 DOI: https://doi.org/10.12345/jxffcxysj.v9i10.39514

Research on the Adaptive Dynamic Integration Mechanism of Domestic Large-scale Multimodal Intelligent Agents

Junjie Guo

Chongqing Electronic Science and Technology Vocational University, Chongqing, 401331, China

Abstract

This paper proposes an adaptive dynamic integration mechanism designed to flexibly orchestrate heterogeneous domestic model clusters into a unified agent with multimodal perception and inference capabilities. The core of the mechanism is a "dual-layer adaptive routing" architecture: The first layer performs task decomposition and modal metacognition, converting composite instructions into directed acyclic graphs of atomic tasks while evaluating modal processing complexity to determine optimal routing strategies; the second layer incorporates lightweight routing adapters for pluggable model integration and real-time adaptive selection, featuring a built-in degradation isolation mechanism to ensure task continuity. Additionally, a natural language-mediated cross-modal inference state transfer method is developed, enabling coherent working memory maintenance during long-range multimodal conversations. Experimental results demonstrate that the proposed dynamic integration mechanism outperforms single-model approaches and fixed pipeline solutions in improving task completion rates, reducing peak-hour response latency, and minimizing GPU resource consumption.

Keywords

multimodal agent; dynamic integration; domestic large-scale model; routing adapter; adaptive routing

国产大模型多模态智能体自适应性动态集成机制研究

郭俊杰

重庆电子科技职业大学，中国·重庆 401331

摘　要

本文提出一种自适应性动态集成机制，旨在将异构国产模型群灵活编排为一个具备统一多模态感知与推理能力的智能体。机制核心为“双层自适应路由”：第一层进行任务拆解与模态元认知，将复合指令转化为原子任务有向无环图，并评估模态处理硬度以决定“捷径策略”；第二层引入轻量路由适配器，实现对候选模型的可插拔集成与实时自适应选择，并内置降级隔离机制保障任务连续性。同时设计了以自然语言为中介的跨模态推理状态迁移方法，使长程多模态对话能够保持连贯的工作记忆。实验结果表明，所提动态集成机制相比于单一模型和固定流水线方案，在提升任务完成率、降低高峰时段响应延迟、降低GPU资源消耗等方面具有优势。

关键词

多模态智能体；动态集成；国产大模型；路由适配器；自适应路由

1 引言

大语言模型的迅猛发展推动人工智能进入智能体时代。以 GPT-4o、Gemini 等为代表的海外模型已展现出原生统一多模态理解和生成能力，同时国内大模型生态也在文本推理领域凭借 DeepSeek、Qwen 等系列跻身第一梯队，但在原生多模态模型层面，仍存在垂直场景细腻度不足、不同模态对齐偏差等问题[1]。同时，国产算力供应紧张且异构化严重，使得简单堆叠一个巨型多模态模型来驱动智能体的方案既不经济、也不现实。此外，“生态碎片化”问题愈发突出，不同厂商发布的多模态模型或视觉专家模型，其输入格式、嵌入维度、标记器乃至输出风格均不统一，智能体只绑定某一款模型易受单点故障和性能波动的影响。因此，如何基于国内已有的多个各有所长的模型（文本推理、视觉理解、语音识别、代码生成等），在不重度改造模型本体的前提下，动态、自适应地集成为一个统一的多模态智能体，成为亟需研究的核心问题。

本文提出一种自适应性动态集成机制，设计了双层自适应路由架构，提出跨模态推理状态迁移方法，旨在构建一

【课题项目】重庆市教育科学规划课题“大模型多模态智能体协作系统设计及信创 ICT 课程教学应用实践”（项目编号：K24YG3090227）。

【作者简介】郭俊杰（1987–），男，中国重庆人，硕士，副教授，从事通信软件技术研究。

个能够感知任务复杂度、模型实时状态及算力约束，并据此动态编排国产异构模型协作的“调度脑”。

2 双层自适应路由与动态集成机制

2.1 任务拆解与模态元认知

当多模态智能体接收到类似“请根据这张 OpenEuler 系统安装截图，指出分区步骤中的错误，并生成修正后的操作手册”的复合指令时，首要任务是对其进行可感知的任务拆解。

本文设计了一个基于大语言模型的规划器，该规划器通过思维链提示，输出一个任务有向无环图（DAG）[2]。节点为原子能力单元，如“OCR 识别截图文字”“理解分区界面元素关系”“逻辑比对最佳实践”“文本生成操作手册”。节点之间标注依赖关系，每个节点携带有模态处理需求（纯文本、视觉理解、代码执行等）。在此基础上引入模态硬度评估，即判断每个视觉相关子任务是否确实需要调用高成本视觉编码器，还是可走“捷径策略”。捷径策略的核心思想是：对于大量结构化或文字为主的截图，只需利用轻量级浅层视觉编码器将图像转化为可用的视觉 token 特征，再交由强推理文本大模型进行“看图说话”式的语义理解。这一做法不仅大幅降低计算开销，还能借助强大的文本推理能力弥补视觉模型在某些领域微调不足的问题。硬度评估由一个轻量分类器完成，该分类器在历史任务数据上训练，能够根据任务描述和缩略图快速判断视觉复杂性，决定分发路径。

此外，规划器会生成一个置信度预言，即评估每一步子任务的预期完成置信度。若某子任务评估为高不确定性（如涉及模糊手写体识别），则预先为其准备降级方案，例如转为请求用户提供更清晰图片，或利用文本模型的常识进行补全。

2.2 基于路由适配器的可插拔集成

为实现对异构国产模型的热插拔式集成，本文提出了路由适配器（Routing Adapter）概念。不同于强制统一接口或重新训练模型，路由适配器是附着在每个候选模型之上的极轻量神经网络模块（参数量 <1M）。其功能为：将路由器输出的任务表征向量映射为该模型擅长的“适配空间”，并输出该模型对于当前子任务的一个胜任力得分。同时，适配器持续收集该模型的实时状态（当前推理延迟、队列长度、错误率）并编码为状态嵌入，与任务表征联合计算。

例如，设当前子任务的任务表征为，模型 m 的适配器输出得分由下式给出：

$$s_m = \sigma(\mathbf{W}_m^{(2)}\mathrm{ReLU}(\mathbf{W}_m^{(1)}[\mathbf{h}_t;\mathbf{e}_m]+\mathbf{b}_m^{(1)})+\mathbf{b}_m^{(2)})$$

式中，为模型 m 当前的状态嵌入（实时捕获推理延迟、队列长度、近期错误率等）；$\mathbf{W}_m^{(1)}$、$\mathbf{W}_m^{(2)}$ 为适配器的可训练权重矩阵，表示将任务表征向量与模型状态嵌入按列拼接，$\mathbf{b}_m^{(1)}$ 为偏置向量，$\mathbf{b}_m^{(2)}$ 为标量偏置；$\sigma(\cdot)$ 为 Sigmoid 函数，将得分映射到 (0,1) 区间。

路由器在选择模型时引入了 ε - 贪心策略 [3]，它的运作逻辑很直接：大概率下（1- ε 的概率），系统会直接调用当前得分最高的那个模型来处理任务；同时保留一个较小的随机概率，去试探其他非最优模型。这样做的优点是，系统不会陷入固化的思维，在反复利用已知靠谱模型的同时，始终保留了对候选池里其他模型的探索机会，让集成系统能在求稳与尝新之间找到一个实际可用的平衡点。

候选池动态感知是自适应性的关键机制。本文建立的候选池包含多种国产模型：云端 Qwen-VL-Max、DeepSeek-VL、Qwen-72B（纯文本）、本地 Qwen-7B、甚至在教师终端运行的 TinyLLaVA 等。服务注册时，给每个模型都打上特定的能力标签，例如“擅长表格理解”、“响应快但细节可能粗糙”、“中文手写体识别效果优秀”等。路由适配器的历史记录会不断累积每个模型在不同的细分任务维度上的真实成功率，这样路由决策就有了长期效果数据做支撑。

同时，本文设计降级与隔离机制，以保证系统的鲁棒性。当所选模型返回结果异常（格式错误）或超时，路由器立即启动两级降级：第一级，转用同模态的备用模型；第二级，若所有相同模态模型不可用，则执行“模态翻译降级”，将视觉任务转化为文本输入，交由纯文本模型结合先验知识处理，同时明确告知用户“当前系统在纯文本模式下推理，可能缺少视觉细节”。该机制有效防止了单点故障导致的交互中断。

2.3 跨模态推理状态迁移与记忆机制

多模态智能体系统工作时，经常需要进行长过程程交互，例如在信创教学场景中，教师可能先上传一张系统架构图进行讲解，接着粘贴终端报错截图要求排错，最后让智能体总结整个实验步骤。在这个过程中，智能体必须维持跨模态的工作记忆。

本文采用“以文本为中枢的模态翻译记忆”策略。具体而言，每次视觉模型产生输出后，一个专门设计的摘要器会将其压缩为一段结构化自然语言文本，存入文本大模型的上下文窗口中。这种做法将不同模态的信息统一转换到最成熟、处理能力最强的文本空间中，使得后续无论调用何种模型，都能通过文本上下文获取先前所有关键模态记忆。

在此基础上，本文进一步设计了反思式重规划闭环。以“生成信创操作系统实验指导 PPT”任务为例，文本模型先生成幻灯片内容和配图描述，随后该描述交由视觉生成模型生成图片，再由视觉理解模型作为“质检员”对生成图片进行评估，检查是否符合国产操作系统风格且没有扭曲组件。若发现问题，生成自然语言反馈，文本模型据此修正，形成一个文本 - 视觉 - 文本的闭环。这一过程完全由路由层调度，无需人工干预，且可并行生成多套方案后择优。

3 实验验证：多模态智能体协同信创 ICT 辅助教学

为全面验证本文所出提动态集成机制在真实场景中的有效性，选取信创 ICT 辅助教学作为实验载体。信创（信

息技术应用创新）产业是国产软硬件生态的核心，涉及麒麟、鸿蒙操作系统、高斯数据库、华为鲲鹏等。在信创人才培训中，常使用智能化辅助工具，其教学材料中包含大量国产系统界面截图、命令行输出、架构图等，适合多模态智能体的性能测试。

3.1 场景描述与系统部署

本文构建了一个“信创 ICT 实训智能助教”系统。教师和学生可通过系统 Web 界面与之交互，典型教学任务包括截图排错、架构图讲解、操作手册生成、混合问答等。

实验时，在云端服务器部署 Qwen-VL-72B、DeepSeek-VL-7B 及 Qwen-72B 文本模型；边缘节点部署 Qwen-7B 和 TinyLLaVA-1.5B；终端计算机运行 Qwen-1.8B 和轻量 OCR 引擎。路由层部署在边缘服务器上，统一接收终端请求，并调度云端或本地模型，路由适配器实时监控各模型状态 [4]。

3.2 实验设置与评估指标

本文收集并标注了 300 个信创 ICT 教学场景下典型多模态任务样本，按照难度分为简单（单步提取）、中等（两步推理）和困难（多步规划与生成）三个级别，每级 100 个。实验时使用的对比包括：

采用单一大模型：固定使用 Qwen-VL-Max 进行端到端回答，不进行拆解和路由转发。

采用固定流水线：先统一调用 OCR/ 视觉模型，再将文本结果交给 Qwen-72B，策略为硬编码。

采用随机路由：随机从候选池中选择可用视觉模型，然后联合文本模型。

采用本文动态集成机制：使用本文描述的双层自适应路由与适配器，灵活选择捷径策略和降级处理。

实验采取的评估指标为：

任务完成率（TCR）：由两名信创讲师对回答进行评测（0-5 打分制，4 分及以上为完成。

平均响应延迟：从请求发出到完整回复的时间。

GPU 资源消耗：记录每次任务消耗的云端 GPU 计算量（以等效 NVIDIA A100 小时为单位）。

降级触发率与恢复成功率 [5]：记录路由降级发生的频率、降级后仍能正确完成任务的比例。

情境记忆保持度：在混合连续对话测试中，由人工评估是否保持上下文一致（0-1 打分制）。

3.3 结果分析与机制验证

将实验结果汇总于表 1 进行性能对比。

结果分析可知，本文机制在任务完成率上提升显著，达到 90%，相比单一模型绝对提升 18.7 个百分点。深入分析发现，提升主要来自两方面：一是捷径策略有效降低了不必要的重模型调用，例如大量清晰截图被分诊至轻量方案处理，速度更快且不牺牲精度；二是降级恢复机制使部分视觉模型失败的任务依然通过文本推理补救完成。例如在一次“OpenEuler 安装分区界面”的模糊截图中，Qwen-VL 识别错误，系统自动降级为提示用户描述界面元素，并由 Qwen-72B 基于最佳实践给出合理建议，最终教师评价仍为有效。

表 1 不同方法在信创 ICT 教学任务上的性能对比

方法	任务完成率 (%)	平均延迟 (s)	GPU 消耗 (A100-h)	降级恢复成功率 (%)	记忆保持度
单一大模型	71.3	3.8	12.7	—	0.63
固定流水线	68.7	4.2	14.1	—	0.51
随机路由	73.5	3.5	11.9	46.2	0.59
本文机制	90.0	2.2	9.0	78.5	0.87

平均延迟降低至 2.2 秒，降幅 43%，得益于边缘模型和本地模型处理的轻量级任务避开了云端排队延迟。尤其在并发教学环境下，动态路由将低延迟压力分散到边缘节点，仅将高难度任务上云，避免洪峰冲击。GPU 资源消耗减少 29%，证明了“大小模型协同”策略的经济性。降级恢复成功率 78.5%，展示出机制在不稳定环境下的鲁棒性。混合连续对话的记忆保持度为 0.87，显著优于固定流水线的 0.51，表明以文本为中枢的模态翻译记忆有效克服了跨模型上下文割裂问题。例如在案例中，教师先问“上次讲过的 OpenEuler 系统服务管理命令是什么”，智能体正确回忆前文截图中的 systemctl 示例并加以解释，保持了教学过程中的连贯性。

综上，信创 ICT 辅助教学场景充分展示了本文所提出机制的实用价值，教师和学员获得了流畅、准确、经济的多模态智能助教，同时系统有效屏蔽国产模型生态的异构性影响。

4 总结

本文针对国产大模型生态下多模态智能体构建中的集成问题，提出了自适应性动态集成机制。通过双层自适应路由实现了任务粒度的按需调度与模型热插拔，通过跨模态状态迁移维护了长程交互的连贯记忆。信创 ICT 辅助教学场景下的实验验证了该方法在完成率、延迟和成本上的综合优势，并证明了大小模型协同和降级恢复的独特价值。

参考文献

[1] Guo D, Yang D, Zhang H, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. arXiv preprint arXiv:2501.12948, 2025.

[2] Chen J, et al. From standalone LLMs to integrated intelligence: A survey of compound AI systems[EB/OL]. arXiv preprint arXiv:2506.04565, 2025.

[3] Dong B, Fan Y, Sun Y, et al. Maximum score routing for mixture-of-experts[C]//Findings of ACL, 2025: 12619-12632.

[4] 徐国愚, 张一丹, 魏笑, 等. 自适应路由与双阈值剪枝的多模态大模型检索增强感知[J]. 计算机科学与探索, 2025.

[5] Han X, et al. Guiding mixture-of-experts with temporal multimodal interactions[EB/OL]. arXiv preprint arXiv:2509.25678, 2025.