计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(9): 2264−2284, 2026

基于多智能体辩论的图文交织生成方法

马 杰[1,2] 张瀚驰[1,2] 渠 宁[2,3] 薛皓苍[2,3] 王鑫平[1,2] 刘 均[2,3]
[1]（西安交通大学电子信息学部网络空间安全学院 西安 710049）
[2]（智能网络与网络安全教育部重点实验室（西安交通大学） 西安 710049）
[3]（西安交通大学电子信息学部计算机学院 西安 710049）
（jiema@xjtu.edu.cn）

Interleaved Text-Image Generation Method Based on Multi-Agent Debate

Ma Jie[1,2], Zhang Hanchi[1,2], Qu Ning[2,3], Xue Haoquan[2,3], Wang Xinping[1,2], and Liu Jun[2,3]
[1]（*School of Cyber Science and Engineering*, *Faculty of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049）
[2]（*Key Laboratory of Intelligent Network and Network Security* (*Xi'an Jiao Tong University*), *Ministry of Education*, *Xi'an* 710049）
[3]（*School of Computer Science and Technology*, *Faculty of Electronic and Information Engineering*, *Xi'an Jiaotong University*, *Xi'an* 710049）

Abstract Multimodal retrieval-augmented generation introduces external textual and visual knowledge, providing important support for complex document understanding, knowledge-based question answering, and step-by-step task guidance. However, existing methods usually delegate the interleaved text-image generation process to a single multimodal large language model. Specifically, the retrieved text chunks, candidate images, and image descriptions are simultaneously fed into the same model, which is required to perform text organization, visual evidence selection, image insertion position determination, and multi-image ordering within a single generation process. In complex long-context scenarios, however, the candidate content often contains weakly relevant text, visually similar images, and redundant visual evidence, which can easily lead to missing key images, selecting irrelevant images, and misordering multiple images. To address these problems, we propose a multimodal retrieval-augmented interleaved text-image generation method based on a multi-agent debate mechanism. The proposed method decomposes the generation process into three collaborative roles: a text reasoning agent, a visual perception agent, and a global judge agent. Through dual-track draft generation, set conflict and ordering conflict detection, multi-round adversarial debate, and global judgment and fusion, the method jointly optimizes image selection, dynamic image supplementation, redundant image removal, and image reordering. Experimental results show that the proposed method achieves competitive performance across the network data, academic document, and lifestyle domains of MRAMG-Bench. In particular, in the lifestyle domain, compared with the strong baseline M2IO-R1, the proposed method improves image recall and ordering consistency by 7.02 and 7.77 percentage points, respectively. In the network data domain, the proposed method improves the average performance by 7.90 percentage points compared with the best complete baseline. These results demonstrate the effectiveness of the proposed method in complex long-document interleaved text-image generation tasks. Anonymous code and reproduction instructions are available at: https://github.com/mira-ai-lab/MRAMG-MultiAgent-Debate

Key words multimodal retrieval-augmented generation; interleaved text-image generation; multi-agent debate; visual evidence selection; image ordering optimization; large language model

收稿日期: 2026-05-29; **修回日期**: 2026-07-15

基金项目: 国家自然科学基金青年科学基金项目(62306229); 陕西省科学技术协会青年人才托举计划项目(20240113)
This work was supported by the National Natural Science Foundation of China for Young Scientists (62306229) and the Young Talent Support Program of Shaanxi Association for Science and Technology (20240113).

通信作者: 刘均(liukeen@mail.xjtu.edu.cn)

摘　要　多模态检索增强生成通过引入外部文本与视觉知识，为复杂文档理解、知识问答和步骤化任务指导等任务场景提供了重要支撑。然而，现有方法通常将图文交织生成过程交由单一多模态大模型统一完成，即把检索得到的文本块、候选图像及其图像描述同时输入同一个模型，并要求模型在一次生成过程中同时承担文本内容组织、视觉证据筛选、图像插入位置判断以及多图顺序规划等任务。但在复杂长上下文场景下，候选内容中常包含弱相关文本、相似图像和冗余视觉证据，容易导致关键图像漏选、无关图像误选及多图顺序错位等问题。针对上述问题，提出一种基于多智能体辩论机制的多模态检索增强图文交织生成方法。该方法将生成过程解耦为文本推理、视觉感知和全局裁决3类协同角色，并通过双轨草稿生成、集合冲突与顺序冲突检测、多轮对抗式辩论及全局裁决与融合机制，对图像选择、动态补图、冗余剔除和顺序重排进行统一优化。实验结果表明，方法在MRAMG-Bench的网络数据、学术文档和生活方式3个领域上均取得较优表现。特别是在生活方式领域，相比当前强基线M2IO-R1，提出方法在图像召回率和顺序一致性指标上分别提升7.02个百分点和7.77个百分点；在网络数据领域，提出方法相比最优完整基线的平均性能提升7.90个百分点。上述结果验证了提出方法在复杂长文档图文交织生成任务中的有效性。为便于结果复现，提供匿名代码仓库与复现说明：https://github.com/mira-ai-lab/MRAMG-MultiAgent-Debate

关键词　多模态检索增强生成；图文交织生成；多智能体辩论；视觉证据选择；图像顺序优化；大语言模型

中图法分类号　TP311.13；TP309

DOI: 10.7544/issn1000-1239.202660338　　**CSTR:** 32373.14.issn1000-1239.202660338

近年来，大语言模型在自然语言理解、知识问答和复杂推理等任务中取得了显著进展[1-4]，但仍面临知识更新滞后、领域知识不足和事实幻觉等问题[5-6]。检索增强生成通过引入外部知识库为模型提供可追溯上下文，已成为提升大模型忠实性与可靠性的重要路径[7-8]。随着实际应用从纯文本问答扩展到网页、论文、操作手册等复杂多模态文档，多模态检索增强生成逐渐成为重要研究方向[9-14]。

在网络数据、学术和生活方式等典型场景中，用户不仅需要语义正确的文本回答，还需要与文本内容对应的视觉证据。例如，网络百科与网页问答强调图文信息匹配，学术论文解析依赖图表与公式说明，菜谱步骤指导和操作手册问答则更关注多步骤图像的顺序组织。随着多模态大模型的发展[15-17]，图文交织生成能够将关键图像插入到相应文本位置，使答案在语义解释、步骤指引和视觉展示之间形成统一表达。如图1所示，相比纯文本生成，图文交织生成更有利于降低用户理解成本并提升信息传达效率。

然而，现有多模态检索增强多模态生成（multimodal retrieval-augmented multimodal generation, MRAMG）方法多依赖单体多模态大模型进行统一生成。这类方法通常将检索得到的文本块、候选图像和图像描述一次性输入模型，再由模型同时完成文本组织、图像选择、插入位置判断和图像顺序规划。在复杂长上下文场景下，大量弱相关文本、相似图像和冗余视觉证据会引入较强噪声；同时，静态Top-N图像筛选虽然降低了视觉Token开销，却可能提前过滤与后续生成强相关的关键图像，导致图像误选、漏选和顺序错位等问题。

针对单体多模态大模型在复杂图文交织生成中容易受到候选图像噪声干扰、遗漏关键视觉证据以及产生多图顺序错位等问题，本文提出一种基于多智能体辩论机制的多模态检索增强图文交织生成方法。受多智能体辩论在复杂推理和决策可靠性方面的启发[18-19]，本文将图文交织生成过程解耦为文本推理、视觉感知和全局裁决3个协同角色，使文本逻辑建模、视觉证据判断和全局一致性控制分别由不同智能体承担。进一步地，本文设计了双轨草稿生成、集合冲突与顺序冲突检测、多轮对抗式辩论以及全局裁决与融合机制。该机制能够将图像误选、漏选和顺序错位等隐式生成偏差转化为结构化冲突，使其能够被显式检测、辩论和裁决。通过上述过程，本文方法提升了复杂长文档场景下图像选择的准确性、视觉证据的完整性和图文顺序的一致性。实验结果表明，本文方法在MRAMG-Bench的网络数据、学术文档和生活方式3个领域上均取得较优表现，尤其在生活方式领域，相比当前强基线M2IO-R1，本文方法在图像召回率和顺序一致性指标上分别提升7.02个百分点和7.77个百分点，验证了其在复杂图文交

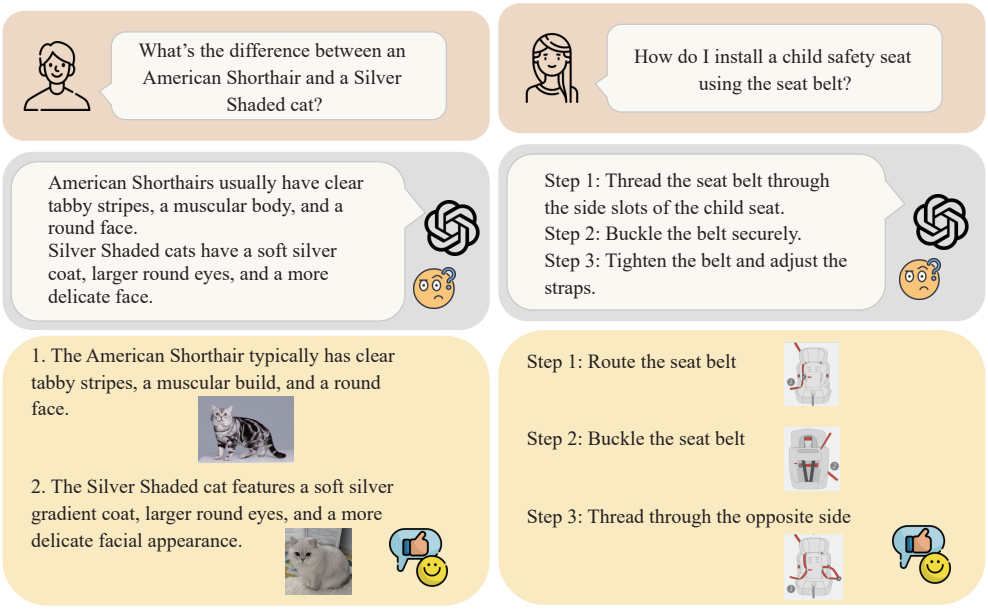


Fig. 1 Comparison of the user interaction effectiveness between pure text generation and interleaved text-image generation

图 1 纯文本生成与图文交织生成的用户交互效果对比

织生成任务中的有效性。本文主要贡献如下：

1)针对单体多模态大模型统一决策过程中任务耦合度高、易受噪声干扰的问题，提出一种基于多智能体辩论机制的 MRAMG 框架。该框架通过文本推理智能体、视觉感知智能体和全局裁决智能体的角色解耦，实现文本逻辑、视觉证据和全局一致性之间的协同建模，降低单模型一次性完成多类决策带来的耦合风险。

2)针对图像误选、关键图像漏选以及多图顺序错位等问题，设计双轨草稿生成、冲突检测与动态视觉证据补充机制。该机制能够显式定位文本路径与视觉路径之间的图像选择差异和顺序分歧，并在辩论过程中按需补充被静态预筛选遗漏的关键图像，从而提升视觉证据覆盖能力与图文排布合理性。

3)在 MRAMG-Bench 数据集上开展系统实验，从主实验、消融实验、动态补图分析、裁决智能体可靠性分析和案例分析等多个角度验证所提方法的有效性。实验结果表明，本文方法能够有效提升图像选择准确性、图像召回率、插入位置合理性和多图顺序一致性，尤其在候选图像密集、步骤依赖较强的生活方式场景下表现出更明显优势。

1 相关工作

1.1 基于多模态检索的单向文本生成方法

早期多模态检索增强研究主要关注输入端的模态扩展，即采用多模态检索输入、单模态文本输出的非对称范式。MuRAG[9] 同时从外部文本库和图像库中检索相关证据，并将视觉特征与文本信息联合输入生成模型，以缓解开放域视觉问答中的事实幻觉问题。RA-CM3[20] 进一步通过检索多模态文档片段增强模型对长尾视觉概念和复杂上下文的理解能力。国内也有研究从检索策略角度提升大语言模型生成可靠性，毕枫林等人[21] 提出基于滑动窗口策略的检索增强生成系统，通过在索引阶段保留更充分的上下文信息，缓解固定上下文窗口和静态检索方法带来的信息丢失问题。该类方法能够提升模型的多模态感知能力或文本生成可靠性，但输出端仍以纯文本为主，难以满足操作手册、教学文档和菜谱步骤等场景对图文联合表达的需求，因此无法直接解决图文交织生成中的图像选择、位置插入和顺序组织问题。

1.2 基于后置检索与精炼的流水线式图文生成方法

为突破纯文本输出限制，部分研究采用后置检索与精炼的流水线式图文生成策略。MuRAR[22] 首先由大语言模型生成纯文本草稿，再根据文本片段检索相关图像，并通过精炼模块将图像插入对应文本位置。该类方法初步实现了从纯文本回答到图文联合输出的扩展，但其图像检索与插入高度依赖初始文本草稿质量，文本组织与视觉选择之间缺乏显式协同。与此同时，含噪多模态信息也会影响后续视觉证据利用。闫萌等人[23] 指出，多模态数据中常夹杂不可避免的噪声，并利用大语言模型过滤低质量多模态信息以提升可信决策能力。对于图文交织生成任务而言，候选图像集合中同样可能包含视觉相

似但语义贡献有限的冗余图像，若缺乏显式筛选与裁决机制，容易导致图文逻辑断裂、插入位置不合理和视觉信息冗余。

1.3 基于多模态大模型的复杂图文交织与动态编排方法

随着多模态大模型在跨模态理解和细粒度视觉感知方面不断发展[15-17]，研究前沿逐渐转向端到端图文交织生成及其系统化评测。MRAMG-Bench[14]正式定义了多模态检索增强多模态生成任务，并从图像选择、插入位置和多图顺序一致性等维度评估模型能力。M2IO-R1[24]将图文编排建模为多步序列决策任务，并通过强化学习优化图像插入器。与此同时，视觉语言大模型仍面临可靠性与鲁棒性问题。李煦等人[25]系统梳理了视觉语言大模型幻觉的成因、评估与治理方法，指出模型可能对图像内容产生错误感知；曾诚等人[26]进一步总结了多模态视觉语言表征学习模型及其对抗样本攻防技术，说明复杂跨模态交互中仍存在安全性与鲁棒性挑战。此外，李国鹏等人[27]从大语言模型驱动智能体角度研究计划复用机制，表明任务分解与计划管理有助于提升智能体系统效率和可靠性。赵博涛等人[28]系统总结了基于多模态大模型的智能体研究进展，指出多模态感知、任务规划和智能决策能力是构建复杂智能体系统的重要基础。丘龙鹏等人[29]进一步面向复杂信息分析任务提出了基于深度研究的多智能体情报分析框架，通过多个角色智能体之间的协同提升复杂任务处理能力。尽管上述研究推动了多模态理解、可信生成与智能体系统的发展，但现有图文交织生成方法仍主要依赖单体模型或单一插入器完成统一决策，在候选图像较多、上下文较长时容易受到检索噪声和静态筛选限制的影响。相比之下，本文从多智能体协同与冲突裁决角度出发，将文本逻辑、视觉感知和全局优化解耦，以提升复杂图文交织生成的可靠性。

2 任务定义与问题分析

2.1 任务定义

在多模态检索增强图文交织生成任务中，系统需要根据用户查询，从外部多模态知识库中检索相关文本与图像信息，并生成包含文本内容与图像内容的交织式答案。与传统文本问答任务不同，该任务不仅要求模型生成语义正确、逻辑连贯的文本内容，还要求模型能够准确选择关键图像、确定图像插入位置，并在多图场景下保持图像顺序与文本叙事逻辑一致。

给定多模态知识库D，其中每个文档通常由文本块和图像块交替组成。形式化地，多模态知识库可表示为：

$$D=\{d_1,d_2,\cdots,d_N\}。\tag{1}$$

其中，第i个文档d_i表示为：

$$d_i=(T_1,I_1,T_2,I_2,\cdots,T_k,I_k,\cdots)，\tag{2}$$

其中，T_k表示第k个文本块，I_k表示与其相邻或相关的第k个图像块。给定用户查询q、多模态知识库D以及基础生成模型M，检索模块F_{ret}首先从知识库中召回与查询相关的图文上下文集合D_q^*，记为：

$$D_q^*=\left\{d_{j_1},d_{j_2},\cdots,d_{j_K}\right\}。\tag{3}$$

在此基础上，模型需要基于用户查询q和检索上下文D_q^*生成最终图文交织答案A^*，其目标可表示为：

$$A^*=\arg\max_A P_M\left(A|q,D_q^*\right)，\tag{4}$$

其中，A^*同时包含文本片段与图像引用。相比传统文本生成任务，多模态检索增强图文交织生成需要同时满足以下目标：1）文本语义正确性；2）图像选择准确性；3）图文位置与顺序一致性。因此，该任务不仅是一个答案生成问题，也是一个复杂的跨模态证据选择与结构化排版问题。

2.2 问题分析

复杂长文档场景下，MRAMG的核心难点在于如何在高噪声候选内容中选择正确图像，并将其插入到合理位置。首先，检索结果往往包含弱相关文本、相似图像和冗余图像，单体多模态大模型在统一生成时容易受到视觉噪声干扰，造成无关图像误选。其次，受视觉Token预算和上下文窗口限制，现有方法通常依赖静态Top-N图像预筛选，但部分关键图像与查询的直接相似度较低，可能在初始阶段被过滤，导致后续生成无法恢复。最后，在操作手册、菜谱步骤和学术图表解析等多图任务中，模型不仅需要选对图像，还需要保证图像插入位置和相对顺序与文本逻辑一致。

因此，本文方法需要满足3个设计目标：1）在高噪声候选上下文中提升关键图像选择准确性，减少误选与漏选；2）在有限视觉上下文预算下缓解静态筛选造成的关键图像遗漏；3）在多图场景下提升图像插入位置与顺序组织的一致性。基于上述目标，下一节介绍本文提出的基于多智能体辩论机制的图文交织生成方法。

3 基于多智能体辩论的图文交织生成方法

3.1 框架总览

针对复杂长文档图文交织生成中候选图像噪声干扰、关键视觉证据遗漏以及多图顺序错位等问题，本文提出一种基于多智能体辩论机制的多模态检索增强图文交织生成框架。该框架打破单体多模态大模型一次性统一生成的范式，将生成过程划分为多模态检索、双轨草稿生成、冲突检测、多轮辩论与全局综合5个阶段，整体流程如图2所示。

首先，系统根据用户问题从多模态知识库中召回相关文本片段和候选图像，为后续图文生成提供外部知识支撑。随后，文本推理智能体和视觉感知智能体分别生成文本主导草稿与视觉主导草稿。其中文本推理智能体更侧重语义逻辑、步骤关系和答案连贯性，视觉感知智能体更侧重候选图像内容、局部视觉证据和图文匹配关系。2类草稿从不同模态视角形成互补判断，为后续冲突检测提供基础。

在获得双轨草稿后，系统进一步比较2条生成路径在图像选择和图像顺序上的差异，并将其转化为结构化冲突。具体而言，图像取舍不一致会形成集合冲突，图像排列或插入逻辑不一致会形成顺序冲突。通过这种方式，原本隐含在生成结果中的图像误选、漏选和顺序错位问题被显式建模为可分析、可辩论和可裁决的冲突实例。

针对检测到的冲突，系统触发多智能体辩论过程，由提出争议的一方进行举证，另一方进行质询，并在存在分歧时由全局裁决智能体进行最终判断。全局裁决智能体综合文本语义、视觉证据和辩论记录，对争议图像的保留、删除、补充以及顺序调整作出决策。最终，系统根据裁决结果对初始草稿进行局部修正，生成图像选择更准确、图文对应关系更清晰、步骤顺序更一致的图文交织答案。

需要说明的是，本节主要介绍整体框架组成、模块关系和信息流转过程。关于文档预处理与索引构建、候选图像预筛选、双轨草稿生成、冲突检测规则以及辩论裁决机制等具体实现细节，将在后续小节中分别展开说明。

3.2 多模态知识库索引构建与检索

为实现面向多模态文档的高效检索，本文首先构建统一的多模态知识库索引。对于原始多模态文档，系统使用特殊占位符⟨*img*⟩标记图像在文本中的物理位置，以保留原始图文结构信息。设预处理后的文档为C，文本切片长度设置$L=256$字符。随后将文档C按照固定长度划分为文本切片集合，记为：

$$C=\{c_1,c_2,\cdots,c_m\}, \tag{5}$$

其中，c_i表示第i个文本块，m表示文本切片总数。

在索引构建阶段，本文采用BGE-M3[30]嵌入模型对文本切片集合C中的所有文本切片进行统一编码。设$f_{\text{BGE-M3}}(\cdot)$表示BGE-M3文本编码函数，则向量数据库索引集合V可表示为：

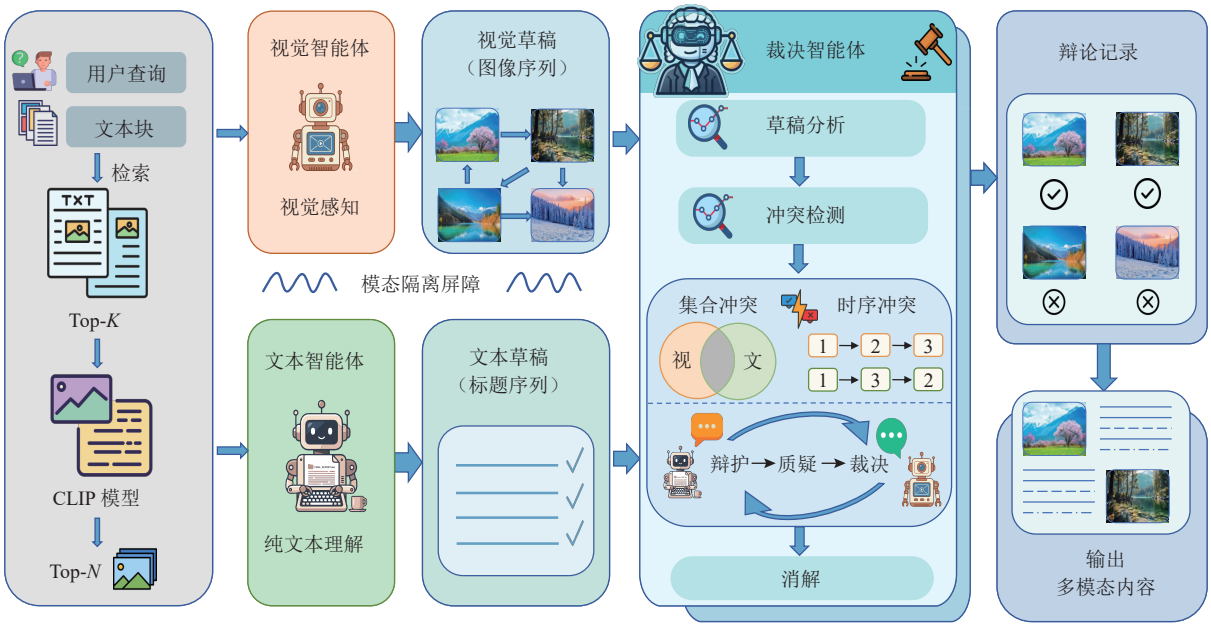


Fig. 2 Flowchart of the multi-agent debate-based interleaved text-image generation method

图2 基于多智能体辩论的图文交织生成方法流程图

$$V=\{f_{\text{BGE-M3}}(c_i)|c_i\in C\}, \tag{6}$$

其中，$f_{\text{BGE-M3}}(c_i)$表示文本切片c_i的向量表示。

在检索阶段，设用户查询问题为q，召回文本切片数量为K。系统首先使用 BGE-M3 对用户问题q进行编码，然后基于余弦相似度从向量索引集合V中召回 Top-K 个最相关文本切片，得到候选文本切片集合R：

$$R=TopK_{c_i\in C}(\cos(f_{\text{BGE-M3}}(q),f_{\text{BGE-M3}}(c_i))), \tag{7}$$

其中，K表示召回文本切片数量，$\cos(\cdot,\cdot)$表示余弦相似度函数，$TopK$算子表示根据余弦相似度由高到低，从候选文本切片集合中选择前K的文本切片。

进一步地，设$Ref(R)$表示由候选文本切片集合R中的⟨img⟩占位符恢复得到的原始图像引用集合，I_j表示第j张候选图像，则检索阶段召回的候选图像集合I_R可表示为：

$$I_R=\{I_j|I_j\in Ref(R)\}, \tag{8}$$

其中，I_R表示与候选文本切片集合R相关联的候选图像集合，I_j表示其中第j张候选图像，$Ref(R)$表示从召回文本切片中恢复图像引用关系的操作。

3.3 双轨独立草稿生成

为降低单体模型在长上下文条件下的跨模态决策耦合风险，本文设计双轨独立草稿生成机制。该机制由文本推理通道和视觉感知通道组成，分别从文本语义推理和视觉内容对齐 2 个角度生成初始图文草稿。

在获得候选文本块集合R与候选图像集合I_R后，系统首先为候选图像集合I_R中的每张图像提取对应的图像描述。设m_i表示候选图像I_i对应的语义描述文本，n表示候选图像数量，则候选图像描述集合M_R记为：

$$M_R=\{m_1,m_2,\cdots,m_n\}, \tag{9}$$

其中，m_i表示第i张候选图像I_i对应的语义描述文本，n表示候选图像数量。

考虑到候选图像数量较多时会造成视觉 Token 开销过大，本文在视觉输入侧引入基于 CLIP[31] 的图像预筛选机制。设$f_{\text{CLIP}}(\cdot)$表示 CLIP 编码函数，N表示预筛选后保留的图像数量，则从候选图像集合I_R中选取 Top-N 图像作为视觉输入集合I_{V}：

$$I_{\text{V}}=TopN_{I_i\in I_R}(\cos(f_{\text{CLIP}}(q),f_{\text{CLIP}}(I_i))), \tag{10}$$

其中，I_{V}表示输入视觉通道的候选图像子集，N表示预筛选后保留的图像数量，$\cos(\cdot,\cdot)$表示用户问题与候选图像之间的跨模态相似度。

随后，文本推理智能体和视觉感知智能体分别生成文本主导草稿D_{T}与视觉主导草稿D_{V}。其中，q表示用户问题，R表示候选文本切片集合，M_R表示候选图像描述集合，I_{V}表示经过 CLIP 预筛选后的视觉输入图像集合。双轨草稿生成过程可表示为：

$$D_{\text{T}}=LLM_{\text{T}}(q,R,M_R), \tag{11}$$

$$D_{\text{V}}=MLLM_{\text{V}}(q,R,I_{\text{V}}), \tag{12}$$

其中$f_{\text{T}}(\cdot)$表示文本推理智能体的生成函数，$f_{\text{V}}(\cdot)$表示视觉感知智能体的生成函数；D_{T}侧重保持文本叙事的逻辑与步骤连贯性；D_{V}侧重利用真实图像内容提升图像选择与局部图文对齐质量。2 类草稿的差异为后续冲突检测与辩论裁决提供依据。

3.4 冲突检测机制

在获得文本草稿D_{T}与视觉草稿D_{V}后，本文进一步设计冲突检测机制，将 2 条生成路径中的图像选择差异和图像顺序差异转化为结构化冲突实例。系统首先从 2 份草稿中分别提取文本通道图像表示序列S_{T}和视觉通道图像标识序列S_{V}，记为：

$$S_{\text{T}}=Extract(D_{\text{T}}), \tag{13}$$

$$S_{\text{V}}=Extract(D_{\text{V}}), \tag{14}$$

其中，$Extract(\cdot)$表示从草稿中提取图像占位符并按照其在答案中出现顺序组成序列的操作，S_{T}和S_{V}既记录图像选择结果，也保留图像在答案中的相对顺序。

基于上述 2 个序列，系统从图像选择一致性和图像时序一致性 2 个层面对潜在冲突进行扫描，并进一步实例化具体的冲突项。

1）集合冲突

集合冲突（set conflict）用于刻画 2 条生成路径在图像取舍上的差异。设$Set(\cdot)$表示将图像标识符序列转换为图像集合的操作，系统首先计算文本通道集合与视觉通道集合之间的单向差集：

$$I_{\text{T}\setminus\text{V}}=Set(S_{\text{T}})\setminus Set(S_{\text{V}}), \tag{15}$$

$$I_{\text{V}\setminus\text{T}}=Set(S_{\text{V}})\setminus Set(S_{\text{T}}), \tag{16}$$

其中，$I_{\text{T}\setminus\text{V}}$表示仅被文本智能体采用的图像集合，$I_{\text{V}\setminus\text{T}}$表示仅被视觉智能体采用的图像集合。

对于差集$I_{\text{T}\setminus\text{V}}\cup I_{\text{V}\setminus\text{T}}$中的每一张争议图像⟨$img$⟩，系统构造一个独立的集合冲突实例$c_{\text{set}}^{img}$。所有集合冲突实例共同构成集合冲突集合$C_{\text{set}}$：

$$C_{\text{set}}=\{c_{\text{set}}^{img}|img\in I_{\text{T}\setminus\text{V}}\cup I_{\text{V}\setminus\text{T}}\}, \tag{17}$$

其中，C_{set}表示集合冲突实例，c_{set}^{img}表示由争议图像⟨img⟩构造的单图集合冲突实例。该设计能够将图像取舍分歧细化到单图粒度，为后续围绕具体图像的存废辩论提供基础输入。

2）顺序冲突

顺序冲突用于刻画双方在图像排列顺序上的分歧。顺序比较需要建立在图像集合一致的基础上。因此，本文采用级联式检测策略。系统首先通过集合冲突辩论确定最终图像集合，然后由文本智能体和视觉智能体基于该共识集合重新生成修订草稿，分别记为D'_{T}与D'_{V}。随后，系统从2份修订草稿中提取完整图像标识符序列S'_{T}与S'_{V}：

$$S'_{\mathrm{T}} = Extract(D'_{\mathrm{T}}), \tag{18}$$

$$S'_{\mathrm{V}} = Extract(D'_{\mathrm{V}}), \tag{19}$$

其中，D'_{T}和D'_{V}分别表示集合冲突修正后的文本通道草稿与视觉通道草稿，S'_{T}和S'_{V}分别表示2份修订草稿中的完整图像标识符序列。

若满足$S'_{\mathrm{T}} \neq S'_{\mathrm{V}}$，则说明双方在图像相对顺序或插入逻辑上仍存在分歧，系统将2条完整序列构造为一个全局顺序冲突实例C_{order}。因此冲突集合C可表示为：

$$C = C_{\mathrm{set}} \cup \{C_{\mathrm{order}}\}, \tag{20}$$

其中，C_{set}表示集合冲突集合，C_{order}表示顺序冲突实例。若满足$S'_{\mathrm{T}} = S'_{\mathrm{V}}$，则说明双方在图像顺序上不存在分歧，此时不生成顺序冲突实例$C = C_{\mathrm{set}}$。最终得到的C_{order}将进一步触发面向时序一致性的对抗式辩论流程，以推动系统在图像排布顺序上达成最终一致。

3.5 多轮对抗式辩论机制

在获得冲突集合C后，系统对每个冲突实例执行多轮对抗式辩论，以逐步消解文本草稿与视觉草稿之间的决策分歧。设$c \in C$表示任意一个冲突实例，t表示当前辩论轮次，T表示最大辩论轮数，$r_c^{(t)}$表示冲突实例c在第t轮辩论后的阶段性结果。第t轮辩论过程可表示为：

$$r_c^{(t)} = F_{\mathrm{debate}}^{(t)}\left(c; D_{\mathrm{T}}, D_{\mathrm{V}}, q, R, M_R, r_c^{(t-1)}\right),\ t = 1, \cdots, T, \tag{21}$$

其中，$F_{\mathrm{debate}}(\cdot)$表示多轮辩论函数，$r_c^{(0)}$初始化为空。

在每轮辩论中，系统首先根据冲突来源确定举证方与挑战方。对于集合冲突，若争议图像仅出现在文本草稿中，则由文本智能体举证；若仅出现在视觉草稿中，则由视觉智能体举证。设$a_{\mathrm{def}}^{(t)}$表示第t轮举证方生成的论证结果，$a_c^{(t-1)}$表示上一轮挑战方的质询意见，则举证过程可表示为：

$$a_{\mathrm{def}}^{(t)} = f_{\mathrm{def}}\left(c, q, R, M_R, a_c^{(t-1)}\right), \tag{22}$$

其中，$f_{\mathrm{def}}(\cdot)$表示举证方的论证生成函数，a_{def}^{t}表示当前轮次的论证结果，$a_c^{(t-1)}$表示上一轮挑战方的质询意见，首轮时a_c^0为空。

特别地，为缓解静态Top-N预筛选可能造成的关键视觉证据遗漏问题，本文在辩论过程中引入动态视觉证据补充机制。对于集合冲突实例C，设$img(c)$表示该冲突对应的争议图像。若该争议图像来自检索阶段召回的候选图像集合，但未进入视觉感知智能体的初始输入集合，则说明该图像可能被静态预筛选过程遗漏。此时，系统允许在当前辩论过程中将其作为额外视觉证据进行动态补充。为刻画该过程，定义触发变量δ_c如下：

$$\delta_c = \begin{cases} 1, & img(c) \in I_R \backslash I_{\mathrm{V}}, \\ 0, & \text{其他}, \end{cases} \tag{23}$$

其中，δ_c表示冲突实例C是否触发动态补图操作。当$\delta_c = 1$时，说明争议图像未进入初始视觉Top-N结果，但仍属于当前检索上下文关联的候选图像，因此系统允许将其作为额外视觉证据补充到当前辩论过程；当$\delta_c = 0$时，不执行动态补图。

进一步地，为明确动态补图后实际引入的视觉证据，刻画当前争议图像是否被加入辩论过程，本文定义表示冲突实例c对应的动态补充视觉证据集合$I_{\mathrm{sup}}(c)$：

$$I_{\mathrm{sup}}(c) = \begin{cases} \{img(c)\}, & \delta_c = 1, \\ \varnothing, & \delta_c = 0。 \end{cases} \tag{24}$$

其中，当$\delta_c = 1$时，$I_{\mathrm{sup}}(c)$仅包含当前冲突实例对应的争议图像$img(c)$；当$\delta_c = 0$时，$I_{\mathrm{sup}}(c)$为空集。该设计使动态补图操作严格围绕当前争议图像展开，既能够补充被静态预筛选遗漏的关键视觉证据，又避免在辩论阶段无约束扩大候选图像范围，从而保证补图过程的针对性和可控性。

动态补充后，当前冲突对应的视觉上下文被扩展为：

$$I_{\mathrm{V}}^{+}(c) = I_{\mathrm{V}} \cup I_{\mathrm{sup}}(c), \tag{25}$$

其中，$I_{\mathrm{V}}^{+}(c)$表示针对冲突实例C扩展后的视觉证据集合。需要说明的是，动态补充图像并不会直接进入最终答案，而是作为当前冲突的额外视觉证据参与后续质询、复核和全局裁决。只有当该图像在辩论或裁决过程中被确认能够有效支撑文本内容、补充关键步骤或消解当前冲突时，才会被记录到全局辩论账本中，并参与最终综合生成。通过上述受限触发、单图补充和裁决复核过程，动态视觉证据补充机制能够在控制视觉噪声和推理成本的同时，恢复被静态Top-N预筛选遗漏的关键图像。随后，挑战方在相同的全局上下文基础上，接收举证方当前轮次生成的辩护意见$a_{\mathrm{def}}^{(t)}$，并生成结构化质询意见$a_c^{(t)}$：

$$a_c^{(t)} = f_{\text{Challenger}}\left(c,q,R,M_R,a_{\text{def}}^{(t-1)}\right), \tag{26}$$

其中，$f_{\text{Challenger}}(\cdot)$表示挑战方的质询生成函数，$a_c^{(t)}$表示第$t$轮挑战方生成的质询意见。若挑战方接受举证方论证，则当前冲突辩论提前结束。若挑战方拒绝该论证且当前轮次尚未达到最大轮数$\mathcal{T}$，系统进入下一轮辩论。若达到最大轮数后双方仍未达成一致，则由全局裁决智能体进行最终判定。

设r_c表示冲突实例c的最终裁决结果，$a_{\text{def}}^{(\mathcal{T})}$和$a_c^{(\mathcal{T})}$分别表示最大轮数下举证方与挑战方的最终论证结果。全局裁决过程可表示为：

$$r_c = f_{\text{Judge}}\left(c,q,R,M_R,a_{\text{def}}^{(\mathcal{T})},a_c^{(\mathcal{T})}\right), \tag{27}$$

其中，$f_{\text{Judge}}(\cdot)$表示全局裁决智能体的判定函数，r_c表示冲突实例c的最终裁决结果。

所有冲突实例的最终裁决结果被记录到全局辩论账本L：

$$L = \{r_c | c \in C\}。 \tag{28}$$

在完成所有冲突裁决后，全局裁决智能体基于辩论账本L执行最终综合生成。设A表示最终生成的图文交织答案，本文选取文本推理智能体生成的草稿D_{T}，或经过集合冲突修正后的草稿D_{T}'作为最终答案生成的基础底本，并根据账本中的裁决结果对其进行局部优化，最终输出：

$$A = f_{\text{Judge}}^{\text{synth}}(D_{\text{T}},L,q,R,M_R), \tag{29}$$

其中，$f_{\text{Judge}}^{\text{synth}}(\cdot)$表示全局综合生成函数，$A$表示最终生成的图文交织答案。

综合生成阶段主要包括3类操作：一是补充经辩论确认但未出现在待修正草稿中的关键图像；二是删除被裁定为冗余或无关的图像；三是根据顺序冲突裁决结果重排图像位置。这种局部一致性优化方式能够保持文本叙事主线的稳定。在此基础上，系统进一步对图像选择、插入位置和顺序关系进行精细化修正。最终，系统生成更加完整、准确且连贯的图文交织答案。

4 实验设置与结果分析

为验证本文所提出方法的有效性，本文在MRAMG-Bench数据集上开展系统实验。实验部分首先介绍数据集、对比方法、评价指标与实施细节，随后从主实验对比、裁判模型交叉验证、消融实验、动态补图机制、冲突触发与辩论行为分析、裁决智能体可靠性、参数敏感性、推理成本与可行性分析以及案例分析等方面进行系统评估。

4.1 数据集介绍

本文采用MRAMG-Bench数据集进行实验验证。该数据集是面向多模态检索增强多模态生成任务的综合性基准数据集，能够评估模型在图像数量判断、视觉证据选择、图像插入位置确定以及多图顺序组织等方面的能力[14]。该数据集共包含4 346篇多模态文档、14 190张图像和4 800个问答对，详细统计信息如表1所示。

Table 1 Statistics of the MRAMG-Bench Dataset

表1 MRAMG-Bench数据集统计信息

项目名称	Wit	Wiki	Web	Arxiv	Recipe	Manual	总计
难度等级	1	1	1	2	3	3	1，2，3
问题数量	600	500	750	200	2 360	390	4 800
文档数量	639	538	1 500	101	1 528	40	4 346
图像数量	639	538	1 500	337	8 569	2 607	14 190
文档平均图像数	1.00	1.00	1.00	3.34	5.61	65.18	3.27
文档平均长度	532.31	865.28	98.42	853.5	485.22	6 365.4	468.37
纯文本问答对	0	0	0	0	92	40	132
单图问答对	600	500	0	184	1 481	92	2 858
双图问答对	0	0	750	12	60	126	948
三图及以上问答对	0	0	0	3	727	132	862

从数据分布来看，MRAMG-Bench覆盖网络数据、学术文档和生活方式3类场景。其中，网络数据领域主要包含单图或双图问答，图文关系相对简单；学术文档领域包含论文图表、公式说明与实验结果分析等内容，对模型的图表理解和跨模态对齐能力提出更高要求；生活方式领域包含Recipe和Manual两个子集，具有更高的图像密度和更强的步骤时序依赖，尤其是Manual子集平均每篇文档包含65.18张图像，是验证复杂多图选择和顺序组织能力的关键场景。

与传统纯文本问答数据集不同，MRAMG-Bench不仅关注答案文本质量，还要求模型在生成过程中正确选择图像、确定插入位置并保持多图顺序一致。因此，该数据集能够较好地检验本文方法在复杂图文交织生成任务中的有效性。

4.2 对比方法与评价指标

为全面评估本文方法的性能，本文选取MRAMG-Bench原始基线及相关代表性方法进行对比。对比方法包括：

1）Rule-Based[14]。该方法基于启发式规则完成图像检索与插入位置确定，主要依赖文本匹配、图像距离和固定模板进行图文排版。

2)LLM-Based[14]。该方法将检索文本和图像描述拼接为长上下文输入大语言模型,由模型直接生成包含图像占位符的图文交织答案。

3)MLLM-Based[14]。该方法进一步引入真实图像输入,通常先利用 CLIP[31] 对候选图像进行 Top-N 筛选,再由多模态大模型完成文本生成和图像插入。

4)MuRAR[22]。该方法是一种两阶段多模态检索与答案精炼框架,先生成纯文本答案,再根据文本片段检索相关多模态内容,并通过大语言模型进行融合。

5)M2IO-R1[24]。该方法是一种强化学习增强的 MRAMG 推理框架,将图像插入过程建模为序列决策任务,并通过奖励函数优化图像选择与插入位置。

本文采用 MRAMG-Bench 官方评价指标进行评估[14]。客观指标包括图像精确率(image precision, Prec.)、图像召回率(image recall, Rec.)、图像 $F1$ 值(Image $F1$-score, $F1$)和图像顺序一致性(image ordering score, Ord.),其中,图像顺序一致性基于预测图像序列与标准图像序列之间的加权编辑距离计算,用于衡量生成图像顺序与真实答案顺序的匹配程度。文本质量指标采用 Rouge-L(R.L.)和 BERT Score(B.S.),分别衡量生成文本与参考答案在结构和语义层面的相似性。

除客观评价指标外,本文还采用大语言模型裁判评价方法[32],从图像相关性(Rel.)、图像有效性(Eff.)、图像位置准确性(Pos.)和整体回答质量(Comp.)对生成结果进行主观评价。其中,图像相关性用于衡量所选图像与问题及答案内容的匹配程度;图像有效性用于评价图像是否能够为答案提供必要的视觉支撑;图像位置准确性用于判断图像插入位置是否符合文本叙事逻辑;整体回答质量用于综合评估答案的完整性、连贯性和信息充分性。最终,本文采用平均性能(Avg.)指标对不同方法在图像选择、文本生成和图文排版等方面的综合表现进行比较。

4.3 实验环境与实施细节

为保证实验结果的公平性和可复现性,本文在统一计算环境下完成索引构建、文本检索、多智能体辩论推理和结果评估。主要计算设备为 NVIDIA A800 GPU。实验代码基于 Python 3.10 实现,模型调用、向量检索、图像预筛选和评价流程均在相同环境下完成。为便于复现,本文将在代码仓库中提供完整的依赖配置、数据处理脚本、模型调用接口和实验运行说明。

模型设置方面,本文选取闭源与开源多模态大模型进行实验。闭源模型包括 GPT-4o 和 GPT-4o-mini,主要通过 API 调用;开源模型采用 Qwen2-VL 系列,包括 Qwen2-VL-7B-Instruct 和 Qwen2-VL-72B-Instruct,并通过本地推理框架进行部署。不同对比方法与本文方法在相同基座模型和相同检索设置下进行评估,以保证实验比较的公平性。

核心超参数设置如表 2 所示。具体而言,文档切片长度设置为 256,文本检索阶段固定召回 Top-K=10 个文本块;视觉输入阶段采用 CLIP[31] 对候选图像进行相似度筛选,并保留 Top-N=10 张图像输入视觉通道;多轮辩论最大轮数设置为 $\mathcal{T}$ = 1;所有基于大语言模型裁判评价方法的主观评价均采用 GPT-4o 作为评估模型。上述设置与 MRAMG-Bench 官方评估流程保持一致,从而保证不同方法之间的可比性。

Table 2 Configuration of Key Hyperparameters

表 2 核心超参数设置

参数	取值	参数说明
Chunk	256	文档切分文本块的固定字符长度
Top-K	10	检索的文本片段数量
Top-N	10	基于 CLIP 筛选的候选图像数量
$\mathcal{T}$	1	辩论的最大迭代次数
LLM Evaluator	GPT-4o	LLM-as-a-Judge 使用的模型

4.4 实验结果

本节通过多维度实验结果分析所提方法的有效性与可行性。首先在 MRAMG-Bench 的网络数据、学术文档和生活方式 3 个领域上进行主实验对比,验证本文方法的整体性能;随后通过裁判模型交叉验证分析评价结果的稳定性;进一步通过消融实验、动态补图机制分析、冲突触发频率与胜方分布统计验证各模块作用;最后从参数敏感性和推理成本角度分析方法在不同配置下的性能变化与实际部署可行性。

需要说明的是,裁判模型交叉验证、辩论轮数 T 敏感性分析以及推理成本统计等实验采用从 MRAMG-Bench 三类场景中抽取的 100 条样本子集进行评估,其中 Web 数据 30 条,Arxiv 数据 30 条,Manual 数据 40 条。该子集覆盖网页内容、学术论文和操作手册 3 类典型图文交织生成场景。

1)对比结果分析

表 3~5 分别展示了不同方法在 MRAMG-Bench 网络数据、学术文档以及生活方式 3 个领域上的性能表现。整体来看,本文方法在 3 个领域中均取得了较优结果,说明所提出的多智能体辩论框架不仅

Table 3 Performance of Different Methods on the Network Data Domain of MRAMG-Bench

表 3 不同方法在 MRAMG-Bench 网络数据领域上的性能表现 %

算法	模型	Prec.	Rec.	Rel.	Comp.	Pos.	Avg.
Rule-Based[14]	GPT-4o	43.54	37.30	38.71	77.15	43.86	50.75
	GPT-4o-mini	38.20	33.08	33.85	76.59	38.57	46.87
	Qwen2-VL-7B-Instruct	37.98	34.81	30.40	70.99	38.28	46.32
	Qwen2-VL-72B-Instruct	34.81	31.49	30.80	71.89	35.14	43.75
MLLM-Based[14]	GPT-4o	82.75	80.57	79.55	85.95	84.65	79.66
	GPT-4o-mini	69.98	86.08	81.11	80.58	72.93	76.29
	Qwen2-VL-7B-Instruct	26.48	32.65	31.59	60.98	30.24	40.19
	Qwen2-VL-72B-Instruct	59.65	60.59	60.59	80.50	61.49	64.16
LLM-Based[14]	GPT-4o	80.86	77.92	75.93	85.74	81.41	77.73
	GPT-4o-mini	69.58	90.95	82.94	81.29	70.96	76.91
MuRAR[22]	GPT-4o	63.75	61.83	80.76	79.60	70.53	68.45
	GPT-4o-mini	61.44	59.69	77.76	78.32	65.21	63.49
M2IO-R1[24]	GPT-4o		96.53	**96.24**			
	GPT-4o-mini		97.47	89.20			
本文	GPT-4o	**97.17**	**99.03**	94.36	**95.66**	**97.25**	**87.56**
	GPT-4o-mini	94.32	98.86	92.64	94.00	96.33	85.92

注：黑体数值表示最优值。

Table 4 Performance of Different Methods on the Academic Document Domain of MRAMG-Bench

表 4 不同方法在 MRAMG-Bench 学术领域上的性能表现 %

算法	模型	Prec.	Rec.	Rel.	Comp.	Pos.	Avg.
Rule-Based[14]	GPT-4o	55.42	63.04	69.10	84.20	75.75	68.02
	GPT-4o-mini	51.71	59.29	67.50	85.20	73.75	66.07
	Qwen2-VL-7B-Instruct	49.17	52.17	60.70	78.90	67.08	60.88
	Qwen2-VL-72B-Instruct	45.42	48.71	60.30	79.60	65.42	59.50
MLLM-Based[14]	GPT-4o	60.39	74.29	89.40	87.50	90.39	76.87
	GPT-4o-mini	36.17	74.79	83.60	83.20	74.66	68.62
	Qwen2-VL-7B-Instruct	1.63	4.00	5.20	49.80	4.46	21.11
	Qwen2-VL-72B-Instruct	31.99	44.87	57.90	84.20	55.16	55.56
LLM-Based[14]	GPT-4o	65.28	76.54	86.00	88.90	84.84	76.91
	GPT-4o-mini	37.69	83.33	85.90	84.50	69.07	69.91
MuRAR[22]	GPT-4o	59.43	64.32	87.20	84.80	77.65	73.69
	GPT-4o-mini	57.48	65.18	83.90	81.60	77.04	71.51
M2IO-R1[24]	GPT-4o		84.20	**97.40**			
	GPT-4o-mini		83.30	96.70			
本文	GPT-4o	**66.95**	**89.75**	93.30	**97.70**	**93.08**	**79.64**
	GPT-4o-mini	64.22	86.99	89.40	97.00	87.76	76.82

注：黑体数值表示最优值。

能够提升图像选择质量，也能够在不同图文结构复杂度下保持较好的泛化能力。需要说明的是，M2IO-R1 原文仅报告了部分图像选择与顺序相关指标，且该方法通过 GRPO 训练专门的图像插入器，相关训练模型与完整复现代码未完全开源，复现成本较高。

具体来看，MRAMG-Bench 的网络数据领域包含 Wit、Wiki 和 Web 三个子集。由于这些子集中的图文关系相对直接且候选图像数量较少，传统的单模

Table 5 Performance of Different Methods on the Lifestyle Domain of MRAMG-Bench

表 5 不同方法在 MRAMG-Bench 生活方式领域上的性能表现 %

算法框架	模型	Prec.	Rec.	Rel.	Ord.	Comp.	Pos.	Avg.
Rule-Based[14]	GPT-4o	47.04	63.54	77.53	43.54	79.17	77.13	65.68
	GPT-4o-mini	49.14	61.52	75.12	43.87	79.32	75.89	64.92
	Qwen2-VL-7B-Instruct	43.79	60.93	77.11	39.52	77.70	77.33	63.87
	Qwen2-VL-72B-Instruct	31.82	49.30	73.02	25.51	72.92	70.94	56.45
MLLM-Based[14]	GPT-4o	43.77	44.68	77.05	32.47	81.36	77.35	61.01
	GPT-4o-mini	29.34	47.71	56.65	21.95	77.19	69.95	56.32
	Qwen2-VL-7B-Instruct	9.66	15.15	18.04	4.26	55.92	16.22	26.69
	Qwen2-VL-72B-Instruct	19.61	26.25	44.08	12.36	74.36	39.94	41.36
LLM-Based[14]	GPT-4o	47.48	62.40	81.87	42.55	80.60	**80.43**	66.11
	GPT-4o-mini	44.36	38.27	52.28	31.16	83.24	54.96	53.33
MuRAR[22]	GPT-4o	50.84	61.17	77.89	42.15	82.61	75.46	65.79
	GPT-4o-mini	50.93	60.54	75.81	40.40	80.77	72.92	64.79
M2IO-R1[24]	GPT-4o		65.98	**88.70**	43.61			
	GPT-4o-mini		65.60	86.09	39.30			
本文	GPT-4o	**57.09**	**73.00**	86.91	**51.38**	**88.11**	70.15	**66.16**
	GPT-4o-mini	54.41	63.89	74.36	44.26	86.15	65.11	62.24

注：黑体数值表示最优值。

型生成方法或规则匹配的方法已经能够取得较好的结果。然而，如表 3 所示，本文方法仍在图像精确率、图像位置得分和平均性能上取得明显优势。其原因在于网络数据虽然图像数量较少，但仍存在视觉相似图像或弱相关候选内容。本文通过文本推理智能体和视觉感知智能体分别从语义一致性和视觉证据匹配 2 个角度生成草稿，再由冲突检测与裁决机制对不一致结果进行修正，从而减少无关图像误选，并提升图像插入位置的合理性。因此，本文方法在相对简单场景中主要体现为对图像选择精度和排版稳定性的提升。

在学术文档领域，文档通常包含图表、公式说明和实验结果分析，文本与图像之间的对应关系更加隐蔽，模型需要从长文本上下文中识别与问题相关的图表证据。如表 4 所示，本文方法在图像召回率、图像位置得分和综合质量上表现突出，这说明多智能体辩论机制能够有效缓解学术文档中关键视觉证据被遗漏的问题。一方面，文本推理智能体能够根据问题和论文上下文判断所需图表的语义作用；另一方面，视觉感知智能体能够从候选图像中补充与图表内容相关的视觉证据。当 2 类草稿存在差异时，集合冲突检测和多轮辩论机制可以进一步对候选图像进行取舍判断，从而提高关键图像召回能力。该结果表明，本文方法在需要跨文本语义与图表内容进行细粒度对齐的场景中具有较强优势。

生活方式领域是 3 个领域中难度最高的场景，尤其是 Recipe 和 Manual 子集通常包含多步骤操作流程，图像之间存在明显的时序依赖。与网络数据和学术领域相比，该领域不仅要求模型选出正确图像，还要求图像插入顺序与操作步骤保持一致。如表 5 所示，本文方法在图像精确率、召回率和顺序一致性等指标上均取得较好表现，说明所提出的冲突检测与多轮辩论机制能够有效处理复杂多图场景中的图像漏选、冗余图像误选和顺序错位问题。特别是顺序一致性指标的提升表明，本文方法并非仅提高了图像覆盖率，而是能够根据文本步骤逻辑和视觉动作线索对图像顺序进行修正，从而增强最终图文答案的流程连贯性。

进一步比较不同类型方法可以发现，Rule-Based 方法依赖固定规则和浅层匹配，虽然在部分简单场景中具有一定稳定性，但难以处理长上下文和复杂图文对应关系；LLM-Based 方法缺少真实视觉输入，容易受到图像描述质量限制；MLLM-Based 方法虽然引入真实图像，但通常由单一模型一次性完成图像选择、位置判断和顺序规划，在候选图像较多时容易受到视觉噪声干扰；MuRAR 等流水线方法则依赖初始文本草稿质量，难以对早期遗漏的关键图像进行有效恢复。相比之下，本文方法通过文本推理、视觉

感知和全局裁决 3 类角色的协同建模，将原本隐式耦合的图像选择和顺序规划过程转化为可检测、可辩论和可裁决的冲突消解过程，因此能够在不同领域中取得更加稳定的性能提升。

综合来看，本文方法的优势并不仅体现在单一指标提升上，而且体现在图像选择、证据覆盖、位置匹配和顺序组织等多个环节的协同优化。网络数据领域的结果验证了方法在低复杂度场景下的稳定性，学术领域的结果体现了方法在图表证据召回和跨模态对齐方面的优势，生活方式领域的结果进一步说明该方法能够处理高图像密度和强步骤依赖场景下的复杂图文排布问题。上述结果充分验证了本文多智能体辩论框架在 MRAMG 任务中的有效性与适用性。

2）裁判模型交叉验证分析

为进一步缓解单一固定模型作为裁决智能体（LLM-as-a-Judge）可能带来的评价偏差，本文在主实验之外补充裁判模型交叉验证实验。具体而言，本文从 MRAMG-Bench 的 3 个领域中随机抽取 100 条样本，其中包含 30 条网络数据样本、30 条学术文档样本和 40 条生活方式样本。在保持生成结果完全一致的前提下，分别采用 GPT-4o、GPT-4o-mini 和 Qwen2-VL-7B-Instruct 作为裁判模型，对图像相关性、图像有效性、图像位置准确性和整体回答质量进行重新评分。实验结果如表 6 所示。

Table 6　Cross-Evaluation Results Under Different Judge Models

表 6　不同裁判模型下的交叉评价结果　%

裁判模型设置	Prec.	Rec.	*F*1	Ord.	R.L.	B.S.	Rel.	Eff.	Comp.	Pos.	Avg.
GPT-4o	**63.65**	**76.22**	**65.59**	**56.40**	**41.06**	**87.99**	86.80	77.00	**94.00**	**95.00**	74.37
GPT-4o-mini	63.65	76.22	65.59	56.40	41.06	87.99	**97.60**	93.20	91.80	95.00	**76.85**
Qwen2-VL-7B-Instruct	63.65	76.22	65.59	56.40	41.06	87.99	97.20	**96.00**	90.80	90.00	76.49

注：黑体数值表示最优值。

从表 6 可以看出，不同裁判模型对同一批生成结果均给出了较高的主观评价，说明本文方法生成的图文交织答案在不同评价模型下具有较好的稳定性。虽然 3 类裁判模型在具体评分尺度上存在一定差异，例如 GPT-4o 在图像有效性和图像相关性方面相对更为保守，而 GPT-4o-mini 和 Qwen2-VL-7B-Instruct 的评分整体更高，但它们对生成质量的整体判断并未出现明显冲突。

进一步来看，GPT-4o、GPT-4o-mini 和 Qwen2-VL-7B-Instruct 在图像位置准确性和整体回答质量上的评价较为一致，表明本文方法生成的答案在图像插入位置、文本连贯性和整体表达完整性方面具有较强稳定性。与此同时，引入与 GPT 系列不同源的 Qwen2-VL-7B-Instruct 后，主观评价结果仍保持较高水平，说明本文方法的实验结论并不依赖于单一 GPT-4o 裁判模型。

值得注意的是，GPT-4o 并未在该交叉验证实验中给出最高主观评价，反而在部分指标上表现得更加严格。因此，若主实验结果主要受到 GPT-4o 同源模型偏好的影响，则 GPT-4o 作为裁判时应更可能给出偏高评价；但实验结果并未呈现这一现象。这说明本文基于 LLM-as-a-Judge 的评价结果具有一定的跨模型一致性，能够在一定程度上缓解单一裁判模型带来的评价偏差问题。

由于本实验保持生成结果不变，仅替换裁判模型，因此图像精确率、召回率、*F*1、顺序一致性、Rouge-L 和 BERT Score 等客观指标不会随裁判模型变化。本文重点比较图像相关性、图像有效性、图像位置准确性和整体回答质量 4 类主观评价指标，以更直接地验证 LLM-as-a-Judge 评价过程的稳定性和可信度。

3）消融实验

为分析各核心模块的独立作用，本文在 MRAMG-Bench 生活方式领域中难度最高的 MRAMG-Manual 子集上开展消融实验。该子集平均每篇文档包含约 65 张候选图像，具有较高图像密度和明显步骤时序依赖，能够有效检验模型在复杂长文档场景下的图像选择、图文对齐与顺序组织能力。实验中，文本推理智能体、视觉感知智能体和全局裁决智能体均采用 GPT-4o-mini 作为基座模型，并设置单模型端到端生成基线 GPT-4o-mini 以及 4 种模块移除变体，结果如表 7 所示。

从整体结果来看，完整框架在图像精确率、召回率、*F*1 值、顺序一致性、图像相关性以及平均性能等关键指标上均取得最优结果，说明本文方法的优势主要体现在关键视觉证据选择和图文结构组织能力上。相比单模型端到端生成方式，本文框架分别利用文本推理智能体和视觉感知智能体建模语义逻辑与视觉证据。在此基础上，系统通过冲突检测与

Table 7 Performance Comparison of Different Modules on MRAMG-Manual Dataset in Ablation Study

表 7 不同模块在 MRAMG-Manual 数据集上性能的消融实验对比 %

设置	Prec.	Rec.	*F*1	Ord.	R.L.	B.S.	Rel.	Eff.	Comp.	Pos.	Avg.
本文	**44.42**	**60.19**	**44.87**	**39.02**	40.86	86.05	**85.08**	62.28	85.91	76.42	**62.51**
GPT-4o-mini	23.43	32.24	25.16	17.33	**43.60**	**91.05**	72.92	**71.13**	75.23	62.22	51.43
w/o Debate	34.06	43.86	33.75	28.08	39.94	87.23	74.56	56.14	83.04	64.47	54.51
w/o CLIP	41.44	59.26	42.56	37.19	39.14	84.10	83.94	60.48	84.05	75.52	60.77
w/o Order	43.20	59.87	44.19	37.14	40.76	85.94	84.23	61.91	**86.19**	**77.87**	62.13
w/o Decoupling	39.04	53.97	42.24	34.06	39.96	87.12	76.56	58.14	84.04	68.88	58.40

注：黑体数值表示最优值。

裁决机制显式处理图像选择和顺序分歧。因此，在高密度候选图像场景下，本文方法能够更有效地区分关键图像、冗余图像和弱相关图像。具体而言，移除多轮辩论机制后，系统虽然仍保留文本与视觉双轨草稿，但缺少对 2 条路径间分歧的显式协商，导致图像召回率、*F*1 值和顺序一致性明显下降，说明双智能体互补信息需要通过辩论与裁决过程才能有效转化为性能提升。移除模态解耦设计后，2 个智能体接收相同输入并产生相似判断，削弱了文本逻辑与视觉证据之间的互补性，使图像精确率、召回率和位置得分均出现下降，表明多智能体框架的有效性并非仅来自多个模型参与，而是来自角色分工和信息差异。

移除 CLIP 预筛选后，视觉智能体直接接收全部候选图像，虽然保留了更多原始视觉信息，但也引入大量冗余和弱相关图像，增加了模型对关键证据判断的干扰。因此，CLIP 预筛选能够在视觉 Token 预算有限的条件下有效控制输入噪声，并与后续动态补图机制共同在效率和完整性之间取得平衡。移除顺序冲突优化模块后，图像精确率和召回率变化相对有限，但顺序一致性下降，说明该模块主要作用于多图排布和步骤逻辑对齐。尽管 w/o Order 在综合回答质量和图像位置准确性上略高于完整框架，但其缺少对多图相对顺序的显式校正，容易在复杂步骤任务中保留潜在时序偏差。

此外，单模型 GPT-4o-mini 在 Rouge-L、BERT Score 和图像有效性指标上表现较高，主要是因为这些指标更偏向文本相似性或局部图像支撑性，并不能充分反映关键图像覆盖率和多图顺序组织能力。相比之下，本文方法在图像精确率、召回率、*F*1 值、顺序一致性和平均性能上更具优势，说明其更适合复杂图文交织生成任务。

综上，消融实验表明，本文框架中各模块承担了不同但互补的作用：模态解耦设计提供差异化的文本与视觉判断，多轮辩论机制负责将差异转化为可裁决的冲突实例，CLIP 预筛选机制控制候选视觉噪声，顺序冲突优化模块进一步提升多图流程组织能力。完整框架在多数核心图像指标和平均性能上的稳定优势说明，本文方法并非依赖单一模块提升结果，而是通过多模块协同实现复杂长文档场景下图像选择、视觉证据覆盖和图文顺序组织的整体优化。

4）动态补图机制

为验证动态视觉证据补充机制的有效性，本文构造 w/o Dynamic Supplement 变体。该变体保留双轨生成、冲突检测与全局裁决流程，但禁止系统在辩论过程中从初始 CLIP 之外补充图像。SIC 表示动态补图数量，SRR 表示动态补图保留率，ESHR 表示有效补图命中率，实验结果如表 8、表 9 和表 10 所示。

整体来看，动态补图机制的作用与任务复杂度密切相关。在网络数据领域中，图文对应关系相对直接，初始 CLIP 通常已能覆盖主要视觉证据，因此动态补图触发较少，性能提升也相对有限。而在学术文档领域和生活方式领域中，候选图像数量更多、图像语义差异更细微，多图之间还存在较强的步骤或论证关系，静态预筛选更容易遗漏关键图像，因此动态补图机制能够发挥更明显的作用。从指标表现看，动态补图主要提升的是关键视觉证据覆盖能力和整体图文组织稳定性，而不是单纯提高某一个局部指标。完整框架在多数场景下取得更高的召回率、*F*1 值和平均性能，说明辩论过程中补充的图像能够有效弥补初始 Top-*N* 候选集的覆盖不足。进一步结合 SRR 和 ESHR 可以发现，动态补图并不是无差别地扩大候选图像集合。

若补充图像只是增加输入噪声，则其最终保留率和命中率应当较低；但实验中学术文档领域和生活方式领域的 SRR 保持在较高水平，说明相当一部

Table 8 Ablation Study on the Dynamic Image Supplementation Mechanisms in Network Data Domain

表 8 在网络数据领域动态补图机制消融实验 %

实验设置	模型	Prec.	Rec.	Rel.	*F*1	Pos.	Avg.	SIC	SRR	ESHR
本文	GPT-4o	**97.17**	**99.03**	**94.36**	**97.33**	97.25	**96.95**	0.02	**0.61**	**0.67**
	GPT-4o-mini	94.32	98.86	92.64	95.13	94.00	94.96	0.11	0.36	0.11
	Qwen2-VL-7B-Instruct	24.75	26.79	87.03	25.10	**99.90**	52.71	**0.93**	0.43	0.17
w/o Dynamic Supplement	GPT-4o	96.13	98.37	66.90	97.12	58.86	80.57	0		
	GPT-4o-mini	92.26	98.41	92.73	94.06	99.29	96.67	0		
	Qwen2-VL-7B-Instruct	27.31	29.19	87.81	27.63	99.79	54.35	0		

注：黑体数值表示最优结果。

Table 9 Ablation Study on the Dynamic Image Supplementation Mechanisms in Academic Document Domain

表 9 在学术文档领域数据动态补图机制消融实验 %

实验设置	模型	Prec.	Rec.	Rel.	*F*1	Pos.	Avg.	SIC	SRR	ESHR
本文	GPT-4o	**66.95**	**89.75**	**93.30**	70.32	93.08	**85.77**	0.11	0.73	**0.31**
	GPT-4o-mini	64.22	86.99	89.40	70.90	87.76	82.10	0.26	**0.79**	0.29
	Qwen2-VL-7B-Instruct	17.69	33.43	76.35	21.20	96.69	49.07	**3.06**	0.38	0.10
w/o Dynamic Supplement	GPT-4o	64.49	89.00	92.10	**72.05**	**100**	85.39	0		
	GPT-4o-mini	62.52	86.01	89.45	69.88	87.94	81.48	0		
	Qwen2-VL-7B-Instruct	6.16	9.15	82.40	6.72	99.45	40.78	0		

注：黑体数值表示最优结果。

Table 10 Ablation Study on the Dynamic Image Supplementation Mechanisms in Lifestyle Domain

表 10 在生活方式领域数据动态补图机制消融实验 %

实验设置	模型	Prec.	Rec.	Rel.	*F*1	Pos.	Avg.	SIC	SRR	ESHR
本文	GPT-4o	**57.09**	**73.00**	**86.91**	**53.19**	70.15	**71.79**	0.66	**0.82**	0.18
	GPT-4o-mini	54.41	63.89	74.36	49.32	65.11	64.44	0.78	0.72	**0.27**
	Qwen2-VL-7B-Instruct	14.03	21.70	47.03	14.63	63.17	32.11	**6.55**	0.28	0.05
w/o Dynamic Supplement	GPT-4o	52.30	66.89	56.75	53.11	52.85	57.20	0		
	GPT-4o-mini	50.23	60.84	83.47	48.76	87.22	70.44	0		
	Qwen2-VL-7B-Instruct	25.28	37.81	75.18	26.35	**100**	52.92	0		

注：黑体数值表示最优结果。

分动态补充图像能够通过后续裁决进入最终答案。同时，ESHR 表明部分补充且保留的图像能够命中参考答案，说明该机制确实能够恢复被初始 CLIP 遗漏的有效视觉证据。因此，动态补图机制缓解了固定 Top-*N* 策略在复杂多图场景中的覆盖不足问题。

需要注意的是，部分单项指标并不总是由 GPT-4o 下的完整框架取得最高结果。这主要是因为不同指标关注的评价维度不同。动态补图会引入更多候选视觉证据，有利于提升召回率和整体完整性，但也会增加图像排序与插入位置判断的难度。因此，某些未补图或较弱补图的设置可能因输出更保守、图像数量更少，在位置得分或局部相关性指标上取得较高结果。然而，这类优势通常不能同时带来更好的召回率、*F*1 值和平均性能，说明其更多体现的是局部评价偏好，而非整体图文交织能力更强。

其次，不同基座模型对补充视觉证据的利用能力存在差异。GPT-4o 具有更强的跨模态理解和全局裁决能力，因此通常能够更有效地区分补充图像中的有效证据与干扰信息；而 GPT-4o-mini 或 Qwen2-VL-7B-Instruct 在面对新增图像时，可能出现证据利用不足或位置规划不稳定的问题。以 Qwen2-VL-7B-Instruct 为例，其动态补图后的 SIC 明显更高，但 Avg.并未同步提升，说明弱模型虽然更容易触发补图需求，但未必能够稳定完成“补充图像筛选—证据判断—顺序组织”的完整推理链条。因此，动态补图机制的收益依赖于模型是否具备足够强的视觉证据甄别和全局

组织能力。

综上, 动态补图机制能够有效缓解静态 CLIP 预筛选带来的关键图像遗漏问题, 尤其适用于候选图像密集、步骤依赖明显的复杂长文档场景。实验结果表明, 该机制通过辩论过程按需恢复有效视觉证据, 从而提升图像覆盖率、综合性能和图文交织生成的稳定性。

5) 冲突触发频率与发起方分析

为进一步分析所提出多智能体辩论框架在不同场景下的运行特性, 本文统计了 3 个领域子集上的平均冲突触发频率及集合冲突发起方比例, 结果如图 3 所示。

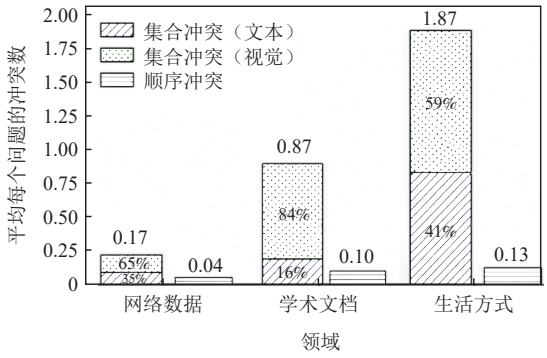


Fig. 3 Distribution of conflict types in MRAMG-Bench

图 3 冲突类型在 MRAMG-Bench 中的分布

从整体冲突触发频率来看, 不同领域之间呈现出显著差异。生活方式领域数据的平均集合冲突数量最高, 达到每个问题 1.87 次, 显著高于学术文档领域的 0.87 和网络数据领域的 0.17; 对应的顺序冲突频率也达到 0.13, 同样高于另外 2 个领域。该结果与前文主实验结论保持一致, 进一步验证了生活方式领域通常包含更加开放且多步骤的任务流程, 对图像插入顺序、步骤依赖关系以及局部排版逻辑提出了更高要求, 因此更容易触发跨模态分歧。

从集合冲突的发起方分布来看, 视觉智能体在 3 个领域中均承担了主要的冲突发起角色。其发起比例在网络数据、学术文档和生活方式 3 个领域分别达到 65%、84% 和 59%。尤其是在学术文档领域, 该比例最高, 说明视觉智能体能够更敏感地发现文本路径中遗漏的图表、流程图或实验结果图像。这一现象进一步表明, 在复杂图文文档场景下, 纯文本推理路径更容易遗漏关键视觉证据, 而视觉智能体能够在后续辩论中有效补充此类信息。

与此同时, 生活方式领域中文本智能体的发起比例提升至 41%, 明显高于另外 2 个领域。这说明在高时序依赖场景下, 文本路径对步骤流程和逻辑顺序更加敏感, 能够更频繁地识别图像插入位置或顺序偏差。

总体来看, 该统计结果进一步验证了双智能体设计的互补性: 视觉智能体更擅长补充缺失视觉信息, 而文本智能体更擅长发现流程逻辑偏差, 为后续多轮辩论提供了有效的冲突来源。

6)裁决智能体可靠性分析

为验证全局裁决智能体在多智能体辩论框架中的作用, 本文设计裁决策略替换实验, 比较完整方法(本文)、去除辩论裁决(w/o Debate)、文本智能体兼任裁决(Text-as-Judge)、视觉智能体兼任裁决(Visual-as-Judge)以及更强裁决(Strong Judge)模型等不同设置下的性能表现, 实验结果如表 11 所示。

从整体结果来看, 去除辩论裁决后模型性能明显下降, 说明在图文交织生成任务中, 文本草稿与视觉草稿之间的分歧不能仅依赖启发式规则进行简单合并。复杂长文档中往往同时存在图像选择冲突、图像顺序冲突和局部文本支撑不足等问题, 如果缺少显式裁决过程, 系统难以判断哪一侧证据更可靠, 也难以及时修正错误的图像选择和排布。因此, 裁决智能体的作用并不是简单“选择一个更好的草稿”, 而是对多智能体生成过程中的冲突进行综合判断和结构化消解。

进一步比较不同裁决主体可以发现, 由单侧智

Table 11 Reliability Experiment of the Judge Agent

表 11 裁决智能体可靠性实验 %

实验设置	裁决智能体设置	Pre.	Rec.	*F*1	Ord.	R.L.	B.S.	Rel.	Eff.	Comp.	Pos.	Avg.
本文	GPT-4o-mini	44.42	60.19	44.87	39.02	40.86	86.05	85.08	62.28	85.91	76.42	62.51
w/o Debate	启发式规则融合	34.06	43.86	33.75	28.08	39.94	**87.23**	74.56	56.14	83.04	64.47	54.51
Text-as-Judge	文本智能体	42.55	59.84	43.26	38.01	41.64	86.03	85.85	61.44	86.72	**80.26**	62.56
Visual-as-Judge	视觉智能体	35.70	55.00	37.45	32.88	**43.86**	86.29	83.28	60.62	87.28	72.31	59.47
Strong Judge	GPT-4o	**45.38**	**60.54**	**44.96**	**39.53**	42.12	86.08	**86.31**	**62.30**	**89.87**	77.64	**63.07**

注: 黑体数值表示最优值。

能体兼任裁决者会带来明显的模态偏置。Text-as-Judge 在位置得分和综合得分上表现较好，说明文本智能体更擅长维持叙事逻辑和图像插入位置的连贯性，但其图像选择和顺序一致性仍低于完整方法，表明单纯依赖文本侧判断容易忽略细粒度视觉证据。Visual-as-Judge 则更关注局部图像内容，但在全局步骤逻辑、文本语义衔接和多图顺序组织方面相对不足，因此整体表现不如独立裁决设置。这说明裁决智能体需要同时具备文本逻辑理解和视觉证据评估能力，而不能完全偏向某一模态。

此外，Strong Judge 的结果进一步说明，裁决模型的综合推理能力会直接影响最终图文交织质量。更强的裁决模型能够更准确地识别文本侧与视觉侧的有效分歧，并在图像筛选、冲突消解和顺序修正之间取得更稳定的平衡。因此，本文框架中的裁决智能体是连接双轨生成结果与最终图文答案的重要中枢，其可靠性决定了多智能体互补信息能否被有效整合。

综上，裁决智能体的核心价值在于对文本推理智能体和视觉感知智能体产生的分歧进行独立、综合的裁决。实验结果表明，独立裁决机制能够有效减少启发式融合带来的误判，同时缓解单侧裁决造成的模态偏置，从而提升复杂图文交织任务中的图像选择质量、顺序组织能力和整体生成稳定性。

7）辩论行为与胜方分布分析

为进一步分析多智能体辩论机制在不同领域中的决策行为，本文统计了不同冲突发起方对应的最终胜方分布，结果如图 4 所示。其中，文本发起冲突表示文本智能体认为当前图像选择或顺序安排与文本语义、步骤逻辑不一致；视觉发起冲突则表示视觉智能体认为当前结果存在视觉证据不足或图像内容不匹配等问题。

从整体结果来看，当冲突由文本智能体发起时，3 个领域中最终均以文本智能体获胜为主。这说明

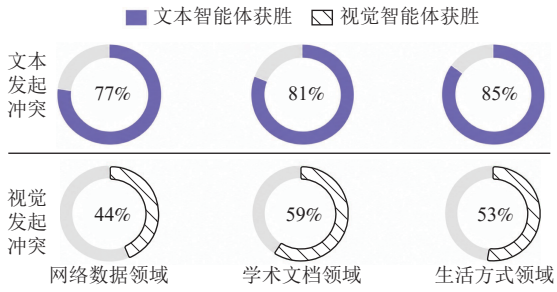


Fig. 4 Distribution of final win rates conducted by different conflict initiators across domains in multi-agent debates

图 4 不同领域的多智能体辩论中由不同冲突发起方主导的最终胜率分布

文本侧发起的冲突通常具有较强的全局逻辑依据。对于图文交织生成任务而言，图像是否应被选入以及应插入何处，不仅取决于局部图像内容，还需要服从文本叙事结构、步骤先后关系和任务回答逻辑。因此，当文本智能体主动指出冲突时，其判断往往反映的是全局语义组织层面的不一致，更容易被全局裁决智能体采纳。

相比之下，当冲突由视觉智能体发起时，最终胜方分布更依赖具体领域特征。网络数据领域的图文关系相对直接，文本描述通常已经能够较完整地表达核心信息，因此视觉侧虽然能够发现部分图像证据问题，但其裁决优势并不明显。学术文档领域中包含较多图表、流程图和实验结果可视化内容，许多关键信息需要依赖图像本身才能准确判断，因此视觉智能体在发起冲突后更容易获得裁决支持。生活方式领域则同时具有明确的文本步骤逻辑和较强的视觉操作依赖，文本与视觉证据的重要性相对均衡，因此视觉发起冲突后的胜方分布也更接近平衡。

该现象说明，多智能体辩论机制并不是简单地让某一类智能体占据主导，而是根据冲突来源和领域特征动态调整裁决依据。文本智能体更适合发现语义结构、步骤顺序和叙事连贯性问题，视觉智能体更适合发现图像内容、局部视觉证据和图文匹配问题。全局裁决智能体通过比较双方证据，能够在不同领域中形成差异化决策，从而避免固定规则融合带来的单侧偏置。

综上，胜方分布结果从行为层面验证了多智能体辩论框架的合理性：文本侧在全局逻辑冲突中具有更强优势，视觉侧在图像证据依赖更高的场景中能够发挥更大作用，而最终裁决结果会随领域特征和冲突类型动态变化。这表明本文框架能够有效利用文本推理与视觉感知之间的互补性，实现更加灵活和可靠的跨模态冲突消解。

8）参数敏感性分析

为进一步分析关键超参数对系统性能与推理效率的影响，本文分别对 CLIP 预筛选候选规模 Top-N 和最大辩论轮数$\mathcal{T}$进行参数敏感性实验。其中，Top-N 敏感性实验在 MRAMG-Manual 子集上进行，用于分析候选图像数量对视觉噪声、关键图像覆盖和推理开销的影响；辩论轮数$\mathcal{T}$敏感性实验在 Web、Arxiv 和 Manual 三类数据中共选取 100 条样本进行评估，用于分析辩论深度对生成质量与推理成本的影响。

① CLIP 预筛选 Top-N 参数敏感性分析

结合图 5(a)性能趋势来看，不同 Top-N 设置下

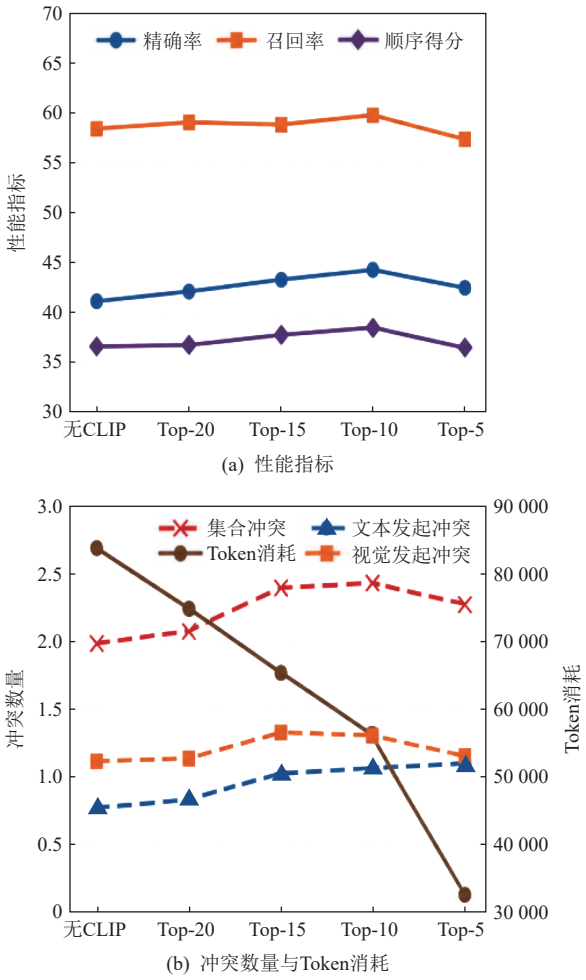


Fig. 5 Sensitivity analysis of the Top-*N* parameter in CLIP pre-filtering

图 5 CLIP 预筛选 Top-*N* 参数敏感性分析

系统表现呈现“先升后降”的变化规律，其中 Top-10 左右能够取得较好的综合效果。这说明 CLIP 预筛选并不是候选图像保留越多越好。当不进行预筛选时，大量弱相关或冗余图像会干扰视觉智能体的判断，使模型在草稿生成和后续辩论阶段更容易受到噪声影响。相反，当候选图像数量过少时，系统虽然减少了干扰信息，但也限制了关键视觉证据的可获取范围，导致部分有效图像无法进入后续推理流程。因此，中等规模的候选图像能够在信息充分性与噪声抑制之间取得更合理的平衡。

进一步结合图 5(b)可以发现，冲突数量与候选图像规模之间存在明显关联。当候选图像较多时，文本智能体与视觉智能体更容易基于不同证据形成分歧，因此集合冲突数量增加；而随着候选图像进一步减少，可供视觉智能体参考的信息变少，双方可形成的有效分歧也随之下降。该现象说明，多智能体辩论机制的有效性依赖于一定程度的信息差异，如果候选集过小，视觉侧能够提供的补充证据有限，则辩论机制难以充分发挥作用。

此外，Token 消耗随候选图像规模增加而显著上升。这是因为更多候选图像不仅会增加草稿生成阶段的视觉输入，还会在后续冲突检测与多轮辩论过程中持续参与比较与筛选。因此，CLIP 预筛选的作用不仅是提升性能，更重要的是控制长文档场景下的推理开销。实验结果表明，中等 Top-*N* 设置能够在保持较好图文生成质量的同时有效降低资源消耗，从而在性能与效率之间取得较优平衡。

② 辩论轮数$\mathcal{T}$参数敏感性分析

为分析辩论深度对生成质量与推理成本的影响，本文设置最大辩论轮数$\mathcal{T}=0,1,2,3$进行实验。其中，$\mathcal{T}=0$表示不执行多轮举证与质询，仅基于初始双轨生成结果进行综合；$\mathcal{T}>0$表示文本推理智能体与视觉感知智能体围绕冲突图像进行多轮举证、质询与裁决。实验结果如表 12 所示。从生成性能来看，当$\mathcal{T}=0$时，系统缺少显式的冲突消解过程，因此整体性能相对较低。随着辩论轮数增加，模型在 Prec.、Rec.、*F*1 和 Ord.等指标上整体呈提升趋势，说明多轮辩论有助于缓解图像选择错误、图像遗漏以及图像顺序不一致等问题。其中，$\mathcal{T}=2$ 和$\mathcal{T}=3$的性能明显优于$\mathcal{T}=0$和$\mathcal{T}=1$，表明更充分的证据交换和裁决过程能够提升图文交织生成质量。

从推理成本来看，引入辩论机制会增加模型调用次数(Calls)、Tokens 消耗和运行时间(Time)。例如，在 GPT-4o 设置下，Calls 从$\mathcal{T}=0$的 3.77 增加到$\mathcal{T}=1$的 5.98，Tokens 也由 14 774.40 增加到 25 694.02。这说

Table 12 Comparison of Performance and Inference Cost Under Different Debate Rounds $\mathcal{T}$

表 12 不同辩论轮数 $\mathcal{T}$ 下的性能与推理成本对比

辩论轮次$\mathcal{T}$	模型	Prec.	Rec.	*F*1	Ord.	Avg.	Calls	Tokens	Time/s
0	GPT-4o	54.22	63.19	64.47	58.09	59.99	3.77	14 774.40	8.60
1	GPT-4o	56.39	70.78	64.22	59.31	62.68	**5.98**	25 694.02	10.94
2	GPT-4o	71.05	78.38	71.91	66.69	72.01	4.78	**31 104.25**	13.78
3	GPT-4o	**76.36**	**84.31**	**77.78**	**72.71**	**77.79**	5.00	30 892.91	**17.52**

注：黑体数值表示最优值。Prec.、Rec.、*F*1 和 Ord.是越大越优，Calls、Tokens 和 Time 是越小越优。

明辩论过程虽然带来额外开销，但这些开销主要用于冲突图像的举证、质询与裁决。

综合来看，$\mathcal{T}=0$成本最低但性能有限；$\mathcal{T}=2$能够在性能和成本之间取得较好平衡；$\mathcal{T}=3$获得最高生成质量，但推理开销相对更高。因此，本文认为多轮辩论机制是有效的，并可根据实际应用需求在$\mathcal{T}=2$和$\mathcal{T}=3$之间进行选择。

综上，参数敏感性实验表明，候选图像规模和辩论轮数都会影响模型在生成质量与推理成本之间的权衡。一方面，CLIP 预筛选阶段存在合理的 Top-N 区间，适中的候选图像数量既能减少冗余视觉噪声，又能保留足够的关键图像，为后续冲突检测、动态补图和辩论裁决提供有效证据。另一方面，增加辩论轮数能够提升图像选择、关键图像召回和顺序一致性，但也会带来额外的模型 Calls、Token 消耗和运行时间。因此，本文方法可以通过调节 Top-N 和 $\mathcal{T}$ 在生成质量与推理效率之间取得灵活平衡，说明所提出框架具有较好的可调节性和实际应用价值。

9）推理成本与可行性分析

为进一步分析多智能体辩论机制带来的额外推理开销，并验证本文方法在实际部署中的可行性，本文统计了不同方法在平均模型 Calls、Token 消耗和运行时间方面的表现。实验数据从 3 类图文交织生成场景中共选取 100 条样本，其中 Web 数据 30 条、Arxiv 数据 30 条、Manual 数据 40 条，用于覆盖网页内容、学术论文和操作手册等不同类型文档。对比方法包括单模型生成方法 MLLM-Based、去除辩论过程的 w/o Debate，以及引入一轮多智能体辩论机制的本文方法。实验结果如表 13 所示，其中 Calls、Tokens 和 Time 均表示每条样本的平均值。

表 13 进一步从推理流程角度分析了不同方法在生成质量与推理开销之间的权衡关系。MLLM-Based 采用单模型一次性生成方式，流程最短，因此在 Calls、Token 消耗和 Time 上均具有明显优势。同时，由于该方法直接使用 GPT-4o 处理检索文本与候选图像，其本身是一个较强的多模态基线，因此在图像选择相关指标上能够取得较高结果。这说明对于部分图文关系较直接的样本，强单模型已经具备较好的图像筛选能力。

然而，单模型生成的优势主要来自其流程简洁性，而不是显式的冲突处理能力。它需要在 1 次推理过程中同时完成文本组织、图像选择、插入位置判断和顺序规划，当候选图像存在相似图像、冗余图像或关键图像被弱化时，模型缺少可解释的中间校正过程。相比之下，w/o Debate 虽然引入了文本推理智能体和视觉感知智能体的双轨生成，但由于没有后续辩论与裁决机制，2 条路径之间的差异无法被充分消解，因此性能并未稳定提升。这说明简单增加生成路径并不必然带来更好结果，关键在于如何将不同智能体产生的分歧转化为有效的修正信号。

在此基础上，本文方法（$\mathcal{T}$=1）进一步引入冲突检测、举证、质询和裁决过程，因此推理开销相应增加。但这种额外开销并非无效冗余，而是用于显式处理文本路径与视觉路径之间的图像选择和顺序分歧。与 w/o Debate 相比，本文方法在图像召回率和顺序一致性上均有所提升，说明辩论机制能够帮助系统恢复部分遗漏的关键视觉证据，并改善多图排列与文本逻辑之间的一致性。换言之，本文方法的核心价值不在于以最低成本完成生成，而在于通过可控的额外推理步骤提升复杂样本中的冲突消解能力。

综合来看，表 13 表明不同推理流程之间存在明显的效率与质量权衡：MLLM-Based 更适合对效率要求较高、图文关系相对简单的场景；w/o Debate 验证了仅有双轨生成仍不足以充分发挥多智能体协同优势；而本文方法通过辩论与裁决机制将额外调用成本转化为更强的视觉证据补充和顺序优化能力。结合表 12 中更高辩论轮数下的结果可以进一步看出，当允许更充分的辩论过程时，本文方法的整体生成质量仍有进一步提升空间。因此，多智能体辩论机制提供了一种可调节的生成框架，能够根据实际应用需求在推理效率和生成质量之间进行灵活折中。

10）案例分析

为进一步展示本文多智能体辩论框架在细粒度

Table 13 Comparison of Performance and Inference Cost across Different Inference Pipelines

表 13 不同推理流程下的性能与推理开销对比

实验设置	模型	Prec.	Rec.	*F*1	Ord.	Calls	Tokens	Time/s
MLLM-Based	GPT-4o	**64.32**	**73.55**	**64.73**	55.94	1.00	4 919.58	6.22
w/o Debate	GPT-4o	54.22	63.19	64.47	58.09	3.77	14 774.40	8.60
本文（$\mathcal{T}$=1）	GPT-4o	56.39	70.78	64.22	**59.31**	**5.98**	**25 694.02**	**10.94**

注：黑体数值表示最优值。Prec.、Rec.、*F*1 和 Ord.越大越优，Calls、Tokens 和 Time 是越小越优。

视觉证据补充与冲突消解中的作用，本文选取MRAMG-Manual子集中的典型案例进行定性分析，如图6所示。问题为：“When using my drill, how can I install keyless single sleeve chuck?”，属于典型的多步骤操作类任务，对图像选择完整性与步骤时序一致性均具有较高要求。

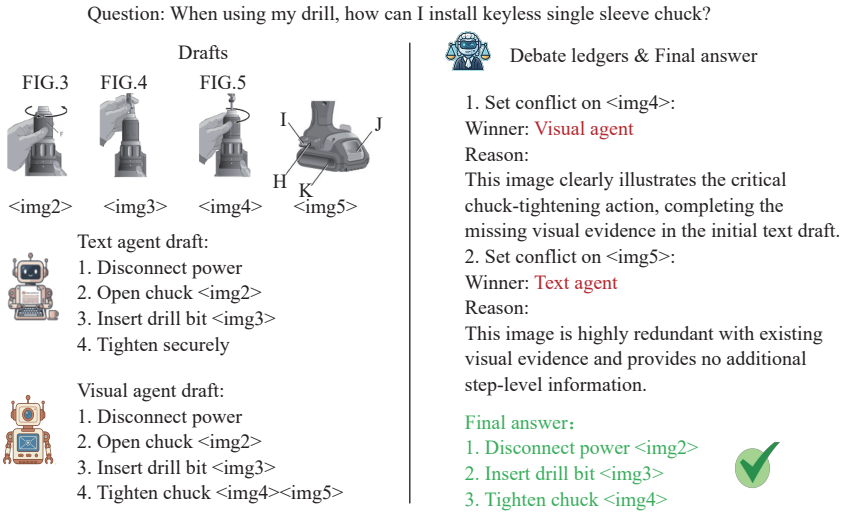


Fig. 6 Case study

图6 案例分析

在初始草稿生成阶段，文本推理智能体能够依据文本上下文构建较清晰的步骤主线，并正确识别“打开夹头”和“插入钻头”等主要操作。然而，文本侧主要依赖语义和步骤逻辑进行推理，对于“旋紧夹头”这一细粒度动作缺少直接视觉支撑，因此遗漏了对应的关键图像。相比之下，视觉感知智能体能够从候选图像中捕捉到更具体的动作细节，补充了与“旋紧夹头”高度相关的图像，但同时也引入了语义相近、信息增益有限的冗余图像。

该案例体现出双轨生成的互补性：文本智能体更擅长保证操作流程的完整性和语言逻辑，视觉智能体更擅长发现文本推理中遗漏的局部视觉证据。但如果仅将两者结果直接合并，容易同时保留关键补充图像和冗余图像，导致最终答案出现视觉重复或步骤干扰。因此，后续的冲突检测与辩论裁决是保证结果质量的关键环节。

在辩论阶段，系统将视觉侧新增的图像识别为集合冲突，并交由全局裁决智能体进行判断。裁决结果保留了能够展示关键旋紧动作的图像，剔除了与已有步骤图像语义重叠的冗余图像。最终答案形成了“断电—插入钻头—旋紧夹头”的完整图文流程，与标准答案保持一致。

该案例说明，本文方法并不是简单增加图像数量，而是通过文本推理、视觉补充和全局裁决之间的协同，实现关键视觉证据的恢复与冗余信息的过滤。由此可见，所提出的多智能体辩论框架能够在复杂操作类图文交织任务中有效提升图像选择的完整性、步骤组织的合理性和最终答案的可解释性。

5 结 论

本文针对多模态检索增强图文交织生成任务中候选图像噪声干扰、关键视觉证据遗漏以及多图顺序错位等问题，提出了一种基于多智能体辩论机制的图文交织生成方法。该方法将传统单体多模态大模型的一次性统一生成过程解耦为文本推理、视觉感知和全局裁决3个协同角色，并通过双轨草稿生成、集合冲突与顺序冲突检测、多轮对抗式辩论、动态视觉证据补充以及全局裁决与融合机制，对图像选择、冗余剔除、关键图像补充和顺序重排进行统一优化，从而将隐式的图文生成偏差转化为可检测、可辩论和可裁决的结构化冲突。在MRAMG-Bench数据集上的实验结果表明，本文方法在网络数据、学术文档和生活方式3个领域均取得较优表现，能够有效提升图像选择、关键图像召回、插入位置合理性和多图顺序一致性。进一步分析表明，方法性能提升主要来源于文本逻辑建模、视觉证据感知和全局裁决机制之间的协同作用，其中多轮辩论与裁决机制对消解跨模态分歧具有重要作用。推理成本分析表明，本文方法相比单模型生成方法会引入额外

模型调用和 Token 消耗，但该开销主要来自冲突检测、举证、质询和裁决过程，整体仍处于可控范围，说明所提出框架具有一定实际部署可行性。未来工作将进一步探索更加轻量化的多智能体协同机制和更加稳定的跨模态裁决策略，以及结合外部知识与长期记忆的开放域图文交织生成方法。

作者贡献声明：马杰负责研究方案设计、方法框架构建、论文部分内容撰写及修改工作；张瀚驰、渠宁参与实验方案制定、算法实现、实验结果分析，并负责论文部分内容撰写及修改工作；薛皓荃、王鑫平负责实验运行、结果整理与分析；刘均负责研究思路指导、论文结构把关与最终修改。所有作者均阅读并同意论文最终版本。张瀚驰和渠宁对本文工作贡献相同，为共同第二作者。

参 考 文 献

[1] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report[EB/OL]. 2023[2026-05-12]. https://arxiv.org/abs/2303.08774

[2] Zhao W X, Zhou Kun, Li Junyi, et al. A survey of large language models[EB/OL]. 2023[2026-05-12]. https://arxiv.org/abs/2303.18223

[3] Shao Zhihong, Wang Peiyi, Zhu Qihao, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models[EB/OL]. 2024[2026-05-12]. https://arxiv.org/abs/2402.03300

[4] Hirschberg J, Manning C D. Advances in natural language processing[J]. Science, 2015, 349(6245): 261−266

[5] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1−38

[6] Liu Zeyuan, Wang Pengjiang, Song Xiaobin, et al. A survey on hallucination in large language models[J]. Journal of Software, 2025, 36(3): 1152−1185 (in Chinese)
(刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025, 36(3): 1152−1185)

[7] Gao Yunfan, Xiong Yun, Gao Xinyu, et al. Retrieval-augmented generation for large language models: A survey[EB/OL]. 2023[2026-05-12]. https://arxiv.org/abs/2312.10997

[8] Yuan Le, Liu Shaohua, Wang Yu, et al. A survey on retrieval-augmented generation optimization technologies for large language models[J]. Chinese Journal of Computers, 2026, 49(2): 383−422 (in Chinese)
(袁乐, 刘绍华, 王禹, 等. 大语言模型检索增强生成优化技术研究综述[J]. 计算机学报, 2026, 49(2): 383−422)

[9] Chen Wenhu, Hu Hexiang, Chen Xi, et al. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 5558−5570

[10] Yu Shi, Tang Chaoyue, Xu Bokai, et al. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents[C/OL]//Proc of the 13th Int Conf on Learning Representations. Singapore: OpenReview. net, 2025[2026-05-12]. https://openreview.net/forum?id=zG459X3Xge

[11] Guo Zirui, Ren Xubin, Xu Lingrui, et al. RAG-Anything: All-in-one RAG framework [EB/OL]. 2025[2026-05-12]. https://arxiv.org/abs/2510.12323

[12] Abootorabi M M, Zobeiri A, Dehghani M, et al. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation[C]//Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA: ACL, 2025: 16776−16809

[13] Yin Shukang, Fu Chaoyou, Zhao Sirui, et al. A survey on multimodal large language models[J]. National Science Review, 2024, 11(12): nwae403

[14] Yu Qinhan, Xiao Zhiyou, Li Binghui, et al. MRAMG-Bench: A comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. [C]//Proc of the 48th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2025: 3616−3626

[15] Zhang Duzhen, Yu Yahan, Dong Jiahua, et al. MM-LLMs: Recent advances in multimodal large language models[C]//Findings of the Association for Computational Linguistics: ACL 2024. Stroudsburg, PA: ACL, 2024: 12401−12430

[16] Liu Haotian, Li Chunyuan, Li Yuheng, et al. Improved baselines with visual instruction tuning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 26296−26306

[17] Wang Peng, Bai Shuai, Tan Sinan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution[EB/OL]. 2024[2026-05-12]. https://arxiv.org/abs/2409.12191

[18] Liang Tian, He Zhiwei, Jiao Wenxiang, et al. Encouraging divergent thinking in large language models through multiagent debate[C]//Proc of the 2024 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2024: 17889−17904

[19] Du Yilun, Li Shuang, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate[C]//Proc of the 41st Int Conf on Machine Learning. Vienna: PMLR, 2024: 11733−11763

[20] Yasunaga M, Aghajanyan A, Shi W, et al. Retrieval-augmented multimodal language modeling[C]//Proc of the 40th Int Conf on Machine Learning. Honolulu: PMLR, 2023: 39755−39769

[21] Bi Fenglin, Zhang Qiming, Zhang Jiarui, et al. A retrieval-augmented generation system based on sliding window strategies in large language models[J]. Journal of Computer Research and Development, 2025, 62(7): 1597−1610 (in Chinese)
(毕枫林, 张豈明, 张嘉睿, 等. 基于滑动窗口策略的大语言模型检索增强生成系统[J]. 计算机研究与发展, 2025, 62(7): 1597−1610)

[22] Zhu Zhengyuan, Lee D, Zhang Hong, et al. MuRAR: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering[C]//Proc of the 31st Int Conf on Computational Linguistics: System Demonstrations. Stroudsburg, PA: ACL, 2025: 126−135

[23] Yan Meng, Xu Cai, Huang Haibin, et al. Large language model-based

trusted multi-modal recommendation algorithm[J]. Journal of Computer Research and Development, 2025, 62(7): 1611−1621 (in Chinese)
(闫萌, 徐偲, 黄海槟, 等. 基于大语言模型的可信多模态推荐算法[J]. 计算机研究与发展, 2025, 62(7): 1611−1621)

[24] Xiao Zhiyou, Yu Qinhan, Li Binghui, et al. M2IO-R1: An efficient RL-enhanced reasoning framework for multimodal retrieval augmented multimodal generation[EB/OL]. 2025[2026-05-12]. https://arxiv.org/abs/2508.06328

[25] Li Xu, Zhu Rui, Chen Xiaolei, et al. Survey of hallucinations in large vision-language models: Causes, evaluations and mitigations[J]. Journal of Computer Research and Development, 2025, 62(12): 2929−2950 (in Chinese)
(李煦, 朱睿, 陈小磊, 等. 视觉语言大模型的幻觉综述: 成因、评估与治理[J]. 计算机研究与发展, 2025, 62(12): 2929−2950)

[26] Zeng Cheng, Ge Yunjie, Zhao Lingchen, et al. Survey of multimodal vision-language representation learning models and their adversarial examples attack and defense techniques[J]. Journal of Computer Research and Development, 2025, 62(9): 2208−2232 (in Chinese)
(曾诚, 葛云洁, 赵令辰, 等. 多模态视觉语言表征学习模型及其对抗样本攻防技术综述[J]. 计算机研究与发展, 2025, 62(9): 2208−2232)

[27] Li Guopeng, Wu Ruiqi, Tan Haisheng, et al. A plan reuse mechanism for LLM-driven agent[J]. Journal of Computer Research and Development, 2024, 61(11): 2706−2720 (in Chinese)
(李国鹏, 吴瑞骐, 谈海生, 等. 面向大语言模型驱动的智能体的计划复用机制[J]. 计算机研究与发展, 2024, 61(11): 2706−2720)

[28] Zhao Botao, Kang Zuheng, Qu Xiaoyang, et al. Research progress and prospects of embodied intelligent agents based on multimodal large models[J]. Big Data, 2025, 11(3): 108−138 (in Chinese)
(赵博涛, 亢祖衡, 瞿晓阳, 等. 基于多模态大模型的具身智能体研究进展与展望[J]. 大数据, 2025, 11(3): 108−138)

[29] Qiu Longpeng, Chen Bo, Zhao Zeya, et al. ZhiShuOSINT: A deep research based multi-agent intelligence analysis framework[J]. Big Data, 2026, 12(1): 61−70 (in Chinese)
(丘龙鹏, 陈波, 赵泽亚, 等. 智枢 OSINT: 基于深度研究的多智能体情报分析框架[J]. 大数据, 2026, 12(1): 61−70)

[30] Chen Jianlv, Xiao Shitao, Zhang Peitian, et al. M3-Embedding: Multi-linguality, multi-functionality, multigranularity text embeddings through self-knowledge distillation[EB/OL]. 2024[2026-05-12]. https://arxiv.org/abs/2402.03216

[31] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. Virtual Event: PMLR, 2021: 8748−8763

[32] Gu Jiawei, Jiang Xuhui, Shi Zhichao, et al. A survey on LLM-as-a-Judge[J]. The Innovation, 2024, 5(4): 100734



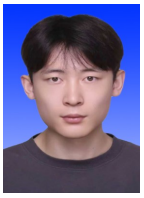
Ma Jie, born in 1993. Associate professor, PhD supervisor. Member of CCF. His main research interests include multimodal content understanding, large model reasoning, and multi-agent systems.

马　杰, 1993 年生。副教授, 博士生导师。CCF 会员。主要研究方向为多模态内容理解、大模型推理、多智能体系统。



Zhang Hanchi, born in 2001. PhD candidate. His main research interest includes self-evolution of multi-agent systems.

张瀚驰, 2001 年生。博士研究生。主要研究方向为多智能体系统自演化。



Qu Ning, born in 2000. Master candidate. His main research interests include knowledge graphs and large language models.

渠　宁, 2000 年生。硕士研究生。主要研究方向为知识图谱、大语言模型。



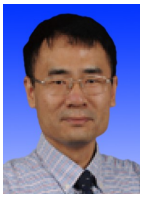
Xue Haoquan, born in 2002. Master candidate. His main research interest includes multimodal RAG.

薛皓荃, 2002 年生。硕士研究生。主要研究方向为多模态 RAG。



Wang Xinping, born in 2004. Undergraduate. Her main research interest includes large language models.

王鑫平, 2004 年生。本科生。主要研究方向为大语言模型。



Liu Jun, born in 1973. Professor, PhD supervisor. Senior member of CCF. His main research interests include natural language understanding, computer vision, and smart education.

刘　均, 1973 年生。教授, 博士生导师。CCF 高级会员。主要研究方向为自然语言理解、计算机视觉、智慧教育。