文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

基于语义理解与文档聚合的多模态协同编辑架构研究：融合 MCP 协议与动态 DAG 编排

刘越杰[1] 甘杜芬[2] 冯雪英[3] 石昌妮[4] 杨瑗维 蓝予阳[6] 黄圣元[7]（通讯作者）
桂林电子科技大学 计算机工程学院 广西北海 536007

DOI：10.32629/ems.v8i2.18491

[摘 要] 针对当前多模态文档协同编辑平台在跨文档知识聚合精度与深层语义理解能力上的不足，本文构建了一套融合模型上下文协议（Model Context Protocol，MCP）与动态有向无环图（DAG）编排的系统架构。该架构通过构建分层式意图感知模型，利用大语言模型将复杂的自然语言指令转化为结构化的 DAG 任务流，并引入 MCP 协议以屏蔽底层异构模型的差异，解决了传统工具在长链路任务中存在的上下文割裂与资源调度难题。同时，结合混合检索增强策略，显著提升了多源异构数据的检索召回率。实证评测显示，该架构在保障数据隐私的前提下，有效提升了复杂协同场景下的语义理解准确率与任务执行效率。

[关键词] 语义理解；模型上下文协议（MCP）；动态 DAG 编排；混合检索增强

1. 引言

当前多模态编辑平台在底层人工智能理解与跨文档聚合方面存在瓶颈。通过融合 MCP 协议 与动态 DAG 编排，提出一种“三层-双通道”架构，旨在解决长物流任务中的上下文割裂问题。

2. 架构设计与理论模型

本研究提出了一种“三层-双总线”的系统架构，旨在解耦意图理解、任务规划与底层执行。整体架构如图 1 所示。

2.1 三层认知模型

2.1.1 语义感知层

该层主要负责处理协同编辑场景下的多模态输入，将非结构化的自然语言指令映射为结构化的意图向量。针对文档编辑中常见的模糊指令，系统引入了语义消歧与置信度评估机制，其具体处理流程如图 2 所示。

利用小参数模型（如 Qwen/DeepSeek）预判意图模糊度，必要时触发反向询问关于自注意力机制 $S_{ambiguity}$ 计算公式如式（1）所示：

$$S_{ambiguity}=1-\max\left(P(intent_i|command,context)\right) \quad (1)$$

2.1.2 编排与决策层

作为系统的核心中枢，本层主要负责实现“文档聚合”的核心逻辑。其工作流程由拓扑规划器与资源调度器协同完成，具体的编排决策逻辑如图 3 所示。

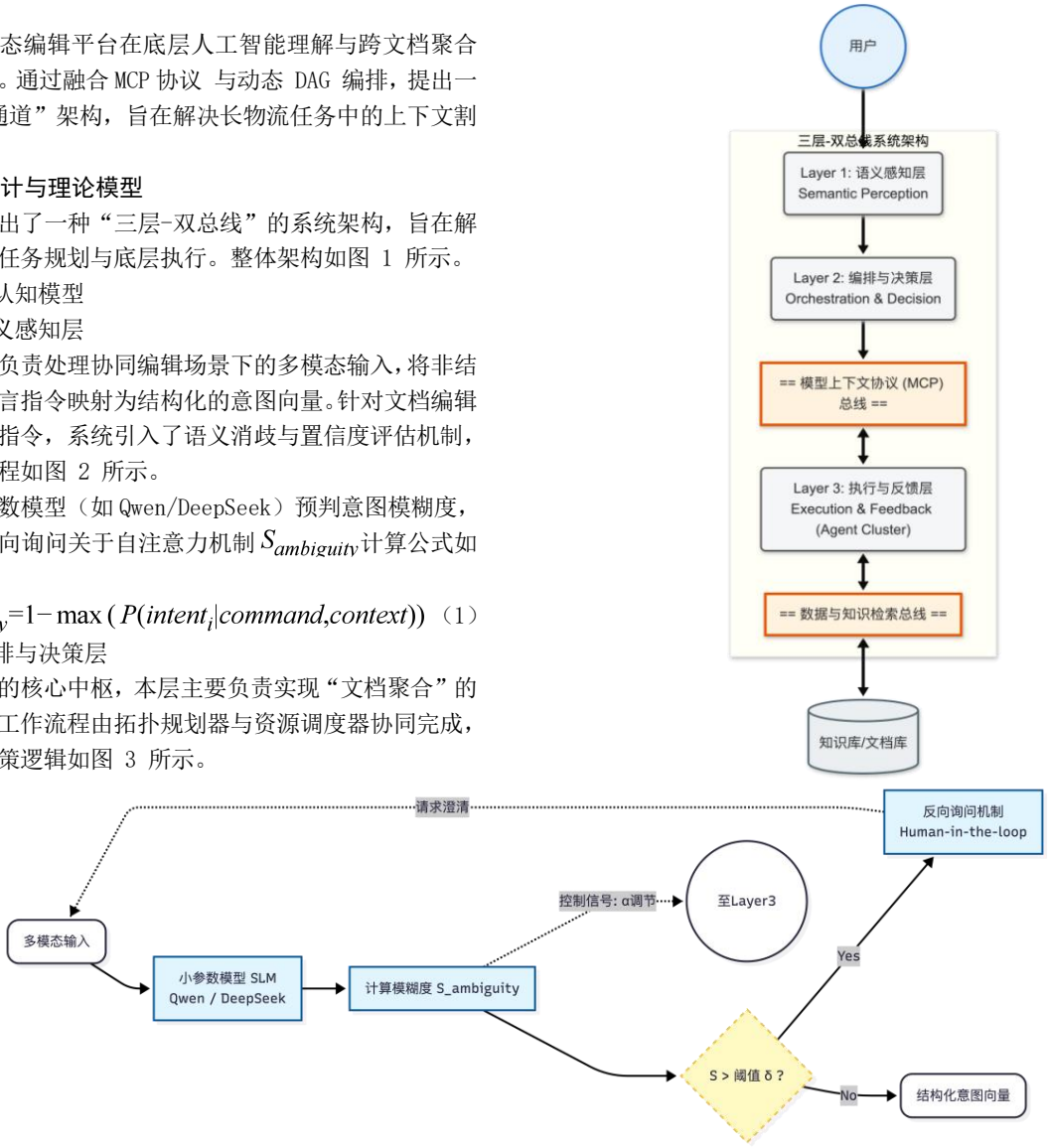

图 2 语义感知处理流程图

文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

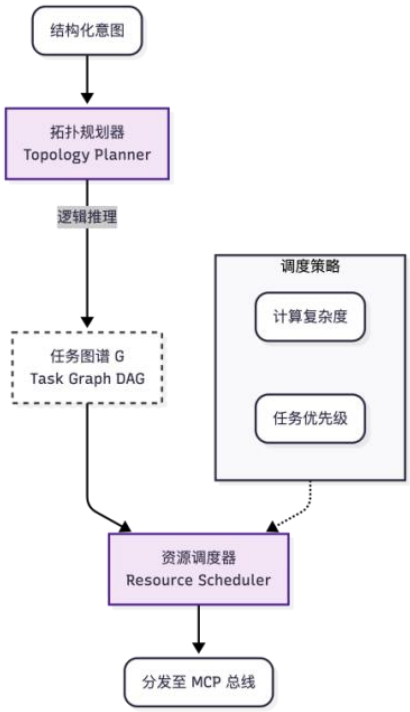

图 3　编排决策逻辑图

拓扑规划器利用利用 LLM 的逻辑推理将复杂聚合任务拆排解为 DAG 任务流，并由调度器根据任务优先级分配计算资源[2]。

2.1.3 执行与反馈层

该层由一组功能原子化的 Agent 集群构成，直接服务于协同编辑的具体功能。集群主要包含：检索、写作、评论。

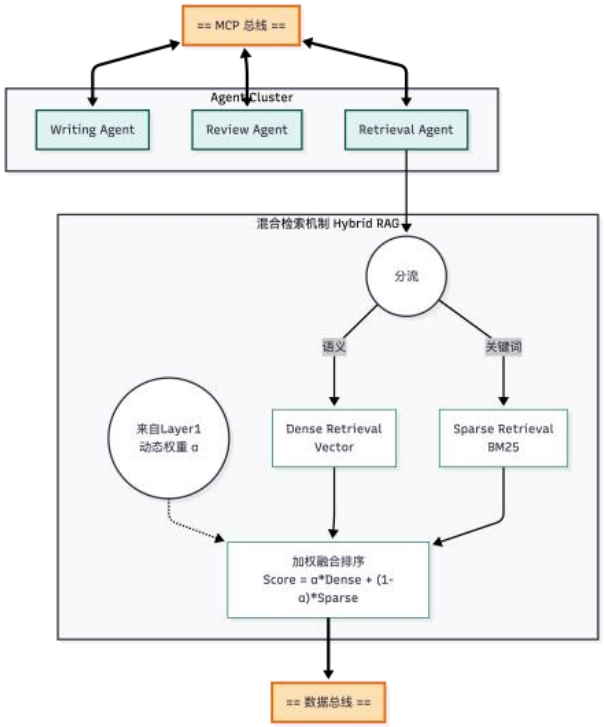

图 4　混合检索策略图

2.2 模型上下文协议

为了解决协同编辑中不同 Agent 间的信息交互难题，本文采用 MCP 协议作为系统内部的“认知总线”。MCP 通过标准化接口解决了 LLMs 与外部工具间的异构性问题[1]。我们定义跨 Agent 传递的标准数据包结构为上下文帧，其结构为通过 MCP 协议定义标准上下文框架结构，并各个 HANDSHAKE 等四个阶段确保了模型间的事务原子性。

此外，典型的 MCP 交互包含 HANDSHAKE、CONTEXT_SYNC、EXECUTE 以及 ACK/NACK 四个阶段，这种严谨的状态流转确保了在多文档聚合过程中，上下文信息能够在异构模型间无损传递且具有事务原子性，有效避免了多智能体协作中的信息丢失问题[1]。

2.3 混合检索增强机制

为了支撑“文档聚合”的高精度需求，本架构引入了混合检索策略，机制如图 4 所示。

单纯的语义检索在处理特定实体时存在局限，而混合检索重排序能显著提升大模型的回答准确率[5]。本系统采用如式（2）所示的混合检索得分计算公式

$$\mathrm{Score}(q,d)=\alpha\cdot\mathrm{Dense}(E_q,E_d)+(1-\alpha)\cdot\mathrm{Sparse}(q,d) \quad (2)$$

其中，根据意图模糊度 $S_{ambiguity}$ 动态调节权重 α，平衡语义相似度与关键词匹配度，以此弥补单一检索模式的不足[5]。

3. 动态任务编排与调度机制

3.1 基于 DAG 的文档聚合任务分解

面对复杂聚合指令，系统构建任务有向无环图 G =（V，E）。这一过程参考了 DAG-LLM[3]的设计思路，通过大模型推理自动识别子任务间的逻辑依赖。例如，在“综合 A、B、C 文档”时，读取和摘要 A、B、C 三个子任务被识别为无依赖关系，系统将其分发至三个不同的 Agent 实例并行处理，显著提升了聚合效率。

3.2 语义熵驱动的模型路由

为了在协同编辑中实现响应速度与智能深度的平衡，本文引入了一种基于语义熵的动态路由算法。对于每个任务节点 v_i，系统预读其输入内容的 Token 分布，计算语义熵 $H(x)$，如式（3）所示：

$$H(x)=-\sum_{i=1}^{N} P(\mathrm{token}_i|\mathrm{context})\log P(\mathrm{token}_i|\mathrm{context})$$

路由策略函数 $R(v_i)$ 根据熵值将任务分发至不同规模的模型，如式（4）所示：

$$R(v_i)=\begin{cases}\mathrm{Model}_{small}(\text{e.g.,Qwen-1.8B}), & \text{if } H(x)<\theta_{low}\ (\text{拼写、格式调整})\\ \mathrm{Model}_{medium}(\text{e.g.,Qwen-14B}), & \text{if } \theta_{low}\le H(x)\le\theta_{high}\ (\text{摘要、简单聚合})\\ \mathrm{Model}_{large}(\text{e.g.,Qwen-72B}), & \text{if } H(x)>\theta_{high}\ (\text{深度推理、创意写作})\end{cases}$$

通过该策略通过动态匹配算力需求，避免了资源浪费，这与边缘计算动态资源分配的思想不谋而合。

3.3 异常处理与自愈机制

在分布式多智能体协同环境中，系统设计了一套基于状态快照与多级熔断的自愈机制。系统基于 MCP 协议，在 DAG 的每个拓扑节点完成时自动触发检查点保存。当任务节点触发异常时，调度器将按照优先级策略介入。

4. 实验与分析

4.1 实验设置与评价体系

本实验在模拟私有化部署的本地集群上进行，采用包含 500 组高复杂度跨文档任务的自建数据集进行验证。评估体系包含语义理解准确率、文档聚合效率和系统响应延迟三个核心

文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

指标。

4.2 案例分析：财报聚合与简报生成

以“根据 Q1 与 Q2 财报分析净利润变化并生成简报”为例。通过 DAG 并行与算力路由，将财报聚合任务耗时缩短至 12 秒。系统时序图如图 5 所示。

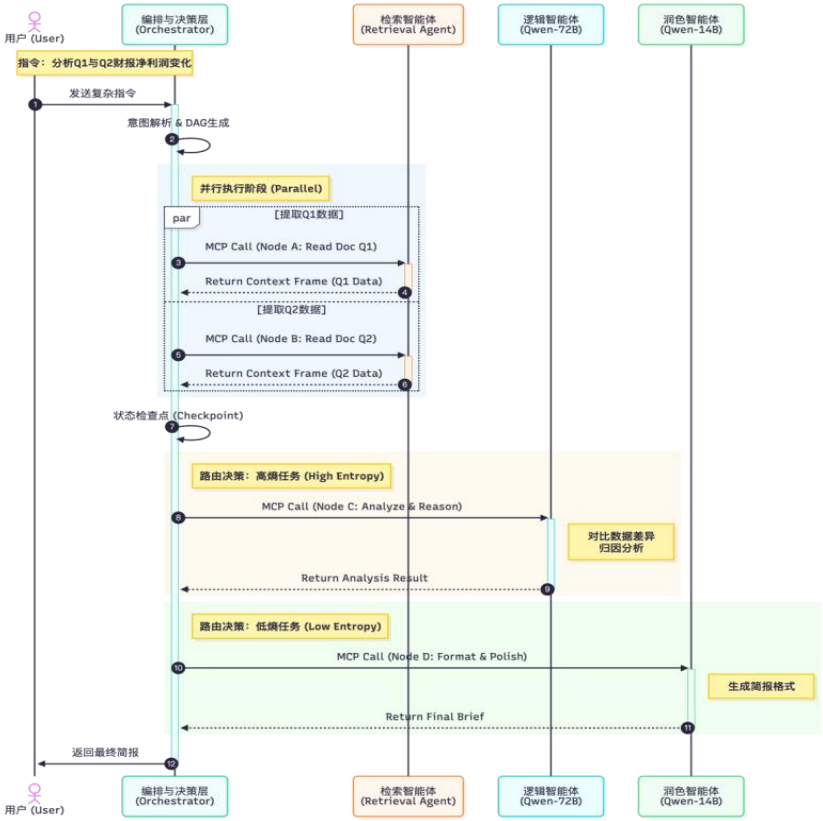

图 5　系统时序图

4.3 性能对比与消融实验

根据不同策略下的性能对比。数据表明，在此架构下在各项指标上均优于静态链式调用与单一云端模型。不同调度策略下的性能对比如表 1 所示。

表 1　不同调度策略下的性能对比

调度策略	语义理解准确率（SR）	聚合任务延迟（ms）	显存利用率
静态链式（Chain-of-Thought）	62.5%	15，400	45%
单一大模型（GPT-4 API）	88.0%	12，200	N/A（Cloud）
本文架构	91.5%	8，900	85%

实验证明，移除 DAG 并行、语义熵路由或混合检索机制，分别会导致延迟增加 60%、负载激增 120% 或准确率下降 15%[5]。

5. 结论

本文针对多模态文档协同编辑场景，设计了一种基于 MCP 协议与动态 DAG 编排的智能体协同架构。该架构通过引入 DAG 任务分解[3]、MCP 标准化通信[1]及混合检索策略[5]，提出了异构模型环境下的语义理解与文档聚合架构。多场景测评证实了该架构在提升任务执行效率与准确率方面的有效性，为未来构建私有化、高可控的智能办公环境提供了理论支撑与实践范式。

[参考文献]

[1]陈星宇，曾泳谕，徐佳沅，王思敏.融合 LLMs 和 MCP 技术的数据分析与知识服务智能体研究与应用[J].北京测绘，2025，39（11）：1700-1706.

[2]梁晓天，姚洁，许梦婕.基于强化学习的 MCP 协议动态资源分配策略在多智能体边缘计算中的应用[J].中国科技信息，2025，（22）：96-98.

[3]林明生，沈立炜，董震.DAG-LLM：基于大语言模型的多机器人任务调度方法[J].计算机工程，1-11.

[4]赵晨，张德.自注意力机制的持续进化：从 Transformer 到 DeepSeek-R1（2025）的深度语义理解跃迁[J].互联网周刊，2025，（16）：20-23.

[5]张健，唐晋韬，王挺，李莎莎.基于混合检索重排序策略的大模型增强方法[J].中文信息学报，2025，39（04）：42-54.

作者简介：刘越杰，男，本科生在读，研究方向为软件工程，智能体架构设计；甘杜芬，女，高级工程师，博士，研究方向为数据处理、算法研究；冯雪英，女，本科生在读，研究方向为自然语言处理，数据分析；

通讯作者：黄圣元，男，本科，研究方向为机电控制及自动化，思政教育。

基金资助：2025 年广西壮族自治区大学生创新训练计划立项项目（项目编号：202510595061）。