Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding

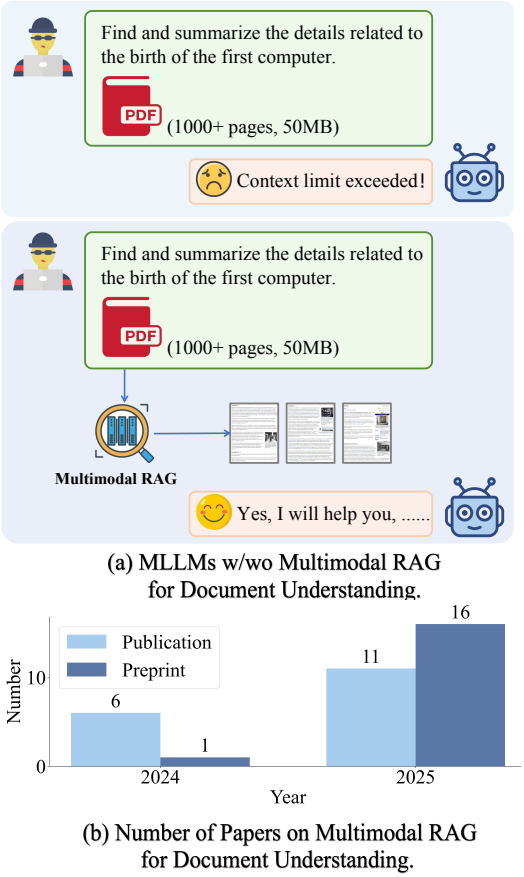
Anonymous ACL submission

Abstract

Document understanding is critical for applications from financial analysis to scientific discovery. Current approaches, whether OCR-based pipelines feeding Large Language Models (LLMs) or native Multimodal LLMs (MLLMs), face key limitations: the former loses structural detail, while the latter struggles with context modeling. Retrieval-Augmented Generation (RAG) helps ground models in external data, but documents' multimodal nature, *i.e.*, combining text, tables, charts, and layout, demands a more advanced paradigm: Multimodal RAG. This approach enables holistic retrieval and reasoning across all modalities, unlocking comprehensive document intelligence. Recognizing its importance, this paper presents a systematic survey of Multimodal RAG for document understanding. We propose a taxonomy based on domain, retrieval modality, and granularity, and review advances involving graph structures and agentic frameworks. We also summarize key datasets, benchmarks, and applications, and highlight open challenges in efficiency, fine-grained representation, and robustness, providing a roadmap for future progress in document AI.

1 Introduction

Document understanding has become a pivotal task in the era of information explosion, as it empowers machines to automatically interpret, organize, and reason over the massive volumes of unstructured and semi-structured documents produced across diverse domains (Subramani et al., 2020; Ding et al., 2024). Early studies primarily focused on text-centric documents, relying on optical character recognition (OCR) techniques (Gu et al., 2021; Appalaraju et al., 2021; Shi et al., 2016) to support layout analysis and key information extraction. However, in real-world scenarios, particularly in scientific domains, documents are often visually rich and contain complex elements such as tables, charts, and images (Park et al., 2019; Ding et al., 2025a). With the rapid progress of Large Language Models (LLMs) and the rising demand for understanding increasingly complex and diverse document types, developing robust and generalizable document understanding frameworks has become an area of growing interest.



(a) MLLMs w/wo Multimodal RAG for Document Understanding.

(b) Number of Papers on Multimodal RAG for Document Understanding.

Figure 1: **Impact and research progress of Multimodal RAG for document understanding:** (a) MLLMs with and without Multimodal RAG for large document comprehension. (b) Growth in related publications from 2024 to 2025.

In visually rich document understanding, different approaches have emerged to address the challenges of integrating layout, text, and structural information. Multimodal LLM (MLLM)-native

methods commonly represent documents as long image sequences, enabling unified learning across modalities with MLLMs (Duan et al., 2025; Xiong et al., 2025; Yu et al., 2025b; Zhou et al., 2024; Nassar et al., 2025; Ye et al., 2023; Hu et al., 2024a,b). While effective, these models struggle with very long documents spanning hundreds or thousands of pages, where sequence length limitations can hinder accurate retrieval and increase the risk of hallucination (Deng et al., 2024; Ma et al., 2024b). To improve modularity and robustness, agent-based approaches introduce specialized agents for subtasks such as layout analysis, content extraction, instruction decomposition, and verification (Liu et al., 2025; Han et al., 2025; Wang et al., 2025a; Wu et al.; Yu et al., 2025c), though such designs often increase system complexity due to coordination overhead. Retrieval-augmented generation (RAG) methods provide another direction by grounding responses with external knowledge, typically retrieving the top-K most relevant pages (see Figure 1 (a)) across one or more documents (Lewis et al., 2020). Importantly, these paradigms are not mutually exclusive: RAG-based systems may employ agents to manage retrieval and verification, while agent-based workflows often incorporate RAG as one of the agent nodes, yielding more flexible hybrid frameworks. These complementary perspectives have shaped the landscape of document understanding, yet among them, RAG has drawn particular attention for its practicality and rapid growth (Arslan et al., 2024; Fan et al., 2024).

Early RAG studies mainly rely on text-centric strategies, extracting text via OCR or combining OCR with MLLM-generated captions for visually rich documents, followed by encoding for retrieval (Wang et al., 2022; Li et al., 2023; Chen et al., 2024c; Khattab and Zaharia, 2020). Despite their effectiveness in certain scenarios, such text-based approaches exhibit fundamental limitations in handling visually rich documents, as they fail to adequately capture cross-modal cues and structural semantics (Abootorabi et al., 2025; Mei et al., 2025). To address these shortcomings, recent efforts have increasingly focused on multimodal RAG frameworks. The growth trend in the number of papers is shown in Figure 1 (b). These methods often represent multi-page documents as image sequences (Faysse et al., 2024; Yu et al., 2024), enabling visual encoders to extract richer representations for retrieval. Recent advances in multimodal RAG have increasingly emphasized finer-grained modeling within individual pages, including tables, charts, and other structured elements, to improve retrieval accuracy and robustness (Wang et al., 2025c; Choi et al., 2025). Extending beyond these coarse-to-fine refinements, recent studies have also investigated graph-based indexing (Yuan et al., 2025) and multi-agent frameworks (Liu et al., 2025), which provide complementary mechanisms for structured reasoning and collaborative coordination in multimodal RAG.

This rapid evolution and increasing complexity in the field have naturally prompted efforts to synthesize the existing literature. However, a closer look reveals a significant gap. Prior surveys have reviewed RAG from multiple perspectives (Arslan et al., 2024; Fan et al., 2024; Gao et al., 2023; Hu and Lu, 2024; Gupta et al., 2024; Zhao et al., 2024; Church et al., 2024). In parallel, recent surveys examining multimodal RAG (Zhao et al., 2023; Abootorabi et al., 2025; Mei et al., 2025) offer limited coverage of document understanding, typically discussing only a few relevant methods. Conversely, while document understanding has been extensively reviewed (Subramani et al., 2020; Ding et al., 2024; Nandi and Sathya, 2024; Van Landeghem et al., 2023; Ding et al., 2025b), existing surveys rarely address multimodal RAG. To bridge this gap, we present the first comprehensive survey that explicitly connects multimodal RAG and document understanding. Unlike prior works that emphasize one aspect while overlooking the other, our survey systematically analyzes their intersection and organizes the most extensive collection of studies in this emerging field. Our contributions can be summarized as follows: (1) We present a comprehensive survey that categorizes existing methods by domain, retrieval modality, granularity, and hybrid enhancements, offering a structured perspective for future research. (2) We compile a broad collection of multimodal RAG datasets, benchmarks, and comparative results for systematic evaluation, and survey evaluation metrics spanning both retrieval and generation. Together, these contributions outline a coherent landscape of multimodal RAG for document understanding, providing both a reference and guidance for future progress.

2 Preliminary

In RAG, a system retrieves a set of relevant document pages and then generates a response conditioned on that evidence. Retrieval can be *closed-*

domain (*e.g.*, grounding to a single source document) or *open-domain* (searching a large corpus). We denote the candidate pool by $D = \{d_i\}_{i=1}^{N}$. Each d_i may include a raster image as well as OCR text T_i. Using modality-specific encoders, we map queries and documents into a shared embedding space. Our notation uses lower-case symbols with subscripts for vectors (*e.g.*, z_i, e_q), and we compute similarity using inner products. Typically, the query q is text, so we compute both text–text and text–image similarities in this shared space (and, if q includes images, e_q^{img} can be defined analogously).

To embed documents and queries, we use image and text encoders: $z_i^{\text{img}} = \text{Enc}_{\text{img}}(d_i)$, $z_i^{\text{text}} = \text{Enc}_{\text{text}}(T_i)$, and $e_q^{\text{text}} = \text{Enc}_{\text{text}}(q)$. Within each modality pair, similarities are inner products (optionally with unit-norm embeddings so the score is cosine similarity): $s_{\text{text}}(e_q, z_i) = \langle e_q^{\text{text}}, z_i^{\text{text}} \rangle$ and $s_{\text{img}}(e_q, z_i) = \langle e_q^{\text{text}}, z_i^{\text{img}} \rangle$.

Vision-only retrieval. When using only the image channel (*i.e.*, for text-image similarity), we rank documents with the score $s_{\text{img}}(e_q, z_i)$ and select those that exceed a threshold τ_{img} (or simply take the K results):

$$X_{\text{img}} = \{\, d_i \in D \mid s_{\text{img}}(e_q, z_i) \geq \tau_{\text{img}} \,\}. \quad (1)$$

Joint vision–text retrieval. We consider two widely used strategies.

(a) Confidence-weighted score fusion. Image and text scores are combined with a convex weight that reflects per-item or per-query confidence. Let $\lambda_i \in [0, 1]$ denote the image confidence for d_i (*e.g.*, from calibration or OCR quality); setting λ_i=1 recovers vision-only and λ_i=0 text-only:

$$\begin{aligned} s_{\text{conf}}(e_q, z_i) &= \lambda_i \, s_{\text{img}}(e_q, z_i) \\ &\quad + (1 - \lambda_i) \, s_{\text{text}}(e_q, z_i), \\ X_{\text{conf}} &= \{\, d_i \in D \mid s_{\text{conf}}(e_q, z_i) \geq \tau_{\text{conf}} \,\}. \end{aligned} \quad (2)$$

(b) Union of modality-specific pages. This strategy involves retrieving pages with each modality independently and then forming the union of the results (optionally followed by deduplication or rank fusion such as Borda or reciprocal-rank fusion (Cormack et al., 2009; Calumby et al., 2017)) using modality-aware thresholds $\tau_{\text{img}}, \tau_{\text{text}}$:

$$\begin{aligned} X_{\text{img}} &= \{\, d_i \in D \mid s_{\text{img}}(e_q, z_i) \geq \tau_{\text{img}} \,\}, \\ X_{\text{text}} &= \{\, d_i \in D \mid s_{\text{text}}(e_q, z_i) \geq \tau_{\text{text}} \,\}, \\ X_{\cup} &= X_{\text{img}} \cup X_{\text{text}}. \end{aligned} \quad (3)$$

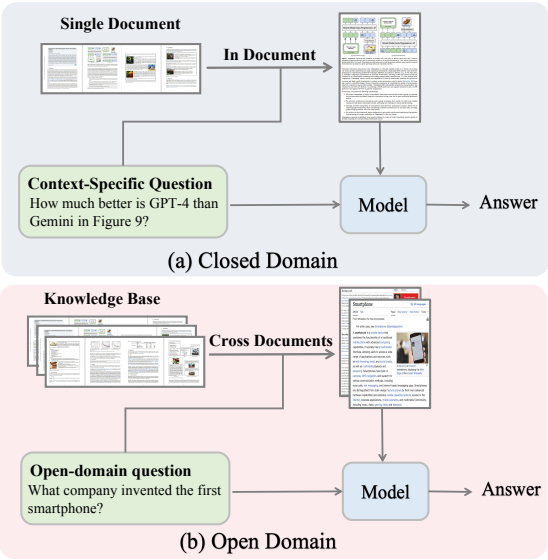


Figure 2: **Comparison between closed-domain and open-domain multimodal RAG**. (a) In the closed domain, the model leverages in-document retrieval from a single document to answer context-specific questions. (b) In the open domain, the model relies on cross-document retrieval from multiple documents to answer open-ended questions.

(Equivalently, one may use top-K per modality and take the union $X_{\cup}^{(K)}$.)

Generation. A generator $\mathcal{G}$ conditions on the original query and the retrieved context chosen as X_{img}, X_{conf}, or $X_{\cup}$ depending on the retrieval regime and produces the final response. The specific mechanism for aggregating multiple pages (*e.g.*, via cross-attention or learned pooling) is left abstract:

$$r = \mathcal{G}(q, X). \quad (4)$$

3 Key Innovations and Methodologies

In this section, we examine the core innovations and methodological advances in recent multimodal RAG approaches for document understanding. Table 1 presents a systematic comparison of representative methods along several key dimensions, including domain openness, retrieval modality, retrieval granularity, graph-based integration, and agent-based enhancement. To provide a structured discussion, we elaborate on each dimension in turn: the distinction between open- and closed-domain settings (Section 3.1), the impact of retrieval modality (Section 3.2), the role of retrieval granularity (Section 3.3), agent and graph based Hybrid Enhancements (Section 3.4).

Method	Venue	LLM/VLM	Vision Encoder	Training	OCR	Domain	Modality	Granularity	Graph	Agent
DSE (2024a)	EMNLP	Phi3V	CLIP-ViT-L/14	✓	✗	Open	Image	Page	✗	✗
ColPali (2024)	ICLR	PaliGemma-3B	SigLIP-SO400M	✓	✗	Open	Image	Page	✗	✗
ColQwen2 (2024)	ICLR	Qwen2-VL-2B	ViT-BigG	✓	✗	Open	Image	Page	✗	✗
CRAEM (2024a)	ACM MM	LLaMA2-7B	Pix2Struct	✓	✓	Closed	Image+Text	Page	✗	✗
VisRAG (2024)	ICLR	MiniCPM-V2.0	SigLIP-SO400M	✓	✗	Open	Image	Page	✗	✗
SV-RAG (2024b)	ICLR	InternVL2-4B	InternViT-300M	✓	✗	Closed	Image	Page	✗	✗
M3DocRAG (2024a)	Preprint	Qwen2-VL-7B	Vit-BigG	✗	✗	Open	Image	Page	✗	✗
VisDoMRAG (2025)	NAACL	Qwen2-VL-2B	ViT-BigG	✗	✓	Open	Image+Text	Page	✗	✗
GME (2025b)	CVPR	Qwen2-VL-7B	ViT-BigG	✓	✓	Open	Image+Text	Page	✗	✗
ViDoRAG (2025b)	EMNLP	Qwen2.5-VL-7B	ViT-BigG	✗	✓	Open	Image+Text	Page	✗	✓
HM-RAG (2025)	ACM MM	Qwen2.5-VL-7B	ViT-BigG	✗	✓	Open	Image+Text	Page	✓	✓
VDocRAG (2025)	CVPR	Phi3V	CLIP-ViT-L/14	✓	✗	Open	Image	Page	✗	✗
FRAG (2025)	Preprint	InternVL2-8B	InternViT-300M	✗	✗	Closed	Image	Page	✗	✗
MG-RAG (2025b)	Preprint	Qwen2.5-VL-3B-Instruct	ViT-BigG	✗	✓	Closed	Image+Text	Element	✗	✗
VRAG-RL (2025c)	Preprint	Qwen2.5-VL-7B-Instruct	ViT-BigG	✓	✗	Open	Image	Element	✗	✗
CoRe-MMRAG (2025)	ACL	Qwen2-VL-7B	ViT-BigG	✓	✓	Open	Image+Text	Page	✗	✗
Light-ColPali (2025)	ACL	PaliGemma	SigLIP-SO400M	✓	✗	Open	Image	Page	✗	✗
MM-R5 (2025a)	Preprint	Qwen2.5-VL-7B	ViT-BigG	✓	✗	Open	Image	Page	✗	✗
SimpleDoc (2025)	Preprint	Qwen2.5-VL-3B-Instruct	ViT-BigG	✗	✗	Open	Image+Text	Page	✗	✗
VisChunk (2025)	Preprint	Gemini-2.5-Pro	–	✗	✓	Closed	Image+Text	Page	✗	✗
DocVQA-RAP (2025a)	ICIC	Qwen2-VL-2B	ViT-BigG	✗	✗	Open	Image	Element	✗	✗
RL-QR (2025)	Preprint	Qwen2.5-VL-3B-Instruct	ViT-BigG	✓	✗	Open	Image	Page	✗	✗
MMRAG-DocQA (2025)	Preprint	Qwen-VL-Plus	ViT-BigG	✗	✓	Closed	Image+Text	Element	✗	✗
Patho-AgenticRAG (2025a)	Preprint	Qwen2-VL-2B	ViT-BigG	✓	✗	Open	Image	Page	✗	✓
M2IO-R1 (2025a)	Preprint	BGE-M3	–	✓	✓	Open	Image+Text	Page	✗	✗
mKG-RAG (2025)	Preprint	LLaMA-3.1-8B	CLIP ViT-L/14	✓	✓	Open	Image+Text	Element	✓	✗
DB3Team-RAG (2025)	Preprint	Llama 3.2-VL	CLIP ViT-L/14	✓	✓	Open	Image+Text	Page	✓	✗
PREMIR (2025)	EMNLP	Qwen2.5-VL-72B	ViT-BigG	✗	✓	Open	Image+Text	Element	✗	✗
ReDocRAG (2025)	ICDAR WML	Qwen2.5-VL-7B-Instruct	ViT-BigG	✓	✗	Closed	Image	Page	✗	✗
CMRAG (2025)	Preprint	Qwen2.5-VL-7B-Instruct	ViT-BigG	✗	✓	Open	Image+Text	Page	✗	✗
MoLoRAG (2025)	EMNLP	Qwen2.5-VL-7B	ViT-BigG	✓	✗	Open	Image	Page	✓	✗
SERVAL (2025)	Preprint	InternVL3-14B	InternViT-300M	✗	✗	Open	Image	Page	✗	✗
MetaEmbed (2025b)	Preprint	Qwen2.5-VL-32B	ViT-BigG	✓	✗	Open	Image	Page	✗	✗
DocPruner (2025)	Preprint	Qwen2.5-VL-3B-Instruct	ViT-BigG	✓	✗	Open	Image	Page	✗	✗

Table 1: **Comparison of recent Multimodal RAG methods for document understanding.** The table summarizes methods along the following dimensions: venue, backbone LLM/VLM, vision encoder, training status, OCR integration, domain scope, retrieval modality, retrieval granularity, graph incorporation, and agent usage.

3.1 Open and Closed Domain

RAG addresses the limitations of LLMs in knowledge acquisition, such as knowledge cut-off, and extends their applicability to specialized domains (Lewis et al., 2020; Joren et al.; Ye et al., 2024; Gupta et al., 2024; Huang and Huang, 2024; Cheng et al., 2025). For document understanding, open-domain multimodal RAG retrieves information from large corpora of domain-specific documents to construct extensive knowledge bases. In contrast, closed-domain multimodal RAG focuses on a single document and selects only the most relevant pages for retrieval, thereby reducing input length and mitigating issues related to limited context windows and hallucination. The distinction between open-domain and closed-domain multimodal RAG is illustrated in Figure 2.

Open-Domain Multimodal RAG. Open-domain multimodal RAG enhances an LLM's knowledge in specialized domains by constructing retrieval databases from large collections of documents. Early approaches typically apply OCR to all documents to build text-based retrieval indices (Wang et al., 2022; Li et al., 2023; Chen et al., 2024c; Khattab and Zaharia, 2020), but this process is computationally expensive and inefficient. To improve scalability, recent methods such as DSE (Ma et al., 2024a) and ColPali (Faysse et al., 2024) leverage vision-language models (VLMs) to encode document pages directly into image embeddings, achieving significant efficiency gains. Despite these advances, most approaches still focus on reasoning within single documents and lack explicit mechanisms for integrating knowledge across sources. Addressing this limitation, M3DocRAG (Cho et al., 2024a) introduces approximate indexing to accelerate large-scale retrieval and establishes the benchmark M3DoCVQA with over 3,000 documents, while VDocRAG (Tanaka et al., 2025) constructs the OpenDocVQA dataset and mitigates page-level information loss by compressing visual content into dense token representations aligned with text.

Closed-Domain Multimodal RAG. Closed-domain multimodal RAG is designed for practical scenarios where MLLMs encounter difficulties with extremely long documents or videos. Current MLLMs remain constrained by limited context windows, and long-context processing often amplifies the risk of hallucination. To address this, closed-

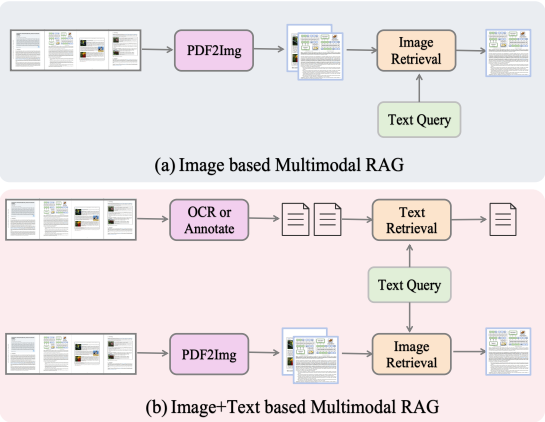


Figure 3: **Comparison of retrieval modality:** (a) image-based RAG retrieves information solely from page images, offering efficiency but limited textual detail; (b) image+text-based RAG integrates OCR/annotations with visual features, enabling richer retrieval at the cost of higher processing complexity.

domain approaches retrieve only the most relevant segments (*e.g.*, pages or frames) from a target document and provide them as input to the MLLM, thereby improving both efficiency and reliability. For single-document visual question answering (DocVQA), SV-RAG (Chen et al., 2024b) employs the MLLM itself as a multimodal retriever, with specialized adapters for page retrieval and evidence-based reasoning. FRAG (Huang et al., 2025), by contrast, independently scores each frame or page, applies a Top-K selection to retain the most informative content, and then delegates answer generation to existing LMMs. CREAM (Zhang et al., 2024a) introduces a coarse-to-fine multimodal retrieval and attention-pooling integration framework, enabling effective cross-page reasoning and multi-page document comprehension for visual question answering. All approaches demonstrate that closed-domain multimodal RAG enables effective comprehension of long documents and videos without extending the model's context length.

3.2 Retrieval Modality

Early text-only RAG methods rely exclusively on textual signals for retrieval, which limits their practical utility: they require time-consuming OCR and underperform on visually rich documents. To address these limitations, current research advances multimodal RAG. One approach treats each page as an image and encodes it with the vision encoder of a VLM. Another adopts hybrid designs that pair page-level images with OCR-extracted text or auxiliary textual annotations generated by MLLMs. The resulting cross-modal representations then support retrieval independently or via score fusion, where similarity scores from different modalities combine to improve performance.

Image-based Retrieval Modality. To handle visually rich documents, most existing methods represent each page as an image and encode it with VLMs, using the VLMs' hidden states as page-level representations (see Figure 3 (a)). In parallel, the query is encoded, and page–query relevance is computed via similarity-based ranking (Ma et al., 2024a; Faysse et al., 2024; Yu et al., 2024; Chen et al., 2024b; Ma et al., 2025; Yu et al., 2025a). Building on image-based embeddings, MM-R5 (Xu et al., 2025a) introduces a reasoning-enhanced reranker that combines supervised fine-tuning and reinforcement learning to strengthen instruction following, elicit explicit reasoning chains, and leverage task-specific rewards for greater precision and interpretability. Complementing this direction, Light-ColPali (Ma et al., 2025) targets memory-efficient visual document retrieval by reducing patch-level embeddings through token merging, achieving near-optimal retrieval with a fraction of the original memory footprint.

Image+Text based Retrieval Modality. Leveraging both image and text for retrieval mitigates the loss of fine-grained textual cues that arise when relying solely on page-level VLM encoders. The text channel is derived from OCR (Zhang et al., 2024a; Suri et al., 2025; Liu et al., 2025; Wang et al., 2025b) or from summary annotations generated by large VLMs (Jain et al., 2025; Choi et al., 2025) (see Figure 2 (b)). VisDomRAG (Suri et al., 2025) and HM-RAG (Liu et al., 2025) adopt a dual-path pipeline: they retrieve and reason within each modality, then summarize and fuse the results into a single answer. By contrast, ViDoRAG (Wang et al., 2025b) and PREMIR (Choi et al., 2025) also retrieve per modality but merge candidates via a simple union before answer generation. Complementing these designs, SimpleDoc (Jain et al., 2025) uses a two-stage scheme for DocVQA: embedding-based candidate selection followed by re-ranking with VLM-generated page summaries, so that the summaries provide richer semantics for more precise evidence aggregation.

3.3 Retrieval Granularity

In document-oriented multimodal RAG, early studies typically treat the page as the atomic retrieval unit, without modeling finer structures such as ta-

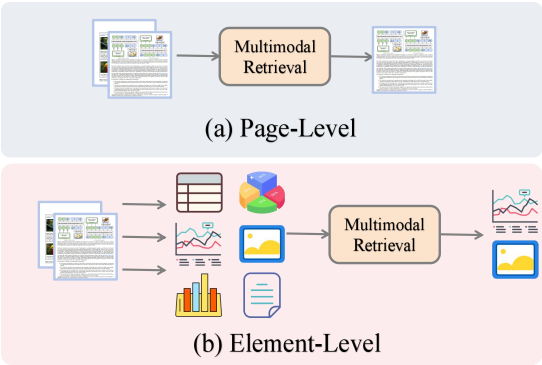


Figure 4: **Comparison of retrieval granularity in multimodal document search.** (a) Page-level: entire pages are encoded and ranked as whole units. (b) Element-level: pages are decomposed into tables, charts, images, and text blocks; retrieval operates over these elements to localize evidence and aggregate results.

bles, charts, or layout cues (see Figure 4). Recent work increasingly focuses on retrieval at a finer, within-page granularity. Some approaches explicitly encode these components to enhance retrieval accuracy, whereas others adopt a two-stage pipeline: first retrieve the most relevant pages, then perform retrieval within those pages to establish fine-grained grounding. This shift toward finer retrieval granularity enables models to deliver more precise and contextually grounded answers.

VRAG-RL (Wang et al., 2025c) equips LLMs with grounding capabilities through reinforcement learning, enabling the model to focus on finer-grained regions within retrieved pages that directly relate to the query. In a complementary way, MG-RAG (Xu et al., 2025b) adopts a multi-granularity multimodal retrieval strategy with hierarchical encoding and layout-aware search, which allows retrieval not only at the page level but also within specific structures such as tables and images. Similarly, DocVQA-RAP (Yu et al., 2025a) introduces a utility-driven retrieval mechanism that prioritizes document segments based on their contribution to answer quality, thereby filtering out fine-grained but redundant evidence. Along this line, MMRAG-DocQA (Gong et al., 2025) employs hierarchical indexing and multi-granularity semantic retrieval to capture fine-grained multimodal associations and long-distance dependencies across pages. Furthermore, mKG-RAG (Yuan et al., 2025) leverages multimodal knowledge graphs and a dual-stage retrieval strategy to refine fine-grained evidence selection by aligning entities and relationships across modalities. Finally, PREMIR (Choi et al., 2025) conducts fine-grained retrieval inside pages by generating predefined QA pairs for tables and charts, and matching them against queries to enhance precision. Together, these approaches illustrate a clear trend toward increasingly fine-grained retrieval in multimodal and document-heavy RAG systems.

3.4 Hybrid Enhancements for Multimodal RAG

Graph-based Multimodal RAG. Graph-based multimodal RAG extends the framework by representing multimodal content as an explicit graph, as shown in Figure 5 (a). Nodes denote modalities or atomic content units such as pages, text spans, images, tables, and layout blocks, while edges encode semantic, spatial, and contextual relations. Retrieval and reasoning over this multimodal graph integrate heterogeneous evidence more effectively, enable finer-grained grounding, and improve the robustness and interpretability of multimodal RAG systems.

HM-RAG (Liu et al., 2025) introduces a hierarchical multi-agent framework where graph-based databases serve as one of the critical modalities for multi-source retrieval, enabling the system to capture structured relations that complement unstructured and web-based evidence. In contrast, mKG-RAG (Yuan et al., 2025) explicitly constructs multimodal knowledge graphs by aligning entities and relationships across vision and text, using them as structured repositories to refine retrieval precision and reliability in knowledge-intensive VQA tasks. Building on this idea, DB3Team-RAG (Xia et al., 2025) incorporates image-indexed knowledge graphs within its domain-specific retrieval pipelines, allowing the system to handle complex multi-turn and ego-centric queries with structured multimodal context. Distinct from these approaches, MoLoRAG (Wu et al., 2025) leverages a page graph that encodes logical connections between document pages, using graph traversal to retrieve semantically and logically relevant evidence for multi-page document understanding. Collectively, these methods underscore the central role of graph structures, whether as retrieval sources, structured repositories, or logical page representations, in advancing fine-grained and reliable retrieval within multimodal RAG systems.

Agent-based Multimodal RAG Agent-based multimodal RAG extends the paradigm by deploying autonomous agents that orchestrate re-

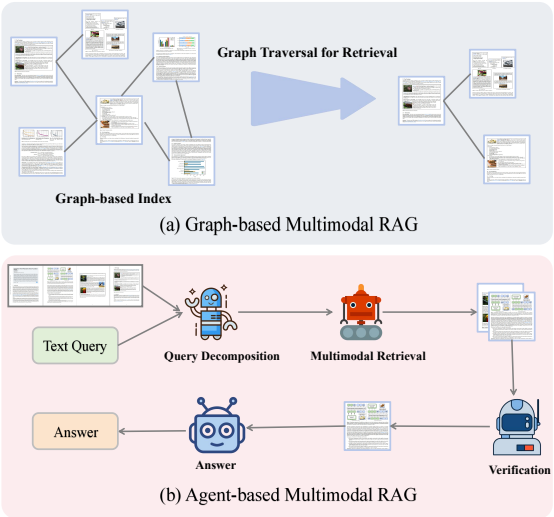


Figure 5: **Hybrid enhancements for multimodal RAG.** (a) Graph-based: documents/elements form a graph index, and retrieval proceeds via graph traversal to surface relevant neighborhoods. (b) Agent-based: an LLM agent decomposes the text query, orchestrates multimodal retrieval, verifies the gathered evidence, and synthesizes the final answer.

Dataset	# Queries	# Documents/Images	Content
TabFQuAD (2020)	210	210 (I)	Tables
PlotQA (2020)	28.9M	224K (I)	Charts
DocVQA (2021)	50K	12,767 (I)	Text, Tables, Charts
VisualMRC (2021)	30,562	10,197 (I)	Text, Tables, Charts
TAT-DQA (2022)	16,558	2,758 (D)	Text, Tables, Charts
InfoVQA (2022)	30K	5.4K (I)	Text, Tables, Charts
ChartQA (2022)	23.1K	17.1K (I)	Charts
ScienceQA (2022)	21K	7,803 (I)	Text, Tables, Charts
DUDE (2023)	41,491	4,974 (D)	Text, Tables, Charts
SlideVQA (2023)	52K	14.5K (I)	Slides
ArXivQA (2024a)	100K	16.6K (D)	Text, Tables, Charts
MMLongBench-Doc (2024b)	1,062	130 (D)	Text, Tables, Charts, Slides
PaperTab (2024)	393	307 (D)	Text, Tables
FetaTab (2024)	1,023	878 (D)	Tables
SPIQA (2024)	27K	25.5K (D)	Tables, Charts
LongDocUrl (2024)	2,325	396 (D)	Text, Tables, Charts
ViDoRe (2024)	3.8K	8.3K (D)	Text, Tables, Charts
VisR-Bench (2024b)	471	226 (D)	Text, Tables, Charts, Slides
M3DoCVQA (2024a)	2,441	3,368 (D)	Text, Tables, Charts
VisDoMBench (2025)	2,271	1,277 (D)	Text, Tables, Charts, Slides
ViDoSeek (2025b)	1,142	300 (D)	Text, Tables, Charts
OpenDocVQA (2025)	206K	43K (I)	Text, Tables, Charts

Table 2: Overview of datasets and benchmarks in multimodal RAG for document understanding. We report the number of queries, dataset size, and covered content types (Text, Tables, Charts, Slides). (D) and (I) indicate that the count refers to documents or images, respectively. **The upper part** covers widely used multimodal document understanding datasets; **the lower part** compiles recent multimodal RAG benchmarks introduced by methods surveyed in this paper to address prior limitations.

trieval–generation interactions across modalities. These agents dynamically formulate queries, choose retrieval strategies, and adaptively fuse information from text, images, tables, and other modalities according to task requirements (see Figure 5 (b)). Multi-agent collaboration further supports iterative reasoning, verification, and evidence refinement, thereby enhancing the accuracy, reliability, and transparency of multimodal RAG systems.

ViDoRAG (Wang et al., 2025b) introduces an iterative agent workflow in which agents are responsible for exploration, summarization, and reflection, thereby enhancing multimodal retrieval and reasoning across visually rich documents. In comparison, HM-RAG (Liu et al., 2025) designs a hierarchical multi-agent architecture composed of a Decomposition Agent for query rewriting and context augmentation, modality-specific Retrieval Agents for parallel evidence collection from diverse sources, and a Decision Agent that integrates results through consistency voting and refinement. Extending this agent-based paradigm into the medical domain, Patho-AgenticRAG (Zhang et al., 2025a) employs agents capable of task decomposition and multi-turn search interactions, enabling accurate retrieval of joint text–image evidence from pathology textbooks while mitigating hallucinations in diagnostic reasoning. Taken together, these frameworks exemplify how specialized agent designs, ranging from iterative reasoning to hierarchical orchestration and domain-tailored search, advance the fine-grained retrieval and reasoning capabilities of multimodal RAG systems.

4 Dataset and Benchmark

Datasets and benchmarks commonly used in multimodal RAG for document understanding typically consist of visually rich document collections. We compile the most widely adopted datasets and benchmarks for this task, reporting their query volume, dataset scale, and data types, such as text, tables, charts, and slides. The representative datasets and benchmarks are presented in the upper part of Table 2. They support the training and evaluation of multimodal models and also serve as essential resources for constructing broader evaluation frameworks. Nevertheless, these resources still exhibit important limitations, motivating the development of more diverse and realistic benchmarks.

Many approaches have also identified various shortcomings in multimodal RAG, which has led to the construction of diverse benchmarks. These newly proposed benchmarks are summarized in the lower part of Table 2. ColPali (Faysse et al., 2024)

constructs a comprehensive and diverse benchmark, ViDoRe, which includes both academic and practical tasks across domains such as energy, government, and healthcare, while SV-RAG (Chen et al., 2024b) collects a human-verified dataset of great diversity to establish VISR-BENCH. Similarly, M3DocVQA (Cho et al., 2024a), VisDoMRAG (Suri et al., 2025) and VDocRAG (Tanaka et al., 2025) address the limitation that most existing benchmarks focus on single-document retrieval rather than cross-document open-domain scenarios by introducing the open-domain benchmarks M3DocVQA, VisDoMBench and OpenDocVQA, respectively. To further overcome the constraint that existing VQA datasets typically pair queries with only one or a few images, whereas real-world applications require reasoning over large-scale corpora, ViDoRAG (Wang et al., 2025b) introduces ViDoSeek, a novel visually rich document dataset specifically tailored for RAG systems, enabling queries with unique answers in realistic retrieval settings and supporting rigorous evaluation of multimodal document understanding.

We also present the performance of different multimodal RAG methods across various benchmarks, along with a detailed explanation of the evaluation metrics and their computation. The full details are provided in the appendix.

5 Application

Multimodal RAG increasingly serves document understanding across finance, scientific research, and survey analysis. In finance, MultiFinRAG (Gondhalekar et al., 2025) improves question answering over reports by jointly modeling narrative text, tables, and figures, while FinRAGBench-V (Zhao et al., 2025) provides a benchmark that emphasizes visual citation for transparent evidence traceability in financial documents. In the scientific domain, HiPerRAG (Gokdemir et al., 2025) enables cross-modal retrieval and reasoning at the scale of millions of research papers, and CollEX (Schneider et al., 2025) supports interactive exploration of multimodal scientific corpora. In the social sciences, a Eurobarometer-based framework embeds RAG with multimodal LLMs (Papageorgiou et al., 2025) to unify text and infographics, improving the interpretability of survey data. Taken together, these applications demonstrate how multimodal RAG strengthens the capacity to understand and leverage complex documents across fields.

6 Challenge and Future Direction

Despite continued advances in multimodal RAG for document understanding, several open challenges direct future research. First, efficiency is a central concern (Ma et al., 2025): integrating high-dimensional visual and textual features incurs substantial computational cost, constraining scalability. Promising directions include lightweight multimodal encoders, adaptive retrieval, and memory-efficient fusion that reduce latency without sacrificing retrieval accuracy. Second, finer-grained document representations are necessary (Wang et al., 2025c; Xu et al., 2025b; Yu et al., 2025a; Gong et al., 2025; Choi et al., 2025): many models operate at the page or paragraph level and overlook tables, figures, footnotes, and layout-specific semantics. Hierarchical encoders and attention mechanisms that capture micro-level structure while preserving collection-level context can improve interpretability and enhance downstream reasoning and decision-making. Finally, security and robustness (Shereen et al., 2025; Cho et al., 2024b; Nazary et al., 2025; Jiang et al., 2024; Xian et al., 2024) rise in importance as adoption expands in finance, healthcare, and law: systems must resist adversarial manipulation, data leakage, and hallucination risk. Privacy-preserving multimodal retrieval, verifiable generation, and trust-calibration mechanisms are essential to ensure reliability and regulatory compliance.

7 Conclusion

This survey provides a systematic overview of multimodal RAG for document understanding. We analyze methodological advances across retrieval modalities, domain settings, retrieval granularity, and the incorporation of graph-based and agent-oriented architectures, highlighting how these developments enhance understanding over visually rich documents. We further consolidate key datasets, benchmarks, and applications in finance, scientific literature, and social analysis, illustrating the broad impact of multimodal RAG. Despite these advances, challenges remain in efficiency, fine-grained multimodal representation, and robustness in real-world deployment. Addressing these issues will be crucial for future advancement, and we hope this work provides a foundation for advancing multimodal RAG toward reliable and generalizable document AI.

Limitations

Although this survey aims to provide a comprehensive synthesis of multimodal RAG for document understanding, several limitations remain. First, while we highlight practical applications, our analysis of real-world deployment challenges such as industry adoption, user-centered evaluation, system integration, and deployment scalability remains preliminary. Broader socio-technical aspects of multimodal RAG systems deserve further exploration in future work. Second, although we summarize major datasets and benchmarks, a more systematic investigation into data quality, annotation consistency, inter-domain transferability, and evaluation alignment across modalities would provide deeper insights into their generalizability and real-world relevance. Furthermore, as multimodal RAG for document understanding is an emerging and rapidly evolving field, newly released datasets, models, and evaluation protocols continue to reshape the landscape. To address this dynamic nature, this survey will be periodically updated and complemented by an open repository to track ongoing progress and facilitate community collaboration.

Ethics Statement

Our work is a survey of existing literature and does not introduce new models, algorithms, or datasets. Therefore, the survey itself does not create new risks. However, we acknowledge that the technologies we review, *i.e.,* multimodal RAG for document understanding, have some potential risks: 1) bias and discrimination inherited from the training data, and 2) the generation of misinformation due to model hallucination. We highlight that addressing these ethical challenges is a critical direction for future research.

The Use of AI assistants. AI assistants (ChatGPT) are used to correct potential grammatical inaccuracies in the manuscript. AI assistants do not participate in research ideation.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. Multi-level information retrieval augmented generation for knowledge-based visual question answering.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, pages 65–72.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Rodrigo Tripodi Calumby, Iago Breno Alves do Carmo Araujo, Felipe Souza Cordeiro, Fabiana Bertoni, Sérgio D Canuto, Fabiano Belém, Marcos André Gonçalves, Ícaro C Dourado, Javier AV Munoz, Lin Li, and 1 others. 2017. Rank fusion and multimodal per-topic adaptiveness for diverse image retrieval. In *MediaEval*.

Sungguk Cha, DongWook Kim, Taeseung Hahn, Mintae Kim, Youngsub Han, and Byoung-Ki Jeon. 2025. Generalized reinforcement learning for retriever-specific query rewriter with unstructured real-world documents. *arXiv preprint arXiv:2507.23242*.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Ryan Rossi, Jiuxiang Gu, and Changyou Chen. 2024a. Mmr: Evaluating reading ability of large multimodal models. *arXiv preprint arXiv:2408.14594*.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A Rossi, Changyou Chen, and Tong Sun. 2024b. Sv-rag: Lora-contextualizing adaptation of mllms for long document understanding. *arXiv preprint arXiv:2411.01106*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024c. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Wang Chen, Guanqiang Qi, Weikang Li, and Yang Li. 2025. Cmrag: Co-modality-based document retrieval and visual question answering. *arXiv preprint arXiv:2509.02123*.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, and 1 others. 2025. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024a. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C Park. 2024b. Typos that broke the rag's back: Genetic attack on rag pipeline by simulating documents in the wild via low-level perturbations. *arXiv preprint arXiv:2404.13948*.

Yejin Choi, Jaewoo Park, Janghan Yoon, Saejin Kim, Jaehyun Jeon, and Youngjae Yu. 2025. Zero-shot multimodal document retrieval via cross-modal question generation. *arXiv preprint arXiv:2508.17079*.

Peter Christen, David J Hand, and Nishadi Kirielle. 2023. A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24.

Kenneth Ward Church, Jiameng Sun, Richard Yue, Peter Vickers, Walid Saba, and Raman Chandrasekar. 2024. Emerging trends: a gentle introduction to rag. *Natural Language Engineering*, 30(4):870–881.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and 1 others. 2024. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. Fquad: French question answering dataset. *arXiv preprint arXiv:2002.06071*.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2024. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*.

Yihao Ding, Soyeon Caren Han, Yan Li, and Josiah Poon. 2025a. Vrd-iu: Lessons from visually rich document intelligence and understanding. *arXiv preprint arXiv:2506.01388*.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. 2025b. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *arXiv preprint arXiv:2507.09861*.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, and 1 others. 2025. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4026–4037.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Ozan Gokdemir, Carlo Siebenschuh, Alexander Brace, Azton Wells, Brian Hsu, Kyle Hippe, Priyanka Setty, Aswathy Ajith, J Gregory Pauloski, Varuni Sastry, and 1 others. 2025. Hiperrag: High-performance retrieval augmented generation for scientific insights. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–13.

Chinmay Gondhalekar, Urjitkumar Patel, and Fang-Chun Yeh. 2025. Multifinrag: An optimized multimodal retrieval-augmented generation (rag) framework for financial question answering. *arXiv preprint arXiv:2506.20821*.

Ziyu Gong, Yihua Huang, and Chengcheng Mai. 2025. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering with hierarchical index and multi-granularity retrieval. *arXiv preprint arXiv:2508.00579*.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and 1 others. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*.

Yucheng Hu and Yuxing Lu. 2024. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. 2025. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *Advances in Neural Information Processing Systems*, 37:67200–67217.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Zhenwen Shao, Qingyun Wu, Huazheng Wang, and 1 others. 2025. Simpledoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. *arXiv preprint arXiv:2506.14035*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. 2024. Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint arXiv:2411.14110*.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, and 1 others. 2024b. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL Workshop*, pages 74–81.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. *arXiv preprint arXiv:2504.12330*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Eric López, Artemis Llabrés, and Ernest Valveny. 2025. Enhancing document vqa models via retrieval-augmented generation. *arXiv preprint arXiv:2508.18984*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and 1 others. 2025. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. *arXiv preprint arXiv:2506.04997*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Kalyan Nandi and S Siva Sathya. 2024. Visual document understanding: A comparative review of modern methods. In *International Conference on Computer Vision and Image Processing*, pages 411–427. Springer.

Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, and 1 others. 2025. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*.

Fatemeh Nazary, Yashar Deldjoo, and Tommaso di Noia. 2025. Poison-rag: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. In *European Conference on Information Retrieval*, pages 239–251. Springer.

Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal learned sparse retrieval with probabilistic expansion control. *Preprint*, arXiv:2402.17535.

Thong Nguyen, Yibin Lei, Jia-Huei Ju, and Andrew Yates. 2025. Serval: Surprisingly effective zero-shot visual document retrieval powered by large vision and language models. *arXiv preprint arXiv:2509.15432*.

George Papageorgiou, Vangelis Sarlis, Manolis Maragoudakis, and Christos Tjortjis. 2025. A multimodal framework embedding retrieval-augmented generation with mllms for eurobarometer data. *AI*, 6(3):50.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Florian Schneider, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, and Chris Biemann. 2025. Collex–a multimodal agentic rag system enabling interactive exploration of scientific collections. *arXiv preprint arXiv:2504.07643*.

Peter H Sellers. 1980. The theory and computation of evolutionary distances: pattern recognition. *Journal of algorithms*, 1(4):359–373.

Ezzeldin Shereen, Dan Ristea, Shae McFadden, Burak Hasircioglu, Vasilios Mavroudis, and Chris Hicks. 2025. One pic is all it takes: Poisoning visual document retrieval augmented generation with a single image. *arXiv preprint arXiv:2504.02132*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text

recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. 2025. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *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)*, pages 6088–6109.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, and 1 others. 2025. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. *arXiv preprint arXiv:2506.02544*.

Vishesh Tripathi, Tanmay Odapally, Indraneel Das, Uday Allu, and Biddwan Ahmed. 2025. Vision-guided chunking is all you need: Enhancing rag with multimodal document understanding. *arXiv preprint arXiv:2506.16035*.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Kesen Wang, Daulet Toibazar, Abdulrahman Alfulayt, Abdulaziz S Albadawi, Ranya A Alkahtani, Asma A Ibrahim, Haneen A Alhomoud, Sherif Mohamed, and Pedro J Moreno. 2025a. Multi-agent interactive question generation framework for long document understanding. *arXiv preprint arXiv:2507.20145*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025b. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025c. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Jingfei Wu, Chaoyuan Shen, Qiyan Deng, Yuping Wang, Jiajun Li, Yuhao Deng, and Minghe Yu. Tabagent: A multi-agent table extraction framework for unstructured documents. *Proceedings of the VLDB Endowment. ISSN*, 2150:8097.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. 2025. Molorag: Bootstrapping document understanding via multi-modal logic-aware retrieval. *arXiv preprint arXiv:2509.07666*.

Yikuan Xia, Jiazun Chen, Yirui Zhan, Suifeng Zhao, Weipeng Jiang, Chaorui Zhang, Wei Han, Bo Bai, and Jun Gao. 2025. Db3 team's solution for meta kdd cup'25. *Preprint*, arXiv:2509.09681.

Xun Xian, Tong Wang, Liwen You, and Yanjun Qi. 2024. Understanding data poisoning attacks for rag: Insights and algorithms.

Zhiyou Xiao, Qinhan Yu, Binghui Li, Geng Chen, Chong Chen, and Wentao Zhang. 2025a. M2io-r1: An efficient rl-enhanced reasoning framework for multimodal retrieval augmented multimodal generation. *arXiv preprint arXiv:2508.06328*.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. 2025b. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. 2025. Docr1: Evidence page-guided grpo for multi-page document understanding. *arXiv preprint arXiv:2508.07313*.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. 2025a. Mm-r5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*.

Mingjun Xu, Zehui Wang, Hengxing Cai, and Renxin Zhong. 2025b. A multi-granularity retrieval framework for visually-rich documents. *arXiv preprint arXiv:2505.01457*.

Yibo Yan, Guangwei Xu, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. 2025. Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. *arXiv preprint arXiv:2509.23883*.

Fuda Ye, Shuangyin Li, Yongqi Zhang, and Lei Chen. 2024. R^2ag: Incorporating retrieval information into retrieval augmented generation. In *EMNLP (Findings)*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, and 1 others. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Bihui Yu, Gaowei Wu, Zhuoya Yao, Huiyang Shi, Qi Chen, Liping Bu, Linzhuang Sun, and Jingxuan Wei. 2025a. Beyond relevance: Utility-driven retrieval for visual document question answering. In *International Conference on Intelligent Computing*, pages 382–393. Springer.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. 2025b. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. *arXiv preprint arXiv:2508.08589*.

Xinlei Yu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Ruolin Shen, Jiangning Zhang, Xiaobin Hu, Yanwei Fu, and Shuicheng Yan. 2025c. Visual document understanding and question answering: A multi-agent collaboration framework with test-time scaling. *arXiv preprint arXiv:2508.03404*.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. 2025. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *arXiv preprint arXiv:2508.05318*.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024a. Cream: coarse-to-fine retrieval and multi-modal efficient tuning for document vqa. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 925–934.

Shuo Zhang, Biao Yang, Zhang Li, Zhiyin Ma, Yuliang Liu, and Xiang Bai. 2024b. Exploring the capabilities of large multimodal models on dense text. In *International Conference on Document Analysis and Recognition*, pages 281–298. Springer.

Wenchuan Zhang, Jingru Guo, Hengzhe Zhang, Penghao Zhang, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. 2025a. Patho-agenticrag: Towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. *arXiv preprint arXiv:2508.02258*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9274–9285.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and 1 others. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *arXiv preprint arXiv:2505.17471*.

Yinan Zhou, Yuxin Chen, Haokun Lin, Shuyu Yang, Li Zhu, Zhongang Qi, Chen Ma, and Ying Shan. 2024. Doge: Towards versatile visual document grounding and referring. *arXiv preprint arXiv:2411.17125*.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

In the appendix, we first present a more detailed introduction to the datasets and benchmarks, along with evaluations of different methods on these benchmarks (Section A). In Section B, we introduce evaluation metrics for multimodal RAG, distinguishing between retrieval and generation. Furthermore, we describe the most commonly used training loss functions for multimodal RAG and provide interpretations of their roles (Section C). Finally, in Section D, we summarize the key contributions of all methods, providing a concise reference for quickly understanding their core ideas.

A Dataset and Benchmark

In the main body, we provide a systematic introduction to the datasets and benchmarks that are widely used for multimodal RAG in document understanding. For each dataset or benchmark, we include a more detailed description, as summarized in Table 4, which lists the data sources and key characteristics. For instance, DocVQA (Mathew et al., 2021) is derived from the UCSF Industry Collections, InfoVQA (Mathew et al., 2022) originates from diverse infographics, and TAT-DQA (Zhu et al., 2021) is constructed from financial reports containing semi-structured tables and text.

In addition, we compile the evaluation results of various multimodal RAG methods on widely used benchmarks, including DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), SlideVQA (Tanaka et al., 2023), and MMLongBench-Doc (Ma et al., 2024b), as presented in Table 3. These results provide a clear comparison of the strengths and weaknesses of different approaches. The evaluation of multimodal RAG performance typically falls into two categories: retrieval and generation, which are presented in the upper and lower parts of Table 3, respectively. Retrieval evaluation focuses on the accuracy of the retrieved pages, whereas generation evaluation measures the correctness of model outputs when the retrieved pages are combined with the user query as input. Since different methods adopt slightly different metrics, we annotate these variations in the table, while aligning comparable metrics to facilitate direct comparison. Detailed explanations of these metrics are provided in Appendix Section B.

B Evaluation Metrics

The evaluation of multimodal RAG methods typically involves two aspects: retrieval evaluation and generation evaluation. Retrieval primarily measures the system's ability to accurately retrieve relevant multimodal information from a large corpus. Generation, on the other hand, evaluates the quality of the model's produced outputs conditioned on the retrieved context. We list the most commonly used metrics along with some newly designed ones that address the limitations in the following.

B.1 Retrieval Evaluation

In the context of multimodal RAG, a variety of metrics are commonly employed to evaluate the performance of the retriever module. Popular measures include Accuracy, Recall, Precision, F1-Score (Christen et al., 2023), Mean Reciprocal Rank (MRR) (Adjali et al.; Nguyen et al., 2024), and Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002).

A widely used measure is Top-K Accuracy, which reflects the hit rate of retrieval.

$$\text{Top-}K \text{ Accuracy} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\Big(\mathrm{Rel}(q) \cap \mathrm{Ret}_K(q) \neq \varnothing\Big), \tag{5}$$

where, for a given query q, $\mathrm{Ret}_K(q)$ denotes the set of the top-K results returned by the retrieval system, $\mathrm{Rel}(q)$ denotes the set of all ground-truth relevant documents or modality segments, and Q denotes the collection of queries. The same symbols appearing in the following formulas carry the same meanings.

Recall@K is usually used to quantify retrieval coverage, measuring how many of the ground-truth relevant items are captured within the top K results:

$$\mathrm{Recall@}K = \frac{1}{|Q|} \sum_{q \in Q} \frac{|\mathrm{Rel}(q) \cap \mathrm{Ret}_K(q)|}{|\mathrm{Rel}(q)|}. \tag{6}$$

Precision@K instead measures accuracy, i.e., the proportion of retrieved items among the top K that are relevant:

$$\mathrm{Precision@}K = \frac{1}{|Q|} \sum_{q \in Q} \frac{|\mathrm{Rel}(q) \cap \mathrm{Ret}_K(q)|}{K}. \tag{7}$$

The F1-Score is often adopted as the harmonic mean of Precision@K and Recall@K, widely used to assess the correctness of retrieved entities or factual fragments in both the retrieval module and the generation process (Li et al., 2024b):

$$\mathrm{F1@}K = \frac{1}{|Q|} \sum_{q \in Q} 2 \cdot \frac{\mathrm{Pr}_K(q) \cdot \mathrm{Re}_K(q)}{\mathrm{Pr}_K(q) + \mathrm{Re}_K(q)}, \tag{8}$$

Method	Metric	DocVQA	SlideVQA	InfoVQA	MMLongBench-Doc
Retrieval Evaluation					
DSE (Ma et al., 2024a)	R@10	–	84.6	–	–
VisRAG (Yu et al., 2024)	R@10	91.20	97.39	97.08	–
CMRAG (Chen et al., 2025)	R@10	–	–	–	64.12
VisRAG (Yu et al., 2024)	MRR@10	75.37	91.85	86.37	–
CMRAG (Chen et al., 2025)	MRR@10	–	–	–	47.64
ColPali (Faysse et al., 2024)	nDCG@5	54.4	–	81.8	–
ColQwen2 (Faysse et al., 2024)	nDCG@5	61.5	–	89.7	–
ColQwen2.5 (Faysse et al., 2024)	nDCG@5	63.6	–	92.5	–
VDocRAG (Tanaka et al., 2025)	nDCG@5	–	77.3	72.9	–
Light-ColPali (Ma et al., 2025)	nDCG@5	53.4	91.7	82.8	73.3
Light-ColQwen2 (Ma et al., 2025)	nDCG@5	56.6	92.9	89.5	77.0
DSE (Ma et al., 2024a)	nDCG@10	–	75.3	–	–
CMRAG (Chen et al., 2025)	nDCG@10	–	–	–	52.10
Generation Evaluation					
VisRAG (Yu et al., 2024)	EM	67.17	60.97	66.43	–
FRAG (Huang et al., 2025)	EM	–	72.7	–	–
SV-RAG (Chen et al., 2024b)	PNLS	87.0	98.8	–	84.8
CRAEM (Zhang et al., 2024a)	ANLS	79.4	–	53.6	–
M3DocRAG (Cho et al., 2024a)	ANLS	84.4	–	–	–
VisDoMRAG (Suri et al., 2025)	ANLS	–	67.2	–	–
VDocRAG (Tanaka et al., 2025)	ANLS	–	56.4	64.6	–
FRAG (Huang et al., 2025)	ANLS	87.4	–	–	–
ReDocRAG (López et al., 2025)	ANLS	73.7	–	63.6	–
M3DocRAG (Cho et al., 2024a)	G-Acc	–	–	–	21.0
FRAG (Huang et al., 2025)	G-Acc	80.5	–	–	37.9
VRAG-RL (Wang et al., 2025c)	G-Acc	–	69.3	–	24.9
SimpleDoc (Jain et al., 2025)	G-Acc	–	–	–	60.58
MMRAG-DocQA (Gong et al., 2025)	G-Acc	–	–	–	52.3
CMRAG (Chen et al., 2025)	G-Acc	–	–	–	43.25
MoLoRAG (Wu et al., 2025)	G-Acc	–	–	–	41.01

Table 3: Evaluation results of RAG methods. The upper block shows **retrieval evaluation** and the lower block shows **generation evaluation**. Different background shades are used to separate the two parts.

where, $\mathrm{Pr}_K(q)$ represents Precision@K, and $\mathrm{Re}_K(q)$ represents Recall@K.

However, the metrics above are insensitive to the ranking order within the top K. In practice, placing highly relevant or informative items at earlier positions is crucial for effective RAG. Adjali et al.; Nguyen et al. (2024) utilize MRR@K to emphasize the position of the first relevant item:

$$\mathrm{MRR@}K = \frac{1}{|Q|} \sum_{q \in Q} \frac{\mathbf{1}\big(\mathrm{rank}_K(q) \leq K\big)}{\mathrm{rank}_K(q)}, \quad (9)$$

where $\mathrm{rank}_K(q)$ denotes the position of the first relevant document within the top-K retrieved results for query q; if no relevant item appears within the top K, the reciprocal rank is set to 0.

Similarly, Zhao et al. (2025); Faysse et al. (2024) employ nDCG@K to penalize relevant items that appear lower in the ranking, thereby rewarding systems that surface high-quality evidence earlier:

$$\mathrm{nDCG@}K = \frac{1}{|Q|} \sum_{q \in Q} \frac{\mathrm{DCG@}K(q)}{\mathrm{IDCG@}K(q)}, \quad (10)$$

where

$$\mathrm{DCG@}K(q) = \sum_{i=1}^{K} \frac{2^{\mathrm{rel}_{q,i}} - 1}{\log_2(i+1)}. \quad (11)$$

Here, $\mathrm{rel}_{q,i}$ represents the graded relevance of the i-th retrieved item for query q. The denominator $\mathrm{IDCG@}K(q)$, called the *ideal DCG*, represents the maximum possible DCG that could be achieved for query q if all relevant items were perfectly ranked at the top of the list.

B.2 Generation Evaluation

In the context of Multimodal RAG, the primary objective of generation quality evaluation is to as-

sess the quality and consistency between model-generated text and reference answers. This involves not only measuring the correctness of the responses but also considering aspects such as fluency, information coverage, and logical coherence. To achieve a comprehensive evaluation, this study examines a wide range of metrics. The earliest are soft matching metrics (*e.g.*, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005)), which rely on n-gram overlap for a soft lexical evaluation that allows partial and flexible matching. They mainly assess fluency and information coverage of generated text. With the rise of question answering and reading comprehension tasks, strict matching metrics (*e.g.*, Exact Match (Rajpurkar et al., 2016), ANLS (Biten et al., 2019), PNLS (Chen et al., 2024a)) are introduced, focusing on exact or near-exact correspondence with reference answers to measure form-level correctness. More recently, driven by the advances in pretrained language models, semantic matching metrics (*e.g.*, BERTScore (Devlin et al., 2019), RoBERTa (Liu et al., 2019), G-Acc (Ma et al., 2024b)) have become prominent, enabling the assessment of deeper semantic consistency through contextual embeddings. By combining these three categories of metrics, generation quality can be evaluated holistically across surface, exact matching, and semantic alignment.

Soft Matching Metrics. The earliest approaches to generation quality evaluation adopt soft matching metrics, which rely on n-gram overlap to provide a soft lexical evaluation that tolerates partial and flexible matching between generated and reference texts. Among them, BLEU (Papineni et al., 2002) is one of the most representative metrics. BLEU evaluates the similarity between generated text and reference text based on n-gram overlap with a brevity penalty (BP). The BLEU score is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad (12)$$

where p_n is the precision for n-grams and w_n is the weight assigned to each n-gram order. The brevity penalty (BP) is given by:

$$\text{BP} = \exp\left(\min\left(0,\ 1 - \frac{r}{c}\right)\right), \quad (13)$$

where c is the candidate (generated) length and r is the reference length.

Compared to BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) evaluates the overlap between generated and reference texts at the n-gram level, and is widely used in summarization tasks. The ROUGE-N score is defined as:

$$\text{ROUGE-}N = \frac{\sum_{\text{ref}} \sum_{n \in \text{ref}} \min\left(S_n, R_n\right)}{\sum_{\text{ref}} \sum_{n \in \text{ref}} R_n}, \quad (14)$$

where S_n and R_n denote the counts of a given n-gram in the system output and reference, respectively. ROUGE-L leverages the Longest Common Subsequence (LCS) between the system output and the reference to capture sentence-level structural similarity. Its recall-oriented form is given by:

$$\text{ROUGE-L} = \frac{\text{LCS}(S, R)}{|R|}, \quad (15)$$

where LCS(S, R) denotes the length of the longest common subsequence between the system output S and the reference R, and |R| is the length of the reference.

Compared to BLEU and ROUGE, METEOR (Banerjee and Lavie, 2005) emphasizes semantic matching beyond exact n-gram overlap. It incorporates stemming, synonym matching, and a penalty for word order differences to better capture the similarity between system outputs and references. The METEOR score is defined as:

$$\text{METEOR} = F_\alpha \cdot (1 - P), \quad (16)$$

where F_α is a weighted harmonic mean of precision (P_{pre}) and recall (P_{rec}), given by:

$$F_\alpha = \frac{P_{rec} \cdot P_{pre}}{\alpha \cdot P_{pre} + (1 - \alpha) \cdot P_{rec}}, \quad (17)$$

and P is a fragmentation penalty based on word order:

$$P = \gamma \left(\frac{ch}{m}\right)^{\beta}, \quad (18)$$

where ch denotes the number of chunks (i.e., contiguous matched word sequences), m is the total number of matched words, and α, β, γ are tunable parameters.

Strict Matching Metrics. In contrast to soft matching metrics, strict matching metrics emphasize exact or near-exact correspondence between generated and reference answers. They assess the consistency and form-level correctness of model outputs, directly reflecting the factual accuracy of the generated responses.

The most representative metric in this category is Exact Match (EM) (Rajpurkar et al., 2016), which computes the percentage of predictions that exactly match one of the reference answers:

$$\mathrm{EM} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(o_i \in A_i), \tag{19}$$

where o_i denotes the predicted answer, A_i is the set of groundtruth answers, and $\mathbf{1}(\cdot)$ is the indicator function.

With the advancement of generative models and their increasing generalization capabilities, more recent metrics have been introduced. Average Normalized Levenshtein Similarity(ANLS) (Biten et al., 2019) is designed to provide a soft evaluation of string-based answers. ANLS is defined as below:

$$\mathrm{NLS}(a_{ij}, o_i) = 1 - \frac{\mathrm{LD}(a_{ij}, o_i)}{\max(|a_{ij}|, |o_i|)}, \tag{20}$$

where o_i is a given prediction, a_{ij} is a groundtruth answer, $\mathrm{LD}(a_{ij}, o_i)$ denotes the standard Levenshtein edit distance (Lcvenshtcin, 1966), and $|\cdot|$ is the string length. The threshold τ controls the minimum similarity required for a prediction to be considered correct.

$$s(a_{ij}, o_i) = \begin{cases} \mathrm{NLS}(a_{ij}, o_i), & \text{if } \mathrm{NLS}(a_{ij}, o_i) \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

$$\mathrm{ANLS} = \frac{1}{N}\sum_{i=1}^{N} \max_j s(a_{ij}, o_i). \tag{22}$$

Moreover, AccANLS (Zhang et al., 2024b) integrates accuracy with ANLS similarity, aiming at addressing the issue of penalizing redundant outputs. Partial Normalized Levenshtein Similarity (PNLS) (Chen et al., 2024a) generalizes ANLS by relaxing the alignment requirement: instead of computing edit distance over the entire strings, it identifies the best-matching substring of the prediction relative to the reference. This design avoids penalizing extra prefixes or suffixes while still accounting for mismatches, insertions, and deletions within the aligned region, making it more suitable for evaluating verbose LLM outputs. Formally, PNLS still follows the NLS formulation but replaces the standard edit distance with a *partial edit distance* $\mathrm{LD}^*(a_{ij}, o_i)$ obtained via approximate string matching (Sellers, 1980). The final score is computed as:

$$\mathrm{PNLS}(a_{ij}, o_i) = 1 - \frac{\mathrm{LD}^*(a_{ij}, o_i)}{\max(|a_{ij}|, |o_i'|)}, \tag{23}$$

where o_i' denotes the optimally aligned substring of the prediction o_i.

Semantic Matching Metrics. Beyond soft and strict matching metrics, semantic matching metrics have emerged to evaluate deeper semantic consistency between generated and reference texts. Metrics such as BERTScore, which leverages contextual embeddings from pretrained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), move beyond simple lexical overlap by capturing semantic similarity between generated and reference texts. This enables a more reliable evaluation of whether the meaning of a response is preserved, even when different phrasings are used. However, while BERTScore provides strong advantages in measuring semantic consistency, it is less suited for scenarios involving long-form, explanatory, or unanswerable responses. To address this gap, Generated Accuracy (G-Acc) (Ma et al., 2024b) has been proposed, which extends evaluation to free-form answers that emphasize reasoning, elaboration, and contextual completeness, thereby offering a more comprehensive assessment of generation quality.

C Training Loss

In multimodal RAG, the most common training objective is a ColBERT-style (Khattab and Zaharia, 2020; Faysse et al., 2024) contrastive loss. The key idea is to represent both queries and documents with multiple contextualized token embeddings and compute their similarity through a *late interaction mechanism*. Formally, given a query q and a document d, we represent them as $\mathbf{H}_q \in \mathbb{R}^{L_q \times D}$ and $\mathbf{H}_d \in \mathbb{R}^{L_d \times D}$, where L_q and L_d denote the number of tokens in the query and document, and D is the embedding dimension. The late interaction similarity is defined as:

$$\mathrm{Sim}(q, d) = \sum_{t=1}^{L_q} \max_{1 \leq m \leq L_d} \langle \mathbf{h}_{q,t}, \mathbf{h}_{d,m} \rangle, \tag{24}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. This operator allows each query token to attend to its most relevant document token, enabling fine-grained matching.

During training, a contrastive objective (Khosla et al., 2020; Wang and Liu, 2021) is optimized over a batch of query–document pairs $\{(x_i, y_i)\}_{i=1}^{B}$. For each query x_i, the paired document y_i is the positive example, while the remaining documents in the batch act as negatives. Let $p_i = \mathrm{Sim}(x_i, y_i)$ and $n_i = \max_{j \neq i} \mathrm{Sim}(x_i, y_j)$ denote the positive and hardest negative similarities, respectively. The loss is defined as:

$$\begin{aligned} \mathcal{L} &= -\frac{1}{B} \sum_{i=1}^{B} \log \left(\frac{\exp(p_i)}{\exp(p_i) + \exp(n_i)} \right) \\ &= \frac{1}{B} \sum_{i=1}^{B} \log \left(1 + \exp(n_i - p_i) \right), \end{aligned} \tag{25}$$

which encourages higher similarity for the positive pair than for any in-batch negative.

This ColBERT-style loss, combining late interaction with contrastive learning, is widely adopted in multimodal RAG systems as it provides effective supervision for aligning queries and documents across both text and vision modalities.

D Key Contribution Summary

Table 5 presents a consolidated overview of the key contributions of existing multimodal RAG approaches for document understanding. By systematically organizing and comparing these methods, this survey highlights the breadth of design choices and research directions in the field. Such a structured summary not only helps researchers quickly grasp the state of the art, but also clarifies common trends, complementary strengths, and open challenges. In doing so, it serves as a reference point for guiding future work and motivating new directions in multimodal retrieval and reasoning for complex document understanding.

Dataset	Features
PlotQA (Methani et al., 2020)	Bridges the gap to real-world plots with a large-scale dataset built from authentic charts and crowd-sourced questions, requiring complex reasoning and out-of-vocabulary answers beyond fixed vocabularies.
TabFQuAD (d'Hoffschmidt et al., 2020)	Evaluates TableQA models in realistic industry settings using a French table question-answering dataset enhanced with GPT-4V generated queries.
DocVQA (Mathew et al., 2021)	Highlights the gap between human and model performance on structured document understanding using a large-scale dataset from UCSF Industry collections.
VisualMRC (Tanaka et al., 2021)	Builds a visual machine reading comprehension dataset from multi-domain webpage documents to advance natural language understanding and generation from document images.
ChartQA (Masry et al., 2022)	Constructs a large-scale chart QA benchmark with human-written and generated questions to evaluate models on complex logical, arithmetic, and visual reasoning over charts.
InfoVQA (Mathew et al., 2022)	Benchmarks models on reasoning over layout, text, and visuals using a diverse infographic QA dataset highlighting the human–machine gap.
TAT-DQA (Zhu et al., 2022)	Samples financial reports with semi-structured tables and text to build a document QA dataset requiring discrete numerical reasoning, highlighting the gap between models and human experts.
ScienceQA (Saikh et al., 2022)	Introduces a multimodal benchmark of diverse science questions with annotated answers, lectures, and explanations to evaluate and enhance models' reasoning through chain-of-thought.
DUDE (Van Landeghem et al., 2023)	Creates a practical benchmark from multi-industry, multi-domain visually-rich documents to evaluate document AI on real-world, multi-task, and low-resource scenarios.
SlideVQA (Tanaka et al., 2023)	Builds a multi-image document QA dataset from slide decks to enable complex single-hop, multi-hop, and numerical reasoning, highlighting the gap between models and human performance.
ArXivQA (Li et al., 2024a)	Builds a scientific QA dataset from ArXiv papers to boost LVLMs' ability in interpreting abstract figures and improving mathematical reasoning.
MMLongBench-Doc (Ma et al., 2024b)	Constructs a long-context multimodal benchmark from lengthy PDFs with cross-page questions to evaluate LVLMs on document understanding.
PaperTab (Hui et al., 2024)	Extracts academic papers in PDF format for extractive, yes/no, and free-form QA.
FetaTab (Hui et al., 2024)	Gathers world knowledge documents in PDF and HTML format for free-form QA.
SPIQA (Pramanick et al., 2024)	Creates a large-scale QA dataset from scientific papers that integrates text with complex figures and tables to evaluate and advance multimodal understanding in research articles.
LongDocUrl (Deng et al., 2024)	Integrates long-document understanding, numerical reasoning, and cross-element locating into a large-scale benchmark to expose critical gaps in current LVLMs.
ViDoRe (Faysse et al., 2024)	Unifies academic tasks with diverse document types and practical tasks across multiple domains and languages to comprehensively evaluate multimodal document retrieval.
VisR-Bench (Chen et al., 2024b)	Selects diverse visually-rich documents with tables, charts, and diagrams, and generate verified QA pairs using GPT-4o to create a benchmark highlighting multimodal reasoning and quality assurance.
M3DoCVQA (Cho et al., 2024a)	Evaluates open-domain DocVQA with M3DoCVQA, a large multi-page PDF benchmark requiring multi-hop, multimodal reasoning across text and visual elements.
VisDoMBench (Suri et al., 2025)	Leverages multiple documents with diverse modalities such as tables, charts, and slides, requiring cross-document reasoning, modality fusion, and verifiable answers.
ViDoSeek (Wang et al., 2025b)	Unifies queries and large corpora of visually rich documents to enable complex reasoning beyond image-based QA, emphasizing multimodal retrieval, cross-document comprehension, and unique answer generation.
OpenDocVQA (Tanaka et al., 2025)	Combines diverse document types, formats, and modalities into a unified open-domain collection to train and evaluate retrieval and QA models on visually-rich documents.

Table 4: Popular datasets and benchmarks in multimodal RAG for document understanding, along with detailed descriptions of their data sources and characteristics.

Method	Key Contribution Summary
DSE (Ma et al., 2024a)	Encodes document screenshots with VLMs for retrieval, avoiding parsing and preserving full multimodal information.
ColPali (Faysse et al., 2024)	Embeds document page images into multi-vector representations with late interaction matching for efficient end-to-end retrieval.
ColQwen2 (Faysse et al., 2024)	Extends Qwen2-VL-2B to generate ColBERT-style multi-vector representations for complex text–image tasks, similar to ColPali.
CRAEM (Zhang et al., 2024a)	Combines coarse-to-fine retrieval with multi-page visual attention pooling, enabling effective integration of multimodal document information.
VisRAG (Yu et al., 2024)	Introduces a VLM-based RAG pipeline that embeds documents as images, preserving multimodal information and avoiding text-parsing loss.
SV-RAG (Chen et al., 2024b)	Introduces a framework where MLLMs act as retriever and generator with two adapters for retrieval and question answering.
M3DocRAG (Cho et al., 2024a)	Unifies retrieval and reasoning across text, charts, and figures, enabling flexible multi-hop DocVQA over single or multi-page documents.
VisDoMRAG (Suri et al., 2025)	Introduces consistency-constrained modality fusion for unified multi-step reasoning across visual and textual modalities in multimodal document QA.
GME (Zhang et al., 2025b)	Advances universal multimodal retrieval by leveraging a synthetic fused-modal training dataset and an MLLM-based dense retriever, achieving state-of-the-art performance on the new UMR Benchmark.
ViDoRAG (Wang et al., 2025b)	Leverages a multi-agent, Gaussian Mixture Model-based hybrid retrieval and iterative reasoning workflow for complex understanding of visually rich documents.
HM-RAG (Liu et al., 2025)	Decomposes queries hierarchically, retrieves from diverse modalities, and integrates results via consistency voting for robust multimodal reasoning.
VDocRAG (Tanaka et al., 2025)	Unifies visually-rich documents into image-based representations and design self-supervised pre-training tasks that compress visual information into dense tokens aligned with textual content for retrieval-augmented generation.
FRAG (Huang et al., 2025)	Selects relevant frames to improve multimodal model generation efficiency and performance.
MG-RAG (Xu et al., 2025b)	Integrates hierarchical encoding, modality-aware retrieval, and VLM-based candidate filtering to effectively handle visually-rich documents.
VRAG-RL (Wang et al., 2025c)	Introduces an RL framework that enables VLMs to reason effectively over documents from pages to fine-grained regions.
CoRe-MMRAG (Tian et al., 2025)	Reconciles inconsistencies between parametric and retrieved multimodal knowledge through a four-stage framework with specialized training for reliable answer generation.
Light-ColPali (Ma et al., 2025)	Reduces memory usage in Visualized Document Retrieval by applying optimized token merging, preserving over 94% effectiveness with as little as 2.8% of the original memory.
MM-R5 (Xu et al., 2025a)	Enhances multimodal document retrieval by integrating supervised fine-tuning and reinforcement learning with reasoning chains and task-specific rewards.
SimpleDoc (Jain et al., 2025)	Combines embedding-based retrieval with summary-based re-ranking, enabling efficient multi-page reasoning with a single VLM agent.
VisChunk (Tripathi et al., 2025)	Leverages multimodal cues to chunk documents while preserving structural and semantic coherence, enhancing downstream RAG performance.
DocVQA-RAP (Yu et al., 2025a)	Proposes a utility-driven retrieval method for VDQA that scores evidence by its predicted contribution to answer quality, reducing reliance on mere semantic relevance.
RL-QR (Cha et al., 2025)	Applies reinforcement learning–based query rewriting without annotations, tailoring rewriters to specific retrievers and boosting RAG performance across text and multimodal databases.
MMRAG-DocQA (Gong et al., 2025)	Leverages hierarchical indexing and multi-granularity retrieval to connect in-page and cross-page multimodal evidence, enabling accurate reasoning over long, modality-rich documents.
Patho-AgenticRAG (Zhang et al., 2025a)	Enables joint text–image retrieval from pathology textbooks with agentic reasoning and multi-turn search, reducing hallucinations and improving diagnostic accuracy.
M2IO-R1 (Xiao et al., 2025a)	Enables multimodal inputs and outputs in RAG with an RL-based framework using an Inserter module for controllable image selection and placement.
mKG-RAG (Yuan et al., 2025)	Enhances RAG-based VQA by constructing multimodal knowledge graphs and employing dual-stage, question-aware retrieval to provide structured, modality-aligned knowledge for more accurate generation.
DB3Team-RAG (Xia et al., 2025)	Integrates domain-specific multimodal retrieval pipelines with unified LLM tuning and refusal training.
PREMIR (Choi et al., 2025)	Boosts multimodal retrieval by generating cross-modal pre-questions, enabling robust token-level matching across domains and languages.
ReDocRAG (López et al., 2025)	Enhances Document VQA by retrieving and reranking key evidence, achieving higher accuracy on multi-page datasets with reduced memory demand.
CMRAG (Chen et al., 2025)	Leverages co-modality representations of text and images for joint retrieval and generation, enabling more effective document visual question answering than text-only or vision-only RAG methods.
MoLoRAG (Wu et al., 2025)	Enhances multi-modal, multi-page DocQA by combining semantic and logic-aware retrieval through page-graph traversal, enabling LVLMs to capture overlooked logical connections for more accurate answers.
SERVAL (Nguyen et al., 2025)	Leverages vision–language models to generate textual descriptions of document images and embed them with a text encoder for scalable zero-shot visual document retrieval.
MetaEmbed (Xiao et al., 2025b)	Employs learnable Meta Tokens to generate compact multi-vector embeddings, enabling scalable test-time trade-offs between retrieval quality and efficiency.
DocPruner (Yan et al., 2025)	Adaptively prunes redundant patch-level embeddings based on intra-document attention, substantially reducing storage costs for multi-vector VDR while preserving retrieval effectiveness.

Table 5: Key contributions of multimodal RAG methods for document understanding.