doi: 10.3969/j.issn.1672-6073.2026.04.013

融合大模型与近端策略优化的列车自动驾驶智能控制研究

高 凯[1,2]，朱 力[1]，苗 佳[1]，刘 磊[1]
（1. 北京交通大学自主运行技术研究所，北京 100044；2. 广西交控智维科技发展有限公司，南宁 530200）

摘 要：在城市轨道交通智能化发展背景下，列车自动驾驶需在节能与准点性之间实现有效权衡，而现有调度规章多以自然语言形式表达，难以直接用于控制建模。为此，本文提出一种融合大语言模型(large language model，LLM)与近端策略优化(proximal policy optimization，PPO)的列车自动驾驶控制方法。该方法利用 LLM 对调度规章与运行经验进行语义解析，并在知识图谱(knowledge graph，KG)约束下自动构建包含状态空间、动作集合、奖励函数与安全约束的马尔可夫决策过程(markov decision process，MDP)模型。在此基础上，引入动作屏蔽机制与奖励权重自适应调节策略，采用 PPO 算法在仿真环境中训练安全可控的自动驾驶策略。仿真结果表明：与传统规则控制、模型预测控制(model predictive control，MPC)和深度 Q 网络(deep Q-network，DQN)方法相比，本文方法的平均牵引能耗分别降低约 2.5%、1.5%和 1.0%，平均晚点时间分别降低约 13.7%、19.6%和 16.3%，且平均 jerk 值降至 0.59m/s^3，验证了 LLM 语义建模与强化学习相结合在列车自动驾驶控制中的有效性。进一步的半实物仿真结果表明，知识图谱校验能够降低语义解析中的规则遗漏与物理不一致风险，动作屏蔽机制可将限速和安全间隔等硬约束嵌入策略执行过程，奖励权重调节机制能够依据能耗、晚点和舒适度指标动态调整多目标权衡。所提方法形成了“语义建模—策略优化—反馈调节”的闭环，为自然语言调度知识参与列车自动驾驶控制建模提供了可解释的实现思路。

关键词：轨道交通；列车自动驾驶；大语言模型；近端策略优化；语义建模；多目标控制
中图分类号：U231 **文献标志码：**A **文章编号：**1672-6073(2026)04-0114-09

An Intelligent Automatic Train OperationTrain Control Method Integrating Large Language Models and Proximal Policy Optimization

Gao Kai[1,2], Zhu Li[1], Miao Jia[1], Liu Lei[1]
(1. Institute of Autonomous Operation Technology, Beijing Jiaotong University, Beijing 100044;
2. Guangxi Traffic Control Zhiwei Technology Development Co., Ltd., Nanning 530200)

Abstract: With advancement of intelligent urban rail transit, automatic train operation needs to achieve an effective balance between energy efficiency and punctuality. However, existing dispatching regulations are mostly expressed in natural language, making them difficult to directly incorporate into control modeling. To address this problem, this paper proposes an automated train operation control method integrating large language models (LLMs) and proximal policy optimization (PPO). The proposed method employs an LLMs to perform semantic parsing of dispatching regulations and operational experience, and automatically constructs a Markov decision process (MDP) model under the constraints of a knowledge graph (KG). The generated MDP includes the state space, action set, reward function, and safety constraints. On this basis, an action masking mechanism and an adaptive reward-weight adjustment strategy are introduced, and PPO is used to train a safe and controllable automated train operation policy in a simulation environment. Simulation results show that, compared with conventional rule-based control, Model Predictive Control (MPC), and Deep Q-Network (DQN), the proposed method reduces the average traction energy consumption by approximately 2.5%, 1.5%, and 1.0%, respectively, and reduces the average delay time by approximately 13.7%, 19.6%, and 16.3%, respectively. In addition, the average jerk is reduced to 0.59 m/s^3. These results verify the effectiveness of combining LLM-based semantic modeling with reinforcement learning for automated train operation control. Further hardware-in-the-loop tests indicate that KG-based consistency checking can reduce rule omissions and physical inconsistencies in semantic

收稿日期：2025-07-21 修回日期：2026-04-13
第一作者：高凯，男，博士研究生，从事智能列车控制、智能运维等研究，25115018@bjtu.edu.cn
基金项目：北京市联合基金(L251011)
引用格式：高凯，朱力，苗佳，等. 融合大模型与近端策略优化的列车自动驾驶智能控制研究[J]. 都市快轨交通，2026，39(4)：114-122.

parsing. The action masking mechanism embeds hard constraints, including speed limits and safe headway distances, into policy execution, while the adaptive reward-weight mechanism adjusts the trade-off among energy consumption, punctuality, and ride comfort according to operational feedback. The proposed method therefore establishes a closed-loop framework of semantic modeling, policy optimization, and feedback adjustment, providing an interpretable approach for incorporating natural-language dispatching knowledge into automated train operation control.

Keywords: rail transit; automated train operation; large language model; proximal policy optimization; semantic modeling; multi-objective control

在城市轨道交通快速发展的背景下，列车自动驾驶(automatic train operation，ATO)系统正由辅助驾驶逐步转向全自动运行，成为提升运输效率、降低能耗和保障行车安全的关键技术。随着线路拓展与客流密度的上升，列车控制系统不仅需确保安全与准点，还需实现能源的高效利用，因此其面临节能与准点目标之间的固有冲突[1]：节能运行倾向于增加惰行、减少牵引频次，而准点控制则要求严格按计划时间运行，二者的权衡使控制问题呈现出多目标、多约束与动态耦合等复杂特征。

当前实际运营中，ATO 多采用“比例-积分-微分(proportional-integral-derivative，PID)控制器+查表策略”模式[2-3]。在调试阶段，系统预先生成目标速度曲线；列车运行过程中，列车自动监控系统(automatic train supervision，ATS)下发时分信息，ATO 根据时分要求查表确定参考速度，并通过 PID 控制器实现实时跟踪控制。该方式实现简便、鲁棒性强，但灵活性与智能化程度有限。近年来，模型预测控制(model predictive control，MPC)在研究层面获得广泛关注，该方法能在多变量约束下优化运行曲线[4]，但受限于建模精度、算力和传感器等因素，难以规模化应用于实际线路。

深度强化学习(deep reinforcement learning，DRL)作为新兴方案[5]，已在仿真环境中展示出较强的策略学习与环境适应能力，以深度 Q 网络(deep Q-network，DQN)[6]、深度确定性策略梯度(deep deterministic policy gradient，DDPG)[7]和近端策略优化(proximal policy optimization，PPO)[8]为代表的强化学习算法，能够自动学习列车节能运行策略，应对复杂扰动。然而，DRL 仍存在样本效率低、奖励函数依赖人工设计、可解释性与安全性欠缺等问题。

针对现有列车自动驾驶方法在多目标冲突建模、调度知识利用以及策略安全性与适应性方面存在的不足，本文提出一种融合大语言模型(large language model，LLM)[9]与近端策略优化(proximal policy optimization，PPO)的智能列车自动驾驶控制方法。该方法以自然语言调度规章和运行经验为输入，在领域知识图谱约束下利用 LLM 进行语义解析，自动构建满足物理与安全约束的马尔可夫决策过程(markov decision process，MDP)模型，从而减少强化学习中状态、奖励与约束的人工设计依赖。在此基础上，引入带有动作屏蔽机制的 PPO 策略优化框架，在保障运行安全的前提下实现节能与准点等多目标协同优化；同时结合运行反馈动态调整奖励权重，使控制策略能够适应不同时段和运行条件的调度需求，最终形成“语义建模—策略优化—反馈调节”的闭环控制框架，为 LLM 与强化学习在轨道交通自动驾驶中的工程化应用提供一种可解释、可部署的解决思路。

1 列车自动驾驶控制问题建模

列车自动驾驶控制需在满足限速、追踪间隔与到站时间窗等安全约束的前提下，兼顾节能效率、准点性与乘坐舒适性。为支持强化学习策略求解，本文对该控制过程进行形式化建模，统一刻画状态、动作、奖励与约束要素[10]。

1.1 自动驾驶控制的马尔可夫建模基础

为便于采用强化学习方法对列车自动驾驶控制策略进行求解，本文将列车运行控制过程建模为 MDP[11]，该决策过程如图 1 所示。s_t 表示当前时刻状态，s_{t+1} 表示下一时刻 t+1 状态；r_t 表示当前时刻 t 的即时奖励，r_{t+1} 表示下一时刻 t+1 的即时奖励；a_t 表示当前时刻执行的动作。在控制周期内，状态转移由当前状态与动作决定，扰动项并入状态向量。

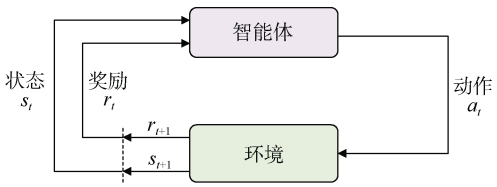

图 1　马尔可夫决策过程

Figure 1　Markov decision process

在此基础上，列车控制过程可抽象为有限时域的马尔可夫决策过程，其五元组定义为

$$M=(S,A,P,R,\mathrm{C}) \tag{1}$$

式中，S 为状态空间；A 为动作集合；$s\in S$ 表示当前状态；$s'\in S$ 表示执行动作后的下一状态；$a\in A$ 表示

当前选择的控制动作；$P(s'\mid s,a)$ 为状态转移概率，表示在状态 s 下执行动作 a 后转移到状态 s'的概率；$R(s,a)$ 为奖励函数，用于评价在状态 s 下执行动作 a 所获得的即时收益；C 为约束条件。

智能体的训练目标是在每一时间步 t 下选择最优动作 $a_t \in A$，使得从初始状态 s_0 出发，累积获得的期望折扣回报最大，目标函数为

$$J(\pi)=\mathbb{E}_{\pi}\left[\sum_{t=0}^{T}\gamma^{t}R(s_t,a_t)\right] \tag{2}$$

式中，π 表示智能体的控制策略，$J(\pi)$表示策略 π 对应的期望累积回报；$\mathbb{E}_{\pi}[\cdot]$表示在策略 π 下对状态转移轨迹求期望；T 为有限时域内的最大决策步数；$\gamma\in[0,1]$ 为折扣因子，用于衡量未来奖励在当前目标中的权重。

1.2 状态、动作与约束设计

在强化学习框架下，状态空间用于刻画列车运行的关键动态与环境信息。本文采用连续状态向量描述列车运行状态，其维度为 5，定义为

$$s_t=[x_t,v_t,\Delta t_t,\rho_t,\lambda_t] \tag{3}$$

式中，x_t 为当前列车所在位置；v_t 为列车当前速度；$\Delta t_t=t_t-t_{\text{schedule}}(x_t)$ 为当前时刻 t_t 与计划时刻表时刻 $t_{\text{schedule}}(x_t)$ 的偏差；ρ_t 为当前牵引变电所电网负荷；λ_t 表示当前车厢客流密度。因此，状态空间可表示为 $S\subset\mathbb{R}^5$，无需离散化处理，通过策略网络进行函数逼近以适应连续状态空间下的优化问题。

在动作空间设计方面，考虑到实际列控系统中牵引与制动通常采用分级控制方式，本文将控制动作 A 离散为 4 类，分别对应无功率、30%牵引、100%牵引与标准制动，表示为

$$A=\{\text{惰行},\text{轻牵引},\text{全牵引},\text{制动}\}\triangleq\{a_0,a_1,a_2,a_3\} \tag{4}$$

上述动作 a_0～a_3 编号仅作为强化学习策略的离散动作编码，不直接表示牵引力或制动力的物理大小。

为约束状态转移的物理一致性，引入简化的一维列车运行动力学模型：

$$v_{t+1}=v_t+\frac{1}{m}\left(F_{\text{traction}}-F_{\text{drag}}-F_{\text{break}}\right)\cdot\Delta t \tag{5}$$

$$x_{t+1}=x_t+v_{t+1}\cdot\Delta t \tag{6}$$

式中，m 为列车总质量；F_{traction} 为由动作 a_t 决定的牵引力；F_{drag} 为空气与滚阻阻力；F_{break} 为纵断面坡度引起的附加阻力。该模型用于状态更新函数 $s_{t+1}=f(s_t,a_t)$。

为保证控制策略的工程可部署性，本文在训练过程中引入 4 个约束。

① 限速约束：$v_t\leqslant v_{\text{limit}}(x_t)$

② 追踪间隔：$x_t^{(i)}-x_t^{(i-1)}\geqslant D_{\text{safe}}$

③ 到站时间窗：$t_{\text{arr}}^{(j)}\in\left[t_{\min}^{(j)},t_{\max}^{(j)}\right]$

④ 动作平滑：$|a_t-a_{t-1}|\leqslant a_{\max\_\text{delta}}$

式中，$v_{\text{limit}}(x_t)$ 表示列车在位置 x_t 处对应的线路限速；i 表示列车编号，$x_t^{(i)}$ 和 $x_t^{(i-1)}$ 分别表示第 i 列车及其前行列车在时刻 t 的位置；D_{safe} 表示最小安全追踪间隔；j 表示车站编号，$t_{\text{arr}}^{(j)}$ 表示列车到达第 j 个车站的实际时刻，$t_{\min}^{(j)}$ 和 $t_{\max}^{(j)}$ 分别表示该车站允许到达时间窗的下限和上限；a_{t-1} 表示上一时间步的控制动作，$a_{\max\_\text{delta}}$ 表示相邻两个时间步之间允许的最大动作变化幅度。

1.3 多目标优化问题构建

在列车自动驾驶场景中，控制策略需在节能、准点性与乘客舒适性等目标之间进行权衡。为此，本文采用加权多目标奖励函数对上述目标进行统一建模，其形式为

$$\begin{aligned}&R(s_t,a_t)=\\&-\left(\alpha\cdot E_{\text{traction}}(t)-\beta\cdot E_{\text{regen}}(t)+\tau\cdot|\Delta t_t|+\delta\cdot J_t\right)\end{aligned} \tag{7}$$

式中，$E_{\text{traction}}(t)$ 为牵引能耗；$E_{\text{regen}}(t)$ 为再生制动回收能量；Δt_t 为运行时刻偏差；J_t 为列车瞬时冲击指标(jerk)，用于间接反映乘客舒适度；α,β,τ,δ 为对应权重参数。

考虑到不同运行时段与调度策略下目标侧重可能变化，本文在任务尺度上统计关键绩效指标(key performance indicators，KPI)，并将其用于奖励权重调整，奖励权重可以表示为

$$\boldsymbol{w_n}=[\alpha_n,\beta_n,\tau_n,\delta_n] \tag{8}$$

其更新由上层 LLM 模块依据运行反馈 KPI 进行调节，对应表达式为

$$\boldsymbol{w_{n+1}}=\text{LLM}_{\text{update}}(\theta_n,\mathbf{KPI}_n) \tag{9}$$

式中，$\mathbf{KPI}_n=\left[K_E^{(n)},K_D^{(n)},K_J^{(n)}\right]^{\mathrm{T}}$ 表示第 n 次评估周期统计得到的 KPI 向量，其中上标 T 表示转置。$K_E^{(n)}$、$K_D^{(n)}$、$K_J^{(n)}$ 分别表示累计牵引能耗、平均晚点偏差和

平均 jerk。

通过“时间步奖励—任务级反馈”的双层机制，策略在安全约束下实现多目标权衡。然而，实际调度规则多以自然语言形式存在，需引入 LLM 进行语义建模。

2 基于 LLM 的语义建模方法

针对自然语言调度规章难以直接用于形式化建模的问题，提出一种 LLM 与领域知识图谱(knowledge graph，KG)[12]的语义建模方法。该方法通过对调度文本进行语义解析与结构化映射，自动生成控制模型所需的状态变量、奖励项与约束条件，为后续控制策略设计提供统一、可解释的建模基础。

2.1 知识图谱辅助的任务语义理解框架

仅依赖 LLM 进行建模可能引入“幻觉”或物理不一致问题。为此，本文引入领域知识图谱作为结构化先验，用于约束和校验 LLM 的语义解析结果。知识图谱形式化定义为有向图，即

$$g=(v,\varepsilon) \tag{10}$$

式中，节点集合 $v=\{v_i\}$表示实体，如站点、轨道区段、速度限制等；边集合 $\varepsilon=\{e_{ij}\}$表示节点间的关系，如站点连通、限速作用于轨段等。通过图结构，知识图谱准确表达了列车运行环境的时空及约束关系。

在语义建模过程中，LLM 与知识图谱通过以下方式协同工作：LLM 从自然语言规章中抽取关键实体与规则；知识图谱用于对抽取结果进行语义对齐、补全与一致性校验。

通过上述机制，LLM 获得了受物理约束引导的语义理解能力，从而避免生成不符合工程实际的模型描述。二者的优缺点对比如表 1 所示。

表 1 大语言模型与知识图谱优缺点对比

Table 1 Comparison of Advantages and Disadvantages of Large Language Models and Knowledge Graphs

方法	优点	缺点
大语言模型	通用知识；语言处理；适应性强；自动建模	知识不全；语言理解局限；无法处理未见过的数据
知识图谱	结构化知识；领域专业知识；准确性；动态推理	隐形知识；局部推理；领域知识不全

2.2 基于 LLM-知识图谱的 MDP 模型自动生成

在完成任务语义理解后，本文进一步将解析结果映射为强化学习所需的 MDP，即

$$M=(S,A,P,R,C)$$

各元组的生成逻辑如下。

① 状态空间 S：由 LLM 从规章文本中抽取关键运行变量(如位置、速度、晚点量)，并由知识图谱提供其物理边界与关联关系，确保状态定义的完备性与可解释性。

② 动作空间 A：根据规章中对牵引与制动行为的描述，结合设备性能约束，自动生成离散控制动作集合。

③ 状态转移概率 P：在简化列车动力学模型基础上，结合知识图谱中的坡度与阻力参数，构建物理一致的状态转移函数，并引入噪声项描述运行不确定性。

④ 奖励函数 R：由 LLM 提取节能、准点与舒适性等目标语义，并映射为多目标加权奖励形式。

⑤ 约束条件 C：由知识图谱提供的限速、安全间隔等硬约束构成，用于限定策略的可行解空间。

整体语义建模流程如图 2 所示。采用分阶段提示实现：先抽取实体/规则，再由知识图谱进行对齐补全与一致性校验，最终输出结构化 MDP 描述。

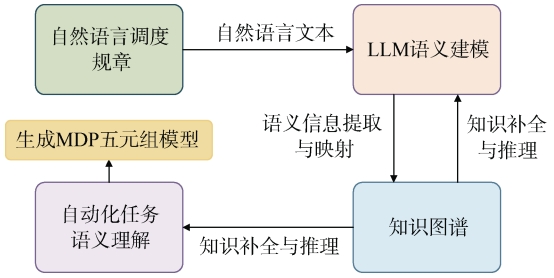

图 2 基于 LLM 的语义建模流程

Figure 2 Semantic modeling framework based on LLMs

3 PPO 驱动的列车自动驾驶策略优化

3.1 基于 LLM 的 PPO 策略优化框架

在 LLM 自动生成的马尔可夫决策过程 MDP 模型基础上，本文引入近端策略优化算法(proximal policy optimization，PPO)[13-14]作为列车控制策略求解器，以实现节能与准点目标的动态平衡。整体框架如图 3 所示。

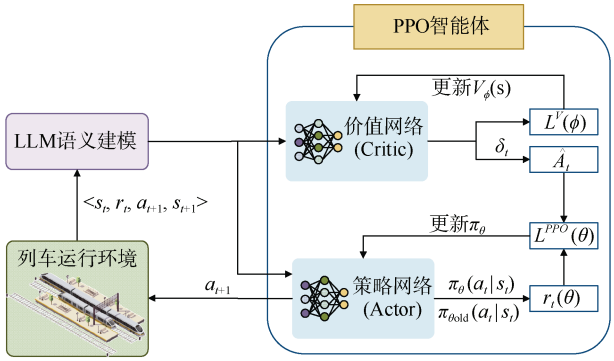

图 3 基于 LLM 的 PPO 策略优化框架

Figure 3 LLM-Based PPO Strategy Optimization Framework

在该框架中，LLM 负责将调度规章和运行经验解析为结构化的环境模型，包括状态变量、奖励项与安全约束；PPO 智能体则在该环境模型约束下与列车运行环境交互，并通过 Actor–Critic 结构完成策略优化。列车运行环境在执行控制动作 a_t 后，向 PPO 智能体反馈状态转移样本 $\langle s_t, a_t, r_t, s_{t+1} \rangle$。

PPO 采用 actor–critic 架构对列车控制策略进行建模。Actor 网络输出在状态 s_t 下各离散控制动作的概率分布，表达式为

$$\pi_\theta(a_t|s_t) = \mathrm{Softmax}\left(f_\theta^{\mathrm{actor}}(s_t)\right) \in \mathbb{R}^4 \tag{11}$$

式中，θ 表述 Actor 网络的参数；$f_\theta^{\mathrm{actor}}(s_t)$ 为 actor 网络对状态 s_t 进行处理的输出；Softmax 函数用于将 actor 网络输出的动作评分归一化为概率分布，使各离散动作的选择概率非负且总和为 1。

critic 网络负责估计状态价值函数 $V_\phi(s_t)$，表达式为

$$V_\phi(s_t) = f_\phi^{\mathrm{critic}}(s_t) \tag{12}$$

式中，ϕ表述 critic 网络的参数；$f_\phi^{\mathrm{critic}}(s_t)$ 为 critic 网络对状态 s_t 进行处理的输出。基于当前状态价值 $V_\phi(s_t)$ 和下一状态价值 $V_\phi(s_{t+1})$，可计算时序差分误差 δ_t，并进一步得到优势函数估计 $\hat{A}_t$，用于衡量动作 a_t 相对于当前策略平均水平的优劣。图 3 中的 $L^V(\phi)$表示 critic 网络的价值损失函数，用于更新价值网络参数；$L^{\mathrm{PPO}}(\theta)$ 表示 actor 网络的策略损失函数，用于更新策略网络参数。

为提升训练稳定性，本文采用基于广义优势估计(generalized advantage estimation，GAE)的 PPO 策略更新机制，其目标函数定义为

$$L^{\mathrm{PPO}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] \tag{13}$$

式中，θ 表示当前策略网络参数；$\hat{A}_t$ 是时刻 t 的优势函数；ϵ 是剪切因子。比率项 $r_t(\theta)$ 定义为新旧策略的比值 $r_t(\theta) = \dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)}$，用于衡量当前策略 $\pi_\theta(a_t \mid s_t)$ 与旧策略 $\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)$ 之间的差异。函数 $\mathrm{clip}(\cdot)$ 表示裁剪操作，其作用是将 $r_t(\theta)$ 限制在 $[1-\epsilon, 1+\epsilon]$ 范围内，以抑制单次策略更新幅度过大，从而提高训练稳定性。

3.2 安全约束建模与策略训练

列车运行任务中存在大量强制性物理与运营约束，如速度曲线限制、追踪间隔约束、停车时间窗与动作平滑性等。本文将这些约束分为“硬约束”与“软约束”，并分别设计了策略层面的处理机制。

对于硬约束(如限速与安全车距)，采用动作屏蔽机制，在每一状态 s 下动态构建可行动作集合 $A_{\mathrm{safe}}(s)$，并在策略输出阶段屏蔽不满足约束的动作，从而保证所有被采样和执行的控制指令均满足基本安全条件，基本安全条件为

$$\pi^*(a|s) = \mathrm{Softmax}\left(f_\theta^{\mathrm{actor}}(s) + \mathbf{mask}(s)\right) \tag{14}$$

式中，$\mathbf{mask}(s)$ 为与动作空间维度相同的屏蔽向量。对于满足安全约束的动作，其对应屏蔽值设为 0；对于不满足约束的动作，其对应屏蔽值设为一个极小值，如足够大的负数。经过 Softmax 归一化后，不可行动作的选择概率趋近于 0，从而避免策略采样到违反硬约束的控制动作。

对于软约束(如节能目标、运行正点性与乘客舒适度)，本文通过多目标奖励函数进行建模，并在策略训练过程中引入基于 KPI 的奖励权重自适应调整机制。通过周期性统计能耗、晚点与舒适性等运行指标，对奖励权重进行更新，使策略在不改变控制模型结构的前提下，能够根据运行反馈动态调整多目标权衡关系，从而提升长期运行性能与策略适应性。

在此基础上，通过 PPO 算法对大语言模型完成语义建模后生成的马尔可夫决策过程进行求解，搭建起语义建模、安全约束、策略优化、KPI 反馈相互衔接的闭环训练流程，具体包括 6 个核心步骤。

步骤 1：LLM 在知识图谱约束下解析调度规章文本，生成 MDP 模型；

步骤 2：初始化策略网络、价值网络与奖励权重；

步骤 3：在动作屏蔽约束下采样交互轨迹；

步骤 4：基于多目标奖励与 PPO 裁剪目标更新策略；

步骤 5：周期性统计运行级 KPI，并更新奖励权重；

步骤 6：重复迭代直至策略收敛。

3.3 安全攸关系统中的确定性与安全性保证机制说明

在 3.2 中，从策略优化层面介绍动作屏蔽、多目标奖励建模等机制，说明如何在 PPO 训练与执行过程中显式引入运行安全约束。然而，对于安全攸关的列

车控制系统而言，仅给出具体的约束建模方法仍不足以回答“该类智能方法在工程上如何保证安全可控性”的问题。为此，有必要进一步从系统结构与责任划分的角度，对本文方法的确定性与安全性保障机制进行说明。

需要指出的是，本文所提出的方法并不试图从理论上对概率性LLM或黑箱强化学习策略本身给出形式化的安全性证明，而更多关注的是在工程实践中，如何通过体系结构设计与多层约束机制，确保智能控制方法在安全攸关的列车控制场景下具备可接受的确定性与运行安全性。

LLM在本文框架中仅承担离线语义建模与模型结构生成的角色，用于将自然语言调度规章解析为结构化的状态、奖励与约束描述。LLM不参与在线决策与控制指令生成，其输出结果在进入策略训练之前需经过领域知识图谱的一致性校验与物理约束过滤。因此，LLM的概率性推理过程不会直接影响列车的实时控制行为。

PPO强化学习策略并非在无约束空间内进行优化。本文通过动作屏蔽机制，将速度限制、安全间隔、动作平滑性等硬安全约束显式嵌入到策略执行层，使得策略在任何状态下仅能选择满足安全条件的控制动作。该机制确保了策略探索与执行始终受限于物理与运营规则定义的安全可行域内，从而避免不安全控制指令被生成或执行。

综上，本文方法的安全性并不依赖于智能算法本身的确定性假设。系统通过融合离线语义建模、知识图谱校验、受约束策略优化及物理规则裁剪，建立了一套多层安全机制来保障工程可控性。这一设计采用结构性约束替代了对算法可证明性的要求，符合当前安全攸关智能系统的主流工程实践。

4 实验与结果分析

为验证所提方法在真实工程约束下的有效性，如图4所示，本文依托列车自动驾驶半实物仿真测试平台(hardware-in-the-loop，HIL)开展测试验证。该平台集成了真实的ATO智能控制单元、列车动力学实时仿真机以及标准的车地通信网络接口，能够在闭环条件下反映控制指令下发、状态反馈、通信延迟及物理执行等过程特性。

4.1 实验设置与评估指标

实验以图4所示的半实物闭环测试平台为基础开展，测试场景配置为北京地铁亦庄线典型区段。平台中加载了线路拓扑、纵断面、限速曲线、站点信息及运行图参数，其中相关场景要素由LLM辅助解析并结构化映射到控制模型中，用于构建列车运行环境。该测试区段全长24.5 km，包含12座车站，并考虑早高峰与平峰两类运行场景。列车动力学参数采用地铁B型车标准配置，主要参数如表2所示。

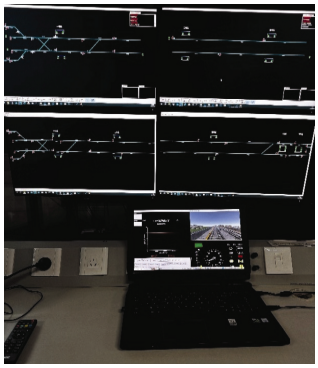

图4　列车自动驾驶半实物仿真测试平台

Figure 4　HIL Simulation Test Platform for ATO

表2　列车动力学参数设置

Table 2　Train Dynamics Parameters

参数名称	数值
列车编组形式	4M2T
最大牵引功率/kW	2 400
最大制动能力/kN	–220
列车质量/t	210
最大速度/(km/h)	80
控制周期/s	2
列车辅助功率/kW	120
Davis 阻力系数 A/kN	2.85
Davis 阻力系数 B/(kN/(km/h))	0.036
Davis 阻力系数 C/(kN/(km/h)^2)	0.001 2
旋转质量系数 $1+\rho$	1.08
牵引系统效率 η_t	0.88
再生制动吸收率 η_b	0.90
车地通信平均延迟/ms	200

在PPO训练中，折扣因子 γ 设为 0.98，GAE系数 λ 设为0.95，裁剪系数 ε 取 0.2；每轮更新采样1280条轨迹，并行环境数为8。上述参数均采用强化学习中常用设置，以保证训练稳定性与结果可复现性。

为体现方法的相对优势，选取规则表控制(Rule-Based)、模型预测控制(MPC)和深度Q网络(DQN)作为对比方法，并在统一运行条件下进行评估。性能指标包括：平均牵引能耗、全程平均晚点时间以及jerk

指标，分别对应节能性、准点性与运行平稳性。

4.2 实验结果与分析

基于实验室完整 ATO 系统的闭环运行对比结果如图 5 所示，即单站间距下本文方法与实际线路历史运行数据的速度-位置曲线对比。由图 5 可知：实际运行数据受限于传统 PID 控制的滞后性与人工经验，存在频繁的牵引与制动切换；而部署在真实 ATO 硬件上的 LLM+PPO 策略，能够利用线路坡度进行长距离惰行，且速度曲线更加平滑，未出现越过列车自动防护系统(automatic train protection，ATP)安全包络的现象。

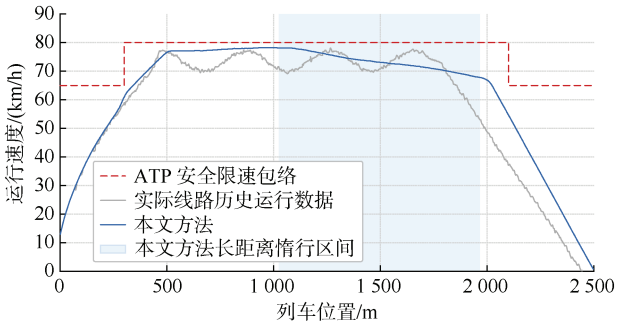

图 5 单站间距下的运行曲线对比

Figure 5 Comparison of operation curves over a single inter-station section

图 6 给出了不同控制策略下的综合性能对比。由图 6 可知：传统规则控制与实际运行数据在能耗与准点性上表现接近，整体能耗较高；MPC 与 DQN 在节能方面有一定改善，但在半实物平台的真实通信延迟扰动下，晚点时间控制不够稳定。本文提出的 LLM+PPO 方法平均牵引能耗为 66.8 kWh，较规则控制、模型预测控制和深度 Q 网络方法分别降低约 2.5%、1.5% 和 1.0%；列车到发时刻平均偏差控制在 8.2 s 以内，表明

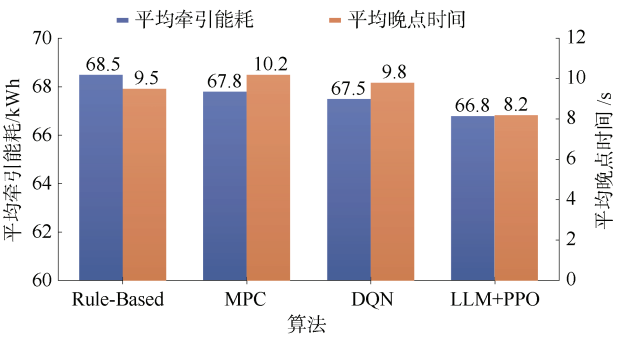

图 6 不同控制策略综合性能对比

Figure 6 Comprehensive performance comparison of different control strategies

该方法在本文设定测试条件下能够保持较稳定的准点性表现。

乘坐舒适性对比结果如图 7 所示。由图 7 可知：本文方法的平均 jerk 值为 0.59 m/s^3，低于实际运行数据的 0.85 m/s^3 与其他对比算法，表明该方法在兼顾节能与准点性的同时，有效抑制了真实物理执行过程中的加速度剧烈波动。综合上述实证结果，本文方法在工程部署环境下依然能够实现多目标约束的协调控制。

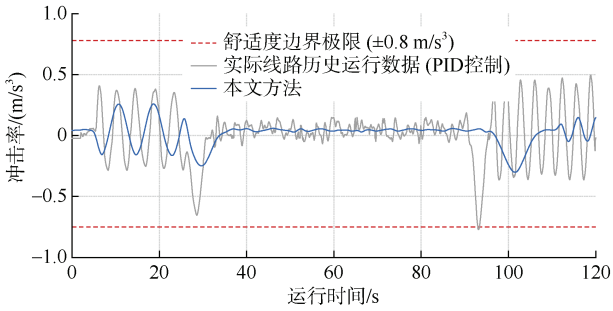

图 7 不同控制策略的乘坐舒适度对比

Figure 7 Comparison of Ride Comfort Under Different Control Strategies

4.3 安全性与可靠性验证

为验证所提方法在安全关键场景下的可靠性，本文从运行行为安全性与语义建模一致性两个层面进行评估。

4.3.1 运行层安全性验证

针对突发约束变化场景，设计安全边界压力测试，考察智能体在阶梯式限速下降条件下的控制行为。突发限速约束下不同方法的安全裕度对比结果如图 8 所示。由图 8 可知：传统软演员-评论家(soft actor-critic，SAC)方法在限速切换点出现了短时包络越界现象，而本文方法能够更早进行减速调整，其速度轨迹始终保持

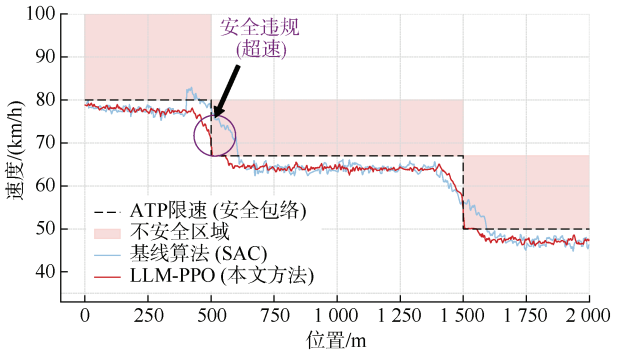

图 8 突发限速约束下不同方法的安全裕度对比

Figure 8 Comparison of Safety Margins under Sudden Speed-Limit Constraints

在 ATP 安全边界内。该结果说明，在本文构造的突发限速测试条件下，所提方法表现出较好的约束响应能力和安全裕度。

4.3.2 建模层可靠性验证

针对 LLM 自动建模可能引入的逻辑或物理不一致问题，引入 KG 对生成的 MDP 结构进行约束校验。不同语义建模配置下 MDP 生成质量对比如表 3 所示。表 3 中：语法正确率表示生成的 MDP 结构描述中可被解析并完整映射为状态、动作、奖励与约束字段的比例；逻辑一致性表示经规则模板与知识图谱关系校验后不存在语义冲突的比例；物理约束满足率表示生成模型在动力学边界、限速约束及动作可执行性检查下通过校验的比例。结果表明：在未引入 KG 约束时，生成结果在逻辑一致性和物理约束满足性方面结果质量不高；引入 KG 后，语法正确率、逻辑一致性和物理约束满足率均有所提升，说明知识图谱约束有助于提高结构化建模结果的规范性、一致性与可执行性。

表 3 不同语义建模配置下 MDP 生成质量对比

Table 3 Comparison of MDP Generation Quality under Different Semantic Modeling Configurations %

模型配置	语法正确率	逻辑一致性	物理约束满足率
LLM	88.5	74.2	62.0
LLM+KG	99.8	99.5	99.7

综上，在本文设置的测试平台和场景下，所提方法在运行层与建模层均表现出较好的安全性和一致性，显示出其在安全关键列车自动驾驶控制场景中的应用潜力。

4.4 消融实验分析

为分析各关键模块对整体性能的影响，从建模方式、优化算法及核心机制 3 个层面开展消融实验。表 4 给出了不同建模方式与强化学习算法的定量对比结果。由表 4 可知：在本文设定的测试条件下，LLM+PPO 在平均牵引能耗和晚点时间两项指标上均取得了相对较低的数值；与人工建模+PPO 相比，其综合表现有所改善，说明基于 LLM 的结构化建模方式有助于提升控制模型的表达完整性与任务适配性；在相同 LLM 建模基础上，PPO 相较于 TRPO 和 SAC 也呈现出较优的综合性能。

表 4 不同建模方式与强化学习算法的性能对比

Table 4 Performance Comparison of Different Modeling Approaches and Reinforcement Learning Algorithms

消融变量	能耗/kWh	晚点/s
人工建模+PPO	67.5	9.6
LLM+PPO	66.8	8.2
LLM+TRPO	67.2	9.1
LLM+SAC	67.0	8.6

不同强化学习算法在训练过程中平均牵引能耗的变化情况如图 9 所示。由练曲线可以观察到，PPO 在本文设置下条件表现出较快的收敛趋势。

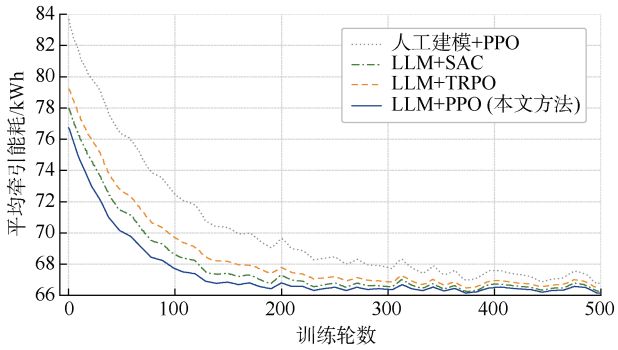

图 9 不同强化学习算法的训练收敛过程对比

Figure 9 Comparison of training convergence among different reinforcement learning algorithms

分别移除动作屏蔽、知识图谱约束和动态权重调优机制，并比较其对训练过程和最终性能的影响，结果如图 10 所示。由图 10 可知：移除动作屏蔽后，训练稳定性和安全约束满足情况均有所下降；移除知识图谱约束后，模型在训练初期的学习效率有所减弱；移除动态权重调优机制后，训练后期的性能提升幅度受到一定影响。

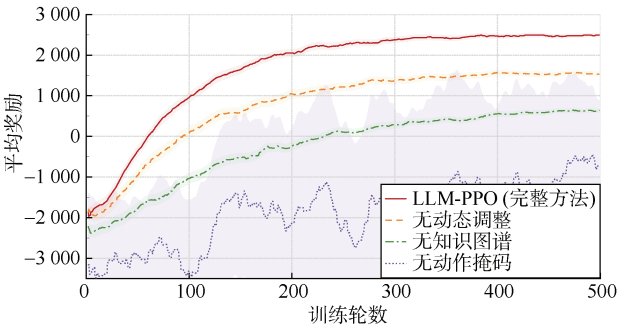

图 10 消融实验中奖励函数的收敛过程

Figure 10 Convergence of Reward Functions in Ablation Experiments

综上，LLM 语义建模、PPO 优化算法及其配套机制在本文方法中具有互补作用，共同影响最终控制效果。

5 结论

本文提出了一种融合 LLM 与 PPO 的列车自动驾驶节能—准点协同控制方法。通过对自然语言调度规章进行语义建模，自动生成马尔可夫决策过程，并结合动作屏蔽与奖励权重自适应调节机制，在满足安全约束的前提下实现多目标优化。基于半实物闭环测试平台及给定运行图条件的验证结果表明：

1) 所提方法的牵引能耗较规则控制方法降低 2.5%，较模型预测控制和深度 Q 网络方法分别降低 1.5%和 1.0%；

2) 列车到发时刻平均偏差控制在 8.2 s 以内，准点性表现稳定；

3) 运行过程中的 jerk 指标保持在 0.59 m/s^3，乘坐舒适性优于对比方法。

在面向未来工程化应用方面，考虑到城市轨道交通对安全性和可靠性的严格要求，在开展更高等级验证之前，仍需围绕硬件接口联调、复杂工况覆盖测试以及与既有系统的适配性开展进一步研究。同时，本文以单列车智能控制为切入点，为后续面向高密度运行条件下的多列车协同控制与网络级调度优化研究提供了可借鉴的建模与优化思路。未来研究将重点关注大模型与强化学习策略的端侧轻量化部署，依托更高级别的硬件在环测试台，持续评估该方法在复杂路网和多场景条件下的适用性与工程可行性。

参考文献

[1] 孟建军, 裴明高, 武福, 等. 城轨列车多目标优化控制算法研究与仿真[J]. 系统仿真学报, 2017, 29(3): 581-588.
Meng Jianjun, Pei Minggao, Wu Fu, et al. Research and simulation on control algorithm for multi-objective optimization of urban rail train[J]. Journal of System Simulation, 2017, 29(3): 581-588.

[2] 侯涛, 兰小斌. 基于改进 PID 搜索算法的列车节能运行优化研究[J]. 重庆交通大学学报(自然科学版), 2025, 44(5): 112-120.
Hou Tao, Lan Xiaobin. Energy-saving operation optimization of train operation based on improved PID search algorithm[J]. Journal of Chongqing Jiaotong University (Natural Science), 2025, 44(5): 112-120.

[3] 徐梓航. 地铁自动驾驶系统优化与控制技术研究[J]. 人民公交, 2025(6): 73-75.

[4] 贾超, 徐洪泽, 王龙生. 基于多质点模型的列车自动驾驶非线性模型预测控制[J]. 吉林大学学报(工学版), 2020, 50(5): 1913-1922.
Jia Chao, Xu Hongze, Wang Longsheng. Nonlinear model predictive control for automatic train operation based on multi-point model[J]. Journal of Jilin University (Engineering and Technology Edition), 2020, 50(5): 1913-1922.

[5] 金则灵. 基于深度强化学习的城轨列车 ATO 智能控制策略研究[D]. 兰州: 兰州交通大学, 2022.
Jin Zelin. Intelligent ATO control strategy for urban rail trains based on deep reinforcement learning[D]. Lanzhou: Lanzhou Jiaotong University, 2022.

[6] 宿帅, 朱擎阳, 魏庆来, 等. 基于 DQN 的列车节能驾驶控制方法[J]. 智能科学与技术学报, 2020, 2(4): 372-384.
Su Shuai, Zhu Qingyang, Wei Qinglai, et al. A DQN-based approach for energy-efficient train driving control[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 372-384.

[7] 武晓春, 金则灵. 基于 DDPG 算法的列车节能控制策略研究[J]. 铁道科学与工程学报, 2023, 20(2): 483-493.
Wu Xiaochun, Jin Zeling. Research on train energy saving control strategy based on DDPG algorithm[J]. Journal of Railway Science and Engineering, 2023, 20(2): 483-493.

[8] 刘清山. 基于模仿与强化协作学习的列车智能驾驶研究[D]. 重庆: 重庆交通大学, 2023.
Liu Qingshan. Research on intelligent train driving based on cooperative imitation and reinforcement learning[D]. Chongqing: Chongqing Jiaotong University, 2023.

[9] Bian R, Geng Y, Yang Z J, et al. AutoMathKG: the automated mathematical knowledge graph based on LLM and vector database[J]. Computational Intelligence, 2025, 41(4): e70096.

[10] Pang H, Wang Z P, Li G Q. Large language model guided deep reinforcement learning for decision making in autonomous driving[EB/OL]. 2024. arXiv: 2412.18511. https://arxiv.org/abs/2412.18511.

[11] 褚端峰, 王如康, 王竞一, 等. 端到端自动驾驶的研究进展及挑战[J]. 中国公路学报, 2024, 37(10): 209-232.
Chu Duanfeng, Wang Rukang, Wang Jingyi, et al. End-to-end autonomous driving: advancements and challenges[J]. China Journal of Highway and Transport, 2024, 37(10): 209-232.

[12] 唐晓晟, 程琳雅, 张春红, 等. 大语言模型在学科知识图谱自动化构建上的应用[J]. 北京邮电大学学报(社会科学版), 2024, 26(1): 125-136.
Tang Xiaosheng, Cheng Linya, Zhang Chunhong, et al. Application of large language models in automated construction of knowledge graphs for university subject doains[J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2024, 26(1): 125-136.

[13] 陈沁鑫, 秦斌, 王欣. 基于改进 PPO 算法的城轨能量管理策略[J]. 电工技术, 2025(1): 41-45.
Chen Qinxin, Qin Bin, Wang Xin. Urban rail energy management strategy based on improved PPO algorithm[J]. Electric Engineering, 2025(1): 41-45.

[14] 金彦亮, 范宝荣, 高塬, 等. 基于元学习和强化学习的自动驾驶算法[J]. 应用科学学报, 2024, 42(5): 795-809.
Jin Yanliang, Fan Baorong, Gao Yuan, et al. Autonomous driving algorithm based on meta-learning and reinforcement learning[J]. Journal of Applied Sciences, 2024, 42(5): 795-809.

（编辑：王艳菊）