Ask in Any Modality:
A Comprehensive Survey on Multimodal Retrieval-Augmented Generation

Mohammad Mahdi Abootorabi[†], **Amirhosein Zobeiri**[◇], **Mahdi Dehghani**[¶], **Mohammadali Mohammadkhani**[§], **Bardia Mohammadi**[§], **Omid Ghahroodi**[†], **Mahdieh Soleymani Baghshah**[§, *], **Ehsaneddin Asgari**[†, *]

[§]Computer Engineering Department, Sharif University of Technology, Tehran, Iran,
[◇]College of Interdisciplinary Science and Technology, University of Tehran, Tehran, Iran,
[¶]Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran,
[†]Qatar Computing Research Institute, Doha, Qatar

Correspondence: mahdi.abootorabi2@gmail.com, soleymani@sharif.edu, easgari@hbku.edu.qa

Abstract

Large Language Models (LLMs) struggle with hallucinations and outdated knowledge due to their reliance on static training data. Retrieval-Augmented Generation (RAG) mitigates these issues by integrating external dynamic information enhancing factual and updated grounding. Recent advances in multimodal learning have led to the development of Multimodal RAG, incorporating multiple modalities such as text, images, audio, and video to enhance the generated outputs. However, cross-modal alignment and reasoning introduce unique challenges to Multimodal RAG, distinguishing it from traditional unimodal RAG. This survey offers a structured and comprehensive analysis of Multimodal RAG systems, covering datasets, metrics, benchmarks, evaluation, methodologies, and innovations in retrieval, fusion, augmentation, and generation. We precisely review training strategies, robustness enhancements, and loss functions, while also exploring the diverse Multimodal RAG scenarios. Furthermore, we discuss open challenges and future research directions to support advancements in this evolving field. This survey lays the foundation for developing more capable and reliable AI systems that effectively leverage multimodal dynamic external knowledge bases. Resources are available at `https://github.com/llm-lab-org/Multimodal-RAG-Survey`.

1 Introduction & Background

In recent years, significant breakthroughs have been achieved in language models, driven primarily by the advent of transformers (Vaswani et al., 2017), enhanced computational capabilities, and the availability of large-scale training data (Naveed et al., 2024). The emergence of foundational Large Language Models (LLMs) (Ouyang et al., 2022; Grattafiori et al., 2024; Touvron et al., 2023; Qwen et al., 2025; Anil et al., 2023) has revolutionized natural language processing (NLP), demonstrating unprecedented capabilities in a wide range of tasks including instruction following (Qin et al., 2024), sophisticated reasoning (Wei et al., 2024c), In-context Learning (Brown et al., 2020), and multilingual machine translation (Zhu et al., 2024a). These advancements have elevated the performance of various NLP tasks, opening new avenues for research and application. Despite their remarkable achievements, LLMs face significant challenges, including hallucination, outdated internal knowledge, and a lack of verifiable reasoning (Huang et al., 2024; Xu et al., 2024b). Their reliance on parametric memory restricts their ability to access up-to-date knowledge, making them less effective for knowledge-intensive tasks compared to task-specific architectures. Moreover, providing provenance for their decisions and updating their world knowledge remain critical open problems (Lewis et al., 2020).

Retrieval-Augmented Generation (RAG) RAG (Lewis et al., 2020) has emerged as a promising solution to these limitations by enabling LLMs to retrieve and incorporate external knowledge, improving factual accuracy and reducing hallucinations (Shuster et al., 2021; Ding et al., 2024a). By dynamically accessing vast external knowledge repositories, RAG systems enhance knowledge-intensive tasks while ensuring responses remain grounded in verifiable sources (Gao et al., 2023).

In practice, RAG systems operate through a retriever-generator pipeline. The retriever leverages embedding models (Chen et al., 2024a; Rau et al., 2024) to identify relevant passages from external knowledge bases and optionally applies re-ranking techniques to improve retrieval precision (Dong et al., 2024a). These retrieved passages are then passed to the generator, incorporating this ex-

[*]These authors contributed equally.

ternal context to produce informed responses. Recent advancements in RAG frameworks, such as planning-guided retrieval (Lee et al., 2024), agentic RAG (An et al., 2024), and feedback-driven iterative refinement (Liu et al., 2024c; Asai et al., 2023), further enhance both retrieval and generation stages. However, traditional RAG architectures are primarily designed for textual information, limiting their ability to address multimodal challenges that require integrating diverse data formats.

Multimodal Learning Parallel to these developments, significant advances in multimodal learning have reshaped artificial intelligence by enabling systems to integrate and analyze heterogeneous data sources for a holistic representation of information. The introduction of CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) marked a pivotal moment in connecting visual and textual information through contrastive learning, inspiring numerous subsequent models and applications (Alayrac et al., 2024; Wang et al., 2023; Pramanick et al., 2023).

These breakthroughs have driven progress in various domains, including sentiment analysis (Das and Singh, 2023) and cutting-edge biomedical research (Hemker et al., 2024), demonstrating the value of multimodal approaches. By enabling systems to process and understand diverse data types such as text, images, audio, and video, multimodal learning plays a key role in advancing artificial general intelligence (AGI) (Song et al., 2025).

Multimodal RAG The extension of LLMs to multimodal LLMs (MLLMs) has further expanded their capabilities, allowing them to process, reason, and generate outputs across diverse modalities (Liu et al., 2023a; Team et al., 2024; Li et al., 2023b). For example, GPT-4 (OpenAI et al., 2024) demonstrates human-level performance in various benchmarks by processing both text and images, marking a significant milestone in multimodal perception and interaction. Building on this foundation, multimodal RAG systems extend traditional RAG frameworks by incorporating multimodal knowledge sources, such as images and audio, to provide enriched context for generation (Hu et al., 2023; Chen et al., 2022a). This integration enhances the precision of generated outputs while leveraging multimodal cues to improve the reasoning capabilities of MLLMs. However, these multimodal systems also present unique challenges, including determining which modalities to retrieve, effectively fusing diverse data types, and addressing the complexities of cross-modal relevance (Zhao et al., 2023a). Figure 1 illustrates the general pipeline of these systems.

Task Formulation A mathematical formulation of the general task for multimodal RAG is presented in this section. These systems generate a multimodal response, denoted as r, in response to a multimodal query q.

Let $D = \{d_1, d_2, ..., d_n\}$ be a multimodal corpus. Each document $d_i \in D$ is associated with a modality M_{d_i} and processed by a modality-specific encoder, yielding:

$$z_i = Enc_{M_{d_i}}(d_i) \tag{1}$$

The set of all encoded representations is denoted by $Z = \{z_1, z_2, ..., z_n\}$. Modality-specific encoders map different modalities into a shared semantic space for cross-modal alignment. A retrieval model R assesses the relevance of each encoded document representation z with respect to the query q, represented as $R(q, z)$. To construct the retrieval-augmented multimodal context, the retrieval model selects the most relevant documents based on a modality-specific threshold:

$$X = \{d_i \mid s(e_q, z_i) \geq \tau_{M_{d_i}}\} \tag{2}$$

where $\tau_{M_{d_i}}$ is a relevancy threshold for the modality of M_{d_i}, e_q is the encoded representation of q in the shared semantic space, and s is a scoring function that measures the relevance between the encoded query and document representations. The generative model G produces the final multimodal response, given the user query q and the retrieved documents X as context, denoted as $r = G(q, X)$.

Related Works As the field of multimodal RAGs is newly introduced and evolving rapidly, especially in recent years, there is a pressing need for a comprehensive survey that explores the current innovations and frontiers of these systems. While more than ten surveys have been published on RAG-related topics such as Agentic RAG (Singh et al., 2025), none provide a detailed and comprehensive overview of advancements in multimodal RAGs. The only related survey to date (Zhao et al., 2023a) categorizes multimodal RAGs by grouping relevant papers based on their applications and modalities. However, our survey provides a more detailed and innovation-driven perspective, offering a detailed taxonomy and exploring emerging

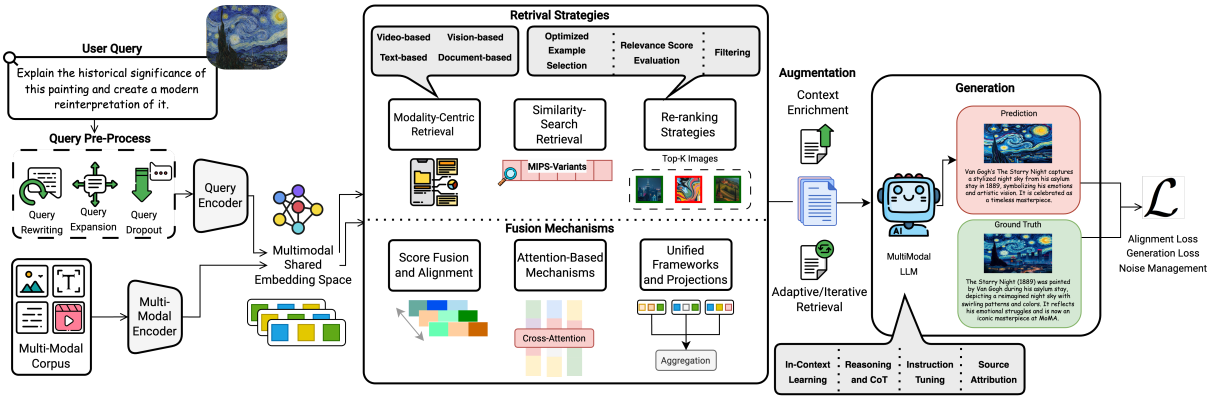

Figure 1: Overview of the multimodal retrieval-augmented generation (RAG) pipeline, highlighting the advancements and techniques employed at each stage. The flow begins with query preprocessing, where user queries are refined and then encoded into a shared embedding space alongside a multimodal database. Retrieval strategies, such as modality-centric retrieval, similarity search, and re-ranking, enhance document selection, while fusion mechanisms align and integrate data from multiple modalities using score fusion or attention-based methods. Augmentation techniques such as iterative retrieval with feedback mechanisms, further refine the retrieved documents for the multimodal LLM. The generation stage incorporates innovations like Chain-of-Thought reasoning and source attribution for better outputs, with loss functions combining alignment loss and generation loss to optimize both retrieval and generation components. Noise management techniques are also applied to improve training stability and robustness.

trends and challenges in depth. Furthermore, significant advancements have been made in the field since its publication, and interest in this topic has grown substantially within the research community. In this survey, we review over 100 papers on multimodal RAGs published in recent years, primarily from the ACL Anthology and other repositories such as the ACM Digital Library.

Contributions In this work, **(i)** we provide a comprehensive review of multimodal RAG, covering task formulation, datasets, benchmarks, tasks and domain-specific applications, evaluation, and key innovations in retrieval, fusion, augmentation, generation, training strategies, and loss functions. **(ii)** We introduce a precise structured taxonomy (Figure 2) that categorizes state-of-the-art models based on their primary contributions, highlighting methodological advancements and emerging frontiers. **(iii)** To support further research, we make resources, including datasets, benchmarks, and key innovations, publicly available. **(iv)** We identify current research trends and knowledge gaps, providing insights and recommendations to guide future advancements in this evolving field.

2 Datasets and Benchmarks

Multimodal RAG research employs diverse datasets and benchmarks to evaluate retrieval, integration, and generation across heterogeneous sources. Image–text tasks, including captioning and retrieval, commonly use MS-COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), and LAION-400M (Schuhmann et al., 2021), while visual question answering with external knowledge is supported by OK-VQA (Marino et al., 2019) and WebQA (Chang et al., 2022). For complex multimodal reasoning, MultimodalQA (Talmor et al., 2021) integrates text, images, and tables, whereas video-text tasks leverage ActivityNet (Caba Heilbron et al., 2015) and YouCook2 (Zhou et al., 2018).

In the medical domain, MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019) facilitate tasks such as medical report generation. It is noteworthy that a number of these datasets are unimodal (e.g., solely text-based or image-based). Unimodal datasets are frequently employed to represent a specific modality and are subsequently integrated with complementary datasets from other modalities. This modular approach allows each dataset to contribute its domain-specific strengths, thereby enhancing the overall performance of the multimodal retrieval and generation processes.

Benchmarks assess multimodal RAG systems on visual reasoning, external knowledge integration, and dynamic retrieval. The M^2RAG (Ma et al., 2024c) benchmark provides a unified evaluation framework that combines fine-grained text-modal and multimodal metrics to jointly assess both the quality of generated language and the effective integration of visual elements Vision-focused evaluations, including MRAG-Bench (Hu et al., 2024c), VQAv2 (Goyal et al., 2017a) and VisDoMBench (Suri et al., 2024), test models on complex visual tasks. Dyn-VQA (Li et al., 2024b), MMBench (Liu et al., 2025), and ScienceQA (Saikh et al., 2022) evaluate dynamic retrieval and multi-hop reasoning across textual, visual, and diagrammatic inputs.

Knowledge-intensive benchmarks, such as TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019), together with document-oriented evaluations such as OmniDocBench (Ouyang et al., 2024), measure integration of unstructured and structured data. Advanced retrieval benchmarks such as RAG-Check (Mortaheb et al., 2025a) evaluate retrieval relevance and system reliability, while specialized assessments such as Counterfactual VQA (Niu et al., 2021) test robustness against adversarial inputs. Additionally, OCR impact studies such as OHRBench (Zhang et al., 2024d) examine the cascading effects of errors on RAG systems. Additional details about datasets, benchmarks, and their categorization are presented in Table 1 and Table 2 in Appendix (§B).

3 Evaluation

Evaluating multimodal RAG models is complex due to their varied input types and complex structure. The evaluation combines metrics from VLMs, generative AI, and retrieval systems to assess capabilities like text/image generation and information retrieval. Our review found about 60 different metrics used in the field. More details, including the formulas for the RAG evaluation metrics, can be found in Appendix (§C). In the following paragraphs, we will examine the most important and widely used metrics for evaluating multimodal RAG.

Retrieval Evaluation Retrieval performance is measured through accuracy, recall, and precision metrics, with an F1 score combining recall and precision. Recall@K, which examines relevant items in top K results, is preferred over standard recall. Mean Reciprocal Rank (MRR) serves as another key metric for evaluation, which is utilized by (Adjali et al., 2024; Nguyen et al., 2024).

Modality Evaluation Modality-based evaluations primarily focus on text and image, assessing their alignment, text fluency, and image caption quality. For text evaluation, metrics include Exact Match (EM), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). MultiRAGen (Shohan et al., 2024) uses Multilingual ROUGE for multilingual settings.
For image captioning, CIDEr (Consensus-Based Image Description Evaluation) (Vedantam et al., 2015) measures caption quality using TF-IDF and cosine similarity (Yasunaga et al., 2023; Zhao et al., 2024; Luo et al., 2024a; Yuan et al., 2024; Sharifymoghaddam et al., 2024; Hu et al., 2023; Rao et al., 2024; Xu et al., 2024a; Kim et al., 2024; Zhang et al., 2024c), while SPICE (Semantic Propositional Image Caption Evaluation) (Anderson et al., 2016) focuses on semantics. SPIDEr (Liu et al., 2017), used in (Zhang et al., 2024c), combines both metrics. For semantic alignment, BERTScore (Zhang et al., 2020) compares BERT embeddings (Sun et al., 2024b; Shohan et al., 2024), and evaluates fluency (Chen et al., 2022a; Zhi Lim et al., 2024; Ma et al., 2024c). CLIP Score (Hessel et al., 2021), used in (Sharifymoghaddam et al., 2024; Zhang et al., 2024c), measures image-text similarity using CLIP (Radford et al., 2021). For image quality, FID (Fréchet Inception Distance) (Heusel et al., 2017) compares feature distributions (Yasunaga et al., 2023; Zhao et al., 2024; Sharifymoghaddam et al., 2024; Zhang et al., 2024c), while KID (Kernel Inception Distance) (Bińkowski et al., 2018) provides an unbiased alternative. Inception Score (IS) evaluates image diversity and quality through classification probabilities (Zhi Lim et al., 2024). For audio evaluation, (Zhang et al., 2024c) uses human annotators to assess sound quality (OVL) and text relevance (REL), while also employing Fréchet Audio Distance (FAD) (Kilgour et al., 2019), an audio-specific variant of FID.

System efficiency is measured through FLOPs, execution time, response time, and retrieval time per query (Nguyen et al., 2024; Strand et al., 2024; Dang, 2024; Zhou, 2024). Domain-specific metrics include geodesic distance for geographical accuracy (Zhou et al., 2024e), and Clinical Relevance for medical applications (Lahiri and Hu, 2024).

4 Key Innovations and Methodologies

4.1 Retrieval Strategy

4.1.1 Efficient Search and Similarity Retrieval

Modern multimodal RAG systems encode diverse input modalities into a unified embedding space

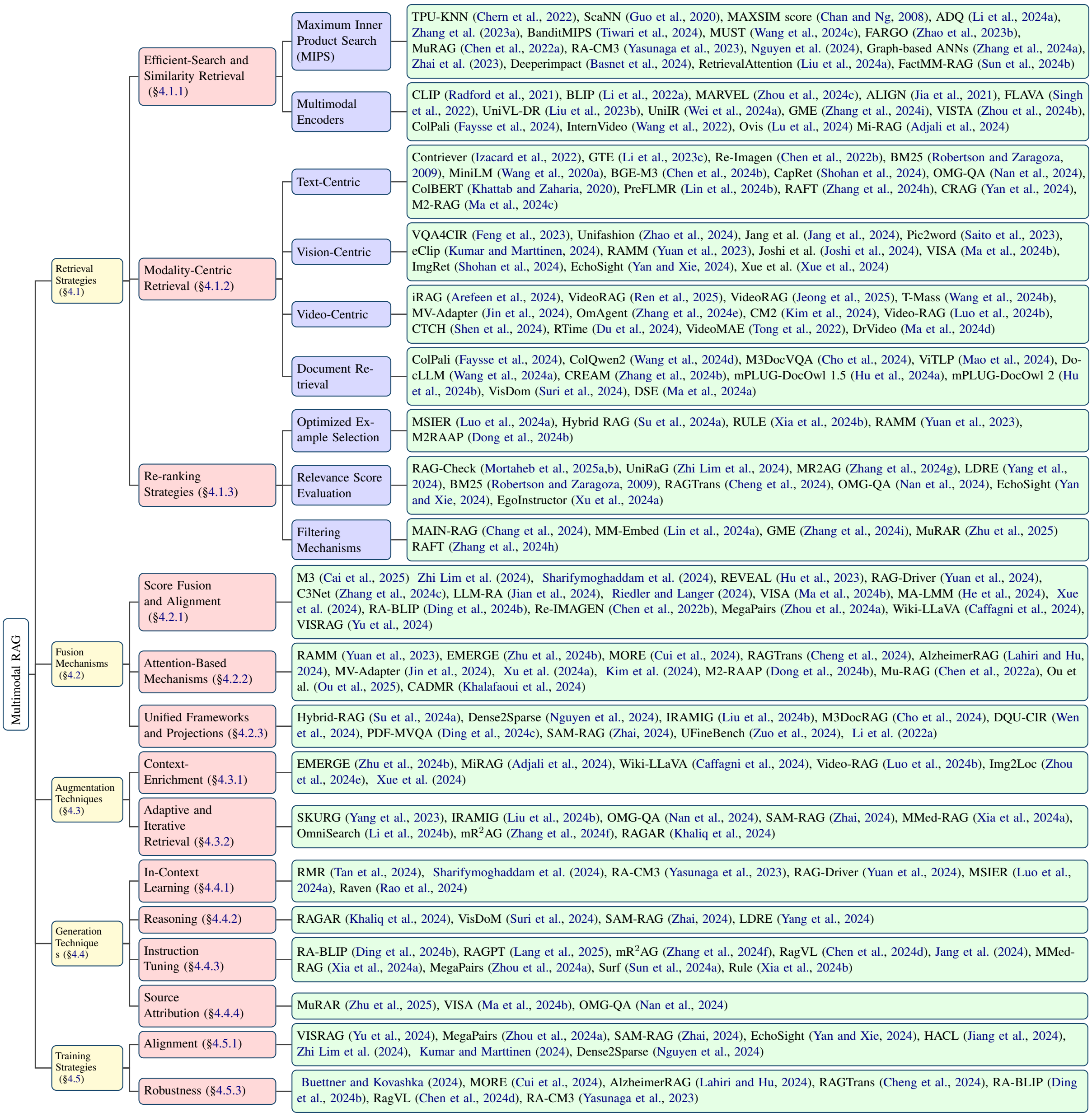

Figure 2: Taxonomy of recent advances in Multimodal RAG. Refer to Appendix (§A) for further details.

to enable direct cross-modal retrieval. Recent advancements in CLIP-based (Radford et al., 2021) and BLIP-inspired (Li et al., 2022a) approaches have driven the evolution of contrastive learning strategies through novel multimodal retrieval architectures and training methodologies (Zhou et al., 2024c; Wei et al., 2024b; Zhang et al., 2024i). As these multi-encoder models project different modalities into a shared latent space, multimodal RAGs rely on efficient search strategies to retrieve relevant external knowledge.

Maximum inner product search (MIPS) variants are widely used for fast and direct similarity comparisons (Tiwari et al., 2024; Wang et al., 2024c; Zhao et al., 2023b). Systems such as MuRAG (Chen et al., 2022a) and RA-CM3 (Yasunaga et al., 2023) employ approximate MIPS to efficiently retrieve the top-k candidates by maximizing the inner product between the query vector and a large collection of image–text embeddings. Large-scale implementations leverage distributed MIPS techniques, such as TPU-KNN (Chern et al., 2022), for high-speed retrieval. Additionally, ScaNN (Scalable Nearest Neighbors) (Guo et al., 2020), MAXSIM score (Chan and Ng, 2008; Cho et al., 2024), and approximate KNN methods (Caffagni et al., 2024) have been adopted for efficient similarity computation.

Recent advancements in MIPS optimization focus on reducing retrieval latency and improving accuracy. These include adaptive quantization (Zhang et al., 2023a; Li et al., 2024a), hybrid sparse-dense representations (Nguyen et al., 2024; Zhang et al., 2024a), and learned index structures (Zhai et al., 2023; Basnet et al., 2024). These techniques help optimize retrieval performance by balancing efficiency and precision in multimodal RAG systems.

4.1.2 Modality-Based Retrieval

Modality-aware retrieval techniques optimize retrieval efficiency by leveraging the unique characteristics of each modality.

Text-Centric Retrieval This remains foundational in multimodal RAG systems, with both traditional methods like BM25 (Robertson and Zaragoza, 2009), and dense retrievers such as MiniLM (Wang et al., 2020a) and BGE-M3 (Chen et al., 2024b) dominate text-based evidence retrieval (Chen et al., 2022b; Suri et al., 2024; Nan et al., 2024). Novel approaches also address the need for fine-grained semantic matching and domain specificity: For instance, ColBERT (Khattab and Zaharia, 2020) and PreFLMR (Lin et al., 2024b) employ token-level interaction mechanisms that preserve nuanced textual details to improve precision for multimodal queries, while RAFT (Zhang et al., 2024h) and CRAG (Yan et al., 2024) enhance retrieval performance by ensuring precise citation of text spans.

Vision-Centric Retrieval This focuses on directly leveraging image representations for knowledge extraction (Kumar and Marttinen, 2024; Yuan et al., 2023). Systems such as EchoSight (Yan and Xie, 2024) and ImgRet (Shohan et al., 2024) retrieve visually similar content by using reference images as queries. In addition, composed Image Retrieval (CMI) models (Feng et al., 2023; Zhao et al., 2024; Jang et al., 2024) enhance retrieval by integrating multiple image features into a unified query representation. Similarly, Pic2word (Saito et al., 2023) maps visual content to textual descriptions, enabling zero-shot image retrieval.

Video-Centric Retrieval These methods extend vision-based techniques by incorporating temporal dynamics and large video-language models (LVLMs), driven by novel frameworks like iRAG (Arefeen et al., 2024), which introduces incremental retrieval for sequential video understanding, and MV-Adapter (Jin et al., 2024), optimizing multimodal transfer learning for video-text retrieval. Recent breakthroughs focus on long-context processing: Video-RAG (Luo et al., 2024b) leverages visually aligned auxiliary texts (OCR/ASR) to enhance retrieval without proprietary models, T-Mass (Wang et al., 2024b) models text as a stochastic embedding to enhance text-video retrieval, and VideoRAG (Ren et al., 2025) -employs dual-channel architectures with graph-based knowledge grounding for extreme-length videos. For temporal reasoning, CTCH (Shen et al., 2024) uses contrastive transformer hashing to model long-term dependencies, while RTime (Du et al., 2024) introduces reversed-video hard negatives to rigorously benchmark temporal causality. Meanwhile, for complex video understanding, OmAgent (Zhang et al., 2024e) adopts a divide-and-conquer framework, and DRVideo (Ma et al., 2024d) addresses long-video understanding with a document-based retrieval approach.

Document Retrieval and Layout Understanding Recent research has advanced beyond traditional uni-modal retrieval methods toward models that directly process entire document pages. These approaches integrate textual, visual, and structural elements like tables, font styles, and page layouts to enhance retrieval performance in complex documents. ColPali (Faysse et al., 2024) pioneers end-to-end document image retrieval by embedding page patches with a vision-language backbone, bypassing OCR entirely. Models like ColQwen2 (Wang et al., 2024d; Khattab and Zaharia, 2020) and M3DocVQA (Cho et al., 2024) extend this paradigm with dynamic resolution handling and holistic multi-page reasoning.

Newer frameworks refine efficiency and layout understanding: ViTLP (Mao et al., 2024) and DocLLM (Wang et al., 2024a) pre-train generative models to align spatial layouts with text, while CREAM (Zhang et al., 2024b) employs coarse-to-fine retrieval with multimodal efficient tuning to balance accuracy and computation costs. Finally, mPLUG-DocOwl 1.5 (Hu et al., 2024a) and 2 (Hu et al., 2024b) unify structure learning across formats (e.g., invoices, forms) without OCR dependencies.

4.1.3 Re-ranking and Selection Strategies

Effective retrieval in multimodal RAG systems requires not only identifying relevant information but also prioritizing retrieved candidates. Re-ranking and selection strategies enhance retrieval quality by optimizing example selection, refining relevance scoring, and applying filtering mechanisms.

Optimized Example Selection These techniques often employ multi-step retrieval, integrating both supervised and unsupervised selection strategies (Luo et al., 2024a; Yuan et al., 2023). For instance, (Su et al., 2024a) enhance multimodal inputs using probabilistic control keywords to improve credibility, RULE (Xia et al., 2024b) calibrates retrieved context selection via statistical methods like the Bonferroni correction to mitigate factuality risks, and clustering-based key-frame selection ensures diversity in video-based retrieval (Dong et al., 2024b).

Relevance Score Evaluation Several methods employ advanced scoring mechanisms to improve retrieval relevance (Mortaheb et al., 2025b,a; Zhi Lim et al., 2024). Multimodal similarity measures, including structural similarity index measure (SSIM) (Wang et al., 2020b), normalized cross-correlation (NCC), and BERTScore (Zhang et al., 2020), aid in re-ranking retrieved documents. Cross-encoders trained on sequence classification tasks refine relevance scoring within retrieval pipelines (Zhi Lim et al., 2024), while hierarchical post-processing techniques integrate retrieval, passage-level, and answer confidence scores for improved ranking (Zhang et al., 2024g; Yan and Xie, 2024; Xu et al., 2024a). LDRE (Yang et al., 2024) employs semantic ensemble methods to adaptively weigh multiple caption features, whereas RAGTrans (Cheng et al., 2024) and OMG-QA (Nan et al., 2024) incorporate traditional ranking functions like BM25 (Robertson and Zaragoza, 2009).

Filtering Mechasnism This ensures high-quality retrieval by eliminating irrelevant data. Hard negative mining, as used in GME (Zhang et al., 2024i) and MM-Embed (Lin et al., 2024a), mitigates modality bias through modality-aware sampling and synthesized negatives. Similarly, consensus-based filtering approaches, seen in MuRAR (Zhu et al., 2025) and ColPali (Faysse et al., 2024), employs source attribution and multi-vector mapping to filter out low-similarity candidates. Moreover, dynamic modality filtering methods, such as RAFT (Zhang et al., 2024h), Img2Loc (Zhou et al., 2024e), and MAIN-RAG (Chang et al., 2024), train retrievers to disregard confusing data, thereby enhancing the discriminative capacity and overall robustness of multi-modal retrieval systems.

4.2 Fusion Mechanisms

4.2.1 Score Fusion and Alignment

Models in this category utilize distinct strategies to align multimodal representations. Zhi Lim et al. (2024) convert text, tables, and images into a single textual format using a cross-encoder trained for relevance scoring. Sharifymoghaddam et al. (2024) introduce interleaved image–text pairs that vertically merge multiple few-shot images (as in LLaVA (Liu et al., 2023a)), while aligning modalities via CLIP score fusion (Hessel et al., 2021) and BLIP feature fusion (Li et al., 2022a). Riedler and Langer (2024), Wiki-LLaVA (Caffagni et al., 2024), C3Net (Zhang et al., 2024c), and MegaPairs (Zhou et al., 2024a), embed images and queries into a shared CLIP space.

VISA (Ma et al., 2024b) employs the Document Screenshot Embedding (DSE) model to align textual queries with visual document representations by encoding both into a shared embedding space. REVEAL (Hu et al., 2023) injects retrieval scores into attention layers to minimize L2-norm differences between query and knowledge embeddings, and MA-LMM (He et al., 2024) aligns video-text embeddings via a BLIP-inspired Query Transformer (Li et al., 2022a). LLM-RA (Jian et al., 2024) concatenates text and visual embeddings into joint queries to reduce retrieval noise, while RA-BLIP (Ding et al., 2024b) employs a 3-layer BERT-based adaptive fusion module to unify visual–textual semantics. Xue et al. (2024) use a prototype-based embedding network (Zheng et al., 2023) to map object-predicate pairs into a shared semantic space, aligning visual features with textual prototypes. Re-IMAGEN (Chen et al., 2022b) balances creativity and entity fidelity in text-to-image synthesis via interleaved classifier-free guidance during diffusion sampling. To improve multimodal alignment, VISRAG (Yu et al., 2024) enhances alignment with position-weighted mean pooling on VLM hidden states, prioritizing later tokens for relevance, and RAG-Driver (Yuan et al., 2024) aligns visual–language embeddings using visual instruction tuning and an MLP projector.

4.2.2 Attention-Based Mechanisms

Attention-based methods dynamically weight cross-modal interactions to support task-specific reasoning. EMERGE (Zhu et al., 2024b), MORE (Cui et al., 2024), and AlzheimerRAG (Lahiri and Hu, 2024) integrate heterogeneous data via cross-attention. RAMM (Yuan et al., 2023) employs a dual-stream co-attention transformer, combining self-attention and cross-attention to fuse retrieved biomedical images/texts with input data. RAG-Trans (Cheng et al., 2024) applies user-aware attention to social media features.

For video-text alignment, MV-Adapter (Jin et al., 2024) leverages Cross Modality Tying to align embeddings, and M2-RAAP (Dong et al., 2024b) enhances fusion through an auxiliary caption-guided strategy that re-weights frames and text captions based on intra-modal similarity. A mutual-guided alignment head then filters misaligned features using dot-product similarity and frame-to-token attention, generating refined frame-specific text representations. Xu et al. (2024a) condition text generation on visual features using gated cross-attention, and Mu-RAG (Chen et al., 2022a) employs intermediate cross-attention for open-domain QA. Kim et al. (2024) leverage cross-modal memory retrieval with pre-trained CLIP ViT-L/14 to map video-text pairs into a shared space, enabling dense captioning through the attention-based fusion of retrieved memories.

4.2.3 Unified Frameworks and Projections

Unified frameworks and projection methods consolidate multimodal inputs into coherent representations. Su et al. (2024a) employ hierarchical cross-chains and late fusion for healthcare data, while IRAMIG (Liu et al., 2024b) iteratively integrates multimodal results into unified knowledge representations. M3DocRAG (Cho et al., 2024) flattens multi-page documents into a single embedding tensor, and PDF-MVQA (Ding et al., 2024c) fuses Region-of-Interest (RoI)-based and patch-based (CLIP) vision-language models (Long et al., 2022).

DQU-CIR (Wen et al., 2024) unifies raw data by converting images into text captions for complex queries and overlaying text onto images for simple ones, then fusing embeddings via MLP-learned weights. SAM-RAG (Zhai, 2024) aligns image-text modalities by generating captions for images, converting the multimodal input into unimodal text for subsequent processing. UFineBench (Zuo et al., 2024) utilizes a shared granularity decoder for ultra-fine text–person retrieval. Nguyen et al. (2024) introduce Dense2Sparse projection, converting dense embeddings from models like BLIP/ALBEF (Li et al., 2022a) into sparse lexical vectors using layer normalization and probabilistic expansion control to optimize storage and interpretability.

4.3 Augmentation Techniques

Basic RAG systems typically follow a single retrieval step and pass retrieved content directly to the generation phase without further refinement, which can lead to inefficiencies and suboptimal outputs. To address this, augmentation techniques have been introduced to refine retrieved data beforehand, improving multimodal interpretation, structuring, and integration (Gao et al., 2023).

4.3.1 Context Enrichment

This focuses on enhancing the relevance of retrieved knowledge by refining or expanding retrieved data. General approaches incorporate additional contextual elements (e.g., text chunks, image tokens, structured data) to provide a richer grounding for generation (Caffagni et al., 2024; Xue et al., 2024). EMERGE (Zhu et al., 2024b) enriches context by integrating entity relationships and semantic descriptions. MiRAG (Adjali et al., 2024) expands initial queries through entity retrieval and reformulation, enhancing subsequent stages for the visual question-answering. Video-RAG (Luo et al., 2024b) enhances long-video understanding through Query Decoupling, which reformulates user queries into structured retrieval requests to extract auxiliary multimodal context. Img2Loc (Zhou et al., 2024e) enhances accuracy by including both the most similar and the most dissimilar points from the database in the prompt, allowing the model to rule out implausible locations for its predictions.

4.3.2 Adaptive and Iterative Retrieval

For more complex queries, dynamic retrieval mechanisms have proven effective. Adaptive retrieval approaches optimize relevance by adjusting retrieval dynamically. SKURG (Yang et al., 2023) determines the number of retrieval hops based on query complexity. SAM-RAG (Zhai, 2024) and mR^{2}AG (Zhang et al., 2024f) dynamically assess the need for external knowledge and filter irrelevant content using MLLMs to retain only task-critical information. MMed-RAG (Xia et al., 2024a) further improves retrieval precision by discarding low-relevance results, while OmniSearch (Li et al., 2024b) introduces a self-adaptive retrieval agent that dynamically decomposes complex multimodal questions into sub-question chains and plans retrieval actions in real-time based on retrieved content.

Iterative approaches refine results over multiple steps by incorporating feedback from prior iterations. IRAMIG (Liu et al., 2024b) improves multimodal retrieval by dynamically updating queries based on retrieved content. Similarly, OMG-QA

(Nan et al., 2024) introduces a multi-round retrieval strategy, where each retrieval step incorporates episodic memory to refine subsequent queries. An evaluation module assesses retrieval effectiveness at each step, guiding the refinement of subsequent retrieval efforts through feedback. RAGAR (Khaliq et al., 2024) further enhances contextual consistency by iteratively adjusting retrieval based on prior responses and multimodal analysis.

4.4 Generation Techniques

Recent advancements in multimodal RAG generation focus on robustness, cross-modal coherence, and task-specific adaptability. These innovations can be broadly categorized into these sections:

4.4.1 In-Context Learning

In-context learning (ICL) with retrieval augmentation enhances reasoning in multimodal RAGs by leveraging retrieved content as few-shot examples without requiring retraining. Models such as RMR (Tan et al., 2024), Sharifymoghaddam et al. (2024), and RA-CM3 (Yasunaga et al., 2023), extend this paradigm to multimodal RAG settings.

RAG-Driver (Yuan et al., 2024) refines ICL by retrieving relevant driving experiences from a memory database, ensuring scenario-specific contextual alignment. MSIER (Luo et al., 2024a) further enhances example selection through a Multimodal Supervised In-Context Examples Retrieval framework, leveraging a foundation MLLM scorer to evaluate both textual and visual relevance. Meanwhile, Raven (Rao et al., 2024) introduces Fusion-in-Context Learning, a novel approach that enriches ICL by incorporating a more diverse set of in-context examples, leading to better performance than standard ICL.

4.4.2 Reasoning

Structured reasoning techniques, such as chain-of-thought (CoT), decompose complex reasoning into smaller sequential steps, enhancing coherence and robustness in multimodal RAG systems. RAGAR (Khaliq et al., 2024) introduces Chain of RAG and Tree of RAG to iteratively refine fact-checking queries and explore branching reasoning paths for more robust evidence generation.

VisDoM (Suri et al., 2024) integrates CoT with evidence curation to ensure logical and contextual accuracy, while SAM-RAG (Zhai, 2024) employs reasoning chains alongside multi-stage answer verification to enhance the relevance, utility, and support of generated responses. Meanwhile, LDRE (Yang et al., 2024) leverages LLMs for divergent compositional reasoning, generating refined captions by incorporating dense captions and modification text.

4.4.3 Instruction Tuning

Several works have fine-tuned or instruct-tuned generation components for specific applications. RA-BLIP (Ding et al., 2024b) leverages the Q-Former architecture from InstructBLIP (Dai et al., 2023) to extract visual features based on question instructions, while RAGPT (Lang et al., 2025) employs a context-aware prompter to generate dynamic prompts from relevant instances. mR^{2}AG (Zhang et al., 2024f) uses instruction tuning with the mR2AG-IT dataset to train MLLMs to adaptively invoke retrieval, identify relevant evidence, and generate accurate answers for knowledge-based VQA tasks.

RagVL (Chen et al., 2024d) employs instruction tuning to enhance the ranking capability of MLLMs, serving them as a re-ranker for filtering the top-k retrieved images. Jang et al. (2024) focus on distinguishing image differences to generate descriptive textual responses. MMed-RAG (Xia et al., 2024a) applies preference fine-tuning to help models balance retrieved knowledge with internal reasoning.

To improve generation quality, MegaPairs (Zhou et al., 2024a) and Surf (Sun et al., 2024a) construct multimodal instruction-tuning datasets from prior LLM errors, while Rule (Xia et al., 2024b) refines Med-LVLM through direct preference optimization to mitigate overreliance on retrieved contexts.

4.4.4 Source Attribution and Evidence Transparency

Ensuring source attribution in multimodal RAG systems is a key focus of recent research. MuRAR (Zhu et al., 2025) integrates multimodal data, fetched by a source-based retriever, to refine LLM's initial response, ensuring informativeness. VISA (Ma et al., 2024b) uses large vision-language models to generate answers with visual source attribution by identifying and highlighting supporting evidence in retrieved document screenshots. Similarly, OMG-QA (Nan et al., 2024) ensures transparency by prompting the LLM to explicitly cite evidence in generated responses.

4.5 Training Strategies

Training multimodal RAG models involves a multi-stage process to effectively handle cross-modal interactions (Chen et al., 2022a). Pretraining establishes the foundation using large paired datasets to learn cross-modal relationships while fine-tuning adapts models to specific tasks by leveraging cross-modal attention (Ye et al., 2019). For instance, REVEAL (Hu et al., 2023) integrates multiple training objectives. Its pretraining phase optimizes Prefix Language Modeling Loss (L_{PrefixLM}), where the model predicts text continuations from a prefix and an image. Supporting losses include Contrastive Loss (L_{contra}) for aligning queries with pseudo-ground-truth knowledge, Disentangled Regularization Loss (L_{decor}) to improve embedding expressiveness, and Alignment Regularization Loss (L_{align}) to align query and knowledge embeddings. During fine-tuning, a cross-entropy objective trains the model for tasks like VQA and image captioning. Details of formulas for widely used RAG loss functions can be found in Appendix (§D).

4.5.1 Alignment

Contrastive learning improves representation quality by pulling positive pairs closer and pushing negative pairs apart in the embedding space. A common objective is the InfoNCE loss (van den Oord et al., 2019), which maximizes the mutual information between positive pairs while minimizing similarity to negatives. Several multimodal RAG models, such as VISRAG (Yu et al., 2024), MegaPairs (Zhou et al., 2024a) and SAM-RAG (Zhai, 2024) utilize InfoNCE loss to improve retrieval-augmented generation. Furthermore, EchoSight (Yan and Xie, 2024) enhances retrieval accuracy by selecting visually similar yet contextually distinct negatives, while HACL (Jiang et al., 2024) improves generation by introducing hallucinative captions as distractors. Similarly, UniRaG (Zhi Lim et al., 2024) strengthens retrieval by incorporating hard negative documents, helping the model distinguish relevant contexts from noise.

The eCLIP loss (Kumar and Marttinen, 2024) extends contrastive training by integrating expert-annotated data and an auxiliary Mean Squared Error (MSE) loss to refine embedding quality. Mixup strategies further augment training by generating synthetic positive pairs, improving generalization in contrastive learning (Kumar and Marttinen, 2024). Dense2Sparse (Nguyen et al., 2024) incorporates two unidirectional losses: an image-to-caption loss $\ell(I \to C)$ and a caption-to-image loss $\ell(C \to I)$. It enforces sparsity through L1 regularization, optimizing retrieval precision by balancing dense and sparse representations.

4.5.2 Generation

A key aspect of multimodal RAG is the generation ability. Autoregressive language models are typically trained using Cross-Entropy Loss (Brown et al., 2020). For image generation, widely used approaches include Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Diffusion Models (Ho et al., 2020). GANs employ various loss functions, such as Binary Cross-Entropy Loss, Minimax Loss, Wasserstein Loss (WGAN), and Hinge Loss. Diffusion Models utilize Mean Squared Error (MSE) Loss for noise prediction, a common approach in Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020).

4.5.3 Robustness and Noise Management

Multimodal training faces challenges such as noise and modality-specific biases Buettner and Kovashka (2024). Managing noisy retrieval inputs is critical for maintaining model performance. MORE (Cui et al., 2024) injects irrelevant results during training to enhance focus on relevant inputs. AlzheimerRAG (Lahiri and Hu, 2024) uses progressive knowledge distillation to reduce noise while maintaining multimodal alignment. RAG-Trans (Cheng et al., 2024) leverages hypergraph-based knowledge aggregation to refine multimodal representations, ensuring more effective propagation of relevant information. RA-BLIP (Ding et al., 2024b) introduces the Adaptive Selection Knowledge Generation (ASKG) strategy, which leverages the implicit capabilities of LLMs to filter relevant knowledge for generation through a denoising-enhanced loss term, eliminating the need for fine-tuning. This approach achieves strong performance compared to baselines while significantly reducing computational overhead by minimizing trainable parameters.

RagVL (Chen et al., 2024d) improves robustness through noise-injected training by adding hard negative samples at the data level and applying Gaussian noise with loss reweighting at the token level, enhancing the model's resilience to multimodal noise. Finally, RA-CM3 (Yasunaga et al., 2023) enhances generalization using Query Dropout, which randomly removes query tokens during retrieval, serving as a regularization method that improves generator performance.

5 Tasks Addressed by Multimodal RAGs

Multimodal RAG systems extend RAG beyond unimodal settings to tasks requiring cross-modal integration, enhancing performance across modalities such as text, images, and audio. In content generation, these models enhance image captioning (Zhi Lim et al., 2024; Hu et al., 2023; Rao et al., 2024) and text-to-image synthesis (Yasunaga et al., 2023; Chen et al., 2022b) by retrieving relevant contextual information. They also improve coherence in visual storytelling and ensure factual alignment in multimodal summarization (Tonmoy et al., 2024). In knowledge-intensive applications, multimodal RAG supports open-domain and knowledge-seeking question answering (Chen et al., 2024d; Ding et al., 2024b; Yuan et al., 2023), video-based QA (Luo et al., 2024b), and fact verification (Khaliq et al., 2024), grounding responses in retrieved knowledge and thereby mitigating hallucinations.

Cross-modal retrieval further advances zero-shot image–text retrieval (Yang et al., 2024; Dong et al., 2024b). Additionally, the recent incorporation of chain-of-thought reasoning (Zhai, 2024; Khaliq et al., 2024) has enhanced its ability to support complex problem solving and logical inference. Finally, integrating multimodal RAG into interactive agents and AI assistants such as Gemini (Team et al., 2024) enables natural language-driven visual search, document understanding, and multimodal reasoning. The taxonomy of application domains can be seen in Figure 3. The following sections explore domain-specific adaptations of these techniques in greater depth.

Healthcare and Medicine Multimodal RAG enhances clinical decision-making through integrated analysis of medical imaging, electronic health records, and biomedical literature. Systems like MMED-RAG (Xia et al., 2024a) address diagnostic uncertainty in medical visual question answering by aligning radiology images with contextual patient data. RULE (Xia et al., 2024b) mitigates hallucinations in automated report generation through dynamic retrieval of clinically similar cases. AsthmaBot (Bahaj and Ghogho, 2024) introduces a multimodal RAG-based approach for supporting asthma patients across multiple languages, enabling structured, language-specific semantic searches. Predictive frameworks such as Realm (Zhu et al., 2024c) demonstrate robust risk assessment by fusing heterogeneous patient data streams, while Hybrid RAG (Su et al., 2024a) advances privacy-preserving architectures for federated clinical data integration. FactMM-RAG (Sun et al., 2024b) automates radiology report drafting by retrieving biomarker correlations from medical ontologies, exemplifying the paradigm's capacity to operationalize expert knowledge at scale.

Software Engineering Code generation systems leverage multimodal RAG to synthesize context-aware solutions from technical documentation and version histories. DocPrompting (Zhou et al., 2023) improves semantic coherence in code completion by retrieving API specifications and debugging patterns. Commit message generation models like RACE (Shi et al., 2022) contextualize code diffs against historical repository activity, while CEDAR (Nashid et al., 2023) optimizes few-shot learning through retrieval-based prompt engineering. REDCODER (Parvez et al., 2021) enhances code summarization via semantic search across open-source repositories, preserving syntactic conventions across programming paradigms.

Fashion and E-Commerce Cross-modal alignment drives advancements in product discovery and design automation. UniFashion (Zhao et al., 2024) enables style-aware retrieval by jointly embedding garment images and textual descriptors, while Dang (2024) reduces search friction through multimodal query expansion. LLM4DESIGN (Chen et al., 2024c) demonstrates architectural design automation by retrieving compliance constraints and environmental impact assessments, underscoring RAG's adaptability to creative domains.

Entertainment and Social Computing Multimedia analytics benefit from RAG's capacity to correlate heterogeneous signals. SoccerRAG (Strand et al., 2024) derives tactical insights by linking match footage with player statistics. MMRA (Zhong et al., 2024) predicts content virality through joint modeling of visual aesthetics and linguistic engagement patterns.

Emerging Applications Autonomous systems adopt multimodal RAG for explainable decision-making, as seen in RAG-Driver's (Yuan et al., 2024) real-time retrieval of traffic scenarios during navigation. ENWAR (Nazar et al., 2024) enhances wireless network resilience through multi-sensor fusion, while Riedler and Langer (2024) streamline equipment maintenance by retrieving schematics during fault diagnosis. Geospatial systems such as Img2Loc (Zhou et al., 2024e) advance image geolocalization through cross-modal landmark correlation.

6 Open Problems and Future Directions

Despite rapid advancements in multimodal RAG systems, fundamental challenges remain in achieving robust, efficient, and human-like reasoning across modalities.

6.1 Generalization, Explainability, and Robustness

Multimodal RAG systems often struggle with domain adaptation and exhibit modality biases, frequently over-relying on text for both retrieval and generation (Winterbottom et al., 2020). Explainability remains a major challenge, as these systems typically fail to attribute answers to precise sources. Current methods often cite entire documents or large visual regions as source attribution rather than identifying the exact part of an image, speech, or other modality that led to the answer (Ma et al., 2024b; Hu et al., 2023).

In addition, the interplay between modalities affects the quality of outcomes produced by these models; for example, answers derived solely from text sources may differ in quality compared to those requiring a combination of text and image inputs (Baltrusaitis et al., 2019). They are also vulnerable to adversarial perturbations, such as misleading images influencing textual outputs, and their performance can degrade when relying on low-quality or outdated sources (Chen et al., 2022b). While the trustworthiness of unimodal RAGs has been studied (Zhou et al., 2024d), enhancing the robustness of multimodal RAGs remains an open challenge and a promising research direction.

6.2 Reasoning, Alignment, and Retrieval Enhancement

Multimodal RAGs struggle with compositional reasoning, where information from different modalities must be logically integrated to generate coherent and contextually rich outputs. While cross-modal techniques such as Multimodal-CoT (Zhang et al., 2023b) have been proposed, further innovations are needed to improve the coherence and contextual relevance of multimodal outputs. Enhancing modality alignment and retrieval strategies, particularly for entity-aware retrieval, is essential. Moreover, despite the potential of knowledge graphs to enrich cross-modal reasoning, they remain largely underexplored in multimodal RAGs compared to text-based RAGs (Zhang et al., 2024f; Procko and Ochoa, 2024).

Retrieval biases such as position sensitivity (Hu et al., 2024c), redundant retrieval (Nan et al., 2024), and biases propagated from training data or retrieved content (Zhai, 2024), pose significant challenges that require further attention. Another promising direction is developing a unified embedding space for all modalities, enabling direct multimodal search without intermediary conversion models (e.g., ASRs). Despite some progress, mapping multimodal knowledge into a unified space remains an open challenge with significant potential.

6.3 Agent-Based and Self-Guided Systems

Recent trends indicate a shift towards agent-based multimodal RAGs that integrate retrieval, reasoning, and generation across diverse domains. Unlike static RAG systems, future multimodal RAGs should integrate interactive feedback and self-guided decision-making to iteratively refine outputs. Existing feedback mechanisms often fail to accurately determine whether errors stem from retrieval, generation, or other stages (Dong et al., 2024b). The incorporation of reinforcement learning and end-to-end human-aligned feedback into multimodal RAGs remains largely unexplored but holds significant potential for enhancing these systems. These methods could enable multimodal RAGs to assess whether retrieval is necessary, evaluate the relevance of retrieved content, and dynamically determine the most suitable modalities for response generation. Achieving robust support for any-to-any modality is essential for adaptability in open-ended tasks (Wu et al., 2024b).

Future multimodal RAGs should incorporate data from diverse real-world sources, such as environmental sensors, alongside traditional modalities like text and images, to enhance situational awareness. This progression aligns with the trend toward embodied AI, where models integrate knowledge with physical interaction, enabling applications in robotics, navigation, and physics-informed reasoning. Bridging retrieval-based reasoning with real-world agency brings these systems closer to AGI.

6.4 Long-Context Processing, Efficiency, Scalability, and Personalization

High computational costs in video frame sampling and memory bottlenecks in processing multi-page documents with images remain key challenges in long-context processing. Fixed extraction rates struggle to capture relevant frames, requiring adaptive selection based on content complexity and movement (Kandhare and Gisselbrecht, 2024). Additionally, retrieval speed-accuracy trade-offs in edge deployments and redundant computations in cross-modal fusion layers emphasize the need for efficient, scalable architectures. Personalization mechanisms, such as adapting retrieval to user-specific contexts like medical history, remain underexplored. As personalization mechanisms evolve, ensuring privacy and mitigating the risks of sensitive data leakage in multimodal outputs remain critical challenges. Lastly, the lack of datasets with complex reasoning tasks and multimodal adversarial examples limits robust evaluation.

7 Conclusion

This study provides a comprehensive review of multimodal Retrieval-Augmented Generation (Multimodal RAG) literature. Specifically, we explore and categorize key advancements across different aspects of multimodal RAG systems, including retrieval, multimodal fusion, augmentation, generation, and training strategies. Additionally, we examine the tasks these systems address, their domain-specific applications, and the datasets, benchmarks, and evaluation methods. We also discuss open challenges and limitations in current approaches, along with promising future directions. We hope this work inspires future research, particularly in enhancing cross-modal reasoning and retrieval, developing agent-based interactive systems, and advancing unified multimodal embedding spaces.

8 Limitations

This study offers a comprehensive examination of multimodal RAG systems. Extended discussions, details of datasets and benchmarks, and additional relevant work are available in the Appendices. While we have made our maximum effort; however, some limits may persist. First, due to space constraints, our descriptions of individual methodologies are necessarily concise. Second, although we curate studies from major venues (e.g., ACL, EMNLP, NeurIPS, CVPR, ICLR, ICML, ACM Multimedia) and arXiv, our selection may inadvertently overlook emerging or domain-specific research, with a primary focus on recent advancements. Additionally, this work does not include a comparative performance evaluation of the various models, as task definitions, evaluation metrics, and implementation details vary significantly across studies, and executing these models requires substantial computational resources.

Furthermore, multimodal RAG is a rapidly evolving field with many open questions, such as optimizing fusion strategies for diverse modalities and addressing scalability challenges. As new paradigms emerge, our taxonomy and conclusions will inevitably evolve. To address these gaps, we plan to continuously monitor developments and update this survey and the corresponding repository to incorporate overlooked contributions and refine our perspectives.

9 Ethical Statement

This survey provides a comprehensive review of research on multimodal RAG systems, offering insights that we believe will be valuable to researchers in this evolving field. All the studies, datasets, and benchmarks analyzed in this work are publicly available, with only a very small number of papers requiring institutional access. Additionally, this survey does not involve personal data or user interactions, and we adhere to ethical guidelines throughout.

Since this work is purely a survey of existing literature and does not introduce new models, datasets, or experimental methodologies, it presents no potential risks. However, we acknowledge that multimodal RAG systems inherently raise ethical concerns, including bias, misinformation, privacy, and intellectual property issues. Bias can emerge from both retrieval and generation processes, potentially leading to skewed or unfair outputs. Additionally, these models may hallucinate or propagate misinformation, particularly when retrieval mechanisms fail or rely on unreliable sources. The handling of sensitive multimodal data also poses privacy risks, while content generation raises concerns about proper attribution and copyright compliance. Addressing these challenges requires careful dataset curation, bias mitigation strategies, and transparent evaluation of retrieval and generation mechanisms.

References

Mohammad Mahdi Abootorabi and Ehsaneddin Asgari. 2024. Clasp: Contrastive language-speech pretraining for multilingual multimodal information retrieval. *Preprint*, arXiv:2412.13071.

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2024. Multi-level information retrieval augmented generation for knowledge-based visual question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16499–16513, Miami, Florida, USA. Association for Computational Linguistics.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, et al. 2019. Nocaps: Novel object captioning at scale. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2024. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhiyu An, Xianzhong Ding, Yen-Chun Fu, Cheng-Chung Chu, Yan Li, and Wan Du. 2024. Golden-retriever: High-fidelity agentic retrieval augmented generation for industrial knowledge base. *Preprint*, arXiv:2408.00798.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. irag: Advancing rag for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4341–4348. ACM.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

A. Author and B. Author. 2018. Coco-cn for cross-lingual image tagging, captioning and retrieval. *arXiv preprint arXiv:1805.08661*.

A. Author and B. Author. 2023a. Fashionpedia: A dataset for fashion understanding. *arXiv preprint arXiv:2301.02560*.

A. Author and B. Author. 2023b. Geode: A dataset for geographic dialogue understanding. *arXiv preprint arXiv:2301.02560*.

Adil Bahaj and Mounir Ghogho. 2024. Asthmabot: Multi-modal, multi-lingual retrieval augmented generation for asthma patient support. *arXiv preprint arXiv:2409.15815*.

Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10.

Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Soyuj Basnet, Jerry Gou, Antonio Mallia, and Torsten Suel. 2024. Deeperimpact: Optimizing sparse learned index structures. *Preprint*, arXiv:2405.17093.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2022. Viquae: A dataset for visual question answering on events. *arXiv preprint arXiv:2204.03485*.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kyle Buettner and Adriana Kovashka. 2024. Quantifying the gaps between translation and native perception in training for multimodal, multilingual retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5863–5870, Miami, Florida, USA. Association for Computational Linguistics.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

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*, pages 1818–1826.

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

Jamie Callan, Matthew Hoy, Anagha Kulkarni, et al. 2022. Clueweb22: 10 billion web documents with visual and semantic information. *arXiv preprint arXiv:2211.15848*.

Yee Seng Chan and Hwee Tou Ng. 2008. Maxsim: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2024. Main-rag: Multi-agent filtering retrieval-augmented generation. *Preprint*, arXiv:2501.00332.

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 (CVPR)*, pages 16495–16504.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335.

Ran Chen, Xueqi Yao, and Xuhui Jiang. 2024c. Llm4design: An automated multi-modal system for architectural and environmental design. *arXiv preprint arXiv:2407.12025*.

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

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2022b. Re-imagen: Retrieval-augmented text-to-image generator. *Preprint*, arXiv:2209.14491.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024d. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 445–455, New York, NY, USA. Association for Computing Machinery.

Felix Chern, Blake Hechtman, Andy Davis, Ruiqi Guo, David Majnemer, and Sanjiv Kumar. 2022. Tpu-knn: K nearest neighbor search at peak flop/s. *Advances in Neural Information Processing Systems*, 35:15489–15501.

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

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via image translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14131–14140.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. More: Multi-modal retrieval augmented generative commonsense reasoning. *Preprint*, arXiv:2402.13625.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Quang-Vinh Dang. 2024. Multi-modal retrieval augmented generation for product query. *Library of Progress-Library Science, Information Technology & Computer*, 44(3).

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.*, 55(13s).

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Damen Dima. 2020. Rescaling egocentric vision. *Comput. Res. Reposit.*, 2006.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024a. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *Preprint*, arXiv:2402.10612.

Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, and Liqiang Nie. 2024b. Ra-blip: Multimodal adaptive retrieval-augmented bootstrapping language-image pre-training. *Preprint*, arXiv:2410.14154.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024c. Pdf-mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *Preprint*, arXiv:2404.12720.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024a. Don't forget to connect! improving rag with graph-based reranking. *Preprint*, arXiv:2405.18414.

Xingning Dong, Zipeng Feng, Chunluan Zhou, Xuzheng Yu, Ming Yang, and Qingpei Guo. 2024b. M2-raap: A multi-modal recipe for advancing adaptation-based pre-training towards effective and efficient zero-shot video-text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2156–2166, New York, NY, USA. Association for Computing Machinery.

Yang Du, Yuqi Liu, and Qin Jin. 2024. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5260–5269, New York, NY, USA. Association for Computing Machinery.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Eli5: Long form question answering. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

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

Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman Khan, Wangmeng Zuo, Xinxing Xu, Rick Siow Mong Goh, and Yong Liu. 2023. Vqa4cir: Boosting composed image retrieval with visual question answering. *Preprint*, arXiv:2312.12273.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, et al. 2017. Audioset: An ontology and human-labeled dataset for audio events. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yash Goyal, Tushar Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Vqa v2: Visual question answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR.

M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. 2024. HEALNet: Multimodal fusion for heterogeneous biomedical data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, Miami, Florida, USA. Association for Computational Linguistics.

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

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024c. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. 2024. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16814.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. Videorag: Retrieval-augmented generation over video corpus. *Preprint*, arXiv:2501.05874.

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

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956, Miami, Florida, USA. Association for Computational Linguistics.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. 2024. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27144–27153.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, et al. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. 2024. Robust multi model rag pipeline for documents containing text, table & images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999. IEEE.

Mahesh Kandhare and Thibault Gisselbrecht. 2024. An empirical comparison of video frame sampling methods for multi-modal rag retrieval. *Preprint*, arXiv:2408.03340.

Yasser Khalafaoui, Martino Lovisetto, Basarab Matei, and Nistor Grozavu. 2024. Cadmr: Cross-attention and disentangled learning for multimodal recommender systems. *Preprint*, arXiv:2412.02295.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *Preprint*, arXiv:1812.08466.

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

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 2024. Do you remember?

dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 554–561.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yogesh Kumar and Pekka Marttinen. 2024. Improving medical multi-modal contrastive learning with expert annotations. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XX*, page 468–486, Berlin, Heidelberg. Springer-Verlag.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Aritra Kumar Lahiri and Qinmin Vivian Hu. 2024. Alzheimerrag: Multimodal retrieval augmented generation for pubmed articles. *Preprint*, arXiv:2412.16701.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *arXiv preprint arXiv:2501.01120*.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. 2024. PlanRAG: A plan-then-retrieval augmented generation for generative large language models as decision makers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6537–6555, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2021. Infoseek: A dataset for multimodal question answering. *arXiv preprint arXiv:2104.06039*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2022b. Webqa: Multimodal question answering on web videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Muquan Li, Dongyang Zhang, Qiang Dong, Xiurui Xie, and Ke Qin. 2024a. Adaptive dataset quantization. *Preprint*, arXiv:2412.16895.

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

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024a. Mm-embed: Universal multimodal retrieval with multimodal llms. *Preprint*, arXiv:2411.02571.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, pages 740–755.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024b. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5294–5316, Bangkok, Thailand. Association for Computational Linguistics.

Di Liu, Meng Chen, Baotong Lu, Huiqiang Jiang, Zhenhua Han, Qianxi Zhang, Qi Chen, Chengruidong Zhang, Bailu Ding, Kai Zhang, Chen Chen, Fan Yang, Yuqing Yang, and Lili Qiu. 2024a. Retrievalattention: Accelerating long-context llm inference via vector retrieval. *Preprint*, arXiv:2409.10516.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

Xingzu Liu, Mingbang Wang, Songhang Deng, Xinyue Peng, Yanming Liu, Ruilin Nong, David Williams, and Jiyuan Li. 2024b. Iterative retrieval augmentation for multi-modal knowledge integration and generation.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024c. RA-ISF: Learning to answer and understand from retrieval augmentation via iterative self-feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4730–4749, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 216–233. Springer.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023b. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proceedings of ICLR*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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. *Preprint*, arXiv:2405.20797.

Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. 2024a. How does the textual information affect the retrieval of multimodal in-context learning? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5321–5335, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024b. Visa: Retrieval augmented generation with visual source attribution. *Preprint*, arXiv:2412.14457.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024c. Multi-modal retrieval augmented multi-modal generation: A benchmark, evaluate metrics and strong baselines. *Preprint*, arXiv:2411.16365.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2024d. Drvideo: Document retrieval based long video understanding. *Preprint*, arXiv:2406.12846.

Zhiming Mao, Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Visually guided generative text-layout pre-training for document intelligence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4713–4730, Mexico City, Mexico. Association for Computational Linguistics.

Kenneth Marino, Xinlei Chen, Abhinav Gupta, Marcus Rohrbach, and Devi Parikh. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

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

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

Matin Mortaheb, Mohammad A. Amir Khojastepour, Srimat T. Chakradhar, and Sennur Ulukus. 2025b. Re-ranking the context for multimodal retrieval augmented generation. *Preprint*, arXiv:2501.04695.

Linyong Nan, Weining Fang, Aylin Rasteh, Pouya Lahabi, Weijin Zou, Yilun Zhao, and Arman Cohan. 2024. OMG-QA: Building open-domain multimodal generative question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1001–1015, Miami, Florida, US. Association for Computational Linguistics.

Noor Nashid, Mifta Sintaha, and Ali Mesbah. 2023. Retrieval-based prompt selection for code-related few-shot learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2450–2462.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

Ahmad M Nazar, Abdulkadir Celik, Mohamed Y Selim, Asmaa Abdallah, Daji Qiao, and Ahmed M Eltawil. 2024. Enwar: A rag-empowered multimodal llm framework for wireless environment perception. *arXiv preprint arXiv:2410.18104*.

Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal learned sparse retrieval with probabilistic expansion control. In *Advances in Information Retrieval*, pages 448–464, Cham. Springer Nature Switzerland.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Weihua Ou, Yingjie Chen, Linqing Liang, Jianping Gou, Jiahao Xiong, Jiacheng Zhang, Lingge Lai, and Lei Zhang. 2025. Cross-modal retrieval of chest x-ray images and diagnostic reports based on report entity graph and dual attention: Cross-modal retrieval of chest x-ray images and diagnostic reports... *Multimedia Syst.*, 31(1).

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2024. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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

Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *EMNLP-Findings*.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. 2023. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *TMLR*.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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*, pages 8748–8763. PMLR.

Varun Nagaraj Rao, Siddharth Choudhary, Aditya Deshpande, Ravi Kumar Satzoda, and Srikar Appalaraju. 2024. Raven: Multitask retrieval augmented vision-language learning. *Preprint*, arXiv:2406.19150.

David Rau, Shuai Wang, Hervé Déjean, and Stéphane Clinchant. 2024. Context embeddings for efficient answer generation in rag. *Preprint*, arXiv:2407.09252.

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

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

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

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Romain Vencu, Richard Beaumont, Robert Kaczmarczyk, Jenia Jitsev, Atsushi Komatsuzaki, et al. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

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*, pages 146–162.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2024. Unirag: Universal retrieval augmentation for multi-modal large language models. *ArXiv*, abs/2405.10311.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Xiaobo Shen, Qianxin Huang, Long Lan, and Yuhui Zheng. 2024. Contrastive transformer cross-modal hashing for video-text retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1227–1235. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ensheng Shi, Yanlin Wang, Wei Tao, Lun Du, Hongyu Zhang, Shi Han, Dongmei Zhang, and Hongbin Sun. 2022. RACE: Retrieval-augmented commit message generation. In *Proceedings of the 2022 Conference* on Empirical Methods in Natural Language Processing, pages 5520–5530, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Faisal Tareque Shohan, Mir Tafseer Nayeem, Samsul Islam, Abu Ubaida Akash, and Shafiq Joty. 2024. XL-HeadTags: Leveraging multimodal retrieval augmentation for the multilingual generation of news headlines and tags. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12991–13024, Bangkok, Thailand. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gunnar A Sigurdsson, Gul Varol, Giovanni Maria Farinella, et al. 2016. Charades: A dataset for multimodal research. *arXiv preprint arXiv:1604.01753*.

Gunnar A Sigurdsson, Gul Varol, Giovanni Maria Farinella, et al. 2018. Charadesego: A dataset for egocentric video understanding. *arXiv preprint arXiv:1804.09626*.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *Preprint*, arXiv:2501.09136.

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*, pages 15638–15650.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*.

Aleksander Theo Strand, Sushant Gautam, Cise Midoglu, and Pål Halvorsen. 2024. Soccerrag: Multimodal soccer information retrieval via natural queries. *Preprint*, arXiv:2406.01273.

Cheng Su, Jinbo Wen, Jiawen Kang, Yonghua Wang, Yuanjia Su, Hudan Pan, Zishao Zhong, and M Shamim Hossain. 2024a. Hybrid rag-empowered multi-modal llm for secure data management in internet of medical things: A diffusion-based contract approach. *IEEE Internet of Things Journal*.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. 2024b. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. *arXiv preprint arXiv:2406.19593*.

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024a. Surf: Teaching large vision-language models to selectively utilize retrieved information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7611–7629.

Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong. 2024b. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *Preprint*, arXiv:2407.15268.

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 preprint arXiv:2412.10704*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Cheng Tan, Jingxuan Wei, Linzhuang Sun, Zhangyang Gao, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z. Li. 2024. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning. *Preprint*, arXiv:2405.20834.

Yansong Tang, Xiaohan Wang, Jingdong Wang, et al. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, and et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Mo Tiwari, Ryan Kang, Jaeyong Lee, Donghyun Lee, Christopher J Piech, Sebastian Thrun, Ilan Shomorony, and Martin Jinye Zhang. 2024. Faster maximum inner product search in high dimensions. In *Forty-first International Conference on Machine Learning*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024b. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16551–16560.

Mengzhao Wang, Xiangyu Ke, Xiaoliang Xu, Lu Chen, Yunjun Gao, Pinpin Huang, and Runkai Zhu. 2024c. Must: An effective and scalable framework for multimodal search of target modality. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 4747–4759. IEEE.

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

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. 2023. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020a. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xin Wang, Jiawei Wu, Junkun Chen, et al. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–10.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. Internvideo: General video foundation models via generative and discriminative learning. *Preprint*, arXiv:2212.03191.

Zhihao Wang, Jian Chen, and Steven C. H. Hoi. 2020b. Deep learning for image super-resolution: A survey. *Preprint*, arXiv:1902.06068.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024a. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024b. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*, page 387–404, Berlin, Heidelberg. Springer-Verlag.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024c. Chain-of-thought prompting elicits reasoning in large language models. In

Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2024. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 229–239. ACM.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. 2020. On modality bias in the tvqa dataset. *Preprint*, arXiv:2012.10210.

Chenyun Wu, Yuting Liu, and Gang Hua. 2019. Fashion iq: A new dataset towards retrieving images by natural language feedback. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11316.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. 2024a. Synthetic multimodal question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA. Association for Computational Linguistics.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024b. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, pages 53366–53397.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *Preprint*, arXiv:2410.13085.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.

D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. 2023. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *Preprint*, arXiv:2306.04362.

Huazhe Xu, Yuan Gao, Fisher Yu, and Trevor Darrell. 2018. Bdd-x: A dataset for explainable driving behavior. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–10.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. 2024a. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Junxiao Xue, Quan Deng, Fei Yu, Yanhao Wang, Jun Wang, and Yuehua Li. 2024. Enhanced multimodal rag-llm for accurate visual question answering. *Preprint*, arXiv:2412.20927.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, Miami, Florida, USA. Association for Computational Linguistics.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5223–5234, New York, NY, USA. Association for Computing Machinery.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 80–90, New York, NY, USA. Association for Computing Machinery.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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

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*, volume 33, pages 9127–9134.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *Preprint*, arXiv:2402.10828.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556.

Jiaqi Zhai, Zhaojie Gong, Yueming Wang, Xiao Sun, Zheng Yan, Fu Li, and Xing Liu. 2023. Revisiting neural retrieval on accelerators. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5520–5531, New York, NY, USA. Association for Computing Machinery.

Wenjia Zhai. 2024. Self-adaptive multimodal retrieval-augmented generation. *Preprint*, arXiv:2410.11321.

Haoyu Zhang, Jun Liu, Zhenhua Zhu, Shulin Zeng, Maojia Sheng, Tao Yang, Guohao Dai, and Yu Wang. 2024a. Efficient and effective retrieval of dense-sparse hybrid vectors using graph-based approximate nearest neighbor search. *Preprint*, arXiv:2410.20381.

Jin Zhang, Defu Lian, Haodi Zhang, Baoyun Wang, and Enhong Chen. 2023a. Query-aware quantization for maximum inner product search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4875–4883.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024b. Cream: Coarse-to-fine retrieval and multi-modal efficient tuning for document vqa. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 925–934, New York, NY, USA. Association for Computing Machinery.

Juntao Zhang, Yuehuai Liu, Yu-Wing Tai, and Chi-Keung Tang. 2024c. C3net: Compound conditioned controlnet for multimodal content generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26886–26895.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2024d. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. *arXiv preprint arXiv:2412.02592*.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024e. OmAgent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10031–10045, Miami, Florida, USA. Association for Computational Linguistics.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfen Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024f. mr2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *ArXiv*, abs/2411.15041.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024g. mr^{2}ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *Preprint*, arXiv:2411.15041.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024h. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024i. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023a. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, Singapore. Association for Computational Linguistics.

Xi Zhao, Bolong Zheng, Xiaomeng Yi, Xiaofan Luan, Charles Xie, Xiaofang Zhou, and Christian S. Jensen. 2023b. Fargo: Fast maximum inner product search via global multi-probing. *Proc. VLDB Endow.*, 16(5):1100–1112.

Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. 2024. Unifashion: A unified vision-language model for multimodal fashion retrieval and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1490–1507, Miami, Florida, USA. Association for Computational Linguistics.

Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. 2023. Prototype-Based Embedding Network for Scene Graph Generation . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22783–22792, Los Alamitos, CA, USA. IEEE Computer Society.

Qi Zhi Lim, Chin Poo Lee, Kian Ming Lim, and Ahmad Kamsani Samingan. 2024. Unirag: Unification, retrieval, and generation for multimodal question answering with pre-trained language models. *IEEE Access*, 12:71505–71519.

Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. 2024. Predicting micro-video popularity via multi-modal retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2579–2583, New York, NY, USA. Association for Computing Machinery.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE.

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

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

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1–10.

Ren Zhou. 2024. Advanced embedding techniques in multimodal retrieval augmented generation a comprehensive study on cross modal ai applications. *Journal of Computing and Electronic Information Management*, 13(3):16–22.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, and Graham Neubig. 2023. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024c. Marvel: unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024d. Trustworthiness in retrieval-augmented generation systems: A survey. volume abs/2409.10102.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024e. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024a. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024b. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3549–3559, New York, NY, USA. Association for Computing Machinery.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and Chengwei Pan. 2024c. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *Preprint*, arXiv:2402.07016.

Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. 2025. Murar: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 126–135.

Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. 2024. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019.

A Taxonomy

In this section, we provide more details regarding the taxonomy of multimodal RAG systems, previously mentioned in Figure 2. Additionally, we present a classification of multimodal RAG application domains in Figure 3.

Figure 2 provides an overview of recent advances in multimodal retrieval-augmented generation (RAG) systems. The taxonomy is organized into several key categories.

- **Retrieval strategies** cover efficient search and similarity retrieval methods (including maximum inner product search (MIPS) variants and different multi-modal encoders) and modality-centric techniques that distinguish between text-, vision-, and video-centric as well as document retrieval models. Reranking strategies further refine these methods via optimized example selection, relevance scoring, and filtering.

- **Fusion mechanisms** are implemented through score fusion and alignment, attention-based techniques, and unified frameworks that project multimodal information into common representations.

- **Augmentation techniques** address context enrichment as well as adaptive and iterative retrieval.

- **Generation methods** include in-context learning, reasoning, instruction tuning, and source attribution.

- **training strategies** are characterized by approaches to alignment and robustness.

Detailed discussions of these categories are provided in the corresponding sections.
Figure 3 presents the taxonomy of application domains for multimodal RAG systems. The identified domains include *healthcare and medicine*, *software engineering*, *fashion and e-commerce*, *entertainment and social computing*, and *emerging applications*. This classification offers a concise overview of the diverse applications and serves as a framework for the more detailed analyses that follow.

B Dataset and Benchmark

Table 1 and Table 2 present a comprehensive overview of datasets and benchmarks commonly

Figure 3: Taxonomy of application domains for Multimodal Retrieval-Augmented Generation systems.

employed in multimodal RAG research. The table is organized into five columns:

- **Category:** This column categorizes each dataset or benchmark based on its primary domain or modality. The datasets are grouped into eight categories: *Image–Text General*, *Video–Text*, *Audio–Text*, *Medical*, *Fashion*, *3D*, *Knowledge & QA*, and *Other*. The benchmarks are grouped into two categories: *Cross-Modal Understanding* and *Text-Focused*. This classification facilitates a clearer understanding of each dataset or benchmark's role within a multimodal framework.

- **Name:** The official name of the dataset or benchmarks is provided along with a citation for reference.

- **Statistics and Description:** This column summarizes key details such as dataset size, the nature of the content (e.g., image–text pairs, video captions, QA pairs), and the specific tasks or applications for which the dataset or benchmarks are used. These descriptions are intended to convey the dataset's scope and its relevance to various multimodal RAG tasks.

- **Modalities:** The modalities covered by each dataset or benchmark are indicated (e.g., Image, Text, Video, Audio, or 3D). Notably, several datasets are unimodal; however, within multimodal RAG systems, these are combined with others to represent distinct aspects of a broader multimodal context.

- **Link:** A hyperlink is provided to direct readers to the official repository or additional resources for the dataset or benchmark, thereby facilitating further exploration of its properties and applications.

C Metrics

Accuracy: Accuracy is typically defined as the ratio of correctly predicted instances to the total instances. In retrieval-based tasks, Top-K Accuracy is defined as:

$$\text{Top-K Accuracy}(y, \hat{f}) = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=1}^{k} \nVdash(\hat{f}_{i,j} = y_i) \tag{3}$$

FID (Fréchet inception distance): FID is a metric used to assess the quality of images created by a generative model. The formula for FID is:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + tr(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \tag{4}$$

where μ_r and Σ_r are the mean and covariance of real images' feature representations, respectively. μ_g and Σ_g are the mean and covariance of generated images' feature representations, respectively. To extract features, InceptionV3 is typically used.

ROUGE-N (N-gram Recall): The ROUGE metric is commonly used to evaluate text summarization and generation. ROUGE-N measures the overlap of N-grams between the generated and reference text. The formula for ROUGE-N is:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}(\text{gram}_N)} \tag{5}$$

ROUGE-L measures the longest common subsequence (LCS) between generated and reference text. The formula for ROUGE-L is:

$$\text{ROUGE-L} = \frac{LCS(X, Y)}{|Y|} \tag{6}$$

BLEU: BLEU is another metric used for assessing text generation. The formula for calculating BLEU is:

$$\text{BLEU}(p_n, \text{BP}) = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

Here, p_n represents the precision of n-grams, w_n denotes the weight assigned to the n-gram precision, and the Brevity Penalty (BP) is defined as:

$$\text{BP} = \begin{cases} 1 & \text{length} > rl \\ \exp\left(1 - \frac{rl}{cl}\right) & \text{length} \leq rl \end{cases} \tag{8}$$

Here, rl represents the reference length and cl represents the candidate length.

BERTScore: BERTScore is a metric for evaluating the quality of text generation, based on the similarity between the contextual embeddings of words in the candidate and reference texts. The formula for calculating BERTScore is:

$$\text{BERTScore}(c, r) = \frac{1}{|c|} \sum_{i=1}^{|c|} \max_{j=1}^{|r|} \cos(\mathbf{e}_i, \mathbf{e}_j) \tag{9}$$

c is the candidate sentence, and r is the reference sentence. e_i and e_j are the embeddings (e.g., from BERT) for words c_i and r_j in the candidate and reference sentences, respectively.

CLIPScore: CLIPScore is a metric that evaluates the similarity between the text and an image by using the CLIP model. The formula for calculating CLIPScore is:

$$\text{CLIPScore} = \frac{\cos(\mathbf{t}, \mathbf{i})}{||\mathbf{t}|| \cdot ||\mathbf{i}||} \tag{10}$$

where t and i are text and image embedding respectively.

Mean Reciprocal Rank (MRR): MRR is a metric used to evaluate the performance of systems that return a list of results, such as search engines or recommendation systems. MRR measures the rank position of the first relevant result in the returned list. The formula for calculating MRR is:

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q} \tag{11}$$

where Q is the total number of queries. $rank_q$ is the rank of the first relevant result for query q.

D Loss Function

InfoNCE (Information Noise Contrastive Estimation): The InfoNCE loss is commonly used in self-supervised learning, especially in contrastive learning methods. The formula for InfoNCE loss is:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{12}$$

where z_i and z_j are the embeddings of a positive pair and τ is a temperature parameter.

GAN (Generative Adversarial Network): The GAN loss consists of two parts: the discriminator loss and the generator loss. The discriminator loss formula is:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{13}$$

where x is a real sample from the data distribution. $G(z)$ is the generated sample from the generator, where z is a noise vector. $D(x)$ is the probability that x is real.
The Generator loss formula is:

$$\mathcal{L}_G = \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{14}$$

Triplet Loss: Triplet Loss is used in metric learning to ensure that similar data points are closer together while dissimilar ones are farther apart in the embedding space. The Triplet loss formula is:

$$\mathcal{L} = \sum_{i=1}^{N} \max(0, ||f(x_a^i) - f(x_p^i)||^2 - ||f(x_a^i) - f(x_n^i)||^2 + \alpha) \tag{15}$$

where x_a^i is the anchor sample. x_p^i and x_n^i are the positive and negative samples respectively. $f(x)$ is the neural network.

Table 1: Overview of Popular Datasets in Multimodal RAG Research.

Category	Name	Statistics and Description	Modalities	Link
Image-Text General	LAION-400M (Schuhmann et al., 2021)	200M image–text pairs; used for pre-training multimodal models.	Image, Text	LAION-400M
	Conceptual-Captions (CC) (Sharma et al., 2018)	15M image–caption pairs; multilingual English–German image descriptions.	Image, Text	Conceptual Captions
	CIRR (Liu et al., 2021)	36,554 triplets from 21,552 images; focuses on natural image relationships.	Image, Text	CIRR
	MS-COCO (Lin et al., 2014)	330K images with captions; used for caption–to–image and image–to–caption generation.	Image, Text	MS-COCO
	Flickr30K (Young et al., 2014)	31K images annotated with five English captions per image.	Image, Text	Flickr30K
	Multi30K (Elliott et al., 2016)	30k German captions from native speakers and human–translated captions.	Image, Text	Multi30K
	NoCaps (Agrawal et al., 2019)	For zero–shot image captioning evaluation; 15K images.	Image, Text	NoCaps
	Laion-5B (Schuhmann et al., 2022)	5B image–text pairs used as external memory for retrieval.	Image, Text	LAION-5B
	COCO-CN (Author and Author, 2018)	20,341 images for cross-lingual tagging and captioning with Chinese sentences.	Image, Text	COCO-CN
	CIRCO (Baldrati et al., 2023)	1,020 queries with an average of 4.53 ground truths per query; for composed image retrieval.	Image, Text	CIRCO
Video-Text	BDD-X (Xu et al., 2018)	77 hours of driving videos with expert textual explanations; for explainable driving behavior.	Video, Text	BDD-X
	YouCook2 (Zhou et al., 2018)	2,000 cooking videos with aligned descriptions; focused on video–text tasks.	Video, Text	YouCook2
	ActivityNet (Caba Heilbron et al., 2015)	20,000 videos with multiple captions; used for video understanding and captioning.	Video, Text	ActivityNet
	SoccerNet (Giancola et al., 2018)	Videos and metadata for 550 soccer games; includes transcribed commentary and key event annotations.	Video, Text	SoccerNet
	MSR-VTT (Xu et al., 2016)	10,000 videos with 20 captions each; a large video description dataset.	Video, Text	MSR-VTT
	MSVD (Chen and Dolan, 2011)	1,970 videos with approximately 40 captions per video.	Video, Text	MSVD
	LSMDC (Rohrbach et al., 2015)	118,081 video–text pairs from 202 movies; a movie description dataset.	Video, Text	LSMDC
	DiDemo (Anne Hendricks et al., 2017)	10,000 videos with four concatenated captions per video; with temporal localization of events.	Video, Text	DiDemo
	Breakfast (Kuehne et al., 2014)	1,712 videos of breakfast preparation; one of the largest fully annotated video datasets.	Video, Text	Breakfast
	COIN (Tang et al., 2019)	11,827 instructional YouTube videos across 180 tasks; for comprehensive instructional video analysis.	Video, Text	COIN
	MSRVTT-QA (Xu et al., 2017)	Video question answering benchmark.	Video, Text	MSRVTT-QA
	MSVD-QA (Xu et al., 2017)	1,970 video clips with approximately 50.5K QA pairs; video QA dataset.	Video, Text	MSVD-QA
	ActivityNet-QA (Yu et al., 2019)	58,000 human–annotated QA pairs on 5,800 videos; benchmark for video QA models.	Video, Text	ActivityNet-QA
	EpicKitchens-100 (Dima, 2020)	700 videos (100 hours of cooking activities) for online action prediction; egocentric vision dataset.	Video, Text	EPIC-KITCHENS-100
	Ego4D (Grauman et al., 2022)	4.3M video–text pairs for egocentric videos; massive–scale egocentric video dataset.	Video, Text	Ego4D
	HowTo100M (Miech et al., 2019)	136M video clips with captions from 1.2M YouTube videos; for learning text–video embeddings.	Video, Text	HowTo100M
	CharadesEgo (Sigurdsson et al., 2018)	68,536 activity instances from ego–exo videos; used for evaluation.	Video, Text	Charades-Ego
	ActivityNet Captions (Krishna et al., 2017)	20K videos with 3.7 temporally localized sentences per video; dense–captioning events in videos.	Video, Text	ActivityNet Captions
	VATEX (Wang et al., 2019)	34,991 videos, each with multiple captions; a multilingual video–and–language dataset.	Video, Text	VATEX
	Charades (Sigurdsson et al., 2016)	9,848 video clips with textual descriptions; a multimodal research dataset.	Video, Text	Charades
	WebVid (Bain et al., 2021)	10M video–text pairs (refined to WebVid-Refined-1M).	Video, Text	WebVid
	Youku-mPLUG (Xu et al., 2023)	Chinese dataset with 10M video–text pairs (refined to Youku-Refined-1M).	Video, Text	Youku-mPLUG
Audio-Text	LibriSpeech (Panayotov et al., 2015)	1,000 hours of read English speech with corresponding text; ASR corpus based on audiobooks.	Audio, Text	LibriSpeech
	SpeechBrown (Abootorabi and Asgari, 2024)	55K paired speech-text samples; 15 categories covering diverse topics from religion to fiction.	Audio, Text	SpeechBrown
	AudioCap (Kim et al., 2019)	46K audio clips paired with human-written text captions.	Audio, Text	AudioCaps
	AudioSet (Gemmeke et al., 2017)	2,084,320 human–labeled 10–second sound clips from YouTube; 632 audio event classes.	Audio, Text	AudioSet
Medical	MIMIC-CXR (Johnson et al., 2019)	125,417 training pairs of chest X–rays and reports.	Image, Text	MIMIC-CXR
	CheXpert (Irvin et al., 2019)	224,316 chest radiographs of 65,240 patients; focused on medical analysis.	Image, Text	CheXpert
	MIMIC-III (Johnson et al., 2016)	Health-related data from over 40K patients (text data).	Text	MIMIC-III
	IU-Xray (Pavlopoulos et al., 2019)	7,470 pairs of chest X–rays and corresponding diagnostic reports.	Image, Text	IU X-ray
	PubLayNet (Zhong et al., 2019)	100,000 training samples and 2,160 test samples built from PubLayNet (tailored for the medical domain).	Image, Text	PubLayNet
Fashion	Fashion-IQ (Wu et al., 2019)	77,684 images across three categories; evaluated with Recall@10 and Recall@50.	Image, Text	Fashion IQ
	FashionGen (Hadi Kiapour et al., 2018)	260.5K image–text pairs of fashion images and item descriptions.	Image, Text	Fashion-Gen
	VITON-HD (Choi et al., 2021)	83K images for virtual try–on; high–resolution clothing items.	Image, Text	VITON-HD
	Fashionpedia (Author and Author, 2023a)	48,000 fashion images annotated with segmentation masks and fine-grained attributes.	Image, Text	Fashionpedia
	DeepFashion (Liu et al., 2016)	Approximately 800K diverse fashion images for pseudo triplet generation.	Image, Text	DeepFashion
3D	ShapeNet (Chang et al., 2015)	7,500 text–3D data pairs; repository for 3D CAD models.	Text, 3D	ShapeNet
Knowledge & QA	VQA (Antol et al., 2015)	400K QA pairs with images for visual question answering.	Image, Text	VQA
	PAQ (Lewis et al., 2021)	65M text–based QA pairs; a large–scale dataset.	Text	PAQ
	ELI5 (Fan et al., 2019)	270K complex and diverse questions augmented with web pages and images.	Text	ELI5
	ViQuAE (Biten et al., 2022)	11.8M passages from Wikipedia covering 2,397 unique entities; knowledge–intensive QA.	Text	ViQuAE
	OK-VQA (Marino et al., 2019)	14K questions requiring external knowledge for VQA.	Image, Text	OK-VQA
	WebQA (Li et al., 2022b)	46K queries that require reasoning across text and images.	Text, Image	WebQA
	Infoseek (Li et al., 2021)	Fine-grained visual knowledge retrieval using a Wikipedia–based knowledge base (6M passages).	Image, Text	Infoseek
	ClueWeb22 (Callan et al., 2022)	10 billion web pages organized into three subsets; a large–scale web corpus.	Text	ClueWeb22
	MOCHEG (Yao et al., 2023)	15,601 claims annotated with truthfulness labels and accompanied by textual and image evidence.	Text, Image	MOCHEG
	VQA v2 (Goyal et al., 2017b)	1.1M questions (augmented with VG–QA questions) for fine-tuning VQA models.	Image, Text	VQA v2
	A-OKVQA (Schwenk et al., 2022)	Benchmark for visual question answering using world knowledge; around 25K questions.	Image, Text	A-OKVQA
	XL-HeadTags (Shohan et al., 2024)	415K news headline-article pairs consist of 20 languages across six diverse language families.	Text	XL-HeadTags
	SEED-Bench (Li et al., 2023a)	19K multiple–choice questions with accurate human annotations across 12 evaluation dimensions.	Text	SEED-Bench
Other	ImageNet (Deng et al., 2009)	14,197,122 images for perspective understanding; a hierarchical image database.	Image	ImageNet
	Oxford Flowers102 (Nilsback and Zisserman, 2008)	102 flower categories with five examples per category; image classification dataset.	Image	Oxford Flowers102
	Stanford Cars (Krause et al., 2013)	Images of different car models (five examples per model); for fine-grained categorization.	Image	Stanford Cars
	GeoDE (Author and Author, 2023b)	61,940 images from 40 classes across 6 world regions; emphasizes geographic diversity in object recognition.	Image	GeoDE

Table 2: Overview of Popular Benchmarks in Multimodal RAG Research.

Category	**Name**	**Statistics and Description**	**Modalities**	**Link**
Cross-Modal Understanding	MRAG-Bench (Hu et al., 2024c)	Evaluates visual retrieval, integration, and robustness to irrelevant visual information.	Images	MRAG-Bench
	M2RAG (Ma et al., 2024c)	Benchmarks multimodal RAG; evaluates retrieval, multi-hop reasoning, and integration.	Images + Text	M2RAG
	Dyn-VQA (Li et al., 2024b)	Focuses on dynamic retrieval, multi-hop reasoning, and robustness to changing information.	Images + Text	Dyn-VQA
	MMBench (Liu et al., 2025)	Covers VQA, captioning, retrieval; evaluates cross-modal understanding across vision, text, and audio.	Images + Text + Audio	MMBench
	ScienceQA (Saikh et al., 2022)	Contains 21,208 questions; tests scientific reasoning with text, diagrams, and images.	Images + Diagrams + Text	ScienceQA
	SK-VQA (Su et al., 2024b)	Offers 2 million question-answer pairs; focuses on synthetic knowledge, multimodal reasoning, and external knowledge integration.	Images + Text	SK-VQA
	SMMQG (Wu et al., 2024a)	Includes 1,024 question-answer pairs; focuses on synthetic multimodal data and controlled question generation.	Images + Text	SMMQG
Text-Focused	TriviaQA (Joshi et al., 2017)	Provides 650K question-answer pairs; reading comprehension dataset, adaptable for multimodal RAG.	Text	TriviaQA
	Natural Questions (Kwiatkowski et al., 2019)	Contains 307,373 training examples; real-world search queries, adaptable with visual contexts.	Text	Natural Questions