国 家 知 识 产 权 局

610000

成都高新区天府大道中段1388号1栋7层747号 成都巾帼知识产权代理有限公司

邢伟(028-86116992)

发文日：

2026年03月26日

申请号：202610003269.8　　**发文序号：**2026032600415100

申请人：香港中文大学（深圳）,深圳市未来智联网络研究院

发明创造名称：基于大模型智能体的跨环境泛化网络流量优化方法

授 予 发 明 专 利 权 通 知 书

1.根据专利法第39条及实施细则第60条的规定，上述发明专利申请经实质审查，没有发现驳回理由，现作出授予专利权的通知。

申请人收到本通知书后，还应当依照办理登记手续通知书的内容办理登记手续。

申请人按期办理登记手续后，国家知识产权局将作出授予专利权的决定，颁发发明专利证书，并予以登记和公告。

期满未办理登记手续的，视为放弃取得专利权的权利。

法律、行政法规规定相应技术的实施应当办理批准、登记等手续的，应依照其规定办理。

2.授予专利权的上述发明专利申请是以下列申请文件为基础的：

☐原始申请文件。☐分案申请递交日提交的文件。☒下列申请文件：

申请日提交的摘要附图； 2026年2月28日提交的说明书第1-97段、说明书附图、说明书摘要； 2026年3月12日提交的权利要求第1-5项。

3.授予专利权的上述发明专利申请的名称：

☒未变更。

☐由_____变更为上述名称。

4. ☐申请人于_____年_____月_____日提交专利号为_____的“放弃专利权声明”，经审查：

☐进入放弃专利权的程序。

☐未进入放弃专利权的程序。

5. ☐审查员依职权对申请文件修改如下：

6. ☐申请人在申请日后补交了实验数据，该数据未包含在授权公告文本中。

注：在本通知书发出后收到的申请人主动修改的申请文件，不予考虑。

国家知识产权局 专利审查业务章 1101081359134

审查员：陈晨　　审查部门：专利审查协作广东中心

联系电话：020-28950737

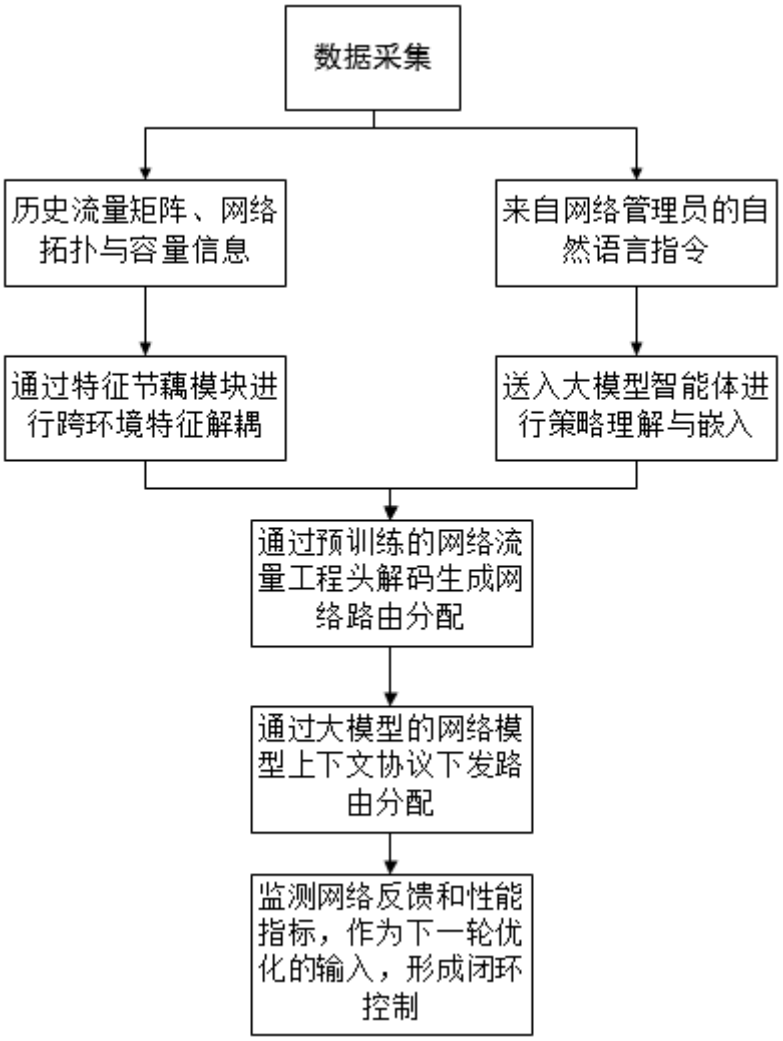
纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路6号 国家知识产权局专利局受理处收

电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

说 明 书 摘 要

本发明涉及网络流量优化领域，具体涉及一种基于大模型智能体的跨环境泛化网络流量优化方法。技术方案包括：数据采集，采集历史流量矩阵、网络拓扑与容量信息以及来自网络管理员的自然语言指令；将采集到的历史流量矩阵、网络拓扑与容量信息送入特征节藕模块进行跨环境特征解耦，同时将自然语言指令送入大模型智能体进行策略理解与嵌入；然后通过预训练的网络流量工程头解码生成网络路由分配；最后通过大模型的网络模型上下文协议下发路由分配，并监测网络反馈和性能指标，作为下一轮优化的输入，形成闭环控制。本发明适用于网络流量优化。

摘 要 附 图


数据采集
历史流量矩阵、网络拓扑与容量信息
来自网络管理员的自然语言指令
通过特征节藕模块进行跨环境特征解耦
送入大模型智能体进行策略理解与嵌入
通过预训练的网络流量工程头解码生成网络路由分配
通过大模型的网络模型上下文协议下发路由分配
监测网络反馈和性能指标，作为下一轮优化的输入，形成闭环控制

权 利 要 求 书

1. 基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，包括：

S1、数据采集与输入，包括历史流量矩阵、网络拓扑与容量信息以及来自网络管理员的自然语言指令；

S2、跨环境特征解耦，将采集到的历史流量矩阵、网络拓扑与容量信息送入特征节藕模块进行跨环境特征解耦；

S3、将自然语言指令送入大模型智能体进行策略理解与嵌入；

S4、通过预训练的网络流量工程头解码生成网络路由分配；

S5、通过大模型的网络模型上下文协议下发路由分配，并监测网络反馈和性能指标，作为下一轮优化的输入，形成闭环控制。

2. 根据权利要求 1 所述的基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，步骤 S1 具体包括：

S101、历史流量矩阵采集，采集历史流量窗口 $H_{sd} = [D_{sd}^{t-L+1}, \dots, D_{sd}^{t}]$，其中 D_{sd} 表示从源节点 s 到目的节点d的流量需求；

S102、网络拓扑与容量信息获取，获取网络拓扑图 $G = (V, E, c)$，其中 V 为节点集合，E 为链路集合，$c: E \to R^{+}$ 为链路容量函数；为每个源-目的对(s, d) 生成候选路径集合$\mathcal{P}_{sd}$ ；

S103、自然语言策略指令接收，接收网络管理员的自然语言指令，指令被送入大模型智能体进行语义解析。

3. 根据权利要求 2 所述的基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，步骤 S2 中，跨环境特征解耦具体包括：

S201、计算相对流量矩阵，对流量矩阵进行归一化处理：

$\widehat{\mathrm{D}}_{\mathrm{sd}} = \frac{D_{sd}}{\max_{s',d' \in V} D_{s'd'}} \quad \forall s, \mathrm{d} \in V$，D 表示流量需求矩阵，归一化后的流量值 $\widehat{D}_{\mathrm{sd}} \in [0,1]$，提取流量分布模式，消除绝对量级影响，使模型对流量绝对量级变化具有鲁棒性；

S202、计算相对容量矩阵，对链路容量进行归一化处理：$\hat{c}(e) = \frac{c(e)}{\max_{e' \in E} c(e')} \quad \forall e \in E$，归一化后的容量值 $\hat{c}(e) \in [0,1]$，提取拓扑无关的容量模式，使模型适应从 Gbps 级广域网到 Tbps 级数据中心的不同规模网络；

S203、生成容量-流量相对系数，计算量级比例因子：$\varphi = \frac{\max_{s,d \in V} D_{sd}}{\max_{e \in E} c(e)}$，保留规模信息以支持最大链路利用率之外的多目标优化。

4. 根据权利要求 1 所述的基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，步骤 S3 具体包括：

S301、网络模型上下文协议驱动的智能体架构初始化；

采用预训练大语言模型作为推理引擎，通过网络模型上下文协议与网络系统进行结构化交互，使智能体具备网络抽象感知、工具调用、多步推理和反馈学习能力；

S302、三维策略嵌入空间映射；

将自然语言指令分解并映射到三维向量$s = [s_h, s_r, s_c] \in [0,1]^3$，三维向量包含三个维度：

历史依赖性s_h，控制对历史流量变化的响应敏感度，$s_h \in [0,1]$，其中 $s_h = 0$表示忽略历史波动，$s_h = 1$表示高度响应历史变化为突发流量预留裕度；

全局鲁棒性s_r，控制对未预见流量突发的弹性能力，$s_r \in [0,1]$，其中 $s_r = 0$为尽力而为策略，$s_r = 1$为最坏情况鲁棒性；

成本敏感性s_c，控制性能与成本之间的权衡偏好，$s_c \in [0,1]$，其中 $s_c = 0$为性能优先不考虑成本，$s_c = 1$为成本优先可接受一定性能折损；

S303、Few-shot 小样本提示工程，为大语言模型提供示例指导策略翻译，通过思维链技术引导逐步推理：识别优化目标、提取约束条件、量化目标权重、映射到三维策略空间以及生成策略嵌入向量；

S304、策略验证与传递，对生成的策略嵌入进行合理性验证，如发现歧义或冲突则触发多轮对话澄清，验证通过后将策略嵌入向量 s 下发给预训练的网络流量工程头，所述合理性验证包含范围检查、冲突检测以及历史一致性检查。

5. 根据权利要求 2 所述的基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，步骤 S4 具体包括：

S401、特征融合与编码；

通过图神经网络进行容量特征编码，首先进行边特征编码 $h_e = \mathrm{EdgeEncoder}(A, \hat{C})$，其中 A 为邻接矩阵，$\hat{C}$ 为归一化容量矩阵，然后进行路径特征聚合 $g_p = \mathrm{PathEncoder}(h_e : e \in p)$，输出解耦特征表示 $h_{decouple} \in R^{d_h}$ ，并融合策略嵌入，解藕流量嵌入以及相对幅度系数作为网络流量工程头中迭代混合专家解码器的输入；

S402、多专家迭代决策生成；

采用多轮迭代路由决策，对于迭代步 $\tau = 1$ 到 $\tau_{\max}$ 执行以下步骤：

编码全局拥塞状态 $b^{(\tau)} = \mathrm{GlobalCongestion}(r^{(\tau-1)}, D, c)$；

对每个源-目的对 (s,d)，构建专家输入 $x_{sd}^{(\tau)} = [g_p, H_{sd}, \varphi, s, b^{(\tau)}]$，其中 g_p 为路径嵌入；

采用 Top-N稀疏激活策略，仅选择权重最高的N个专家，对选中专家的权重进行重归一化：$\widetilde{g_k}$，激活的专家并行推理 $a_{p,k}^{(\tau)} = \mathrm{Expert}_k^{(\tau)}\left(x_{sd}^{(\tau)}\right) \quad \forall p \in \mathcal{P}_{sd}, k \in \mathrm{Top\text{-}N}$；

加权融合专家输出 $a_p^{(\tau)} = \sum_k^{\text{Top-N}} \widetilde{g_k} \cdot a_{p,k}^{(\tau)}$；

Softmax 归一化产生路由概率 $r_p^{(\tau)} = \frac{\exp\left(a_p^{(\tau)}\right)}{\sum_{p' \in \mathcal{P}sd} \exp\left(ap'^{(\tau)}\right)}$，输出 $r_p^{(\tau_{max})}$最终路由决策；

S403、约束满足与输出生成；

流量守恒约束通过 Softmax 自动满足：$\sum_{p \in \mathcal{P}sd} r_p = 1 \quad \forall (s,d) \in V \times V$，计算每条链路的实际负载，令最大链路利用率 MLU 作为主要优化目标：$\text{MLU} = \max_{e \in E} \frac{f_e}{c(e)}$，输出最终的流量分配矩阵 $R = [r_p]$，对于每个 SD 对 (s,d) 和路径 $p \in \mathcal{P}_{sd}$，$r_p \in [0,1]$ 表示该路径承载的流量比例。

6. 根据权利要求 2 所述的基于大模型智能体的跨环境泛化网络流量优化方法，其特征在于，步骤 S5 中，通过大模型网络模型上下文协议下发路由分配，并监测网络反馈和性能指标具体包括：

S501、通过网络模型上下文协议下发指令；

通过网络模型上下文协议将配置下发到网络控制器，控制器通过 OpenFlow 或 NETCONF 协议下发到网络设备；

S502、实时性能监测；

持续采集性能指标，包含链路利用率、端到端延迟、丢包率、最大链路利用率、平均链路利用率，异常事件检测包括拥塞检测、链路故障检测以及流量突变检测；

S503、异步策略调整与持续优化；

采用双时间尺度控制架构：设置快速控制循环的频率与延迟，慢速控制循环的频率与延迟，快速控制循环和慢速控制循环独立运行解耦时间尺度，自适应策略调整为定期触发或事件触发，大语言模型分析反馈数据生成新策略嵌入，所述事件触发包含性能异常、拓扑变化以及流量漂移。

说 明 书

基于大模型智能体的跨环境泛化网络流量优化方法

技术领域

本发明涉及网络流量优化领域，具体涉及一种基于大模型智能体的跨环境泛化网络流量优化方法。

背景技术

在实际应用中，大模型智能体在下游任务上的表现常常受限于高质量数据的稀缺性。特别是在网络领域，网络流量数据多分布于本地且具有高度私密性，即使存在部分数据中心的开源数据，由于网络拓扑结构和数据表现形式存在较大差异，难以实现跨环境迁移和有效利用，限制了大模型通过丰富训练数据提升性能的空间。此外，网络流量状态空间庞大且动态变化，设计兼容高效且实用的大模型输入输出架构成为极具挑战的课题。

当前网络流量工程面临的主要挑战在于，现有方法普遍缺乏足够的智能化能力及架构上的灵活性，导致在提升路由质量、保障系统可扩展性以及实现跨环境泛化能力三者之间难以取得有效平衡。这种智能化不足限制了流量工程控制器对复杂网络拓扑的感知与适应，也影响了其在动态、多变的网络环境中快速响应和迁移能力，进而制约了整体网络性能的提升和应用范围的扩展。

传统的基于优化器的流量工程控制器（如基于线性规划或混合整数规划的方法）虽然能够进行全局流量优化，但缺乏灵活适应复杂多变业务需求的能力，且其在大规模网络上计算复杂度高且收敛缓慢，难以满足大规模网络的扩展要求与实时响应。基于机器学习的方法则通过模式学习提升性能，然而往往需要针对特定网络环境重复训练，泛化能力弱，难以在新拓扑或异构环境中保持效果，缺乏"开箱即用"能力，并且需要持续性监控与维护。同时，现有系统对自然语言指令的处理能力不足，无法直接智能解析和执行复杂的语义命令，增加了网络管理员的操作难度并降低了整体工作效率。

因此，如何突破上述瓶颈，融合大模型智能体的优势，提升网络流量工程的智能化和适应性，是满足未来大规模、动态网络环境需求的关键所在。

发明内容

本发明的目的在于克服现有技术的缺点，提供一种基于大模型智能体的跨环境泛化网络流量优化方法，提升了网络流量工程的智能化和适应性。

本发明采取如下技术方案实现上述目的，本发明提供一种基于大模型智能体的跨环境泛化网络流量优化方法，包括：

S1、数据采集与输入，包括历史流量矩阵、网络拓扑与容量信息以及来自网络管理员的

自然语言指令；

S2、跨环境特征解耦，将采集到的历史流量矩阵、网络拓扑与容量信息送入特征节藕模块进行跨环境特征解耦；

S3、将自然语言指令送入大模型智能体进行策略理解与嵌入；

S4、通过预训练的网络流量工程头解码生成网络路由分配；

S5、通过大模型的网络模型上下文协议下发路由分配，并监测网络反馈和性能指标，作为下一轮优化的输入，形成闭环控制。

进一步的是，步骤 S1 具体包括：

S101、历史流量矩阵采集，采集历史流量窗口 $H_{sd} = [D_{sd}^{t-L+1}, \dots, D_{sd}^{t}]$，其中 D_{sd} 表示从源节点 s 到目的节点d的流量需求；

S102、网络拓扑与容量信息获取，获取网络拓扑图 $G = (V, E, c)$，其中 V 为节点集合，E 为链路集合，$c: E \to R^{+}$ 为链路容量函数；为每个源-目的对(s, d) 生成候选路径集合$\mathcal{P}_{sd}$ ；

S103、自然语言策略指令接收，接收网络管理员的自然语言指令，指令被送入大模型智能体进行语义解析。

进一步的是，步骤 S2 中，跨环境特征解耦具体包括：

S201、计算相对流量矩阵，对流量矩阵进行归一化处理：

$\widehat{\mathrm{D}}_{\mathrm{sd}} = \frac{D_{sd}}{\max_{s',d' \in V} D_{s'd'}} \quad \forall s, \mathrm{d} \in V$，D 表示流量需求矩阵，归一化后的流量值 $\widehat{D}_{\mathrm{sd}} \in [0,1]$，提取流量分布模式，消除绝对量级影响，使模型对流量绝对量级变化具有鲁棒性；

S202、计算相对容量矩阵，对链路容量进行归一化处理：$\hat{c}(e) = \frac{c(e)}{\max_{e' \in E} c(e')} \quad \forall e \in E$，归一化后的容量值 $\hat{c}(e) \in [0,1]$，提取拓扑无关的容量模式，使模型适应从 Gbps 级广域网到 Tbps 级数据中心的不同规模网络；

S203、生成容量-流量相对系数，计算量级比例因子：$\varphi = \frac{\max_{s,d \in V} D_{sd}}{\max_{e \in E} c(e)}$，保留规模信息以支持最大链路利用率之外的多目标优化。

进一步的是，步骤 S3 具体包括：

S301、网络模型上下文协议驱动的智能体架构初始化；

采用预训练大语言模型作为推理引擎，通过网络模型上下文协议与网络系统进行结构化交互，使智能体具备网络抽象感知、工具调用、多步推理和反馈学习能力；

S302、三维策略嵌入空间映射；

将自然语言策略分解并映射到三维向量$s = [s_h, s_r, s_c] \in [0,1]^3$，三维向量包含三个维度：

历史依赖性s_h，控制对历史流量变化的响应敏感度，$s_h \in [0,1]$，其中 $s_h = 0$表示忽略历

史波动，$s_h = 1$表示高度响应历史变化为突发流量预留裕度；

全局鲁棒性s_r，控制对未预见流量突发的弹性能力，$s_r \in [0,1]$，其中 $s_r = 0$为尽力而为策略，$s_r = 1$为最坏情况鲁棒性；

成本敏感性s_c，控制性能与成本之间的权衡偏好，$s_c \in [0,1]$，其中 $s_c = 0$为性能优先不考虑成本，$s_c = 1$为成本优先可接受一定性能折损；

S303、Few-shot 小样本提示工程，为大语言模型提供示例指导策略翻译，通过思维链技术引导逐步推理：识别优化目标、提取约束条件、量化目标权重、映射到三维策略空间以及生成策略嵌入向量；

S304、策略验证与传递，对生成的策略嵌入进行合理性验证，如发现歧义或冲突则触发多轮对话澄清，验证通过后将策略嵌入向量 s 下发给预训练的网络流量工程头，所述合理性验证包含范围检查、冲突检测以及历史一致性检查。

进一步的是，步骤 S4 具体包括：

S401、特征融合与编码；

通过图神经网络进行容量特征编码，首先进行边特征编码 $h_e = \mathrm{EdgeEncoder}(A, \hat{C})$，其中 A 为邻接矩阵，$\hat{C}$ 为归一化容量矩阵，然后进行路径特征聚合 $g_p = \mathrm{PathEncoder}(h_e : e \in p)$，输出解耦特征表示 $h_{decouple} \in R^{d_h}$ ，并融合策略嵌入，解藕流量嵌入以及相对幅度系数作为网络流量工程头中迭代混合专家解码器的输入；

S402、多专家迭代决策生成；

采用多轮迭代路由决策，对于迭代步 $\tau = 1$ 到 $\tau_{\max}$ 执行以下步骤：

编码全局拥塞状态 $b^{(\tau)} = \mathrm{GlobalCongestion}(r^{(\tau-1)}, D, c)$；

对每个源-目的对 (s,d)，构建专家输入 $x_{sd}^{(\tau)} = [g_p, H_{sd}, \varphi, s, b^{(\tau)}]$，其中 g_p 为路径嵌入；

采用 Top-N稀疏激活策略，仅选择权重最高的N个专家，对选中专家的权重进行重归一化：$\widetilde{g_k}$，激活的专家并行推理 $a_{p,k}^{(\tau)} = \mathrm{Expert}_k^{(\tau)}\left(x_{sd}^{(\tau)}\right) \quad \forall p \in \mathcal{P}_{sd}, k \in \text{Top-N}$；

加权融合专家输出 $a_p^{(\tau)} = \sum_k^{\text{Top-N}} \widetilde{g_k} \cdot a_{p,k}^{(\tau)}$；

Softmax 归一化产生路由概率 $r_p^{(\tau)} = \frac{\exp\left(a_p^{(\tau)}\right)}{\sum_{p' \in \mathcal{P}sd} \exp\left(ap'^{(\tau)}\right)}$，输出 $r_p^{(\tau_{max})}$最终路由决策；

S403、约束满足与输出生成；

流量守恒约束通过 Softmax 自动满足：$\sum_{p \in \mathcal{P}sd} r_p = 1 \quad \forall (s,d) \in V \times V$，计算每条链路的

实际负载，令最大链路利用率 MLU 作为主要优化目标：$\mathrm{MLU} = \max_{e \in E} \frac{f_e}{c(e)}$，输出最终的流量分配矩阵 $R = [r_p]$，对于每个 SD 对 (s,d) 和路径 $p \in \mathcal{P}_{sd}$，$r_p \in [0,1]$ 表示该路径承载的流量比例。

进一步的是，步骤 S5 中，通过大模型网络模型上下文协议下发路由分配，并监测网络反馈和性能指标具体包括：

S501、通过网络模型上下文协议下发指令；

通过网络模型上下文协议将配置下发到网络控制器，控制器通过 OpenFlow 或 NETCONF 协议下发到网络设备；

S502、实时性能监测；

持续采集性能指标，包含链路利用率、端到端延迟、丢包率、最大链路利用率、平均链路利用率，异常事件检测包括拥塞检测、链路故障检测以及流量突变检测；

S503、异步策略调整与持续优化；

采用双时间尺度控制架构：设置快速控制循环的频率与延迟，慢速控制循环的频率与延迟，快速控制循环和慢速控制循环独立运行解耦时间尺度，自适应策略调整为定期触发或事件触发，大语言模型分析反馈数据生成新策略嵌入，所述事件触发包含性能异常、拓扑变化以及流量漂移。

本发明的有益效果为：

本发明基于大模型智能体进行动态语义理解和路由策略控制，实现自然语言指令到路由策略的智能转换。

本发明创新设计特征解耦模块，实现网络流量需求与链路容量的分离建模，生成相对流量、相对容量及容量-流量相对系数，提升模型应对异构数据场景的鲁棒性和跨环境泛化能力。

本发明提出预训练流量工程头异步融合策略嵌入与解耦特征，高效形成智能流量分配决策，使用混合专家架构提升模型面对异构路由策略的兼容能力和决策质量。

附图说明

图 1 是本发明提供的基于大模型智能体的跨环境泛化网络流量优化方法流程图；

图 2 是本发明提供的基于大模型智能体的跨环境泛化网络流量优化系统架构图。

具体实施方式

为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述。

本发明提供一种基于大模型智能体的跨环境泛化网络流量优化方法，如图 1 所示，具体

包括：

S1、数据采集与输入，包括历史流量矩阵、网络拓扑与容量信息以及来自网络管理员的自然语言指令；

S101、流量矩阵采集，采集流量需求矩阵$D \in R^{|V|\times|V|}$，可选地采集历史流量窗口 $H_{sd} = [D_{sd}^{t-L+1}, \ldots, D_{sd}^{t}]$，其中 D_{sd} 表示从源节点s到目的节点d的流量需求，V 代表节点集合，R 代表实数域，D 是一个节点数量乘节点数量大小的矩阵，t 是时刻索引，L 是历史流量窗口长度；

S102、网络拓扑与容量信息获取，获取网络拓扑图 $G = (V, E, c)$，其中 V 为节点集合，E 为链路集合，$c: E \to R^{+}$ 为链路容量函数；为每个源-目的对 (s, d) 生成候选路径集合 $\mathcal{P}_{sd}$；

S103、自然语言策略指令接收，接收网络管理员的自然语言指令（如：″优先保证低延迟，同时避免链路拥塞″），指令被送入大模型智能体进行语义解析。

S2、跨环境特征解耦，将采集到的历史流量矩阵、网络拓扑与容量信息送入特征节藕模块进行跨环境特征解耦；

S201、计算相对流量矩阵，对流量矩阵进行归一化处理：$\widehat{\mathrm{D}}_{\mathrm{sd}} = \frac{D_{sd}}{\max_{s',d' \in V} D_{s'd'}} \quad \forall s, \mathrm{d} \in V$，归一化后的流量值 $\widehat{D}_{\mathrm{sd}} \in [0,1]$，提取流量分布模式，消除绝对量级影响，使模型对流量绝对量级变化具有鲁棒性；

S202、计算相对容量矩阵，对链路容量进行归一化处理：$\hat{c}(e) = \frac{c(e)}{\max_{e' \in E} c(e')} \quad \forall e \in E$，归一化后的容量值 $\hat{c}(e) \in [0,1]$，提取拓扑无关的容量模式，使模型适应从 Gbps 级广域网到 Tbps 级数据中心的不同规模网络；

S203、生成容量-流量相对系数，计算量级比例因子：$\varphi = \frac{\max_{s,d \in V} D_{sd}}{\max_{e \in E} c(e)}$，保留规模信息以支持 MLU（max link utilization，最拥塞的链路的利用率）之外的多目标优化（如成本敏感、时延优化），实践中通常 $\varphi \in (0,1)$。

S3、将自然语言指令送入大模型智能体进行策略理解与嵌入；

S301、网络模型上下文协议驱动的智能体架构初始化，采用预训练大语言模型（如 Claude-3.5-Sonnet、GPT-4）作为推理引擎，通过网络模型上下文协议与网络系统进行结构化交互，使智能体具备网络抽象感知、工具调用、多步推理和反馈学习能力。

S302、三维策略嵌入空间映射，将自然语言策略分解并映射到三维向量 $s = [s_h, s_r, s_c] \in [0,1]^3$，包含三个维度：

历史依赖性s_h，控制对历史流量变化的响应敏感度，$s_h \in [0,1]$，其中 $s_h \to 0$ 表示忽略历史波动，$s_h \to 1$ 表示高度响应历史变化为突发流量预留裕度；

全局鲁棒性s_r，控制对未预见流量突发的弹性能力，$s_r \in [0,1]$，其中 $s_r \to 0$ 为尽力而

为策略，$s_r \to 1$ 为最坏情况鲁棒性（接近等价多路径路由）；

成本敏感性s_c，控制性能与成本之间的权衡偏好，$s_c \in [0,1]$，其中 $s_c \to 0$ 为性能优先不考虑成本，$s_c \to 1$ 为成本优先可接受一定性能折损；

S303、Few-shot 小样本提示工程，为 LLM 提供少量示例指导策略翻译，通过思维链（Chain-of-Thought）技术引导逐步推理：识别优化目标、提取约束条件、量化目标权重、映射到三维策略空间、生成策略嵌入向量；

S304、策略验证与传递，对生成的策略嵌入进行合理性验证（范围检查、冲突检测、历史一致性检查），如发现歧义或冲突则触发多轮对话澄清，验证通过后将策略嵌入向量 s 下发给网络流量工程头。

S4、通过预训练的网络流量工程头解码生成网络路由分配；

S401、特征融合与编码，通过图神经网络进行容量特征编码：首先进行边特征编码 $h_e = \mathrm{EdgeEncoder}(A, \hat{C})$，其中 A 为邻接矩阵，$\hat{C}$ 为归一化容量矩阵；然后进行路径特征聚合 $g_p = \mathrm{PathEncoder}(h_e : e \in p)$，输出解耦特征表示 $h_{decouple} \in R^{d_h}$ ，并融合策略嵌入，解藕流量嵌入以及相对幅度系数作为 TE-Head 中迭代混合专家解码器的输入；

S402、多专家迭代决策生成，系统采用多轮迭代（通常 3-5 轮）逐步精炼路由决策：对于迭代步 $\tau = 1$ 到 $\tau_{\max}$ 执行以下步骤：

编码全局拥塞状态 $b^{(\tau)} = \mathrm{GlobalCongestion}(r^{(\tau-1)}, D, c)$；

对每个源-目的对 (s,d)，构建专家输入 $x_{sd}^{(\tau)} = [g_p, H_{sd}, \varphi, s, b^{(\tau)}]$，其中 g_p 为路径嵌入；采用 Top-N稀疏激活策略（通常 $N = 2$），仅选择权重最高的 N 个专家，对选中专家的权重进行重归一化：$\widetilde{g_k}$，激活的专家并行推理 $a_{p,k}^{(\tau)} = \mathrm{Expert}_k^{(\tau)}\left(x_{sd}^{(\tau)}\right) \quad \forall p \in \mathcal{P}_{sd}, k \in \text{Top-N}$；

加权融合专家输出 $a_p^{(\tau)} = \sum_k^{\text{Top-N}} \widetilde{g_k} \cdot a_{p,k}^{(\tau)}$；

Softmax 归一化产生路由概率 $r_p^{(\tau)} = \frac{\exp\left(a_p^{(\tau)}\right)}{\sum_{p' \in \mathcal{P}sd} \exp\left(ap'^{(\tau)}\right)}$，输出 $r_p^{(\tau_{max})}$最终路由决策。

S403、约束满足与输出生成，流量守恒约束通过 Softmax 自动满足：$\sum_{p \in \mathcal{P}sd} r_p = 1 \quad \forall (s,d) \in V \times V$，计算每条链路的实际负载，令最大链路利用率作为主要优化目标：$\mathrm{MLU} = \max_{e \in E} \frac{f_e}{c(e)}$，输出最终的流量分配矩阵 $R = [r_p]$，对于每个 SD 对 (s,d) 和路径 $p \in \mathcal{P}_{sd}$，$r_p \in [0,1]$ 表示该路径承载的流量比例。

S5、通过大模型的网络模型上下文协议下发路由分配，并监测网络反馈和性能指标，作

为下一轮优化的输入，形成闭环控制。

步骤 5 中，通过大模型的网络模型上下文协议下发路由分配，并监测网络反馈和性能指标包括：

S501、通过网络模型上下文协议下发指令，通过网络模型上下文协议将配置下发到网络控制器（如 ONOS、OpenDaylight），控制器通过 OpenFlow/NETCONF 协议下发到网络设备；

S502、实时性能监测，持续采集关键性能指标：链路利用率、端到端延迟、丢包率、最大链路利用率、平均链路利用率，异常事件检测包括拥塞检测（链路利用率持续超过 0.9）、链路故障检测（丢包率超过 50%或无流量）、流量突变检测（流量变化超过 50%阈值）；

S503、异步策略调整与持续优化，系统采用双时间尺度控制架构：快速控制循环频率为 10-30 秒、延迟小于 2 秒，慢速控制循环频率为分钟到小时级、延迟 5-20 秒，快速控制循环和慢速控制循环独立运行解耦时间尺度；自适应策略调整为定期触发（每约 30 分钟）或事件触发（性能异常、拓扑变化、流量漂移），大语言模型分析反馈数据生成新策略嵌入。

如图 2 所示，本发明提供一种基于大模型智能体的跨环境泛化网络流量优化系统，包含：

特征解耦模块，接收历史流量矩阵、网络拓扑与容量信息数据，将其转换为与具体网络拓扑和绝对量无关的相对特征（如相对流量、相对容量、容量-流量相对系数），从而消除环境特异性，显著提高模型的跨环境泛化能力。

大模型智能体，负责接收并解析网络管理员的自然语言指令，通过大语言模型的逻辑推理和工具使用能力，将复杂的语义指令转化为可量化的路由策略嵌入向量；

预训练网络流量头，作为核心决策单元，它异步融合由特征解耦模块输出的解耦特征和由大模型智能体输出的策略嵌入，高效生成初步的流量分配意图。为提升系统对异构路由策略的兼容性，流量工程头采用混合专家架构。混合专家架构路由器根据当前网络状态和策略需求，动态地选择或加权激活不同的“专家”（例如，延迟优化专家、支出优化专家、抖动最小化专家），从而实现灵活且高质量的流量决策。

相比于传统方案，本发明在流量工程的关键指标上实现了显著提升。通过在多种真实网络拓扑和流量模式下的广泛实验，本发明验证了系统的有效性。

具体而言，本发明在零样本泛化能力上取得了突破性进展。与现有的基于机器学习的流量工程控制器（如 DOTE、FIGRET、HARP、AETHER 等）相比，本发明首次实现了跨规模网络的零样本部署能力：在≤50 节点网络拓扑上训练的单一模型，可直接部署到 700+节点的大规模网络，最优性损失仅为 1.3%（相对于线性规划求解器的最优解），而传统机器学习方法在跨环境部署时通常需要针对目标网络重新训练，否则性能显著下降。在与部分泛化方法（如 HARP）的直接对比中，本发明在多个测试拓扑上的归一化 MLU 指标平均低 0.25 至 1.5，展现出更强

的泛化能力。

在计算效率方面，相比基于线性规划的优化器（如 Gurobi 求解器），本发明在 700 节点规模网络上的响应时间减少 123 倍，从数分钟级降至秒级（<2 秒），同时保持接近最优的决策质量。通过混合专家的稀疏激活机制，TELLM 仅需激活 20-30%的模型参数即可完成推理，在保证决策质量的同时显著降低了计算开销。在包含 754 个节点的大规模网络测试中，本发明的推理时间仍保持在秒级范围内，验证了系统的可扩展性。

在实际网络部署方面，基于 13 节点物理网络原型（21 条链路，1 Gbps 带宽）的测试表明，本发明在 80 Mbps 流量负载下相比传统 ECMP 方案实现了丢包率的显著降低（从 16.61%降至 3.24%）、平均延迟的改善（从 1616 μs 降至 562 μs）以及抖动的减少（从 2.55 降至 2.27），证明了系统在真实网络环境中的有效性。

在泛化能力的多维度验证中，本发明展现出良好的鲁棒性：对于 2 倍流量缩放和 10%稀疏噪声扰动，系统性能保持稳定；在拓扑结构差异较大的网络（节点数从 4 到 754）上，单一预训练模型均能直接部署并保持合理的性能水平。系统支持通过自然语言指令进行策略配置，在测试的多种策略偏好（历史响应、全局鲁棒性、成本敏感等）下均能生成符合预期的流量分配方案。

以上所述仅是本发明的优选实施方式，应当理解本发明并非局限于本文所披露的形式，不应看作是对其他实施例的排除，而可用于各种其他组合、修改和环境，并能够在本文所述构想范围内，通过上述教导或相关领域的技术或知识进行改动。而本领域人员所进行的改动和变化不脱离本发明的精神和范围，则都应在本发明所附权利要求的保护范围内。

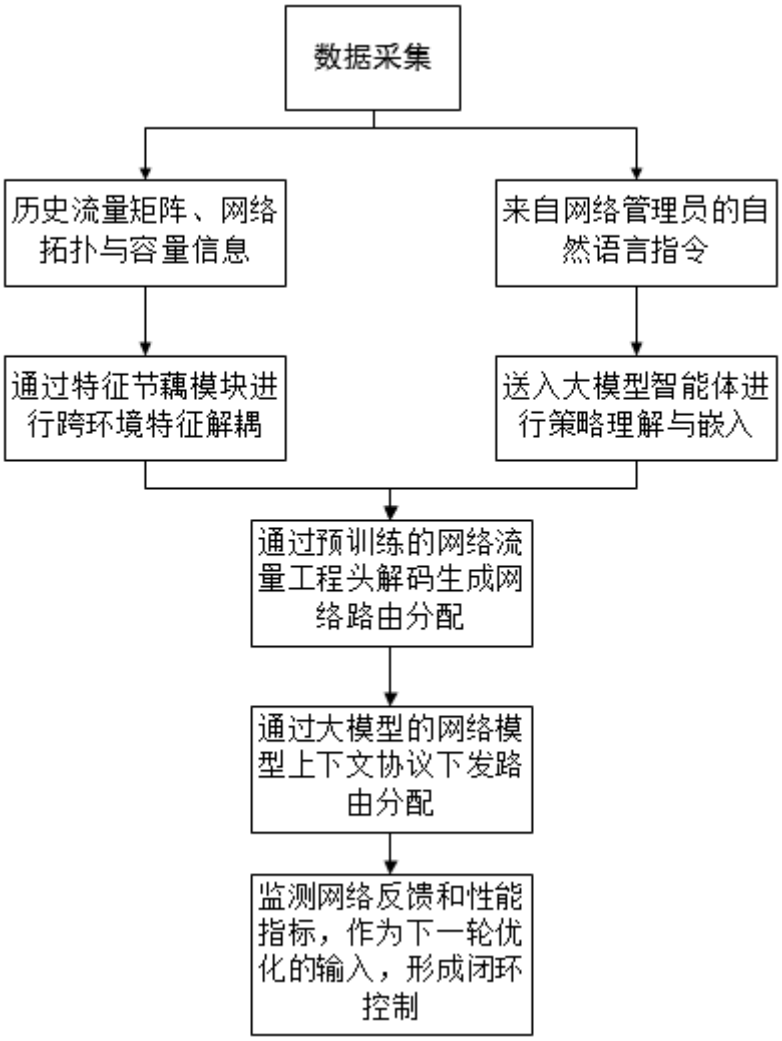

数据采集
历史流量矩阵、网络拓扑与容量信息
来自网络管理员的自然语言指令
通过特征节藕模块进行跨环境特征解耦
送入大模型智能体进行策略理解与嵌入
通过预训练的网络流量工程头解码生成网络路由分配
通过大模型的网络模型上下文协议下发路由分配
监测网络反馈和性能指标，作为下一轮优化的输入，形成闭环控制

图 1

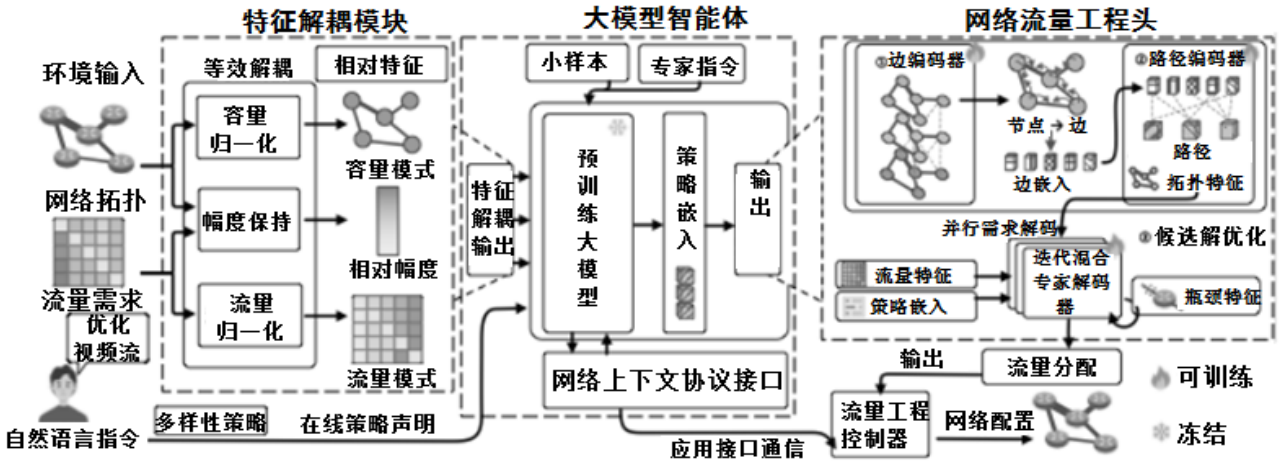

特征解耦模块
大模型智能体
网络流量工程头
环境输入
网络拓扑
流量需求
优化视频流
自然语言指令
等效解耦
相对特征
容量归一化
幅度保持
流量归一化
容量模式
相对幅度
流量模式
多样性策略
在线策略声明
特征解耦输出
小样本
专家指令
预训练大模型
策略嵌入
输出
网络上下文协议接口
应用接口通信
边编码器
节点→边
边嵌入
路径编码器
路径
拓扑特征
并行需求解码
候选解优化
流量特征
策略嵌入
迭代混合专家解码器
瓶颈特征
输出
流量分配
流量工程控制器
网络配置
可训练
冻结

图 2