

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：基于强化学习的移动悬吊大时滞前瞻补偿方法
作者：屈冠行，王安良，吴坤
DOI：10.13700/j.bh.1001-5965.2026.0215
收稿日期：2026-04-18
网络首发日期：2026-09-01
引用格式：屈冠行，王安良，吴坤．基于强化学习的移动悬吊大时滞前瞻补偿方法[J/OL]．北京航空航天大学学报．https://doi.org/10.13700/j.bh.1001-5965.2026.0215



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。

基于强化学习的移动悬吊大时滞前瞻补偿方法

屈冠行[1,2]，王安良[3,✉]，吴坤[4]

（1. 北京航空航天大学 北航学院，北京 102206；2. 北京航空航天大学 机械及自动化学院，北京 102206； 3. 北京航空航天大学 宇航学院，北京 102206；4. 北京航空航天大学 飞行学院，北京 102206）

*通信作者 E-mail：wanganliang@buaa.edu.cn

摘要 针对大跨度移动跟随式微重力模拟系统中柔性索单向受力、多体非线性耦合及物理大时滞导致的传统控制失效难题，提出融合地形前瞻感知与近端策略优化算法的深度强化学习控制架构，设计物理优先级软门控奖励函数与历史状态堆叠机制，辅以“时滞-地形”二维课程学习策略实现前瞻预测式补偿。全速域数字孪生实验表明：该架构单步网络推理耗时缩减至 0.35 ms；在 20.83 ms 通讯时滞、连续陡坡以及高达 95%极端丢包率工况下，动态张力跟随误差严格约束在 ±15% 以内，任务存活率达 100%，综合性能显著优于传统改进型比例积分微分控制器。研究揭示了强化学习在变阻尼物理场中的非对称泛化特性，证明中高阻尼域训练的模型可无损迁移至低阻尼域。结合 12t 级轮式起重机原型进行极限功率核算，为大型航天器微重力模拟系统向物理实机的零样本跨域部署提供了理论依据。

关键词 微重力模拟；主动悬吊；深度强化学习；时滞前瞻补偿；非对称物理泛化；零样本迁移

中图分类号 V476.3；TP273.4 文献标志码 A DOI: 10.13700/j.bh.1001-5965.2026.0215

Time-delay preview control of mobile active suspension via deep reinforcement learning

QU Guanxing[1,2], WANG Anliang[3,✉], WU Kun[4]

（1. Beihang School, Beihang University, Beijing 102206, China；

2. School of Mechanical Engineering and Automation, Beihang University, Beijing 102206, China；

3. School of Astronautics, Beihang University, Beijing 102206, China；

4. School of Aviation, Beihang University, Beijing 102206, China）

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

Abstract To address the traditional control failure problems caused by the unidirectional force of flexible cables, multi-body nonlinear coupling, and large physical time delays in large-span mobile-following microgravity simulation systems, a deep reinforcement learning control architecture integrating terrain look-ahead perception with the Proximal Policy Optimization algorithm is proposed. Specifically, a physical-priority soft-gated reward function and a historical frame stacking mechanism are carefully designed. Supplemented by a two-dimensional "time-delay and terrain" curriculum learning strategy, the system successfully achieves look-ahead predictive compensation. Comprehensive full-speed-range digital twin experiments demonstrate that the proposed architecture drastically reduces the single-step network inference time to merely 0.35 ms. Under extreme operating conditions characterized by a severe communication time delay of 20.83 ms, continuous steep slopes, and an extreme packet loss rate of up to 95%, the dynamic tension following error is strictly constrained within a tight bound of ±15%. With a task survival rate reaching exactly 100%, the comprehensive dynamic performance significantly outperforms that of traditional improved proportional-integral-derivative controllers. More importantly, this study explicitly reveals the asymmetric physical generalization characteristic of reinforcement learning models operating within variable damping physical fields, proving that a model trained in medium-to-high damping domains can be migrated to low damping domains with lossless transferability. Based on the peak power evaluation conducted on a 12t-class wheeled crane prototype,

收稿日期：2026-04-18

网络首发时间：2026-09-01 15:35:34 网络首发地址：https://link.cnki.net/urlid/11.2625.V.20260901.1430.001

this research provides a theoretical foundation for the zero-shot cross-domain deployment of large-scale spacecraft microgravity simulation systems to physical hardware.
Key words microgravity simulation; active suspension; deep reinforcement learning; time-delay preview compensation; asymmetric physical generalization; zero-shot transfer

近年来，我国稳步推进月球科研站的建设与载人登月目标[1-2]。然而月球 1/6g 的低重力环境与覆盖松软月壤的崎岖地貌会深刻改变巡视器的轮地力学特性[3-4]。因此在地面构建高保真的等效微重力实验环境至关重要。目前，多索悬吊、气浮混合等卸载平台[5-8]已得到广泛应用。邱雪松等人提出主被动复合式磁浮-悬吊重力补偿系统有效克服了传统机构的机械摩擦，提升了恒力补偿的稳态精度[9]。但以上方案多受限于固定导轨的静态物理边界，难以满足大范围全地形的灵活测试需求。为此，引入基座可大范围自主追踪的移动跟随卸载正成为该领域的发展前沿。

尽管移动悬吊架构扩展了测试包络，但 BIHL 等研究者指出，多体物理耦合极易引入多余的惯性与侧向载荷干扰[10]。在经典反馈架构中，误差重算法[11]、LPV-PID 观测器[12]以及基于频域综合的主动恒力卸载策略[13]能有效降低测量噪声与稳态力控误差，但面对极端地貌突变时，由于缺乏前瞻补偿，极易因滞后响应引发柔性索的破坏性冲击载荷。为应对这一多体耦合与物理硬约束问题，模型预测控制（model predictive control ，MPC）及其基于观测器的鲁棒扩展方案被引入该领域[14-15]。

为缓解 MPC 高昂的在线算力与系统大时滞痛点，Yan 等研究者提出了结合神经网络观测器的预定时间鲁棒控制，实现了收敛时间可控的高精度力控[16]；还提出了引入气动人工肌肉等柔性执行器，并结合神经网络与事件触发 MPC 进行控制[17-18]。除此之外，还有研究通过叠加被动恒力机构与主动音圈电机来构建六自由度复合测试台，以缓解重力补偿的不稳定性[19]。上述研究有效提升了系统的鲁棒表现。然而，尽管基于高阶解算器的非线性模型预测控制在处理多维物理硬约束上极具优势，但其高昂的在线求解耗时往往受限于实际系统的车载底层算力。为此，探索了一种兼顾极速在线推理与前瞻时滞补偿的轻量化数据驱动控制范式，成为该领域的重要工程补充。

面对传统依靠精确解析建模与滞后反馈的固有瓶颈，强化学习凭借其端到端的极速推理能力展现出巨大潜力[20-21]。最新研究证实了融合深度强化学习与预测经验机制在悬架时滞补偿中的显著优势[22]。然而，直接将纯数据驱动策略迁移至高刚度、大时滞的悬吊系统时，仍面临奖励稀疏、高频动作抖振与抗冲击的落地难题。为此，本文打破传统控制范式，构建了一套大跨度移动跟随式主动悬吊物理验证系统，并提出了一种融合地形前瞻感知与近端策略优化（proximal policy optimization, PPO）的深度强化学习控制架构。本文的主要贡献如下：

1）设计了大跨度移动跟随式主动悬吊架构，为航天器全地形微重力模拟提供了新硬件范式。

2）设计物理优先级软门控奖励函数与多帧状态堆叠机制，将单步推理耗时缩减至 0.35 ms，依托地形前瞻实现由“滞后反应”向“主动预测”的跨越，显著抑制了 20 ms 级物理时滞与柔性索冲击。

3）揭示了强化学习模型在变阻尼物理场中的“非对称泛化规律”。有效抑制了纯数据驱动算法的底层稳态驻留噪声，展现出了卓越的 Sim-to-Real 迁移潜力。

1 动力学建模

1.1 系统物理架构与坐标系定义

如图 1 所示，系统的物理架构主要由三部分组成，具体参数如表 1 所示。

1）随动大车：作为系统的基座，在平坦路面上运动，负责跟随月球车的水平位移，保持绳索垂直。

2）卷扬机构：安装于大车的吊杆上，可在吊杆上左右移动，通过收放绳索控制月球车的垂直受力，实现重力补偿。

3）月球车：作为被控对象（负载），在崎岖地形上自主运动。

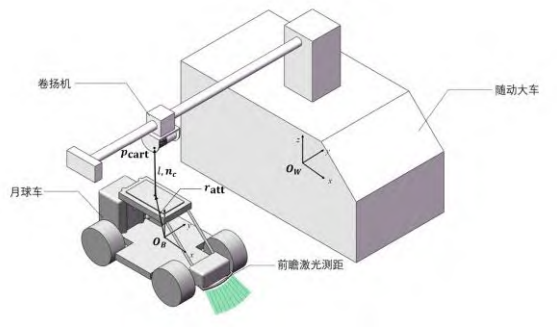

图 1 月球车微重力模拟测试系统物理架构
Fig. 1. Physical architecture of the lunar rover microgravity simulation test system

表 1 微重力悬吊系统核心参数
Table 1 Core parameters of the microgravity suspension system

物理量	数值
月球车质量m	600kg
大车质量M	10000kg
月球车目标速度$\boldsymbol{v}_{\mathbf{target}}$	5.6m/s≈20km/h
大车最大速度$\boldsymbol{v}_{\mathbf{max_gantry}}$	12m/s=43.3km/h
吊绳劲度系数k_s	15000N/m
卷扬机阻尼系数c_s	6000kg/s

为精确描述系统的多体空间运动，本文定义了两个全局独立坐标系（如图 1 所示）：

1）世界坐标系 $\{W\}$（$O_W - X_W Y_W Z_W$）：原点 O_W 固定于大车导轨起点所在水平地面。X_W 轴平行于主导轨指向大车前进方向，Z_W 轴垂直于水平地面向上，Y_W 轴由右手定则确定。大车出绳点的位置向量 p_{cart}、月球车质心位置向量 p_r 均在 $\{W\}$ 中表达。

2）机体坐标系 $\{B\}$（$O_B - X_B Y_B Z_B$）：原点 O_B 固定于月球车质心。X_B 轴位于车辆纵对称面内指向正前方，Z_B 轴垂直于车体底盘向上，Y_B 轴指向车体左侧。绳索挂点相对于质心的偏置向量定为 r_{att}，该向量固连于 $\{B\}$ 系中，且为常数向量。

两个坐标系之间的空间位姿转换通过旋转矩阵 R_{WB} 表达。

1.2 多体动力学方程推导

随动大车被限制在单自由度水平导轨上运动。其动力学方程表示为

$$M\ddot{\boldsymbol{x}}_{\mathbf{c}} = \boldsymbol{u}_{\mathbf{cart}} - \mu_{\mathrm{c}}\dot{\boldsymbol{x}}_{\mathbf{c}} - (\boldsymbol{F}_{\mathbf{c}} \cdot \boldsymbol{e}_{\boldsymbol{x}}) \tag{1}$$

式中：M 为大车与卷扬机构的总质量；x_{c} 为大车的水平位置；u_{cart} 为大车电机的驱动力；μ_{c} 为粘性摩擦系数；$\boldsymbol{F}_{\mathbf{c}}$ 为绳索作用在月球车上的拉力矢量。

卷扬机通过改变滚筒转速来控制绳索释放长度。将其简化为旋转惯量系统，其动力学方程为

$$J_{\mathrm{w}}\ddot{\boldsymbol{\vartheta}} = \boldsymbol{u}_{\mathbf{winch}} - TR_{\mathrm{w}} - b_{\mathrm{w}}\dot{\boldsymbol{\vartheta}} \tag{2}$$

绳索释放速度与滚筒转速的运动学关系为：

$$\dot{l_0} = R_{\mathrm{w}}\dot{\vartheta} \tag{3}$$

式中：J_{w} 为卷扬机滚筒的转动惯量；$\boldsymbol{\vartheta}$ 为滚筒转角；R_{w} 为滚筒半径；u_{winch} 为卷扬电机的输出力矩；T 为绳索内部张力标量，产生的负载力矩为 TR_{w}；b_{w} 为电机及传动系统的阻尼系数；l_0 为绳索的自然长度。

月球车在非结构化地形上运动时，具有六个自由度（3 个平动 + 3 个转动）。采用牛顿-欧拉方程进行描述。

1）平动方程。

$$m_{\mathrm{r}}\ddot{\boldsymbol{p}}_{\mathbf{r}} = \boldsymbol{F}_{\mathbf{c}} + m_{\mathrm{r}}\boldsymbol{g} + \sum_{i=1}^{4}\boldsymbol{F}_{\boldsymbol{tire},i}(\boldsymbol{p}_{\mathbf{r}},\mathrm{Terrain}) \tag{4}$$

式中：m_{r} 为月球车质量；$\boldsymbol{p}_{\mathbf{r}}$ 为月球车质心的位置向量；$\boldsymbol{g} = [0,0,-g]^T$ 为重力加速度矢量；$\boldsymbol{F}_{\boldsymbol{tire},i}$ 为第 i 个车轮受到的地面接触力。

2）转动方程。在月球车坐标系下，欧拉方程为

$$\boldsymbol{I}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega} \times (\boldsymbol{I}\boldsymbol{\omega}) = \boldsymbol{\tau}_{\mathbf{total}} \tag{5}$$

式中: $\boldsymbol{I}$ 为月球车的惯性张量，$\boldsymbol{\omega}$ 为角速度矢量。关键在于合外力矩 $\boldsymbol{\tau_{total}}$ 的计算，它包含了绳索偏置拉力产生的力矩。为精确描述轮地交互，将车轮接触力在局部接触坐标系下进行正交分解，合外力矩的具体展开形式为

$$\boldsymbol{\tau}_{\text{total}} = \underbrace{(\boldsymbol{R}_{\text{WB}}\boldsymbol{r}_{\text{att}}) \times \boldsymbol{F}_c}_{\text{Cable Torque}} + \sum_{i=1}^{4}(\boldsymbol{R}_{\text{WB}}\boldsymbol{r}_{c,i}) \times \left(\boldsymbol{R}_{C_i}^{W} \begin{bmatrix} F_{x,i} \\ F_{y,i} \\ F_{z,i} \end{bmatrix} \right) \tag{6}$$

式中：$\boldsymbol{R_{WB}}$ 为从机体坐标系到世界坐标系的旋转矩阵；$\boldsymbol{r_{att}}$ 为绳索挂点相对于月球车质心的偏置向量。；$\boldsymbol{r}_{c,i}$ 为机体坐标系下质心到第 i 个车轮实际接地点的三维位置向量；$\boldsymbol{R}_{C_i}^{W}$ 为第 i 个车轮局部接触坐标系到世界坐标系的旋转矩阵；$F_{x,i}$、$F_{y,i}$ 和 $F_{z,i}$ 分别为第 i 个车轮在接触点处受到的纵向力（牵引/制动摩擦力）、侧向力（转向摩擦力）以及法向力（地面支持力）。该式表明，若挂点不在质心正上方，或非结构化地形导致各车轮三向受力不均时，将对车体产生倾覆力矩， 这是控制系统必须抑制的扰动。

为了模拟绳索的柔性及松弛现象，采用开尔文-沃伊特（Kelvin-Voigt）黏弹性理论来描述柔性索的力学行为[23]。

$$\boldsymbol{T} = \max\left(0, k_{\text{s}}(l - l_0) + c_{\text{s}}\left(\dot{\boldsymbol{l}} - \dot{\boldsymbol{l}}_{\boldsymbol{0}}\right)\right) \tag{7}$$

式中：k_{s} 和 c_{s} 分别为绳索的刚度系数和阻尼系数。

绳索的实际长度 l 和方向向量 $\boldsymbol{n_c}$ 由大车与月球车的几何关系决定：

$$l = |\boldsymbol{p}_{\mathbf{cart}} - (\boldsymbol{p}_{\mathbf{r}} + \boldsymbol{R}_{\mathbf{WB}}\boldsymbol{r}_{\mathbf{att}})|_2 \tag{8}$$

$$\boldsymbol{n}_{\mathbf{c}} = \frac{\boldsymbol{p}_{\mathbf{cart}} - (\boldsymbol{p}_{\mathbf{r}} + \boldsymbol{R}_{\mathbf{WB}}\boldsymbol{r}_{\mathbf{att}})}{l} \tag{9}$$

式中：$\boldsymbol{p}_{\mathbf{cart}}$为绳索与大车连接处的出绳点位置向量。绳索作用于月球车的拉力矢量为 $\boldsymbol{F}_{\mathbf{c}} = T \cdot \boldsymbol{n}_{\mathbf{c}}$。

值得指出的是，开尔文-沃伊特模型通常适用于小变形与低应变率工况。在本文设定的测试工况中，月球车水平最高速度为 5.6 m/s，最大坡度 $\theta \approx 15°$，绳索轴向最大相对伸缩速度$v_{\text{axial}} = 5.6 \times \sin(15°) \approx 1.45$ m/s。以系统典型参数（绳长 $L = 12$ m，小应变上限 $\varepsilon_{\max} = 5\%$）核算，系统维持线性应变的临界轴向速度 $v_{\max} = 1.5$ m/s。实际最大轴向速度未突破该临界值。因此，采用该黏弹性模型既能在理论上保证足够的动力学保真度。

2 控制策略设计

综合前文动力学建模可知，移动悬吊系统不仅包含柔性索单向受力引发的刚度突变，其大车位移与绳索空间偏角之间亦存在深度的乘性非线性耦合。此外，物理大时滞在系统高刚度特性下极易导致位移误差被急剧放大。传统依赖精确解析模型与即时误差反馈的线性控制框架难以兼顾边界约束与时滞补偿。

针对以上困境，本文提出一种融合地形前瞻信息的深度强化学习控制框架[24]。本文将详细阐述控制架构的设计、马尔可夫决策过程的构建以及奖励函数的优化设计。为保证稳态精度的同时提高对突发地形扰动的响应速度，本文设计了如图 2 所示的前馈-反馈融合控制架构。

该架构包含两条并行的数据通路：

1）反馈通路：传感器组实时采集绳索张力、偏角、大车速度与绳长等物理响应，反映系统“误差状态”以保证闭环稳定性；

2）前馈通路：前视激光测距扫描前方地形生成高度序列，作为“环境预判”直接输入策略网络，主动补偿地形起伏引发的张力波动。其中，前视激光测距仪为安装在月球车车头，呈扇形排列，向斜下方发射激光测距的装置。

两路信号在状态融合模块进行时空对齐与特征融合，形成高维状态向量 s_t，输入至 PPO 智能体

生成控制指令。

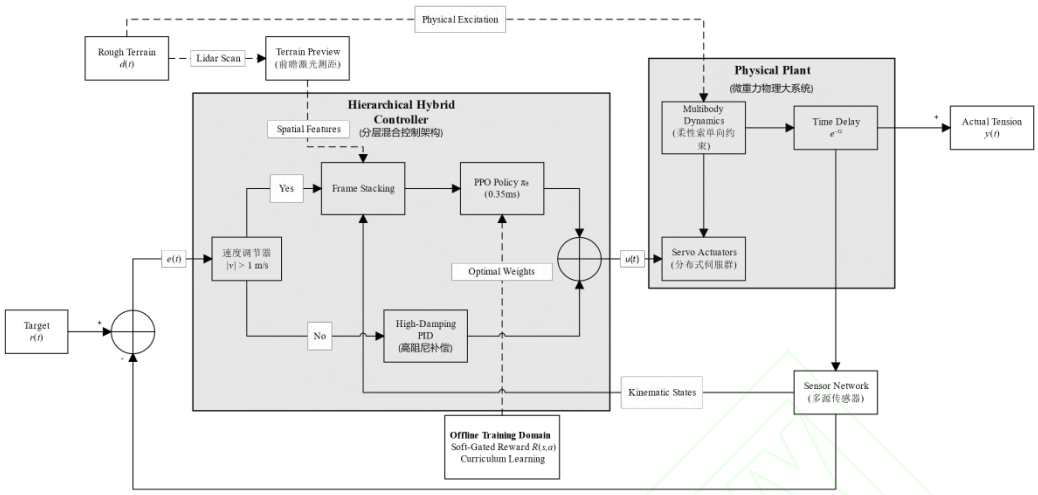

图 2 融合地形前瞻的微重力主动悬吊控制架构

Fig. 2 Control architecture of the active suspension system for microgravity simulation based on terrain preview

2.1 强化学习环境构建

2.1.1 状态空间

为了实现对复杂多体系统的精准控制，状态向量$\boldsymbol{s_t} \in \mathbb{R}^{25}$设计为混合特征向量，包含物理状态反馈与环境感知信息：

各项定义如下：

$$\boldsymbol{s_t} = [\boldsymbol{\Delta p}, \boldsymbol{\xi}_{\mathbf{cable}}, v_{\mathrm{cart}}, y_{\mathrm{slide}}, T, \boldsymbol{\Theta}_{\mathbf{rover}}, \dot{z}_{\mathrm{rover}}, \boldsymbol{D}_{\mathbf{lidar}}]^T \tag{10}$$

式中：Δp为吊钩与月球车相对位置向量；$\boldsymbol{\xi}_{\mathbf{cable}}$为绳索向量；$v_{\mathrm{cart}}$为大车速度；$y_{\mathrm{slide}}$为吊钩横移位置；$T$为绳索拉力；$\boldsymbol{\Theta}_{\mathbf{rover}}$为月球车姿态；$\dot{z}_{\mathrm{rover}}$为月球车$z$方向速度；$\boldsymbol{D}_{\mathbf{lidar}}$为激光测距数据。

2.1.2 动作空间

PPO 策略网络输出的动作向量$\boldsymbol{a_t} \in \mathbb{R}^3$对应系统的三个独立执行机构，分别控制纵向跟随、横向跟随及垂直重力补偿。

定义动作向量为

$$\boldsymbol{a_t} = \left[u_{\mathrm{throttle}}, v_{\mathrm{lat}}, v_{\mathrm{rope}}\right]^T \tag{11}$$

式中：u_{throttle}为大车油门，控制随动大车沿主导轨（纵向 x 轴）的驱动力矩与牵引力；v_{lat}为吊钩横移速度；v_{rope}为卷扬机绳索收放速度。

2.1.3 奖励函数设计

1）基于物理优先级的软门控奖励机制。总体奖励函数 R_t：

$$R_t = \mathcal{G}_{gate}(\theta) \cdot (w_{\mathrm{ten}} r_{\mathrm{ten}}) + w_{\mathrm{vert}} r_{\mathrm{vert}} + w_{\mathrm{aux}} r_{\mathrm{aux}} + P_{\mathrm{penalty}} \tag{12}$$

式中：$\mathcal{G}_{gate}$为门控系数 $r_{\mathrm{vert}} = exp(-\frac{(1.0-v_{dot})^2}{\sigma_{vert}})$，其中$v_{\mathrm{dot}}$为绳索向量与垂直方向夹角的余弦值；$r_{\mathrm{ten}} = exp(-\frac{|\Delta T|}{\sigma_{ten}})$,其中$\Delta T$为当前实际张力与目标恒定张力之间的误差值；$r_{\mathrm{aux}} = r_{xy} = \exp\left(-\frac{|\boldsymbol{p}_{\mathbf{cart}}-\boldsymbol{p}_{\mathbf{rover}}|_{xy}^2}{\sigma_{pos}^2}\right)$，$|\cdot|_{xy}$ 表示水平面上的欧氏距离范数；$\sigma_{\mathrm{vert}} = 0.05$，$\sigma_{\mathrm{ten}} = 0.2$，$\sigma_{\mathrm{pos}} = 1.0$，为对应项的敏感度系数；$w_{\mathrm{vert}} = 5$，$w_{\mathrm{tension}} = 4, w_{\mathrm{aux}} = 1$ 为对应的奖励权重系数；P_{penalty}为翻车惩罚，当月球车姿态角超过一定阈值时判定为翻车，给予较大惩罚并重置状态。

在构建门控系数 $\mathcal{G}_{gate}$ 时，本研究未采用简单的分段函数。因为硬截断会导致奖励信号不连

续，而在强化学习中学习信号与策略梯度的连续平滑性对于复杂动作空间的收敛至关重要[25]。为此，本文设计了基于 Sigmoid 函数的连续平滑软门控系数：

$$\mathcal{G}_{gate}(\theta)=\frac{1}{1+e^{-\beta(r_{\mathrm{vert}}-\theta_{\mathrm{th}})}} \tag{13}$$

式中：θ_{th} 为门控阈值，本次训练设定为 0.93；β 为门控斜率系数，本次训练设定为 80。

最终，该奖励函数通过“软门控”机制，成功在神经网络中植入了“垂直度 ≫ 拉力”的严格物理逻辑，彻底解决了策略退化问题。

2）广域位置引导与全局梯度重塑。在强化学习的训练初期，或者当系统遭受剧烈地形扰动导致绳索极度偏离垂直位置时，前两级奖励（几何垂直 r_{vert} 与恒张力 r_{ten}）往往面临严重的奖励稀疏问题。

由于 r_{vert} 和 r_{ten}有效响应域狭窄（σ_{vert}与σ_{ten}较小）。当大车与月球车的水平偏差超出一定范围时，$r_{\mathrm{vert}}\approx 0$；此时由于几何构型限制，张力往往无法维持恒定，导致 $r_{\mathrm{ten}}\approx 0$，进而导致策略网络的梯度趋近于零，智能体失去优化的方向指引，训练无法收敛。为了解决次问题，本文引入了基于大车与月球车相对欧氏距离的广域位置引导奖励 r_{aux}，为整个状态空间构建了一个全局收敛势场。r_{aux} 的权重 w_{aux} 最低，但它确保了智能体始终在物理可行的解空间内进行探索，为后续的高精度力控创造了必要的几何前提。

2.2 仿真平台搭建与训练策略

本研究基于高性能物理引擎 PyBullet 构建了移动悬吊系统的数字孪生环境[26]。近年来，得益于物理仿真器与深度强化学习的结合，在纯仿真环境中训练复杂机械系统并直接部署到物理世界的 Sim-to-Real 技术路线已得到广泛验证[27]。因此，本文在仿真环境中完整集成了大车-绳索-月球车多体动力学方程，并模拟了真实物理约束。

2.2.1 仿真平台搭建

为模拟真实的实验条件与物理约束，仿真参数依据实际硬件系统的物理特性进行设置，并开发了观察延迟缓冲区。传感器数据生成后，存入长度为 N_{delay} 的队列，策略网络读取队列尾部即其前 n 帧的历史状态。为了测试算法对非结构化地形的适应能力，仿真环境采用了 K. Perlin 提出的分形噪声生成算法（Perlin Noise）[28]，生成了连续起伏的随机地形。

2.2.2 基于“模型演进”的渐进式课程学习策略

本文引入了由 Y. Bengio 等人提出的“课程学习”训练范式[29]。基于此思想，本文设计了包含基准模型构建与最终模型适配的训练流程，并逐步改进模型训练方式。

1）基准模型训练。为验证 PPO 处理多体耦合动力学的可行性，并在消融实验中提供基准对照，本文在无时滞（$T_d=0$）环境下采用常规线性加权奖励函数训练了基准模型。该模型在无时滞平地及坡道工况下展现出优异的跟随精度，证实了数据驱动方案的有效性。然而，置入含时滞（$T_d>0$）环境后，其性能发生断崖式下跌并引发剧烈等幅震荡，凸显了引入时序架构升级与高阶课程学习的必要性。

2）最终模型训练。针对系统通讯链路中存在的显著物理延迟，导致系统面临严重的非马尔可夫问题。为解决这一问题，本文引入了历史状态堆叠机制[30]，将过去 k 帧历史状态序列联合作为当前观测输入。在架构升级后，本文采用了如表 2 所示时滞与地形交替递进的课程学习路径。

表 2 “时滞-地形”二维课程学习训练策略

Table 2 Training strategy of the "delay-terrain" two-dimensional curriculum learning

课程阶段	通讯时滞	地形特征
S1	0 帧 (0ms)	平坦地面
S2	2 帧 (8.33ms)	平坦地面
S3	2 帧 (8.33ms)	平地 + 坡道
S4	4 帧 (16.67ms)	平地 + 坡道
S5	5 帧 (20.83ms)	平地 + 坡道

单次回合的坡道训练地形被参数化设计为“下坡-水平过渡-上坡”复合凹陷特征，其坡度值 $\tan\theta$

（θ 为物理坡角）于 [0.15, 0.22] 区间内进行均匀随机采样。

本实验基于 PyTorch 深度学习框架实现 ，超参数设定参考 PPO 原始文献中的连续控制任务经验配置[24]，并结合移动悬吊系统大惯量特性进行微调。策略网络与价值网络均采用 Adam 优化器进行参数更新[31]。具体参数如表 3 所示。

表 3 PPO 深度强化学习算法超参数设置

Table 3 Hyperparameter settings of the PPO deep reinforcement learning algorithm

参数	学习率	批量大小	折扣因子 γ	剪切阈值
基准模型	3×10^{-4}（线性衰减）	2048	0.995	0.25
最终模型	3×10^{-4}（线性衰减）	4096	0.995	0.25

3　仿真对比与分析

3.1 对比实验

3.1.1 算法评估指标

在时滞与崎岖地形耦合的测试中，系统极易发生动力学失稳。本文定义“任务失败”为：在测试序列中，绳索张力发生的极端跳变（张力低于 1000N 或高于 8500N）。任务成功率即为未触发上述失效条件的测试样本占总测试样本的百分比。为排除系统崩溃样本与启动初期物理调节过程的干扰，评价指标均截取启动 3 s 后的成功样本（$t\in[3.0,T_{\mathrm{end}}]$）进行统计。

本文选取以下归一化均方根误差（Normalized Root Mean Square Error, NRMSE）、阈值截断相对标准差（Threshold-Clipped Standard Deviation, TCSD）与归一化实际波动标准差（Normalized Standard Deviation, NSD）作为参考指标。针对悬吊系统对“安全性”的特殊要求，传统均方误差无法区分“允许范围内的调节波动”与“危险的失稳震荡”。为此，本文提出了阈值截断指标 $\delta_{\mathrm{th}}=15\%$，仅计算超出该阈值的恶性偏差，并将其转化为相对标准差形式。TCSD 定义为

$$\mathrm{TCSD}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\mathbb{I}(e_t)\cdot(e_t)^2} \tag{14}$$

式中：定义相对误差序列 $e_t=\frac{|F_t-F_{ref,t}|}{F_{ref,t}}$；定义截断函数 $\mathbf{1}(e_t)=\begin{cases}0, & if\ \ e_t\le 0.15\\ 1, & if\ \ e_t>0.15\end{cases}$。

3.1.2 对比实验分析

为评估本文提出的 PPO-前瞻控制策略性能，本研究将其与比例-积分-微分（proportional-integral-derivative, PID）控制器进行对比[20]。在悬吊式重力卸载系统中，引入鲁棒控制思想以对抗绳索非线性与系统大惯量是当前的共识[14]，因此本研究在标准 PID 的基础上引入了一阶低通滤波的噪声抑制与针对绳索松弛的智能积分抗饱和[11]，以代表传统线性控制在处理此类非线性系统时的性能上限作为高标准基准。在无时滞状态下，将基准模型与优化 PID 在含随机噪声的平地及连续陡坡地形中进行对比，表 4 展示了量化结果。

表 4 不同控制策略的动态跟踪性能量化评估对比

Table 4 Quantitative performance evaluation comparison among different control strategies

评估指标	优化 PID	基准模型	优化 PID	基准模型
	平地工况		坡道工况	
成功率	53%	95%	69%	51%
均方根误差	0.0292	0.0810	0.0273	0.0588
截断方差	0.0071	0.0047	0.0068	0.0018
方差	0.0292	0.0219	0.0273	0.0270

数据分析：

1）平地工况下基准 PPO 模型成功率远超 PID，但坡道工况下却骤降至 51%。表明缺乏“物理软门控”的常规线性奖励函数无法解决重力突变时“张力维持”与“摆角控制”的强耦合冲突。而采用软门控奖励函数设置的最终模型在平地与坡道上的成功率则高达 96%以上。

2）PID 幸存样本的均方根误差（0.0292, 0.0273）虽低于基准模型（0.0810, 0.0588），但基准模型方差更小（0.0219, 0.0270），其均方根偏高系主动引入稳态偏置以换取下坡缓冲的安全裕度策略所致，并非控制震荡。

3）在衡量边界失稳的 TCSD 指标上，基准模型（0.0047, 0.0018）全面压制 PID（0.0071, 0.0068）。

3.1.3 实际工况下的时滞敏感性与控制机理分析

为了探究不同控制策略在面对通讯与计算时滞时的鲁棒性边界，本节在相同的坡道混合地形工况下，测试了三种模型在不同延迟帧数（N_{d}）下的性能表现。仿真步长设为$\Delta t = 1/240s$，测试范围覆盖 $0 \sim 5$ 帧（约 0ms ~ 21ms）。

图 3 展示了三种模型归一化均方根误差随延迟帧数的变化趋势，结合实验数据，我们可以得出以下深层控制机理分析：

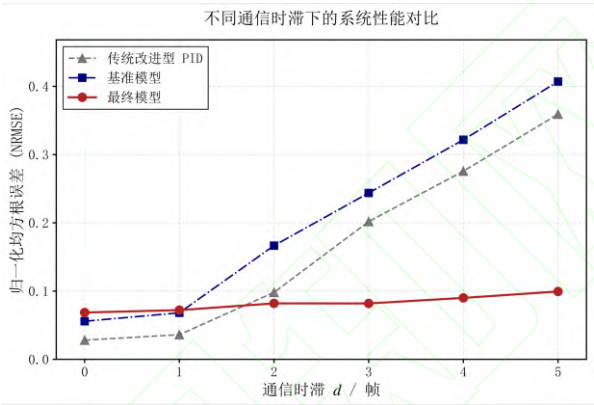

图 3 不同通信时滞下各控制策略的 NRMSE 变化趋势对比

Fig. 3 Comparison of NRMSE trends for different control strategies under varying communication latencies

1）基准模型与 PID 的“反应式”同构退化。如图 3 所示，PID 与基准模型的 NRMSE 随延迟增加均呈指数级发散，趋势高度一致。这揭示了无时滞温床训练的底层缺陷：策略网络为快速获取即时奖励而陷入局部最优，无视了输入层的前瞻地形数据。其控制本质退化为仅依赖当前反馈的“非线性 PID”。这种纯“反应式”控制在面对物理链路的绝对滞后时，必然走向动态崩溃。

2）最终模型的“预测式”时滞补偿。与前两者的崩溃形成反差，最终模型在 $0 \sim 5$ 帧的延迟跨度下，其 NRMSE 仅从 0.0686 微增至 0.0994，抗扰曲线近乎水平。这定量证实了本文“时空融合感知架构”的有效性：包含大时滞注入的课程学习促使策略网络利用前瞻测距信息，并依托历史状态序列形成时序记忆，实现了从“滞后反馈”向“前瞻补偿”的控制范式跨越。

3.1.4 算法算力与实时性对比分析

为定量评估算力消耗，本文在单核 CPU（模拟典型工控机）下对 PPO 策略网络进行了 1000 次蒙特卡洛测试。实测单步推理均时仅为 0.3556 ms（峰值 1.6158 ms，标准差$0.0857ms$），理论控制频率高达 2811.9 Hz。相比之下，根据 S. Sun 等人在最新多体线缆悬吊系统研究中的基准测试，即便采用学术界当前顶级的开源 C++ 代码生成求解器，在配备高性能 Intel Core i7 处理器的平台上，其非线性模型预测控制（Nonlinear Model Predictive Control, NMPC）的单步在线求解耗时仍高达数十毫秒[16]。结合前文时滞鲁棒性的测试结论，PPO 算法的超低推理耗时为底层传感器信号传输与电机响应预留了充裕的时间裕度，有效保障了系统在 20 ms 延迟约束下的高精度控制。

3.2 关键超参数敏感性分析

为排除纯数据驱动策略中常见的超参数偶然性，并深度探究物理优先级软门控机制在底层的演化规律，本节在剥离地形与时滞扰动的基础平地工况下，对关键奖励参数展开了系统的网格化敏感性分析。

测试的基准参数为最终模型所使用的参数（$w_{vert}=5, w_{ten}=4, k=80, \theta_{th}=0.93, \sigma_{\mathrm{vert}}=0.05, \sigma_{ten}=0.2$），测试工况为附带随机噪声的水平路面，无延迟，训练轮数为 100,0000.

3.2.1 奖励权重配比与动态归一化特性

针对垂直度奖励权重 w_{vert} 与张力奖励权重 w_{ten} 的博弈关系，本研究采用了“恒定总和微扰法”（$w_{vert}+w_{ten}=9$)进行一维灵敏度测试。如图 4 所示，系统在极宽的权重分布空间内均保持了 90% 以上的成功率，且 c-NRMSE 总体维持在较低水平。

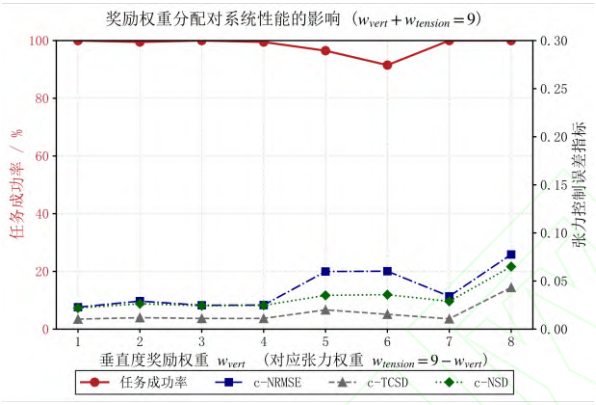

图 4 不同垂直度与张力奖励权重分配下的系统控制性能对比

Fig. 4 Comparison of system control performance under different verticality and tension reward weight allocations

这一现象揭示了物理软门控机制在解决多目标冲突中的核心优势。不同于传统的线性加权奖励易因权重失衡导致智能体为了追求单一高分项而牺牲物理约束；在本文架构中，门控机制强制植入了“先保证垂直，再优化拉力”的因果逻辑。确保随着模型收敛，智能体能稳定控制在门控开启的安全域内，同时获取垂直度与拉力的奖励。

在总奖励幅值守恒的情况下，智能体达到稳态时获取的总体收益与策略梯度方向已高度一致，权重的相对大小对系统最终收敛结果的影响被极大削弱。这从算法底层证明，软门控机制剥离了对权重超参数的过度依赖。本文最终沿用 5:4 作为基准参数以提供最平滑的梯度过渡带，为应对后续全地形、大时滞等恶劣工况奠定了泛化基石。

3.2.2 门控形态的边界试探与时空适配

物理软门控的开启时机由阈值 θ_{th} 控制，平滑度由斜率 k 决定。由表 5 的二维网格数据可知，随着门控阈值 θ_{th} 的逐渐逼近绝对垂直（即向 0.98 逼近），系统的 c-NRMSE 呈非线性急剧恶化，最终走向大面积发散（训练无法收敛）。在动态跟随中，严苛的几何约束会导致拉力奖励过低，策略网络因缺乏连续的力控梯度指引而发散。

表 5 物理软门控参数二维网格扫描下的控制均方根误差

Table 5 Control root mean square errors under two-dimensional grid scanning of physical soft-gated parameters

k/θ_{th}	0.83	0.88	0.93	0.98
40	0.0517	0.2321	0.0552	0.4415
80	0.0759	0.0785	0.0599	0.2513
120	0.0604	0.0374	0.1294	0.1648
160	0.0439	0.0427	0.0463	/

另一方面，虽然门控斜率 k 在无延迟平地工况下的横向对比中未表现出剧烈的性能级差，但其对于算法的时序平滑演化至关重要。考虑到真实物理验证中客观存在的 20ms 级通信时滞与非结构化地形，本文前瞻性地选取 $k=80$ 与 $\theta_{th}=0.93$ 作为中庸组合，以确保控制梯度在时空维度的平滑性。

3.2.3 容差窗口的对角线耦合演化规律

由表 6 可知，垂直容差 σ_{vert} 与张力容差 σ_{ten} 的高性能收敛区呈现出显著的对角性。随着垂直敏感度界限 σ_{vert} 的放宽，与之匹配的最佳张力敏感度 σ_{ten} 必须呈正相关同步放大。这一对角线分

布可能表明，当任一维度的容差显著偏离对角线时，奖赏梯度的不平衡将导致智能体的学习路径被单一因素主导，进而陷入局部最优。

表 6 垂直与张力容差参数二维网格扫描下的控制均方根误差

Table 6 Control root mean square errors under two-dimensional grid scanning of vertical and tension tolerance parameters

$\sigma_{vert}/\sigma_{ten}$	0.1	0.2	0.3	0.4
0.025	0.0679	0.0631	0.4444	/
0.05	0.1663	0.0599	0.3484	0.3370
0.075	/	0.0770	/	0.0747
0.1	0.1740	0.0604	0.0365	0.0658

分析表明，本文选定的基准容差参数（$\sigma_{vert} = 0.05, \sigma_{ten} = 0.2$）是位于对角线的平滑过渡段。该参数集在防止梯度消失与抑制局部最优之间达到了平衡，确证了本文强化学习架构具备极强的参数抗摄动能力与理论合理性。

3.2.4 基础参数向复杂工况推广的适用性说明

值得说明的是，上述参数敏感性分析均在平地无时滞的基准工况下完成，但其结论可无损推广至本文核心的大时滞与坡道地形耦合工况。其底层物理与算法依据在于：

1)这些核心参数本质上共同定义了系统“垂直度”与“张力”控制的相对物理优先级、响应敏感度与容许边界，作为客观物理底线与外在地形起伏或通讯时滞完全解耦。

2)在本文的控制架构中，应对大时滞与地形冲击的核心机制是“10 帧历史状态堆叠”与“前瞻地形感知”，而非奖励函数的二次重塑。复杂工况考验的是策略网络对高维时序状态的决策表征能力，底层的评价体系维持恒定。

3)依托“时滞-地形”二维课程学习机制，复杂工况的训练必定基于平地基准工况。最终模型在 20ms 级时滞与连续陡坡工况下实现 100% 成功率，反向印证了本节标定的超参数在复杂外推域中具备较强的鲁棒性。

3.3 模型鲁棒性验证与物理泛化机理分析

为系统性评估最终模型在复杂边界条件下的鲁棒性与跨域泛化能力，本节设计并开展了一系列多维工况交叉验证实验。如无特殊说明，后续测试的基准仿真环境均默认引入 $T_{\mathrm{d}} = 5$ 帧的系统通信时滞，并在地形中叠加连续随机噪声扰动。

3.3.1 脉冲扰动与解耦能力分析

为剥离地形耦合因素，单纯考察控制器对系统剧烈动能突变的暂态响应特性，本节在理想无噪声水平路面上开展了极端速度阶跃实验。系统初始以 7.0 m/s 匀速巡航，于 t=20 s 处注入阶跃降速指令，目标速度骤降至 4.0 m/s，以此模拟通信丢包重连或紧急避障等极端高动态场景 。

如图 5 所示，在 t=20s 的速度突变点，尽管系统经历了高达 3.0 m/s 的瞬时速度差，但全程平滑过渡，严格收敛于 ±15 的动态安全边界内。

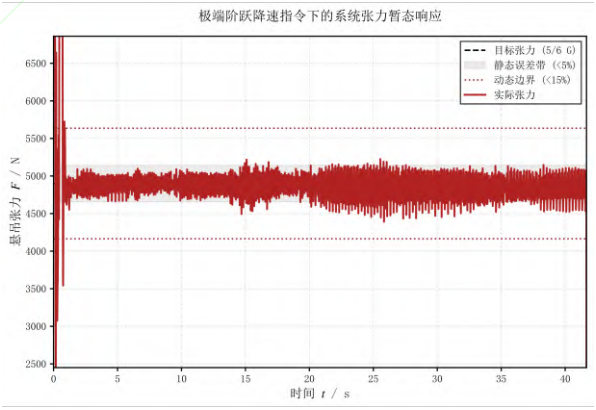

图 5 极端阶跃降速指令下的系统张力暂态响应

Fig. 5 Transient tension response of the system under extreme step deceleration commands

3.3.2 极端工况下的前馈感知鲁棒性分析

在真实深空探测与野外物理验证中，受扬尘遮挡或通讯链路高频抖动影响，前瞻传感器常出现数据随机缺失。为科学评估本控制架构的鲁棒边界，本文引入基于概率的随机丢包模型，设定前视激光数据在每一控制帧有概率 p 发生丢失；当触发丢包时，系统保持上一有效帧的数据（即零阶保持机制）。本研究在复合坡道，20.83ms 时滞工况下，对不同丢包率（p = 0.0 ~ 0.99）进行了全面的蒙特卡洛测试，量化评估如表 7 所示。

表 7 不同感知丢包率下的系统动态跟踪性能评估

Table 7 Evaluation of system dynamic tracking performance under different perception dropout rates

p 值	成功率	NRMSE	TCSD	NSD
0.0	100%	0.0447	0.0197	0.0413
0.1	100%	0.0453	0.0203	0.0420
0.3	100%	0.0441	0.0195	0.0410
0.5	100%	0.0446	0.0196	0.0410
0.7	100%	0.0450	0.0200	0.0415
0.9	100%	0.0446	0.0198	0.0414
0.95	100%	0.0448	0.0198	0.0415
0.98	99%	0.0459	0.0208	0.0421
0.99	96%	0.0461	0.0208	0.0425

测试数据表明，最终模型对极端稀疏感知展现出了卓越的鲁棒性。在丢包率攀升至 95% 的广阔区间内，系统 NRMSE 波动极小，任务成功率始终维持 100%。其底层机理在于，系统 240 Hz 的控制基频使得在 95% 丢包工况下，智能体仍有约 12 Hz 的有效地形更新，并依托零阶保持机制与历史状态堆叠成功重构了连续的地形梯度趋势。当且仅当丢包率触及 0.98 以上的极端边缘时，任务成功率才出现实质性衰减，这揭示了该前瞻补偿架构在极端工况下的物理失效临界点。

3.4 泛化机制与边界适应性验证

3.4.1 速度工况的泛化性测试

最终模型展现出显著的跨域“零样本迁移”特性。尽管智能体仅在单一基准速度（v = 5.6 m/s）及特定域随机化坡度（$\tan\theta \in [0.15,0.22]$）的工况下进行训练，但在测试中能够迁移至未曾经过训练的速度域与超出分布边界的地形中。

为量化评估策略网络在多维未知扰动下的泛化边界，本文开展了跨速域与跨坡度的交叉测试，稳态均方根误差如表 8 所示。数据表明，面对 2.8 m/s 至 7 m/s 的大跨度定速巡航和大于训练上限的陡坡耦合工况，系统均能将绳索张力误差约束在容许精度带内，全局未发生柔性索松弛失效。

表 8 跨速域与坡度泛化测试下的张力跟踪均方根误差

Table 8. Root mean square error of tension tracking under cross-speed and cross-slope generalization tests

测试工况	v=2.8m/s	v=5.6m/s	v=7m/s
$\tan\theta$=0	0.0434	0.0630	0.0764
$\tan\theta$=0.22	0.0472	0.0679	0.0860
$\tan\theta$=0.3	0.0422	0.0633	0.0899

为考察高频动态响应极限，本文在附带随机噪声的水平工况下开展了正弦连续变速跟随测试，月球车速度变化满足$v = 4 + \sin t(m/s)$。如图 6 曲线所示，在目标速度连续变化的瞬态过程中，系统张力维持稳定，未激发同频震荡干扰。验证了模型具备极强的主动阻尼抑制与抗扰动能力。

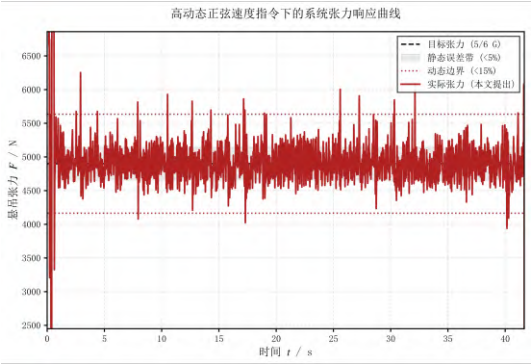

图 6 高动态正弦速度指令下的系统张力响应曲线
Fig. 6. System tension response curves under highly dynamic sinusoidal speed commands

3.4.2 全速域非单调泛化特征与静止抖振溯源

如图 7 所示，在水平多速域测试（$c_s = 3000$）中，最终模型的归一化均方根误差随速度变化呈现出反直觉的“U 型”非单调趋势。

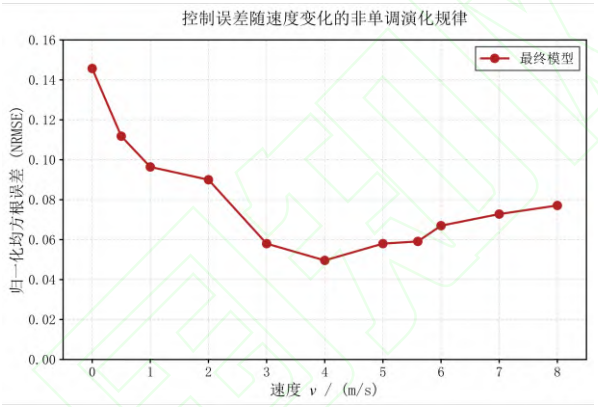

图 7 控制误差随速度变化的非单调演化规律
Fig. 7 Non-monotonic evolution law of control error under varying velocities

最终模型在高速区间（$4.0 \sim 8.0m/s$）展现出卓越的“动态自稳定”特性，误差低且增长极缓（最优点 $v = 4.0m/s$ 时仅为 0.0496）。而当系统处于静止或极低速（$v < 1.0m/s$）时，NRMSE 陡升至全域峰值（0.1457）。这种“低速劣化”的底层诱因在于系统信噪比的急剧衰减。在静止或低速状态下，系统真实宏观运动趋势趋于零，此时传感器底噪或仿真误差在观测空间中占据主导。高度敏感的 PPO 策略网络将高频随机噪声误判为真实的状态偏移并进行过度补偿。该现象在物理底层映射了非线性控制中的“抖振”机理，在算法表层则体现为深度强化学习在极微小状态偏差下极易触发的“动作震荡”通病[32]。随着速度跨越低速阈值，真实运动学特征重新覆盖底噪，系统迅速恢复并收敛至高精度状态。

为弥补这一盲区，本方案建议在实体化部署时使用 “基于速度调度的分层混合控制架构”：

1）死区锁定模式（$|v| \leq 1.0$ m/s）：系统切断策略网络输出，强制降维为“高阻尼 PID 控制”。

2）前瞻巡航模式（$|v| > 1.0$ m/s）：当月球车建立起宏观运动趋势并跨越信噪比阈值后，控制权平滑过渡至 PPO 智能体。

3.4.3 变阻尼场中的非对称泛化与向下兼容机制

为实现策略的零样本跨域部署，Tobin 等人首次提出了域随机化技术，证明了通过在仿真环境中向物理及视觉参数注入宽泛的随机噪声，可以显著提升模型对未建模误差的鲁棒性[33]。

本文在宽域变阻尼工况（$0 \leq c_s \leq 9000$）下，对仅在单一固定阻尼（$c_s = 6000$）训练的最终模型进行了零样本迁移测试。如图 8 所示，实验揭示了一种高度反直觉的“非对称物理泛化特性”：随着阻

尼增加，系统的 NRMSE 从 $c_s = 0$ 时的 0.0315 单调恶化至 $c_s = 9000$ 时的 0.1165。这意味着，模型对未见过的低阻尼域展现出了完美的“向下兼容”，却在更高阻尼域的泛化受阻。

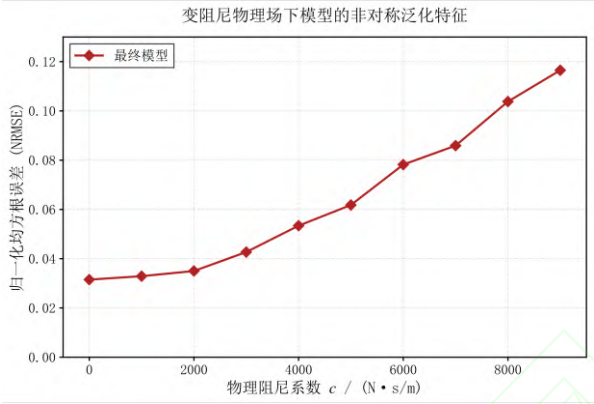

图 8 变阻尼物理场下强化学习模型的非对称泛化特征

Fig. 8 Asymmetric generalization characteristics of the reinforcement learning model under varying physical damping fields

该现象的底层逻辑在于动力学特征的显隐性与信噪比差异。低阻尼工况下，系统表现为显著的欠阻尼振荡，高信噪比信号清晰地放大了由刚度与惯性主导的物理演化规律，使智能体极易捕捉其状态转移特征；而在高阻尼域，强烈的耗散项迅速掩盖了惯性作用，系统能量极速衰减至稳态。此时动力学特征极度“隐性化”，导致物理特征区分度骤降，预测相对误差因而被显著放大。

上述发现为传统域随机化理论提供了物理学补充[27]。面向实际样机部署时，无需在全参数空间内进行低效的均匀随机采样；仅需在“中高阻尼”的边界区进行重点训练，即可以极低的训练成本实现对低阻尼物理域的泛化。

3.4.4 面向工程应用的建议与系统工作边界

在真实的物理部署中，系统的全局性能必然受限于底层机电执行机构的绝对物理饱和特性。为明确本系统由理论向工程实体化跨越的可行性，基于 12t-16t 级轮式重载起重机原型核算表明，系统牵引与卷扬的峰值机械响应功率分别需达到工程配置的 300 kW 与 22 kW 级。

鉴于深度强化学习算法的“黑盒”不可解释性，在面向航天级高可靠性测试需求部署时，须构建理论上的安全底线。建议在实体样机控制器的底层驱动层引入基于控制屏障函数或传统运动学包络的安全过滤器。当且仅当 PPO 策略网络输出的动作指令处于有效补偿的安全边界内时予以执行；反之，底层控制权将瞬间被硬限幅接管。

此外，当前验证主要依托高保真数字孪生环境。物理系统中未建模的减速机非线性摩擦、真实电机死区特性等情况，均可能导致策略的跨域衰减。未来工作将聚焦于引入硬件在环半物理仿真验证，以进一步弥合 Sim-to-Real 的物理鸿沟。

4 结 论

1）聚焦于移动跟随式微重力模拟系统的控制瓶颈，针对架构中高刚度柔性索单向受力、多体非线性耦合以及通讯时滞导致的传统控制失效难题，本文提供了一种端到端的数据驱动解决方案。实现了非结构化月面地形下的高频动态重力补偿。

2）设计了物理软门控奖励函数与历史状态堆叠观测空间，并辅以“时滞-地形”二维课程学习策略。该架构将单步推理耗时极度压缩至 0.35ms，打破了 NMPC 40ms+ 的算力枷锁，且各项数据优于传统的 PID 控制模型。

3）深入探究智能体在全速域及变阻尼场中的泛化机理，揭示了强化学习在多体系统中的“阻尼非对称泛化特性”：在中高阻尼工况下训练的模型，能够零样本无损迁移至低阻尼域。此规律突破了传统域随机化的均匀采样范式，证明本模型具备极强的抵抗未建模阻尼误差的能力，为实体样机的高

效、低成本零样本跨域部署提供了理论参考。

致谢 本文的研究工作源于国家航天任务的真实工程痛点。特别感谢中国航天科技集团公司第八研究院在月球车微重力地面模拟系统的前期预研与概念验证阶段，为本课题提供的工程需求及修改建议。

参考文献（References）

[1] 吴伟仁，刘继忠，唐玉华，等．中国探月工程[J]．深空探测学报，2019，6（5）：405-416． WU W R，LIU J Z，TANG Y H，et al．China lunar exploration program[J]．Journal of Deep Space Exploration，2019，6（5）：405-416（in Chinese）．
[2] LIN Y T,YANG W,ZHANG H,et al.Return to the Moon: new perspectives on lunar exploration[J].Science Bulletin,2024,69(16):2136-2148.
[3] SUTOH M,WAKABAYASHI S,HOSHINO T.Traveling and abrasion characteristics of wheels for lunar exploration rover in vacuum[J].Journal of Terramechanics,2016,68:37-49.
[4] NISHIDA S I,WAKABAYASHI S.A mobility system for lunar rough terrain[C]//Proceedings of ICROS-SICE International Joint Conference.Piscataway:IEEE,2009:16-21.
[5] 朱战霞，袁建平．航天器操作的微重力环境构建[M]．北京：中国宇航出版社，2013：23-229． ZHU Z X，YUAN J P．Microgravity environment construction for spacecraft operation[M]．Beijing：China Astronautic Publishing House，2013：23-229（in Chinese）．
[6] WEI Y,YANG X,XU Z,et al.Novel ground microgravity experiment system for a spacecraft-manipulator system based on suspension and air-bearing[J].Aerospace Science and Technology,2023,141:108587.
[7] JIAO B,HE B,ZHANG Y,et al.Reconfigurable modular gravity offloading system for multi-body suspension target[J].Aerospace Science and Technology,2026,168:110898.
[8] BUSE F,BURKHARD L,BORGSMÜLLER N,et al.Mobility test means of the MMX Rover IDEFIX at DLR - Prototypix[C]//Proceedings of 2025 International Conference on Space Robotics(iSpaRo).Piscataway:IEEE,2025:454-460.
[9] 邱雪松，吴承霖，邵梓康，等．月球车磁浮-悬吊重力补偿恒力控制系统设计[J]．华中科技大学学报（自然科学版），2025，53（6）：139-144．DOI：10．13245/j．hust．250619．
[10] BIHL T.A review of gravity offloading[C]//Proceedings of 2025 IEEE National Aerospace and Electronics Conference(NAECON).Piscataway:IEEE,2025:11235450.
[11] DA SILVA L R,FLESCH R C C,NORMEY-RICO J E.Analysis of anti-windup techniques in PID control of processes with measurement noise[J].IFAC-PapersOnLine,2018,51(4):948-953.
[12] PHAM T P,SENAME O,DUONG-THI T H.PID LPV observer for a semi-active suspension system[J].IFAC-PapersOnLine,2025,59(15):73-78.
[13] SURESHKUMAR V,HARIB K.Simulation of low gravitational field using gravitational offloading and force control[C]//Proceedings of 2024 Advances in Science and Engineering Technology International Conferences(ASET).Piscataway:IEEE,2024:10708663.
[14] SUN S,FRANCHI A.Nonlinear MPC for full-pose manipulation of a cable-suspended load using multiple UAVs[J/OL].arXiv preprint arXiv:2301.08545,2023.
[15] YAN H,LU H,ZHANG Z,et al.Robust model predictive control based on ESO for suspension gravity offloading system[C]//Proceedings of the 44th Chinese Control Conference.Piscataway:IEEE,2025:3099-3104.
[16] YAN H,LU H,YANG Y,et al.Predefined-time robust control for a suspension-based gravity offloading system[J].Aerospace,2025,12(6):495.
[17] YAN H,LU H,WANG H,et al.Neural-network-based nonlinear model predictive control of suspension gravity offload system[C]//Proceedings of the 3rd Conference on Fully Actuated System Theory and Applications(FASTA).Piscataway:IEEE,2024:10595198.
[18] YAN H,LU H,YANG Y,et al.Suspension gravity offloading system with event-triggered ESO-based robust MPC[J].ISA Transactions,2025,167:849-863.
[19] JIN D,HUO R,NIE Y,et al.A six-degree-of-freedom microgravity test-bed to provide gravity compensation for asteroid exploration robots[C]//Proceedings of 2025 International Conference on Mechatronics,Robotics,and Artificial Intelligence(MRAI).Piscataway:IEEE,2025:269-276.
[20] HYLA P.The crane control systems: a survey[C]//Proceedings of 2012 17th International Conference on Methods and Models in Automation and Robotics(MMAR).Piscataway:IEEE,2012:505-509.
[21] 宋健，孙浩轩，李德勇，等．大型空间机构展开卸载系统垂向主动力控仿真[J/OL]．系统仿真学报，（2026-03-25）[2026-04-01]．https：//doi．org/10．16182/j．issn1004731x．joss．26-0023． SONG J，SUN H X，LI D Y，et al．Vertical active force control simulation of deployment and offloading system for large space mechanism[J/OL]．Journal of System Simulation，（2026-03-25）[2026-04-01]．https：//doi．org/10．16182/j．issn1004731x．joss．26-0023（in Chinese）．
[22] ZHAN L,XU X,BAI Z,et al.Intelligent compensation for uncertain time delay in vehicle magnetorheological suspension control using predictive experience[J].Control Engineering Practice,2026,169:106796.
[23] WU C,ZHAO P,CHEN P,et al.Design,dynamic modelling and experimental study on a tether-net system for active debris removal[J].Mechanism and Machine Theory,2025,208:105958.
[24] SCHULMAN J,WOLSKI F,DHARIWAL P,et al.Proximal policy optimization algorithms[J/OL].arXiv preprint arXiv:1707.06347,2017.[2026-04-01].https://arxiv.org/abs/1707.06347.
[25] TSURUMINE Y,CUI Y,UCHIBE E,et al.Deep reinforcement learning with smooth policy update: application to robotic cloth manipulation[J].Robotics and Autonomous Systems,2019,112:72-83.
[26] COUMANS E,BAI Y.PyBullet,a Python module for physics simulation for games,robotics and machine learning[EB/OL].(2016-01-01)[2024-04-01].http://pybullet.org.
[27] ANDRYCHOWICZ M,BAKER B,CHOCIEJ M,et al.Learning dexterous in-hand manipulation[J].The International Journal of Robotics Research,2020,39(1):3-20.
[28] PERLIN K.An image synthesizer[J].ACM SIGGRAPH Computer Graphics,1985,19(3):287-296.
[29] BENGIO Y,LOURADOUR J,COLLOBERT R,et al.Curriculum learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning(ICML).New York:ACM,2009:41-48.
[30] MNIH V,KAVUKCUOGLU K,SILVER D,et al.Human-level control through deep reinforcement learning[J].Nature,2015,518(7540):529-533.
[31] KINGMA D P,BA J.Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning

Representations(ICLR).San Diego:ICLR,2015:1-15.
[32] CHEN C,TANG H,HAO J,et al.Addressing action oscillations through learning policy inertia[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Palo Alto:AAAI Press,2021,35(8):7041-7049.
[33] TOBIN J,FONG R,RAY A,et al.Domain randomization for transferring deep neural networks from simulation to the real world[C]//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).Piscataway:IEEE,2017:23-30.