, 0.021]-0.008, 0.019, 0.012]0.000, 0.001, 0.002]</控制>

这种结构化格式使大语言模型能够生成对控制策略的可解释推理，以及精确的数值控制序列。 <REASONING>部分展现模型的决策过程，提高其透明度；<CONTROLS> 部分则包含可供执行的控制输入。

E. 专家示范生成

高质量的专家演示是监督微调的基础。每个控制系统都采用合适的最优控制方法 [12]：线性系统（双积分器）采用线性二次调节器，非线性系统（范德波尔）采用序列二次规划，旋转动力学系统（航天器）则采用基于庞特里亚金极大值原理的边值问题求解器。

数据生成过程在状态空间边界内（$s_0 \sim \mathcal{U}(s_{\min}, s_{\max})$）为每个系统均匀采样 2000 个初始状态，并将其划分为 1800 个训练样本和 200 个评估样本。系统专用求解器计算使矩阵 $Q = I$、$R = 0.1 \cdot I$、$Q_f = 10 \cdot I$ 对应的二次代价最小化的最优控制序列。为增强泛化能力，数据集纳入多种策略：最优解（40%）、优先考虑能效或时间最优性的替代策略（30%）、次优但实用的方法（20%），以及针对困难情况的恢复场景（10%）。每个控制序列均使用适当的数值积分方法（线性系统采用欧拉法，范德波尔刚性非线性动力学采用四阶龙格－库塔法）通过真实系统动力学进行仿真，以验证约束满足情况；同时为轨迹标注性能指标，包括代价、最终状态误差、控制 effort、平滑度和约束违反情况。

推理文本并非使用固定模板，而是通过分析特定初始状态的特征动态生成，包括与目标的距离、速度方向、系统特定挑战、明确权衡取舍的策略依据，以及预测的轨迹行为。例如，对于高速状态，推理会强调采取强力制动以防止超调；对于低能量状态，则会强调以最小努力平缓收敛。这种依赖状态的方法确保模型真正理解控制原理，而非记忆固定模式。最终数据集包含每个系统 1800 个训练样本和 200 个评估样本，推理文本平均长度为 150 至 250 词，控制序列为 10 步。

F. 两阶段训练方法

该方法采用两阶段方案，将格式学习与控制优化分开，确保模型先学会正确的输出结构，再通过环境交互进行策略优化。

1) 阶段 *1*：监督微调（*SFT*）：第一阶段通过专家演示，使模型掌握控制问题格式。SFT 阶段训练大语言模型，以最小化专家演示的负对数似然：$\mathcal{L}_{\text{SFT}} = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta(y_i|x_i)$，其中 x_i 是输入提示，y_i 是专家响应。训练目标包括：生成格式规范的响应，并保持<推理> 和 <控制> 部分的一致性；从最优示例中培养基本的控制直觉；确保控制输出具备数值精度；并学会解析自然语言形式的控制规范。

训练采用 AdamW 优化器（8 位），学习率为 2×10^{-4}，先进行 50 步线性预热，再采用余弦衰减；批次大小为 4 个序列，训练 500 步，序列长度最长为 1024 个 token，并将梯度裁剪的最大范数设为 1.0。在留出评估集上评估模型，以验证格式合规率（目标：> 95 %）、解析成功率、通过比较 LQR 代价评估的基本控制性能，以及评估轨迹上的约束满足情况。只有达到格式合规性和基本性能阈值的模型才会进入 GRPO训练。

2）第二阶段：组相对策略优化（*GRPO*）：第二阶段通过与环境交互并基于奖励进行学习，来优化控制策略，并实现式（7）中定义的 GRPO 目标函数。其目标是利用轨迹仿真产生的奖励信号优化控制性能，保持监督微调阶段确立的输出格式一致性，探索可能优于初始示范的策略，并平衡探索与利用。

训练采用 AdamW 优化器，学习率为 3×10^{-6}，使用 100 步预热的余弦退火，每步批次大小为 1 个状态；针对每个状态生成 $N = 4$ 个候选控制序列，温度设为 1.0，并采用 min-p 筛选（0.1），共训练 500 步。裁剪参数为 $\epsilon = 0.2$，KL 惩罚系数为 $c = 0.01$。每个训练步骤执行以下流程：从训练分布中采样初始状态；使用当前策略通过温度采样生成 $N = 4$ 个候选控制序列；利用 RK45 积分，根据系统动力学仿真每个候选序列；计算每个候选序列的轨迹奖励；计算组平均奖励 $\bar{R}_i = \frac{1}{N}\sum_{j=1}^{N} R(s_i, a_{ij})$ 和优势值 $A_{ij} = R(s_i, a_{ij}) - \bar{R}_i$；计算重要性采样比 $r_{ij}(\theta) = \frac{\pi_\theta(a_{ij}|s_i)}{\pi_{\theta_{\text{old}}}(a_{ij}|s_i)}$；根据式（7）计算 GRPO 损失；并通过梯度下降更新策略参数，同时监控格式合规性。

基于组的优势估计通过将每个候选项与平均性能进行比较来提供稳定的学习信号，从而无需单独的价值函数。裁剪机制可防止破坏性的策略更新通过将 $r_{ij}(\theta)$ 约束为 $[1-\epsilon, 1+\epsilon]$。

G. 奖励函数设计

奖励函数通过加权求和来平衡对有效控制至关重要的多个目标：$R(s, a) = w_1 R_{\text{LQR}} + w_2 R_{\text{terminal}} + w_3 R_{\text{constraint}} + w_4 R_{\text{format}} + w_5 R_{\text{auxiliary}}$。主要性能指标是 LQR 代价 $R_{\text{LQR}} = -\sum_{t=0}^{T} s_t^T Q s_t + u_t^T R u_t$，其代价矩阵为 $Q = I$，$R = 0.1 \cdot I$，$Q_f = 10 \cdot I$。根据最终状态误差$\|s_T\|$，通过指数奖励和分级奖励来奖励终端精度。通过违规惩罚和有效性奖励来确保约束满足。格式合规性奖励确保输出结构的一致性，并对格式违规施加惩罚。辅助组件包括控制持续性惩罚，用于抑制接近目标时过度控制；以及收敛奖励，用于奖励误差的单调下降。

各组件权重遵循课程学习调度：GRPO训练早期（步骤 0-200）采用较高的格式奖励权重和较低的控制权重；训练中期（步骤 200-400）采用均衡权重；训练后期（步骤 400-500）采用较低的格式权重和较高的控制权重。这一过程逐步将重点从维持输出结构转向优化控制性能，同时防止格式崩溃。每个系统的具体权重值和奖励范围见第四节。

IV. 实验结果

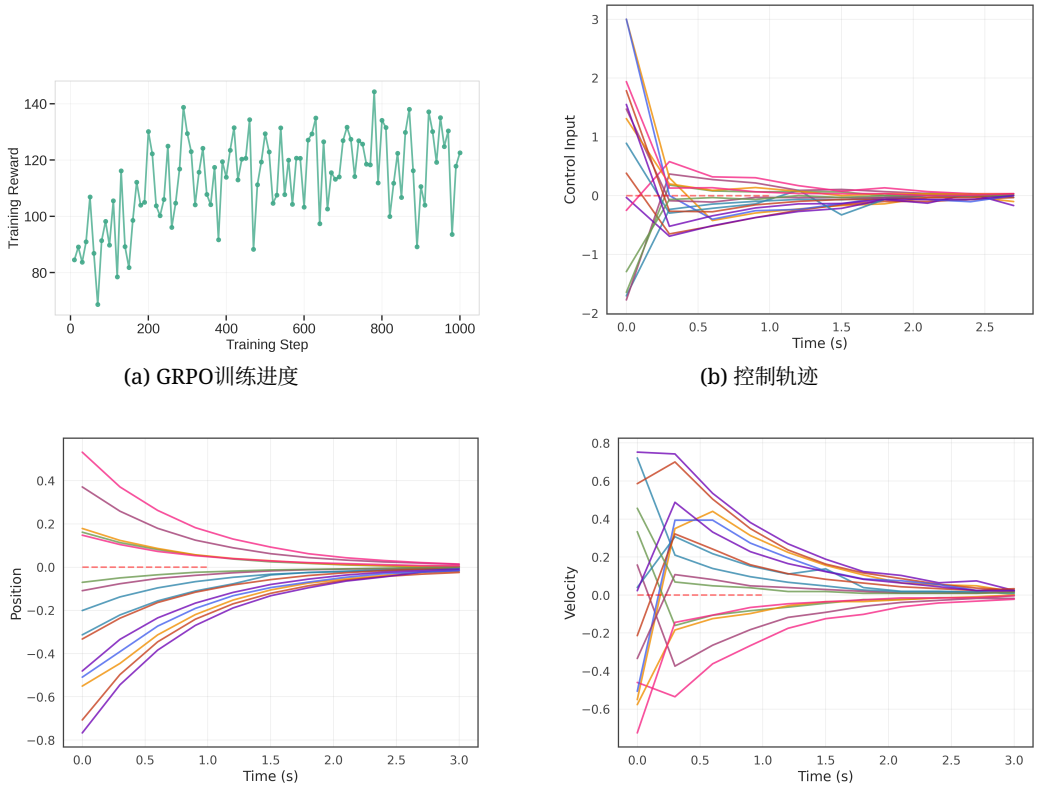
在确立训练方法后，本节评估该框架在第二节-B所介绍的四个控制系统中的有效性，研究过程从线性动力学逐步扩展到日益复杂的航空航天应用。对于每个系统，本文展示GRPO训练过程，分析控制性能，并给出示例推理输出，以说明模型的决策过程。

表 I 和表 II 总结了实验配置。该框架采用 Unsloth 进行高效节省显存的 LoRA微调，并使用 TRL库 实现监督微调和 GRPO。

A. 双积分器系统

评估首先从双积分器开始，该系统作为一个典型的线性控制问题，用于验证LLM-GRPO框架的基本能力，然后再推进到更具挑战性的非线性系统。图2a展示了训练进程和控制性能。图2a展示了1000步训练中的GRPO训练收敛情况，奖励呈现出明显的上升趋势：从接近于零开始，在200步后达到约10—12，此后持续保持稳定性能。初始100步后，方差显著降低，表明尽管随机优化过程带来了轻微波动，策略仍在稳定改进。

图 2b–2d 中的控制和状态轨迹展示了在多个评估



(a) GRPO训练进度　(b) 控制轨迹

(c) 位置轨迹　(d) 速度轨迹

图2. 双积分器控制性能：(a) GRPO训练进度，展示奖励收敛；(b) 控制输入轨迹；(c) 位置状态演化；(d) 速度状态演化。

表I 实验参数

类别	参数	值
模型配置		
	基础模型	Qwen3-4B-Base
	微调方法	LoRA（秩 8）
	可训练参数	∼40M (< 1%)
数据集		
	每个系统的训练样本	1800
	每个系统的评估样本	200
监督微调训练		
	学习率	2×10^{-4}
	批次大小	4
*GRPO*训练		
	学习率	1×10^{-6}
	批次大小	1
	每个状态的候选项	8
	温度	1.0
	裁剪（ϵ）	0.2
	KL 系数（c）	0.05

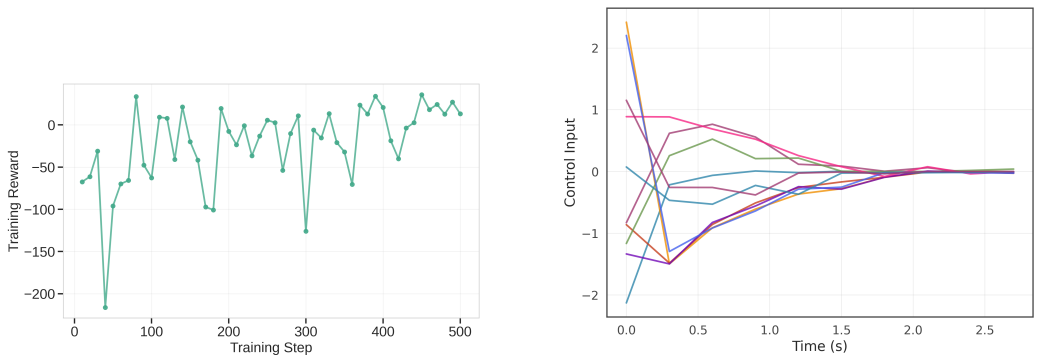
表 II 系统特定参数和约束

系统	类别	数值
双重积分器	状态：$x, \dot{x}$	$[-1, 1]$
	控制：u	$[-3, 3]$
	训练：监督微调/GRPO	1000/1000 步
范德波尔	状态：x，$\dot{x}$	$[-1, 1]$
	控制：u	$[-3, 3]$
	参数：μ	1.0
	训练：监督微调/GRPO	1000/500 步
轨道抬升	状态：r，u，v	$[0.5, 2.0]$, $[-0.5, 0.5]$, $[0.5, 1.5]$
	控制：ϕ	$[0, 2\pi]$ 弧度
	参数：μ，T	1.0，0.1405 牛顿
	质量：$m\ (t)$	$m_0 = 1.0$，$m_1 = -0.075$ 千克/秒
	训练：监督微调/GRPO	1000/1000 步
航天器去翻滚	状态：$\omega_i\ (i = 1,2,3)$	$[-1.0, 1.0]$ 弧度/秒
	控制：$u_i\ (i = 1,\ 2,\ 3)$	$[-4, 4]$ N·m
	惯性：J	diag$([14, 10, 8])$ 千克·米^2
	训练：监督微调/GRPO	1000/3000 步

回合中实现的有效稳定。控制输入在初始阶段出现瞬态响应，幅值最高达 ± 2 N；随后在前 2 秒内呈现调制振荡，之后逐渐趋近于零，且始终保持在 $[-3, 3]$ N 的限值范围内。位置和速度状态从不同的初始条件（位置为 $[-0.4, 0.3]$，速度为 $[-0.4, 0.4]$ m/s）收敛至原点，并在 3–4 秒，呈现出单调或近似单调的良好阻尼特性，这是最优线性二次调节器型策略的典型表现。

B. 范德波尔振荡器

在线性动力学上验证了该框架后，分析现在考察它能否处理非线性系统。范德波尔振荡器引入了极限环动力学，其中系统会自然维持振荡，需要通过主动阻尼来实现稳



(a) GRPO训练进度 (b) 控制轨迹

(c) 位置轨迹 (d) 速度轨迹

图 3. 范德波尔振荡器控制性能：(a) 非线性动力学下的 GRPO 训练进程，(b) 用于极限环抑制的控制输入轨迹，(c) 位置状态演化，以及 (d) 速度状态演化。

定。图 3 展示了训练和控制性能。图 3a 显示，训练奖励起始值接近 -50，在约第 200 步时显著提升，达到接近零的水平，并逐渐优化，到第 600 步时略微转为正值。与双积分器相比，训练过程的方差更大，在前 400 步中偶尔会低于 -100，这反映了非线性动力学和极限环抑制带来的额外复杂性。

图 3b–3d 中的控制轨迹和状态轨迹展现出比线性情形更复杂的行为。控制输入幅度接近 $[-2.5, 2.5]$ N，呈非单调振荡曲线；在 2 秒后仍持续施加控制，随后趋近于零，且始终满足 $[-3, 3]$ N 的约束。来自不同初始条件的位置和速度状态在 4–6 秒内收敛，并表现出瞬态振荡以及高达 ± 0.6m/s 的速度峰值，这是极限环抑制的典型特征。收敛路径的差异凸显了非线性动力学对状态的依赖性。

C. 升轨机动

该框架已在学术领域的线性和非线性控制问题上得到验证，现将其应用于航空航天相关场景。轨道提升问题带来了先前系统中未曾涉及的领域特定挑战：

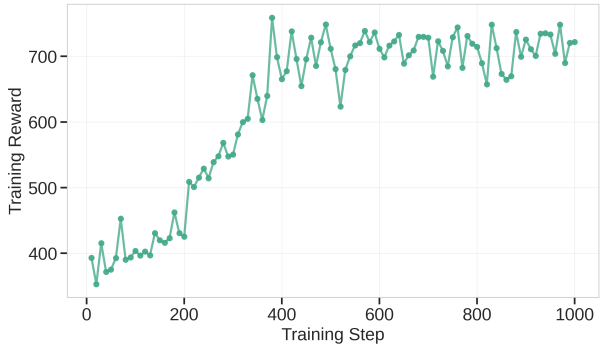


图 4. 轨道转移控制的升轨 GRPO 训练进展及奖励收敛情况。

由开普勒动力学支配的非线性轨道力学、需要进行连续角度优化的推力矢量控制，以及推进剂消耗导致的时变质量。控制器必须在固定推力大小为 $T = 0.1405$ N 的条件下，处理连续推力角输入（$\phi \in [0, 2\pi]$ 弧度），使航天器从初始轨道转移至目标半径，同时遵循轨道动力学。图 4 和图 5 展示了训练和控制结果。

图 4 展示了轨道转移控制的 GRPO 训练奖励变化过程。训练呈现出明显的几个阶段：

初始探索阶段的奖励接近零（步骤 0-200），随后奖励迅速提升至约 300（步骤 200-400），并在步骤 400-800 期间进入高方差的精细调整阶段。奖励峰值达到 400-500，表明成功完成了轨道转移；探索过程中偶尔出现的奖励下降则说明尝试失败。步骤 600 之后持续保持高奖励的平台期且方差受控，这表明策略学习趋于收敛。与先前系统相比，奖励的绝对值更高，反映出多组件奖励结构纳入了终端奖励和轨道转移目标。

图5a中的径向距离轨迹表明，在多种情景下均成功实现了轨道抬升。从接近 $r = 1.0$ 的初始半径出发，所有轨迹均平滑且单调地增大，在持续4秒的机动过程中，最终半径接近$r = 1.7 - 1.9$ 。轨迹紧密聚集，表明转移性能稳定；平滑的曲线则反映出连续低推力推进下高效的能量管理。没有出现振荡或反向变化，证实了轨道扩张过程稳定。

图5b展示了轨道转移过程中径向速度分量的变化。所有轨迹的径向速度初始值均接近零，在约 $u \approx 0.45$ –0.5 的峰值处达到最大值，时间约为 $t = 2.0$ s；随后，随着航天器接近目标轨道，速度逐渐降回零附近。这种典型的钟形曲线符合最优低推力轨道转移的特征：转移过程中需要逐渐增大径向速度，而接近最终轨道时则需将其降低。各轨迹的峰值速度在时间和幅值上同步，表明系统学会了协调推力矢量控制与轨道力学。

图5c中的切向速度轨迹起始于约 $v \approx 1.0$ ，并呈现出典型的先升后降模式。速度在约 $t = 1.0 - 1.5$ s时升至接近 $v \approx 1.1$的峰值，随后下降至$v \approx 0.65 - 0.75$的最终值。这一变化体现了轨道动力学的耦合特性：航天器转移至更高轨道时，随着径向距离增大，切向速度必须降低，以满足角动量守恒。切向速度平滑、协调的变化表明，控制器隐式地理解了轨道力学约束。

图5d展示了推力角控制序列，它们决定了推力的施加方向。接近$\phi \approx 0.5 - 1.0$ rad的初始推力角意味着推力主要沿切向施加，以增加轨道能量。在约 $t = 1.5 - 2.0$ s时，大多数轨迹都迅速转向更大的推力角（$\phi \approx 5$–5.5 rad，约为 285°–315°），表明策略性地转向径向或组合式推力矢量控制。一些轨迹在稳定于最终角度之前，会在 $\phi \approx 4.5$ rad附近进行中间调节。这些并非简单的推力矢量控制序列，与恒定角度策略大相径庭，展现了模型学会优化耦合的轨道动力学，在径向扩展与切向速度管理之间取得平衡。

D. 航天器消旋

本次评估中的最后一个系统也是最具挑战性的系统，它综合了此前所有难点，包括非线性动力学、多维控制，以及陀螺效应造成的状态变量之间的强耦合。

航天器消旋问题是本研究中最复杂的控制挑战，需要在非线性欧拉动力学和陀螺耦合条件下同时实现三轴旋转稳定。控制器必须协调三个独立的力矩输入（u_1，u_2，$u_3 \in [-4, 4]$N·m），以制止所有轴上的翻滚运动，同时满足航天器非对称惯性张量 J =diag([14, 10, 8]) kg·m^{2}的要求。图6和图7展示了训练过程和控制性能。

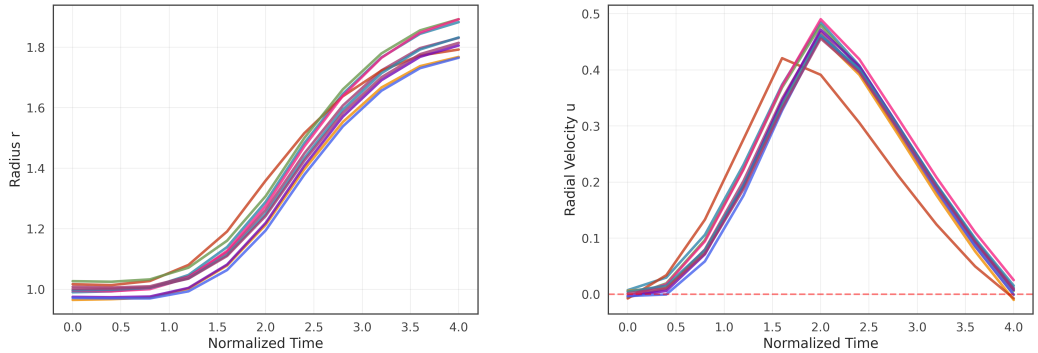
图6展示了去自旋任务中GRPO训练的演变过程。训练开始时，奖励值非常低，约为 -60 至 -80，表明模型最初未能稳定耦合旋转动力学。在第50至150步之间，性能显著提升，奖励值升至 -40以上。整个训练过程中波动较大，奖励值在 -80 至 $+20$之间起伏，即使在后期也偶尔会骤降至 -80 以下。尽管存在这种波动，总体仍呈上升趋势：后期训练步骤（400至500步）的基线性能有所改善，达到约 -20至0，偶尔还能取得 $+20$的成绩。持续的大幅波动反映了在非线性陀螺耦合条件下学习协调多轴控制的固有难度，因为微小的控制误差也可能导致行为发散。

图7a、图7b和图7c中的角速度轨迹表明，在多种初始条件下均实现了成功的多轴稳定。图7a展示了从 $[-1.0, 1.0]$ rad/s完整范围内起始的 ω_1 轨迹，这些轨迹均在约2.5至3.0秒内平稳收敛至零。轨迹呈单调或近似单调收敛，振荡极小，表明绕第一主轴的控制具有良好的阻尼。

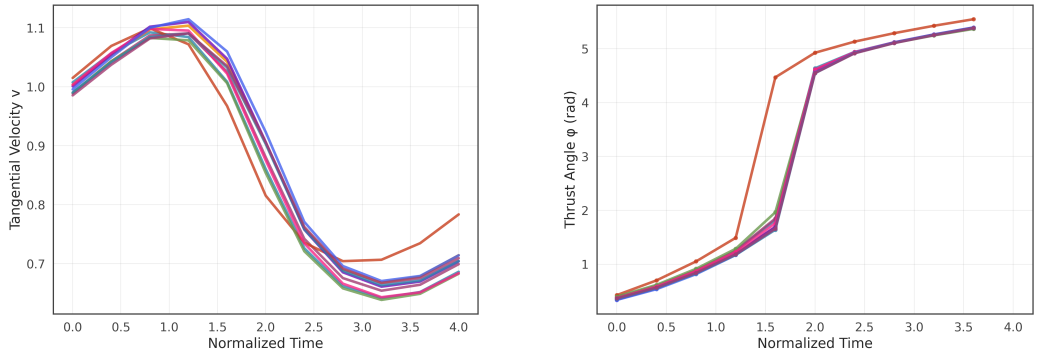
图 7b 展示了从跨度为 $[-1.0, 1.0]$ rad/s 的初始条件出发的 ω_2 演化。与第一轴相似，所有轨迹都在 2.5-3.0 秒内收敛至接近零的角速度。与 ω_1 相比，若干轨迹在收敛过程中呈现出稍大的曲率，这可能反映了中间惯性值（$J_2 = 10$kg·m^2）所产生的交叉耦合效应。

图 7c 展示了第三轴从 $[-1.0, 1.0]$ rad/s 之间的初始角速度开始的稳定过程。其收敛特性与其他轴相当，所有轨迹都在相同的 2.5-3.0 秒时间范围内达到接近零的角速度。三个轴协调一致地同时收敛，证实学习策略能够成功处理耦合的非线性动力学，而不是依次稳定各轴。

图7d、图7e和图7f中的控制力矩轨迹揭示了对陀螺耦合的精妙利用。图7d显示，u_1 控制输入主要处于 $[-2, 2]$ N·m范围内，大多数轨迹在整个机动过程中都保持相对恒定的力矩数值，而不是



(a) 径向距离轨迹　　(b) 径向速度轨迹



(c) 切向速度轨迹　　(d) 推力角轨迹

图 5. 轨道转移轨迹：(a) 径向距离演化，(b) 径向速度分量，(c) 切向速度分量，以及 (d) 推力角控制输入。

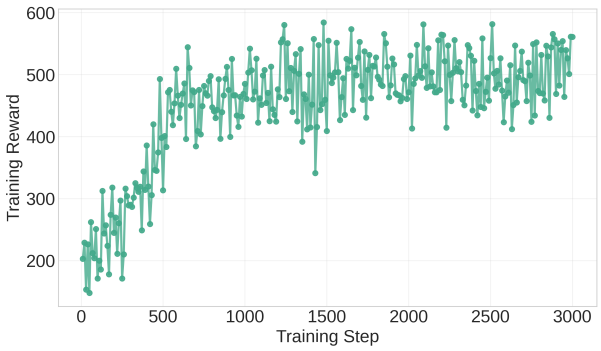


图 6. 航天器消旋 GRPO 训练进程，展示了存在陀螺耦合时三维旋转控制的奖励收敛情况。

平稳衰减至零。多条轨迹表现出约为 $+1.5$ 至 $+2.0$ N·m的近似恒定正转矩，而其他轨迹则在接近 -0.5 至 -2.0 N·m的范围内保持负转矩。这种持续的控制effort 模式与单轴系统或非耦合系统中观察到的衰减型控制明显不同。

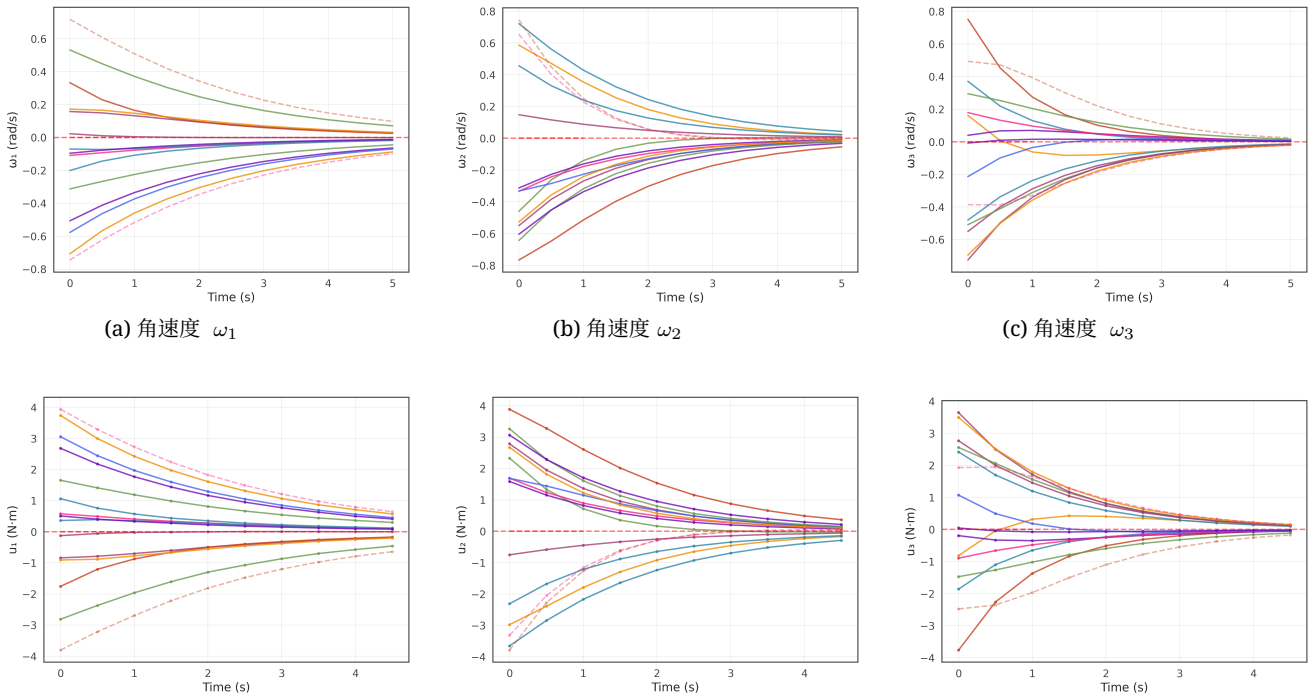
图7e展示了跨越约 $[-2, 2]$ N·m的 u_2 控制轨迹。与 u_1 类似，许多轨迹的扭矩数值相对稳定，并明显聚集在 $+1.0$ 至 $+1.5$N·m和 -0.5至 -1.5N·m附近。一些轨迹呈现出随时间逐渐漂移或轻微调制的情况，但主要表现为持续施加扭矩。

图 7f 显示， u_3 控制输入范围约为 $[-1.5, 1.5]$ N·m，比其他轴的范围略窄。控制模式的变化幅度大于u_1 和 u_2，有几条轨迹显示，在机动过程中，转矩水平逐渐转换。

三个轴上持续且不衰减的控制转矩模式，有力地表明学习策略利用了陀螺耦合效应。控制器并未施加会随着角速度降低而自然衰减的直接阻尼转矩，而是维持策略性的转矩组合，利用欧拉方程中的惯性积实现轴间的动量传递。这种复杂行为在所有角速度同时趋于零的同时，仍维持看似稳定的控制输入，表明 LLM-GRPO框架已隐式学会耦合旋转动力学的底层物理规律。在实现协调稳定的同时，所有控制输入均远低于 $[-4, 4]$N·m 约束。

E. 跨系统综合

在全部四个系统中，都呈现出一致的模式：两阶段训练方法通过监督微调成功建立控制能力，随后 GRPO 通过



(a) 角速度 ω_1 (b) 角速度 ω_2 (c) 角速度 ω_3

(d) 控制力矩 u_1 (e) 控制力矩 u_2 (f) 控制力矩 u_3

图 7. 航天器去翻滚性能：(a-c) 三轴角速度轨迹显示了协调的多轴稳定过程，(d-f) 控制力矩序列展示了对陀螺耦合效应的利用。

环境交互对其进行优化。该框架无需修改架构，即可从二维线性动力学扩展到三维非线性航天器控制，仅需进行特定系统奖励调优和提示工程。收敛时间保持在相近水平（稳定任务为 2–4 秒，轨道转移为 4 秒），这表明学习策略在不同复杂度水平下实现了相似的相对性能。值得注意的是，训练方差会随系统复杂度增加：从双积分器中的稳定收敛，到航天器消旋中的较高方差，反映出在非线性耦合条件下学习协调多轴控制的内在难度。

F. 推理输出分析

该基于大语言模型的控制框架的一项显著特点是，会为每个控制序列生成配套的人类可读推理。与仅产生数值输出的传统最优控制方法不同，该方法会对控制策略提供可解释说明，从而便于验证并建立对自主系统的信任。以下展示训练后模型生成的代表性推理输出，以说明其所学领域知识的深度。

航天器去自旋示例： 对于初始角速度为 $[\omega_1 = -0.507, \omega_2 = -0.313, \omega_3 = 0.040]$ rad/s 的航天器，经过 GRPO 训练的模型生成了：

大语言模型响应

<推理>对于这项航天器消旋机动，初始角速度为[omega_1=-0.507、omega_2=-0.313、omega_3=0.040]rad/s，我将采用边值问题最优控制，使航天器在 5.00 秒内静止。

航天器动力学遵循欧拉转动方程：omega_dot = -J^(-1)(omega x J*omega) + J^(-1)*u，其中惯性矩阵 J = diag([14.0, 10.0, 8.0]) kg*m^2。

分析：
- 初始动量大小：0.594 rad/s - 主要翻滚轴：X (omega_1) - 耦合常数：K_1=0.143, K_2=-0.600, K_3=0.500

策略：最优
- 在 10 个步骤中施加 3D 力矩序列 - 每步持续时间：0.5s - 目标：零角速度（消旋状态） - 约束：|omega_i| <= 0.8 rad/s, |u_i| <=4.0 N*m

这种方法利用轴间的非线性耦合，同时最大限度地减少控制 effort 并遵守物理约束。</推理>

<控制>[1.973, 1.133, -0.159]
[1.679, 0.897, -0.100]1.
432, 0.715, -0.058]1.228,
0.577, -0.030]...</控制>

此示例的最终状态误差为 0.006，成功在指定时间范围内使航天器完成去翻滚。

升轨示例：对于从初始状态 $[r = 1.008, u = -0.006, v = 0.989]$ 开始的轨道转移，模型展示了基于能量的推理：

大语言模型响应

<推理>基于能量的升轨分析 当前轨道能量：epsilon_0 = -0.5034 - 动能分量：(u^2+v^2)/2 =0.4887 - 势能分量：-mu/r = -0.9922 - 总比能量：-0.5034 目标轨道能量：epsilon_target = -mu/(2* r_target) = -0.3333 - 所需能量增量：Delta_epsilon =0.1701 能量传递策略：必须通过推力做功增加 0.1701 个单位的比能量 最优推力方向：- 切向推力分量：直接增加轨道能量（效率最高） - 径向推力分量：尽量减小，以保持轨道近似圆形 物理约束：- 当前半径处的重力加速度：g = mu/r^2 = 0.9845 - 离心效应：v^2/r = 0.9699 - 净加速度平衡决定推力需求 轨迹演化：在 4.0 秒内，10 次推力脉冲将逐步增加能量，同时保持轨道近似圆形，使轨道逐渐抬升。</推理><控制>0.395, 0.614, 0.912, 1.282, ...</控制>

推理输出展现了通过两阶段训练过程习得的几项关键能力：正确识别特定系统的物理规律（旋转动力学中的欧拉方程、航天器机动中的轨道力学），定量分析初始条件和目标状态，进行明确考量约束的策略规划，以及为控制方法提供基于物理原理的论证。值得注意的是，平均推理长度会随系统复杂度而增加（双积分器问题约为 620 个字符，而轨道抬升问题超过 1050 个字符），这表明模型会针对物理要求更高的问题投入更多比例的推理精力。

除可解释性之外，GRPO训练还带来了可量化的性能提升。双积分器的平均最终状态误差为 0.054，范德波尔振荡器为 0.039，航天器消旋的误差为 0.034 rad/s。这些量级为 10^{-2} 的误差表明，系统均成功收敛，涵盖从线性动力学到陀螺耦合作用下的非线性多轴稳定等各类系统。结果证实，通过 GRPO 进行强化学习优化，能够在保留监督微调阶段所建立的推理能力的同时，生成有效的控制策略。

V. 结论

本文提出了一种将大语言模型与组相对策略优化相结合的框架，以实现由推理驱动的自主控制。该两阶段方法先进行监督微调，再通过 GRPO 进行强化学习，结果表明，大语言模型能够生成有效的控制策略，同时提供便于人类理解的决策说明。针对复杂度逐步增加的四种系统所得结果验证了该方法：双积分器实现了快速的状态收敛；范德波尔振荡器成功抑制了极限环动力学；升轨机动展示了通过学习推力矢量控制实现轨道转移的能力；航天器消旋则在存在非线性陀螺耦合的情况下，实现了协调的三轴稳定。

生成的推理轨迹展现出对系统物理学的深刻理解，涵盖非线性振荡器中的极限环抑制，以及航天器姿态控制中的陀螺耦合利用。这种可解释性代表着一项重大进展，可应用于决策过程必须透明的安全关键型应用。

未来的工作将优先在航天器姿态控制试验台上进行硬件验证，并研究多系统迁移学习，使单个大语言模型能够控制不同系统，而无需分别进行微调。更长期的研究方向包括与模型预测控制相结合以实现实时重新规划，以及开发能够提供安全保证的形式化验证方法。本文建立的框架为将具备推理能力的大语言模型应用于自主控制奠定了基础，并为下一代航天器自主能力的发展带来了广阔前景。

致谢

本研究由美国空军部人工智能加速器资助，并依据合作协议编号FA8750-19-2-1000完成。本文所述观点和结论均为作者所有，不应被解读为代表美国空军部或美国政府明示或暗示的官方政策。美国政

府有权为政府用途复制和分发重印件，尽管本文中有任何版权声明。

参考文献

[1]AI Engineering Academy。《GRPO背后的理论》。*AIEngineering Academy*，2024年。网址：https://aiengineering.academy/LLM/TheoryBehindFinetuning/GRPO/.[2] Jinze Bai 等人。《通义千问技术报告》。*arXiv*预印本 *arXiv:2309.16609*，2024年。[3]Alejandro Carrasco、Victor Rodriguez-Fernandez 和 Richard Linares。用于自主航天器控制的大语言模型微调：使用坎巴拉太空计划的案例研究。*arXiv*预印本 *arXiv:2408.08676*，2024年。网址：https: //arxiv.org/abs/2408.08676.[4]DeepSeek AI、Guo, Dian、Zhu, Zhihong、Cheng, Wenhao、Zhang, Zhaoye、Yuan, Yibin、Li, Lei、Shen, Yichi、Kong, Lv、Wen, Yixin、Niu, Qingyan、Jiang, Xin 和 Zhu, Hao。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*，2025年。网址：https://arxiv.org/abs/arXiv:2501.12948.[5]DeepSeek AI、Zhu, Zhihong、Guo, Dian、Cheng, Wenhao、Fang, Zeming、Yuan, Yibin、Li, Lei、Shen, Yichi、Kong, Lv、Wen, Yixin、Zhang, Zhaoye、Jiang, Xin、Niu, Qingyan 和 Zhu, Hao。DeepSeekMath：推动开放语言模型数学推理的极限。*arXiv*预印本 *arXiv:2402.03300*，2024年。网址：https://arxiv.org/ abs/2402.03300.[6] Brian Gaudet、Richard Linares 和 Roberto Furfaro。采用强化元学习的自适应制导与综合导航。《航天学报》，169：180–190，2020年。[7] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等人。LoRA：大语言模型的低秩适配。*ICLR*，1(2)：3，2022年。[8] A mit Jain 和 Richard Linares。微型递归控制：用于高效最优控制的迭代推理。*arXiv*预印本 *arXiv:2512.16824*，2025年。[9] Amit Jain、Roshan Eapen 和 Puneet Singla。带终端约束的最优反馈控制的稀疏近似哈密顿-雅可比解。载于*2023*年第*62*届*IEEE*决策与控制会议（*CDC*），第1269–1274页。IEEE，2023年。[10]Amit Jain、Victor Rodriguez-Fernandez 和 Richard Linares。通过基于 Transformer 的强化学习进行多阶段航天器轨迹优化。*arXiv*预印本 *arXiv:2511.11402*，2025年。[11] Yeseung Kim 等人。大语言模型与智能机器人融合的综述。*arXiv*预印本 *arXiv:2404.09228*，2024年。网址：https://arxiv.org/abs/arXiv:2404. 09228.[12] Donald E Kirk。《最优控制理论：导论》。Dover Publications，2004年。[13] David Miller 和 Richard Linares。低推力最优控制：通过强化学习实现。载于第*29*届*AAS/AIAA*空间飞行力学会议，2019年。[14]Charles Oestreich、Richard Linares 和 Ravi Gondhalekar。通过强化学习实现自主六自由度航天器对接机动。《航空航天信息系统杂志》，18(7)：417–428，2021年。[15] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017年。[16] Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。麻省理工学院出版社，2018年。[17] Unsloth AI。在本地训练自己的 r1推理模型（GRPO）。Unsloth AI博客，2025年1月。网址：https://unsloth. ai/blog/r1-reasoning.[18] David A Vallado。《天体动力学基础与应用》。施普林格，2001年。[19] Bong Wie。《航天器动力学与控制》。美国航空航天学会，2008年。[20]Enrico Zucchelli 等人。微调语言模型作为空间系统控制器。*arXiv*预印本 *arXiv:2501.16588*，2025年。网址：https://arxiv.org/abs/arXiv:2501. 16588。