Roles of MLLMs in Visually Rich Document Retrieval for RAG: A Survey

Xiantao Zhang
Beihang University
zhangxiantao@buaa.edu.cn

Abstract

Visually rich documents (VRDs) challenge retrieval-augmented generation (RAG) with layout-dependent semantics, brittle OCR, and evidence spread across complex figures and structured tables. This survey examines how Multimodal Large Language Models (MLLMs) are being used to make VRD retrieval practical for RAG. We organize the literature into three roles: *Modality-Unifying Captioners*, *Multimodal Embedders*, and *End-to-End Representers*. We compare these roles along retrieval granularity, information fidelity, latency and index size, and compatibility with reranking and grounding. We also outline key trade-offs and offer some practical guidance on when to favor each role. Finally, we identify promising directions for future research, including adaptive retrieval units, model size reduction, and the development of evaluation methods.

1 Introduction

Visually rich documents (VRDs), such as PDFs, scanned pages, slide decks, reports, forms, and infographics, encode meaning through the interplay of text, layout, figures, and graphics. As retrieval-augmented generation (RAG) (Lewis et al., 2020) becomes a default pattern for grounding large language models (LLMs) (Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Nakano et al., 2022), many real-world deployments are moving beyond plain text to these document types (Ma et al., 2024; Faysse et al., 2025; Yu et al., 2025; Suri et al., 2025; Tanaka et al., 2025). This shift strains classical text-only RAG pipelines, motivating the broader development of multimodal RAG (MM-RAG) systems designed to retrieve and reason over varied data types, including images and tables (Chen et al., 2022; Yasunaga et al., 2023).

However, VRDs represent a uniquely difficult case for MM-RAG. Unlike retrieving standalone images or text, VRD retrieval must contend with meaning derived from the *fusion* of layout, embedded text, and graphics. Consequently, traditional preprocessing steps like optical character recognition (OCR) and layout parsing remain brittle and lossy, fine-grained visual cues vanish in textual proxies, and evidence may span multiple pages or views. Recent surveys in document understanding (Ding et al., 2025) echo this shift, underscoring both the opportunity and difficulty of learning over text–layout–vision jointly.

At the same time, a new generation of methods argues for *seeing* pages directly. Document Screenshot Embedding (DSE) (Ma et al., 2024) treats a page screenshot as the unit of indexing, avoiding preprocessing choices that introduce error and latency. Its premise is pragmatic: keep all information available at retrieval time. Likewise, ColPali (Faysse et al., 2025) fuses Vision–Language Models (VLMs) with late-interaction matching, and results show that learning directly over page images can simplify pipelines and improve effectiveness. Beyond retrieval alone, VisRAG (Yu et al., 2025) integrates vision-based retrieval with generation, adopting the page-as-image abstraction end-to-end to mitigate conversion loss during both retrieval and answer synthesis.

VRD-centric evaluation has also matured. Beyond classic DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), and SlideVQA (Tanaka et al., 2023), newer resources now increasingly stress chart reasoning and multi-slide evidence aggregation reflecting practical needs like finding a single number inside a plot or tracing an argument across a deck (Tamber et al., 2025; Yang et al., 2025; Liu et al., 2025; Chen et al., 2025b; Peng et al., 2025). These datasets collectively highlight why retrieval must respect both layout and visual semantics, not only text.

Scope and goal This survey focuses specifically on visually rich document retrieval for RAG. We

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 19–36
December 20-24, 2025 ©2025 Association for Computational Linguistics

analyze how Multimodal Large Language Models (MLLMs) are used to index and retrieve pages, page regions, tables, figures, and slide content for RAG over documents. Our goal is to distill design patterns, compare empirical trends, and surface trade-offs that matter for building reliable, cost-aware systems.

Contributions This survey makes the following contributions:

1. **Role-based taxonomy of VRD–RAG.** We organize how MLLMs enter the pipeline into three roles tailored to documents.

2. **Comparative analysis of key trade-offs.** We contrast these roles in terms of retrieval unit, robustness to OCR and layout errors, latency and indexing cost, and compatibility with reranking and grounding, summarizing evidence from recent VRD-focused work.

3. **Practical takeaways and open challenges.** We discuss when to favor caption-first vs. image-first retrieval, how to balance page-level recall with element-level precision, how to budget compute and storage for multimodal indices, and where evaluation lags behind given the current benchmarks.

Organization §2 reviews background on RAG, multimodal retrieval, and MLLMs. §3 develops the three-role framework and contrasts representative approaches. §4 examines trade-offs and open challenges. §5 concludes with takeaways and future directions.

2 Related Work

2.1 Retrieval-Augmented Generation

RAG combines a retriever and a generator to bridge retrieval-based and generative models. This hybrid approach dynamically retrieves documents to condition generation, enhancing factual accuracy and access to knowledge beyond training data (Shuster et al., 2021; Gao et al., 2024; Lewis et al., 2020; Asai et al., 2024; Shi et al., 2024; Izacard et al., 2023; Borgeaud et al., 2022; Li et al., 2024). Recent research expanded RAG to open-domain QA (Guu et al., 2020; Mao et al., 2021; Siriwardhana et al., 2023), dialogue systems (Thulke et al., 2021; Komeili et al., 2022; Li et al., 2022b), and multimodal tasks (Chen et al., 2022; Yasunaga et al., 2023; Hu et al., 2023; Luo et al., 2024; Ren et al., 2025; Jeong et al., 2025), highlighting its potential for integrating diverse knowledge into NLP pipelines.

2.2 Multimodal Retrieval

Recent studies have demonstrated that multimodal retrieval and RAG significantly enhance LLMs by integrating diverse data modalities, such as text, images, and audio. The seminal work of MuRAG (Chen et al., 2022) inaugurated the era of end-to-end multimodal retrieval-augmented transformers, a pioneering innovation that has since been shown to enhance performance in a range of tasks, including question answering, by leveraging external multimodal memory. In a similar manner, RA-CM3 (Yasunaga et al., 2023) was the first to demonstrate the capabilities of joint retrieval and text and image generation, achieving superior performance compared to models such as DALL-E (Ramesh et al., 2021), while being more efficient. Wei et al. (2024) proposed UniIR, a universal multimodal retrieval model designed to handle a wide range of tasks. Subsequent advancements include GENIUS (Kim et al., 2025), a universal generative framework for multimodal search, and UMaT (Bi and Xu, 2025), which unifies video and audio data via textual representations for long-form question-answering. A comprehensive survey by Zhao et al. (2023) further systematizes these approaches, highlighting improvements in factuality, robustness, and cross-modal reasoning. Collectively, these works emphasize the transformative potential of multimodal RAG in scaling LLM capabilities across domains.

2.3 Multimodal Large Language Models

MLLMs have emerged as a transformative advancement in the field of artificial intelligence, extending the capabilities of LLMs by integrating multiple data modalities, such as text, images, audio, and videos. LLaVA (Liu et al., 2023, 2024a) has been at the forefront of visual instruction tuning, achieving this through the alignment of a vision encoder with a language model via a cross-modal connector. Subsequent developments like the Qwen-VL Series (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025) and the InternVL Series (Chen et al., 2024c,b, 2025d; Zhu et al., 2025; Wang et al., 2025) have demonstrated significant progress in multimodal understanding and reasoning, including specialized alignment techniques for complex domains like mathematical reasoning (Zhuang et al., 2025).

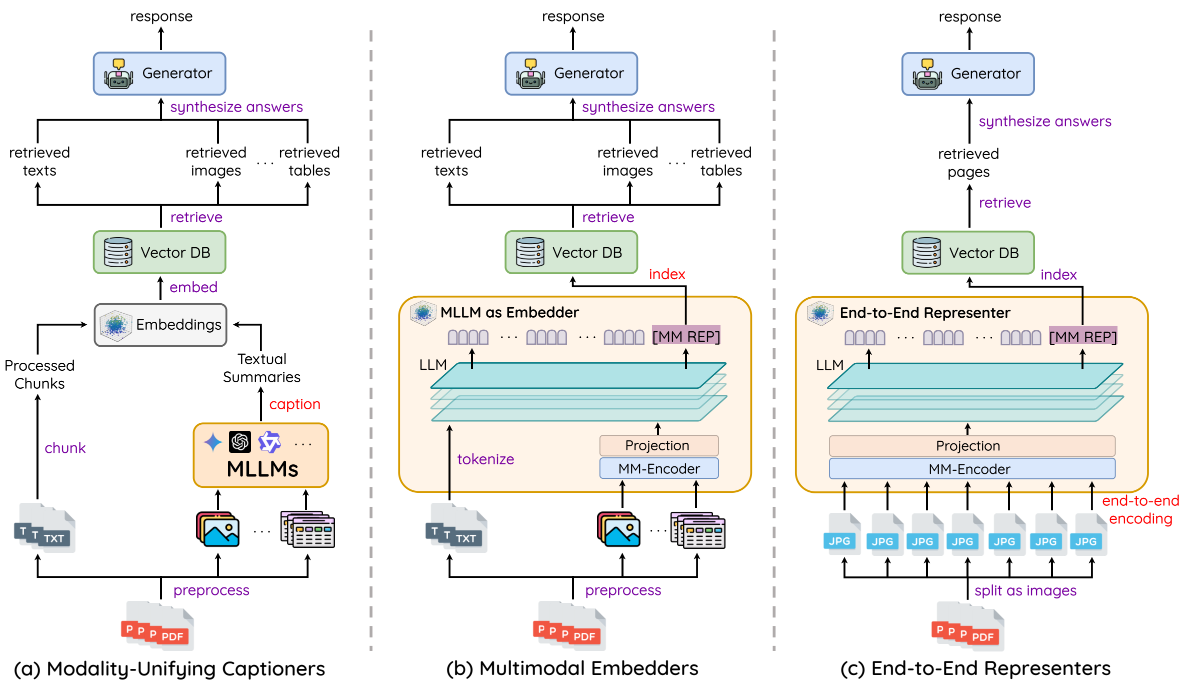

Figure 1: Overview of how MLLMs enter VRD retrieval for RAG across three roles. **Left:** *Modality-Unifying Captioners* (§3.1); **Middle:** *Multimodal Embedders* (§3.2); **Right:** *End-to-End Representers* (§3.3). Each panel sketches the pipeline from document intake to retrieval and answer synthesis, highlighting typical retrieval units and index types.

2.4 Related Surveys

A growing body of surveys maps the RAG landscape from text-only pipelines to fully multimodal systems. Early overviews on RAG (Gao et al., 2024; Fan et al., 2024a) consolidate architectures, training strategies, and evaluation, motivating retrieval as a remedy for hallucinations and stale knowledge. Xu et al. (2025b) survey the evolution of model architectures in information retrieval (IR). For multimodality, Zhao et al. (2023) provide one of the first broad treatments across images, tables, and audio. More recent efforts expand the scope and depth: Abootorabi et al. (2025) organize the full multimodal RAG pipeline together with datasets and training strategies; Mei et al. (2025) synthesize definitions and components with an emphasis on cross-modal alignment; Zheng et al. (2025) bridge RAG with visual understanding and generation and discuss embodied settings; and Gao et al. (2025) review multimodal RAG approaches for document understanding and compile a collection of multimodal RAG datasets. Additionally, Ding et al. (2025) provide a comprehensive overview of deep learning–based VRD. Compared with these works, our survey narrows the focus to visually rich documents and contributes a role-based taxonomy for how MLLMs enter the pipeline while foregrounding practical trade-offs specific to document-centric RAG.

3 Three Roles of MLLMs in VRD RAG

We introduce the *Emergent Large-Scale Paradigm* of multimodal RAG: the systematic use of MLLMs to move beyond text-only pipelines by treating page images, layout, and visual structure as first-class retrieval signals. Rather than a single recipe, this paradigm appears in practice through three complementary roles that MLLMs can play in VRD pipelines: *Modality-Unifying Captioners*, which translate non-text elements into textual surrogates for conventional indexing; *Multimodal Embedders*, which map heterogeneous inputs into a shared representation space for cross-modal search; and *End-to-End Representers*, which encode whole pages directly without explicit OCR or layout parsing. Viewing the literature through these roles provides a concrete basis for analyzing retrieval granularity, information fidelity, and system cost in §4.

3.1 MLLMs as Modality-Unifying Captioners

As sketched in Figure 1 (left), this role converts non-textual elements into textual surrogates for conventional indexing and retrieval. In the *Modality-Unifying Captioner* role, systems translate non-textual inputs into textual surrogates so that retrieval and generation can proceed in the *text* modality. For VRDs, this typically means (i) OCR- and layout-aware textualization of pages and regions, and (ii) higher-level natural-language descriptions that summarize figures, tables, and UI screenshots. The resulting text is embedded with standard text encoders and indexed alongside native document text, enabling drop-in multimodal support for existing text-only RAG stacks.

From captioning to document textualization Early captioners established language as a universal interface for vision. Vinyals et al. (2015) and Xu et al. (2015) demonstrated global and attention-grounded image descriptions; Johnson et al. (2016) introduced region-level captions, inspiring fine-grained retrieval in VRDs, where figure panels or table regions should be independently retrievable. OCR-aware captioning such as TextCaps (Sidorov et al., 2020) explicitly reads in-image text, crucial for charts and slides where on-image text encodes semantics. In VRD pipelines, LayoutLM (Xu et al., 2020, 2021b; Huang et al., 2022) unified OCR tokens with 2D coordinates for forms and invoices, while the DocVQA (Mathew et al., 2021) benchmark standardized OCR-first evaluation. Beyond OCR, Donut (Kim et al., 2022) mapped document images directly to target text to reduce error propagation, and Pix2Struct (Lee et al., 2023) turned UI/web screenshots into simplified HTML, making both approaches practical captioners that emit structured proxies well-suited for text indexing.

Captions as textual proxies The same *text proxy* pattern recurs across modalities and offers lessons for VRDs. In video and audio, Miech et al. (2019) leveraged narration transcripts for supervision; Xu et al. (2021a) aligned video with text via contrastive pretraining; Lei et al. (2018) operationalized subtitle-centric QA with temporal localization. Error cascades in speech-to-text QA were documented by Spoken SQuAD (Li et al., 2018), underscoring the brittleness of ASR-first pipelines. For environmental audio, AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2019) showed that textual captions are effective surrogates for downstream retrieval and clustering. For structured vision, Johnson et al. (2015) treated a scene graph as a structured caption to drive semantic retrieval. In clinical imaging, R2Gen (Chen et al., 2020) cast images into long textual reports, indexable as evidence in text-first RAG.

Practical Deployments Production systems generally follow a consistent approach to captioning. They first introduce an upstream captioning layer that generates page or region summaries, verbalizes tables, and produces figure descriptions. After this conversion stage, a mature text retriever and reader are applied to the resulting text. Practical tutorials by LangChain Team (2023) and Surla et al. (2024) describe this convert-first-then-index workflow for slide and PDF question answering. Comparable industrial deployments and discussions of potential limitations are provided by Riedler and Langer (2024). Evidence from these studies suggests that using stronger captioners consistently improves recall and answer quality, even when the retrieval model remains unchanged.

Video-RAG as a mirror for VRDs The captioning approach can be extended to long videos by treating time-aligned text as the primary index. Recent systems refine this concept by explicitly captioning long videos through the extraction of automatic speech recognition (ASR), optical character recognition (OCR), and object detection outputs, which are converted into textual fragments used as retrievable evidence. Video-RAG (Luo et al., 2024) represents one such example, where ASR, OCR, and detection results are transformed into retrievable text aligned with sampled video frames. SceneRAG (Zeng et al., 2025) incorporates ASR transcripts with timestamps and scene segmentation, together with a scene-level knowledge graph to enable multi-hop retrieval. VideoAgent (Fan et al., 2024b) further unifies these modalities into a memory structure that combines two-second subtitle segments with tables describing object states. Subsequent research, including works by Ren et al. (2025) and Jeong et al. (2025), extends these ideas to very long videos by advocating dual-channel architectures that preserve both textual proxies and visual context. This approach mirrors the best practices established in VRD tasks, where captions are paired with region crops to improve reranking and grounding. Resources devoted to video chaptering, such as Chapter-Llama (Ventura et al., 2025) and the VidChapters-7M dataset (Yang et al., 2023), illustrate how ASR transcripts combined with visual features can yield robust segment-level indices. These insights are directly applicable to VRD pipelines, where similar methods can strengthen section- or figure-level retrieval.

Conversion to a dominant modality Outside text, *proxy conversion* is a common way to reuse strong tooling. In 3D perception, MV3D (Chen et al., 2017), PIXOR (Yang et al., 2018), and PointPillars (Lang et al., 2019) project LiDAR point clouds to BEV or pseudo-images to leverage 2D detectors and infra. While their target modality is vision, the strategy is analogous to VRD captioners: convert heterogeneous inputs into the most mature stack. For enterprise VRDs, the most mature stack is text retrieval, hence captioning and structural textualization is the natural endpoint.

Where modern MLLMs fit Modern MLLMs enable seamless integration of captioning within VRD pipelines. These models generate both page-level and region-level descriptions, convert chart and table content into text by describing units, axes, trends, and outliers, and can even produce structured representations in formats such as HTML, Markdown, or JSON, following the design principles of systems like Pix2Struct and Donut. Empirical evidence shows that employing more capable captioners leads to measurable gains in recall and answer accuracy for slide and PDF question answering tasks (LangChain Team, 2023; Riedler and Langer, 2024), even when the retrieval process at query time remains entirely text-based.

Advantages Across modalities, the strategy of first converting heterogeneous inputs into a dominant or well-supported modality reflects a shared set of motivations. By transforming diverse signals into text, practitioners can leverage decades of progress in indexing, retrieval, and evaluation. Cascaded architectures allow modular replacement and incremental upgrading of components such as OCR or retrievers. This design enhances interpretability, facilitates debugging, and eases deployment and optimization in production environments. Additionally, online query latency remains unaffected, as all processing of charts and tables by VLMs is confined to the preprocessing stage.

Disadvantages Despite these advantages, the paradigm carries inherent risks. Captioning or transcription inevitably compresses the source signal, risking the omission of fine-grained visual or temporal information. Highly structured visual elements, such as charts, diagrams, or tables, often lose numerical precision or relational cues when summarized in free-form text, a weakness that motivates corrective research like chart-to-text generation. Recognition errors from ASR or OCR can cascade, significantly degrading downstream retrieval or QA accuracy. The Spoken-SQuAD dataset quantified this impact for speech QA (Li et al., 2018), while models such as Donut explicitly sought to eliminate OCR-induced error chains through end-to-end document decoding (Kim et al., 2022). Furthermore, preprocessing for large-scale captioning with this approach can be costly. Processing vast repositories of documents, each potentially containing numerous images and tables, requires substantial computational resources and time for the MLLM to generate descriptions for every non-textual element. This upfront cost can be a major bottleneck, especially for dynamic datasets where new multimodal content is frequently added.

This suggests that while the modality-unifying captioner role offers an accessible path to multimodal RAG, it may be best suited for applications where the non-textual elements are relatively simple, where some information loss is tolerable, or where the scale of data does not make preprocessing costs insurmountable.

3.2 MLLMs as Multimodal Embedders

As shown in Figure 1 (middle), this role embeds heterogeneous inputs into a shared space to enable cross-modal search and matching.

While the captioner role (§3.1) is practical, restricting MLLMs solely to text conversion has inherent limitations. In response to these constraints, the research community has increasingly focused on leveraging the advanced representation capabilities of MLLMs to enhance multimodal RAG. A prominent direction within this effort involves utilizing MLLMs as *Multimodal Embedders*. In this role, MLLMs function directly as powerful embedding models, transforming data from diverse modalities into a shared, rich semantic feature space.

The Core Mechanism Instead of converting modalities to text, the MLLM learns to map inputs from different modalities into a common high-dimensional vector space. In this shared space, the embeddings of semantically related items from different modalities are expected to be close to each other, allowing for direct comparison, similarity search, and retrieval across modalities. For example, an image query could retrieve relevant textual passages, or a textual query could retrieve relevant images and text.

Historical Roots This fundamental idea, unifying disparate modalities into a common representational space to facilitate joint reasoning and retrieval, has deep historical roots. The most canonical instantiation is perhaps CLIP (Radford et al., 2021) and its numerous successors (Jia et al., 2021; Zhai et al., 2022; Li et al., 2022a, 2023; Zhai et al., 2023; Yao et al., 2021; Yu et al., 2022), which align text and image representations via contrastive learning, establishing CLIP as the de facto standard embedding backbone in early multimodal RAG systems. Earlier precursors include DeViSE (Frome et al., 2013), which projected visual features into word2vec semantic space for zero-shot recognition; Deep CCA (Andrew et al., 2013) and its deep extensions DGCCA (Benton et al., 2019), which learned shared subspaces via canonical correlation analysis; VSE++ (Faghri et al., 2018), which emphasized hard negative mining for improved alignment; and SCAN (Lee et al., 2018), which introduced stacked cross-attention to enable fine-grained word-region alignment for stronger image–text matching. More recently, ImageBind (Girdhar et al., 2023) unified six modalities into a single embedding space, achieving cross-modal alignment using only image-paired data as a bridging signal.

The Shift to MLLM-based Embedders Nevertheless, recent studies (Zhou et al., 2024b) have indicated that the text embedding capabilities of these vision-language models (e.g., CLIP) are comparatively inferior to those of specialized text embedding models. This limitation may hinder their effectiveness in tasks involving text-intensive multimodal documents. Drawing inspiration from pioneering work such as LLM2Vec (BehnamGhader et al., 2024), there has been a surge in research efforts aimed at repurposing MLLMs as embedding models (Jiang et al., 2025a,b; Meng et al., 2025; Zhou et al., 2024a; Lin et al., 2025a; Zhang et al., 2025c; Lan et al., 2025; Liu et al., 2024b; Chen et al., 2025a; Lee et al., 2025b; Lin et al., 2025b). Representative approaches include: VLM2Vec (Jiang et al., 2025b; Meng et al., 2025), which endows VLMs with instruction-aware embedding capabilities and reports consistent gains across the Massive Multimodal Embedding Benchmark (MMEB); MM-Embed (Lin et al., 2025a), which identifies and mitigates modality bias via modality-aware hard negative mining and continual text-to-text fine-tuning; and E5-V (Jiang et al., 2025a), which surprisingly achieves state-of-the-art multi-modal retrieval by training exclusively on text pairs, leveraging prompting to bridge modalities and drastically reduce annotation and training costs.

Training Strategy This methodology, mirroring LLM2Vec, transforms MLLMs into CLIP-analogous representation models by embedding diverse modalities into a shared feature space. For instance, in the case of VLMs, the process involves aggregating extensive datasets similar to CLIP's training corpus. During training, the EOS token serves as the representative token, and contrastive learning employs InfoNCE loss (van den Oord et al., 2019). Through this approach, textual and visual modalities are seamlessly integrated into a unified feature space. Leveraging the MLLM's world knowledge from multimodal next-token prediction (Chen et al., 2024a), this method demonstrates exceptional representational capacity across diverse data types. Moreover, it offers the flexibility to replace existing embedding models with minimal disruption.

In addition to CLIP-style training, MoCa (Chen et al., 2025a) converts causal VLMs into bidirectional multimodal embedders via continual pre-training and heterogeneous contrastive finetuning. Vision-centric contrastive learning (VC^{2}L) (Lin et al., 2025b) renders mixed text–image content into pixels to avoid OCR misalignment. General training advances include a generalized contrastive loss (GCL) (Lee et al., 2025b) that jointly contrasts text, image, and fused representations within a batch.

The Role of Data and Synthetic Supervision Beyond algorithmic design, data composition and synthetic supervision play pivotal roles. Zhang et al. (2025c) target universal multimodal retrieval over visually rich documents, emphasizing balanced modality mixing and efficient generation of fused-modality training pairs. Zhou et al. (2024a) scale synthetic supervision by generating instruction-style queries over image pairs, significantly enhancing zero-shot generalization.

Optimization On the optimization front, LLaVE (Lan et al., 2025) introduces hardness-weighted contrastive learning to better separate ambiguous negatives. Complementing standard bi-encoder frameworks, LamRA (Liu et al., 2024b) attaches lightweight LoRA (Hu et al., 2022) heads to generative MLLMs, unifying retrieval and reranking and enabling strong transfer to unseen retrieval tasks.

Empirical Evidence and Performance The superior impact of these MLLM-derived embeddings is evident in downstream performance, with extensive evaluations confirming better representational quality and metrics. Table 2, compiling results from the MMEB introduced by Jiang et al. (2025b), illustrates this trend. Notably, employing MLLMs as Multimodal Embedders yields substantial performance enhancements compared to RAG systems using conventional multimodal embedding models like CLIP. In VQA-related tasks, for example, MLLMs as Multimodal Embedders leverage their inherent advanced visual reasoning capabilities, highlighting their distinct advantage.

Complementary advances in reranking Several concurrent efforts investigate VLM-based reranking methods to complement retrieval with reasoning-aware relevance modeling (Xu et al., 2025a; Wasserman et al., 2025a; Chen et al., 2025c; Gong et al., 2025).

3.3 MLLMs as End-to-End Representers

As illustrated in Figure 1 (right), the *End-to-End Representers* role encodes whole pages directly to retrieve at page granularity. It uses MLLMs to generate holistic representations directly from entire multimodal inputs, such as treating full document pages as single images. Instead of breaking a document down into its constituent parts and processing them separately or converting them, the MLLM takes a more holistic view. For example, an entire PDF page, with its complex layout of text, images, and graphical elements, might be fed as a single image input to the MLLM, which then generates a unified representation for that entire page.

A key characteristic of this approach is that it bypasses intermediate steps like explicit OCR or layout parsing. Traditional document processing pipelines often rely on separate modules for OCR, layout analysis, and then subsequent processing of these extracted elements. Each of these stages can introduce errors.

Rationale To illustrate this methodology, consider the example of VLMs. In this instance, specific components within traditional RAG pipelines are replaced with VLMs, thereby enabling direct end-to-end representation generation. This approach is motivated by two key factors. Firstly, previous research has demonstrated that the process of OCR introduces noise into RAG systems, degrading their performance (Zhang et al., 2025b; Xie et al., 2025). Secondly, the advanced visual comprehension capabilities of contemporary VLMs render separate identification of layouts, tables, images, and other discrete elements unnecessary. Instead, an entire PDF page can be treated as a single image input to a VLM, thereby facilitating the production of a holistic representation.

Exemplary Models Significant contributions have been made in this domain, with DSE (Ma et al., 2024), ColPali (Faysse et al., 2025) and VisRAG (Yu et al., 2025) being particularly noteworthy examples. DSE has the capacity to convert document screenshots directly into dense vectors for retrieval. ColPali incorporates the late-interaction matching mechanism of ColBERT (Khattab and Zaharia, 2020), embedding document page images into high-dimensional vector spaces for retrieval. This method excels at capturing intricate visual details and is simple, fast, and end-to-end trainable. Similarly, VisRAG directly encodes and retrieves document pages, mitigating information loss while fully exploiting the visual content present in documents. These approaches adopt InfoNCE loss for training, aligning with the training approach of the *Multimodal Embedders* role.

Beyond Single-Page Beyond page-level late-interaction encoders, multi-page representers decouple retrieval and reasoning. DREAM (Zhang et al., 2025a) integrates hierarchical multimodal retrieval and a multi-page VLM with global and token-level cross-page attention. ColMate (Masry et al., 2025), while primarily an embedder, inherits ColBERT-style end-to-end matching over page images and masked text. Industry efforts such as Docopilot (Duan et al., 2025) study non-RAG, end-to-end multi-round document understanding over the Doc-750K corpus, complementary to retrieval-centric approaches.

Advantage This end-to-end Representer methodology capitalizes on the advanced representational capabilities of MLLMs while concurrently reducing the overall latency of the pipeline (Faysse et al., 2025; Yu et al., 2025). In traditional multimodal RAG, predominant latency sources are initial layout analysis, segmentation, and OCR, not embedding itself. Employing MLLMs for end-to-end recognition, despite a slight increase in embedding duration, results in a substantial reduction in total processing time. This is demonstrated in Table 1, which compares the latency of an OCR-reliant pipeline with an MLLM-based end-to-end representer, showing a reduction in total offline latency for the MLLM-based approach due to the elimination of parsing overhead.

This approach can also reduce the noise caused by imperfect parsing. OCR errors, misinterpretations of document layout, or failures to correctly segment different content blocks can degrade the quality of information fed into a RAG system. An MLLM that directly sees the entire page might learn to be more robust to such variations or low-quality inputs, as it can leverage the global context of the page. This holistic processing can be particularly advantageous for ingesting large volumes of complex documents, such as scanned PDFs or documents with unconventional layouts, where traditional parsing tools might struggle.

Furthermore, end-to-end training integrates the previously acquired world knowledge and inherent capabilities of MLLMs, and thus elevates the performance ceiling of multimodal RAG systems.

4 Trade-offs and Future Directions

While integrating MLLMs into RAG systems offers significant benefits, this paradigm also presents challenges and is not universally optimal. Key limitations involve retrieval granularity, information fidelity, and computational and storage demands.

4.1 Retrieval Granularity and Interpretability

4.1.1 Coarse vs. Fine-Grained Retrieval

The *End-to-End Representer* role, despite preprocessing benefits, often yields coarser retrieval granularity. For example, both ColPali and VisRAG adopt the page as the retrieval unit (Faysse et al., 2025; Yu et al., 2025). While representing a whole page with one vector identifies relevant pages, it obscures fine-grained details, forcing a secondary search for specific facts, unlike text-based RAG systems that retrieve individual paragraphs or sentences. This highlights a fundamental tension: holistic processing improves robustness but sacrifices retrieval precision, whereas fine-grained retrieval enhances precision but risks losing global context or suffering from error propagation.

4.1.2 Information Loss in Conversion

Similarly, the Modality-Unifying Captioner role, which converts non-textual elements to text, inherently suffers from information loss, as textual descriptions rarely capture the full richness of images, tables, or diagrams. This preprocessing information loss directly degrades fidelity: if the LLM generator receives incomplete or oversimplified context, the final output will lack nuance and accuracy, undermining the RAG system's purpose.

Ideal granularity and acceptable information loss are application-dependent. For instance, general summarization may tolerate coarser granularity, whereas fact-checking demands high-fidelity, fine-grained retrieval. This tension highlights the need for adaptive, task-aware systems rather than a single, universally optimal strategy.

Recent studies (Gong et al., 2025; Chen et al., 2025c; Zhang et al., 2025a) further demonstrate that adaptive hierarchical and co-modality retrieval strategies can effectively recover fine-grained evidence and improve cross-page reasoning in visually rich documents.

4.2 Computational Overhead and Costs

4.2.1 Increased Latency

Generating rich multimodal embeddings or detailed textual captions using MLLMs is computationally intensive. Figure 2 shows MLLMs as multimodal embedders incur substantially higher latency during both offline encoding and online searching compared to CLIP-based models. This starkly illustrates why model miniaturization is essential for broader applicability.

4.2.2 Substantial Storage Demands

MLLM-RAG systems also face substantial storage demands. MLLMs in the *Multimodal Embedder* role produce high-dimensional embeddings, often significantly larger than traditional text vectors, to capture rich cross-modal information. Storing these vectors for large corpora can become prohibitive. For instance, Lin et al. (2025a) report that its index storage demands exceed those of CLIP-based models by a factor of five or more.

This increased storage footprint not only incurs direct hardware costs but also degrades efficiency by slowing index loading and vector searches, compounding latency issues.

Potential solutions include model miniaturization via higher-quality data or knowledge distillation (Hinton et al., 2015), which could produce compact Multimodal Small Language Models to address these root challenges.

Another promising avenue is the adoption of a Matryoshka-style multimodal learning framework (Sturua et al., 2024; Cai et al., 2025), which learns representations across multiple granularities. By dynamically selecting inference modes, this approach could offer a scalable performance-cost gradient tailored to downstream tasks.

Recent works (Rajendran et al., 2025; Yan et al., 2025; Günther et al., 2025; Masry et al., 2025) have also explored efficiency-oriented solutions that balance accuracy and cost through adaptive routing, vector pruning, and lightweight embedding designs.

4.3 Challenges in Evaluation Metrics

Evaluating multimodal RAG remains fundamentally difficult because traditional metrics, largely developed for text-only settings, cannot fully capture the fidelity and interpretability of cross-modal reasoning. While frameworks such as RAGAs (Es et al., 2024) and ARES (Saad-Falcon et al., 2024) provide initial measures for faithfulness and relevance, multimodal scenarios introduce new failure sources, including misaligned visual grounding and inconsistencies between retrieved and generated evidence (Mortaheb et al., 2025). Recent benchmarks (Wasserman et al., 2025b; Peng et al., 2025) highlight that current systems often underperform on real-world, document-heavy, and paraphrase-variant data, underscoring a persistent gap between laboratory metrics and practical robustness. Human-centered datasets can also help narrow this pragmatic gap (Zhang, 2025a).

A more holistic evaluation paradigm is needed, combining end-to-end performance with modality-aware diagnostics such as table and figure grounding accuracy, cross-page evidence localization, and paraphrase robustness, aligning with broader calls for benchmarks that prioritize safety and real-world user needs (Zhang, 2025b). Progress in this direction will enable fairer comparison across retrieval granularity levels and provide actionable signals for improving factual alignment and interpretability in visually rich document RAG systems.

5 Conclusion

This survey has chartered the evolving landscape of Retrieval-Augmented Generation for visually rich documents, focusing on the critical roles played by MLLMs. We have structured this emergent field by proposing a taxonomy of three primary roles: *Modality-Unifying Captioners*, *Multimodal Embedders*, and *End-to-End Representers*.

Our analysis reveals that there is no single, universally optimal solution. Instead, practitioners face a distinct set of trade-offs. The *Captioner* role offers a pragmatic path to multimodal support by integrating with mature, text-based RAG pipelines, but at the risk of information loss and error cascades from imperfect textual conversion. The *Embedder* role enables true cross-modal search by unifying modalities in a shared vector space, but this power often comes at the cost of significant computational and storage overhead. Finally, the *Representer* role provides robustness by bypassing brittle OCR and parsing steps, but this simplicity typically sacrifices retrieval precision by operating at a coarse, page-level granularity.

These findings highlight a tension in the field: a balancing act between retrieval granularity, information fidelity, computational cost, and pipeline simplicity. As the field matures, we anticipate future research will focus on three challenges. First, the development of adaptive and hierarchical retrieval methods to dynamically blend coarse-grained and fine-grained retrieval to get the best of both. Second, the need for model miniaturization and efficiency, producing smaller, faster MLLMs that make these advanced techniques practical for real-world latency and storage budgets. Finally, the design of next-generation evaluation benchmarks that move beyond simple text-based metrics to holistically measure factual accuracy, cross-modal grounding, and the interpretability of RAG systems handling complex, visually-grounded evidence.

Limitations

This survey has limitations. Firstly, its scope is constrained by available literature on MLLMs in multimodal RAG. The generalizability of the synthesized findings may be limited by the datasets, MLLMs, and tasks predominantly featured in these studies. Secondly, while performance and latency are discussed based on reported figures, this survey does not account for the variability in hardware configurations or deployment environments used in those studies, which could impact real-world applicability comparisons. Lastly, the reviewed literature often focuses more on technical and performance aspects, with less emphasis on user-centric evaluation metrics such as nuanced interpretability and usability. This survey reflects that focus, leaving broader user-centric analyses for future work or dedicated studies.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA. PMLR.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2019. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 1–6, Florence, Italy. Association for Computational Linguistics.

Xiaowei Bi and Zheyuan Xu. 2025. Everything can be described in words: A simple unified multi-modal framework with semantic and temporal alignment. *Preprint*, arXiv:2503.09081.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2025. Matryoshka multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. 2025a. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. *Preprint*, arXiv:2506.23115.

Jian Chen, Ming Li, Jihyung Kil, Chenguang Wang, Tong Yu, Ryan Rossi, Tianyi Zhou, Changyou Chen, and Ruiyi Zhang. 2025b. Visr-bench: An empirical study on visual retrieval-augmented generation for multilingual long document understanding. *Preprint*, arXiv:2508.07493.

Liang Chen, Zekun Wang, Shuhuai Ren, Lei Li, Haozhe Zhao, Yunshui Li, Zefan Cai, Hongcheng Guo, Lei Zhang, Yizhe Xiong, Yichi Zhang, Ruoyu Wu, Qingxiu Dong, Ge Zhang, Jian Yang, Lingwei Meng, Shujie Hu, Yulong Chen, Junyang Lin, and 8 others. 2024a. Next token prediction towards multimodal intelligence: A comprehensive survey. *Preprint*, arXiv:2412.18619.

Wang Chen, Wenhan Yu, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. 2025c. Cmrag: Co-modality-based visual document retrieval and question answering. *Preprint*, arXiv:2509.02123.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi

Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2025. Deep learning based visually rich document content understanding: A survey. *Preprint*, arXiv:2408.01287.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset. *Preprint*, arXiv:1910.09387.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, Jifeng Dai, and Wenhai Wang. 2025. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4026–4037.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. *Preprint*, arXiv:1707.05612.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024a. 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*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024b. Videoagent: A memory-augmented multimodal agent for video understanding. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXII*, page 75–92, Berlin, Heidelberg. Springer-Verlag.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Sensen Gao, Shanshan Zhao, Xu Jiang, Lunhao Duan, Yong Xien Chng, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, Jia-Wang Bian, and Mingming Gong. 2025. Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding. *Preprint*, arXiv:2510.15253.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Ziyu Gong, Chengcheng Mai, and Yihua Huang. 2025. Mhier-rag: Multi-modal rag for visual-rich document question-answering via hierarchical and multi-granularity reasoning. *Preprint*, arXiv:2508.00579.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, Suzhou, China. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the*

37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, and 5 others. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298, Vienna, Austria. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Ting Jiang, Shaohan Huang, Minghui Song, Zihan Zhang, Haizhen Huang, Liang Wang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, deqing wang, and Fuzhen Zhuang. 2025a. E5-v: Universal embeddings with multimodal large language models.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025b. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, page 498–517, Berlin, Heidelberg. Springer-Verlag.

Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. 2025. Genius: A generative framework for universal multimodal search. *Preprint*, arXiv:2503.19868.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *Preprint*, arXiv:2503.04812.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

LangChain Team. 2023. Multi-modal rag template.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. 2025b. Generalized contrastive learning for universal multimodal retrieval. *Preprint*, arXiv:2509.25638.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Preprint*, arXiv:1804.00320.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *Preprint*, arXiv:2403.10446.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022b. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025a. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*.

Yiqi Lin, Alex Jinpeng Wang, Linjie Li, Zhengyuan Yang, and Mike Zheng Shou. 2025b. Exploring a unified vision-centric contrastive alternatives on multimodal web documents. *Preprint*, arXiv:2510.18703.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2024b. Lamra: Large multimodal model as your advanced retrieval assistant. *Preprint*, arXiv:2412.01720.

Zhenghao Liu, Xingsheng Zhu, Tianshuo Zhou, Xinyi Zhang, Xiaoyuan Yi, Yukun Yan, Yu Gu, Ge Yu, and Maosong Sun. 2025. Benchmarking retrieval-augmented generation in multi-modal contexts. *Preprint*, arXiv:2502.17297.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Preprint*, arXiv:2411.13093.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. *Preprint*, arXiv:2009.08553.

Ahmed Masry, Megh Thakkar, Patrice Bechard, Sathwik Tejaswi Madhusudhan, Rabiul Awal, Shambhavi Mishra, Akshay Kalkunte Suresh, Srivatsava Daruru, Enamul Hoque, Spandana Gella, Torsten Scholak, and Sai Rajeswar. 2025. ColMate: Contrastive late interaction and masked text for multimodal document retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2071–2080, Suzhou (China). Association for Computational Linguistics.

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

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

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

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Preprint*, arXiv:2507.04590.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Matin Mortaheb, Mohammad A. Amir Khojastepour, Srimat T. Chakradhar, and Sennur Ulukus. 2025. Rag-check: Evaluating multimodal retrieval augmented generation performance. *Preprint*, arXiv:2501.03995.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *Preprint*, arXiv:2510.03663.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ravi K. Rajendran, Biplob Debnath, Murugan Sankaradass, and Srimat Chakradhar. 2025. EcoDoc: A cost-efficient multimodal document processing system for enterprises using LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1530–1537, Vienna, Austria. Association for Computational Linguistics.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *Preprint*, arXiv:2502.01549.

Monica Riedler and Stefan Langer. 2024. Beyond text: Optimizing rag with multimodal inputs for industrial applications. *Preprint*, arXiv:2410.21943.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. *Preprint*, arXiv:2003.12462.

Shamane Siriwardhana, Rivindu Weerasekera, Tharindu Kaluarachchi, Elliott Wen, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Trans. Assoc. Comput. Linguistics*, 11:1–17.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Annie Surla, Aditi Bodhankar, and Tanay Varshney. 2024. An easy introduction to multimodal retrieval-augmented generation.

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. 2025. Benchmarking llm faithfulness in rag with evolving leaderboards. *Preprint*, arXiv:2505.04847.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13636–13645.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *Preprint*, arXiv:2102.04643.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2025. Chapter-llama: Efficient chaptering in hour-long videos with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18947–18958.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Navve Wasserman, Oliver Heinimann, Yuval Golbari, Tal Zimbalist, Eli Schwartz, and Michal Irani. 2025a. Docrerank: Single-page hard negative query generation for training multi-modal rag rerankers. *Preprint*, arXiv:2505.22584.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025b. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31660–31683, Vienna, Austria. Association for Computational Linguistics.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Peijin Xie, Shun Qian, Bingquan Liu, Dexin Wang, Lin Sun, and Xiangzheng Zhang. 2025. Textlessrag: End-to-end visual document rag by speech without text. *Preprint*, arXiv:2509.07538.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021a. Video-CLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta

Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

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. *Preprint*, arXiv:2506.12364.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021b. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Zhichao Xu, Fengran Mo, Zhiqi Huang, Crystina Zhang, Puxuan Yu, Bei Wang, Jimmy Lin, and Vivek Srikumar. 2025b. A survey of model architectures in information retrieval. *Preprint*, arXiv:2502.14822.

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. *Preprint*, arXiv:2509.23883.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vidchapters-7m: Video chapters at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 49428–49444. Curran Associates, Inc.

Bin Yang, Wenjie Luo, and Raquel Urtasun. 2018. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *Preprint*, arXiv:2502.14864.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *Preprint*, arXiv:2111.07783.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Retrieval-augmented multimodal language modeling. *Preprint*, arXiv:2211.12561.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Preprint*, arXiv:2205.01917.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Nianbo Zeng, Haowen Hou, Fei Richard Yu, Si Shi, and Ying Tiffany He. 2025. Scenerag: Scene-level retrieval-augmented generation for video understanding. *Preprint*, arXiv:2506.07600.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133.

Jinxu Zhang, Qiyuan Fan, Yongqi Yu, and Yu Zhang. 2025a. Dream: Integrating hierarchical multimodal retrieval with multi-page multimodal language model for documents vqa. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 4213–4221, New York, NY, USA. Association for Computing Machinery.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025b. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. *Preprint*, arXiv:2412.02592.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. *Preprint*, arXiv:2403.19651.

Xiantao Zhang. 2025a. AuraDial: A large-scale human-centric dialogue dataset for Chinese AI psychological counseling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2847–2863, Suzhou, China. Association for Computational Linguistics.

Xiantao Zhang. 2025b. The escalator problem: Identifying implicit motion blindness in ai for accessibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6635–6643.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025c. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. *Preprint*, arXiv:2303.10868.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. *Preprint*, arXiv:2503.18016.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024a. Megapairs: Massive data synthesis for universal multimodal retrieval. *Preprint*, arXiv:2412.14475.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024b. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2025. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):26183–26191.

A Supplemental Data

This section provides supplementary empirical data referenced in the main survey, offering a more detailed view of the performance and cost trade-offs discussed.

Model	Offline (ms)			Online (ms)		
	P.	E.	Total	E.	S.	Total
MiniCPM	284	28	312	28	26	54
VisRAG-Ret	–	121	121	28	26	54

Table 1: Latency comparison between an OCR-reliant pipeline (MiniCPM (Hu et al., 2024)) and an MLLM-based end-to-end representer (VisRAG-Ret (Yu et al., 2025)) during offline and online processing stages. Abbreviations: P. - Parsing; E. - Encoding; S. - Searching.

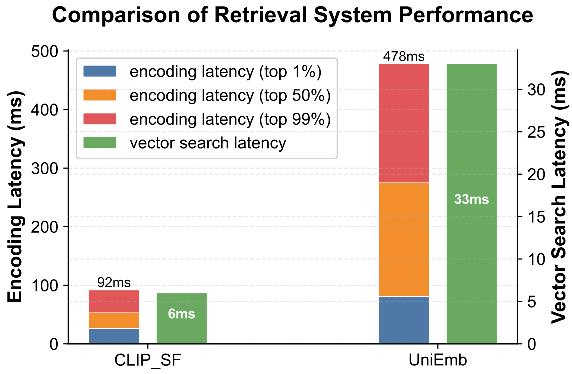

Figure 2: Comparison of encoding latency (displaying top 1%, 50%, and 99th percentiles) and vector search latency for the CLIP_{SF} (Wei et al., 2024) and UniEmb (Lin et al., 2025a) models. Measurements were based on 100 randomly sampled queries from each of the 16 M-BEIR (Wei et al., 2024) tasks.

- **Table 2** presents a comprehensive comparison on the MMEB benchmark, substantiating the performance gains of the *MLLM as Multimodal Embedder* role (§3.2) over traditional baselines.
- **Table 3** details retrieval performance (MRR@10) across various VQA datasets, comparing *End-to-End Representers* (§3.3) with baseline methods.
- **Table 1** and **Figure 2** provide specific latency measurements, illustrating the computational overhead and costs discussed in §4.2.

Model	**Per Meta-Task Score**				**Average Score**
	Classification	VQA	Retrieval	Grounding	Overall
# of datasets →	10	10	12	4	36
Baselines					
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	37.8
BLIP2 (Li et al., 2023)	27.0	4.2	33.9	47.0	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	34.8
OpenCLIP (Cherti et al., 2023)	**47.8**	10.9	52.3	53.3	39.7
UniIR ($BLIP_{FF}$) (Wei et al., 2024)	42.1	<u>15.0</u>	<u>60.1</u>	<u>62.2</u>	<u>42.8</u>
UniIR ($CLIP_{SF}$) (Wei et al., 2024)	<u>44.3</u>	**16.2**	**61.8**	**65.3**	**44.7**
Magiclens (Zhang et al., 2024)	38.8	8.3	35.4	26.0	27.8
Baseline Average	40.4	10.3	46.9	52.2	36.1
MLLMs as Multimodal Embedders					
VLM2Vec (Phi-3.5-V-4B) (Jiang et al., 2025b)	54.8	54.9	62.3	79.5	60.1
VLM2Vec (LLaVA-1.6-7B) (Jiang et al., 2025b)	61.2	49.9	67.4	<u>86.1</u>	62.9
VLM2Vec (Qwen2-VL-7B) (Jiang et al., 2025b)	<u>62.6</u>	57.8	<u>69.9</u>	81.7	<u>65.8</u>
MMRet-MLLM (LLaVA-1.6-7B) (Zhou et al., 2024a)	56.0	57.4	<u>69.9</u>	83.6	64.1
GME (Qwen2-VL-2B) (Zhang et al., 2025c)	56.9	41.2	67.8	53.4	55.8
LLaVE-2B (Lan et al., 2025)	62.1	<u>60.2</u>	65.2	84.9	65.2
LLaVE-7B (Lan et al., 2025)	**65.7**	**65.4**	**70.9**	**91.9**	**70.3**
MLLM-based Average	59.9	55.3	67.6	80.2	63.5
Average Improvement (Δ = MLLM-based - Baselines)	+19.5	+45.0	+20.7	+28.0	+27.4

Table 2: Performance comparison of multimodal embedding models on the MMEB benchmark, compiled from (Jiang et al., 2025b) and other cited works. Scores are averaged per meta-task, and an overall average score is also provided. Within each model category, the best reported performance for each task is marked in **bold**, and the second-best is <u>underlined</u>. This table synthesizes results to highlight the contrast between these model categories and summarizes the average improvement reported for MLLMs over the baselines.

Model	**ArxivQA**	**ChartQA**	**DocVQA**	**InfoVQA**	**PlotQA**	**SlideVQA**	**Average**
Baselines							
BM25 (2009) (OCR)	<u>43.65</u>	61.47	<u>75.27</u>	66.94	57.28	86.78	65.23
bge-large (2023) (OCR)	39.29	59.64	75.04	72.38	51.33	81.38	59.13
MiniCPM (2024) (OCR)	58.43	77.74	72.54	<u>83.45</u>	**64.78**	<u>91.74</u>	<u>74.78</u>
NV-Embed-v2 (2025a) (OCR)	**59.39**	<u>80.47</u>	**75.46**	**84.24**	59.36	**92.49**	**75.24**
SigLIP (2023)	<u>59.16</u>	**81.34**	64.60	74.59	<u>61.32</u>	89.08	71.68
MLLMs as End-to-End Representers							
ColPali (2025)	<u>72.50</u>	73.49	**82.79**	81.15	55.32	**93.99**	<u>76.54</u>
VisRAG-Ret (2025)	**75.11**	76.63	75.37	**86.37**	<u>62.14</u>	91.85	**77.91**

Table 3: Overall retrieval performance (MRR@10) across multiple Visual Question Answering (VQA) datasets, summarizing results from cited studies. This table synthesizes and compares reported performances of traditional baselines with *MLLMs as End-to-End Representers*. In each model category, the best reported performance is marked in **bold**, and the second-best is <u>underlined</u>.