arXiv:submit/6310504 [cs.AI] 26 Mar 2025

A Survey on Multimodal Retrieval-Augmented Generation

LANG MEI, Huawei Cloud BU, China
SIYU MO, Huawei Cloud BU, China
ZHIHAN YANG, Huawei Cloud BU, China
CHONG CHEN*, Huawei Cloud BU, China

Multimodal Retrieval-Augmented Generation (MRAG) represents a significant advancement in enhancing the capabilities of large language models (LLMs) by integrating multimodal data, such as text, images, and videos, into the retrieval and generation processes. Traditional Retrieval-Augmented Generation (RAG) systems, which primarily rely on textual data, have shown promise in reducing hallucinations and improving response accuracy by dynamically incorporating external knowledge. However, these systems are limited by their reliance on text-only modalities, which restricts their ability to leverage the rich, contextual information available in multimodal data. MRAG addresses this limitation by extending the RAG framework to include multimodal retrieval and generation, thereby enabling more comprehensive and contextually relevant responses. In MRAG, the retrieval step involves locating and integrating relevant knowledge from diverse modalities, while the generation step utilizes multimodal large language models (MLLMs) to produce answers that incorporate information from multiple data types. This approach not only enhances the quality of question-answering systems but also significantly reduces the incidence of hallucinations by grounding responses in factual, multimodal knowledge. Recent research has demonstrated that MRAG outperforms traditional text-modal RAG, particularly in scenarios where visual and textual information are both critical for understanding and responding to queries. This survey systematically reviews the current state of MRAG research, focusing on four key aspects: essential components and technologies, datasets, evaluation methods and metrics, and existing limitations. By analyzing these dimensions, we aim to provide a comprehensive understanding of how MRAG can be effectively constructed and improved. Additionally, we highlight current challenges and propose future research directions, encouraging further exploration into this promising paradigm. Our work underscores the potential of MRAG to revolutionize multimodal information retrieval and generation, offering a forward-looking perspective on its development and applications.

CCS Concepts: • **Information systems → Multimedia and multimodal retrieval**; **Language models**; • **Computing methodologies → Natural language processing**.

Additional Key Words and Phrases: Multimodal Retrieval-Augmented Generation, Multimodal Large Language Model, Multimodal Document Parsing and Indexing, Multimodal Search Planning

ACM Reference Format:
Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2018. A Survey on Multimodal Retrieval-Augmented Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 80 pages. https://doi.org/XXXXXXX.XXXXXXX

*Chong Chen is the corresponding author.

Authors' Contact Information: Lang Mei, Huawei Cloud BU, Beijing, China, meilang1@huawei.com; Siyu Mo, Huawei Cloud BU, Beijing, China, mosiyu@huawei.com; Zhihan Yang, Huawei Cloud BU, Beijing, China, yangzhihan4@huawei.com; Chong Chen, Huawei Cloud BU, Beijing, China, chenchong55@huawei.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large language models (LLMs), especially the Transformer-based variants, have achieved extraordinary success in many language-related tasks. Through pre-training on extensive, high-quality instruction datasets, LLMs can learn a wide range of language patterns, structures, and factual knowledge. These pre-trained LLMs can generate human-like text with high degrees of fluency and coherence, and attain strong performance on question-answering tasks, which demonstrates their ability to understand and respond to a wide range of queries. However, despite their impressive capabilities, LLMs still face significant limitations. One of the primary challenges lies in their performance within specific domains or knowledge-intensive tasks. While these models are often trained on diverse and extensive datasets, such datasets may not cover the depth of knowledge required for highly specialized fields or real-time information updates. This can be particularly problematic in areas like medicine, law, finance, and other technical fields where precision and up-to-date knowledge are to be prioritized. When handling queries that extend beyond the scope of their training knowledge or require the most current information, LLMs may generate responses that are speculative or based on patterns they have learned, rather than on verified facts. This can result in misleading, incorrect, or even entirely fabricated answers, a phenomenon known as "hallucination". Minimizing the incidence of hallucinations is important for enhancing the reliability of LLMs in providing accurate and context-relevant information across different domains.

Recently, Retrieval-Augmented Generation (RAG) has emerged as an effective solution to mitigate hallucinations, by enhancing the generation capabilities of large language models (LLMs) through the retrieval of relevant external knowledge. Existing RAG systems typically operate through a two-step process: retrieval and generation. In the retrieval step, the goal is to quickly locate relevant knowledge that is semantically similar to the query from a large-scale document collection. Since the relevant knowledge is often scattered across various parts of documents, each document is pre-processed into multiple chunks. Additional chunks may be created through manual or automated methods. This process, known as document chunkerization, ensures that fine-grained knowledge can be retrieved more efficiently. In the generation step, the retrieved document chunks are combined with the query to form an augmented input. This augmented input provides the LLM with context that includes external knowledge. Furthermore, RAG allows LLMs to dynamically integrate the latest information during the inference stage. This capability ensures that the model's responses are not only based on static, pre-trained knowledge but are continuously updated with current and relevant data. By retrieving and referencing external knowledge, RAG grounds the generated responses in factual information, thereby significantly reducing the occurrence of hallucinations. However, previous research on RAG systems has primarily focused on knowledge bases built from plain text and LLMs pre-trained on plain text, ignoring other rich sources of knowledge available for query responses in the real world, such as videos and images, referred to as "multimodal data".

Multimodal data refers to data that comes from multiple sources or formats. This can include text, images, audio, video, and other types of data. In real-world scenarios, humans naturally interact with multimodal data, such as browsing web pages that combine text, images, and videos in mixed layouts. By analyzing images or videos alongside text, the user can better understand the context of the content, and thus improve the satisfaction with the quality of the answers. For example, if a passenger inquires about how to store luggage while flying, it will be clearer that the system provides relevant graphic guides or instructional videos. However, transferring the capabilities of LLMs to the domain of multimodal text and images remains an active area of research, as plain-text LLMs are typically trained only on textual corpora and lack perceptual abilities for visual signals. How to effectively incorporate multimodal data is important to enhance the capability of

LLMs. In recent years, the development of multimodal generative models has showcased additional application possibilities. Apart from textual generative models, multimodal generative models have been increasingly applied in fields such as human-computer interaction, robot control, image search, and speech generation. Similarly, based on multimodal generative models and multimodal data, how to effectively process Multimodal Retrieval-Augmented Generation (**MRAG**) is an issue that needs to be explored.

Recently, some research have demonstrated that MRAG with multimodal data outperforms traditional text-modal RAG. By enhancing the generation capabilities of multimodal large language models (MLLMs) through the retrieval of external multimodal knowledge, MRAG system can further enhance question answering capabilities and quality, thereby further reducing hallucination issues. The main differences between text-modal RAG and MRAG lie in retrieval and generation. In the retrieval step, the former only needs to consider retrieving relevant textual knowledge from a large document collection, while the latter needs to consider how to retrieve and integrate the relevant knowledge under different modalities, as well as the relationships between knowledge in different modalities. In the generation step, the former only needs to consider the input text query and relevant textual knowledge, and output a text answer based on the LLM. The latter, however, needs to consider how to utilize the input query from different modalities and multimodal retrieval knowledge, and output an answer that includes information from different modalities based on the MLLM.

Considering the immense potential of MRAG in this field, this survey aims to systematically review and analyze the current state and main challenges of MRAG. We discuss existing research from several key perspectives: 1) *What important components and technologies are involved in MRAG?* 2) *What types of datasets can be used for the evaluation of MRAG?* 3) *What methods and metrics are used to evaluate MRAG?* 4) *What limitations exist in the different aspects of MRAG?* We explore the main challenges faced by MRAG, and hope to provide clearer guidelines for their future development. In summary, the main contributions of this paper are as follows:

- **Comprehensive and Timely Survey:** We conducted an extensive survey on the emerging paradigm of multimodal Retrieval-Augmented Generation, systematically reviewing the current state of research and development in this field.
- **Systematic Analysis from Four Key Perspectives:** Our survey is organized around four key aspects: essential components and technologies, datasets, evaluation methods and metrics, and limitations. This structured approach allows for a detailed understanding of how MRAG can be efficiently constructed, its reliability issues, and how it can be further improved.
- **Current Challenges and Future Research Directions:** We discuss the existing challenges of MRAG, highlight potential research opportunities and directions, and provide a forward-looking perspective on the future development of this paradigm, encouraging researchers to delve deeper into this exciting field.

We have provided an overall introduction to this survey paper. The section 2 presents a comprehensive overview of multimodal retrieval-augmented generation, covering multiple developmental stages. The section 3 delves into the technical details of multimodal retrieval-augmented generation, focusing on key components such as multimodal retrieval, multimodal generation, etc. In section 4, we discuss how to comprehensively evaluate multimodal retrieval-augmented generation systems using datasets, including specialized assessments for different competency areas. The section 5 introduces relevant metrics for evaluating multimodal retrieval-augmented generation systems. In section 6, we outline the current technical challenges associated with multimodal retrieval-augmented generation. In section 7, based on previous investigation of MRAG, we summarize

future work in this field, and provide some suggestions. Finally, we give the conclusion of the paper in section 8.

2 Overview of MRAG

Multimodal Retrieval-Augmented Generation (**MRAG**) represents a significant evolution of the traditional Retrieval-Augmented Generation (**RAG**) framework, building upon its foundational structure while extending its capabilities to process diverse data modalities. While RAG is limited to processing plain text, MRAG integrates multimodal data, including images, audio, video, and text, enabling it to address more complex and diverse real-world applications where information spans multiple modalities.

In the early stages of MRAG development, researchers converted multimodal data into unified textual representations. This approach allowed for a seamless transition from RAG to MRAG by leveraging existing text-based retrieval and generation mechanisms. Although this strategy simplified multimodal data integration and improved the end-to-end user experience, it introduced significant limitations. For instance, the conversion process often resulted in the loss of modality-specific information, such as visual details in images or tonal nuances in audio, restricting the system's ability to fully exploit the potential of multimodal inputs. Subsequent research has focused on addressing these limitations by developing more advanced methods to optimize MRAG systems. These advancements have substantially enhanced MRAG's performance and versatility, achieving state-of-the-art results across various multimodal tasks. This paper categorizes the evolution of MRAG into three distinct stages:

2.1 MRAG1.0

The initial stage of the MRAG framework, commonly termed "pseudo-MRAG", emerged as a straightforward extension of the highly successful RAG paradigm. This stage was rapidly adopted due to its adherence to RAG's core principles, with modifications to support multimodal data. As illustrated in Figure 1, the MRAG1.0 architecture consists of three key components: Document Parsing and Indexing, Retrieval, and Generation.

- **Document Parsing and Indexing:** This component is responsible for processing multimodal documents in formats such as Word, Excel, PDF, and HTML. It extracts textual content using Optical Character Recognition (OCR) or format-specific parsing techniques. A document layout detection model is then utilized to segment the document into structured elements, including titles, paragraphs, images, videos, tables, and footers. For textual content, a chunking strategy is applied to segment or group semantically coherent passages. For multimodal data, specialized models are used to generate captions describing images, videos, and other non-textual elements. These chunks and captions are encoded into vector representations using an embedding model and stored in a vector database. The choice of embedding model is crucial, as it significantly impacts the performance and effectiveness of downstream retrieval tasks.
- **Retrieval:** This component processes user queries by encoding them into vector representations using the same embedding model applied during indexing. The query vectors are then utilized to retrieve the top-k most relevant chunks and captions from the vector database, typically employing cosine similarity as the relevance metric. Duplicate or overlapping information from chunks and captions is merged to create a consolidated set of external knowledge, which is subsequently integrated into the prompt for the generation phase. This ensures the system retrieves contextually relevant information to deliver accurate and informed responses.
- **Generation:** In the Generation phase, the MRAG system synthesizes the user's query and retrieved documents into a coherent prompt. A large language model (LLM) generates a response

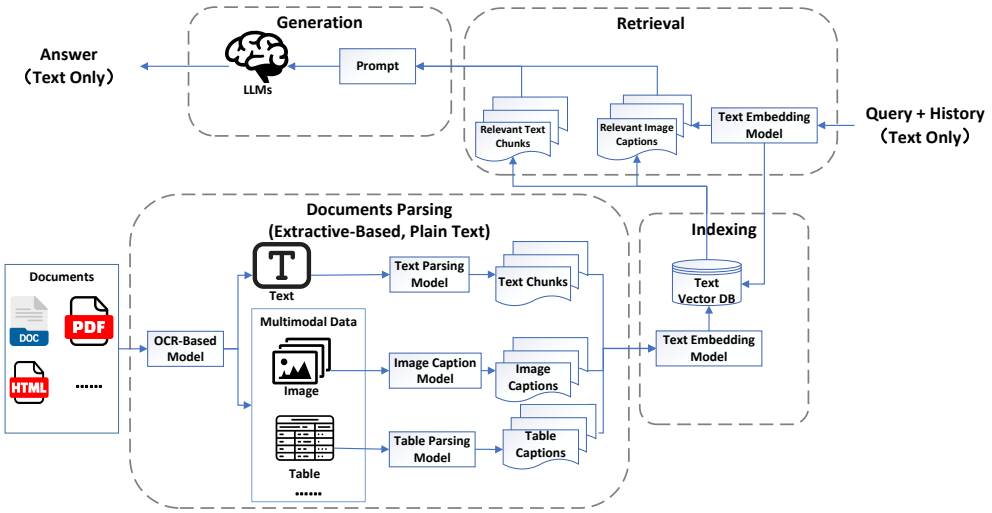

Fig. 1. The architecture of MRAG1.0, often termed "pseudo-MRAG", closely resembles traditional RAG, consisting of three modules: Document Parsing and Indexing, Retrieval, and Generation. While the overall process remains largely unchanged, the key distinction lies in the Document Parsing stage. In this stage, specialized models are employed to convert diverse modal data into modality-specific captions. These captions are then stored alongside textual data for utilization in subsequent stages.

by integrating its parametric knowledge with the retrieved external information. This approach enhances response accuracy and timeliness, particularly in domain-specific contexts, while reducing the risk of hallucinations common in LLM outputs. In multi-turn dialogues, the system incorporates conversational history into the prompt, enabling contextually aware and seamless interactions.

Despite its initial success, MRAG1.0 exhibited several notable limitations that constrained its effectiveness:

- **Cumbersome Document Parsing:** Converting multimodal data into textual captions introduced substantial complexity to the system. This necessitated distinct models for processing different data modalities, increasing both computational overhead and system intricacy. Additionally, the conversion process frequently often to multimodal information loss. For instance, image captions typically provided only coarse-grained descriptions, failing to capture fine-grained details essential for accurate retrieval and generation.
- **Bottleneck of Retrieval:** While text vector retrieval technology is well-established, MRAG1.0 encountered challenges in achieving high recall accuracy. Similar to traditional RAG, the chunking strategy for text segmentation often fragmented keywords, making some content irretrievable. Additionally, transforming multimodal data into text, while enabling non-textual data retrieval, introduced additional information loss. These issues collectively created a bottleneck, limiting the system's ability to retrieve comprehensive and accurate information.
- **Challenges in Generation:** Unlike traditional RAG, MRAG1.0 required processing not only text chunks but also image captions and other multimodal data. Effectively organizing these diverse elements into coherent prompts while minimizing redundancy and preserving relevant information posed a significant challenge. Additionally, the "Garbage In, Garbage Out" (GIGO)

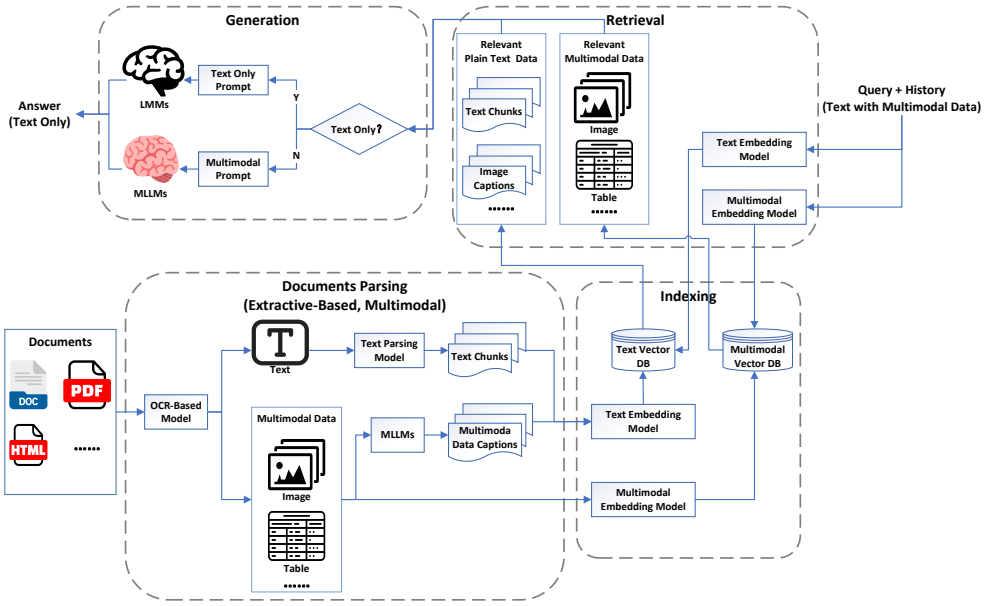

Fig. 2. The architecture of MRAG2.0 retains multimodal data through document parsing and indexing, while introducing multimodal retrieval and MLLMs for answer generation, truly entering the multimodal era.

principle highlighted the sensitivity of LLMs to input quality. Information loss during parsing and retrieval increased the risk of incorporating irrelevant data, compromising the robustness and reliability of the generated responses.

The limitations of MRAG1.0 created a performance ceiling, highlighting the need for more advanced technological solutions. The system's reliance on text-based representations for multimodal data, along with inherent challenges in retrieval and generation, revealed critical gaps in multimodal understanding, retrieval efficiency, and generation robustness. Subsequent iterations of MRAG must address these issues by adopting more sophisticated models, enhancing information retention during parsing, and improving the integration of multimodal data into retrieval and generation processes.

2.2 MRAG2.0

With the rapid evolution of multimodal technologies, MRAG has transitioned into a "true multimodal" era, termed MRAG2.0. Unlike its predecessor MRAG1.0, MRAG2.0 not only supports user queries with multimodal inputs but also preserves the original multimodal data within the knowledge base. By leveraging the capabilities of MLLMs, the generation module can now process multimodal data directly, minimizing information loss during data conversion. As illustrated in Figure 2, the MRAG2.0 architecture incorporates several key optimizations:

- **MLLMs Captions:** The representational capabilities of MLLMs have significantly advanced, especially in captioning tasks. MRAG2.0 leverages a single, unified MLLM—or multiple MLLMs—to extract captions from multimodal documents. This approach replaces the conventional paradigm of using separate models for different modalities, simplifying the document parsing module and reducing its complexity.

- **Multimodal Retrieval:** MRAG2.0 enhances its retrieval module to support multimodal user inputs by preserving original multimodal data and enabling cross-modal retrieval. This allows text-based queries to directly retrieve relevant multimodal data, combining caption-based recall with cross-modal search capabilities. The dual retrieval approach enriches data sources for downstream tasks while minimizing data loss, improving accuracy and robustness for downstream tasks.
- **Multimodal Generation:** To fully leverage original multimodal data, the generation module in MRAG2.0 has been enhanced by integrating MLLMs, enabling the synthesis of user queries and retrieval results into a coherent prompt. When retrieval results are accurate and the input comprises original multimodal data, the generation module mitigates information loss typically associated with modality conversion. This enhancement has significantly improved the accuracy of question-answering (QA) tasks, especially in scenarios involving interrelated multimodal data.

Despite these advancements, MRAG2.0 encounters several emerging challenges: 1) Integrating multimodal data inputs may reduce the accuracy of traditional textual query descriptions. Furthermore, current multimodal retrieval capabilities remain inferior to text-based retrieval, potentially limiting the overall accuracy of the retrieval module. 2) The diversity of data formats presents new challenges for the generation module. Efficiently organizing these diverse data forms and clearly defining inputs for generation are critical areas requiring further exploration and prioritization.

2.3 MRAG3.0

As illustrated in Figure 3, the MRAG3.0 system represents a significant evolution from its predecessors, introducing structural and functional innovations that enhance its capabilities across multiple dimensions. This new paradigm shift is characterized by three key advancements: 1) Enhanced Document Parsing: A novel approach retains document page screenshots during parsing, minimizing information loss in database storage. 2) True End-to-End Multimodality: While earlier versions emphasized multimodal capabilities in knowledge base construction and system input, MRAG3.0 introduces multimodal output capabilities, completing the end-to-end multimodal framework. 3) Scenario Expansion: Moving beyond traditional focus on understanding capabilities—primarily applied in VQA (Visual Question Answering) scenarios reliant on knowledge bases, the new paradigm integrates understanding and generation capabilities through module adjustments and additions. This unification significantly broadens the system's applicability. In the following sections, we will detail the scenarios supported by MRAG3.0 and the specific module modifications enabling these advanced capabilities.

2.3.1 Scenario for MRAG.

- **Retrieval-Augmented Scenario:** This scenario addresses cases where LLMs or MLLMs alone cannot adequately answer user queries. MRAG3.0 retrieves relevant content from external knowledge bases to provide accurate answers, leveraging its enhanced retrieval capabilities.
- **VQA Scenario:** This scenario serves as a critical test for evaluating the fundamental capabilities of MLLMs, which generate responses directly from user inputs containing text and multimodal queries without retrieval. The new MRAG paradigm introduces a search planning module, enabling dynamic routing and retrieval to minimize unnecessary searches and the inclusion of irrelevant information.
- **Multimodal Generation Scenario:** This primarily pertains to multimodal generation tasks, such as text-to-image or text-to-video generation. While the original MRAG framework primarily addressed understanding tasks, the new MRAG paradigm extends its capabilities by modifying multiple generation modules, unifying the solutions for both understanding and generation tasks within a single framework. Following integration, the generation scenarios are further

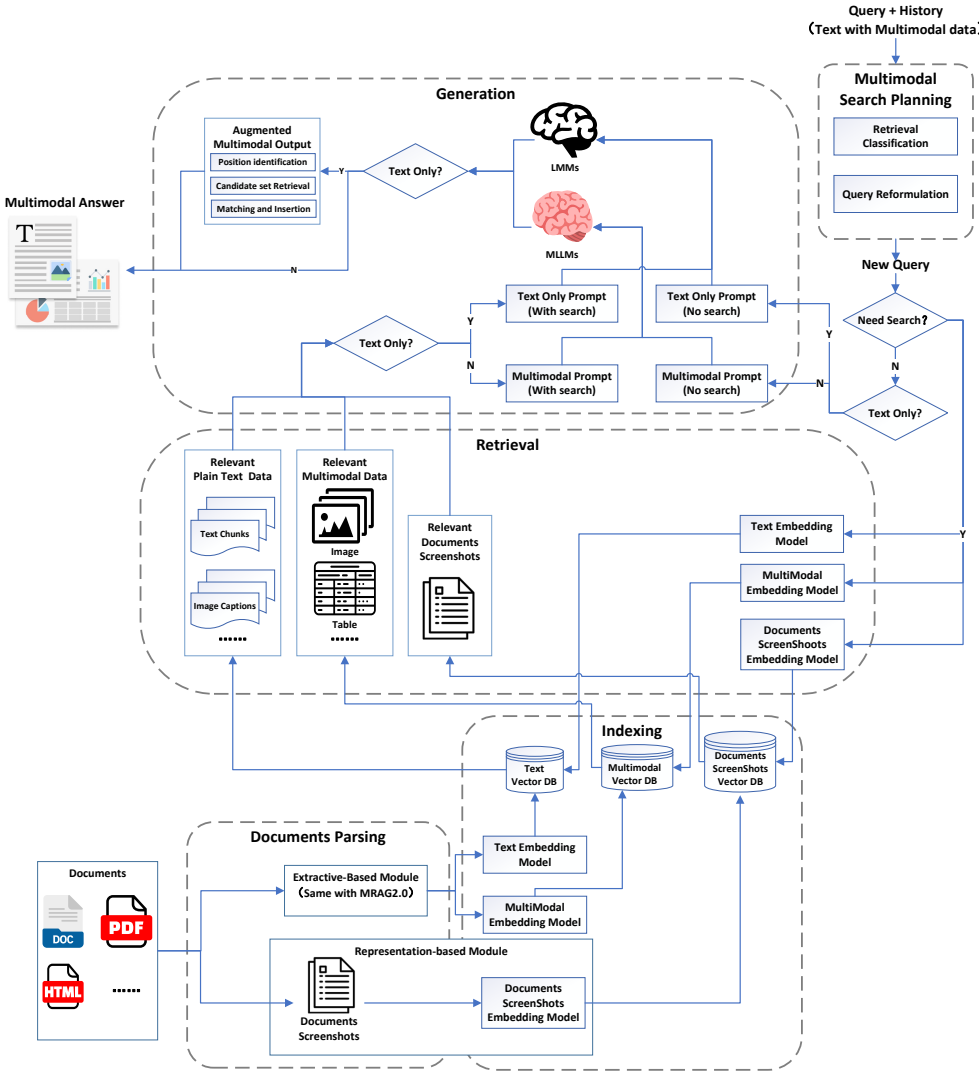

Fig. 3. MRAG3.0 architecture integrates document screenshots during the document parsing and indexing stages to minimize information loss. At the input stage, it incorporates a Multimodal Search Planning module, unifying Visual Question Answering (VQA) and Retrieval-Augmented Generation (RAG) tasks while refining user query precision. At the output stage, the Multimodal Retrieval-Augmented Composition module enhances answer generation by transforming plain text into multimodal formats, thereby enriching information delivery.

enhanced by Retrieval-Augmentation (RA), which significantly improves the overall performance of generation tasks (see Figure 4).

- **Fusion Multimodal Output Scenario:** This scenario is distinct from those previously mentioned but represents a significant aspect of the new paradigm, warranting separate discussion. In traditional settings, the final output is typically a plain text response. However, the new paradigm enhances the generation module to produce outputs that integrate multiple modalities within a

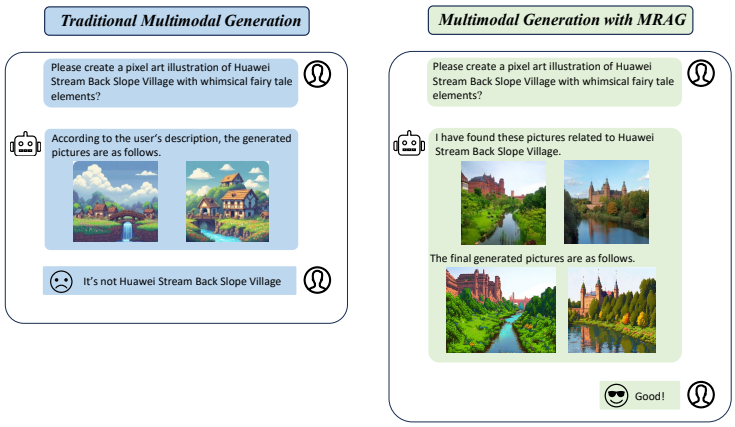

Fig. 4. The user aims to generate images depicting "Huawei Stream Back Slope Village." Due to the location's obscurity and the model's limited knowledge, it may produce inaccurate representations, such as images of houses by a stream. By integrating retrieval-augmented capabilities, the model can access relevant information beforehand, enabling the generation of precise and contextually accurate images.

single response (e.g., combining text, images, or videos). This can be further categorized into three sub-scenarios (see Figure 5).

- **Multimodal Data is Answer:** The query can be answered directly through multimodal data without any text, as the adage "a picture is worth a thousand words" suggests.
- **Multimodal Data Enhances Accuracy:** The integration of multimodal data enhances the accuracy of responses, particularly in instructional contexts such as "How to register for a Gmail account.". By generating answers that interweave text and image, users can more effectively comprehend and follow the required operations.
- **Multimodal Data Enhances Richness:** While multimodal data is not essential, its inclusion can significantly enhance user experience. For instance, when responding to a query such as "Please introduce the Eiffel Tower.", supplementing the textual explanation with relevant images or a brief video can offer users a more engaging and visually enriched experience.

2.3.2 *Modified Modules.*

- **Documents Parsing and Indexing:** To minimize information loss and enhance the accuracy of document retrieval, the document parsing and indexing module has been upgraded with innovative technologies. This new approach preserves document screenshots during parsing, addressing the information loss issues inherent in previous methods. By utilizing fine-tuned MLLMs, the system vectorizes and indexes these document screenshots, enabling efficient retrieval of relevant document screenshots based on user queries. This optimization not only improves the reliability of the knowledge base but also paves the way for advanced multimodal retrieval capabilities.
- **Generation:** Previously, the generation module relied exclusively on large models with understanding capabilities. The new paradigm integrates large models with generation capabilities, unifying reasoning and generation scenarios at the system architecture level. Additionally, by incorporating a multimodal output enhancement submodule, it facilitates a shift from text-based answers to mixed multimodal answers. The implementation methods can be categorized into two types:

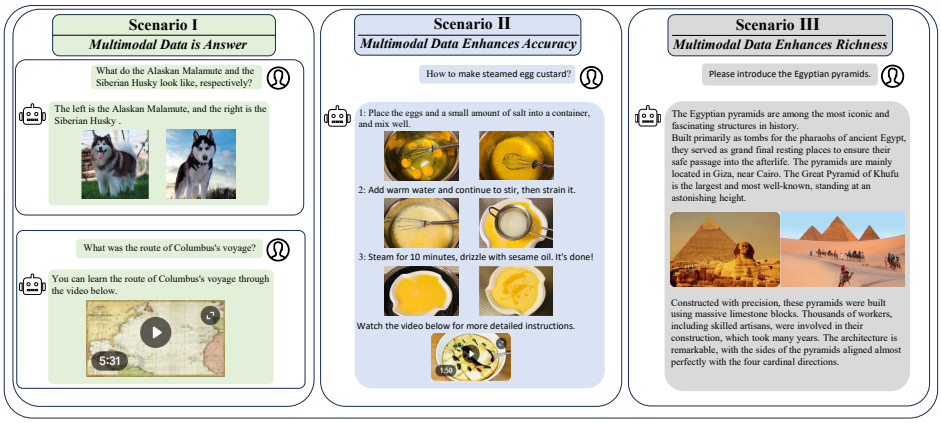

Fig. 5. Multimodal output in QA scenarios can be categorized into three distinct types. In sub-scenario I, the user's query can be fully addressed using only images or videos, without requiring supplementary textual information. Sub-scenario II involves a step-by-step explanation that combines text and images to ensure clarity and precision; omitting the images may lead to user confusion at specific steps. In sub-scenario III, supplementary images enrich the information conveyed in the answer, but their removal does not compromise the answer's accuracy.

- **Native MLLM-Based Output:** In this task, the generation of multimodal data is entirely model-driven, eliminating the need for external data sources to supplement the model responses. The most straightforward approach involves using a unified MLLM to produce the desired multimodal output in a single step, ensuring seamless integration of diverse data types, such as text, images, or audio, within a cohesive framework.
- **Augmented Multimodal Output:** This method utilizes pre-existing multimodal data to enhance textual responses. After generating the text, the system executes three sequential subtasks to create the final multimodal output: 1) Position Identification: The system determines optimal insertion points within the text where multimodal elements (e.g., images, videos, graphs) can be integrated to complement or clarify the content. This step ensures that the multimodal data aligns contextually with the text. 2) Candidate Set Retrieval: Relevant multimodal data is retrieved from external sources, such as the web or a knowledge base, by querying and filtering potential candidates that best match the text's context and intent. 3) Matching and Insertion: The system selects the most appropriate multimodal element from the retrieved candidate set based on relevance, quality, and coherence. The chosen data is then seamlessly integrated into the identified positions, producing a cohesive and enriched multimodal answer.

2.3.3 *New Modules.*

- **Multimodal Search Planning:** This module tackles key decision-making challenges in MRAG systems by focusing on two core tasks: retrieval classification and query reformulation. Given a multimodal query $Q = (q, v)$, where q is the textual component and v is the visual component, the module is designed to optimize information acquisition. Specifically, retrieval classification involves determining the relevance and category of the multimodal query to guide the search toward the most appropriate data sources. Query reformulation, on the other hand, refines the

query by integrating textual and visual cues to improve retrieval accuracy and comprehensiveness. By combining these tasks, the module strengthens the system's ability to handle complex multimodal inputs, ensuring more effective and contextually relevant information retrieval.

– **Retrieval Classification:** This task determines the optimal retrieval strategy a^* from the action space $\mathcal{A} = \{a_{none}, a_{text}, a_{image}\}$ based on the current query and optionally the retrieved historical documents. The decision process is formulated as:

$$a^* = \underset{a \in \mathcal{A}}{argmax} \mathcal{F}_{RC}\left(a \mid Q, \mathcal{D}\right) \tag{1}$$

where the retrieval control module $\mathcal{F}_{RC}$ evaluates the utility of retrieval actions by considering query characteristics, the MLLM's inherent capabilities, and, when available, the retrieved documents $\mathcal{D}$ from previous iterations. For example, in multi-hop scenarios, after retrieving visual information in the initial round, the module may leverage accumulated knowledge to determine subsequent actions, such as text-based retrieval or direct generation. Existing MRAG frameworks typically follow a rigid pipeline with predetermined retrieval actions, which poses significant limitations. Recent studies [125] have shown that compulsive image-to-image retrieval can be counterproductive, as retrieved images may introduce misleading information, degrading MLLM performance. This highlights the necessity of dynamic retrieval strategy selection.

– **Query Reformulation:** In scenarios where external information is required ($a^* \neq a_{none}$) for queries, the task of query reformulation involves generating an enhanced query Q^* by integrating visual information and, when applicable, retrieved documents from previous iterations. This process can be formulated as:

$$Q^* = \mathcal{F}_{QR}\left(Q, \mathcal{D}\right) \tag{2}$$

where $\mathcal{F}_{QR}$ denotes the query enhancement function, which utilizes visual cues and, if available, historical retrieval results to refine the query's precision. This task is particularly critical in real-world human interactions, where queries often rely heavily on visual context and frequently employ anaphoric references. The inherent challenges of visual incompleteness and textual ambiguity pose significant obstacles to retrieving relevant information through straightforward search mechanisms. For complex queries that necessitate multi-hop reasoning, the enhanced query Q^* may be further decomposed into a series of atomic sub-queries $\{q_1^*, ..., q_n^*\}$. Each sub-query is meticulously formulated by considering both textual and visual contexts, as well as the accumulated knowledge from previous iterations, when relevant. This decomposition allows for a more granular and precise retrieval process, addressing the nuanced dependencies and ambiguities present in real-world queries.

This dual-task approach optimizes information acquisition by minimizing unnecessary retrievals while maximizing the relevance of retrieved content. The structured planning framework significantly enhances the MRAG system's ability to gather comprehensive and accurate information, ensuring computational efficiency.

3 Components & Technologies of MRAG

In this section, we will sequentially introduce the details of the five key technical components of MRAG: Multimodal Document Parsing and Indexing (section3.1), Multimodal Search Planning (section 3.2), Multimodal Retrieval (section 3.3), Multimodal Generation (section 3.4)

3.1 Multimodal Document Parsing and Indexing

MRAG systems significantly enhance the reliability and quality of generated answers, by integrating target multimodal knowledge from external multimodal knowledge bases. Target multimodal knowledge can be derived from various granularity in knowledge bases, including localized segments within a single document, cross-segment references within a document, or even cross-document knowledge collections. Thus, how to effectively parse, index, and organize the multimodal documents in external knowledge bases, can largely affect the model's utilization of target multimodal knowledge, thereby determining end-to-end performance. In this section, we first classify documents in multimodal knowledge bases according to their structure, then we provide a detailed introduction to the parsing methods and the evolution of these methods for different types of multimodal documents. Specifically, multimodal documents can be categorized into the following three types:

- **Unstructured Multimodal Data:** refers to various multimodal information that does not have a specific format or schema, such as text, images, videos, and audio. Among the unstructured data, documents with images are widely studied in MRAG. For example, SlideVQA [347] is a typical dataset for visual question-answering task, where all documents are input as images.
- **Semi-structured Multimodal Data:** mainly refers to multimodal information that lacks the rigid schema of traditional relational databases but retains some organizational features, such as PDFs, HTML, XML, and JSON. In such documents, rule-based methods can directly extract structural characteristics. For instance, in HTML, the title can be identified using the <title> tag. A common challenge in processing these documents is that their inherent structure, easily interpretable by humans, is often lost during parsing, resulting in information loss.
- **Structured Multimodal Data:** refers to multimodal information arranged in a predefined format, typically following a fixed schema, such as relational databases and knowledge graphs. The primary challenge in handling such data is formulating an accurate structured query language corresponding to natural language.

In MRAG scenarios, the primary focus is on processing and leveraging unstructured and semi-structured documents. Document parsing methods in MRAG are broadly categorized into two approaches: extraction-based and representation-based. Each approach has distinct advantages and limitations, with the choice depending on task-specific requirements such as scalability or computational efficiency. Extraction-based methods involve a two-step process: first, multimodal information is extracted from documents, and second, the extracted data is parsed and structured for storage and downstream use. In contrast, representation-based methods do not require explicit extraction of multimodal information. Instead, these methods focus on storing document content holistically, often employing representation techniques for document segments. This approach enables a more comprehensive processing of document content.

3.1.1 Extraction-based. Early document parsing solutions were entirely extractive. They evolved gradually from plain text extraction to multimodal data extraction, depending on the type of content being extracted. This subsection will present the process in this sequence.

- **Plain Text Extraction.** In this phase, only textual information from all modal data in the document was extracted. For example, for tables and images, only their textual content was captured. Semi-structured documents, such as PDFs, XML, and HTML, can be parsed directly according to their structural rules. Numerous open-source tools support such capabilities, including pymupdf [3] and pdfminer [2] for PDF parsing, and jsoup [1] for HTML extraction. While this approach enables simple and efficient document parsing, it has limitations: it cannot extract multimodal

information (e.g., text within images) and struggles with complex document formats. Additionally, the parsed results often suffer from significant loss of document structure information.

To enhance document parsing accuracy and address the limitations of rule-based methods in handling complex real-world documents, such as in Visual Document Understanding (VDU) tasks, OCR (Optical Character Recognition)-based approaches have become widely adopted. The traditional OCR-based document parsing pipeline consists of three main stages: text detection, text recognition, and text parsing. Text Detection involves locating and extracting text regions from documents. Early methods primarily relied on Connected Component Analysis (CCA) [26, 363] and Edge Detection algorithms [261]. For more complex layouts, techniques such as Contour Analysis and Stroke Width Transform (SWT) [73, 344] were employed to handle multi-oriented text. With advancements in machine learning, hybrid models combining regression-based object detection frameworks (e.g., Faster R-CNN) with semantic segmentation networks were developed to address arbitrary-shaped text instances [15, 206, 481]. This stage outputs precise bounding boxes or polygon coordinates around text elements, serving as the basis for subsequent processing. Text Recognition converts visual text representations into machine-readable text, playing a critical role in digitizing unstructured data. Its evolution can be divided into three phases: The classical phase relied on handcrafted features [290] and statistical models [22], but faced challenges with fragmented processing and limited robustness. The deep learning phase introduced CNNs (Convolutional Neural Networks) for feature extraction and CTC (Connectionist Temporal Classification)/RNNs (Recurrent Neural Networks) for sequence modeling, with breakthroughs like CRNN enabling unified pipelines and improved accuracy on irregular text. The modern phase leverages transformer architectures [191], achieving global context awareness and robustness to arbitrary-shaped text. Text Parsing reconstructs semantic relationships through three key steps: Layout Analysis segments documents into logical components using rule-based heuristics and graph models based on spatial and typographic cues. Syntactic Parsing extracts structured data from unstructured text using regular expressions and finite-state machines. Post-processing corrects recognition errors through contextual algorithms like language model interpolation (e.g., n-gram models and dictionary lookups). This comprehensive process ensures accurate semantic reconstruction from complex documents.

However, the OCR-dependent approach has critical problems: It is not conducive to parallelization and occupies a large amount of computing resources, besides, errors in the pipeline will propagate downward through the system, affecting the overall performance. In recent years, with the development of Transformer architectures, the aforementioned issues have been effectively addressed. It enhances global context modeling through the self-attention mechanism, significantly improves processing efficiency by leveraging parallel computing, and directly maps images to structured text in an end-to-end training mode. This effectively eliminates the cumulative error issues associated with the multi-stage cascading of traditional OCR systems. LayoutLM [412] uses the BERT architecture as the backbone and adds two new input embeddings: a 2-D position embedding and an image embedding to jointly model interactions between text and layout information across scanned document images. LayoutLMv2 [413] and LayoutLMv3 [133] further propose a new single multimodal framework to model the interaction among text, layout, and image. DocFormer [12] based on the multimodal transformer architecture proposes a novel multimodal attention layer to fuse text, vision, and spatial features in a document, thereby achieving end-to-end document parsing.

- **Multimodal Extraction.** In this phase, the original format of multimodal data is preserved during extraction, allowing downstream tasks to autonomously determine subsequent operations. For semi-structured documents, extraction can be performed similarly using rule-based methods. Relevant multimodal data is identified through specific tags, such as extracting images from

HTML files using the "<img>" tag. However, this approach faces similar challenges to plain text extraction.

The pipeline for multimodal document parsing based on OCR consists of three steps: page segmentation, text recognition, and text parsing. Page segmentation, similar to text detection in plain text extraction, locates and extracts target regions while annotating them with semantic labels (e.g., title, table, footnote). This subtask of semantic segmentation commonly employs CNN-based methods, categorized into region-based, FCN-based, and weakly supervised approaches [111]. Text recognition, similar to plain text extraction, focuses on parsing text data such as titles and page text. Text parsing involves layout analysis and other operations, processing multimodal data according to downstream task requirements. In the era of LLMs, multimodal data is often converted into text for utilization, as seen in models like TableNet [286] for tables and UniChart [265] for charts. This necessitates distinct models for extracting captions from different modalities. With the advancement of MLLMs, there is a trend toward unifying these models into a single MLLM framework, leveraging their robust representation capabilities [173, 227]. Further developments in MLLMs enable the direct retention and input of original multimodal data during generation [312, 437, 474].

3.1.2 Representation-based. Although extractive-based methods have been widely adopted, they suffer from several inherent limitations: (1) The parsing process is time-consuming, involves multiple steps, and requires different models for different document types; (2) Critical information, such as document structure, may be lost during extraction; and (3) Parsing errors can propagate to downstream tasks. Recent advancements in MLLMs [6, 20, 218] have enabled a novel approach that directly uses document screenshots as primary data for metadata indexing, addressing these issues [78, 170, 253, 342, 463]. To capture both global and local information, DSE [253] processes the document screenshot along with its sub-images through a unified encoding framework. Additionally, a late interaction mechanism, inspired by ColBERT [163], has been introduced to improve recall efficiency [78]. However, page-level document splitting may hinder the model's ability to capture full context and inter-part relationships. To address this problem, a holistic document representation method has been proposed [170], which segments large documents into passages within the token limit of MLLMs. Empirical studies reveal a performance gap between multimodal and text-only retrieval, highlighting differences in effectiveness when using raw multimodal data versus text or combined modalities [312, 463]. Consequently, a new paradigm has emerged that leverages OCR for text indexing, document screenshots for multimodal indexing, and executes textual and visual RAG in parallel. The results from both streams are then fused through modality integration to produce the final answer [342].

3.2 Multimodal Search Planning

Multimodal search planning refers to the strategies employed by MRAG systems, to effectively retrieve and integrate information from multiple modalities to address complex queries. The planning can be broadly categorized into two main approaches: fixed planning and adaptive planning.

3.2.1 Fixed Planning. Early MRAG systems typically adopt fixed planning strategies for handling multimodal queries, characterized by predetermined processing pipelines that lack flexibility in adapting to diverse query requirements. These approaches can be broadly categorized based on their retrieval modality choices:

- **Planning for Single-modal Retrieval.** Early fixed planning strategies usually focus on a single modality for retrieval, despite the multimodal nature of input queries. These approaches can be

broadly classified into text-centric and image-centric paradigms, reflecting initial efforts to adapt traditional IR query processing techniques [171, 184] to the multimodal domain.

- **Text-centric** planning approaches prioritize textual retrieval by transforming multimodal queries into text-only formats. For instance, Plug-and-Play [357] employs vision-language models to convert the visual component of a query into textual descriptions, followed by text-based retrieval planning. This strategy simplifies the multimodal problem into a traditional text-based RAG pipeline, leveraging established multi-stage query processing techniques from conventional IR systems. However, this approach often introduces a semantic gap between the user's original intent and the generated textual descriptions. The conversion of visual queries to text may fail to precisely capture the user's specific information needs, leading to the retrieval of irrelevant or noisy documents that diverge from the query's focus.
- **Image-centric** planning strategies rely solely on image-based retrieval regardless of the query characteristics. Systems such as Wiki-LLaVA [24] demonstrate this paradigm by consistently triggering image retrieval from knowledge bases for multimodal queries. While this approach ensures visual information preservation, it presents practical limitations. Recent empirical studies [125] highlight that compulsive image retrieval can be counterproductive, particularly when textual information suffices or when retrieved images introduce misleading visual contexts, impairing MLLM performance.

The inflexibility of single-modality planning strategies highlights their inherent limitations: they cannot adapt to the diverse information needs of real-world scenarios. For example, while a text-centric approach may be suitable for queries referencing visual content but focused on factual information, an image-centric strategy is more effective for queries requiring detailed visual comparisons.

- **Planning for Multimodal Retrieval.** Recent studies have begun investigating the use of multimodal information retrieval to enhance the performance of MRAG systems. Unlike single-modality approaches, these methods integrate both textual and visual knowledge sources, albeit through fixed processing pipelines. For instance, MMSearch [147] employs a rigid multimodal planning pipeline, mandating Google Lens image searches for all image-containing queries. This is followed by a "Requery" phase, where LLMs reformulate the search query using the original query, image, and Google Lens results. While this structured approach ensures systematic information retrieval, its inflexible design often leads to unnecessary image searches, increasing computational overhead when visual information is irrelevant to the query.

Fixed pipeline approaches, whether single-modality or multimodality, exhibit several critical limitations. First, their rigid retrieval strategies struggle to adapt to the diverse nature of real-world queries, where the optimal retrieval modality depends on specific information needs. Second, mandatory retrieval operations often introduce redundant or irrelevant information, particularly when certain knowledge types are unnecessary for addressing the query. Third, these approaches incur significant computational overhead, especially in multimodal pipelines handling large-scale knowledge bases. As highlighted by mR^{2}AG [474], a more fundamental issue is that not all queries require external knowledge retrieval. Current MRAG systems frequently perform retrieval indiscriminately, resulting in unnecessary computational costs and potential noise in response generation. These limitations emphasize the need to transition from predetermined pipelines to adaptive planning mechanisms that dynamically adjust retrieval strategies based on query characteristics and intermediate results.

3.2.2 Adaptive Planning. Recent studies have highlighted two key limitations in fixed pipeline approaches [197]: 1) Non-adaptive Retrieval Queries: inflexible retrieval strategies that fail to adjust to evolving contexts or intermediate results; and 2) Overloaded Retrieval Queries: concatenating

visual content descriptions with input questions into a single query, leading to ambiguous retrievals and irrelevant knowledge. To address these issues, OmniSearch [197] introduces a self-adaptive planning agent for multimodal retrieval, mimicking human problem-solving behavior. Instead of relying on a fixed pipeline, the system dynamically breaks down complex multimodal questions into sub-question chains with retrieval actions. At each step, the agent adapts its next action based on the problem-solving state and retrieved content, enabling deeper understanding of retrieved information and adaptive refinement of retrieval strategies. CogPlanner [440] iteratively refines queries and selects retrieval strategies, enabling both parallel and sequential modeling approaches.

3.3 Multimodal Retrieval

In this section, we present a comprehensive overview of the three critical components of multimodal retrieval in the MRAG system: retriever (section 3.3.1), reranker, and refiner. Each component plays a distinct yet interconnected role in enhancing the quality and relevance of information retrieval and utilization for LLMs. We summarize the taxonomy of multimodal retrieval research in Figure 6.

3.3.1 RETRIEVER. The retriever is a core component that sources relevant documents from a large external knowledge base using advanced indexing and search algorithms. It retrieves candidate information aligned with user queries, aiming to provide a broad yet relevant set of documents to support high-quality LLM responses. Its performance is crucial, as it directly influences the quality of the downstream retrieval pipeline. As shown in Figure 7, existing retrieval methods fall into two categories based on architecture: Single/Dual-stream Structure and Generative Structure, each involves single-modal (e.g., text, images) and cross-modal information retrieval.

- **Single/Dual-stream Structure**: The Single-stream Structure integrates multimodal fusion modules to model image-text relationships in a unified semantic space, capturing fine-grained interactions but incurring higher computational costs and slower inference, limiting scalability for large-scale multimodal retrieval tasks in real-world applications. In contrast, the Dual-stream Structure uses separate vision and language streams, leveraging contrastive learning to align global features in a shared semantic space efficiently. However, it lacks explicit multimodal interaction and struggles with feature alignment due to information imbalance, exacerbated by the brevity of dataset captions.
  - **Retrieval for Single-Modal.** In MRAG systems, single-modal retrieval focuses on text and image retrieval. Text retrieval uses NLP techniques to extract relevant information from datasets, identifying contextually aligned documents. Image retrieval employs computer vision algorithms and feature extraction methods to encode visual data into high-dimensional vectors for similarity matching. Both modalities are essential for enhancing MRAG system performance.
    - **Text-centric.** Text retrieval, a core component of information retrieval (IR), identifies relevant textual information from large corpora or web resources in response to user queries. It is widely used in downstream applications such as question answering [161, 304], dialogue systems [334, 436, 456], web search [92, 269, 270], and retrieval-augmented generation systems [33, 45, 102, 330]. Recent advancements categorize text retrieval methods into two types: sparse retrieval and dense retrieval.
      - **Sparse Text Retrieval.** Early research in text retrieval focused on extracting representative terms from documents, leading to the development of vector space models [320] based on the "bag-of-words" assumption, which represents documents and queries as sparse term vectors, ignoring term order. Term weighting methods like tf-idf [9, 314, 319] and BM25 models [315, 316] were introduced to assign weights based on term importance within and across corpora, while inverted indexes [513] improved retrieval efficiency by organizing corpora into term-document ID pairs. Statistical language modeling [452]

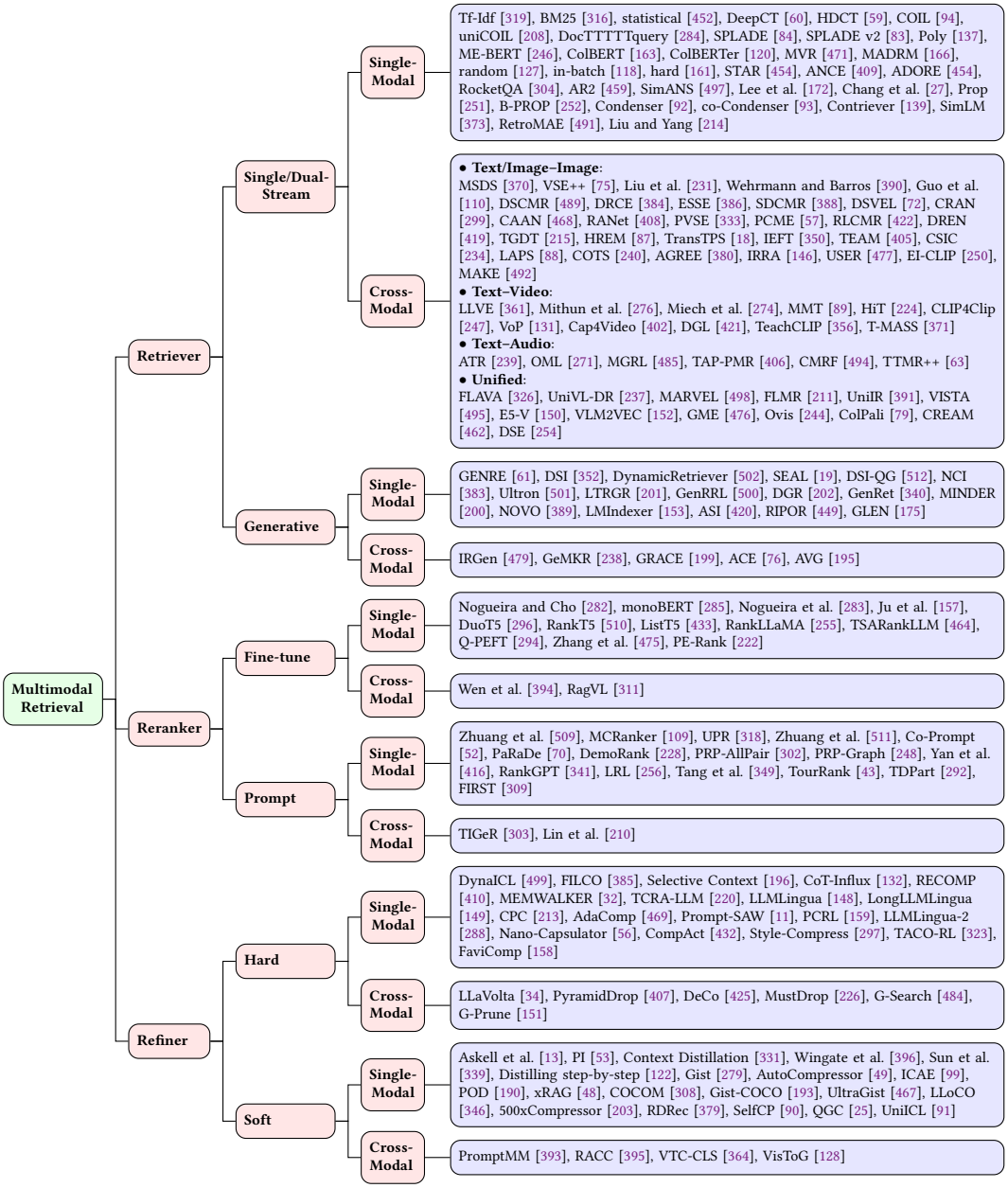

Fig. 6. Taxonomy of recent advancements in multimodal retrieval research.

further advanced retrieval by estimating term probability distributions for probabilistic ranking. However, early sparse retrieval methods face limitations, such as assuming term independence and relying on lexical matching, which hinders their ability to capture contextual term importance or semantic relationships between terms. Consequently, these

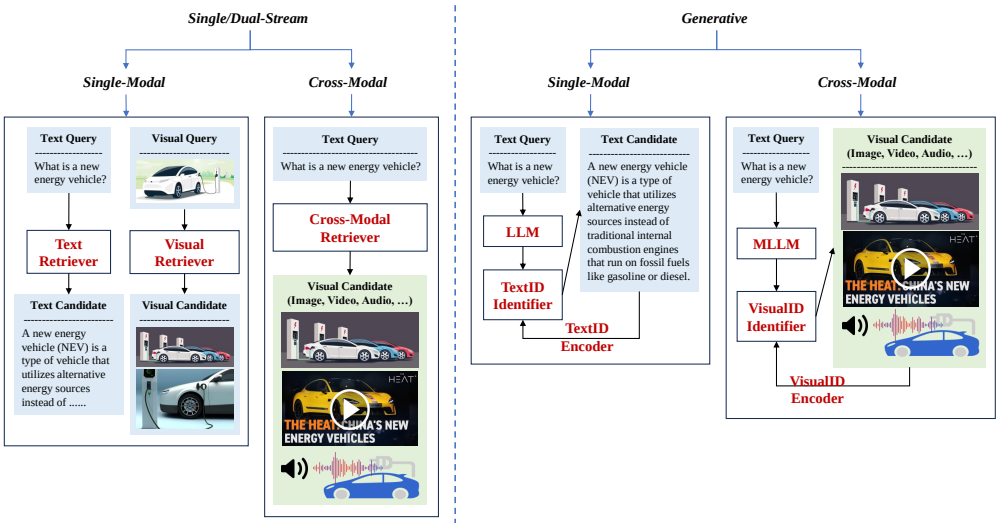

Fig. 7. The architectures of retriever in multimodal retrieval.

methods struggle to understand deeper textual meanings and contextual relevance between queries and documents.

Recent advancements in sparse retrieval models have been driven by the integration of pre-trained language models (PLMs). While these approaches leverage PLMs, they remain fundamentally rooted in lexical matching, enabling the reuse of traditional sparse index structures by incorporating auxiliary information such as contextualized embeddings [94, 208] and extended tokens [83, 84, 284]. This research domain focuses on two main approaches: term weighting and term expansion. Term weighting enhances relevance estimation by leveraging context-specific token representations. DeepCT [60] and HDCT [59] use learned token representations to estimate the context-specific importance of terms within passages, while COIL [94] and uniCOIL [208] employ contextualized token representations of exact matching terms to compute relevance via dot products and summed similarity scores. Term expansion mitigates vocabulary mismatch by expanding queries or documents using PLMs. For instance, DocTTTTTquery [284] predicts relevant queries for documents to enrich the document's content, while SPLADE [84] and SPLADEv2 [83] project terms onto vocabulary-sized weight vectors derived from masked language model logits. These vectors, aggregated via methods like summing or max pooling, effectively expand content by incorporating absent terms. Sparsity regularization ensures efficient sparse representations for inverted index usage.

In summary, sparse retrieval models achieve an optimal balance in cross-domain transfer, retrieval efficiency, and overall effectiveness.

- **Dense Text Retrieval.** Recent advancements in deep learning [5, 50, 51, 105, 119, 167, 169], particularly pre-trained language models (PLMs) [23, 62, 229] based on the Transformer architecture [80, 362], have increasingly adopted dense vector embeddings in low-dimensional Euclidean spaces for modeling semantic relationships between queries and documents. These embeddings enable relevance measurement through Euclidean distances or inner products. Dense retrieval methods have demonstrated strong performance

across various information retrieval tasks [161, 163, 270]. Additionally, Approximate Nearest Neighbor Search (ANNS) algorithms [98, 142, 156], particularly quantization-based methods [98, 142] and their retrieval-oriented variants [415, 453, 455, 461], enable efficient retrieval of top-ranked documents from large collections using precomputed ANNS indices. Dense retrieval techniques primarily focus on two key aspects: model architecture and training methods.

For model architecture, dense retrieval methods employ a two-tower architecture to balance retrieval efficiency and effectiveness by modeling semantic interactions between queries and documents through their representations. These methods vary in representation granularity, primarily falling into two categories: single-vector and multi-vector representations. Then, the relevance scores are computed using similarity functions (e.g., cosine similarity, inner product) between these embeddings. A common technique involves placing a special token (e.g., "[CLS]") at the beginning of a text sequence, with its learned representation capturing the overall semantics. The existing dense retrieval models learn the query and document representations by fine-tuning PLMs like BERT [62], RoBERTa [229], or Mamba Gu and Dao [106], Zhang et al. [458], or large language models (LLMs) like RepLLaMA [255] on annotated datasets (e.g., MSMARCO [281], BEIR [354]). However, single-vector bi-encoders struggle to model fine-grained semantic interactions between queries and documents. To address this limitation, multi-vector representation enhance text representation and semantic interaction by employing multiple-representation bi-encoders. The Poly-encoder [137] generates multiple context codes to capture text semantics from multiple views. ME-BERT [246] produces m representations for a candidate text using the contextualized embeddings of the first m tokens. ColBERT [163] maintains per-token contextualized embeddings with a late interaction mechanism. ColBERTer [120] extends ColBERT by combining single- ("[CLS]") and multi-representation (per-token) mechanisms for better performance. MVR [471] introduces multiple "[VIEW]" tokens to learn diverse representations, with a local loss to identify the best-matched view. MADRM [166] learns multiple aspect embeddings for queries and texts, supervised by explicit aspect annotations.

For training method, to achieve optimal retrieval performance, dense retrieval models are typically trained using two key techniques: negative sampling and pretraining. Negative sampling focuses on selecting high-quality negatives to compute the negative log-likelihood loss used for training dense retrieval models. Basic methods include random sampling [127] and in-batch negatives [118, 161, 304], which increase the number of negatives within memory limits but do not guarantee the inclusion of hard negatives, i.e., irrelevant texts with high semantic similarity to the query. Hard negatives are critical for improving the model's ability to distinguish relevant from irrelevant texts. Various approaches have been proposed to incorporate hard negatives. BM25-retrieved documents are used as static hard negatives [95, 161]. STAR [454] combines static hard negatives with random negatives, while ANCE [409] retrieves hard negatives using a warm-up dense retrieval model and refreshes the document index during training. ADORE [454] employs an adaptive query encoder to retrieve top-ranked texts as hard negatives, keeping the text encoder and document index fixed. However, hard negatives may include false negatives, introducing noise that can degrade performance. RocketQA [304] addresses this by using a cross-encoder to filter out likely false negatives. AR2 [459] integrates a dual-encoder retriever with a cross-encoder ranker, jointly optimized through a minimax adversarial objective to produce harder negatives and improve the retriever. SimANS [497] introduces the concept of sampling ambiguous negatives, i.e., texts ranked near positives

with moderate similarity to the query. These negatives are more informative and less likely to be false negatives, further enhancing model performance.

Pretraining aims to learn universal semantic representations that generalize to downstream dense retrieval tasks. To enhance the modeling capacity of PLMs, self-supervised pretraining tasks, such as those proposed by Lee et al. [172] (selecting random sentences as queries) and Chang et al. [27] (leveraging hyperlinks for constructing query-passage pairs), mimic retrieval objectives. Prop [251] and B-PROP [252] use document language models (e.g., unigram, BERT) to sample word sets, training PLMs to predict pairwise preferences. To enhance dense retrieval models, studies focus on improving the "[CLS]" token embedding. Condenser [92] aggregates global text information for masked token recovery, while co-Condenser [93] adds a query-agnostic contrastive loss to cluster related text segments while distancing unrelated ones. Contriever [139] generates positive pairs by sampling two spans from the same text and negatives using in-batch and cross-batch texts. Following with an unbalanced architecture (strong encoder, simple decoder), SimLM [373] pretrains the encoder and decoder with replaced language modeling, recovering original tokens after replacement. It further optimizes the retriever through hard negative training and cross-encoder distillation. RetroMAE [491] utilizes a high masking ratio for the decoder and a standard ratio for the encoder, incorporating an enhanced decoding mechanism with two-stream and position-specific attention masks. Liu and Yang [214] introduces a two-stage pretraining approach, combining general-corpus pretraining with domain-specific continual pretraining, achieving strong benchmark performance.

However, single-modal retrieval is inherently limited by its inability to capture cross-modal relationships, which underscores the importance of integrating multimodal retrieval strategies to bridge textual and visual semantics for more comprehensive information retrieval and generation.

- **Retrieval for Cross-modal.** Cross-modal retrieval enables the identification of relevant data in one modality (e.g., images) using a query from another (e.g., text). It enhances MRAG systems by facilitating the retrieval and generation of information across diverse modalities, including text, images, audio, and video.
  - **Text–Image Retrieval.** Text–Image Retrieval aims to match images with corresponding textual queries by leveraging multimodal data co-occurrence, such as paired text-image instances or manual annotations, to capture semantic correlations. Existing methods can be categorized into three groups: CNN/RNN-based approaches, Transformer-based techniques, and Vision-Language Pretraining (VLP) model-based methods.

    Early CNN/RNN-based methods [75, 110, 174, 231, 370, 390] extract features from each modality separately using MLP, CNN, and RNN, enforcing cross-modal constraints through positive/negative sample construction. MSDS [370] uses CNN with a maximum likelihood-based scheme for image-text relevance. VSE++ [75] combines CNN and RNN with hard sample mining in ranking loss. Advances include residual learning [231], character-level convolution [390], and disentangled representation [110] for improved feature mapping and retrieval. DSCMR [489] maps multimodal data into a shared space using modality-specific networks and fully connected layers, leveraging label constraints and pairwise loss for discriminant learning. Recent CNN/RNN-based methods improve image-text matching by addressing key challenges. DRCE [384] enhances rare content representation and association using a dual-path structure, adaptive fusion, and reranking to mitigate long-tail issues. ESSE [386] tackles one-to-many correspondence by projecting data as sectors with uncertainty apertures. SDCMR [388] employs diverse CNNs for multimodal feature extraction and a dual adversarial mechanism to isolate semantic-shared features, ensuring retrieval consistency.

These methods collectively advance cross-modal retrieval robustness and accuracy. Spatial attention [72, 135, 299, 408, 468] is widely used in CNN/RNN-based cross-modal retrieval to uncover fine-grained associations by generating weighted masks for local regions, enhancing key features while suppressing irrelevant ones. DSVEL [72] employs spatial-aware pooling to align image regions with text, while CRAN [299] and CAAN [468] improve global-local alignment through relation alignment and context-aware selection. RANet [408] refines attention mechanisms with reference attention to reduce incorrect scores and adaptive aggregation to amplify relevant information and minimize redundancy.

Transformer-based methods [18, 57, 87, 215, 333, 350, 419, 422, 451] leverage multi-head self-attention to encode multimodal relationships and optimize modality-specific encoders, demonstrating superior performance in multimodal modeling and cross-modal retrieval tasks. Recent advancements in multimodal representation learning have focused on enhancing Transformer architectures and feature alignment. PVSE [333] integrates self-attention and residual learning, while PCME [57] uses probabilistic embeddings to model one-to-many and many-to-many correlations. RLCMR [422] tokenizes multimodal data and trains with a unified Transformer encoder for cross-modal semantic correlation. DREN [419] refines feature representation through character-level and context-driven augmentation. TGDT [215] unifies coarse- and fine-grained learning with multimodal contrastive loss for feature alignment. HREM [87] improves image-text matching by capturing multi-level intra- and inter-modal relationships. TransTPS [18] extends Transformers with cross-modal multi-granularity matching and contrastive loss for better feature distinction. IEFT [350] models text-image pairs as unified entities to model their intrinsic correlation.

With the rapid advancement of pretraining paradigms, Vision-Language Pretraining (VLP) models [46, 62, 68, 107, 136, 144, 183, 194, 305], including both single- and dual-stream architectures, have leveraged large-scale visual-linguistic datasets for joint pretraining. Researchers have utilized the strong representational capabilities [88, 146, 234, 240, 250, 380, 405, 477, 492] of VLP models to significantly enhance cross-modal retrieval performance. Single-stream models like TEAM [405] align multimodal token embeddings for token-level matching, while dual-stream approaches such as COTS [240] integrate contrastive learning with token- and task-level interactions. Methods like CSIC [234] and LAPS [88] improve multimodal alignment by quantifying semantic significance and associating patch features with words, respectively. AGREE [380] fine-tunes and reranks cross-modal entities to harmonize their alignment. IRRA [146] employs text-specific mask mechanism to capture fine-grained intra- and inter-modal relationships. USER [477], EI-CLIP [250], and MAKE [492] leverage CLIP [305] or ALIGN [144] to integrate contrastive learning and keyword enhancement for enriching representations. Overall, VLP models, through strategies such as fine-tuning, reranking, and follow-up training, have become essential for improving cross-modal alignment and interaction.

* **Text–Video Retrieval.** Text-video retrieval involves matching textual descriptions with corresponding videos, requiring spatiotemporal representations to address temporal dynamics, scene transitions, and precise text-video alignment. This task is more complex than text-image retrieval due to the need to model both visual and sequential information effectively.

Early CNN/RNN-based methods [64, 273, 274, 276, 361, 441] encode videos and texts into a shared latent space for similarity measurement. LLVE [361] employs CNNs and LSTMs to extract latent features from images and texts, with LSTMs further capturing temporal relationships between video frames. Subsequent studies [274, 276] apply mean/max pooling to frame sequences to generate compact video-level representations, prioritizing efficiency

over granularity. Later advancements incorporate additional modalities, such as audio and motion, to enhance video semantics [273]. For text encoding, simpler methods like Word2Vec, LSTMs, or GRUs are commonly used [274, 276, 441], with evidence suggesting that combining multiple text encoding strategies improves retrieval performance [64].

Transformer-based methods [89, 224] utilize self-attention mechanisms to jointly encode videos and texts, enabling cross-modal interaction. MMT [89] employs mutual attention between video and text modalities, integrating temporal information to enhance feature representation. Inspired by MoCo [116], HiT [224] introduces hierarchical cross-modal contrastive matching at both feature and semantic levels. Additionally, these methods [89, 224] encode diverse modalities in video data, such as audio and motion, further enriching video representations.

Recently, VLP-based models [131, 247, 356, 371, 402, 421] utilize pretrained models like CLIP [305] to enhance text-video tasks in text-video retrieval tasks by capturing cross-modal and temporal dependencies. CLIP4Clip [247] adapt CLIP for text-video retrieval and captioning, analyzing temporal dependencies. VoP [131] introduces prompt tuning and fine-tunes CLIP to model spatiotemporal video aspects, while Cap4Video [402] leverages zero-shot captioning with CLIP and GPT-2 [306] for auxiliary captions. DGL [421] proposes dynamic global-local prompt tuning, emphasizing intermodal interaction and global video information through shared latent spaces and attention mechanisms. TeachCLIP [356] improves CLIP4Clip by integrating fine-grained cross-modal knowledge from advanced models, and refining text-video similarity with an frame-feature aggregation block. T-MASS [371] addresses dataset limitations by enriching text embeddings with stochastic text modeling.

* **Text–Audio Retrieval.** Text-audio retrieval involves matching textual queries with corresponding audio content, requiring alignment of semantic text information with dynamic acoustic patterns in speech, music, or environmental sounds. The challenge lies in bridging the gap between discrete text and continuous audio signals.

  Early CNN/RNN-based approaches [239, 271, 485] focus on encoding text and audio separately and aligning them in a shared space for similarity measurement. ATR [239] uses pretrained CNN-based audio networks with NetRVLAD pooling [143] to aggregate features into a unified representation. OML [271] employs CNNs for robust audio feature extraction and metric learning to enhance audio-text alignment. MGRL [485] leverages CNNs for localized audio features and introduces adaptive aggregation to handle varying text–audio granularities.

  Furthermore, Transformer-based methods [63, 406, 494] utilize multi-head attention mechanisms and fine-tuning to enhance cross-modal interactions. TAP-PMR [406] employs scaled dot-product attention to enable text to focus on relevant audio frames, reducing misleading information, while its prior matrix revised loss optimizes dual matching by addressing similarity inconsistencies. CMRF [494] enhances audio-lyrics retrieval through directional cross-modal attention and reinforcement learning to refine multimodal embeddings and interactions. TTMR++ [63] integrates fine-tuned LLMs and rich metadata to generate detailed text descriptions, improving retrieval by addressing musical attributes and user preferences.

* **Unified-Modal Retrieval.** Unified-Modal Retrieval aims to process diverse hybrid-modal data (e.g., text, images, videos) within a unified model architecture, such as transformer-based PLMs, to encode all modalities into a shared feature space. This enables efficient cross-modal retrieval between any pairwise combination of hybrid-modal data. With the growing demand for multimodal applications, there is an increasing need for unified multimodal retrieval models tailored to complex scenarios. Current approaches leverage pre-trained models like CLIP [305], BLIP [186], and ALIGN [144] for multimodal embedding. For instance, FLAVA

[326] integrates multiple modalities into a unified framework, leveraging joint pretraining on multimodal data with cross-modal alignment and fusion objectives. Similarly, UniVL-DR [237] encodes queries and multimodal resources into a shared embedding space, employing a universal embedding optimization strategy with modality-balanced hard negatives and an image verbalization method to bridge the gap between images and texts. MARVEL [498] addresses the modality gap between images and texts by incorporating visual features into the encoding process. FLMR [211] enhances image representations by using a visual model aligned with existing text-based retrievers to supplement the image representation of image-to-text transforms. UniIR [391] introduces a unified instruction-guided multimodal retriever, achieving robust generalization through instruction tuning on diverse multimodal-IR tasks. VISTA [495] extends image understanding capability by integrating visual token embeddings into a text encoder, supported by high-quality composed image-text data and a multi-stage training algorithm. E5-V [150] fine-tunes MLLMs on single-text or vision-centric relevance data, outperforming traditional image-text pair training. VLM2VEC [152] proposes a contrastive training framework to convert vision-language models into embedding models using the MMEB dataset [152]. To address modality imbalance, GME [476] trains an MLLM-based dense retriever on the large-scale UMRB dataset[476]. Ovis [244] aligns visual and textual embeddings by integrating a learnable visual embedding table, enabling probabilistic combinations of indexed embeddings for rich visual semantics. ColPali [79] leverages Vision Language Models and the ViDoRe benchmark [79] to index documents from their visual features, facilitating efficient query matching with late interaction mechanisms. CREAM [462] employs a coarse-to-fine retrieval and ranking approach, combining similarity calculations with large language model-based grouping and attention pooling for MLLM-based multi-page document processing. DSE [254] fine-tunes a large vision-language model on 1.3 million Wikipedia web page screenshots, enabling direct encoding of document screenshots into dense representations.

- **Generative Structure**: Traditional information retrieval (IR) methods, which rely on similarity matching to return ranked lists of documents, have long been a cornerstone of information acquisition, dominating the field for decades. However, with the advent of pre-trained language models, generative retrieval (GR) has emerged as a novel paradigm, garnering increasing attention in recent years. GR primarily consists of two fundamental components: model training and document identifier. Model Training aims to train generative models to effectively index and retrieve documents, while enhancing the model's capacity to memorize information from the document corpus. This is typically achieved through sequence-to-sequence (seq2seq) training, where the model learns to map queries to their corresponding Document Identifiers (DocIDs). The training process emphasizes optimizing the model's understanding of semantic relationships between queries and documents, thereby improving retrieval accuracy. Document Identifiers (DocIDs) serve as the target output for the generative retrieval model, and unique representations of each document in the corpus. The quality of these identifiers is crucial, as they directly impact the model's ability to memorize and retrieve document information. Effective DocIDs are often generated using dense, low-dimensional embeddings or structured representations that capture the essential content and context of documents, enabling the model to distinguish between documents more accurately and enhancing retrieval performance. By overcoming the limitations of traditional IR in terms of content granularity and relevance matching, GR offers enhanced flexibility, efficiency, and creativity, better aligning with practical demands.
  - **Retrieval for Text-modal.** The recent advancements in generative language models have demonstrated their ability to memorize knowledge from documents and recall knowledge to respond to user queries effectively, which focuses on the use of document identifiers (DocIDs)

and their optimization for retrieval tasks. The approaches can be categorized into static DocID-based methods and learnable DocID-based methods.

Static DocID-based methods rely on pre-defined, fixed document identifiers. They often use unique names, numeric formats, or structured identifiers to represent documents. GENRE [61] generates entity names via constrained beam search using a prefix tree, with document titles serving as DocIDs. DSI [352] introduces numeric DocID formats, including unstructured, naively structured, and semantically structured identifiers, trained through indexing and retrieval strategies. DynamicRetriever [502] uses unstructured atomic DocIDs and enhances memorization with pseudo queries. SEAL [19] representing documents with N-gram sub-string identifiers, leveraging FM-Index [82] for retrieval. DSI-QG [512] represents documents with generated queries, re-ranked by a cross-encoder. NCI [383] generates document identifiers using a seq2seq network with a prefix-aware decoder. It is trained on both labeled and augmented pseudo query-document pairs. Ultron [501] combines URLs and titles as DocIDs to uniquely identify web documents. It encodes documents into a latent semantic space using BERT [62] and compresses vectors via Product Quantization (PQ) [98, 142], with PQ codes serving as semantic identifiers. Additional digits ensure DocID uniqueness. LTRGR [201] focuses on learning to rank passages directly using generative retrieval models, optimizing autoregressive models via rank loss. GenRRL [500] integrates reinforcement learning for aligning token-level DocID generation with document-level relevance estimation. DGR [202] enhances generative retrieval through knowledge distillation, using a cross-encoder as a teacher model to provide fine-grained ranking supervision. Despite these innovations, most approaches rely on static DocIDs, which are not optimized for retrieval tasks, limiting their ability to capture document semantics and relationships, thereby hindering retrieval performance.

To address this limitation, Learnable DocID-based methods introduce learnable document representations, where DocIDs are optimized during training to better capture document semantics and improve retrieval performance. GenRet [340] employs a discrete autoencoder to encode documents into compact DocIDs, minimizing reconstruction error. MINDER [200] enhances document representations using multi-view identifiers, including pseudo-queries, titles, and sub-strings. NOVO [389] introduces learnable continuous N-gram DocIDs, refining embeddings through query denoising and retrieval tasks. LMIndexer [153] generates neural sequential discrete IDs via progressive training and contrastive learning, addressing semantic mismatches. ASI [420] automates DocID learning, assigning similar IDs to semantically close documents and optimizing end-to-end retrieval using an generative model. RIPOR [449] improves relevance scoring during sequential DocID generation using dense encoding and Residual Quantization [264]. GLEN [175] employs a dynamic lexical identifier with a two-phase index learning strategy. Firstly, the keyword-based DocID are defined by extracting keywords from documents using self-supervised signals. Secondly, dynamic DocIDs are refined by integrating query-document relevance, enabling efficient inference. The field of generative text retrieval is evolving from static, pre-defined DocIDs to dynamic, learnable DocIDs that better capture document semantics and relationships. Learnable DocIDs, combined with advanced techniques like reinforcement learning, knowledge distillation, and contrastive learning, are driving improvements in retrieval performance.

– **Retrieval for Cross-modal.** Similarly, MLLMs are considered to memorize and retrieve multimodal content, such as images and videos, within their parameters. When presented with a user query for visual content, the MLLM is expected to "recall" the relevant image from its parameters as a response. Achieving this capability presents significant challenges, particularly in developing effective visual memory and recall mechanisms within MLLMs. IRGen [479] employs a seq2seq model to predict discrete visual tokens (image identifiers) from

query images. Its key innovation is a semantic image tokenizer that encodes global features into discrete visual tokens, enabling end-to-end differentiable search for improved accuracy and efficiency. GeMKR [238] integrates LLMs with visual-text features through a generative multimodal knowledge retrieval framework. It first guides multi-granularity visual learning using object-aware prefix tuning techniques to align visual features with LLMs' text feature space, then adopts a two-step retrieval process: generating knowledge clues relevant to the query and retrieving documents based on these clues. GRACE [199] assigns unique identifier strings to represent images, training MLLMs to memorize and retrieve image identifiers from textual queries. ACE [76] combines K-Means and RQ-VAE to construct coarse and fine tokens as multimodal data identifiers, aligning natural language queries with candidate identifiers. AVG [195] introduces autoregressive voken (i.e., visual token) generation, tokenizing images into vokens that serve as image identifiers while preserving visual and semantic alignment. By framing text-to-image retrieval as a token-to-voken generation task, AVG bridges the gap between generative training and retrieval objectives through discriminative training, refining the learning direction during token-to-voken generation.

3.3.2 RERANKER. Reranker, as a critical second-stage component in multimodal retrieval, is designed to re-rank a multimodal document list initially retrieved by a first-stage retriever. It achieves this by employing advanced relevance scoring mechanisms, such as cross-attention models, which enable more contextual interactions between queries and documents. Based on the utilization of large models, including LLMs and MLLMs, existing reranking methods can be categorized into two primary paradigms: fine-tuning-as-reranker and prompting-as-reranker.

- **Fine-tuning-as-Reranker**: The fine-tuning-as-reranker paradigm adapts PLMs to domain-specific reranking tasks through supervised fine-tuning on domain-specific datasets, addressing their inherent lack of ranking awareness and inability to effectively measure query-document relevance.
  - **Reranking for Text-Modal**: According the development of large models' architecture, reranker can be divided to three categories: encoder-only, encoder-decoder, and decoder-only.

    Encoder-only rerankers have advanced document ranking by fine-tuning PLMs (e.g., BERT [62]) to achieve precise relevance estimation. Key examples include Nogueira and Cho [282] and monoBERT [285], which format query-document pairs as query-document sequences. The relevance score is derived from the "[CLS]" token's representation via a linear layer, with optimization achieved through negative sampling and cross-entropy loss.

    Existing research on encoder-decoder rerankers primarily formulates document ranking as a generation task [157, 283, 296, 510], fine-tuning models like T5 to generate classification tokens (e.g., "true" or "false") for query-document pairs, with relevance scores derived from token logits [283]. Extensions include multi-view learning approaches [157] that simultaneously generate classification tokens for query-document pairs and queries conditioned on documents, and DuoT5 [296], which compares the classification tokens of document pairs to determine relative relevance. Beyond these approaches, studies have explored alternative training losses and architectures. Contrast with previous methods that rely on text generation losses, RankT5 [510] directly produces numerical relevance scores for each query-document pair, optimizing with ranking losses instead of generation losses. ListT5 [433] further advances this by processing multiple documents simultaneously, directly generating reranked lists using the Fusion-in-Decoder architecture.

    Recent studies [222, 255, 294, 464, 475] have explored fine-tuning decoder-only models like LLaMA for document reranking. RankLLaMA [255] formats query-document pairs into

prompts and uses the last token representation for relevance scoring. TSARankLLM [464] employs a two-stage training approach: continuous pretraining on web-sourced relevant text pairs to align LLMs with ranking tasks, followed by fine-tuning with supervised data and tailored loss functions. Q-PEFT [294] introduces query-dependent parameter-efficient fine-tuning to generate accurate queries from documents. In contrast, listwise approaches like those in [475] and PE-Rank [222] focus on directly outputting reranked document lists. Zhang et al. [475] highlight the limitations of point-wise datasets with binary labels, and instead use ranking outputs from existing systems as gold standards to train a listwise reranker. PE-Rank [222] compresses documents into single embeddings, reducing input length and improving reranking efficiency.

- **Reranking for Cross-Model**: The multi-modal reranking uses the multi-modal question and multi-modal knowledge items to obtain the relevance score, as reranking have already shown its importance in various knowledge-intensive tasks. Wen et al. [394] fine-tunes a pretrained MLLM to facilitate cross-item interaction between questions and knowledge items. The reranker is trained on the same dataset as the answer generator, using distant supervision by checking whether answer candidates appear in the knowledge text. RagVL RETRIEVAL et al. [311] introduces a novel framework featuring knowledge-enhanced reranking and noise-injected training. The approach involves instruction-tuning the MLLM with a simple yet effective template to enhance its ranking capability, enabling it to serve as a reranker for accurately filtering the top-k retrieved images.

In summary, these approaches leverages the representational capacity of large models while optimizing them for task-specific relevance signals, often achieving high reranking accuracy. However, it requires substantial computational resources and labeled training data, resulting in increased costs.

- **Prompting-as-Reranker**: In contrast, the prompting-as-reranker paradigm leverages large models in a zero-shot or few-shot manner by designing prompts that direct the model to generate relevance scores or rankings directly. This approach exploits the inherent knowledge and reasoning capabilities of large models, eliminating the need for extensive fine-tuning and offering greater flexibility and resource efficiency. Researchers have explored prompting LLMs and MLLMs to perform ranking tasks on multimodal documents, with prompting strategies generally categorized into three types: point-wise, pair-wise, and list-wise methods.
  - **Reranking for Text-Model**: LLMs are increasingly employed in text-modal reranking tasks, leveraging their advanced capabilities to optimize the ranking of textual documents.

    Point-wise methods evaluate the relevance between a query and individual documents, reranking them based on relevance scores.. Zhuang et al. [509] integrates fine-grained relevance labels into prompts for better document distinction. MCRanker [109] addresses biases in existing point-wise rerankers by generating relevance scores based on multi-perspective criteria. UPR [318] re-scores retrieved passages using a zero-shot question generation model. Zhuang et al. [511] show that LLMs pre-trained without supervised instruction fine-tuning (e.g., LLaMA) also exhibit strong zero-shot ranking capabilities. Despite their effectiveness, these methods often rely on suboptimal handcrafted prompts. To improve prompts for ranking tasks, Co-Prompt [52] introduces a discrete prompt optimization method for improving prompt generation in reranking tasks. PaRaDe [70] proposes a difficulty-based approach to select the most challenging in-context demonstrations for prompts, though experiments reveal that this method does not significantly outperform random selection. To improve demonstration selection, DemoRank [228] advances demonstration selection with a dependency-aware demonstration reranker, optimizing top-ranked examples through efficient training sample construction and a novel list-pairwise loss.

Pair-wise methods involve presenting LLMs with a query and a document pair, instructing them to identify the more relevant document. PRP-AllPair [302] generates all possible pairs, assigns discrete relevance judgments, and aggregates these into a final relevance score per document. PRP-Graph [248] improves this by using judgment generation probabilities and a graph-based aggregation for scoring relevance. Additionally, a post-processing technique [416] refines LLM-generated labels by aligning them with pairwise preferences while minimizing deviations from original values.

Listwise methods directly rank document lists by incorporating queries and documents into prompts, instructing LLMs to output reranked document identifiers. RankGPT [341] introduces instructional permutation generation and a sliding window strategy to address context length limits, while LRL [256] reorders document identifiers for candidate documents. However, these methods face challenges: (1) performance is highly sensitive to document order, revealing positional bias, and (2) the sliding window strategy limits the number of documents ranked per iteration. Recent advancements have attempted to address these issues: Tang et al. [349] propose permutation self-consistency to mitigate bias. TourRank [43] introduces a tournament mechanism, parallelizing reranking to minimize the impact of initial document order. TDPart [292] employs a top-down partitioning algorithm, which processes documents to depth using a pivot element. FIRST [309] leverages the output logits of the first generated identifier to directly obtain a ranked ordering of candidates.

– **Reranking for Cross-Model**: Prompt-Based Multimodal Reranker uses prompts to guide a MLLM in reranking items. TIGeR [303] proposes a framework leveraging multimodal LLMs for zero-shot reranking via a generative retrieval approach. However, their method is limited to text-only query retrieval tasks. In contrast, Lin et al. [210] extends this scope by utilizing multimodal LLMs to address diverse multimodal reranking tasks, supporting queries and documents in text, image, or interleaved text-image formats.

In summary, these approaches leverages the pre-existing knowledge and reasoning capabilities of LLMs, reducing the need for extensive task-specific fine-tuning. Consequently, it provides greater flexibility and resource efficiency, particularly in scenarios with limited labeled data or computational resources. However, its effectiveness depends heavily on the quality and design of the prompts, as well as the model's ability to generalize its pre-trained knowledge to the specific demands of the target task.

3.3.3 REFINER. Theoretically, LLMs improves with more comprehensive task-relevant knowledge in the retrieved and reranked context. However, unlimited input length poses practical deployment challenges: (1) Limited Context Window: LLMs have a fixed input length determined during pre-training, and any text exceeding this limit is truncated, leading to loss of contextual semantics. (2) Catastrophic Forgetting: Insufficient cache space can cause LLMs to forget previously learned knowledge when processing long sequences. (3) Slow Inference Speed. Consequently, refined prompts are crucial for optimizing LLM performance.

The refiner is an optional yet highly impactful component that optimizes retrieved and reranked information before its utilization by the LLM. It performs advanced processing tasks, such as summarization, distillation, or contextualization, to condense and refine content into a more digestible and actionable format. By extracting key insights, eliminating redundancies, and aligning information with the query's context, the refiner enhances the utility of the retrieved data, enabling the LLM to generate more coherent, accurate, and contextually relevant responses.

Prompt refinement can be achieved through two primary approaches: hard prompt methods and soft prompt methods. Hard prompt methods involve filtering out unnecessary or low-information content, still using natural language tokens and resulting in less fluent but generalizable prompts

that can be used across LLMs with different embedding configurations. Soft prompt methods, in contrast, encode prompt information into continuous representations, producing latent vectors (special tokens) that are not human-readable but optimized for model performance.

- **Hard Prompt Refiner**: Hard prompts consist of natural language tokens from the LLM/MLLM's vocabulary, representing specific words or sub-words, and can be generated by humans or models.
  - **Refining for Text-Model**: Recent advancements in prompt compression and context distillation aim to optimize the efficiency of LLMs. DynaICL [499] employs a meta controller to dynamically allocate in-context demonstrations based on input complexity and computational constraints. FILCO [385] distills retrieved documents using lexical and information-theoretic methods—String Inclusion, Lexical Overlap, and CXMI—training both context filtering and generation models for RAG tasks. CPC [213] preserves semantic integrity by using a context-aware encoder to remove irrelevant sentences, while AdaComp [469] dynamically selects optimal documents via a compression-rate predictor. LLMLingua [148] introduces a coarse-to-fine approach, compressing prompt components (instructions, questions, demonstrations) using a small language model (SLM) to measure token informativeness via perplexity (PPL). LongLLMLingua [149] extends this to long documents, employing a linear scheduler, reordering mechanism, and contrastive perplexity to retain question-relevant tokens while ensuring key information integrity. CoT-Influx [132] compresses GPT-4-generated Chain-of-Thought (CoT) prompts using a shot-pruner and token-pruner, both implemented as MLPs trained via reinforcement learning. These methods collectively improve performance while reducing useless CoT examples and redundant tokens. Selective Context [196] evaluates lexical unit informativeness using a causal language model and a percentile-based filtering method to remove redundancy. It calculates token self-information by predicting next-token probabilities, aggregating these at phrase and sentence levels. Prompt-SAW [11] preserves syntactic and semantic structures by extracting key tokens via relation-aware graphs, integrating them into compressed prompts. PCRL [159] treats prompt compression as a binary classification task, using a frozen pre-trained policy language model with trainable MLP layers. The compression policy labels tokens as include or exclude, optimizing a reward function that balances faithfulness and prompt length reduction. LLMLingua-2 [288] employs a bidirectional encoder-only model with a linear classification layer for compression, determining token retention or removal. RECOMP [410] employs extractive and abstractive compressors to generate query-focused summaries, leveraging contrastive learning and knowledge distillation. Nano-Capsulator [56] optimizes compression using reward feedback from response differences and enforces strict length constraints. MEMWALKER [32] uses interactive prompting to build and navigate a memory tree for context summarization. CompAct [432] sequentially compresses document segments for long-context question-answering, achieving high compression rates. Style-Compress [297] iteratively refines prompts using diverse styles and task-specific examples, evaluated by larger LLMs. TCRA-LLM [220] combines summarization and semantic compression to reduce token size. TACO-RL [323] employs reinforcement learning for task-aware prompt compression, ensuring low latency. FaviComp [158] enhances evidence familiarity by combining token probabilities from compression and target models, reducing perplexity.
  - **Refining for Cross-Model**: Recent advancements in visual token compression for MLLMs focus on enhancing efficiency without significant performance loss. LLaVolta [34] introduces a method to reduce the number of visual tokens, enhancing training and inference efficiency without compromising performance. To minimize information loss during compression while maintaining training efficiency, it employs a lightweight, staged training scheme. This scheme

progressively compresses visual tokens from heavy to light compression during training, ensuring no information loss during testing. PyramidDrop [407] is a visual redundancy reduction strategy for MLLMs, designed to improve efficiency in both inference and training with negligible performance loss. It partitions the MLLM into several stages and drops a predefined ratio of image tokens at the end of each stage. DeCo [425] proposes the principle of "Decouple Compression from Abstraction," which involves compressing visual tokens at the patch level using projectors while allowing the LLM to handle visual semantic abstraction entirely. MustDrop [226] measures the importance of each token throughout its lifecycle, including the vision encoding, prefilling, and decoding stages. During vision encoding, it merges spatially adjacent tokens with high similarity and establishes a key token set to retain vision-critical tokens. In the prefilling stage, it further compresses vision tokens guided by text semantics using a dual-attention filtering strategy. In the decoding stage, an output-aware cache policy reduces the size of the KV cache. By employing tailored strategies across these stages, MustDrop achieves an optimal balance between performance and efficiency. G-Search [484] proposes a greedy search algorithm to determine the minimum number of vision tokens to retain at each layer, from shallow to deep. Based on this strategy, a parametric sigmoid function (P-Sigmoid) is designed to guide token reduction at each layer of the MLLM, with parameters optimized using Bayesian Optimization. G-Prune [151] introduces a graph-based method for training-free visual token pruning. It treats visual tokens as nodes and constructs connections based on semantic similarities. Information flow is propagated through weighted links, and the most important tokens are retained for MLLMs after iterations.

Although interpretable and transparent, the inherent ambiguity of hard prompts often hinders the precise expression of intent, limiting their effectiveness in diverse or complex scenarios. Crafting accurate and impactful hard prompts demands significant human effort and may require model-based refinement or optimization. Moreover, even minor variations in hard prompts can lead to inconsistent LLM performance for identical tasks.

- **Soft Prompt Refiner**: Soft prompts are trainable, continuous vectors that match the dimensionality of token embeddings in LLM's vocabulary. Unlike hard prompts, which rely on discrete tokens from a predefined vocabulary, soft prompts are optimized through training to capture nuanced meanings that discrete tokens cannot express. When fine-tuned on diverse datasets, soft prompts enhance the LLM's performance across various tasks.
  - **Refining for Text-Model**: Language models convert text prompts into vectors for denser representation, enabling compression of discrete text into continuous vectors within the model. These vectors can serve as internal parameters (internalization) or additional soft prompts (encoding). Such compression extends the context window and enhances inference speed, particularly with repeated prompt usage.

    Early work focused on system prompt internalization. Askell et al. [13] used Knowledge Distillation to align models with human values, while Choi et al. [53] introduced Pseudo-Input Generation, generating pseudo-inputs from prompts and distilling knowledge between teacher and student models to avoid redundant inference computations. Later research compressed user prompt contexts. Snell et al. [331] distilled abstract instructions, reasoning, and examples into prompts with distinct distribution differences, enabling task execution without explicit prompts. Sun et al. [339] internalized ranking techniques for zero-shot relevance tasks, while Distilling Step-by-Step [122] improved reasoning tasks by distilling rationales as additional supervision. In retrieval-augmented generation, xRAG [48] integrated compressed document embeddings via a plug-and-play projector, using self-distillation for robustness. For context compression, COCOM [308] reduced long contexts to few embeddings, balancing trade-offs between decoding time and answer quality. LLoCO [346] learned offline compressed representations for efficient

QA retrieval. QGC [25] retained key information under high compression using query-guided dynamic strategies. UniICL [91] unified demonstration selection, compression, and generation within a single frozen LLM, projecting demonstrations and inputs into virtual tokens for semantic-based processing.

Recent advancements in prompt compression for LLMs focus on encoding hard prompts into reusable soft prompts to enhance efficiency and generalization across tasks. Early work by Wingate et al. [396] distilled complex hard prompts into concise soft prompts by minimizing output distribution differences, reducing inference costs. A series of works aim to enhance generalization across diverse prompts. Gist [279] used meta-learning to encode multi-task instructions into gist tokens, while Gist-COCO [193] employed an encoder-decoder architecture to compresses original prompts into shorter gist prompts, via the Minimum Description Length principle. UltraGist [467] optimized cross-attention for compressing ultra-long contexts into near-lossless UltraGist tokens. AutoCompressor [49] iteratively compressed contexts segments into summary vectors using a Recurrent Memory Transformer, reducing computational load. Other approaches, like ICAE [99] and 500xCompressor [203], fine-tuned LoRA-adapted LLMs for context encoding and prompt compression. For LLM-based recommendations, POD [190] distilled discrete prompt templates into continuous prompt vectors with an whole-word embedding to integrate the item ID, while RDRec [379] synthesizes training data and internalizes rationales into a smaller model. SelfCP [90] balances training cost, inference efficiency, and generation quality by compressing over-limit prompts asynchronously using frozen LLMs as the compressor and generator and trainable linear layers to project hidden states into LLM-acceptable memory tokens.

– **Refining for Cross-Model**: PromptMM [393] tackles overfitting and side information inaccuracies in multi-modal recommenders by using Multi-modal Knowledge Distillation with prompt-tuning. It compresses models by distilling user-item relationships and multi-modal content from complex teacher models to lightweight student models, eliminating extra parameters. Soft prompt-tuning bridges the semantic gap between multi-modal context and collaborative signals, enhancing robustness. Additionally, a disentangled multi-modal list-wise distillation with modality-aware re-weighting addresses multimedia data inaccuracies. RACC [395] compresses and aggregates retrieved knowledge for image-question pairs, generating a compact Key-Value (KV) cache modulation to adapt downstream frozen MLLMs for efficient inference. VTC-CLS [364] uses the prior knowledge of the association between the [CLS] token and visual tokens in the visual encoder to evaluate visual token importance, enabling Visual Token Compression and shortening visual context. VisToG [128] introduces a grouping mechanism using pretrained vision encoders to group similar image segments without segmentation masks. Semantic tokens represent image segments after linear projection and before input into the vision encoder. Isolated attention identifies and eliminates redundant visual tokens, reducing computational demands.

However, as dataset size increases, so do the computational resource requirements. Additionally, soft prompts are less interpretable than hard prompts, as their continuous vectors are not directly readable or explainable by humans.

3.4 Multimodal Generation

Multimodal generation based on Multimodal Large Language Models (MLLMs) represents a significant advancement, enabling the generation of content across multiple modalities such as text, images, audio, and video. These models leverage the strengths of large language models (LLMs) and extend them to handle and integrate diverse data types, creating rich, coherent, and contextually

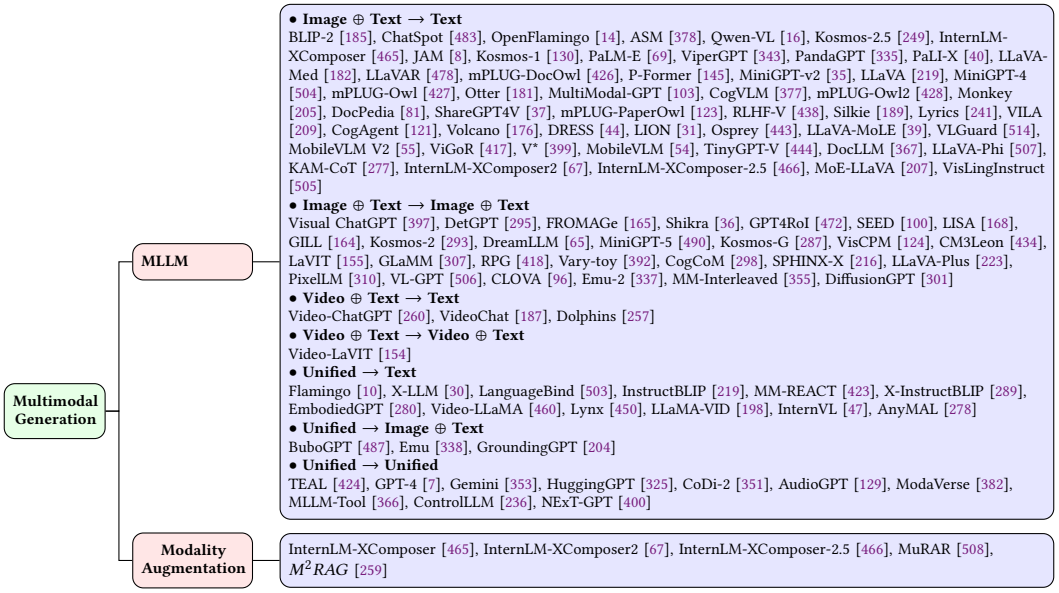

Fig. 8. Taxonomy of recent advancements in multimodal generation research.

relevant outputs. We classify MLLMs from generative perspectives of inputs and outputs, and summarize the related researches in Figure 8.

3.4.1 MODALITY INPUT. With the rapid advancement of large language models in the domain of textual knowledge comprehension and question-answering, researchers try to explore how to enable these models to understand and process inputs from a broader range of modalities, thereby facilitating more extensive multimodal question-answering tasks. Initial efforts focused on incorporating image modality into the input of large models. For instance, Blip-2 [185] proposes a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. Internlm-xcomposer2 [66] proposes a vision-language model excelling in free-form text-image composition and comprehension. This model goes beyond conventional vision-language understanding, adeptly crafting interleaved text-image content from diverse inputs like outlines, detailed textual specifications, and reference images, enabling highly customizable content creation. DiffusionGPT [301] leverages Large Language Models (LLM) to offer a unified generation system capable of seamlessly accommodating various types of prompts and integrating domain-expert models. DiffusionGPT constructs domain-specific Trees for various generative models based on prior knowledge. When provided with an input, the LLM parses the prompt and employs the Trees-of-Thought to guide the selection of an appropriate model.

As the variety of modal data continues to expand, more complex modalities, such as video, have been integrated into the inputs of large models. For instance, Video-ChatGPT [260] proposes a multimodal model that merges a video-adapted visual encoder with an LLM. The resulting model is capable of understanding and generating detailed conversations about videos. Video-LaVIT [154] address spatiotemporal dynamics limitations in video-language pretraining with an efficient video

decomposition that represents each video as keyframes and temporal motions. These are then adapted to an LLM using well-designed tokenizers that discretize visual and temporal information as a few tokens, thus enabling unified generative pre-training of videos, images, and text. At inference, the generated tokens from the LLM are carefully recovered to the original continuous pixel space to create various video content. The proposed framework is both capable of comprehending and generating image and video content.

Recently, the input for multimodal large models has evolved from specialized modal data to a unified input that can handle arbitrary modal data. For instance, InstructBLIP [289] conduct a vision-language instruction tuning based on the pretrained BLIP-2 models. Additionally, we introduce an instruction-aware Query Transformer, which extracts informative features tailored to the given instruction. InternVL [47] design a large-scale vision-language foundation model (InternVL) which scales up the vision foundation model to 6 billion parameters and progressively aligns it with the LLM using web-scale image-text data from various sources. GPT-4 [7] proposes a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks. HuggingGPT [325] proposes an LLM-powered agent that leverages LLMs (eg, ChatGPT) to connect various AI models in machine learning communities (eg, Hugging Face) to solve AI tasks. Specifically, we use ChatGPT to conduct task planning when receiving a user request, select models according to their function descriptions available in Hugging Face, execute each subtask with the selected AI model, and summarize the response according to the execution results. By leveraging the strong language capability of ChatGPT and abundant AI models in Hugging Face, HuggingGPT can tackle a wide range of sophisticated AI tasks spanning different modalities and domains. NExT-GPT [400] present an end-to-end general-purpose any-to-any MM-LLM system. NExT-GPT connect an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, image, video, and audio. By leveraging the existing well-trained high-performing encoders and decoders, NExT-GPT is tuned with only a small amount of parameter (1%) of certain projection layers, which not only benefits low-cost training but also facilitates convenient expansion to more potential modalities. Moreover, we introduce a modality-switching instruction tuning (MosIT) and manually curate a high-quality dataset for MosIT, based on which NExT-GPT is empowered with complex cross-modal semantic understanding and content generation.

3.4.2 MODALITY OUTPUT. With the explosive growth in the capabilities of MLLMs, the ability to answer questions based on multimodal inputs and generate multimodal outputs has also seen a qualitative improvement. There is also increasing attention from researchers on VQA scenarios that shift from generating text results to generating multimodal results that include text.In this section, we are discussing multimodal outputs that are not scenarios like text-to-image or text-to-video, which only generate a single modality, but rather scenarios where the answers includes text and at least one other modality of data, such as text-image output, or image-video output.In the basic VQA task, MIMOQA[328] was the first to propose the concept of multimodal output, which achieved the capability of multimodal output by transforming questions into an image-text matching task. It constructed a dual-tower model called MExBERT. The text stream, based on BERT, takes in the query and related documents to output the final text answer. The visual stream, based on VGG-19, receives images related to the query and documents, outputting a relevance score between the image and text. The final insertion of the image is determined by this relevance score. Its groundbreaking introduction to multimodal output research has, however, certain limitations: 1) It is necessary to screen out images related to the question. The model only needs to select and output images from the small number of screened ones. The task is relatively simple. 2) Multimodality is

still limited to the image modality. 3) To simplify the issue, it is still limited to scenarios where the input images must include at least one relevant image. Based on the aforementioned limitations, the latest research has made corresponding improvements[67, 259, 401, 465, 466, 508].

A common workflow paradigm for implementing multimodal output is to first conduct position identification after generating a text answer to determine where to insert multimodal data. Subsequently, based on the surrounding context of the corresponding positions, candidate multimodal data is retrieved. Finally, a relevance matching model is utilized to determine the final data to be inserted. InternLM-XComposer [465] achieves multimodal output of text and images. After generating each paragraph of text, it calls a model to determine whether to insert an image. If it is determined that an image needs to be inserted, it will generate a caption of the image to be inserted and search the web for candidate images, eventually allowing the model to select the most relevant image from candidate set for insertion. InternLM-XComposer2 and 2.5 [67, 466] allow users to directly input a set of candidate images on the basis of the above. MuRAR [508] has also implemented multi - modal output in RAG scenarios based on this paradigm, but it has innovated the methods of position identification and candidate set recall in RAG scenarios. It uses source attribution to confirm the correspondence between the generated snippet and the retrieved snippet from the large model input, thereby determining the insertion point, and the candidate set directly uses the multimodal data associated with the retrieved snippet, simplifying the recall operation. In addition, it has expanded the multimodal data from images to include tables and videos. M^2RAG [259] employs an alternative paradigm to achieve multimodal output in the RAG scenario. It uses the user's query to simultaneously recall associated text elements and images. Then, based on the associations of the images and text elements in the original document, they are refined. Subsequently, MLLMs are employed to vectorize the images or convert them into descriptions, which are input into the generative model in the form of placeholders. The output generates answer text and a simple description placeholder for the associated image. Finally, through a chain-of-thought(COT) process, the placeholders are converted into actual images. NExT-GPT [401] employs an entirely different and novel paradigm. It directly trains a unified multimodal large model, unifying the reasoning and generation process, and directly generates multimodal data including text, images, videos, etc., through the model [401].

4 Dataset for MRAG

To evaluate the general capabilities of MRAG systems in real-world multimodal understanding and knowledge-based question-answering tasks, we curated a collection of existing datasets designed to comprehensively evaluate the MRAG pipeline. These datasets are categorized into two classes: (1) Retrieval & Generation-Joint Components, which evaluate the synergy of retrieval and generation by requiring systems to retrieve external knowledge and generate accurate responses; and (2) Generation, focusing solely on the model's ability to produce contextually accurate outputs without external retrieval. This categorization enables a detailed evaluation of MRAG systems' strengths and limitations in diverse scenarios.

4.1 Dataset for Retrieval & Generation

Datasets for Retrieval & Generation in MRAG are designed to evaluate end-to-end systems capable of retrieving relevant knowledge from multimodal sources (e.g., text, images, videos) and generating accurate responses. These datasets evaluate the synergistic integration of retrieval and generation, focusing on the system's ability to dynamically utilize external knowledge to improve response quality and relevance. In this section, we introduce key benchmarks designed for diverse evaluation of Retrieval & Generation tasks. Figure 9 provides an overview of existing benchmarks, while Table 1 summarizes the statistics of selected representative datasets.

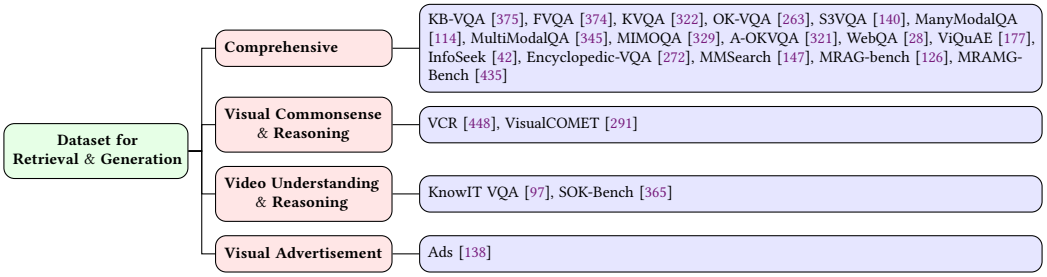

Fig. 9. Categories of MRAG dataset for retrieval & generation.

Table 1. Summary of dataset for retrieval & generation.

Dataset	Time	Statistics
Comprehensive		
KB-VQA [375]	2015	2,402 questions, 700 images, 1 knowledge bases.
FVQA [374]	2017	5,826 questions, 2,190 images, 3 knowledge bases.
KVQA [322]	2019	183,007 question-answer pairs about 18,880 unique entities contained within 24,602 images.
OK-VQA [263]	2019	14,055 questions, 10 scenarios.
S3VQA [140]	2021	6,765 question-image pairs.
ManyModalQA [114]	2020	10,190 questions with 2,873 image, 3,789 text, and 3,528 table.
MultiModalQA [345]	2021	29,918 questions that requires knowledge from text, tables, and images (35.7% require cross-modality reasoning).
MIMOQA [329]	2021	56,693 QA pairs, with 401,182 images.
A-OKVQA [321]	2022	24,903 multiple-choice questions.
WebQA [28]	2022	24,929 image-based and 24,343 text-based questions.
ViQuAE [177]	2022	3.7K questions paired with images. A Knowledge base composed of 1.5M Wikipedia articles paired with images.
InfoSeek [42]	2023	8.9K human-written and 1.3M semi-automated questions, 9 image classification and retrieval datasets.
Encyclopedic-VQA [272]	2023	1M Image-Question-Answer triplets derived from 221k textual QA pairs from 16.7k different categories. Each QA pair is combined with (up to) 5 images. 514k unique images. 15k textual single-hop questions, 25k multi-answer questions, and 22k two-hop questions.
MMSearch [147]	2024	2,901 unique images, 300 manually collected queries spanning 14 subfields.
MRAG-bench [126]	2024	1,353 multiple-choice questions, 16,130 images, 9 scenarios.
MRAMG-Bench [435]	2025	4,800 QA pairs across three distinct domains, containing 4,346 documents and 14,190 images, with tasks categorized into three difficulty levels.
Visual Commonsense Reasoning		
VCR [448]	2019	290k multiple choice QA problems derived from 110k movie scenes.
VisualCOMET [291]	2020	1,465,704 commonsense inferences over 59,356 images, and 139,377 distinct events.
Video Understanding & Commonsense Reasoning		
KnowIT VQA [97]	2020	24,282 human-generated QA pairs about a popular sitcom.
SOK-Bench [365]	2024	44K QA pairs covers over 12 types of questions, sourcing from about 10K situations. Each question is accompanied by two types of answers: a direct answer and a set of four multiple-choice options.
Visual Advertisement		
Ads [138]	2017	202,090 questions from 64,832 image ads and 3,477 video ads.

Early knowledge-based datasets include KB-VQA [375] and FVQA [374], which rely on closed knowledge. FVQA, for instance, uses a fixed knowledge graph, making questions straightforward once the knowledge is known, with minimal reasoning required. KVQA [322] focuses on images in Wikipedia articles, primarily testing named entity recognition and Wikipedia knowledge retrieval rather than commonsense reasoning. OK-VQA [263] and A-OKVQA [321] evaluate multimodal reasoning using external knowledge, with A-OKVQA introducing "rationale" annotations to better evaluate knowledge acquisition and reasoning. S3VQA [140] extends OK-VQA by requiring object detection and web queries, but like OK-VQA, it often reduces to single retrieval tasks rather than complex reasoning. MultiModalQA [345] pioneers complex questions requiring reasoning across snippets, tables, and images, focusing on cross-modal knowledge extraction. However, its template-based questions simplify the task to filling in blanks with modality-specific answering mechanisms.

ManyModalQA [114] also uses snippets, images, and tables but emphasizes answer modality choice over knowledge aggregation. MIMOQA [329] introduces "Multimodal Input Multimodal Output", requiring both text and image selections to enhance understanding. WebQA [28] is a manually crafted, multi-hop multimodal QA dataset that retrieves visual content but provides only textual answers, relying solely on MLLMs for reasoning, making it unsuitable for models dependent on linguistic context. ViQuAE [177] focuses on answering questions about named entities grounded in a visual context using a Knowledge Base. InfoSeek [42] and Encyclopedic-VQA [272] target knowledge-based questions beyond common sense knowledge, with Encyclopedic-VQA using model-generated annotations. MMSearch [147] evaluates MLLMs as multimodal search engines, focusing on image-to-image retrieval. Compared with previous works, MRAG-bench [126] evaluates MLLMs in utilizing vision-centric retrieval-augmented knowledge, identifying scenarios where visual knowledge outperforms textual knowledge. MRAMG-Bench [435] evaluates answers combining text and images, leveraging multimodal data within a corpus. Additionally, VCR [448] and VisualCOMET [291], derived from movie scenes, evaluate Visual Commonsense Reasoning. KnowIT VQA [97] and SOK-Bench [365] focus on video understanding and reasoning task, combining visual, textual, and temporal reasoning with knowledge-based questions. Ads [138] proposes an automatic advertisement understanding task, featuring rich annotations on topics, sentiments, and persuasive reasoning.

4.2 Dataset for Generation

The Generation category evaluates a model's intrinsic capacity to generate contextually accurate outputs based solely on its pre-trained knowledge and internal reasoning, without external retrieval. This evaluation isolates the generation component, providing insights into the model's foundational language understanding capabilities. It enables a detailed analysis of MRAG systems' strengths and limitations across diverse scenarios. In this section, we provide an overview of representative benchmarks developed for various evaluation of Generation tasks. The existing benchmarks are systematically organized in Figure 10, and the statistics of selected representative benchmarks are summarized in Table 2.

4.2.1 Comprehensive. To rigorously evaluate the capabilities of MLLMs, a diverse range of evaluation benchmarks has been developed. These benchmarks are designed to test various dimensions of model performance. By utilizing these benchmarks, researchers can systematically quantify the strengths and limitations of MLLMs, ensuring their alignment with real-world applications and user expectations. This evaluation framework not only supports the iterative improvement of MLLMs but also provides a standardized basis for comparing models in terms of perceptual and reasoning abilities.

VQA v2 [104], an early benchmark with 453K annotated QA pairs, focuses on open-ended questions with concise answers. VizWiz [112], introduced around the same time, includes 8K QA pairs from visually impaired individuals' daily lives, addressing real-world needs of disabled users. NLVR2 [336] explores multi-image vision capabilities by evaluating captions against image pairs. However, these benchmarks often fail to assess modern MLLMs' emergent capabilities, such as advanced reasoning. Recent efforts like LVLM-eHub [411], MDVP [212], and LAMM [430] compile extensive datasets for comprehensive evaluation, revealing that while MLLMs excel in commonsense tasks, they lag in image classification, OCR, VQA, large-scale counting, fine-grained attribute differentiation, and precise object localization. Fine-tuning can mitigate some of these limitations.

Researchers are developing specialized benchmarks to address the limitations of traditional evaluations for MLLMs. Notable examples include MME [29], which covers 14 perception and

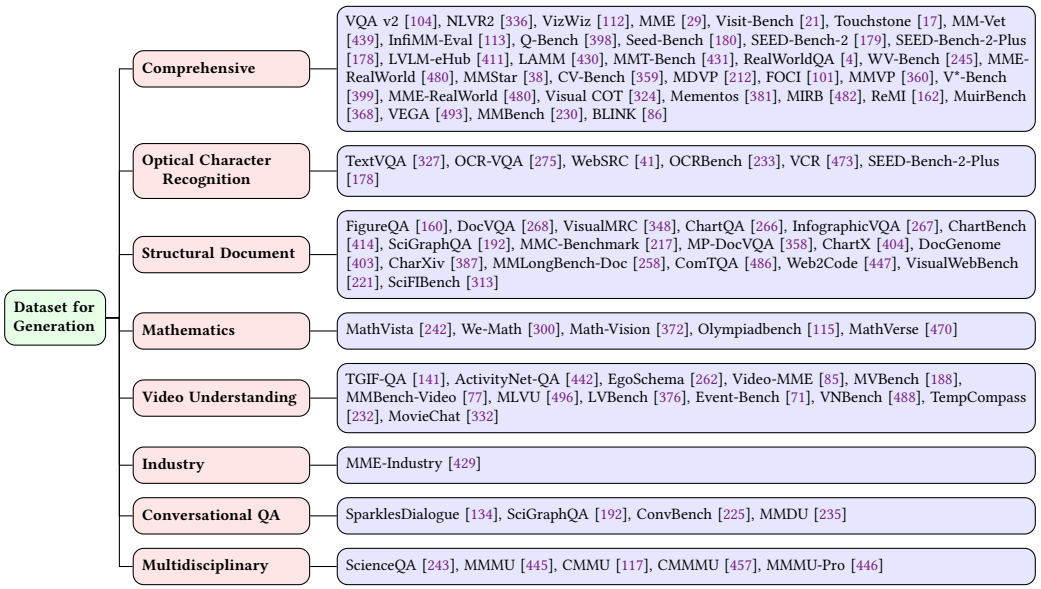

Fig. 10. Categories of MRAG dataset for generation.

cognition tasks; MMBench [230], featuring 20 ability dimensions, including object localization and social reasoning; and SEED-Bench [180], which focuses on multiple-choice questions. SEED-Bench-2 [179] expanded the scope to 24K QA pairs, including the evaluation of both text and image generation. MMT-Bench [431] further scaled up to 31K QA pairs across diverse scenarios. Common findings reveal that model performance improves with scale, but challenges persist in fine-grained perception tasks (e.g., spatial localization), chart and visual mathematics comprehension, and interleaved image-text understanding. Open-source MLLMs have shown rapid progress, often matching or surpassing closed-source models.

Real-world usage scenarios are critical for evaluating model performance in practical applications. Benchmarks like RealWorldQA [4] evaluates spatial understanding capabilities sourced from real-life scenarios, while BLINK [86] highlights tasks such as visual correspondence and multi-view reasoning that challenge current MLLMs despite being intuitive for humans. WV-Bench [245] and Visit-Bench [21] emphasize human preferences and instruction-following capabilities, whereas V*-Bench [399] evaluates high-resolution image processing and correct visual details through attribute recognition and spatial reasoning tasks. MME-RealWorld [480] enhances quality and difficulty with extensive annotated QA pairs and high-resolution images. These benchmarks reveal that fine-grained perception tasks remain challenging for models, while artistic style recognition and relative depth perception are relatively stronger. Although closed-source models like GPT-4o outperform others, human performance still surpasses general models significantly.

Many studies simplify evaluation into binary or multi-choice problems for easier quantification, but this approach overlooks the importance of the reasoning process, which is critical for understanding model capabilities. To address this, some works use open-ended generation and LLM-based evaluators, though these face challenges with inaccurate LLM scoring. For instance, MM-Vet [439] employs diverse question formats to assess integrated vision-language capabilities, while Touchstone [17] emphasizes real-world dialogue evaluation, arguing that multiple-choice questions are insufficient for evaluating multimodal dialogue capabilities. InfiMM-Eval [113] evaluates models on

deductive, abductive, and analogical reasoning across tasks, including intermediate reasoning steps, aligning with practical scenarios like mathematical problem-solving. These benchmarks highlight the strengths and limitations of MLLMs in complex tasks. Closed-source models excel in reasoning but struggle with complex localization, structural relationships, charts, and visual mathematics. High-resolution data improves recognition of small objects, dense text, and fine-grained details. While Chain-of-Thought (CoT) strategies significantly boost reasoning in closed-source models, their impact on open-source models remains limited.

The development of multimodal benchmarks emphasizes continuous refinement to accurately assess model capabilities. MMStar [38] addresses data leakage by curating 1.5K visually-dependent QA pairs, while CV-Bench [359] tackles the scarcity of vision-centric benchmarks with 2.6K manually-inspected samples for 2D/3D understanding. FOCI [101] evaluates MLLMs using domain-specific subsets and supplementary classification datasets, revealing challenges in fine-grained perception. MMVP [360] identifies 9 distinct patterns in CLIP-based models, showing MLLMs' struggles with visual details, with only Gemini and GPT-4V performing above random guessing. Q-Bench [398] evaluates low-level attribute perception, highlighting GPT-4V's near-human performance. Visual-COT [324] introduces visual chain-of-thought prompts to enhance MLLMs' focus on specific image regions. To further upgrading vision capabilities on multiple image understanding, Mementos [381] evaluates sequential image understanding, while MIRB [482] focuses on multi-image reasoning across perception, visual knowledge, and multi-hop reasoning tasks. ReMI [162] designs 13 tasks with diverse image relationships and input formats, and MuirBench [368] includes 12 multi-image understanding tasks with unanswerable variants for robust assessment. VEGA [493] is specifically designed to evaluate interleaved image-text comprehension. The task requires models to identify relevant images and text while filtering out irrelevant information to arrive at the correct answer. Evaluation results reveal that even advanced proprietary MLLMs, such as GPT-4V and Gemini 1.5 Pro, achieve only modest performance, highlighting significant room for improvement in interleaved information processing capabilities.

4.2.2 Optical Character Recognition (OCR). Multimodal benchmarks are increasingly focusing on the evaluation of Optical Character Recognition (OCR) tasks, driving progress in document understanding. Early benchmarks like TextVQA [327] and OCR-VQA [275] evaluated standard text recognition, while WebSRC [41] introduces advanced structural reasoning tasks like web page layout interpretation. SEED-Bench-2-Plus [178] and OCRBench [233] expanded evaluation to diverse data types, including charts, maps, and web pages, showing models achieving near state-of-the-art performance in recognizing various OCR text. VCR [473] addresses OCR task with partially obscured text embedded in images, requiring content reconstruction. Despite advancements, many MLLMs struggle with fine-grained OCR tasks. While models like GPT-4V perform well, they lag behind specialized OCR models. Performance varies significantly by data type, with knowledge graphs and maps posing greater challenges than simpler formats like charts, suggesting potential improvements through data-specific optimization or dedicated OCR integration.

4.2.3 Structural Document. Structural documents, including charts, HTML web content, and various document formats, play a critical role in practical applications due to their ability to efficiently convey complex information. These data types are characterized by their highly structured nature and information density, distinguishing them from natural images. Unlike images, which rely on visual patterns and textures, structural documents require models to comprehend intricate layouts, spatial relationships, and semantic connections between embedded elements such as text, tables, and graphical components.

To advance models capable of understanding and reasoning with such data, several benchmarks have been proposed for different types of structural documents. Early dataset FigureQA [160]

introduces a visual reasoning corpus with synthetic images and scientific-style figures, focusing on relationships between plot elements. ChartQA [266] emphasizes VQA with charts, ranging from tasks that require both data extraction and math reasoning. ChartX [404] collects a comprehensive dataset with 22 topics, 18 chart types, and 7 tasks, incorporating multiple modalities. VisualMRC [348] targets visual machine reading comprehension, emphasizing natural language understanding and generation. ChartBench [414] evaluates chart comprehension and data reliability through complex reasoning. MMC-Benchmark [217] provides a human-annotated benchmark to assess MLLMs on visual chart understanding tasks like chart information extraction, reasoning, and classification. Web2Code [447] introduces a webpage-to-code dataset for instruction tuning and an evaluation framework to assess MLLMs' webpage understanding and HTML code translation capabilities. VisualWebBench [221] evaluates MLLMs on various web tasks at website, element, and action levels. Many charts lack data point annotations, necessitating MLLMs to infer values using chart elements. ComTQA [486] introduces a table VQA benchmark for perception and comprehension tasks, while DocVQA [268] focuses on document image QA with an emphasis on information extraction tasks. InfographicVQA [267] targets understanding infographics images, which are designed to present information concisely. Infographics exhibit diverse layouts and structures, requiring basic reasoning and arithmetic skills. As MLLMs advance, benchmarks now focus on complex chart and document understanding. For instance, DocGenome [403] analyzes scientific papers, covering tasks like information extraction, layout detection, VQA, and code generation. CharXiv [387] targets challenging charts from scientific papers, while MP-DocVQA [358] extends DocVQA to multi-page scenario, where questions are constructed based on multi-page documents instead of single page. MMLongBench-Doc [258] focuses on long document understanding, averaging 47.5 pages. SciGraphQA [192] is a synthetic dataset with 295K QA dialogues about academic graphs, generated using Palm-2 from CS/ML ArXiv papers. SciFIBench [313] benchmarks scientific figure interpretation, using adversarial filtering for negative examples and human verification for quality assurance.

Despite advancements, a performance gap persists between proprietary and open-source models on conventional benchmarks. Current MLLMs continue to face challenges in reasoning tasks and long-context document comprehension, particularly in interpreting extended multimodal contexts, which remains a critical limitation.

4.2.4 Mathematics. Visual math problem-solving is key to evaluating MLLMs, leading to the development of specialized benchmarks. MathVista [242] pioneered this effort by aggregating 28 existing datasets and introducing 3 new ones, featuring diverse tasks like logical, algebraic, and scientific reasoning with various visual inputs. Subsequent benchmarks, such as Math-Vision [372] and OlympiadBench [115], introduced more complex tasks and fine-grained evaluation methods. We-Math [300] decomposes problems into sub-problems to assess fundamental understanding, while MathVerse [470] further evaluates MLLMs' comprehension of math diagrams by transforming problems into six versions with varying proportions of visual and textual content.

Despite promising results from MLLMs, significant challenges remain. existing MLLMs often struggle with interpreting complex diagrams, rely heavily on textual cues, and address composite problems through memorization rather than underlying reasoning. These limitations highlight the need for further development in MLLM capabilities.

4.2.5 Video Understanding. Traditional video-QA benchmarks like TGIF-QA [141] and ActivityNet-QA [442] are domain-specific, focusing on tasks related to human activities. With advancements in MLLMs, new benchmarks have emerged to address more complex video understanding challenges. Video-MME [85] explores diverse video domains with multimodal inputs and manual annotations, while MVBench [188] reannotates existing datasets using ChatGPT. MMBench-Video [77] features

free-form questions for short to medium-length videos. Benchmarks like MLVU [496], LVBench [376], Event-Bench [71], and VNBench [488] emphasize long-video understanding, testing models on extended multimodal contexts. VNBench [488] introduces a synthetic framework for evaluating tasks like retrieval and ordering, by inserting irrelevant images or text into videos. Specialized benchmarks like EgoSchema [262] focus on egocentric videos. TempCompass [232] evaluates fine-grained temporal perception, and MovieChat [332] targets long videos but often reduces tasks to short-video problems. Current MLLMs, especially open-source ones, face challenges with long-context processing and temporal perception, underscoring the need for improved capabilities in these areas.

4.2.6 Industry. The absence of a comprehensive benchmark for evaluating MLLMs across diverse industry verticals has limited understanding of their applicability in specialized real-world scenarios. To address this gap, MME-Industry [429] was developed specifically for industrial applications, covering over 21 industrial sectors such as power generation, electronics manufacturing, textile production, steel, and chemical processing. Domain experts from each sector meticulously annotated and validated test cases, ensuring the benchmark's reliability, accuracy, and practical relevance. MME-Industry thus serves as a robust tool for assessing MLLMs in industrial contexts.

4.2.7 Conversational QA. Current MLLMs are primarily designed for multi-round chatbot interactions, yet most benchmarks focus on single-round QA tasks. To better align with real-world conversational scenarios, multi-round QA benchmarks have been developed to simulate human-AI interactions with extended contextual histories. SparklesDialogue [134] evaluates conversational proficiency across multiple images and dialogue turns, featuring flexible text-image interleaving with two rounds and four images per instance. SciGraphQA [192] constructs multi-turn QA conversations based on scientific graphs from Arxiv papers, emphasizing complex scientific discourse. ConvBench [225] assesses perception, reasoning, and creation capabilities across individual rounds and overall conversations, revealing that MLLMs' reasoning and creation failures often stem from inadequate fine-grained perception. MMDU [235] engages models in multi-turn, multi-image conversations, with up to 20 images and 27 turns, highlighting that the performance gap between open-source and closed-source models is largely due to limited conversational instruction tuning data. These benchmarks collectively enhance the evaluation of MLLMs in complex, real-world interaction scenarios.

4.2.8 Multidisciplinary. The mastery of multidisciplinary knowledge is a key indicator of a model's expertise, and several benchmarks have been developed to evaluate this capability. ScienceQA [243] comprises scientific questions annotated with lectures and explanations, designed to facilitate chain-of-thought evaluation. It spans grade-level knowledge across diverse domains. MMMU [445] presents a more challenging college-level benchmark across diverse subjects, including engineering, art and design, business, science, humanities, social science, and medicine. Its question format extends beyond single image-text pairs to include interleaved text and images. Similarly, CMMU [117] and CMMMU [457] provide Chinese domain-specific benchmarks for grade-level and college-level knowledge, respectively. MMMU-Pro [446] enhances MMMU with a more robust version for advanced evaluation.

Table 2. Summary of dataset for generation.

Dataset	Time	Statistics
Comprehensive		
VQA v2 [104]	2017	contains more than 443K training, 214K validation and 453K test image-question pairs.

NLVR2 [336]	2018	contains 107,292 examples of English sentences paired with web photographs, including 29,680 unique sentences and 127,502 images. The task is to determine whether a natural language caption is true about a pair of photographs.
VizWiz [112]	2018	contains 20,000 training, 3,173 validation, and 8,000 test sets of visual questions originating from blind people.
MME [29]	2023	measures both perception and cognition abilities on a total of 14 subtasks
Visit-Bench [21]	2023	comprising 592 instances and 1,159 public images. The instances are either from 45 newly assembled instruction families or reformatted from 25 existing datasets. 10 instruction families cater to multi-image query scenarios.
Touchstone [17]	2023	908 questions covering 27 subtasks. The highest proportion of questions pertains to recognition, accounting for about 44.1%, followed by comprehension questions at 29.6%. The proportions of the other categories are 15.3% for basic descriptive ability, 7.4% for visual storytelling ability, and 3.6% for multi-image analysis ability.
MM-Vet [439]	2023	defines 6 core vision-language capabilities and examines the 16 integrations of interest derived from their combinations. It contains 200 images, and 218 questions (samples), all paired with their respective ground truths.
InfiMM-Eval [113]	2023	It consists of 279 manually curated reasoning questions, associated with a total of 342 images. The dataset is categorized into three reasoning paradigms: deductive, abductive, and analogical reasoning. 49 questions pertain to abductive reasoning, 181 require deductive reasoning, and 49 involve analogical reasoning. Furthermore, the dataset is divided into two folds based on reasoning complexity, with 108 classified as "High" reasoning complexity and 171 as "Moderate" reasoning complexity.
Q-Bench [398]	2023	consists of 2,990 diverse-sourced images, each equipped with a human-asked question focusing on its low-level attributes.
Seed-Bench [180]	2023	consists of 19K multiple-choice questions with accurate human annotations, which spans 12 evaluation dimensions including the comprehension of both the image and video modality.
SEED-Bench-2 [179]	2024	comprises 24K multiple-choice questions with accurate human annotations, which span 27 dimensions, including the evaluation of both text and image generation.
LVLM-eHub [411]	2024	contains 42 datasets in our LVLM-eHub. The sizes of specific datasets are 109.8K, 29.5K, 177.2K, 67.3K, and 8.9K for visual perception, knowledge acquisition, reasoning, commonsense, and object hallucination, respectively.
LAMM [430]	2024	evaluate 9 common image tasks, using a total of 11 datasets with over 62,439 samples, and 3 common point cloud tasks, by utilizing 3 datasets with over 12,788 data samples.
MMT-Bench [431]	2024	comprises 31,325 meticulously curated multi-choice visual questions from various multimodal scenarios, covering 32 core meta-tasks and 162 subtasks in multimodal understanding.
RealWorldQA [4]	2024	consists of 765 images, with a question and easily verifiable answer for each image.
WV-Bench [245]	2024	constructed by selecting 500 high-quality samples from 8,000 user submissions in WV-ARENA.
MME-RealWorld [480]	2024	constructed by collecting more than 300K images from public datasets and the Internet, filtering 13,366 high-quality images for annotation and contributing to 29,429 question-answer pairs that cover 43 subtasks across 5 real-world scenarios.
MMStar [38]	2024	contains 1,500 challenging samples, each rigorously validated by humans. It identify 6 core capabilities (i.e., coarse perception, fine-grained perception, instance reasoning, logical reasoning, science & technology, mathematics) along with 18 specific dimensions.
CV-Bench [359]	2024	provides 2,638 manually-inspected examples, and formulate natural language questions that evaluates 2D understanding via spatial relationships & object counting, and 3D understanding via depth order & relative distance.
MDVP [212]	2024	contains 1.6M unique image-visual prompt-text instruction-following samples, including natural images, document images, OCR images, mobile screenshots, web screenshots, and multi-panel images.
FOCI [101]	2024	constructed from 5 popular classification datasets for different domains: 1) aircraft contains images of 100 different aircraft types; 2) flowers contains images of 102 different flower species; 3) food covers 101 dishes; 4) pets contains images of 37 cat and dog breeds. 5) cars covers 196 car models. Additionally, FOCI create 4 domain subsets for animals (1322 classes), plants (957 classes), food (563 classes), and man-made objects (2631 classes).
MMVP [360]	2024	summarizes 9 prevalent patterns of the CLIP-blind pairs, such as "orientation", "counting", and "viewpoint". Utilizing the collected CLIP-blind pairs, MMVP design 150 pairs with 300 questions.
V*-Bench [399]	2024	It is built based on 191 high-resolution images with an average image resolution of 2246×1582. V*-Bench contains two sub-tasks: attribute recognition and spatial relationship reasoning. The attribute recognition task has 115 samples. The spatial relationship reasoning task has 76 samples.
MME-RealWorld [480]	2024	contains 29,429 question-answer pairs that cover 43 subtasks across 5 real-world scenarios, where each one has at least 100 questions.
Visual COT [324]	2024	438k visual chain-of-thought question-answer pairs spans across five distinct domains, each consisting of a question, an answer, and an intermediate bounding box as CoT contexts. About 98k question-answer pairs include extra detailed reasoning steps.
Mementos [381]	2024	It consists of 4,761 image sequences with varying episode lengths, encompassing diverse scenarios from dailylife, robotics tasks, and comic-style storyboards. Each sequence is paired with a human-annotated description of the primary objects and their behaviors within the sequence.
MIRB [482]	2024	comprises 925 samples with average image number of 3.78, constructed from four distinct categories of multi-image understanding: perception, visual world knowledge, reasoning, and multi-hop reasoning.

ReMI [162]	2024	It consists of 13 tasks that span a range of domains and properties. The tasks require reasoning over up to six images, with all tasks requiring reasoning over at least two images. The images comprise a variety of heterogeneous image types.
MuirBench [368]	2024	It consists of 12 distinctive multi-image understanding tasks that involve 10 categories of multi-image relations, comprising 11,264 images and 2,600 multiple-choice questions, with an average of 4.3 images per instance.
VEGA [493]	2024	contains 50k scientific literature entries, over 200k question-and-answer pairs, and a rich trove of 400k images. It includes the Interleaved Image-Text Comprehension subset, which is segmented into two categories based on token length: one supports up to 4,000 tokens, while the other extends to 8,000 tokens. Here, images are equated to 256 tokens each. Both categories offer roughly 200k training instances and approximately 700 test samples.
MMBench [230]	2025	contains 3,217 multiple-choice questions covering a diverse spectrum of 20 fine-grained skills.
BLINK [86]	2025	reformats 14 classic computer vision tasks, and contains 3,807 multiple-choice questions across 7.3K images, paired with single or multiple images and visual prompting.
Optical Character Recognition		
TextVQA [327]	2019	contains 45,336 questions asked by humans on 28,408 images that require reasoning about text to answer. Each question-image pair has 10 ground truth answers provided by humans.
OCR-VQA [275]	2019	comprises of 207,572 images of book covers and contains more than 1 million question-answer pairs about visual question answering by reading text in images.
WebSRC [41]	2021	It consists of 400K question-answer pairs, which are collected from 6.4K web pages. Along with the QA pairs, corresponding HTML source code, screenshots, and metadata are also provided in the dataset. Each question in WebSRC requires a certain structural understanding of a web page to answer.
OCRBench [233]	2024	includes 1000 question-answer pairs, which is consist of five components: text recognition, scene text-centric VQA, document-oriented VQA, key information extraction, and handwritten mathematical expression recognition.
VCR [473]	2024	It comprise 2.11M English and 346K Chinese entities, featuring captions in both languages across 'easy' and 'hard' difficulty levels.
SEED-Bench-2-Plus [178]	2024	comprises 2.3K multiple-choice questions with precise human annotations, spanning three broad categories: Charts, Maps, and Webs.
Structural Document		
FigureQA [160]	2017	its training set contains 100,000 images with 1.3 million questions; the validation and test sets each contain 20,000 images with over 250, 000 questions. The images are synthetic, scientific-style figures from five classes: line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts.
DocVQA [268]	2021	contains 50,000 question-answer pairs with 12,767 document images sourced from documents in UCSF Industry Documents Library.
VisualMRC [348]	2021	It contains 30562 pairs of a question and an abstractive answer for 10,197 document images sourced from multiple domains of webpages.
ChartQA [266]	2022	consists of 20,882 charts curated from four different online sources. It covers 9,608 human-written questions focusing on logical and visual reasoning questions, and generates another 23,111 questions automatically from human-written chart summaries.
InfographicVQA [267]	2022	comprises 30,035 questions over 5,485 images. Questions in the dataset include questions grounded on tables, figures and visualizations and questions that require combining multiple cues.
ChartBench [414]	2023	includes over 68k charts and more than 600k high-quality instruction data, covering 9 major categories and 42 subcategories of charts. 5 chart question-answering tasks to assess the models' cognitive and perceptual abilities.
SciGraphQA [192]	2023	generate 295K samples of open-vocabulary multi-turn question-answering dialogues about the graphs. As context, it provided the text-only Palm-2 with paper title, abstract, paragraph mentioning the graph, and rich text contextual data from the graph itself, obtaining dialogues with an average 2.23 question-answer turns for each graph.
MMC-Benchmark [217]	2023	consists of 2k QA pairs with 1,063 unique images, accompanied by 1,275 multiple-choice questions and 851 free-form questions. The average length of the questions is 15.6.
MP-DocVQA [358]	2023	contains 46K questions and 6K documents, with a total of 48K pages (images). On average, each question is associated with 8.27 pages.
ChartX [404]	2024	collected 48K multi-modal chart data covering 22 topics, 18 chart types, and 7 tasks. Each chart data within this dataset includes 4 modalities, including image, Comma-Separated Values (CSV), python code, and text description. 7 chart tasks is classified into perception tasks and cognition tasks.
DocGenome [403]	2024	constructed by annotating 500K scientific documents from 153 disciplines in the arXiv open-access community. It contains structure data from all modalities including 13 layout attributes along with their LATEX source codes. It provides 6 logical relationships between different entities within each scientific document. It covers various document-oriented tasks.
CharXiv [387]	2024	involves 2,323 real-world charts handpicked from scientific papers spanning 8 major subjects published on arXiv, and produces more than 10K questions.
MMLongBench-Doc [258]	2024	comprising 1,082 expert-annotated questions. It is constructed upon 135 lengthy PDF-formatted documents with an average of 47.5 pages and 21,214 textual tokens. 494 questions are single-page questions (with one evidence page). 365 questions are cross-page questions requiring evidence across multiple pages. 223 questions are designed to be unanswerable for detecting potential hallucinations.
ComTQA [486]	2024	comprises a total of 9,070 QA pairs across 1,591 images. It contains challenging questions, such as multiple answers, mathematical calculations, and logical reasoning.
Web2Code [447]	2024	contains a total of 1179.7k webpage based instruction-response pairs.

VisualWebBench [221]	2024	consists of 7 tasks, and comprises 1.5K human-curated instances from 139 real websites, covering 87 sub-domains.
SciFIBench [313]	2024	consists of 2000 multiple-choice scientific figure interpretation questions split between two tasks across 8 categories. The questions are curated from arXiv paper figures and captions.
Mathematics		
MathVista [242]	2023	incorporates 28 existing multimodal datasets, including 9 math-targeted question answering (MathQA) datasets and 19 VQA datasets. In addition, it creates three new datasets (i.e., IQTest, FunctionQA, PaperQA) which are tailored to evaluating logical reasoning on puzzle test figures, algebraic reasoning over functional plots, and scientific reasoning with academic paper figures, respectively. It consists of 6,141 examples, with 736 of them being newly curated.
We-Math [300]	2024	It collect and categorize 6.5K visual math problems, spanning 67 hierarchical knowledge concepts and 5 layers of knowledge granularity.
Math-Vision [372]	2024	comprises 3,040 mathematical problems within visual contexts across 12 grades, selected from 19 math competitions. It contains 1,532 problems in an open-ended format and 1,508 in a multiple-choice format. All problems encompass 16 subjects over 5 levels of difficulty.
Olympiadbench [115]	2024	an Olympiad-level bilingual multimodal scientific benchmark, featuring 8,476 problems from Olympiad-level mathematics and physics competitions, including the Chinese college entrance exam. Each problem is detailed with expert-level annotations for step-by-step reasoning.
MathVerse [470]	2025	It contains 2,612 math problems from three fundamental math subjects, i.e., plane geometry (1,746), solid geometry (332), and functions (534). Each problem is then transformed by human annotators into six distinct versions, each offering varying degrees of information content in multimodality, contributing to 15K test samples in total.
Video Understanding		
TGIF-QA [141]	2017	consists of 103,919 QA pairs collected from 56,720 animated GIFs. TGIF-QA includes four task types: repetition count, repeating action, state transition, frame QA.
ActivityNet-QA [442]	2019	It exploits 5,800 videos from the ActivityNet dataset, which contains about 20,000 untrimmed web videos representing 200 action classes. Each video is annotated with ten question-answer pairs using crowdsourcing to finally obtain 58,000 question-answer pairs. The maximum question length is 20 and the maximum answer length is 5. The average question length is 8.67 and average answer length is 1.85.
EgoSchema [262]	2023	consists of over 5000 human curated multiple choice question answer pairs, spanning over 250 hours of real video data. For each question, it requires the correct answer to be selected between five given options based on a three-minute-long video clip.
Video-MME [85]	2024	It contains a annotated set of 2,700 high-quality multiple-choice questions (3 per video) from 900 videos, 744 subtitles and 900 audio files across various scenarios. For diversity in video types, it spans 6 visual domains, with 30 subfields. For duration in temporal dimension, it encompasses both short-, medium-, and long-term videos, ranging from 11 seconds to 1 hour.
MVBench [188]	2024	covers 20 video temporal understanding tasks that cannot be effectively solved with a single frame. Each task produces 200 multiple-choice QA pairs by leveraging ChatGPT to automatically reannotate existing video datasets with their original annotations.
MMBench-Video [77]	2024	incorporates approximately 600 web videos from YouTube, spanning 16 major categories. Each video ranges in duration from 30 seconds to 6 minutes. The benchmark includes roughly 2,000 original question-answer pairs, contributed by volunteers, covering a total of 26 fine-grained capabilities.
MLVU [496]	2024	consists of 3,102 questions across 9 categories with 2,593 questions for dev set and 509 questions for test set. It is made up of videos of diversified lengths, spanning from 3 min to more than 2 hours. Besides, each video is further partitioned as incremental segments, e.g., the first 3 min, the first 6 min, and the entire video.
LVBench [376]	2024	gathers an initial collection of 500 videos, each with a minimum duration of 30 minutes. Finally, these videos is annotated to select a subset of 103 videos.
Event-Bench [71]	2024	includes 6 event-related tasks and 2,190 test instances.
VNBench [488]	2024	1,350 samples with 9 sub-tasks.
TempCompass [232]	2024	collects a total of 410 videos and 500 pieces of meta-information, with 9 content categories.
MovieChat [332]	2024	1K long videos and 13K manual question-answering pairs.
Industry		
MME-Industry [429]	2025	encompasses 21 distinct domain, comprising 1050 question-answer pairs with 50 questions per domain.
Conversational QA		
SparklesDialogue [134]	2023	SparklesDialogueCC comprises 4.5K dialogues, each consisting of at least two images spanning two conversational turns. SparklesDialogueVG includes 2K dialogues, each with at least three distinct images across two turns.
SciGraphQA [192]	2023	selected 290,000 Computer Science or Machine Learning ArXiv papers, and then used Palm-2 to generate 295K samples of open-vocabulary multi-turn question-answering dialogues about the graphs. As context, it provided the text-only Palm-2 with paper title, abstract, paragraph mentioning the graph, and rich text contextual data from the graph itself, obtaining dialogues with an average 2.23 question-answer turns for each graph.
ConvBench [225]	2024	comprises 577 image-instruction pairs tailored for multi-round conversations. Each pair is structured around three sequential instructions, each targeting a distinct cognitive skill—beginning with perception, followed by reasoning, and culminating in creation. Encompassing 215 tasks, the benchmark is divided into 71 tasks focused on perception, 65 on reasoning, and 79 on creation.

MMDU [235]	2024	comprises 110 multi-image multi-turn dialogues with more than 1600 questions, each accompanied by detailed long-form answers. The questions in MMDU involve 2 to 20 images, with an average image&text token length of 8.2k tokens, a maximum turn length of 27, and a maximum image&text length reaching 18K tokens.
Multidisciplinary		
ScienceQA [243]	2022	multiple-choice science question dataset containing 21,208 examples. It covers diverse topics across three subjects: natural science, social science, and language science.
MMMU [445]	2024	includes 11.5K multimodal questions from college exams, quizzes, and textbooks, covering 6 core disciplines. These questions span 30 subjects and 183 subfields, comprising 30 highly heterogeneous image types.
CMMU [117]	2024	It consists of 3,603 questions in 7 subjects, covering knowledge from primary to high school. The questions can be categorized into 3 types: multiple-choice, multiple-response, and fill-in-the-blank.
CMMMU [457]	2024	A Chinese Multi-discipline multimodal Understanding, including 12k manually collected multimodal questions from college exams, quizzes, and textbooks, covering 6 core disciplines. These questions span 30 subjects and comprise 39 highly heterogeneous image typesbenchmark.
MMMU-Pro [446]	2024	3460 questions in total (1730 samples are in the standard format and the other 1730 are in the screenshot or photo form)

5 Evaluation Metrics of MRAG

Multimodal RAG systems generally consist of four core components: document parsing, search planning, retrieval, and generation, which collectively influence their end-to-end performance. Accurate and comprehensive evaluation of these components is essential, leveraging available multimodal benchmarks. In practice, three common evaluation strategies are typically employed: human evaluation, rule-based evaluation, and LLM/MLLM-based evaluation. Each strategy offers distinct advantages and disadvantages in calculating evaluation metrics.

- ***Human evaluation***: Human evaluation is widely regarded as the gold standard for assessing MRAG systems, as their effectiveness is ultimately determined by human users. This method is extensively used in research to ensure the reliability and relevance of model outputs. For instance, Bingo [58] employs human annotators to assess the accuracy of GPT-4V's responses, with a focus on identifying and analyzing model biases. In hallucination detection, M-HalDetect [108] demonstrates that human evaluation outperforms model-based methods in detecting subtle inaccuracies, highlighting its precision. Additionally, WV-Arena [245] uses a human voting system combined with Elo ratings to rank and compare multiple models, providing a robust benchmarking framework. However, human evaluation presents challenges, including increased time and labor costs, which limit its scalability for large-scale assessments. The reliability of results can also be affected by the limited number of evaluators, as individual biases may influence outcomes. To address these issues, some studies employ diverse evaluator pools and cross-validation techniques to enhance the balance and representativeness of assessments. Nonetheless, the trade-off between evaluation accuracy and resource expenditure remains a critical consideration in designing RAG model evaluation methodologies.
- ***Rule-based evaluation***: Rule-based evaluation metrics [41, 430, 473] are essential for assessing the performance of MRAG systems. These metrics rely on standardized evaluation tools, enabling objective, reproducible assessments with minimal human intervention. Compared to subjective human evaluations, deterministic metrics offer significant advantages, including reduced time consumption, lower susceptibility to bias, and greater consistency across multiple assessments. Such consistency is particularly crucial for large-scale evaluations or when comparing different systems or model iterations.
- ***LLM/MLLM-based evaluation***: For evaluation of MRAG systems, LLMs/MLLMs are employed to compare reference answers with generated outputs or to directly score responses. For example, MM-Vet [439] uses GPT-4 to automate evaluation, generating scores for each sample based on the input question, ground truth, and model output. Similarly, TouchStone [17] and LLaVA-bench [219] leverage GPT-4 to directly compare generated answers with reference answers, simplifying

the evaluation process. While integrating LLMs/MLLMs in evaluation reduces human effort, it has limitations. This approach is prone to systematic biases, such as sensitivity to the order of response presentation. Additionally, evaluation outcomes are heavily influenced by the inherent capabilities and limitations of the LLMs/MLLMs themselves, leading to potential inconsistencies, as different models may produce divergent results for the same task. These challenges underscore the need for careful model selection and evaluation design to mitigate biases and ensure reliable assessments.

5.1 Metrics of Retrieval & Generation

The evaluation of MRAG systems is essential for ensuring their effectiveness and reliability in processing complex, multimodal data. Evaluation metrics can be broadly classified into rule-based and LLM/MLLM-based approaches.

- **Rule-based Metrics**: Rule-based metrics evaluate the performance of MRAG systems using predefined criteria and heuristics. These metrics are generally interpretable, transparent, and computationally efficient, making them well-suited for tasks with well-defined benchmarks. Examples of common rule-based metrics include:
  - **Exact Match (EM)**: This metric evaluates whether the model's output exactly matches the ground truth, offering a clear and unambiguous performance measure. It is especially valuable in tasks requiring high accuracy and fidelity to reference data, such as question answering, fact verification, and information retrieval. While exact match (EM) provides a straightforward and interpretable evaluation, it may fall short in scenarios where semantically equivalent but lexically divergent responses are acceptable.
  - **ROUGE-N (N-gram Recall)**: The ROUGE metric is a widely used framework for evaluating text summarization and generation tasks. ROUGE-N measures the overlap of N-grams (contiguous sequences of N words) between generated text and one or more reference texts, with a strong emphasis on recall. This metric assesses how well the generated text captures the essential content of the reference. For example, ROUGE-1 evaluates unigram overlap, ROUGE-2 focuses on bigrams, and higher-order N-grams (e.g., ROUGE-3) capture more complex linguistic structures. While ROUGE-N provides a quantitative measure of lexical similarity, it is often supplemented by other metrics to account for semantic coherence, fluency, and relevance, particularly in multimodal contexts where textual and non-textual data interact.
  - **BLEU**: BLEU is a widely used metric in NLP for evaluating the quality of machine-generated text by assessing its similarity to one or more reference texts. Initially designed for machine translation, BLEU has been adapted to various NLP tasks, including multimodal generation. In multimodal settings, BLEU can evaluate the alignment between generated text and associated modalities (e.g., images, videos) by comparing the output to reference descriptions or captions. However, while BLEU offers a quantitative measure of n-gram overlap, it has limitations in capturing semantic depth, contextual coherence, and multimodal consistency, which are essential for comprehensive evaluation in MRAG systems.
  - **Mean Reciprocal Rank (MRR)**: MRR is a widely used metric for evaluating the performance of systems that produce ranked lists of results, such as search engines, recommendation systems, or retrieval-augmented models. MRR measures the rank position of the first relevant item in the returned list, reflecting the system's ability to surface correct or useful information quickly. It is calculated as the average of the reciprocal ranks of the first relevant result across multiple queries or tasks. A higher MRR indicates better performance, as it demonstrates the system's effectiveness in prioritizing relevant results at the top of the list.

- **CIDEr (Consensus-based Image Description Evaluation)**: CIDEr is specifically designed to measure the agreement between machine-generated captions and human-authored reference captions. It utilizes a TF-IDF weighting mechanism to quantify the similarity between generated and reference texts.
- **SPICE (Semantic Propositional Image Caption Evaluation)**: The evaluation of MRAG systems frequently utilizes the SPICE metric to assess the quality of generated captions. SPICE prioritizes semantic fidelity by parsing captions into structured scene graphs, which depict objects, attributes, and relationships within the text. These generated scene graphs are subsequently compared to reference graphs derived from ground-truth captions. By emphasizing semantic similarity over lexical overlap, SPICE offers a robust measure of how well the generated content aligns with the intended meaning. This makes it particularly well-suited for evaluating multimodal systems that integrate visual and textual information, ensuring a nuanced and contextually accurate assessment of MRAG outputs.
- **BERTScore**: Evaluation of MRAG focuses on assessing the quality and relevance of outputs in contexts integrating both textual and non-textual data (e.g., images, audio). A key metric for evaluating textual components is BERTScore, which utilizes contextual embeddings from BERT to measure semantic similarity between generated and reference texts. Unlike traditional metrics such as BLEU or ROUGE, which depend on exact word matches or n-gram overlap, BERTScore captures deeper semantic relationships by aligning tokens based on their contextual embeddings.
- **Perplexity**: It measures the model's ability to predict the next word in a sequence, with lower perplexity values indicating greater confidence and accuracy in predictions. This reflects a stronger understanding of the underlying data distribution.

Rule-based metrics offer objective and reproducible outcomes but frequently lack the adaptability needed to capture nuanced semantic or contextual understanding, especially in multimodal environments where text, images, and other data types interact.

- **LLM/MLLM-based Metrics**: The emergence of LLMs and MLLMs has transformed evaluation paradigms, enabling the use of their advanced reasoning and comprehension capabilities. LLM/MLLM-based metrics now provide more holistic and context-aware assessments of MRAG systems, with key approaches including:
  - **Answer Precision**: This metric measures the degree to which the knowledge in a model-generated answer is supported or entailed by the ground truth. It assesses the accuracy and relevance of retrieved information by evaluating the overlap between the model's output and the factual or contextual basis provided by the ground truth. High answer precision indicates that the model effectively utilizes retrieved multimodal data to produce responses aligned with the expected factual content. This metric is crucial for evaluating the reliability and factual consistency of multimodal RAG systems, ensuring that generated outputs are both contextually appropriate and informationally accurate.
  - **Ground Truth Recall**: This metric measures the degree to which the knowledge in the ground truth is accurately captured and reflected in the model-generated response. It assesses the model's ability to retrieve and integrate relevant information from the provided knowledge base or multimodal sources, ensuring the output aligns with the factual or contextual details in the reference data. It is particularly crucial for evaluating retrieval-augmented systems, as it directly quantifies the fidelity of the model's output to the intended knowledge. Higher scores indicate stronger alignment with the ground truth, reflecting enhanced retrieval and generation capabilities.
  - **Retrieved Context Precision**: This metric measures the alignment between the knowledge in the retrieved context and the information in the ground truth response. It evaluates the

proportion of relevant and accurate information in the retrieved context that is directly supported or entailed by the ground truth, assessing the retrieval system's precision and contextual appropriateness in generating accurate responses. This metric is especially vital in multimodal RAG systems, where integrating diverse data types (e.g., text, images, audio) requires robust evaluation of relevance and precision across modalities.

- **Retrieved Context Recall**: This metric measures the degree to which the retrieved context aligns with and encompasses the knowledge necessary to generate ground truth responses. It evaluates the proportion of relevant information from the ground truth captured within the retrieved context, serving as a key indicator of the retrieval system's effectiveness in supporting accurate and comprehensive response generation. High values indicate that the retrieval mechanism effectively identifies and incorporates essential knowledge, thereby enhancing the overall performance of the MRAG system.
- **Faithfulness**: This metric evaluates the extent to which generated text maintains factual consistency with the information in the retrieved documents, ensuring the output accurately reflects the source material and minimizes hallucinations or deviations from the evidence. In MRAG systems, it also ensures alignment with multimodal retrieved content, including textual, visual, and auditory elements, maintaining consistency across modalities.
- **Hallucination**: This metric measures the proportion of generated outputs containing hallucinated content, such as unsupported claims, fabricated information, or inaccuracies not substantiated by the retrieved data. It is essential for evaluating the reliability and factual consistency of the model's responses.

LLM/MLLM-based metrics are highly effective at capturing complex semantic relationships and contextual nuances, making them particularly suitable for multimodal RAG systems. However, they may inherit biases from the underlying models and demand substantial computational resources.

- **Metric Calculation**: When evaluating multimodal retrieval-augmented generation systems, implementation methods for the same metric can vary significantly, primarily categorized into coarse-grained and fine-grained approaches. These methodologies differ in their granularity and the depth of analysis applied to assess the quality of model-generated responses against reference answers.
  - **Coarse-Grained Evaluation:** Coarse-grained evaluation utilizes LLMs or MLLMs to compare model-generated responses with reference answers. This method involves inputting both the generated output and the reference into the LLM/MLLM, which evaluates the overall semantic alignment, coherence, and relevance between the two. The model assesses whether the generated content captures the core meaning and intent of the reference, providing a holistic score or qualitative feedback. This approach is computationally efficient and scalable, making it suitable for rapid benchmarking and high-level quality checks in large-scale applications. However, its broad focus may overlook fine-grained inaccuracies, such as subtle factual errors, logical inconsistencies, or nuanced contextual mismatches. Consequently, coarse-grained evaluation is best used as an initial screening tool or in scenarios where high-level semantic fidelity is prioritized over detailed precision. For more rigorous evaluation, it is often supplemented by fine-grained metrics that address specific aspects of content quality. In summary, coarse-grained evaluation offers a pragmatic balance between efficiency and effectiveness, particularly in applications requiring quick assessments or large-scale model comparisons.
  - **Fine-Grained Evaluation:** Fine-grained evaluation, such as RAGChecker [317] and RAGAS [74], offers a nuanced and detailed approach to assessing MRAG systems, surpassing the limitations of coarse-grained methods. This approach involves decomposing both model-generated

responses and reference answers into granular knowledge points or semantic units, which are individually evaluated based on criteria such as accuracy, relevance, and alignment with the reference. By analyzing responses at this level of detail, the method enables precise identification of a model's strengths and weaknesses, particularly in capturing and reproducing intricate information. The fine-grained approach is especially valuable for diagnosing performance issues in handling complex or nuanced content. However, it is computationally intensive, requiring robust mechanisms for extracting, matching, and evaluating multiple semantic units. Careful design of these mechanisms is essential to ensure evaluation consistency and reliability. Despite its challenges, this method provides a rigorous and comprehensive framework for advancing the development and refinement of MRAG systems, making it a critical tool in the field.

The choice between coarse-grained and fine-grained evaluation depends on the assessment objectives. Coarse-grained methods are ideal for obtaining quick, high-level insights, whereas fine-grained approaches are better suited for detailed analysis and iterative model refinement. Integrating both strategies can provide a balanced perspective, combining the efficiency of coarse-grained evaluation with the precision of fine-grained analysis to comprehensively assess MRAG systems.

6 Challenges of MRAG

In this section, we delineate the challenges associated with various modules in a MRAG system. These challenges span multiple critical components, including document parsing and indexing, search planning, retrieval, generation, dataset, and evaluation. Each module presents unique complexities that must be addressed to ensure the system's effectiveness and robustness.

6.1 Document Parsing and Indexing

Document parsing and indexing has established the data foundation based on MRAG, which plays a crucial role in the entire system. The relevant technologies extensively studied even before the advent of LLMs, have seen significant advancements in the LLM era. However, they continue to face several challenges that necessitate further exploration and refinement.

- **Challenges in Data Accuracy and Completeness:** As the primary input source, the accuracy and completeness of upstream data are critical. Errors or omissions in the upstream data can propagate and amplify downstream, significantly degrading system performance. For example, while MRAG systems have enhanced document information preservation—such as capturing per-page screenshots—they still face challenges in maintaining inter-page relationships. This limitation is particularly problematic in long documents with associated segments. Preserving these relationships is essential for ensuring contextually accurate outputs.
- **Balancing Multimodal and Textual Data:** The document parsing module has grown increasingly complex as modern MRAG systems must handle multimodal data, including images, tables, and text. To address this, contemporary approaches preserve the original multimodal inputs to minimize information loss, while also converting them into textual captions or descriptions. Although retaining the original data reduces information loss, relying solely on it has proven suboptimal. Recent studies highlight the benefits of leveraging textual representations derived from multimodal data. For example, Riedler and Langer [312] showed that models generate higher-quality responses using textual captions from images rather than processing raw images directly. Similarly, Ma et al. [259] found that LLMs outperform MLLMs in text generation tasks, revealing a performance gap between multimodal and text-focused systems. This gap highlights

the limitations of current multimodal systems in effectively integrating diverse data types, necessitating additional components in document parsing pipelines. These enhancements, while improving functionality, increase system complexity and expand the volume of data requiring processing, storage, and management.

6.2 Multimodal Search Planning

The challenges in multimodal search planning can be more effectively understood through a hierarchical framework similar to leveled RAG systems, where queries span a spectrum from simple factual retrievals to complex, creative tasks. This framework highlights three critical challenges that must be addressed to advance the field.

- **Intelligent Adaptive Planning Mechanisms:** The primary challenge is developing intelligent adaptive planning mechanisms that can dynamically adjust to the diversity and complexity of queries. Current systems often rely on predetermined pipelines, which fail to accommodate variations in query characteristics or computational constraints, leading to inefficient resource allocation and suboptimal performance [125, 474]. While fixed strategies may suffice for homogeneous query types, real-world applications handle heterogeneous query patterns that demand dynamic adjustment of retrieval strategies. For example, complex queries involving multi-hop reasoning or creative problem-solving could greatly benefit from a multi-agent collaborative approach [369]. In such a framework, specialized agents could explore parallel reasoning paths, propose complementary retrieval strategies, and collaboratively synthesize findings to construct comprehensive search plans. This collaborative paradigm not only simulates diverse perspectives but also facilitates intricate interactions between knowledge sources and reasoning steps. By evaluating search plans from multiple angles, such systems can balance effectiveness and efficiency, ensuring robust performance across diverse query types.
- **Query Reformulation and Semantic Alignment:** A second major challenge is query reformulation, particularly in maintaining semantic alignment between the original multimodal query intent and the reformulated queries [197]. As queries become more sophisticated, accurately capturing and maintaining their intent grows increasingly complex. This challenge is amplified in multimodal contexts, where queries may integrate text, images, audio, or other data types, each requiring precise interpretation. To address this, multi-perspective reformulation strategies could be employed, leveraging diverse interpretations of the query to generate reformulations that better align with the original intent. Such strategies might integrate contextual understanding, domain-specific knowledge, and cross-modal alignment techniques to ensure semantic consistency with the user's intent.
- **Comprehensive Evaluation Benchmarks:** The third critical challenge is the absence of comprehensive evaluation benchmarks capable of assessing planning mechanisms across diverse query complexities and scenarios. Existing benchmarks often focus on narrow performance aspects, failing to capture the full spectrum of real-world applications. To address this gap, future benchmarks should evaluate systems across multiple dimensions, including adaptability to query diversity, robustness in handling complex queries, and efficiency in resource utilization. These benchmarks should incorporate a wide range of query types, from simple factual retrievals to multi-hop reasoning and creative tasks, ensuring rigorous testing under realistic conditions. Additionally, benchmarks should incorporate metrics for semantic alignment in query reformulation, computational efficiency, and scalability.

These interconnected challenges highlight the need for future research to develop adaptive planning mechanisms capable of addressing both query diversity and complexity. This could involve multi-agent coordination for advanced cases, alongside robust query reformulation and comprehensive evaluation frameworks.

6.3 Retrieval

Multimodal retrieval has made significant progress but continues to face challenges that can be categorized into methodological and practical issues. These challenges arise from the inherent complexity of integrating and retrieving information across diverse data modalities such as text, images, audio, and video. Below, we outline the key challenges in this field:

- **Heterogeneity of Cross-Modal Data**: The heterogeneity of data across modalities poses a significant challenge in multimodal retrieval and representation learning. Text, being sequential and discrete, relies on syntactic and semantic structures best captured by language models, while images, being spatial and continuous, require convolutional or transformer-based architectures to extract hierarchical visual features. This structural divergence complicates cross-modal alignment and comparison, as each modality demands specialized feature extraction techniques tailored to its unique characteristics. Extracting meaningful and comparable features from each modality is non-trivial, requiring domain-specific expertise and sophisticated models capable of capturing nuanced data properties. For instance, while transformers excel in processing sequential data like text, their adaptation to spatial data like images often necessitates architectural modifications, such as vision transformers (ViTs), to handle pixel arrays. Aligning these features into a unified representation space that preserves cross-modal semantic relationships remains a major challenge. Current approaches, including cross-modal transformers and MLLMs, often fail to create a common embedding space that adequately captures the semantic richness of each modality while ensuring inter-modal consistency.
- **Cross-modal components (reranker, refiner)**: While the dual-tower architecture has made significant strides in first-stage retrieval by efficiently encoding and aligning multimodal data (e.g., text and images), developing advanced reranking models that enable fine-grained multimodal interaction remains a challenge. Additionally, refining external multimodal knowledge post-retrieval and reranking remains underexplored, despite its potential to enhance result accuracy and relevance. Addressing these gaps requires innovative methodologies that leverage MLLMs and LLMs to enable sophisticated cross-modal understanding and reasoning.

6.4 Generation

The multimodal module in MRAG achieves human-aligned sensory representation through diversified modality integration, which significantly enhances user experience and system usability. However, achieving these enhancement objectives entails addressing shared challenges across both QA systems and multimodal generation:

- **Multimodal Input**: Multimodal systems face the challenge of integrating diverse data structures and representations across modalities such as text, images, audio, and video. As multimodal models evolve, they are increasingly required to process arbitrary combinations of modalities (e.g., text+image, text+video, image+audio). This necessitates a highly flexible and adaptive framework capable of dynamically accommodating diverse input configurations. Such frameworks must be modality-agnostic, enabling seamless integration of any input combination without predefined structures or extensive retraining. Achieving this flexibility involves designing architectures that generalize across modalities, extract relevant features, and fuse them meaningfully, regardless of input composition.

- **Multimodal Output**:
  - **Coherent and Contextually Relevant Generation**: Ensuring consistency across different modalities in the output presents a significant challenge. For instance, in a text-image pair, the image must accurately reflect the scene or object described in the text, while the text should precisely convey the visual content.
  - **Positioning and Integration of Multimodal Elements**: In multimodal outputs, such as text with embedded images or videos, the model must intelligently determine where to integrate non-textual elements. This requires an understanding of the narrative flow and the identification of optimal insertion points to enhance coherence and readability. Additionally, the model should dynamically generate or retrieve relevant multimodal content based on context. For instance, when creating a text-image pair, the model may need to generate an image caption, search for relevant images, and select the most appropriate one. This process must be efficient and seamless to ensure the final output is both relevant and high-quality.
  - **Diversity of Outputs**: In some applications, generating diverse outputs—such as multiple images or videos corresponding to a given text description—is essential. However, balancing diversity with relevance and quality poses a significant challenge. The model must explore a broad range of possibilities while ensuring each output remains contextually appropriate and adheres to high-quality standards.

6.5 Dataset & Evaluation

The advancement of MLLMs has heightened the need for comprehensive evaluation. Despite the introduction of over a hundred benchmarks by both academic and industrial communities, several challenges remain in the current evaluation landscape. First, there is a lack of a universally accepted, standardized capability taxonomy, with existing benchmarks often defining their own disparate ability dimensions. Second, current benchmarks exhibit significant gaps in critical areas such as instruction following, complex multimodal reasoning, multi-turn dialogue, and creativity assessment. Third, task-specific evaluations for MLLMs are insufficient, particularly in commercially relevant domains like invoice recognition, multimodal knowledge base comprehension, and UI understanding and industry. Finally, while existing multimodal benchmarks primarily focus on image and video modalities, there is a notable deficit in assessing capabilities related to audio and 3D representations. Addressing these challenges is essential for developing more robust and comprehensive evaluation methodologies for MLLMs in the future.

Despite rapid advancements, current evaluations of MLLMs remain insufficiently comprehensive, primarily focusing on perception and reasoning abilities through objective questions. This creates a significant gap between evaluation methodologies and real-world applications. Moreover, optimizing models based on objective assessments often leads developers to prioritize objective question corpora during instruction tuning, potentially degrading the quality of dialogue experiences. Although subjective multimodal evaluation platforms like WildVision and OpenCompass MultiModal Arena have emerged, further research is needed to develop assessment methods that better align with practical usage scenarios. Current evaluation strategies predominantly rely on curated or crafted questions to assess specific capabilities, yet complex multimodal tasks typically require the integration of multiple skills. For instance, a chart-related question may involve OCR, spatial relationship recognition, reasoning, and calculations. The absence of decoupled assessments for these distinct capabilities represents a major limitation in existing frameworks. Additionally, crucial abilities such as instruction following remain under-evaluated. Multiturn dialogue, the primary mode of human interaction with multimodal models, remains a weakness for most models, and corresponding evaluations, are still in their infancy. In the realm of complex multimodal reasoning, current evaluations predominantly focus on mathematical and examination problems, necessitating

improvements in both difficulty and relevance to everyday use cases. Notably, the evaluation of multimodal creative tasks, a key application area for these models—such as text generation based on image and textual prompts—remains largely unexplored, highlighting a critical gap in the current evaluation landscape.

MLLMs are still in the early stages of development, with limited business applications to date. As a result, current evaluations primarily focus on assessing foundational capabilities rather than real-world performance. Moving forward, it is critical to develop evaluation frameworks that measure MLLM performance on specific tasks with commercial value, such as large-scale document processing, multimodal knowledge base comprehension, anomaly detection, and industrial visual inspection. When designing task-specific evaluations, it is essential to consider not only performance metrics but also computational costs and inference speeds, benchmarking them against traditional computer vision methods like OCR, object detection, and action recognition to determine practical applicability. Additionally, a key potential of MLLMs lies in their ability to plan and interact with environments as agents to solve complex problems. Developing diverse virtual environments for MLLMs to demonstrate agent-based problem-solving capabilities will likely become a critical component of future evaluations. Current efforts in this domain remain nascent, highlighting a promising area for future research in multimodal AI assessment.

7 Future Directions

In this chapter, we propose several suggestions to the future development of multimodal Retrieval-Augmented Generation (MRAG) systems, informed by related research and identified challenges. These recommendations collectively aim to overcome existing limitations and unlock the full potential of MRAG in complex, real-world scenarios.

7.1 Documents Parsing

Multimodal document parsing has become a crucial element in MRAG systems, particularly with the emergence of large language models (LLMs) and multimodal large models (MLLMs). The fusion of text, images, and other data types into a cohesive framework presents both transformative opportunities and notable challenges. This paper provides a detailed analysis of future directions in this evolving field.

- **Enhancing Data Accuracy and Completeness:**
  - **Contextual Relationship Preservation:** To improve the accuracy and coherence of multimodal document parsing, especially for long and complex documents, advanced algorithms are needed to capture and preserve both inter-page and intra-document relationships. Techniques such as graph-based representations and hierarchical document modeling can help maintain contextual coherence across the document. These methods enable the system to understand structural and semantic dependencies between sections, tables, figures, and other elements, ensuring the preservation of the document's logical flow. Additionally, cross-referencing mechanisms are essential for linking related content across pages. These mechanisms dynamically connect sections, tables, and figures, facilitating seamless retrieval and utilization of contextual relationships in downstream tasks like information extraction, summarization, or question answering. By integrating these approaches, the system can better handle the complexities of long documents, ensuring accurate maintenance and leveraging of contextual relationships for enhanced performance in multimodal document understanding tasks. This is particularly relevant when combining Optical Character Recognition (OCR), LLMs, and MLLMs to process and interpret documents with diverse content types.

  - **Error Detection and Correction:** To improve the accuracy and reliability of multimodal document parsing, integrating advanced error detection and correction mechanisms is crucial. Leveraging LLMs and MLLMs, systems can validate extracted text against the original document, identifying and correcting inaccuracies or omissions. These models can be enhanced with consistency-checking algorithms to ensure coherence and accuracy across multimodal data, including text, images, and tables. For critical documents, a human-in-the-loop (HITL) approach is advisable. This involves human reviewers verifying and refining parsed data, especially in cases where systems may struggle with complex layouts, ambiguous content, or domain-specific nuances. By combining the strengths of LLMs, MLLMs, and human expertise, this hybrid approach ensures high accuracy and reliability, making it suitable for precision-demanding applications such as legal, medical, or financial document processing.
- **Improving Multimodal Data Integration:**
  - **Unified Multimodal Representation:** Advancing multimodal document parsing requires the development of unified representation frameworks that seamlessly integrate diverse data types, such as text, images, and tables, into a cohesive structure. Such frameworks enable robust, context-aware analysis by leveraging multimodal transformers—like CLIP, Flamingo, or other state-of-the-art models—to encode disparate modalities into a shared embedding space. This interoperability enhances downstream tasks, including information extraction, question answering, and summarization. A promising approach involves hybrid strategies that combine raw multimodal data with textual representations. For instance, raw images can support visual tasks (e.g., object detection or layout analysis), while textual captions or OCR-derived text can improve text generation tasks (e.g., summarization or translation). This dual methodology leverages the strengths of each modality, ensuring both accuracy and efficiency in processing complex documents, as demonstrated by recent research. Additionally, integrating LLMs and MLLMs with Optical Character Recognition (OCR) systems can enhance the parsing of scanned or image-based documents. By aligning OCR outputs with multimodal embeddings, these systems improve the handling of noisy or unstructured data, enabling more accurate interpretation and contextual understanding.
  - **Advanced Captioning and Description Generation:** To improve multimodal data integration, particularly in document parsing, enhancing automated captioning and description generation for non-textual elements like images, tables, and charts is critical. Leveraging state-of-the-art vision-language models (VLMs) and MLLMs can boost the accuracy and contextual relevance of textual descriptions. These models bridge the gap between visual and textual data, enabling more comprehensive document understanding. Integrating domain-specific knowledge into captioning models is essential for generating accurate and contextually tailored descriptions. This can be achieved by fine-tuning pre-trained models on domain-specific datasets or incorporating external knowledge bases. Such an approach ensures that descriptions align with the document's content, enhancing the utility of multimodal data integration.
- **Leveraging LLMs and MLLMs for Enhanced Parsing:**
  - **LLM/MLLM-Driven Parsing and Indexing:** LLMs and MLLMs can be fine-tuned on domain-specific corpora to improve their ability to parse and interpret complex document structures. Leveraging their advanced multimodal understanding, these models can accurately identify and extract key information—such as legal clauses, scientific hypotheses, or technical specifications—even from dense or unstructured text. Fine-tuning enhances their proficiency in recognizing domain-specific terminology, relationships, and contextual nuances. Furthermore, LLMs and MLLMs can generate metadata, tags, and summaries. By automatically annotating documents with relevant keywords, classifications, or concise summaries, these models streamline the organization and accessibility of large document repositories. This capability

is particularly valuable in applications like legal case management, academic research, and enterprise knowledge bases. In multimodal contexts, MLLMs extend these capabilities by integrating and interpreting data from diverse sources, such as text, images, tables, and diagrams. This enables a more comprehensive parsing process, where visual and textual elements are jointly analyzed to extract richer, more accurate information. For example, in scientific documents, MLLMs can parse and correlate data from textual descriptions and accompanying charts, facilitating a deeper understanding of the content.

- **Bridging the Gap Between LLMs and MLLMs:** A promising approach involves hybrid architectures that combine the strengths of MLLMs and LLMs. MLLMs process raw multimodal inputs (e.g., images, audio, video) to extract meaningful representations, while LLMs generate coherent and contextually accurate textual outputs. This division of labor optimizes performance, as MLLMs excel in multimodal feature extraction and LLMs in linguistic precision. For example, in document parsing, MLLMs analyze visual layouts, tables, or embedded graphics, while LLMs synthesize this information into structured textual formats.

7.2 Multimodal Search Planning

The future of multimodal search planning should focus on addressing three key challenges within the hierarchical framework: intelligent adaptive planning mechanisms, query reformulation and semantic alignment, and comprehensive evaluation benchmarks. Below are targeted suggestions for advancing each area.

- **Intelligent Adaptive Planning Mechanisms:**
  - **Multi-Agent Collaborative Systems:** To address the challenges of multimodal search and complex query resolution, multi-agent collaborative systems can be designed to leverage specialized agents working in tandem. These systems enhance efficiency, adaptability, and robustness in handling multi-hop, creative, or cross-modal queries. Key mechanisms include: 1) Parallel Reasoning Paths: Specialized agents can simultaneously explore multiple reasoning trajectories, enabling faster and more comprehensive solutions. This approach is particularly effective for multi-hop queries, where intermediate reasoning steps are critical, or for creative tasks requiring diverse perspectives. By evaluating multiple pathways in parallel, the system can identify optimal solutions while mitigating the risk of local optima. 2) Complementary Retrieval Strategies: Agents can employ diverse retrieval methodologies, such as keyword-based, semantic, or cross-modal retrieval, to address different aspects of a query. For instance, one agent might focus on extracting structured data, while another leverages semantic embeddings or visual-textual alignment for multimodal contexts. The synthesis of these strategies ensures robust and contextually relevant search outcomes, enhancing the system's ability to handle heterogeneous data sources. 3) Dynamic Resource Allocation: Agents can monitor computational resources and system constraints in real-time, dynamically adjusting retrieval strategies to optimize performance. For example, under limited computational bandwidth, agents might prioritize lightweight retrieval methods or redistribute tasks to balance load. This adaptive mechanism ensures efficient resource utilization while maintaining high-quality query resolution. 4) Integration with MLLMs and LLMs: The collaborative multi-agent framework can be seamlessly integrated with MLLMs and LLMs to enhance their capabilities. MLLMs can serve as central orchestrators, interpreting multimodal inputs and coordinating agent tasks, while LLMs provide deep contextual understanding and reasoning support. This integration enables the system to handle complex, multimodal queries with greater precision and adaptability.
  - **Hierarchical Planning Frameworks:** To address the challenges of ambiguous or highly creative queries in multimodal search and planning, integrating human feedback into the

decision-making process is essential. Human-in-the-loop (HITL) systems facilitate iterative refinement by leveraging user expertise to guide and validate intermediate results. These interactive systems enable users to dynamically adjust search parameters, prioritize modalities, or correct misinterpretations, ensuring more accurate and contextually relevant outcomes. By combining the strengths of multimodal large language models (MLLMs) with human intuition, HITL systems enhance adaptability, build trust, and improve the robustness of intelligent planning frameworks. This collaborative approach is particularly valuable in domains requiring nuanced understanding, creativity, or domain-specific knowledge.

  - **Reinforcement Learning for Adaptation:** Intelligent adaptive planning mechanisms can be developed using reinforcement learning (RL) to enable dynamic, context-aware decision-making. By modeling the search process as a sequential decision problem, RL agents can be trained to optimize resource allocation and retrieval accuracy. These agents adapt their strategies by receiving rewards for minimizing computational overhead, reducing latency, and delivering precise results tailored to query characteristics, such as modality, complexity, and user intent.

- **Query Reformulation and Semantic Alignment:**
  - **Multi-Perspective Reformulation:** To address the complexity of multimodal search, query reformulation strategies must generate diverse interpretations while preserving the original intent across modalities. This involves: 1) Contextual Understanding: Leveraging contextual embeddings (e.g., from transformer-based models) to capture semantic nuances and contextual dependencies, ensuring reformulated queries retain the richness of the original input. 2) Cross-Modal Alignment: Employing advanced techniques like contrastive learning to align representations across text, images, and audio modalities. By embedding queries and multimodal data into a shared latent space, this ensures consistent interpretation and retrieval across diverse data types. 3) Domain-Specific Knowledge Integration: Incorporating domain-specific ontologies or knowledge graphs to enhance reformulation accuracy, particularly in specialized fields. This leverages structured domain knowledge to improve the relevance and precision of reformulated queries.
  - **Interactive Query Refinement:** A pivotal strategy is interactive query refinement, which allows users to iteratively adjust queries based on intermediate results. Intelligent systems can facilitate this process by suggesting alternative query formulations, identifying ambiguities, and providing contextual feedback to better align queries with user intent. By integrating user feedback loops and real-time semantic analysis, these systems dynamically bridge the gap between user input and multimodal data, ensuring more precise and contextually relevant search outcomes.
  - **Explainable Reformulation:** A critical aspect of query reformulation is its explainability. By offering clear and concise explanations for how queries are transformed, users gain insight into the system's reasoning and decision-making processes. For example, when a user submits a vague or ambiguous query, the system can generate a reformulated version while detailing the rationale behind the changes, such as term disambiguation, incorporation of contextual cues, or alignment with multimodal data (e.g., text, images, or audio). This transparency fosters user trust, enables validation of the system's interpretation, and enhances user control and satisfaction. Furthermore, explainable reformulation underscores the importance of semantic alignment, where the system bridges the gap between user intent and the underlying data representation, ensuring the reformulated query accurately reflects the user's needs.
- **Comprehensive Evaluation Benchmarks:**
  - **Diverse Query Datasets:** It is crucial to establish robust benchmarks. These benchmarks should incorporate diverse query datasets spanning a wide range of query types, from simple

factual retrievals to complex multi-hop reasoning and creative tasks. The datasets must reflect real-world heterogeneity in query patterns and modalities, capturing the intricacies of user interactions across text, image, audio, and video inputs. By integrating such diversity, benchmarks can more accurately assess model performance, generalization capabilities, and adaptability to varied real-world applications.

- **Multi-Dimensional Metrics:** To ensure the effectiveness and reliability of Multimodal Search Planning systems, robust evaluation frameworks must be established. These frameworks should employ multi-dimensional metrics to comprehensively assess system performance across diverse operational scenarios. Key dimensions include: 1) Adaptability: The system's ability to handle a broad spectrum of query types, from simple to highly complex, while integrating multiple modalities (e.g., text, images, audio). This metric evaluates the model's flexibility in addressing varied user needs and its capacity to generalize across domains. 2) Robustness: The system's resilience under challenging conditions, such as computational constraints, noisy or incomplete inputs, and adversarial scenarios. Robustness ensures consistent performance in real-world applications, where ideal conditions are seldom present. 3) Efficiency: The optimization of resource utilization (e.g., memory, processing power) and response time. This metric is critical for scalability and user satisfaction, especially in time-sensitive or resource-constrained environments. 4) Semantic Alignment: The system's accuracy in preserving and interpreting the intent of user queries during reformulation or multimodal integration. This ensures that outputs remain contextually and semantically aligned with the user's original request.

7.3 Retrieval

The challenges in multimodal retrieval underscore the complexity of integrating and retrieving information across diverse data types. To address these issues and advance the field, future research should prioritize the following directions:

- **Unified Cross-Modal Representation Learning**: The primary objective is to develop robust and unified representation learning frameworks that effectively align and compare data across diverse modalities, including text, images, audio, and video. A key element of this framework is the enhancement of cross-modal attention mechanisms to better model complex interactions between modalities. Cross-modal attention layers, inspired by transformer architectures, are central to capturing fine-grained relationships. These mechanisms enable one modality to focus on relevant features in another, allowing the model to dynamically prioritize the most informative aspects of the data. For example, text can guide attention over visual regions, or audio cues can emphasize relevant temporal segments in video data. Techniques such as multi-head cross-modal attention and hierarchical attention further refine this process, ensuring robust and context-aware representations.
- **Cross-Modal Context**: The primary objective is to improve the ability of models to perform fine-grained interactions between modalities, particularly in the reranking and refinement stages.
  - **Reranker:** Multi-Modal Reranking Models rerank retrieved document list by incorporating detailed cross-modal interactions, such as text-image, text-audio, or video-text relationships. By integrating LLMs and MLLMs, reranking models enhance their ability to capture nuanced semantic alignments between modalities.
  - **Refiner:** Leveraging their cross-modal reasoning abilities, MLLMs refine retrieval results through knowledge-enhanced refinement, yielding more accurate, contextually relevant, and semantically rich outputs. This refinement process utilizes MLLMs' contextual understanding and multimodal alignment to re-rank, filter, or augment retrieved content.

7.4 Generation

The future of multimodal generation should focus on overcoming existing challenges on multimodal input and output. Below are key suggestions for advancing multimodal generation systems.

- **Flexible and Adaptive Multimodal Input Frameworks:** To address the growing complexity of multimodal data, it is crucial to develop modality-agnostic architectures that dynamically adapt to diverse and arbitrary input modality combinations, such as text+image, text+video, or image+audio. These frameworks should process inputs without relying on predefined structures or extensive retraining.
- **Coherent and Contextually Relevant Multimodal Output:** Achieving coherent and contextually relevant outputs in multimodal generation necessitates the development of advanced models capable of maintaining consistency across modalities. For example, in text-image generation tasks, the generated image must precisely align with the textual description, and the text should accurately reflect the visual content of the image. This cross-modal consistency is essential for ensuring the reliability and usability of multimodal systems.
- **Intelligent Positioning and Integration of Multimodal Elements:** To seamlessly integrate non-textual elements (e.g., images, videos, audio) into a narrative, models must be trained to identify optimal insertion points. This requires a deep understanding of the content's structure, flow, and contextual nuances to ensure coherence, readability, and enhanced user engagement. Advanced techniques, such as attention mechanisms, can analyze the narrative's semantic and syntactic structure, enabling the model to determine where multimodal elements can complement or enrich the text. Modern multimodal systems must dynamically retrieve or generate contextually relevant non-textual content. For example, when generating a text-image pair, the model should use cross-modal alignment techniques to either retrieve an existing image from a database or synthesize a new one that aligns with the textual context. This relies on robust multimodal representation learning, where embeddings from different modalities (text, image, video) are mapped into a shared latent space, enabling precise cross-modal retrieval or generation.
- **Diversity in Multimodal Outputs:** Achieving a balance between diversity, relevance, and quality in multimodal generation requires controlled mechanisms. For instance, in text-to-image generation, models should produce diverse yet faithful representations of textual descriptions. Techniques like conditional sampling can guide models to explore varied latent spaces while adhering to input constraints.

7.5 Dataset & Evaluation

The future direction of datasets and evaluation in MRAG should focus on addressing current gaps and challenges while harnessing the unique capabilities of MLLMs. Below are refined suggestions for advancing datasets and evaluation methodologies in this field.

- **Comprehensive Benchmark Development**: To enhance the evaluation of MLLMs and LLMs in retrieval-augmented generation, it is crucial to develop comprehensive benchmarks that address key limitations in current assessment frameworks. These benchmarks should focus on the following areas: 1) Instruction Following: Create tasks to evaluate the model's ability to comprehend and execute complex, multi-step instructions across diverse modalities. This includes assessing precision in adhering to nuanced directives and handling ambiguous or incomplete inputs. 2) Multiturn Dialogue: Develop datasets that simulate real-world conversational dynamics, emphasizing the model's capacity for context retention, coherence, and adaptability over extended interactions. Scenarios should include cross-modal references and long-term memory challenges. 3) Complex Multimodal Reasoning: Design tasks requiring the integration of multiple modalities

(e.g., text, images, audio) to solve real-world problems, such as interpreting charts, maps, or combining visual and textual data for decision-making. 4) Creativity Evaluation: Introduce benchmarks to assess generative capabilities in creative tasks, such as composing stories, poems, or designing visual artifacts from multimodal inputs. These tasks should measure originality, relevance, and the ability to synthesize diverse inputs into coherent outputs. 5) Diverse Modalities: Expand evaluation frameworks to include emerging modalities like audio, 3D models, and sensor data, ensuring robustness and versatility in handling a wide range of input types.

- **Multimodal Retrieval-Augmented Generation**: The development of robust metrics for evaluating retrieval and generation in multimodal systems requires assessing relevance, precision, diversity, and cross-modal alignment to ensure semantic consistency and contextual appropriateness. Metrics should quantify the system's ability to filter noise and redundancy, delivering concise and meaningful outputs. For generation quality, coherence, fluency, creativity, and adaptability are essential, alongside factual accuracy and consistency with retrieved data and external knowledge. Effective multimodal integration is crucial to unify diverse inputs into contextually rich outputs. Comprehensive benchmarks must simulate real-world scenarios, incorporating varied queries, multimodal sources, and differing complexity levels to evaluate the end-to-end performance of retrieval-augmented generation (RAG) pipelines.

8 Conclusion

In conclusion, this survey comprehensively examines the emerging field of Multimodal Retrieval-Augmented Generation (MRAG), highlighting its potential to enhance the capabilities of large language models (LLMs) by integrating multimodal data such as text, images, and videos. Unlike traditional text-based RAG systems, MRAG addresses the challenges of retrieving and generating information across different modalities, thereby improving the accuracy and relevance of responses while reducing hallucinations. The survey systematically analyzes MRAG from four key perspectives: essential components and technologies, datasets, evaluation methods and metrics, and existing limitations. It identifies current challenges, such as effectively integrating multimodal knowledge and ensuring the reliability of generated outputs, while also proposing future research directions. By providing a structured overview and forward-looking insights, this survey aims to guide researchers in advancing MRAG, ultimately contributing to the development of more robust and versatile Multimodal Retrieval-Augmented Generation.

References

[1] [n. d.]. jsoup. https://jsoup.org/
[2] [n. d.]. pdfminer. https://github.com/pdfminer/pdfminer.six
[3] [n. d.]. PyMuPDF. https://github.com/pymupdf/PyMuPDF
[4] [n. d.]. realworldQA. https://huggingface.co/datasets/visheratin/realworldqa
[5] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing* 22, 10 (2014), 1533–1545.
[6] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase,

Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[8] Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. 2023. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564* (2023).

[9] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[10] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[11] Muhammad Asif Ali, Zhengping Li, Shu Yang, Keyuan Cheng, Yang Cao, Tianhao Huang, Guimin Hu, Weimin Lyu, Lijie Hu, Lu Yu, et al. 2024. Prompt-saw: Leveraging relation-aware graphs for textual prompt compression. *arXiv preprint arXiv:2404.00489* (2024).

[12] 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*. 993–1003.

[13] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).

[14] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).

[15] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* 1, 2 (2023), 3.

[17] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890* (2023).

[18] Liping Bao, Longhui Wei, Wengang Zhou, Lin Liu, Lingxi Xie, Houqiang Li, and Qi Tian. 2023. Multi-Granularity Matching Transformer for Text-Based Person Search. *IEEE Transactions on Multimedia* (2023).

[19] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems* 35 (2022), 31668–31683.

[20] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. arXiv:2407.07726 [cs.CV] https://arxiv.org/abs/2407.07726

[21] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595* (2023).

[22] Chinmoy B Bose and Shyh-Shiaw Kuo. 1994. Connected and degraded text recognition using hidden Markov model. *Pattern Recognition* 27, 10 (1994), 1345–1363.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[24] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1818–1826.

[25] Zhiwei Cao, Qian Cao, Yu Lu, Ningxin Peng, Luyang Huang, Shanbo Cheng, and Jinsong Su. 2024. Retaining key information under high compression ratios: Query-guided compressor for llms. *arXiv preprint arXiv:2406.02376* (2024).

[26] Bing-Bing Chai, Jozsef Vass, and Xinhua Zhuang. 1999. Significance-linked connected component analysis for wavelet image coding. *IEEE Transactions on Image processing* 8, 6 (1999), 774–784.

[27] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932* (2020).

[28] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16495–16504.

[29] Yunhang Shen Yulei Qin Mengdan Zhang Xu Lin Jinrui Yang Xiawu Zheng Ke Li Xing Sun Yunsheng Wu Rongrong Ji Chaoyou Fu, Peixian Chen. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394* (2023).

[30] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160* (2023).

[31] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26540–26550.

[32] Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029* (2023).

[33] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.

[34] Jieneng Chen, Luoxin Ye, Ju He, Zhaoyang Wang, Daniel Khashabi, and Alan L Yuille. 2024. Efficient large multi-modal models via visual context compression. *Advances in Neural Information Processing Systems* 37 (2024), 73986–74007.

[35] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* (2023).

[36] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195* (2023).

[37] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. Springer, 370–387.

[38] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330* (2024).

[39] Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160* (2024).

[40] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565* (2023).

[41] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465* (2021).

[42] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713* (2023).

[43] Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiaxin Mao, and Dawei Yin. 2024. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. *arXiv preprint arXiv:2406.11678* (2024).

[44] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14239–14250.

[45] Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025. Improving Retrieval-Augmented Generation through Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2501.15228* (2025).

[46] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
[47] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. 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*. 24185–24198.
[48] Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. *arXiv preprint arXiv:2405.13792* (2024).
[49] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. *arXiv preprint arXiv:2305.14788* (2023).
[50] Kyunghyun Cho. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
[51] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[52] Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, and Jong C Park. 2023. Discrete prompt optimization via constrained generation for zero-shot re-ranker. *arXiv preprint arXiv:2305.13729* (2023).
[53] Eunbi Choi, Yongrae Jo, Joel Jang, and Minjoon Seo. 2022. Prompt injection: Parameterization of fixed inputs. *arXiv preprint arXiv:2206.11349* (2022).
[54] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* 1, 2 (2023), 3.
[55] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766* (2024).
[56] Yu-Neng Chuang, Tianwei Xing, Chia-Yuan Chang, Zirui Liu, Xun Chen, and Xia Hu. 2024. Learning to compress prompt in natural language formats. *arXiv preprint arXiv:2402.18700* (2024).
[57] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8415–8424.
[58] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287* (2023).
[59] Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*. 1897–1907.
[60] Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1533–1536.
[61] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904* (2020).
[62] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[63] SeungHeon Doh, Minhee Lee, Dasaem Jeong, and Juhan Nam. 2024. Enriching Music Descriptions with A Finetuned-LLM and Metadata for Text-to-Music Retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 826–830.
[64] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.
[65] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499* (2023).
[66] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420* (2024).
[67] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. arXiv:2401.16420 [cs.CV] https://arxiv.org/abs/2401.16420

[68] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[69] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).
[70] Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler, et al. 2023. PaRaDe: Passage ranking using demonstrations with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 14242–14252.
[71] Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. 2024. Towards Event-oriented Long Video Understanding. *arXiv preprint arXiv:2406.14129* (2024).
[72] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2018. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3984–3993.
[73] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. 2010. Detecting text in natural scenes with stroke width transform. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2963–2970.
[74] 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*. 150–158.
[75] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).
[76] Minghui Fang, Shengpeng Ji, Jialong Zuo, Hai Huang, Yan Xia, Jieming Zhu, Xize Cheng, Xiaoda Yang, Wenrui Liu, Gang Wang, et al. 2024. Ace: A generative cross-modal retrieval framework with coarse-to-fine semantic modeling. *arXiv preprint arXiv:2406.17507* (2024).
[77] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding. *arXiv preprint arXiv:2406.14515* (2024).
[78] 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:2407.01449 [cs.IR] https://arxiv.org/abs/2407.01449
[79] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.
[80] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.
[81] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences* 67, 12 (2024), 1–14.
[82] Paolo Ferragina and Giovanni Manzini. 2000. Opportunistic data structures with applications. In *Proceedings 41st annual symposium on foundations of computer science*. IEEE, 390–398.
[83] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086* (2021).
[84] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.
[85] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075* (2024).
[86] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*. Springer, 148–166.
[87] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15159–15168.
[88] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. 2024. Linguistic-Aware Patch Slimming Framework for Fine-grained Cross-Modal Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26307–26316.
[89] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 214–229.

[90] Jun Gao, Ziqiang Cao, and Wenjie Li. 2024. SelfCP: Compressing over-limit prompt via the frozen large language model itself. *Information Processing & Management* 61, 6 (2024), 103873.
[91] Jun Gao, Ziqiang Cao, and Wenjie Li. 2024. Unifying demonstration selection and compression for in-context learning. *arXiv preprint arXiv:2405.17062* (2024).
[92] Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253* (2021).
[93] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).
[94] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186* (2021).
[95] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*. Springer, 146–160.
[96] Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. 2024. Clova: A closed-loop visual assistant with tool usage and update. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13258–13268.
[97] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 10826–10834.
[98] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2013. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence* 36, 4 (2013), 744–755.
[99] Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945* (2023).
[100] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041* (2023).
[101] Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification. *arXiv preprint arXiv:2406.14496* (2024).
[102] Peiyuan Gong, Jiamian Li, and Jiaxin Mao. 2024. Cosearchagent: a lightweight collaborative search agent with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2729–2733.
[103] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).
[104] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.
[105] Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.
[106] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).
[107] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* 35 (2022), 26418–26431.
[108] Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18135–18143.
[109] Fang Guo, Wenyu Li, Honglei Zhuang, Yun Luo, Yafu Li, Qi Zhu, Le Yan, and Yue Zhang. 2024. Generating diverse criteria on-the-fly to improve point-wise LLM rankers. *arXiv preprint arXiv:2404.11960* (2024).
[110] Weikuo Guo, Huaibo Huang, Xiangwei Kong, and Ran He. 2019. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1712–1720.
[111] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. 2018. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval* 7 (2018), 87–93.
[112] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.
[113] Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, et al. 2023. InfiMM-Eval: Complex Open-Ended Reasoning Evaluation For Multi-Modal Large Language Models. *arXiv e-prints* (2023), arXiv–2311.

[114] Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7879–7886.

[115] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008* (2024).

[116] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[117] Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. CMMU: A Benchmark for Chinese Multi-modal Multi-type Question Understanding and Reasoning. *arXiv preprint arXiv:2401.14011* (2024).

[118] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652* (2017).

[119] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[120] Sebastian Hofstätter, Omar Khattab, Sophia Althammer, Mete Sertkan, and Allan Hanbury. 2022. Introducing neural bag of whole-words with colberter: Contextualized late interactions using enhanced reduction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 737–747.

[121] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14281–14290.

[122] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301* (2023).

[123] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6929–6938.

[124] Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038* (2023).

[125] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models. *arXiv preprint arXiv:2410.08182* (2024).

[126] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models. *arXiv preprint arXiv:2410.08182* (2024).

[127] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2553–2561.

[128] Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. 2024. Efficient Multi-modal Large Language Models via Visual Token Grouping. *arXiv preprint arXiv:2411.17773* (2024).

[129] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23802–23804.

[130] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems* 36 (2023), 72096–72109.

[131] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6565–6574.

[132] Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, and Mao Yang. 2023. Fewer is more: Boosting LLM reasoning with reinforced context pruning. *arXiv preprint arXiv:2312.08901* (2023).

[133] 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*. 4083–4091.

[134] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463* (2023).

[135] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2310–2318.

[136] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12976–12985.

[137] S Humeau. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969* (2019).

[138] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1705–1715.

[139] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* 2, 3 (2021).

[140] Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2491–2498.

[141] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2758–2766.

[142] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.

[143] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3304–3311.

[144] 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 *International conference on machine learning*. PMLR, 4904–4916.

[145] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2023. Bootstrapping vision-language learning with decoupled language pre-training. *Advances in Neural Information Processing Systems* 36 (2023), 57–72.

[146] Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2787–2797.

[147] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, et al. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959* (2024).

[148] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736* (2023).

[149] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839* (2023).

[150] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580* (2024).

[151] Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025. What Kind of Visual Tokens Do We Need? Training-free Visual Token Pruning for Multi-modal Large Language Models from the Perspective of Graph. *arXiv preprint arXiv:2501.02268* (2025).

[152] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160* (2024).

[153] Bowen Jin, Hansi Zeng, Guoyin Wang, Xiusi Chen, Tianxin Wei, Ruirui Li, Zhengyang Wang, Zheng Li, Yang Li, Hanqing Lu, et al. 2023. Language models as semantic indexers. *arXiv preprint arXiv:2310.07815* (2023).

[154] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. 2024. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161* (2024).

[155] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669* (2023).

[156] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[157] Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multi-view learning for passage re-ranking. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1803–1807.
[158] Dongwon Jung, Qin Liu, Tenghao Huang, Ben Zhou, and Muhao Chen. 2024. Familiarity-aware evidence compression for retrieval augmented generation. *arXiv preprint arXiv:2409.12468* (2024).
[159] Hoyoun Jung and Kyung-Joong Kim. 2024. Discrete prompt compression with reinforcement learning. *IEEE Access* (2024).
[160] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300* (2017).
[161] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).
[162] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, et al. 2024. ReMI: A Dataset for Reasoning with Multiple Images. *arXiv preprint arXiv:2406.09175* (2024).
[163] 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*. 39–48.
[164] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. *Advances in Neural Information Processing Systems* 36 (2023), 21487–21506.
[165] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*. PMLR, 17283–17300.
[166] Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-aspect dense retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3178–3186.
[167] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
[168] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9579–9589.
[169] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 1995.
[170] Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. 2024. Unified Multimodal Interleaved Document Representation for Retrieval. arXiv:2410.02729 [cs.CL] https://arxiv.org/abs/2410.02729
[171] Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2020. Learning dense representations of phrases at scale. *arXiv preprint arXiv:2012.12624* (2020).
[172] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).
[173] 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 *International Conference on Machine Learning*. PMLR, 18893–18912.
[174] 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)*. 201–216.
[175] Sunkyung Lee, Minjin Choi, and Jongwuk Lee. 2023. GLEN: Generative retrieval via lexical index learning. *arXiv preprint arXiv:2311.03057* (2023).
[176] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362* (2023).
[177] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3108–3120.
[178] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790* (2024).
[179] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13299–13308.
[180] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).

[181] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425* (2023).

[182] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2023), 28541–28564.

[183] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11336–11344.

[184] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110* (2022).

[185] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[186] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[187] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

[188] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.

[189] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665* (2023).

[190] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1348–1357.

[191] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 13094–13102.

[192] Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349* (2023).

[193] Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yukun Yan, Shuo Wang, and Ge Yu. 2024. Say more with less: Understanding prompt learning behaviors through gist compression. *arXiv preprint arXiv:2402.16058* (2024).

[194] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 121–137.

[195] Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. 2024. Revolutionizing Text-to-Image Retrieval as Autoregressive Token-to-Voken Generation. *arXiv preprint arXiv:2407.17274* (2024).

[196] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201* (2023).

[197] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S. Yu, Fei Huang, and Jingren Zhou. 2024. Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset and Self-adaptive Planning Agent. (2024). arXiv:2411.02937 [cs.CL] https://arxiv.org/abs/2411.02937

[198] Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*. Springer, 323–340.

[199] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. *arXiv preprint arXiv:2402.10805* (2024).

[200] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Multiview identifiers enhanced generative retrieval. *arXiv preprint arXiv:2305.16675* (2023).

[201] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8716–8723.

[202] Yongqi Li, Zhen Zhang, Wenjie Wang, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Distillation Enhanced Generative Retrieval. *arXiv preprint arXiv:2402.10769* (2024).

[203] Zongqian Li, Yixuan Su, and Nigel Collier. 2024. 500xCompressor: Generalized Prompt Compression for Large Language Models. *arXiv preprint arXiv:2408.03094* (2024).

[204] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. 2024. LEGO: language enhanced multi-modal grounding model. *arXiv e-prints* (2024), arXiv–2401.

[205] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26763–26773.

[206] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. 2016. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. arXiv:1611.06779 [cs.CV] https://arxiv.org/abs/1611.06779

[207] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947* (2024).

[208] Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807* (2021).

[209] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26689–26699.

[210] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571* (2024).

[211] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems* 36 (2023), 22820–22840.

[212] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271* (2024).

[213] Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane Luke. 2024. Prompt compression with context-aware sentence encoding for fast and improved llm inference. *arXiv preprint arXiv:2409.01227* (2024).

[214] Alexander Liu and Samuel Yang. 2022. Masked autoencoders as the unified learners for pre-trained sentence representation. *arXiv preprint arXiv:2208.00231* (2022).

[215] Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. 2023. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing* 32 (2023), 3622–3633.

[216] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935* (2024).

[217] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774* (2023).

[218] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] https://arxiv.org/abs/2304.08485

[219] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[220] Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. *arXiv preprint arXiv:2310.15556* (2023).

[221] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955* (2024).

[222] Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2024. Leveraging passage embeddings for efficient listwise reranking with large language models. In *THE WEB CONFERENCE 2025*.

[223] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*. Springer, 126–142.

[224] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11915–11925.

[225] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. 2024. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194* (2024).

[226] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. 2024. Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model. *arXiv preprint arXiv:2411.10803* (2024).

[227] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. 2023. MMHQA-ICL: Multimodal In-context Learning for Hybrid Question Answering over Text, Tables and Images. *arXiv preprint arXiv:2309.04790* (2023).

[228] Wenhan Liu, Yutao Zhu, and Zhicheng Dou. 2024. Demorank: Selecting effective demonstrations for large language models in ranking task. *arXiv preprint arXiv:2406.16332* (2024).

[229] Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* 364 (2019).

[230] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player?. In *European conference on computer vision*. Springer, 216–233.

[231] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE international conference on computer vision*. 4107–4116.

[232] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. [n. d.]. Tempcompass: Do video llms really understand videos?, 2024c. *URL https://arxiv. org/abs/2403.00476* ([n. d.]).

[233] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences* 67, 12 (2024), 220102.

[234] Zejun Liu, Fanglin Chen, Jun Xu, Wenjie Pei, and Guangming Lu. 2022. Image-text retrieval with cross-modal semantic importance consistency. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 5 (2022), 2465–2476.

[235] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024. MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs. *arXiv preprint arXiv:2406.11833* (2024).

[236] Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Ziheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, et al. 2024. Controlllm: Augment language models with tools by searching on graphs. In *European Conference on Computer Vision*. Springer, 89–105.

[237] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv preprint arXiv:2209.00179* (2022).

[238] Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18733–18741.

[239] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. 2022. Audio-text retrieval in context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4793–4797.

[240] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*. 15692–15701.

[241] Junyu Lu, Dixiang Zhang, Songxin Zhang, Zejian Xie, Zhuoyang Song, Cong Lin, Jiaxing Zhang, Bingyi Jing, and Pingjian Zhang. 2023. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278* (2023).

[242] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* (2023).

[243] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[244] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797* (2024).

[245] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences. *arXiv preprint arXiv:2406.11069* (2024).

[246] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics* 9 (2021), 329–345.

[247] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[248] Jian Luo, Xuanang Chen, Ben He, and Le Sun. 2024. Prp-graph: Pairwise ranking prompting to llms with graph aggregation for effective text re-ranking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5766–5776.

[249] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419* (2023).
[250] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18051–18061.
[251] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. Prop: Pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 283–291.
[252] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Yingyan Li, and Xueqi Cheng. 2021. B-PROP: bootstrapped pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1513–1522.
[253] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying Multimodal Retrieval via Document Screenshot Embedding. arXiv:2406.11251 [cs.IR] https://arxiv.org/abs/2406.11251
[254] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251* (2024).
[255] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2425.
[256] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156* (2023).
[257] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2024. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*. Springer, 403–420.
[258] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523* (2024).
[259] Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024. Multi-modal Retrieval Augmented Multi-modal Generation: A Benchmark, Evaluate Metrics and Strong Baselines. arXiv:2411.16365 [cs.CL] https://arxiv.org/abs/2411.16365
[260] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).
[261] Raman Maini and Himanshu Aggarwal. 2009. Study and comparison of various image edge detection techniques. *International journal of image processing (IJIP)* 3, 1 (2009), 1–11.
[262] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.
[263] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.
[264] Julieta Martinez, Holger H Hoos, and James J Little. 2014. Stacked quantizers for compositional vector compression. *arXiv preprint arXiv:1411.2173* (2014).
[265] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761* (2023).
[266] 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* (2022).
[267] 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*. 1697–1706.
[268] 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*. 2200–2209.
[269] Lang Mei, Jiaxin Mao, Gang Guo, and Ji-Rong Wen. 2022. Learning Probabilistic Box Embeddings for Effective and Efficient Ranking. In *Proceedings of the ACM Web Conference 2022*. 473–482.
[270] Lang Mei, Jiaxin Mao, Juan Hu, Naiqiang Tan, Hua Chai, and Ji-Rong Wen. 2023. Improving first-stage retrieval of point-of-interest search by pre-training models. *ACM Transactions on Information Systems* 42, 3 (2023), 1–27.
[271] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang. 2022. On metric learning for audio-text cross-modal retrieval. *arXiv preprint arXiv:2203.15537* (2022).
[272] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3113–3124.

[273] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516* (2018).

[274] 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*. 2630–2640.

[275] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*. IEEE, 947–952.

[276] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*. 19–27.

[277] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18798–18806.

[278] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2024. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1314–1332.

[279] Jesse Mu, Xiang Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems* 36 (2023), 19327–19352.

[280] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2023), 25081–25094.

[281] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).

[282] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[283] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713* (2020).

[284] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6, 2 (2019).

[285] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[286] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2019. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 128–133.

[287] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992* (2023).

[288] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. 2024. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968* (2024).

[289] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799* (2023).

[290] Yanwei Pang, Yuan Yuan, Xuelong Li, and Jing Pan. 2011. Efficient HOG human detection. *Signal processing* 91, 4 (2011), 773–781.

[291] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 508–524.

[292] Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Top-down partitioning for efficient list-wise ranking. *arXiv preprint arXiv:2405.14589* (2024).

[293] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[294] Zhiyuan Peng, Xuyang Wu, Qifan Wang, Sravanthi Rajanala, and Yi Fang. 2024. Q-peft: Query-dependent parameter efficient fine-tuning for text reranking with large language models. *arXiv preprint arXiv:2404.04522* (2024).

[295] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167* (2023).

[296] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667* (2021).
[297] Xiao Pu, Tianxing He, and Xiaojun Wan. 2024. Style-Compress: An LLM-Based Prompt Compression Framework Considering Task-Specific Styles. *arXiv preprint arXiv:2410.14042* (2024).
[298] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236* (2024).
[299] Jinwei Qi, Yuxin Peng, and Yuxin Yuan. 2018. Cross-media multi-level alignment with relation attention network. *arXiv preprint arXiv:1804.09539* (2018).
[300] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284* (2024).
[301] Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. 2024. DiffusionGPT: LLM-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061* (2024).
[302] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563* (2023).
[303] Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2024. Unified text-to-image generation and retrieval. *arXiv preprint arXiv:2406.05814* (2024).
[304] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191* (2020).
[305] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[306] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[307] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13009–13018.
[308] David Rau, Shuai Wang, Hervé Déjean, and Stéphane Clinchant. 2024. Context embeddings for efficient answer generation in rag. *arXiv preprint arXiv:2407.09252* (2024).
[309] Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: Faster Improved Listwise Reranking with Single Token Decoding. *arXiv preprint arXiv:2406.15657* (2024).
[310] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26374–26383.
[311] TIMODAL RETRIEVAL, KNOWLEDGE-ENHANCED RERANKING, and NOISE-INJECTED TRAINING. [n. d.]. MLLM IS A STRONG RERANKER: ADVANCING MUL. ([n. d.]).
[312] Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. arXiv:2410.21943 [cs.CL] https://arxiv.org/abs/2410.21943
[313] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807* (2024).
[314] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60, 5 (2004), 503–520.
[315] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.
[316] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.
[317] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems* 37 (2024), 21999–22027.
[318] Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. *arXiv preprint arXiv:2204.07496* (2022).
[319] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[320] Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (1975), 613–620.

[321] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*. Springer, 146–162.

[322] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8876–8884.

[323] Shivam Shandilya, Menglin Xia, Supriyo Ghosh, Huiqiang Jiang, Jue Zhang, Qianhui Wu, and Victor Rühle. 2024. TACO-RL: Task Aware Prompt Compression Optimization with Reinforcement Learning. *arXiv preprint arXiv:2409.13035* (2024).

[324] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems* 37 (2024), 8612–8642.

[325] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36 (2023), 38154–38180.

[326] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15638–15650.

[327] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8317–8326.

[328] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal Input Multimodal Output Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 5317–5332. doi:10.18653/v1/2021.naacl-main.418

[329] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 5317–5332.

[330] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics* 11 (2023), 1–17.

[331] Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context. *arXiv preprint arXiv:2209.15189* (2022).

[332] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18221–18232.

[333] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1979–1988.

[334] Yiping Song, Rui Yan, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, and Dongyan Zhao. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. (2018).

[335] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355* (2023).

[336] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491* (2018).

[337] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14398–14409.

[338] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222* (2023).

[339] Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Instruction distillation makes large language models efficient zero-shot rankers. *arXiv preprint arXiv:2311.01555* (2023).

[340] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).

[341] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542* (2023).

[342] Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2024. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation. arXiv:2412.10704 [cs.CL] https://arxiv.org/abs/2412.10704

[343] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11888–11898.

[344] Adiba Tabassum and Shweta A Dhondse. 2015. Text detection using MSER and stroke width transform. In *2015 Fifth International Conference on Communication Systems and Network Technologies*. IEEE, 568–571.

[345] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039* (2021).

[346] Sijun Tan, Xiuyu Li, Shishir Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E Gonzalez, and Raluca Ada Popa. 2024. Lloco: Learning long contexts offline. *arXiv preprint arXiv:2404.07979* (2024).

[347] 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. arXiv:2301.04883 [cs.CL] https://arxiv.org/abs/2301.04883

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

[349] Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *arXiv preprint arXiv:2310.07712* (2023).

[350] Xu Tang, Yijing Wang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. 2023. Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–15.

[351] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27425–27434.

[352] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.

[353] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[354] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663* (2021).

[355] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. 2024. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208* (2024).

[356] Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. 2024. Holistic Features are almost Sufficient for Text-to-Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17138–17147.

[357] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773* (2022).

[358] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition* 144 (2023), 109834.

[359] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860* (2024).

[360] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9568–9578.

[361] Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124* (2016).

[362] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
[363] Thorsten Wagner and Hans-Gerd Lipinski. 2013. IJBlob: an ImageJ library for connected component analysis and shape analysis. *Journal of Open Research Software* 1, 1 (2013).
[364] Ao Wang, Fengyuan Sun, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. [CLS] Token Tells Everything Needed for Training-free Efficient MLLMs. *arXiv preprint arXiv:2412.05819* (2024).
[365] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024. SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13384–13394.
[366] Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, Zhengxin Li, Lin Ma, Shenghua Gao, et al. 2024. Tool-lmm: A large multi-modal model for tool agent learning. *arXiv e-prints* (2024), arXiv–2401.
[367] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023. DocLLM: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908* (2023).
[368] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. *arXiv preprint arXiv:2406.09411* (2024).
[369] Jinyu Wang, Jingjing Fu, Rui Wang, Lei Song, and Jiang Bian. 2025. PIKE-RAG: sPecIalized KnowledgE and Rationale Augmented Generation. arXiv:2501.11551 [cs.CL] https://arxiv.org/abs/2501.11551
[370] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. 2015. Image-text cross-modal retrieval via modality-specific feature learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. 347–354.
[371] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024. Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16551–16560.
[372] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804* (2024).
[373] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Simlm: Pre-training with representation bottleneck for dense passage retrieval. *arXiv preprint arXiv:2207.02578* (2022).
[374] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2413–2427.
[375] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).
[376] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. 2024. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035* (2024).
[377] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems* 37 (2024), 121475–121499.
[378] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. 2023. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907* (2023).
[379] Xinfeng Wang, Jin Cui, Yoshimi Suzuki, and Fumiyo Fukumoto. 2024. Rdrec: Rationale distillation for llm-based recommendation. *arXiv preprint arXiv:2405.10587* (2024).
[380] Xiaodan Wang, Lei Li, Zhixu Li, Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao. 2023. Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 456–464.
[381] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529* (2024).
[382] Xinyu Wang, Bohan Zhuang, and Qi Wu. 2024. Modaverse: Efficiently transforming modalities with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26606–26616.
[383] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems* 35 (2022), 25600–25614.
[384] Yan Wang, Yuting Su, Wenhui Li, Jun Xiao, Xuanya Li, and An-An Liu. 2023. Dual-path rare content enhancement network for image and text matching. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 10 (2023), 6144–6158.

[385] Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377* (2023).
[386] Zheng Wang, Zhenwei Gao, Mengqun Han, Yang Yang, and Heng Tao Shen. 2024. Estimating the Semantics via Sector Embedding for Image-Text Retrieval. *IEEE Transactions on Multimedia* (2024).
[387] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521* (2024).
[388] Zheng Wang, Xing Xu, Jiwei Wei, Ning Xie, Yang Yang, and Heng Tao Shen. 2024. Semantics disentangling for cross-modal retrieval. *IEEE Transactions on Image Processing* 33 (2024), 2226–2237.
[389] Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023. NOVO: learnable and interpretable document identifiers for model-based IR. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2656–2665.
[390] Jônatas Wehrmann and Rodrigo C Barros. 2018. Bidirectional retrieval made simple. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7718–7726.
[391] 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 *European Conference on Computer Vision*. Springer, 387–404.
[392] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Small language model meets with reinforced vision vocabulary. *arXiv preprint arXiv:2401.12503* (2024).
[393] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. Promptmm: Multi-modal knowledge distillation for recommendation with prompt-tuning. In *Proceedings of the ACM Web Conference 2024*. 3217–3228.
[394] Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. 2024. Multimodal reranking for knowledge-intensive visual question answering. *arXiv preprint arXiv:2407.12277* (2024).
[395] Weixi Weng, Jieming Zhu, Xiaojun Meng, Hao Zhang, Rui Zhang, and Chun Yuan. 2024. Learning to Compress Contexts for Efficient Knowledge-based Visual Question Answering. *arXiv preprint arXiv:2409.07331* (2024).
[396] David Wingate, Mohammad Shoeybi, and Taylor Sorensen. 2022. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. *arXiv preprint arXiv:2210.03162* (2022).
[397] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).
[398] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181* (2023).
[399] Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13084–13094.
[400] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.
[401] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519 [cs.AI] https://arxiv.org/abs/2309.05519
[402] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10704–10713.
[403] Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. DocGenome: An Open Large-scale Scientific Document Benchmark for Training and Testing Multi-modal Large Language Models. *arXiv preprint arXiv:2406.11633* (2024).
[404] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185* (2024).
[405] Chen-Wei Xie, Jianmin Wu, Yun Zheng, Pan Pan, and Xian-Sheng Hua. 2022. Token embeddings alignment for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4555–4563.
[406] Yifei Xin, Dongchao Yang, and Yuexian Zou. 2023. Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[407] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247* (2024).
[408] Guoxin Xiong, Meng Meng, Tianzhu Zhang, Dongming Zhang, and Yongdong Zhang. 2024. Reference-Aware Adaptive Network for Image-Text Matching. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[409] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[410] Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408* (2023).

[411] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[412] 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*. 1192–1200.

[413] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740* (2020).

[414] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915* (2023).

[415] Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. 2021. Efficient passage retrieval with hashing for open-domain question answering. *arXiv preprint arXiv:2106.00882* (2021).

[416] Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, and Harrie Oosterhuis. 2024. Consolidating Ranking and Relevance Predictions of Large Language Models through Post-Processing. *arXiv preprint arXiv:2404.11791* (2024).

[417] Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*. Springer, 37–53.

[418] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*.

[419] Song Yang, Qiang Li, Wenhui Li, Xuanya Li, and An-An Liu. 2022. Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 8037–8050.

[420] Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Auto search indexer for end-to-end document retrieval. *arXiv preprint arXiv:2310.12455* (2023).

[421] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6540–6548.

[422] Yang Yang, Chubing Zhang, Yi-Chu Xu, Dianhai Yu, De-Chuan Zhan, and Jian Yang. 2021. Rethinking Label-Wise Cross-Modal Retrieval from A Semantic Sharing Perspective.. In *IJCAI*. 3300–3306.

[423] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[424] Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. 2023. Teal: Tokenize and embed all for multi-modal large language models. *arXiv preprint arXiv:2311.04589* (2023).

[425] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985* (2024).

[426] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499* (2023).

[427] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).

[428] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 13040–13051.

[429] Dongyi Yi, Guibo Zhu, Chenglin Ding, Zongshu Li, Dong Yi, and Jinqiao Wang. 2025. MME-Industry: A Cross-Industry Multimodal Evaluation Benchmark. *arXiv preprint arXiv:2501.16688* (2025).

[430] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems* 36 (2024).

[431] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006* (2024).

[432] Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. Compact: Compressing retrieved documents actively for question answering. *arXiv preprint arXiv:2407.09014* (2024).

[433] Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung-won Hwang. 2024. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. *arXiv preprint arXiv:2402.15838* (2024).

[434] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591* 2, 3 (2023), 3.

[435] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025. MRAMG-Bench: A BeyondText Benchmark for Multimodal Retrieval-Augmented Multimodal Generation. *arXiv preprint arXiv:2502.04176* (2025).

[436] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*. 829–838.

[437] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594* (2024).

[438] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13807–13816.

[439] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* (2023).

[440] Xiaohan Yu, Zhihan Yang, and Chong Chen. 2025. Unveiling the Potential of Multimodal Retrieval Augmented Generation with Planning. *arXiv preprint arXiv:2501.15470* (2025).

[441] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3165–3173.

[442] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9127–9134.

[443] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28202–28211.

[444] Zhengqing Yuan, Zhaoxu Li, Weiran Huang, Yanfang Ye, and Lichao Sun. 2023. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862* (2023).

[445] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9556–9567.

[446] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813* (2024).

[447] Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, et al. 2024. Web2Code: A Large-scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs. *arXiv preprint arXiv:2406.20098* (2024).

[448] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6720–6731.

[449] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM on Web Conference 2024*. 1441–1452.

[450] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. 2024. What matters in training a gpt4-style language model with multimodal inputs?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7930–7957.

[451] Zhixiong Zeng and Wenji Mao. 2022. A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval. *arXiv preprint arXiv:2201.02772* (2022).

[452] ChengXiang Zhai et al. 2008. Statistical language models for information retrieval a critical review. *Foundations and Trends® in Information Retrieval* 2, 3 (2008), 137–213.
[453] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Jointly optimizing query encoder and product quantization to improve retrieval performance. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2487–2496.
[454] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1503–1512.
[455] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2022. Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1328–1336.
[456] Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2687–2692.
[457] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944* (2024).
[458] Hanqi Zhang, Chong Chen, Lang Mei, Qi Liu, and Jiaxin Mao. 2024. Mamba Retriever: Utilizing Mamba for Effective and Efficient Dense Retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4268–4272.
[459] Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. *arXiv preprint arXiv:2110.03611* (2021).
[460] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023).
[461] Han Zhang, Hongwei Shen, Yiming Qiu, Yunjiang Jiang, Songlin Wang, Sulong Xu, Yun Xiao, Bo Long, and Wen-Yun Yang. 2021. Joint learning of deep retrieval model and product quantization based embedding index. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1718–1722.
[462] Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024. CREAM: coarse-to-fine retrieval and multi-modal efficient tuning for document VQA. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 925–934.
[463] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2024. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. arXiv:2412.02592 [cs.CV] https://arxiv.org/abs/2412.02592
[464] Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. A two-stage adaptation of large language models for text ranking. *arXiv preprint arXiv:2311.16720* (2023).
[465] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023. InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition. arXiv:2309.15112 [cs.CV] https://arxiv.org/abs/2309.15112
[466] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output. arXiv:2407.03320 [cs.CV] https://arxiv.org/abs/2407.03320
[467] Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024. Compressing lengthy context with ultragist. *arXiv preprint arXiv:2405.16635* (2024).
[468] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3536–3545.
[469] Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2024. AdaComp: Extractive Context Compression with Adaptive Predictor for Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2409.01579* (2024).
[470] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?. In *European Conference on Computer Vision*. Springer, 169–186.
[471] Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. *arXiv preprint arXiv:2203.08372* (2022).
[472] Shilong Zhang, Peize Sun, Shoufa Chen, Minn Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. [n. d.]. GPT4roi: Instruction tuning large language model on regionof-interest, 2024. In *URL https://openreview.*

net/forum.

[473] Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. 2024. VCR: Visual Caption Restoration. *arXiv preprint arXiv:2406.06462* (2024).

[474] Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, et al. 2024. mR^2AG: Multimodal Retrieval-Reflection-Augmented Generation for Knowledge-Based VQA. *arXiv preprint arXiv:2411.15041* (2024).

[475] Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2023. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. *arXiv preprint arXiv:2312.02969* (2023).

[476] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855* (2024).

[477] Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. 2024. USER: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing* (2024).

[478] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107* (2023).

[479] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, et al. 2024. Irgen: Generative modeling for image retrieval. In *European Conference on Computer Vision*. Springer, 21–41.

[480] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2024. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? *arXiv preprint arXiv:2408.13257* (2024).

[481] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai. 2016. Multi-Oriented Text Detection with Fully Convolutional Networks. arXiv:1604.04018 [cs.CV] https://arxiv.org/abs/1604.04018

[482] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking Multi-Image Understanding in Vision and Language Models: Perception, Knowledge, Reasoning, and Multi-Hop Reasoning. *arXiv preprint arXiv:2406.12742* (2024).

[483] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. 2023. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474* (2023).

[484] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. 2024. Accelerating Multimodel Large Language Models by Searching Optimal Vision Token Reduction. *arXiv preprint arXiv:2412.00556* (2024).

[485] Shengwei Zhao, Linhai Xu, Yuying Liu, and Shaoyi Du. 2023. Multi-grained representation learning for cross-modal retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2194–2198.

[486] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Wengang Zhou, et al. 2024. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *arXiv preprint arXiv:2406.01326* (2024).

[487] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581* (2023).

[488] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. 2024. Needle In A Video Haystack: A Scalable Synthetic Framework for Benchmarking Video MLLMs. *arXiv preprint arXiv:2406.09367* (2024).

[489] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10394–10403.

[490] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239* (2023).

[491] Liu Zheng and Shao Yingxia. 2022. RetroMAE: Pre-training retrieval-oriented transformers via masked auto-encoder. *arXiv: 2205.12035* (2022).

[492] Xiaoyang Zheng, Zilong Wang, Sen Li, Ke Xu, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. 2023. Make: Vision-language pre-training based product retrieval in taobao search. In *Companion Proceedings of the ACM Web Conference 2023*. 356–360.

[493] Chenyu Zhou, Mengdan Zhang, Peixian Chen, Chaoyou Fu, Yunhang Shen, Xiawu Zheng, Xing Sun, and Rongrong Ji. 2024. VEGA: Learning Interleaved Image-Text Comprehension in Vision-Language Large Models. *arXiv preprint arXiv:2406.10228* (2024).

[494] Dong Zhou, Fang Lei, Lin Li, Yongmei Zhou, and Aimin Yang. 2024. Cross-Modal Interaction via Reinforcement Feedback for Audio-Lyrics Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[495] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. VISTA: visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292* (2024).
[496] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264* (2024).
[497] Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, Nan Duan, et al. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. *arXiv preprint arXiv:2210.11773* (2022).
[498] Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2023. MARVEL: unlocking the multi-modal capability of dense retrieval via visual module plugin. *arXiv preprint arXiv:2310.14037* (2023).
[499] Wangchunshu Zhou, Yuchen Eleanor Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2023. Efficient prompting via dynamic in-context learning. *arXiv preprint arXiv:2305.11170* (2023).
[500] Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12481–12490.
[501] Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, Peitian Zhang, and Ji-Rong Wen. 2022. Ultron: An ultimate retriever on corpus with a model-based indexer. *arXiv preprint arXiv:2208.09257* (2022).
[502] Yu-Jia Zhou, Jing Yao, Zhi-Cheng Dou, Ledell Wu, and Ji-Rong Wen. 2023. DynamicRetriever: a pre-trained model-based IR system without an explicit index. *Machine Intelligence Research* 20, 2 (2023), 276–288.
[503] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852* (2023).
[504] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).
[505] Dongsheng Zhu, Xunzhu Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. 2024. Vislinginstruct: Elevating zero-shot learning in multi-modal language models with autonomous instruction optimization. *arXiv preprint arXiv:2402.07398* (2024).
[506] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. 2023. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251* (2023).
[507] Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. 2024. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*. 18–22.
[508] Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. 2024. MuRAR: A Simple and Effective Multimodal Retrieval and Answer Refinement Framework for Multimodal Question Answering. arXiv:2408.08521 [cs.IR] https://arxiv.org/abs/2408.08521
[509] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122* (2023).
[510] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2308–2313.
[511] Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243* (2023).
[512] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128* (2022).
[513] Justin Zobel and Alistair Moffat. 2006. Inverted files for text search engines. *ACM computing surveys (CSUR)* 38, 2 (2006), 6–es.
[514] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207* (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009