多模态检索增强生成驱动的文档问答综述(特邀)

李泽鸣,王树良,尚子贺,盛明
(北京理工大学计算机学院,北京 100081)

摘 要: 传统检索增强生成(RAG)方法主要面向纯文本场景,其检索与生成机制难以有效建模多模态文档中普遍存在的视觉元素、空间布局与结构语义,在图文混合、长文档及跨文档推理任务中表现受限。为此,多模态检索增强生成(MRAG)通过联合建模文本、图像与版式结构,在生成过程中引入多模态证据检索与调度,已然发展为视觉富文档问答与推理的核心技术范式。本文系统综述 MRAG 在文档问答任务中的研究进展。首先,围绕多模态文档理解的实际需求,分析 MRAG 在多模态对齐、长上下文建模、证据可追溯性及系统鲁棒性等面临的关键挑战。其次,立足 MRAG 系统支持生成过程的方式,分别从嵌入范式、文档检索范围、布局感知机制与多模态检索策略 4 个维度,梳理对比代表性方法,聚焦讨论不同设计选择对生成稳定性、推理精度与系统复杂度的影响。再次,总结现有多模态文档问答数据集与评测体系的特点与不足,分析当前评测在多模态证据粒度与推理可解释性方面的局限。最后,指出 MRAG 正由面向静态相似度匹配的检索机制,演进为以生成与推理需求为中心的动态证据规划范式,应通过多模态、多粒度协同建模,持续提升复杂文档问答系统的可靠性与可解释性。

关键词: 多模态文档;多模态检索增强生成;文档问答;生成驱动检索;布局感知建模;多模态推理
中图分类号: TP391 **文献标志码:** A **DOI:** 10.19678/j.issn.1000-3428.0260043

Review of Document Q&A Driven by Multimodal Retrieval-Augmented Generation (Invited)

LI Zeming, WANG Shuliang, SHANG Zihe, SHENG Ming
(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

【Abstract】 Traditional Retrieval-Augmented Generation (RAG) methods predominantly focus on pure-text scenarios. In these scenarios, their retrieval and generation mechanisms encounter difficulties in effectively modeling common visual elements, spatial layouts, and structural semantics within multimodal documents. This drawback restricts their performance in tasks related to text-image hybridization, long documents, and cross-document reasoning. To tackle this issue, Multimodal Retrieval Augmented Generation (MRAG), by integrating text, image, and layout structure modeling, and incorporating multimodal evidence retrieval and scheduling during the generation process, has already developed into a core technical paradigm for Question & Answer (Q & A) and reasoning in visually-rich documents. This paper conducts a systematic review of research progress in MRAG applications for document Q & A tasks. Firstly, based on the practical requirements for multimodal document understanding, we analyze the key challenges in MRAG implementation, including multimodal alignment, long-context modeling, evidence traceability, and system robustness. Secondly, from the perspective of how MRAG systems support the generation process, we compare representative methods across four dimensions: embedding paradigms, document retrieval scope, layout-aware mechanisms, and multimodal retrieval strategies. We focus on how design choices influence generation stability, reasoning accuracy, and system complexity. Thirdly, we summarize the characteristics and limitations of existing multimodal document Q & A datasets and evaluation frameworks, and analyze the current constraints in evidence granularity and reasoning explainability. Finally, we point out that MRAG is evolving from static similarity-matching retrieval mechanisms to dynamic evidence planning paradigms centered on generation and reasoning needs, and should continuously enhance the reliability and explainability of complex document Q & A systems through collaborative multimodal modeling with multi-granularity approaches.

【Key words】 multimodal document; Multimodal Retrieval-Augmented Generation (MRAG); document Question & Answer (Q&A); generation-driven retrieval; layout-aware modeling; multimodal reasoning

基金项目: 国家自然科学基金(42371480, 62306033)。
作者简介: 李泽鸣,男,博士研究生,主研方向为数据智能;王树良(通信作者),教授;尚子贺,本科生;盛明,博士研究生。
收稿日期: 2026-01-09 **修回日期:** 2026-02-12 **E-mail:** slwang2011@bit.edu.cn

0 引言

随着预训练生成模型在大规模数据上的快速发展,如何提升模型在知识密集型任务中的实时性与可靠性成为自然语言处理领域的重要研究方向。检索增强生成(RAG)通过引入外部知识库,使生成模型能够在推理过程中利用检索得到的证据,有效缓解参数化模型在复杂问答(Q&A)任务中的信息缺失与幻觉问题[1]。早期研究主要面向文本模态,通过构建稠密检索向量空间、优化检索排序策略等方式提升检索与生成之间的协同效果,逐渐形成了由检索器、知识库与生成器组成的典型体系结构[2-3]。相关研究还探索了检索与模型预训练的深度结合,使模型在训练阶段学习如何主动利用外部知识库,提高知识获取能力。

然而,真实世界的信息载体并不限于纯文本,报告、报表、合同、学术论文等视觉富文档占据了大量知识密集型任务的核心场景。这类文档同时包含文本、图形图像、表格与复杂版式结构,仅依赖文本检索难以充分表达其语义特征。为此,多模态检索增强生成(MRAG)成为新的研究趋势[2]。在 MRAG 中,模型需要同时建模视觉特征、文本内容及空间版式结构,并在检索阶段对多模态证据进行排序,在生成阶段融合跨模态信息,显著提升了文档理解与推理的能力。一般的多模态检索框架从输入的自然语言查询出发,通过多模态嵌入对文档中的文本、图像与版式信息进行表示,并在检索阶段从大规模视觉文档库中筛选与查询相关的多模态证据。随后,生成模块在融合检索结果的基础上执行跨模态推理,生成最终答案。该统一框架为后续不同方法在嵌入方式、检索粒度、布局感知与推理机制上的设计差异提供了共同的分析视角。

在视觉文档理解研究方向,视觉文档问答(VQA)任务及其数据集的提出推动了该领域的发展。以 DocVQA[3]等基准为代表的研究体系,使模型能够处理文档图像中的文字识别、版式理解及跨区域推理等任务。随后,研究者提出了一系列联合文本与版式的预训练模型,如 LayoutLM[4]、LayoutLMv2[5],通过引入文本框位置编码、视觉特征与跨模态自注意力机制,使模型在表格识别、文本理解等场景中取得显著性能提升。与此同时,端到端文档理解模型逐渐兴起,通过用图像编码器替代传统的光学字符识别(OCR),模型能够更直接地从文档图像中学习视觉-文本对齐关系,代表性方法包括 DocFormer、Donut 等[6-7]。这些研究推动了文档理解从“文本管道式处理”向“统一多模态建模”方向演进。

在推理范式方面,链式思维(CoT)提示、自问式推理等方法为复杂任务提供了分解机制[8]。基于智能体思想的框架进一步将检索、推理及工具调用整合为多步骤过程,实现了可解释的决策链条。近年来,相关思想被引入视觉文档领域。代表性工作 ViDoRAG[9]提出在视觉文档检索与生成任务中引入多智能体结构,通过多模态检索、证据筛选与迭代推理的协同,实现跨页、多文档的推理过程。该类方法表明,MRAG 在真实复杂文档理解场景中具有显著潜力。

尽管 MRAG 取得了快速进展,但在视觉文档场景下仍面临若干关键挑战。首先,多模态对齐难度较高,如何在检索阶段有效结合视觉、文本与版式特征仍值得深入研究。其次,实际文档往往跨越多页甚至多文件,模型需要处理长上下文、多跳推理及跨文档整合等任务,现有方法在效率与准确性之间仍难以兼顾。此外,检索增强模型的证据可追溯性与鲁棒性问题尚未得到充分解决,多模态融合后的推理链条仍缺乏统一评测体系[10-11]。这些问题限制了 MRAG 在法律、金融、医疗等高可靠性场景中的直接应用。

本文旨在围绕“视觉文档智能问答与推理”方向开展系统综述,其主要目的与贡献包括:一是界定并整理该研究领域的概念边界与发展脉络;二是按方法学体系化分类并比较主流方法(包含端到端联合编码方法、OCR+布局管道、RAG 方法与智能体多轮推理方法);三是总结现有基准与评测指标的优缺点并指出衡量证据可追溯性的评测需求;四是提出若干可操作的未来研究方向与实验建议以引导后续工作。本文结构安排如下:引言说明研究背景与问题;第 1 章分析视觉文档问答/推理的主要问题与挑战;第 2 章按方法类别系统梳理并比较现有工作;第 3 章提出未来研究方向;第 4 章总结多模态检索目前的应用领域;第 5 章给出总结与展望。通过该综述,本团队希望为研究者与工程实践者提供一个清晰的方法图谱与可复现的比较维度,促进视觉文档 MRAG 在学术与工业应用中的稳健发展。

1 问题与挑战

MRAG 通过引入非参数化的外部知识库,缓解大模型在处理长文档时出现的显存瓶颈、上下文窗口受限以及生成幻觉等问题。与纯文本检索不同,图文混合文档往往同时包含自然图像、行文文本、表

格结构及复杂的空间排布，使得模型不仅要理解跨模态信息间的联系，还要在多种异构证据之间实现稳健的一致性推理。这种多模态交互的特性显著提升了语义对齐、结构理解与证据整合的难度，因此也使得该任务在建模层面面临更加严苛的挑战。本章将从形式化定义出发，系统分析该任务的关键环节与技术挑战。

1.1 文献检索策略与统计分析

为保证综述的系统性与可复现性，本文对MRAG驱动的文档问答相关研究进行了系统文献检索与筛选。检索数据库主要包括 Web of Science、Scopus、IEEE Xplore、ACM Digital Library 以及 ACL Anthology，时间跨度主要覆盖 2019 年至 2025 年。检索关键词组合包括“multimodal retrieval-augmented generation”、“visual document question answering”、“multimodal RAG”、“layout-aware retrieval”和“vision-language large models for document QA”等，并结合“retrieval”、“reasoning”和“evidence grounding”等扩展词进行布尔组合检索。

初始检索共获得相关文献约 300 篇，经过去重、题录筛选与全文相关性判定后，保留与多模态文档问答和 RAG 模型紧密相关的代表性论文 45 篇，涵盖计算机视觉、自然语言处理与多模态大模型三个研究方向。进一步按照方法类型（嵌入范式、布局建模、检索粒度、生成驱动机制）与应用任务（DocVQA、InfographicsVQA、WebQA、跨文档推理等）进行分类统计。

从年度分布及技术演进路径来看，图 1 展示了相关研究在 2022 年以后呈现出的显著增长趋势（彩色效果见《计算机工程》官网 html 版，下同）。尤其随着多模态大语言模型（LLM）与 RAG 框架的发展，面向复杂文档理解的 MRAG 方法成为研究热点。上述统计结果表明，该领域正由以 OCR 与静态特征匹配为主的早期方法，逐步演进为以多模态语义对齐、证据可追溯推理和生成驱动检索为核心的新范式。

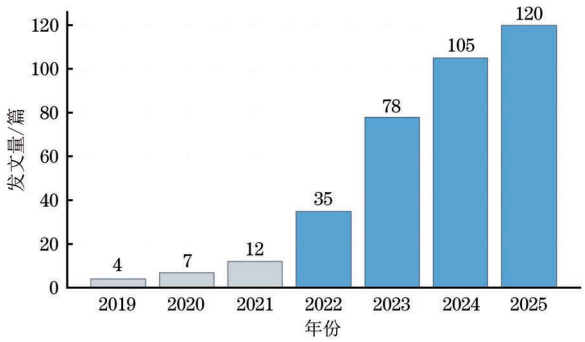


图 1 2019—2025 年 MRAG 文献年度分布

Fig. 1 Annual distribution of MRAG literature from 2019 to 2025

1.2 概念定义与形式化表述

MRAG 的核心目标是构建一个能够利用外部证据库辅助推理的生成模型。形式化地，给定用户查询 q 和包含 N 个文档的多模态知识库 $D=\{d_1,d_2,\cdots,d_N\}$，系统的目标是生成最优回复 y，并使得后验概率 $P(y|q,D)$ 最大化。然而，直接在全量知识库的基础上进行推理不仅计算代价极高，而且输入序列长度会随着文档数量急剧增长。在多模态场景中，图表信息、表格单元、OCR 文本等会进一步扩大输入长度，使得注意力计算的复杂度与显存开销呈指数式上升。因此，实际系统普遍采用“检索-生成”两阶段框架：在生成之前从大规模知识库中筛选出一个相对精炼的候选文档子集，使模型能够在可控的输入长度下进行推理。

检索阶段需要根据相关性评分函数 $S(q,D)$ 从全量文档 D 中筛选出与查询 q 相关的子集 R，这一过程的关键在于兼顾召回与精排，既不能漏掉重要证据，又必须避免噪声文档进入生成阶段以干扰推理。检索阶段的核心任务是根据相关性度量函数，从 D 中筛选出一个与查询 q 相关的子集 R。常见做法包括以下 2 种选取策略。

第 1 种是基于置信度的阈值截断策略，通过引入超参数 τ，仅保留相似度高于该阈值的文档，从而过滤掉低质量的噪声：

$$R=\{d\in D\mid S(q,d)>\tau\} \tag{1}$$

但这种选取策略过于依赖超参数 τ，并且对于不同的数据集超参数 τ 的变化可能较大。

第 2 种方法是基于排序截断的策略，即按照相似度排序选择最高的 k 个文档（Top-k）。这是更为通用的表述方式：

$$R=\{d\in D\mid \mathrm{rank}(S(q,d))\leqslant k\} \tag{2}$$

此方法的超参数的含义更容易被人们理解，但依旧逃不开对超参数的强依赖。当 k 设置过小时，系统可能无法召回关键证据；当 k 设置过大时，输入长度随之爆炸式增长，导致显存占用、推理延迟显著增加，同时大量无关信息涌入会稀释真正有用的上下文线索，从而降低生成质量。

为了避免手工调参带来的不稳定性，近年来的研究开始采用机器学习[如高斯混合模型（GMM）]对相似度分布进行建模，以实现对阈值 τ 和选取数量 k 的自适应确定。对于检索任务而言，通常将得分分布划分为“相关文档分布”与“噪声文档分布”两类，通过求取这两类分布的交点即可自动得到更合理的阈值。同时，利用每个文档属于“相关”分布的

后验概率，还可以动态确定需保留的文档数量，使得系统在简单查询上保持紧凑、高效，而在复杂查询上自动扩大候选范围，从而在召回率、输入长度和推理成本之间取得更平衡的优化效果。

1.3 典型的检索方式

在 MRAG 体系中，检索模块承担从海量文档中筛选候选证据的关键职责，其设计既影响文档召回的准确性，又决定生成阶段可处理的上下文规模。完整的检索过程一般包括预处理、特征构建、索引建立、候选召回以及短列表重排等环节。自然语言查询首先被编码成查询向量，文档的视觉、文本和版面结构经过各自的编码器转换为可检索的向量表示；随后系统通过建立稠密向量索引或倒排索引从知识库中高效召回候选文档；对于召回的短列表，系统再执行重排，以提高最终进入生成器的证据质量。

根据不同的模态使用方式与融合深度，检索机制大体可分为 3 类：将整页视为单一路径输入的视觉检索，以文本与视觉通道分别召回并在集合层面合并的并行检索，以及通过特征融合或打分融合实现细粒度对齐的多模态联合检索。下面给出 3 类方法的结构、数学表述与工程实现。

第 1 类是单一路径的视觉检索。该策略将每个文档页面视为图像，通过跨模态对齐模型将自然语言查询投影到视觉语义空间，计算用户查询向量与页面向量之间的相似度，从而完成排序。设查询向量 $\boldsymbol{q}$ 与文档库 $\boldsymbol{D}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_N\}$，视觉编码函数为 $f_{\text{vis}}(\ )$，查询编码函数为 $f_{\text{q}}(\ )$，相似度函数为余弦相似度 sim( )。检索得分定义为：

$$S(\boldsymbol{q},\boldsymbol{d})=\text{sim}(f_{\text{q}}(\boldsymbol{q}),f_{\text{vis}}(\boldsymbol{d})) \tag{3}$$

系统按照该得分对所有文档排序，并按照式(2)取前 k 个文档得到检索子集。

在工程实现中，视觉向量通常存储在稠密向量索引结构内。主流方案包括基于图结构的近似最近邻检索，以及基于倒排分簇与乘积量化的向量压缩检索。该类方法易于大规模部署，且对图像整体布局和跨页的视觉一致性较为敏感，但由于整页信息被压缩到单一向量，局部文本、细节图标、表格单元等语义信号往往受到弱化，对精确事实类查询的召回能力有限。

第 2 类是模态独立的并行检索。该策略为文本与视觉各自构建独立的检索通道：文本侧采用稀疏倒排索引或文本向量索引，以充分利用 OCR 字符信息在关键词匹配上的优势；视觉侧继续使用视觉向量索引，侧重图表、版式与非文本要素的辨识。设视觉编码函数为 $f_{\text{vis}}(\ )$，文本编码函数为 $f_{\text{text}}(\ )$，文档的视觉模态和文本模态信息分别为 $\boldsymbol{d}_{\text{vis}}$ 和 $\boldsymbol{d}_{\text{text}}$，则两路相似度计算分别为：

$$S_{\text{vis}}(\boldsymbol{q},\boldsymbol{d})=\text{sim}(f_{\text{q}}(\boldsymbol{q}),f_{\text{vis}}(\boldsymbol{d}_{\text{vis}})) \tag{4}$$

$$S_{\text{text}}(\boldsymbol{q},\boldsymbol{d})=\text{sim}(f_{\text{q}}(\boldsymbol{q}),f_{\text{text}}(\boldsymbol{d}_{\text{text}})) \tag{5}$$

系统按照式(2)分别对两种相似度取 Top-k 的文档集合，得到 R_{vis} 和 R_{text}，最终候选集通过集合并集得到：

$$R=R_{\text{vis}}\cup R_{\text{text}} \tag{6}$$

这种方法在工程上极具灵活性：两路索引结构将两种模态检索过程解耦，文本信息缺失时可由视觉补偿，视觉结构模糊时文本检索仍可维持可靠性。并行检索的主要挑战在于并集规模往往较大，需要额外的去重、排序融合和短列表精排，以控制生成阶段的输入长度并保证证据一致性。

第 3 类是多模态联合检索。该策略试图在特征层或得分层对视觉与文本进行更深度的融合，以获得更稳定、更细粒度的相关性评估。常见形式之一是得分融合，分别计算文本相似度和视觉相似度，再根据文档质量、OCR 稳定性或模态可信度设定融合权重，将两类得分统一到一个排序序列中，即：

$$S(\boldsymbol{q},\boldsymbol{d})=\lambda S_{\text{vis}}(\boldsymbol{q},\boldsymbol{d})+(1-\lambda)S_{\text{text}}(\boldsymbol{q},\boldsymbol{d}) \tag{7}$$

式中：λ 是超参数，可以在检索过程中学习或动态调整。最终检索结果依旧按照式(2)得到。

得分融合对模态间尺度一致性要求较高，需要保证文本与视觉得分在同一语义区间可比。另一类更强的联合检索方法采用晚期交互结构，将文档的词元或图像分块分别保留为局部向量序列，由查询侧向量与文档侧局部向量逐一匹配，从而实现细粒度的跨模态对齐。晚期交互检索通常不依赖简单的单向量相似度，而是需要专门的局部交互索引与重排机制，计算成本更高，但在复杂结构文档与跨模态细节理解任务中具有明显优势。

1.4 生成方式

在检索增强体系中，生成阶段承担着“最终落笔”的角色。生成器以用户查询和已选定的跨模态证据为输入，通过多轮推理逐步形成最终回应。随着模型能力与任务规模的共同扩张，生成方式正从“拼接式”向“调度式”演化，即由被动处理上下文转向主动调用证据，围绕问题形成完整的推理链[12]。

最基础的生成范式只要求模型在文本输入中看到检索片段，然后在同一上下文中直接生成答案。然而，当文档规模扩展到数百页乃至上千页时，长序列带来的载荷、注意力稀释和表征干扰都会削弱模型辨别关键证据的能力[13]，并使幻觉风险迅速升高。更复杂的图文混排材料进一步挑战了这种单一

序列式生成，使模型难以在噪声和跨模态结构中保持稳定理解。因此，生成方式逐渐被重新设计为一个由检索调度、跨模态对齐、证据过滤和答案生成组成的递进式过程[14]。

这一趋势的核心是让生成器在不同时间点对证据进行不同形式的聚合。例如，当检索模块返回多页候选时，生成器需要在生成过程中主动调用与问题相关的图像区域、文本片段或结构化内容。聚合方式可以由交叉注意力驱动，也可以通过学习到的池化结构在多页中提取核心内容。模型并不依赖统一的全局表示，而是通过局部视角反复对证据进行查阅，让推理过程更贴近真实的“阅读”[15]。这种生成阶段的动态检索策略，与前文的晚期交互设计相互呼应，使模型能够在长文档环境中保持高效运行，又不牺牲推理的精细度。

为了提升生成质量并避免长序列带来的退化，一些研究进一步引入更具模块化的生成框架。例如，多智能体系统通过设置布局分析、内容抽取、任务分解、事实核验等不同角色[9]，使生成器能够专注于最终语言组织，而由其他模块负责结构解析与证据筛选。这类方法提高了生成的鲁棒性，但也带来了系统协调成本。与之相辅相成的是检索增强框架，它通过外部知识的实时调取减少模型记忆负担，在多文档场景中以更低的代价获得可控推理。同样值得注意的是，这两类范式并非彼此独立，通过互相嵌套的方式已经形成新的混合体系[12]。生成器可以通过代理机制管理检索与验证，而代理系统也常常将检索增强作为其关键的节点功能，使整体流程在可控性、灵活性和性能之间实现新的平衡。

在文本为主的早期阶段，生成往往依赖 OCR 提取的内容，或结合大模型对图像的描述来补全文档信息[16]。这在特定场景下具有一定效果，却难以充分捕捉复杂文档中的视觉结构、跨模态线索和隐含关系[4]。近年来，MRAG 的兴起推动生成进入新的阶段，模型直接在多页图像序列上执行视觉编码，从而保留版式、结构、表格、图表等多种信息源，使生成更能依托真实文档的视觉逻辑进行推理[12]。这一范式进一步催生了更细粒度的生成方式，包括对单页内部各区域、表格单元、图形要素的独立建模，使生成器在最终生成阶段能够调取更准确、更稳定的跨模态证据。

随着技术演进，一些研究也开始探索图结构索引、多级指针网络等更结构化的方式[17-18]，让生成器在推理时不仅基于证据本身，还基于证据之间的拓扑关系开展推断。这类方法与多智能体框架一起，为生成阶段提供了更强的协同能力，使最终回答不仅精准可依，也具备更高的可追溯性与一致性。总体来看，生成方式的演化体现出相同的趋势，即从“直接接收输入”到“面向任务的证据调度”，并利用跨模态检索和结构化建模，让模型在复杂的文档理解任务中保持可靠性、可解释性和泛化能力[12]。

2 研究现状分析

为了从系统层面对现有 MRAG 方法进行统一分析与比较，本文结合图 2 所示的整体架构，从预处理、检索增强、生成与评测闭环 4 个阶段出发，抽象出影响 MRAG 性能的若干关键设计维度，包括文档语义表征与嵌入粒度、文档检索范围、布局与结构建模机制、多模态检索策略。基于上述维度，本文对近年来具有代表性的相关工作进行了系统性梳理与归纳，对比结果汇总于表 1。

MRAG 的快速发展呈现出多线并进的格局：一方面，文档理解任务本身不断扩展，从传统的文本型页面扩展到高度结构化、多模态并存的大型文档集[19-20]；另一方面，从嵌入、检索、布局建模到多模态融合，整个技术链条正在经历从粗粒度到细粒度、从单页到跨文档、从被动检索到主动推理的系统性跃迁[21-22]。大量工作显示，MRAG 已成为解决海量、多模态文档推理任务的最具实用性的范式之一，覆盖金融、法律、医疗与政务等多个高风险应用场景[23-24]。本章将从“任务挑战-技术路线-评测体系”3 个层面构建统一分析视角：从跨模态语义对齐、推理可解释性等关键挑战出发；分析现有嵌入建模与检索机制如何针对上述挑战进行设计；结合典型数据集与评价指标，讨论其对相关能力的覆盖程度及局限性，从而梳理 MRAG 系统在嵌入、布局感知、检索策略以及多模态融合等环节的典型做法与发展趋势。

2.1 嵌入方式

嵌入方式作为 MRAG 系统的核心基础设施，它直接决定了文档的语义深度和跨模态对齐的质量，进而影响后续检索和生成阶段的准确性与效果[19-20]。

在视觉文档理解任务中，文档往往同时包含文本、图像、表格及复杂的空间布局结构，如何将这些异构信息映射为可检索、可比较的向量表示，是构建高效 MRAG 系统的关键问题之一[21]。结合图 2 中检索增强模块的设计，可以观察到，现有方法在嵌入建模上的差异并不仅体现在所使用的模态类型上，更深层次地体现在表示粒度的选择以及其与后续索

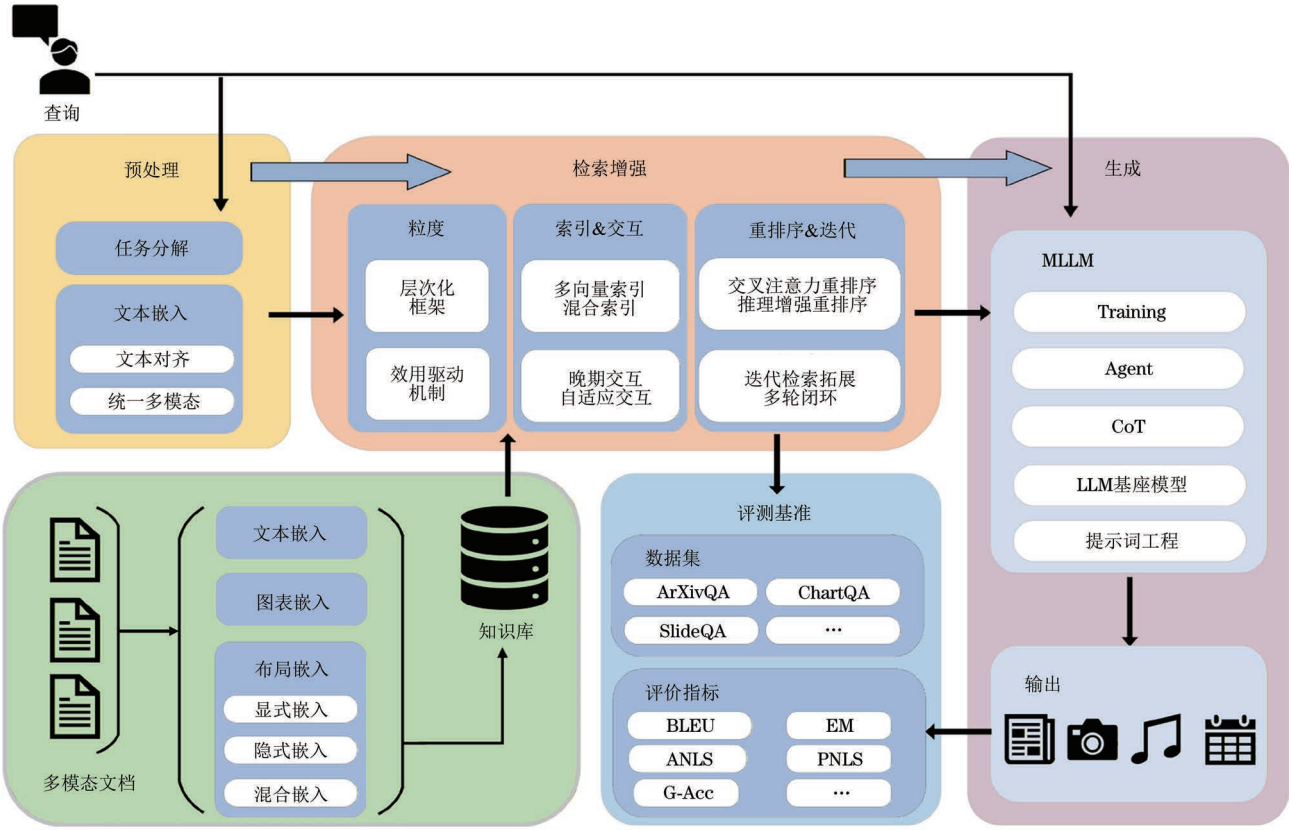


图 2 MRAG 技术路线图

Fig. 2 MRAG technology line map

表 1 多模态视觉文档 RAG 方法的系统设计对比

Table 1 Comparison of system design of multimodal visual document RAG methods

文献方法	文本嵌入	布局建模机制	检索与融合范式	证据参与推理的粒度	是否训练	是否具备 OCR 功能
ViDoRAG[9]	统一多模态嵌入	混合	分离检索后融合	页面级	×	√
DSE[25]	统一多模态嵌入	隐式	统一检索	页面级	√	×
VisRAG[26]	统一多模态嵌入	隐式	统一检索	页面级	√	×
VDocRAG[27]	统一多模态嵌入	隐式	统一检索	页面级	√	×
ColPali[28]	统一多模态嵌入	隐式	统一检索	区域级	√	×
ColQwen2[29]	统一多模态嵌入	隐式	统一检索	区域级	√	×
Light-ColPali[30]	统一多模态嵌入	隐式	统一检索	区域级	√	×
SV-RAG[31]	统一多模态嵌入	隐式	统一检索	页面级	√	×
M3DocRAG[32]	统一多模态嵌入	隐式	统一检索	页面级	×	×
SimpleDoc[33]	统一多模态嵌入	隐式	统一检索	页面级	×	×
FRAG[34]	文本对齐	隐式	分离检索后融合	页面级	×	×
GME[35]	统一多模态嵌入	混合	分离检索后融合	页面级	√	√
HM-RAG[36]	混合编码	混合	分离检索后融合	页面级	×	√
CoRe-MMRAG[37]	混合编码	混合	分离检索后融合	页面级	√	√
VisDoMRAG[38]	混合编码	混合	分离检索后融合	页面级	×	√
MGRAG[39]	混合编码	显式	分离检索后融合	元素级	×	√
VRAG-RL[40]	统一多模态嵌入	隐式	统一检索	元素级	√	×

引与检索机制的协同方式上。基于这一视角，本文将现有方法概括为统一多模态嵌入、文本对齐以及混合编码 3 类多模态嵌入范式，其整体框架与核心差异如图 3 所示。这些范式在表示粒度、检索行为及适用任务上存在显著差异，其对应关系已在表 1 中进行了归纳。

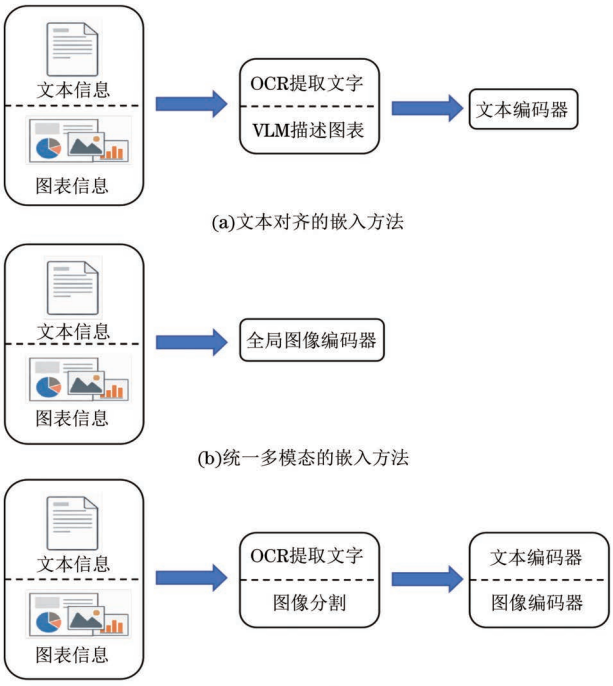


图 3　3 种多模态嵌入表示方法

Fig. 3　Three multimodal embedding representation methods

2.1.1　文本对齐

文本对齐是 MRAG 系统早期采用的一类多模态嵌入范式，旨在将文档中的异构信息统一映射为文字表述，从而便于使用文本编码器进行向量化表示[25]。在这一方法中，文档的文本内容首先通过 OCR 提取，而非文本模态的信息，如图表或图片，通过视觉语言模型(VLM)生成文字描述，形成可统一处理的文本序列。这样文档中所有信息均可通过文本编码器生成向量，用于检索和推理。文本对齐方法实现简单、工程成熟，适合处理文本为主、视觉结构相对简单的文档，如报告、合同等。在 OCR 质量较高且图表数量有限的情况下，这种方法能够有效提取文档核心信息[20]。然而，随着文档复杂度的增加，尤其当出现复杂表格、图示、数学公式及层次化标题时，文本对齐方法的局限性逐渐显现。由于视觉信息被转换为文本描述，可能丢失布局、位置或颜色等细粒度信息，跨页推理以及多模态信息的聚合能力受到一定限制[19]。因此，尽管文本对齐在特定简单场景下仍然实用，但在处理高度视觉化或结构复杂的文档时，其通用性和精细语义表达能力有限[21]。

2.1.2　统一多模态嵌入

随着视觉处理能力的不断提升，MRAG 系统逐渐采用统一多模态嵌入的方式，特别适用于文档中包含大量图像、表格及复杂版式的场景[22,26,28]。该方法通过 VLM 或全局图像编码器，将整页或局部图像直接映射到多模态向量空间，使文本和视觉信息能够在同一共享空间中进行比较和检索。与文本对齐方法不同，统一多模态嵌入无需先将图像转换为文字描述，而是直接利用视觉编码器提取特征，同时结合文档中提取的文本信息，实现跨模态统一表示。

这种方法在保留文档布局信息、文本块位置关系以及图示关键视觉提示方面具有天然优势[19,31]。通过多模态编码，系统能够在跨页推理中保持文档信息的全局一致性，同时缓解长文本输入带来的计算瓶颈，并增强多页跳转和复杂查询中的定位能力[32]。代表性方法包括 DSE[25]、ColPali[28]、VisRAG[26]等，它们均通过 VLM 或统一编码器将图文信息映射到共享向量空间，从而实现任意模态间的检索与推理。

然而，统一多模态嵌入的主要局限在于粒度较粗。整页级或块级的视觉表示难以直接处理文档中更细粒度的信息，如段落级问题、表格单元解析或图表细节推理[20,39]。在这些任务中，单一视觉表示可能不足以提供所需的语义精度，从而影响跨模态检索的召回率和准确度。因此，对于需要精细定位和多模态细节理解的任务，可能需要结合混合编码或后期融合策略，以提升系统的推理能力[37-38]。

2.1.3　混合编码

混合编码是 MRAG 系统中另一类多模态嵌入范式，其核心思想是对不同模态使用独立编码器，并在检索或生成阶段融合模块整合信息[9,36-37]。在这种方法中，文本、图像或其他视觉内容各自通过最适合的编码器生成向量表示，例如文本采用 BERT、E5 等文本编码器，图像采用 ResNet、ViT 等视觉编码器，而表格或结构化元素则可能通过专门的结构化编码器进行表示。各模态向量通常存储在不同的向量库中，检索时可分别执行召回，再通过后期融合或 LLM 进行跨模态推理。

混合编码方法在处理视觉信息和文本信息差异较大或需要高精度粒度推理的场景中表现出较高的灵活性[38-39]。它能够充分利用各模态的特长，实现针对不同信息类型的最优表示，同时在融合阶段保证多模态信息的协同使用。代表性方法包括 ViDoRAG[9]、HM-RAG[36]、CoRe-MMRAG[37] 和 MGRAG[39]，这些方法均采用模块化检索和后期融合策略，有效兼顾视觉细节与跨模态推理能力。

然而，混合编码也存在一定局限性。由于各模态独立编码并需要后期融合，系统设计较为复杂，对

计算资源和实时性要求较高。同时,不同模态向量的整合效果在很大程度上依赖融合模块的设计,跨模态语义对齐和召回精度可能受到影响[38]。因此,混合编码适合处理结构复杂或高度多模态化的文档任务,但在资源受限或需要低延迟检索的场景中,其应用可能受限。

总体而言,3 类嵌入范式在表示能力、系统复杂度与适用任务上形成了清晰分工。文本对齐方法实现成本最低、工程成熟,但对视觉结构与空间信息的表达能力有限,更适用于文本主导、版式相对简单的文档场景;统一多模态嵌入在跨模态对齐与页面级检索中表现最为稳健,是当前大规模视觉文档 RAG 系统的主流选择,但其表示粒度较粗,难以支撑精细证据定位;混合编码在细粒度建模和复杂推理任务中具有明显优势,能够同时保留文本精度与视觉结构信息,但系统设计复杂、计算开销较高。因此,从整体趋势看,统一多模态嵌入在通用性与可扩展性上占据主导地位,而混合编码更适合作为面向高复杂度推理任务的增强方案。

2.2 文档检索范围

在 MRAG 模型中,如何选择合适的文档范围进行信息检索与处理,对于文档理解的效果至关重要[1,40-43]。文档范围的划分通常可以分为“单一文档范围”和“开放域范围”两种模式,这两者各自有其独特的处理方法和应用场景,代表了不同的文档理解策略和技术路径。

单一文档范围侧重于从单一文档中提取信息,通常应用于需要深入理解某一特定文档内容的任务。在这种模式下,系统会将目标文档作为检索的主要对象,通常使用多模态信息(如文本、图像等)对文档的不同部分进行编码,从而构建精确的检索数据库[3,44-45]。这类方法的优势在于能够在有限的文档内集中精力进行细致的分析,尤其适用于那些文档较长但相对封闭的任务。

然而,开放域范围则更多聚焦于从多个文档中提取和整合信息。这种方法适用于开放域的任务,尤其是当任务需要涉及大量外部知识或多个领域的信息时[46]。例如,系统可能需要在一个广泛的文档库中检索相关知识,构建一个覆盖特定领域的大型知识库。这类方法的挑战在于如何有效地整合来自不同文档的信息,并确保信息的一致性和高效性[1]。在技术上,开放域检索通常会面临检索效率、信息冗余以及多文档信息融合等问题。因此,越来越多的研究尝试使用 VLM 或其他高效的多模态表示方法来提高检索的精度和速度,进而提升跨文档知识获取的能力[47]。

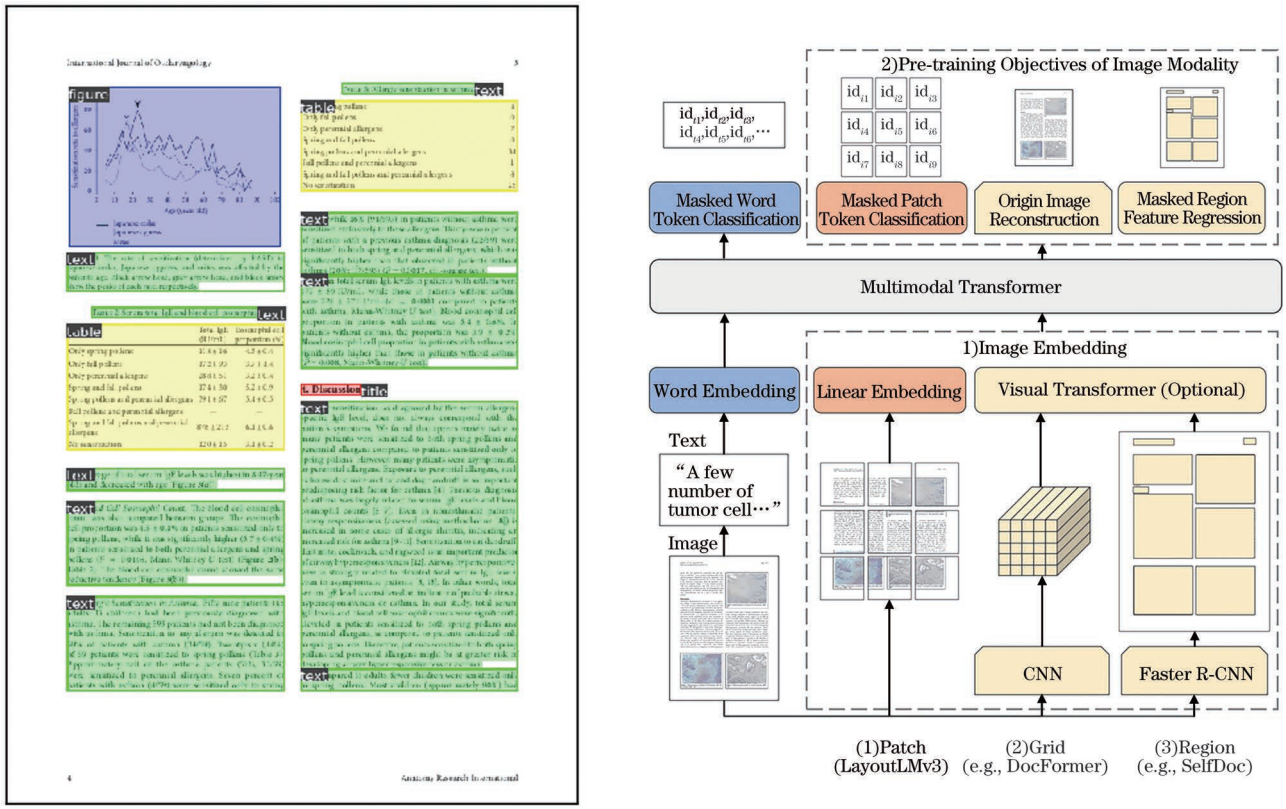
两种文档范围的差异在于:单一文档范围主要集中于深入理解和推理某一文档中的内容,强调信息的精确提取和细致推断;开放域范围则侧重于如何从大规模文档中获取和融合多来源的信息,强调信息的广度和多文档的关联性[48]。这种差异使得两者在具体应用中有所不同,尤其是在应对长文档或涉及多个领域的复杂任务时,如何选择合适的文档范围将直接影响任务的完成质量和效率。

在研究和应用中,单一文档范围和开放域范围并不是互斥的,它们可以根据具体的应用需求进行灵活切换。比如在某些情况下,系统可以先从一个大规模的文档库中检索相关文档,然后通过聚焦单一文档的内容进行深入分析,以期达到最优的理解效果[49]。

从方法效果与系统可扩展性的角度看,单一文档范围与开放域范围并不存在绝对优劣,而是服务于不同层级的理解目标。单一文档范围能够在受控上下文内实现更稳定的证据定位与精细推理,适用于合同分析、报表解读等封闭场景;开放域范围在跨文档知识整合和复杂问题覆盖性上更具优势,是多模态 RAG 走向真实应用不可或缺的能力。然而,其检索噪声、上下文膨胀和推理一致性问题更为突出。总体而言,当前研究趋势正在由“纯开放域”转向“开放域召回+单文档精炼”的混合范式,以在覆盖性与推理可靠性之间取得平衡。

2.3 布局感知方式

在 MRAG 系统中,布局感知方式是处理视觉丰富文档(如 PDF、表单、报告)的关键环节,它旨在保留文档的空间结构、元素位置关系以及图文混排特征,避免传统文本主导方法导致的语义丢失[50]。随着文档复杂度的提升,布局感知已成为提升检索精度和生成质量的核心技术。早期方法主要依赖 OCR 提取文字以及利用图像分割技术提取边界框(BBox)坐标进行显式编码[47-48],而近年来,随着 VLM 的兴起,隐式布局感知通过图像分块的二维位置编码逐渐占据主导,尤其在检索阶段表现出色。从现有方法的整体实现来看,不同系统在布局建模机制上的取舍具有显著差异,其主要实现路径如图 4 所示。这些路径既包括基于 OCR/VLM 生成边界框的显式布局建模,又包括完全基于视觉编码的隐式方式,具体的对应关系在表 1 中进行了系统对比。



(a)显式布局方法(边界框标定结果)　　(b)隐式布局方法(文档分块嵌入)

图 4　两种文档布局嵌入方法

Fig. 4　Two methods for document layout embedding

2.3.1　显式布局感知

显式布局感知主要通过 OCR 工具提取文本块、表格、图像等元素的边界框坐标(通常归一化为$[x_0, y_0, x_1, y_1]$),并将其直接融入模型表示之中[9]。这种方法特别适合高度结构化的文档,能够精确捕捉元素间的相对位置和层次关系。典型代表包括 LayoutLM 系列及其后续版本,它们在 Transformer 架构中引入二维位置嵌入(将横纵坐标分别嵌入后相加)或相对位置偏置,显著提升了文档理解任务的表现[4-5]。DocLLM 进一步创新,通过解耦的空间注意力机制将边界框编码为独立的空间嵌入,并与文本嵌入进行交互,支持轻量级多模态推理,而无需完整的图像编码器[51]。UDOP 采用统一的视觉-文本-布局 Transformer 架构,结合布局标记和相对位置偏置,实现多任务文档处理[28]。

显式方法的优势在于计算高效、对结构化元素(如表格、表单)敏感,但高度依赖 OCR 的准确性,在噪声较多或版式复杂的文档中容易出错。此外,将边界框坐标作为特殊标记插入序列的简单做法也被广泛采用,虽然实现便捷但融合深度有限。

2.3.2　隐式布局感知

隐式布局感知避免了显式 OCR 和边界框提取,直接将文档页面作为图像输入,利用视觉编码器的内置位置机制捕捉布局。这种方法在保留完整视觉线索(如颜色、字体、间距)方面具有天然优势,尤其适用于图文密集或跨页文档[52]。

主流路径以基于视觉 Transformer 的模型为主,将页面划分为图像分块,通过视觉 Transformer 的二维位置编码隐式保留布局,并采用晚期交互机制生成多向量嵌入,用于高效的页面级检索[53]。Qwen2-VL 引入多模态旋转位置嵌入,支持任意分辨率图像的二维位置编码,进一步增强了对变尺度布局的感知能力[54]。类似地,PaliGemma 等模型通过 SigLIP 视觉骨干捕捉空间关系[55]。

隐式方法的优势在于端到端处理,能够避免 OCR 误差,并支持更丰富的视觉语义(如图表细节),但整体粒度较粗、计算成本较高,在需要特定单元格提取等细粒度任务中往往需结合重排序或混合策略。

2.3.3　混合与发展趋势

在实际 MRAG 系统中,单一的布局感知策略往往难以同时兼顾检索效率与生成精度,因此在研究与工程实践中逐渐形成了混合式布局感知框架。具体而言:在检索阶段,系统更倾向于采用隐式、视

觉主导的布局建模方式，通过视觉编码器或跨模态嵌入将页面整体结构压缩为高维语义表示；在生成阶段，引入显式布局信息，如边界框坐标、阅读顺序或区域层级关系，以辅助 VLM 对关键证据进行精确定位与细粒度理解[9,53]。这种“粗到细”的两阶段设计在复杂文档问答、财务报表分析及临床记录解读等任务中表现出较好的鲁棒性与可扩展性。

从发展趋势来看，布局感知 MRAG 正呈现出若干新的研究方向。首先，多智能体（Multi-agent）框架逐渐被引入文档级与跨文档推理场景，不同代理分别负责页面解析、证据检索与答案生成，通过协作机制实现对复杂布局信息的分解与整合[56]。其次，基于图结构的索引与推理方法受到关注，将文档中的文本块、视觉区域与表格单元建模为节点，并以空间邻接、语义依赖或引用关系构建边，从而支持跨页面、跨文档的布局级推理与证据链追踪[57-58]。此外，视觉源归因技术的发展使得模型不仅能够生成答案，还可以显式指出答案所依赖的页面区域与视觉证据，提升系统在高风险应用场景中的可解释性与可信度[59-60]。

综合现有研究可以看出，显式与隐式布局感知在性能与鲁棒性之间取得了权衡。显式布局建模在表格解析、表单填充等结构化任务中具有更高的空间精度和计算效率，但其性能高度依赖 OCR 与版面分析质量，在复杂或低质量文档中稳定性不足；隐式布局感知通过端到端视觉建模显著增强了系统对复杂版式和视觉语义的鲁棒性，已成为当前多模态检索阶段的主流方案，但在细粒度空间关系建模和推理可控性方面仍存在不足。因此，现阶段更具实用价值的方案并非二选一，而是采用隐式布局进行粗粒度检索，并在生成或重排序阶段引入显式结构信息进行精炼。

2.4 多模态检索策略

在 MRAG 系统中，检索策略构成了连接多模态表示空间与生成推理阶段的关键枢纽，其设计直接影响系统对视觉丰富文档的召回覆盖、证据定位精度以及生成答案的可靠性[61]。相较于早期沿用文本 RAG 的单模态粗粒度范式，当前 MRAG 的检索策略已发生结构性转变：检索不再仅承担被动相似度匹配的角色，而是逐步演化为面向证据组织与推理需求的主动规划过程[62-63]。

这一演进得益于 VLM 的快速发展，尤其是基于 ViT 的视觉编码器、多模态位置编码以及多向量检索机制的成熟，使系统能够在不依赖显式 OCR 或规则切分的前提下，直接在视觉空间中建模布局、结构与语义之间的复杂关系[64]。整体来看，当前 MRAG 检索策略呈现出 3 个相互协同的特征：检索粒度由页面级向层次化与动态自适应演进，索引与交互机制更加高效灵活，重排序与迭代策略逐步与生成模型的推理过程深度耦合。以下将从这 3 个维度，对主流技术路径及其发展趋势进行系统梳理。

2.4.1 检索粒度

检索粒度的设计决定了 MRAG 系统能否从页面级相关性判断过渡到面向推理需求的精准证据链构建。相较于对粗粒度与细粒度的静态划分，近期研究更倾向于从层次化框架与效用驱动机制两个更高维度理解粒度问题，使检索过程能够随任务与上下文动态调整[65]。

层次化检索框架通过多阶段或多层索引结构，将全局匹配与局部精炼有机结合。初始阶段侧重于快速缩小候选空间，随后在页面内部对表格、图表、标题或特定视觉区域进行更精细的定位与筛选。MGRAG[66]基于多粒度多模态检索架构，在统一编码框架下同时支持页面级匹配与结构内局部搜索，并显式建模元素间的空间关系；MMRAG-DocQA[67]通过层次化索引进一步扩展到跨页与跨文档依赖，适用于长文档推理场景；mKG-RAG[68]引入多模态知识图谱作为结构化中介，先在实体层面对齐，再在关系与属性层面精炼证据，实现跨模态、多粒度的高效检索[69]。

效用驱动检索机制则将粒度选择与最终生成质量直接关联，通过显式或隐式评估每个候选片段对答案的贡献度，引导系统优先保留高价值证据。VRAG-RL[40]利用强化学习训练模型在召回页面中自动聚焦最相关区域，显著提升了证据针对性与解释性；DocVQA-RAP[70]通过量化片段效用过滤冗余上下文，缓解上下文过载问题；PREMIR[71]针对表格与图表等高信息密度结构预生成问答对，作为细粒度索引项进行精确匹配，进一步提高页面内检索精度。

值得注意的是，以 ColPali[53]为代表的视觉主导多向量检索架构，在机制层面天然模糊了传统粗细粒度的边界。其晚期交互方式允许查询在检索阶段动态选择与之最相关的视觉分块，而无需显式结构切分或规则定义。这种“隐式粒度自适应”能力，使视觉主导检索逐渐成为当前 MRAG 系统中最具通用性和扩展潜力的粒度实现路径，也为后续将检索过程进一步融入生成模型的主动证据规划奠定了基础。

2.4.2 索引与交互机制

索引结构与交互机制为多模态多粒度检索提供

了高效的底层支撑，直接影响系统的扩展性和实时响应能力。传统文本 RAG 常用的单向量稠密索引（如基于 FAISS 的近似最近邻搜索）已难以满足 MRAG 对视觉布局和细粒度匹配的需求，取而代之的是多向量索引、层次索引以及混合索引的综合应用。

多向量索引是当前图像主导检索的主流选择，以 ColBERT 及其衍生架构为代表[53,72]。该机制为每个文档页面或分块生成多个独立向量（通常对应图像分块或文本标记），存储时保留向量集合而不提前聚合，检索阶段通过晚期交互动态计算查询向量与文档向量集合的最大相似度后求和得到总分。这种设计天然支持细粒度匹配，同时显著提升召回精度，已被 Milvus、Qdrant、Pinecone 等向量数据库广泛支持[73]。层次索引则进一步扩展了多粒度能力，例如 MMRAG-DocQA[67]构建的树状索引结构，先在页面级粗索引中快速定位候选文档，再切换到子页面级细索引进行局部搜索；MGRAG[66]结合布局感知的层次编码，将表格、图表等结构化元素作为独立子节点索引，支持并行多路径检索。

混合索引则在实际企业级系统中更为常见，通常结合稠密向量索引（用于语义匹配）、稀疏索引（如 BM25，用于精确关键词匹配）以及图索引（如多模态知识图谱，用于实体关系推理）。mKG-RAG[68]利用 Neo4j 或类似图数据库存储跨模态实体节点与关系边，与向量索引联合查询，实现从粗粒度实体召回到细粒度关系精炼的完整流程。此外，部分工作引入多智能体索引规划机制，由代理根据查询类型动态选择索引路径，进一步提升系统灵活性。

在交互机制方面，晚期交互已成为图像主导 MRAG 的标准配置，避免了早期向量聚合导致的空间信息压缩。同时，新兴的自适应交互策略开始出现，例如基于查询复杂度的动态分块调整或注意力引导的向量加权，进一步优化计算效率。

2.4.3 重排序与迭代策略

初次召回往往存在噪声和冗余，重排序与迭代策略通过二次精炼显著提升证据质量，是 MRAG 从粗放到精准的关键步骤。

重排序机制通常利用更强大的 VLM 或专用排序器对 Top-k 候选进行重新评估。常见路径包括交叉注意力重排序（模型对查询与候选页面执行多头注意力计算相关性）和推理增强重排序。MM-R5[74]是典型代表，它结合监督微调和强化学习训练专用重排序器，促使模型严格遵循指令、产生显式推理链，并根据任务特定奖励（如答案一致性、证据覆盖度）优化排序结果，从而提升精度与可解释性。DocVQA-RAP[70]的效用驱动重排序则从答案生成角度量化每个候选的贡献度，优先保留高价值片段并过滤低相关或冗余证据。SimpleDoc[34]利用 VLM 生成的页面摘要作为重排序信号，提供比原始嵌入更丰富的语义上下文。

迭代检索策略进一步将重排序扩展为多轮闭环过程，代表了 MRAG 向主动推理的演进。受 Self-RAG[75]和多智能体框架启发，部分系统根据初步生成结果或中间推理状态反馈缺失信息，触发二次或多次检索。例如，某些多智能体 MRAG 设计专职“检索代理”与“推理代理”协作：推理代理发现证据不足时，检索代理调整查询或切换模态/粒度重新召回。VRAG-RL[40]的强化学习框架也可视为一种隐式迭代，通过奖励信号引导模型在多轮交互中逐步聚焦最相关区域。

此外，融合策略（如多模态得分加权、候选多样性采样）常与重排序结合使用，确保最终上下文既全面又精炼。当前，多模态检索策略正通过模态深度融合、粒度动态调节、高效索引交互以及智能重排序迭代的全面协同，构建起高度鲁棒和自适应的检索体系。实际部署中，系统往往集成图像主导的快速粗召回、混合模态的语义增强、多粒度层次索引以及强化学习驱动的重排序等多项技术，形成闭环优化的完整管道。未来，随着 VLM 位置编码的持续优化、多智能体规划的深入应用以及向量数据库对晚期交互的原生支持进一步成熟，MRAG 检索策略将更趋向于查询自适应、证据主动规划的智能范式。这不仅将大幅提升系统在长文档、多文档以及跨领域推理任务中的表现，还将显著增强其在金融、法律、医疗等文档密集型高风险场景中的实用性和可信度，推动 MRAG 技术向更高级的结构化文档智能方向演进[47]。

从整体演进路径来看，多模态检索策略已由静态相似度匹配，转向以推理需求为中心的主动证据规划机制。粗粒度页面级检索在效率和召回覆盖上仍具有不可替代的基础作用，但难以满足复杂推理对精确证据解读的需求；细粒度与多向量检索显著提升了定位能力，却对索引结构和计算资源提出更高要求。重排序与迭代检索进一步弥补了初次召回的不确定性，是提升系统稳定性与可解释性的关键组件。总体而言，当前最具优势的 MRAG 系统普遍采用“多向量粗召回＋推理感知重排序＋迭代精炼”的组合策略，而单一粒度或单轮检索已难以支撑真实复杂文档场景。

3 数据集和测试基准

多模态视觉文档问答与推理领域的发展在很大程度上依赖于高质量、多样化的数据集以及系统化、可复现的评测基准。围绕视觉文档理解、知识密集型问答与跨模态综合推理等核心任务，现有研究构建了大量数据资源，并逐步形成了一套相对成熟但仍在持续演进的评测体系。本章对主流数据集类型及其评价指标进行系统梳理，重点分析不同评测基准在任务设定、信息侧重与能力覆盖范围上的差异。与此同时，数据集与评测基准为不同方法提供了统一且可比的量化尺度，为后续章节中各类 MRAG 模型的定量评测与批判性分析奠定基础。

3.1 多模态数据集分类与特点

从任务形式与数据来源的角度来看，现有多模态数据集大致可以归纳为 5 类，其划分依据主要体现在模态构成、推理深度以及是否引入外部知识等方面。

从任务设定与评测目标的角度来看，现有多模态数据集可划分为 5 类。尽管各类数据集均涉及视觉与语言信息的联合建模，但其核心评测侧重点存在明显差异，分别偏向于综合多模态能力、视觉内容对齐、视觉富文档理解、语义与逻辑推理以及时序与交互建模能力。

第 1 类为 MRAG 综合评测数据集。该类数据集侧重于对模型综合多模态能力的整体评估，而非单一子能力的测试。模型不仅需要完成跨模态检索，还需在多文档和长上下文条件下进行信息整合与答案生成，涉及检索、对齐、推理和生成等多个环节的协同表现。因此，该类数据集更适合作为检验多模态系统端到端能力与工程实用性的综合性基准。

第 2 类为知识密集型视觉问答数据集。相较于其他类别，该类数据集更偏向于考察模型对视觉内容与外部知识的对齐能力。虽然问题通常需要借助常识或结构化知识作答，但评测重点并不在于复杂文档检索或长链推理，而在于模型能否正确理解图像语义，并将其与相关知识进行有效匹配与融合，从而完成基于视觉语义的知识推断。

第 3 类为半结构化文档理解与问答数据集。该类数据集主要用于评估模型对图表、文档和网页等结构化视觉载体的理解能力。其核心挑战不在于外部知识推理，而在于复杂版面下的文本识别、结构解析以及表格、图像等多种信息形式的联合建模。模型需要具备对文档结构和视觉布局的精细感知能力，以支持跨区域的信息定位与整合。

第 4 类为视觉理解与推理评测数据集。该类数据集更强调模型对输入内容本身进行深层语义推理和逻辑推断的能力。在该类任务中，视觉信息通常作为问题背景或条件约束，而评测重点在于模型是否能够基于给定内容完成多步推理、数学计算或因果分析，用以检验多模态模型在推理层面的泛化能力。

第 5 类为视频理解与多轮交互的数据集。该类数据集引入时间维度和交互过程，主要考查模型在动态场景中的时序建模与状态跟踪能力。与静态图像问答不同，该类任务要求模型持续整合跨时间的信息，并在多轮交互中保持上下文一致性，是对多模态模型长期记忆和连续推理能力的补充评测。

上述各类数据集在模态组合、任务设定与应用场景上存在显著差异，其整体对比情况如表 2 所示。由表 2 可以看出，面向多模态 RAG 的综合评测数据集在规模和文档复杂度上仍相对有限，这也在一定程度上制约了现有方法在真实长文档与跨文档场景下的系统性评估。表 2 从类别、数据集大小与关键多模态内容(图像，文本，文档，表格，视频)等维度，对不同类别数据集进行了系统整理，为后续方法比较提供了统一参照。

从任务适用性的角度来看，表 2 中不同类别的数据集在 MRAG 研究中承担着不同的评测功能。MRAG 综合评测数据集更适用于评估模型在“检索-对齐-推理-生成”全流程下的端到端能力，尤其适用于多文档与长上下文场景中的系统级性能分析；知识密集型视觉问答数据集主要用于考查模型在视觉语义理解基础上融合外部知识的能力，更适合作为视觉-知识对齐机制的验证基准，而非复杂文档检索评测；文档、图表与网页理解类数据集侧重评估模型对视觉富文档的结构解析、文本定位与跨区域信息整合能力，是当前 MRAG 在真实文档应用中的核心评测资源；视觉理解与推理类数据集主要用于检验模型在多模态条件约束下的逻辑推理、数学计算与因果分析能力，更适用于分析模型的推理泛化性能；视频理解与多轮交互数据集则引入时间维度与交互过程，用于评估模型在动态场景中的时序建模、状态跟踪与长期上下文保持能力。

3.2 多模态视觉文档问答评价指标

模型性能评测是多模态视觉文档问答研究中的关键环节。现有评价指标体系大体可分为基于字面匹配的自动化指标、面向文档特性的编辑距离指标、基于深度语义表示的相似度指标，以及人工或 LLM

表 2 MRAG 相关数据集概览

Table 2 Overview of related datasets for MRAG

类别	数据集名称	数据集容量	关键多模态内容
MRAG 综合评测	ViDoSeek[9]	1 142 QA，300 文档	
	MRAG-Bench[10]	1.35K QA，16.13K 图像	
	MRAMG-Bench[19]	4.40K QA，4.35K 文档，14.19K 图像	
	M3DocVQA[32]	2 441 QA，3 868 文档	
	ViDoRe[53]	3.8K QA，8.3K 图像	
	WebQA[76]	24.93K（图像）24.34K（文本）QA	
	ViQuAE[77]	3.7K QA 知识库：1 500K 文章	
	MIMOQA[78]	56.69K QA，401.18K 图像	
	VisDoMBench[38,79]	2 271 QA，1 277 文档	
知识密集型视觉问答	OK-VQA[80]	14.06K QA	
	A-OKVQA[81]	24.90K QA	
	KVQA[82]	183.01K QA，24.60K 图像	
	FVQA[83]	5.83K QA，2.19K 图像	
	S3VQA[84]	6.77K 图像-问题对	
文档、图表与网页理解	DocVQA[3]	50K QA，12.77K 图像	
	InfographicVQA[85]	30.04K QA，5.49K 图像	
	SlideVQA[86]	52K QA，14.5K 图像	
	ChartQA[87]	～20.88K 图表，～32.0K QA	
	TAT-DQA[88]	16 558 QA，2 758 文档	
	DUDE[89]	41 491 QA，4 974 文档	
	SPIQA[90]	27K QA，25.5K 文档	
	MP-DocVQA[91]	46.0K QA，6.0K 文档(48.0K 页)	
	TextVQA[92]	45.34K QA，28.41K 图像	
	OCR-VQA[93]	1 000.0K＋ QA，207.57K 图像	
	WebSRC[94]	400.0K QA，6.4K 网页	
	SEED-Bench-2-Plus[95]	2.3K QA	
视觉理解与推理	VQA v2[96]	1 110.0K＋ 图像-问题对	
	VCR[97]	290.0K QA	
	MMBench[98]	3.22K QA	
	ScienceQA[99]	47.1K QA，7.3K 图像	
	MathVista[100]	6.14K 样例	
	FigureQA[101]	120.0K 图像，1 550.0K 问题	
视频理解与多轮交互	Video-MME[102]	2.70K QA，0.90K 视频	
	ActivityNet-QA[103]	58.0K QA，5.80K 视频	
	EgoSchema[104]	0.50K＋ QA	
其他	MME-Industry[105]	1.05K QA	
	OCRBench[106]	1.00K QA	

参与的整体判别指标等类别，不同指标在鲁棒性、计算成本与适用任务方面各有侧重。

传统基于文本匹配的自动化指标，如 BLEU、ROUGE 与 METEOR，主要通过 n-gram 重合度或序列相似性对生成结果进行量化。这类指标隐含的前提假设是答案表达具有较强的规范性与稳定性，因此更适用于事实型问答、模板化生成或答案空间受限的任务场景。在跨模态重述、自由生成或多证据融合场景中，该类指标往往难以准确反映模型对关键信息的理解程度。

针对视觉文档问答中普遍存在的 OCR 识别噪声与字符级偏差问题，编辑距离类指标被广泛采用。以 ANLS 与 PNLS 为代表的指标通过放宽字符匹配约束，提高了对识别误差和文本边界不确定性的容忍度，更适合评估以信息抽取与局部文本定位为主的任务。然而，这类指标本质上仍聚焦于最终答案字符串的相似性，其评估信号主要反映“是否读对了内容”，而非“是否通过合理的检索与对齐过程获得内容”。因此，在包含显式检索、证据聚合或跨文档推理的 MRAG 场景中，单独依赖编辑距离指标难以刻画模型在证据选择与多模态融合方面的真实能力。

基于深度语义表示的评价方法，如 BERTScore，则更强调语义一致性而非表面匹配。这类指标在同义改写、信息重组及抽象表述等场景下具有更强的鲁棒性，适用于开放式生成与跨模态语义对齐任务，但其隐含假设是语义相似性能够代表答案质量，在高自由度生成场景中，可能对语义连贯但证据支撑不足的回答给予较高评分，从而弱化对事实可验证性的约束。

此外，对于生成式与多步推理任务，人工评测或引入 LLM 的整体判别指标逐渐成为重要补充。这类方法通常从答案正确性、信息完整性与推理合理性等维度进行综合评估，更适合衡量复杂推理链条的整体质量，但同时也面临评测成本与主观性的问题。

本文将不同评价范式的核心计算思想、优势与局限以对比图表形式呈现。表 3 通过横向区分字面匹配、编辑距离、语义嵌入与整体判别等评测层级，纵向对比其鲁棒性、计算开销与适用任务类型，从而更加直观地展示不同测试基准在多模态视觉文档问答场景下的适配性差异。

表 3 多模态视觉文档问答常用评价指标对比

Table 3 Comparison of evaluation metrics for multimodal visual document Q&A

指标	核心计算方式	优点	局限
BLEU	n-gram 精确度匹配	计算快，标准统一	忽略语义，不适用于同义词
nDCG	基于位置权重的对数衰减增益归一化计算	侧重检索排序质量	仅反映检索精度，不代表生成的答案质量
ROUGE	n-gram 或 LCS 召回率	关注信息覆盖度	重词形，无法应对语序变化
METEOR	引入同义词与语序惩罚	比 BLEU 更接近人工判断	计算复杂，依赖词典
EM	答案完全一致判断	标准严格，无歧义	过于苛刻，不答错字
ANLS	归一化编辑距离评分	抗 OCR 错误，领域标准	仍为字面匹配，忽略语义
PNLS	部分子串编辑距离	容忍答案冗余内容	计算复杂，未普及
BERTScore	预训练模型嵌入匹配	评估语义一致性	计算开销大，无法应对流利废话
G-Acc	人工或 LLM 整体正确性	评估复杂推理与阐述	成本高，依赖评判者

现有数据集与评测基准为多模态视觉文档问答研究提供了重要支撑，但仍存在以下不足：第一，多数数据集仍偏向于封闭域或特定文档类型，面向开放域、跨行业复杂文档的基准尚显不足；第二，评测指标多集中于答案本身的匹配度，对检索证据的质量、多模态融合的有效性以及推理过程的合理性缺乏细粒度评估；第三，缺乏对系统在对抗性样本、长尾分布等场景下鲁棒性的系统化测试。

3.3 主流 MRAG 模型量化评测

为了客观评估前述各类 MRAG 方法的实际效能，本节选取了具有高度代表性的 3 个评测基准进行量化对比。选择 ArXivQA、SlideVQA 与 InfoVQA 作为核心指标依据，主要基于其在学术界与工业界的公认度与互补性：这 3 个数据集共同构成了当前多模态文档检索基准 ViDoRe 的核心，分别覆盖了学术论文的细粒度公式解析、幻灯片的复杂版式排布以及信息图表的视觉语义理解，能够较为全面地反映模型在真实视觉富文档场景下的鲁棒性。

然而，表 4 的量化结果也显示，不同模型范式的优势具有明显的任务依赖性，其中“—”表示原文献中没有该指标值。需要说明的是，表 4 统一采用 nDCG@10 作为评价指标，是因为该指标能够同时

衡量检索结果的排序质量与相关文档的等级差异，相比单纯的 BLEU 或 EM，nDCG 更适用于多模态文档检索中“部分相关”和“跨页相关”等非二值相关性场景。在视觉文档问答任务中，证据往往具有不同程度的相关性与信息密度，因此采用 nDCG 能够更准确反映模型对高价值证据的排序能力，同时保证不同数据集之间的可比性。整体来看，多模态检索技术正在经历从传统“OCR＋文本嵌入”向更强调原生视觉交互与跨模态联合建模的范式迁移。其中，以 ColQwen2 和 ColPali 为代表的视觉延迟交互架构在视觉结构复杂、版式依赖显著的任务中表现尤为突出，尤其是在包含微小文字与复杂排布的 SlideVQA 场景下，其性能相较传统方案具有明显优势。值得注意的是，ColQwen2 作为当前该领域的代表性模型，在涉及高难度学术图表解析的 ArXivQA 任务中取得了显著领先，这表明更强的底座视觉感知能力对于精细文档结构的建模具有重要作用。

在视觉语义相对宏观、版面结构约束较弱的 InfoVQA 任务中，VisRAG 等单向量检索架构与复杂交互模型之间的性能差距明显缩小，说明在该类场景下，增加交互复杂度并不必然带来成比例的性能收益。与此同时，Light-ColPali 在性能轻微下降的情况下显著降低了存储与索引成本，进一步揭示了多向量与延迟交互范式在性能提升与系统代价之间所面临的现实权衡。相比之下，早期或未针对文档视觉结构进行专门优化的模型（如 UDOP 和 SigLIP）在 InfoVQA 等视觉信息高度密集的任务中表现受限，其瓶颈更多源于模型假设与文档视觉特性的错位，而非单纯的模型规模不足。

表 4 主流 MRAG 模型检索性能对比

Table 4 Comparison of retrieval performance of mainstream MRAG models %

模型名称	核心架构	ArXivQA（nDCG@10）	SlideVQA（nDCG@10）	InfoVQA（nDCG@10）
ColQwen2	基于 Qwen2-VL	86.20	95.80	87.50
VisRAG	单向量检索	75.11	91.85	86.37
ColPali	多向量检索	72.50	93.99	81.15
Light-ColPali	存储优化	70.50	92.10	78.80
SigLIP	原生视觉编码器	59.16	89.08	74.59
VDocRAG	文档级检索框架	—	77.30	72.90
UDOP	统一文档预训练	—	64.70	47.40

4 主要应用领域

随着多模态文档理解技术的快速发展，越来越多的研究开始关注如何将这些技术应用于实际的社会和工业场景。本章聚焦于时空数据分析、医疗健康和科研领域，展示多模态技术在这些现实应用中的潜力与实践效果。在时空领域，多模态技术主要用于融合空间和时间信息，以提高数据处理和预测的精度，尤其在交通管理和应急响应系统中，已展现出显著的应用价值。在医疗健康领域，多模态理解技术通过整合影像数据、基因组学数据等，推动了智能诊断、个性化医疗和临床辅助决策的发展。此外，科研领域也在逐步采用多模态技术进行知识图谱构建和科研数据分析，加速了学术研究的创新和跨学科的合作。

4.1 时空领域

在时空数据分析领域，多模态理解技术的应用尤为突出。时空数据涉及空间信息和时间信息的有机结合，如何在此基础上建立准确的语义模型，是多模态理解技术的一项关键任务。通过构建语境图谱，这些技术可以有效地整合时间、空间和实体等多维数据，为时序变化分析和地理事件推理提供支持。对于智能城市的建设，时空数据的有效融合已成为优化交通管理、能源调度以及城市环境监控的核心技术。

例如，在智能交通系统中，结合城市交通数据、天气数据、实时监控图像等多模态信息，可以精准预测交通流量，减少交通堵塞，提升道路使用效率[107]。此外，在灾害应急响应中，基于多模态的时空数据分析不仅可以通过卫星图像和气象数据实时预测自然灾害的发生，还能在灾后帮助决策者调度资源，进行迅速的灾后恢复[108]。这些系统的成功应用大大提高了应急管理的效率和准确性，为政府和企业在处理突发公共事件时提供了宝贵的支持。

随着大数据和云计算技术的成熟，时空数据分析已经成为环境监测和气候变化研究中不可或缺的工具[109]。如全球气候变化研究中，结合遥感数据与地面气象数据能够帮助科学家更好地理解气候模

式变化及其对生态系统的影响。这些多模态数据的整合与分析,为全球变暖、海平面上升等环境问题的研究提供了坚实的理论基础 。

4.2 医疗健康领域

医疗健康行业是多模态技术应用最为广泛的领域之一。随着医疗数据类型的多样化和信息量的增加,单一模态的信息处理方式已无法满足智能医疗系统的需求。因此,多模态医疗图像检索和智能诊断系统成为研究的热点。通过将 CT、X 射线等影像数据与诊断报告等文本信息进行融合,医疗系统能够更全面地理解患者的健康状况,从而实现个性化医疗服务。例如,基于图像和文本数据的联合检索系统已在多个医疗影像数据库中得到了应用[110]。研究表明,这种方法能够显著提高影像数据检索的效率和精度,为临床医生提供更为精准的辅助决策支持。

此外,随着人工智能和大数据技术的进步,医疗领域越来越依赖于基于多模态数据的智能病历分析系统。这些系统通过融合患者的影像、遗传数据、电子病历等多模态信息,不仅能够提供更精确的诊断结果,还能够辅助医生进行个性化治疗方案的设计[111]。相关研究表明,基于多模态模型的智能诊断系统相比传统方法在准确性和效率上有了显著提高,尤其在复杂疾病(如癌症、心血管疾病)的早期诊断中,展现出了巨大的潜力。

4.3 科研数据与辅助研究

科研领域本身就是一个信息密集、数据复杂的领域,随着研究数据量和复杂度的不断增加,科研人员亟需一种能够有效整合不同来源数据的智能系统。多模态技术在科研数据集成、文献检索和科研成果分析中起到了至关重要的作用。例如,在科研知识图谱的构建中,文献文本、实验结果、图像数据和时间序列等信息常常以不同的模态形式存在[112]。通过将这些模态数据进行融合,可以更好地表示科研领域中的实体关系,并推动知识发现和科研创新。多模态深度学习模型能够从大规模科研数据中挖掘潜在的科研热点,为学术界和行业界提供有价值的洞察。

特别是在疾病关系提取(REMAP)的研究中[113],多模态学习方法已被用来构建医学领域的知识图谱。这些图谱不仅能够在不同的学术文献中提取疾病之间的关系,还能帮助研究人员发现新的疾病相关性,为药物研发和临床试验提供理论支持。相关研究表明,结合图谱和多模态数据的疾病关系提取模型显著提升了疾病相关性推理的准确性和效率。

此外,科研领域中越来越多的系统开始使用多模态方法来进行文献推荐和自动化标注[114]。这些系统通过对文献中的文本、图表和附加数据进行分析,能够自动化地为科研人员推荐相关文献,并辅助研究人员进行数据整合和分析。这种方法在减少文献检索时间和提升科研效率方面发挥了重要作用。

随着通用多模态大模型的发展,科研数据理解与辅助研究正由以任务定制模型为主的判别式范式,逐步演进为以生成式推理和证据驱动分析为核心的新范式。以 GPT-5、Gemini 等为代表的新一代多模态大模型在大规模图文语料上进行联合预训练,具备跨模态语义对齐、复杂指令理解与多步推理能力,使其不仅能够作为科研问答和文献分析的结果生成器,还可在检索阶段参与证据重排序与跨文献、跨模态信息整合。在 MRAG 框架中,这类模型通过引入上下文感知的生成式推理机制,有助于缓解传统科研信息系统在长文献理解、隐式关联建模及推理链构建方面的局限,从而在科研知识图谱构建、疾病关系发现以及文献智能检索与推荐等场景中,进一步提升信息整合的深度与科研辅助决策的可靠性。

5 未来展望

MRAG 为大模型引入外部多模态知识提供了一条重要技术路径,在复杂信息理解、跨模态推理与生成任务中展现出广阔应用前景。然而,从整体研究进展来看,现有工作仍主要集中于局部方法或模块层面的改进,对于检索目标如何与任务需求对齐、多模态证据如何参与推理,以及模型与数据系统如何协同工作等关键问题,尚未形成统一而系统的解决框架。面向未来,MRAG 的发展有必要从更宏观的角度出发,推动其由“增强生成方法”向“多模态智能系统范式”演进。

1)面向任务目标的多模态语义感知检索机制。现有 MRAG 方法普遍以模态内或跨模态相似度作为核心检索依据,检索策略更多关注语义相关性本身,而对下游任务的实际需求考虑不足。在复杂应用中,不同任务对证据类型、粒度及组织方式具有显著差异,仅依赖相似度往往难以保证检索结果能够有效支撑后续推理。未来研究应从任务目标出发,探索语义感知的多模态检索机制,通过显式建模问题类型与检索策略之间的关系,根据诊断、预测或解释等不同任务动态调整检索模态与粒度;同时,引入条件化与多跳检索方式,在文本语义约束下逐步筛

选关键证据，构建更符合推理需求的证据获取流程，从而提升检索结果对下游任务的支撑能力。

2）面向推理过程的多模态信息协同建模方法。尽管现有 MRAG 系统能够检索到来自不同模态的相关信息，但多模态证据在生成模型中的利用方式仍较为粗糙，多数方法采用简单拼接或隐式融合，难以刻画证据之间的结构关系与相互约束。未来研究需进一步关注多模态信息在推理过程中的组织方式，通过引入更显式的中间表示或分步推理机制，使模型能够围绕不同模态证据进行逐步分析、验证与综合判断。该方向的核心在于从“提供更多信息”转向“支持更清晰的推理过程”，以提升推理结果的稳定性与一致性。

3）面向统一数据处理的模型-数据库协同架构。在系统层面，当前 MRAG 方法中大模型与数据库、检索系统之间的协作仍然较为松散，模型往往通过有限接口被动调用数据资源，难以充分利用数据库在结构化查询、约束执行与中间结果验证方面的优势。未来研究有必要从整体架构层面对这一问题进行重构，将多模态数据处理过程中反复出现的基本功能进行整理与抽象，形成一组可复用的数据操作算子，并以数据智能体的形式对不同类型数据进行统一管理与调用。在该框架下，大模型主要负责任务理解、推理规划与决策生成，而具体的数据查询、过滤、聚合与验证操作则由相应的数据智能体完成，并通过明确的交互机制形成闭环协作。该方向代表了一种从“模型中心”向“模型与数据协同”的转变，有助于提升 MRAG 系统在复杂数据场景下的效率、可控性与可扩展性。

6 结束语

本文以多模态文档为核心研究对象，系统综述了 MRAG 在视觉富文档问答与推理任务中的研究进展。针对传统 RAG 方法主要面向纯文本场景、难以有效建模文档中视觉元素与空间结构的问题，本文从方法范式层面对 MRAG 的技术特征与发展脉络进行了系统梳理。首先，本文界定了多模态文档问答任务的基本概念，分析了其在多模态对齐、长上下文建模与复杂推理组织等方面相较于文本问答所面临的关键挑战，并概括了以 MRAG 为核心的统一系统框架。其次，围绕 MRAG 支持生成与推理过程的关键设计要素，从嵌入范式、布局感知机制、文档检索范围以及多模态检索策略等方面，对代表性方法进行了分类总结与比较分析，揭示了不同设计选择在生成稳定性、推理精度与系统复杂度之间的权衡关系。最后，结合现有研究成果，本文从检索目标建模、生成驱动的证据组织以及多模块协同系统设计等角度，对 MRAG 的未来发展方向进行了展望，指出该领域正由静态相似度检索向面向推理需求的动态证据规划演进。本文的综述工作旨在为后续多模态文档问答与推理研究提供系统化的方法视角与参考框架，推动 MRAG 在复杂真实文档场景中的进一步发展。

参考文献

[1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[2] MEI L, MO S Y, YANG Z H, et al. A survey of multimodal retrieval-augmented generation [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2504.08748.

[3] MATHEW M, KARATZAS D, JAWAHAR C V. DocVQA: a dataset for VQA on document images [C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Washington D.C., USA: IEEE Press, 2021: 2199-2208.

[4] XU Y H, LI M H, CUI L, et al. LayoutLM: pre-training of text and layout for document image understanding [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2020: 1192-1200.

[5] XU Y, XU Y H, LÜ T C, et al. LayoutLMv2: multi-modal pre-training for visually-rich document understanding [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Philadelphia, USA: Association for Computational Linguistics, 2021: 2579-2591.

[6] KIM G, HONG T, YIM M, et al. OCR-free document understanding Transformer [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2111.15664.

[7] APPALARAJU S, JASANI B, KOTA B U, et al. DocFormer: end-to-end Transformer for document understanding [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2021: 1-12.

[8] 李子骏, 肖辉, 李雪峰. 面向知识密集型任务的检索增强生成技术综述[J]. 微电子学与计算机, 2025, 42(10): 48-65.
LI Z J, XIAO H, LI X F. Survey on retrieval-augmented generation techniques for knowledge-intensive tasks [J]. Microelectronics & Computer, 2025, 42(10): 48-65. (in Chinese)

[9] WANG Q C, DING R X, CHEN Z H, et al. ViDoRAG: visual document retrieval-augmented generation via dynamic iterative reasoning agents[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2502.18017.

[10] HU W B, GU J C, DOU Z Y, et al. MRAG-Bench: vision-centric evaluation for retrieval-augmented multimodal models[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2410.08182.

[11] LIU Z H, ZHU X S, ZHOU T S, et al. Benchmarking retrieval-augmented generation in multi-modal contexts [C]// Proceedings of the 33rd ACM International Conference on Multimedia. New York, USA: ACM Press, 2025: 4817-4826.

[12] ABOOTORABI M M, ZOBEIRI A, DEHGHANI M, et al. Ask in any modality: a comprehensive survey on multimodal

retrieval-augmented generation[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2502.08826.
[13] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document Transformer[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2004.05150.
[14] HUANG Y P, LÜ T C, CUI L, et al. LayoutLMv3: pre-training for document AI with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: ACM Press, 2022: 4083-4091.
[15] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. Journal of Machine Learning Research, 2019, 21: 140: 1-140: 67.
[16] JAUME G, EKENEL H K, THIRAN J P. FUNSD: a dataset for form understanding in noisy scanned documents[C]//Proceedings of the International Conference on Document Analysis and Recognition Workshops (ICDARW). Washington D.C., USA: IEEE Press, 2019: 1-6.
[17] ZHANG X K, SONG D J, CHEN Y X, et al. Topology-aware embedding memory for continual learning on expanding networks[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2024: 4326-4337.
[18] LI H P, WEI G C, XU H C, et al. DocPointer: a parameter-efficient pointer network for key information extraction[C]//Proceedings of the 6th ACM International Conference on Multimedia in Asia. New York, USA: ACM Press, 2024: 1-7.
[19] YU Q H, XIAO Z Y, LI B H, et al. MRAMG-Bench: a comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation[C]//Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2025: 3616-3626.
[20] CHEN Z L, ZHANG P, XU M Y, et al. LocatingGPT: a multi-modal document retrieval method based on retrieval-augmented generation[C]//Proceedings of the IEEE 9th International Conference on Data Science in Cyberspace (DSC). Washington D.C., USA: IEEE Press, 2025: 232-239.
[21] MARTIN R, WALDEN W, KRIZ R, et al. Seeing through the MiRAGE: evaluating multimodal retrieval augmented generation[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2510.24870.
[22] SHEN Z X, YU J F, WANG W Y, et al. Global question-aware multimodal retrieval-augmented generation for multimedia multi-hop question answering[C]//Proceedings of the 7th ACM International Conference on Multimedia in Asia. New York, USA: ACM Press, 2025: 1-8.
[23] WANG J H, ASHRAF T, HAN Z Y, et al. MIRA: a novel framework for fusing modalities in medical RAG[C]//Proceedings of the 33rd ACM International Conference on Multimedia. New York, USA: ACM Press, 2025: 6307-6315.
[24] YILMAZ R E, TAYSI M A, ÖZMEN A İ, et al. Grounded answer generation over multimodal financial records via semantic indexing[C]//Proceedings of the 10th International Conference on Computer Science and Engineering. Washington D.C., USA: IEEE Press, 2025: 160-165.
[25] MA X G, LIN S C, LI M H, et al. Unifying multimodal retrieval via document screenshot embedding[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2406.11251.
[26] YU S, TANG C Y, XU B K, et al. VisRAG: vision-based retrieval-augmented generation on multi-modality documents[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2410.10594.
[27] TANAKA R, IKI T, HASEGAWA T, et al. VDocRAG: retrieval-augmented generation over visually-rich documents[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2025: 24827-24837.
[28] TANG Z N, YANG Z Y, WANG G X, et al. Unifying vision, text, and layout for universal document processing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2023: 19254-19264.
[29] YAMANO M, FUKUOKA K, MIYAMORI H. Two-stage approach using pretrained language models for question answering on Japanese document images[C]//Proceedings of the 33rd ACM International Conference on Multimedia. New York, USA: ACM Press, 2025: 13791-13796.
[30] MA Y B, LI J S, ZANG Y H, et al. Towards storage-efficient visual document retrieval: an empirical study on reducing patch-level embeddings[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2506.04997.
[31] CHEN J, ZHANG R Y, ZHOU Y F, et al. SV-RAG: LoRA-contextualizing adaptation of MLLMs for long document understanding[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2411.01106.
[32] CHO J, MAHATA D, IRSOY O, et al. M3DocRAG: multi-modal retrieval is what you need for multi-page multi-document understanding[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2411.04952.
[33] JAIN C, WU Y R, ZENG Y F, et al. SimpleDoc: multi-modal document understanding with dual-cue page retrieval and iterative refinement[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2506.14035.
[34] ZHAO D F. FRAG: toward federated vector database management for collaborative and secure retrieval-augmented generation[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2410.13272.
[35] CHEN C, PETTERSON S, PHILLIPS R L, et al. Toward Graduate Medical Education (GME) accountability: measuring the outcomes of GME institutions[J]. Academic Medicine, 2013, 88(9): 1267-1280.
[36] LIU P, LIU X, YAO R Y, et al. HM-RAG: hierarchical multi-agent multimodal retrieval augmented generation[C]//Proceedings of the 33rd ACM International Conference on Multimedia. New York, USA: ACM Press, 2025: 2781-2790.
[37] TIAN Y, LIU F, ZHANG J Y, et al. CoRe-MMRAG: cross-source knowledge reconciliation for multimodal RAG[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2506.02544.
[38] SURI M, MATHUR P, DERNONCOURT F, et al. VisDoM: multi-document QA with visually rich elements using multimodal retrieval-augmented generation[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Philadelphia, USA: Association for Computational Linguistics, 2025: 6088-6109.
[39] XU M J, WANG Z H, CAI H X, et al. A multi-granularity retrieval framework for visually-rich documents[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2505.01457.
[40] WANG Q C, DING R X, ZENG Y, et al. VRAG-RL: empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2505.22019.
[41] YOHANNES H M, MAHMOUD Y, NAZEERUDDIN M, et al. Multimodal Retrieval and Fusion Framework

(MRaFF)[C]// Proceedings of the 8th International Conference on Information and Computer Technologies (ICICT). Washington D. C., USA: IEEE Press, 2025: 186-191.
[42] 王合庆, 魏杰, 景红雨, 等. Meta-RAG: 基于元数据驱动的电力领域检索增强生成框架[J]. 计算机工程, 2026, 52(2): 383-392.
WANG H Q, WEI J, JING H Y, et al. Meta-RAG: a metadata-driven retrieval-augmented generation framework for the power industry[J]. Computer Engineering, 2026, 52(2): 383-392. (in Chinese)
[43] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2312.10997.
[44] BENCHAREF R, RAHICHE A, CHERIET M. DIVE-Doc: downscaling foundational image visual encoder into hierarchical architecture for DocVQA[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Washington D. C., USA: IEEE Press, 2026: 7597-7606.
[45] YU W H, CHEN W, QI G Q, et al. BBox DocVQA: a large scale bounding box grounded dataset for enhancing reasoning in document visual question answer[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2511.15090.
[46] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, USA: Association for Computational Linguistics, 2020: 6769-6781.
[47] GAO S S, ZHAO S S, JIANG X, et al. Scaling beyond context: a survey of multimodal retrieval-augmented generation for document understanding[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2510.15253.
[48] ZHAO S Y, YANG Y Q, WANG Z L, et al. Retrieval Augmented Generation (RAG) and beyond: a comprehensive survey on how to make your LLMs use external data more wisely[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2409.14924.
[49] SHAO Z W, YU Z, WANG M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2023: 14974-14983.
[50] CUI L, XU Y, LÜ T, et al. Document AI: benchmarks, models and applications[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2111.08609.
[51] WANG D S, RAMAN N, SIBUE M, et al. DocLLM: a layout-aware generative language model for multimodal document understanding[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Philadelphia, USA: Association for Computational Linguistics, 2024: 8529-8548.
[52] LEE C Y, LI C L, DOZAT T, et al. FormNet: structural encoding beyond sequential modeling in form document information extraction[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2203.08411.
[53] FAYSSE M, SIBILLE H, WU T, et al. ColPali: efficient document retrieval with vision language models[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2407.01449.
[54] WANG P, BAI S, TAN S N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2409.12191.
[55] BEYER L, STEINER A, PINTO A S, et al. PaliGemma: a versatile 3B VLM for transfer[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2407.07726.
[56] WU Q Y, BANSAL G, ZHANG J Y, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2308.08155.
[57] LOCKARD C, SHIRALKAR P, DONG X L, et al. ZeroShotCeres: zero-shot relation extraction from semi-structured webpages[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2005.07105.
[58] WANG J P, JIN L W, DING K. LiLT: a simple yet effective language-independent layout Transformer for structured document understanding[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2202.13669.
[59] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2404.16130.
[60] NGUYEN T, CHIN P, TAI Y W. MA-RAG: multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2505.20096.
[61] PINEDA V, DAYAN N, DE LARA E. Poster: leveraging geo-spatiality in geo-distributed vector databases[C]// Proceedings of the 10th ACM/IEEE Symposium on Edge Computing. New York, USA: ACM Press, 2025: 1-3.
[62] WAGLE S, MUNIKOTI S, MEYUR R, et al. Leveraging multimodal AI for efficient data discovery in wind energy research[C]// Proceedings of Practice and Experience in Advanced Research Computing 2025: the Power of Collaboration. New York, USA: ACM Press, 2025: 1-3.
[63] MOON J, HONG C. Multimodal clinical decision support for melanoma diagnosis using retrieval-augmented generation and vision-language models[C]//Proceedings of the IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS). Washington D. C., USA: IEEE Press, 2025: 1-6.
[64] KEITA M, HAMIDOUCHE W, EUTAMENE H B, et al. REVEAL: a retrieval-augmented generation approach for contextual identification of synthetic visual content[C]// Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media. New York, USA: ACM Press, 2025: 12-20.
[65] ZENG Q X. Retrieval augmented 3D garment generation from single image[C]// Proceedings of the 33rd ACM International Conference on Multimedia. New York, USA: ACM Press, 2025: 9648-9656.
[66] MAO J B, ZHENG C F, LIU W L, et al. MGRAG: Multimodal grid-aware retrieval augmentation generation framework for power grid work tickets[J]. Pattern Recognition, 2026, 169: 111845.
[67] GONG Z Y, MAI C C, HUANG Y H. MHier-RAG: multi-modal RAG for visual-rich document question-answering via hierarchical and multi-granularity reasoning [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2508.00579.
[68] YUAN X, NING L B, FAN W Q, et al. mKG-RAG: multimodal knowledge graph-enhanced RAG for visual question answering[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2508.05318.
[69] YU S, TANG C Y, XU B K, et al. VisRAG: vision-based multi-modal document retrieval-augmented generation[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2410.10594.
[70] YU B H, WU G W, YAO Z Y, et al. Beyond relevance: utility-driven retrieval for visual document question

answering[C]//Proceedings of International Conference on Intelligent Computing. Singapore: Springer, 2025: 382-393.
[71] CHOI Y, PARK J, YOON J, et al. Zero-shot multimodal document retrieval via cross-modal question generation[C]// Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics 2025: 26068-26083.
[72] KHATTAB O, ZAHARIA M. ColBERT: efficient and effective passage search via contextualized late interaction over BERT[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2020: 39-48.
[73] WANG S N, ZHAO Y J, XIE Y L, et al. Towards reliable vector database management systems: a software testing roadmap for 2030[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2502.20812.
[74] XU M J, DONG J H, HOU J, et al. MM-R5: MultiModal reasoning-enhanced ReRanker via reinforcement learning for document retrieval[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2506.12364.
[75] ASAI A, WU Z Q, WANG Y Z, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2310.11511.
[76] CHANG Y S, CAO G H, NARANG M, et al. WebQA: multihop and multimodal QA[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2022: 16474-16483.
[77] LERNER P, FERRET O, GUINAUDEAU C, et al. ViQuAE, a dataset for knowledge-based visual question answering about named entities[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2022: 3108-3120.
[78] SINGH H, NASERY A, MEHTA D, et al. MIMOQA: multimodal input multimodal output question answering[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: Association for Computational Linguistics, 2021: 5317-5332.
[79] DU Y X, SONG J R, ZHOU Y F, et al. G^2-Reader: dual evolving graphs for multimodal document QA[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2601.22055
[80] MARINO K, RASTEGARI M, FARHADI A, et al. OK-VQA: a visual question answering benchmark requiring external knowledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 3190-3199.
[81] SCHWENK D, KHANDELWAL A, CLARK C, et al. A-OKVQA: a benchmark for visual question answering using world knowledge[C]//Proceedings of ECCV'22. Berlin, Germany: Springer, 2022: 146-162.
[82] SHAH S, MISHRA A, YADATI N, et al. KVQA: knowledge-aware visual question answering [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 8876-8884.
[83] WANG P, WU Q, SHEN C H, et al. FVQA: fact-based visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(10): 2413-2427.
[84] JAIN A, KOTHYARI M, KUMAR V, et al. Select, substitute, search: a new benchmark for knowledge-augmented visual question answering[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2021: 2491-2498.
[85] MATHEW M, BAGAL V, TITO R, et al. InfographicVQA[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Washington D.C., USA: IEEE Press, 2022: 2582-2591.
[86] TANAKA R, NISHIDA K, NISHIDA K, et al. SlideVQA: a dataset for document visual question answering on multiple images [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 13636-13645.
[87] MASRY A, LONG D X, TAN J Q, et al. ChartQA: a benchmark for question answering about charts with visual and logical reasoning[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Philadelphia, USA: Association for Computational Linguistics, 2022: 2263-2279.
[88] ZHU F B, LEI W Q, FENG F L, et al. Towards complex document understanding by discrete reasoning [C]// Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: ACM Press, 2022: 4857-4866.
[89] VAN LANDEGHEM J, POWALSKI R, TITO R, et al. Document Understanding Dataset and Evaluation (DUDE)[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2024: 19471-19483.
[90] CHELLAPPA R, PRAMANICK S, VENUGOPALAN S. SPIQA: a dataset for multimodal question answering on scientific papers[C]//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 118807-118833.
[91] TITO R, KARATZAS D, VALVENY E. Hierarchical multimodal Transformers for multipage DocVQA [J]. Pattern Recognition, 2023, 144: 109834.
[92] SINGH A, NATARAJAN V, SHAH M, et al. Towards VQA models that can read[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 8309-8318.
[93] MISHRA A, SHEKHAR S, SINGH A K, et al. OCR-VQA: visual question answering by reading text in images [C]// Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). Washington D.C., USA: IEEE Press, 2020: 947-952.
[94] CHEN X Y, ZHAO Z H, CHEN L, et al. WebSRC: a dataset for Web-based structural reading comprehension [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL Press, 2021: 4173-4185.
[95] LI B H, GE Y Y, CHEN Y, et al. SEED-Bench-2-Plus: benchmarking multimodal large language models with text-rich visual comprehension [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2404.16790.
[96] GOYAL Y, KHOT T, AGRAWAL A, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering[J]. International Journal of Computer Vision, 2019, 127(4): 398-414.
[97] ZELLERS R, BISK Y, FARHADI A, et al. From recognition to cognition: visual commonsense reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 6713-6724.

[98] LIU Y, DUAN H D, ZHANG Y H, et al. MMBench: is your multi-modal model an all-around player? [C] // Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2025: 216-233.
[99] LU P, MISHRA S, XIA T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2209.09513.
[100] LU P, BANSAL H, XIA T, et al. MathVista: evaluating mathematical reasoning of foundation models in visual contexts[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2310.02255.
[101] KAHOU S E, MICHALSKI V, ATKINSON A, et al. FigureQA: an annotated figure dataset for visual reasoning [EB/OL]. [2025-12-27]. https://arxiv.org/abs/1710.07300.
[102] FU C Y, DAI Y H, LUO Y D, et al. Video-MME: the first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2025: 24108-24118.
[103] YU Z, XU D J, YU J, et al. ActivityNet-QA: a dataset for understanding complex Web videos via question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 9127-9134.
[104] MANGALAM K, AKSHULAKOV R, MALIK J. EgoSchema: a diagnostic benchmark for very long-form video language understanding [EB/OL]. [2025-12-27]. https://arxiv.org/abs/2308.09126.
[105] YI D Y, ZHU G B, DING C L, et al. MME-industry: a cross-industry multimodal evaluation benchmark[EB/OL]. [2025-12-27]. https://arxiv.org/abs/2501.16688.
[106] LIU Y L, LI Z, HUANG M X, et al. OCRBench: on the hidden mystery of OCR in large multimodal models[J]. Science China Information Sciences, 2024, 67(12): 220102.
[107] HAN X, LI Z, CAO H, et al. Multimodal spatio-temporal data visualization technologies for contemporary urban landscape architecture: a review and prospect in the context of smart cities[J]. Land, 2025, 14(5): 1069.
[108] RHAIEM M A B, SELMI M, FARAH I R, et al. Leveraging volunteered geographical information and spatio temporal big data in disaster management: opportunity and challenges[J]. International Journal of Data Science and Analytics, 2025, 21(1): 25.
[109] KUMAR R, BHANU M, MENDES-MOREIRA J, et al. Spatio-temporal predictive modeling techniques for different domains: a survey[J]. ACM Computing Surveys, 2025, 57(2): 1-42.
[110] CAO Y, STEFFEY S, HE J B, et al. Medical image retrieval: a multimodal approach[J]. Cancer Informatics, 2014, 13(3): 125-136.
[111] SHAIK T, TAO X H, LI L, et al. A survey of multimodal information fusion for smart healthcare: mapping the journey from data to wisdom [J]. Information Fusion, 2024, 102: 102040.
[112] CHEN Y, GE X K, YANG S L, et al. A survey on multimodal knowledge graphs: construction, completion and applications[J]. Mathematics, 2023, 11(8): 1815.
[113] WEN J, ZHANG X, RUSH E, et al. Multimodal representation learning for predicting molecule-disease relations[J]. Bioinformatics, 2023, 39(2): btad085.
[114] KREUTZ C K, SCHENKEL R. Scientific paper recommendation systems: a literature review of recent publications[J]. International Journal on Digital Libraries, 2022, 23(4): 335-369.

文字编辑 陆燕菲
栏目编辑 宋 圆