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[†,*]

[†]Qatar Computing Research Institute, [‡]Saarland University, [♣] Zuse School ELIZA, [◇]University of Tehran,
[♠]Max Planck Institute for Software Systems, [¶]K.N. Toosi University of Technology, [§]Sharif University of Technology

Correspondence: soleymani@sharif.edu and easgari@hbku.edu.qa

https://multimodalrag.github.io

Abstract

Large Language Models (LLMs) suffer from hallucinations and outdated knowledge due to their reliance on static training data. Retrieval-Augmented Generation (RAG) mitigates these issues by integrating external dynamic information for improved factual grounding. With advances in multimodal learning, Multimodal RAG extends this approach by incorporating multiple modalities such as text, images, audio, and video to enhance the generated outputs. However, cross-modal alignment and reasoning introduce unique challenges beyond those in unimodal RAG. This survey offers a structured and comprehensive analysis of Multimodal RAG systems, covering datasets, benchmarks, metrics, evaluation, methodologies, and innovations in retrieval, fusion, augmentation, and generation. We review training strategies, robustness enhancements, loss functions, and agent-based approaches, while also exploring the diverse Multimodal RAG scenarios. In addition, we outline open challenges and future directions to guide research 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. All resources are publicly available [1].

1 Introduction & Background

Recent advancements in transformer architectures (Vaswani et al., 2017), coupled with increased computational resources and the availability of large-scale training datasets (Naveed et al., 2024), have significantly accelerated progress in the development of language models. 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), excelling in tasks such as instruction following (Qin et al., 2024), reasoning (Wei et al., 2024b), in-context learning (Brown et al., 2020), and multilingual translation (Zhu et al., 2024a). Despite these achievements, LLMs face challenges such as hallucinations, outdated knowledge, and a lack of verifiable reasoning (Huang et al., 2024; Xu et al., 2024b). Their reliance on parametric memory limits access to up-to-date information, reducing their effectiveness in knowledge-intensive tasks.

Retrieval-Augmented Generation (RAG) RAG (Lewis et al., 2020) addresses 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 external knowledge sources, RAG enhances knowledge-intensive tasks while grounding responses in verifiable sources (Gao et al., 2023). In practice, RAG systems follow a retriever-generator pipeline: the retriever uses embedding models (Chen et al., 2024a; Rau et al., 2024) to identify relevant passages from external knowledge bases and may apply re-ranking techniques to improve precision (Dong et al., 2024a). The retrieved passages are then provided to the generator, which leverages this contextual information to produce more informed and coherent 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), have further improved both the retrieval and generation components of these systems.

Multimodal Learning In parallel with advances in language modeling, multimodal learning has emerged as a transformative area in artificial intelli-

[1]`https://github.com/llm-lab-org/Multimodal-RAG-Survey`

*These authors contributed equally.

♣This author is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16776–16809

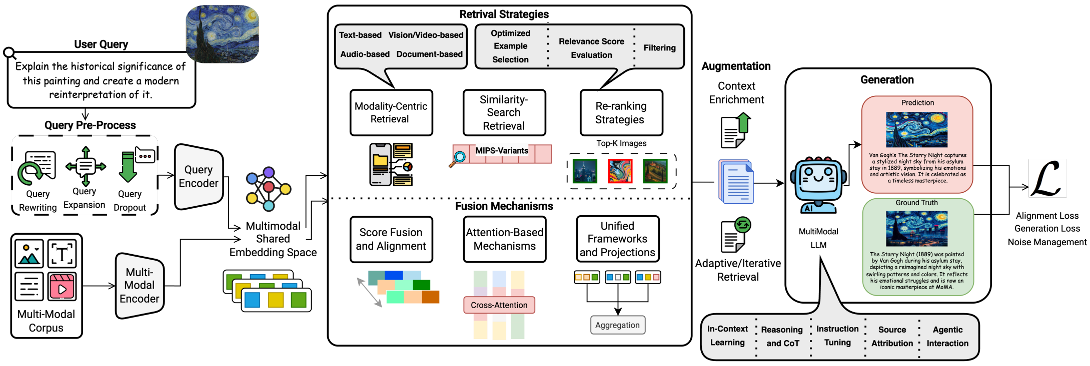

Figure 1: Overview of the multimodal RAG pipeline, illustrating key techniques and recent advancements.

gence, enabling systems to integrate and reason over heterogeneous data sources for more comprehensive representations. A pivotal breakthrough was the introduction of CLIP model (Radford et al., 2021), which aligned visual and textual modalities through contrastive learning and inspired a wave of subsequent models (Alayrac et al., 2024; Wang et al., 2023a; Pramanick et al., 2023). These developments have catalyzed progress across diverse domains, including sentiment analysis (Das and Singh, 2023) and biomedical applications (Hemker et al., 2024), highlighting the effectiveness of multimodal approaches. By facilitating the joint processing of text, images, audio, and video, multimodal learning is increasingly recognized as a critical enabler of progress toward artificial general intelligence (AGI) (Song et al., 2025).

Multimodal RAG The extension of large language models to multimodal LLMs (MLLMs) has significantly broadened their capabilities, enabling reasoning and generation across multiple data modalities (Liu et al., 2023a; Team et al., 2024; Li et al., 2023b). Notably, GPT-4 (OpenAI et al., 2024) demonstrates human-level performance by jointly processing text and images, marking a milestone in multimodal understanding. Building on this progress, multimodal RAG incorporates diverse sources, such as images, audio, and structured data, to enrich contextual grounding and enhance generation quality (Hu et al., 2023; Chen et al., 2022a). This approach improves reasoning by leveraging cross-modal cues, but also introduces challenges, including modality selection, effective fusion, and managing cross-modal relevance (Zhao et al., 2023a). Figure 1 illustrates the general pipeline of these systems.

Multimodal RAG Formulation We present a mathematical formulation of multimodal RAG. These systems aim to generate a multimodal response r given a multimodal query q. Let $D = \{d_1, d_2, \ldots, d_n\}$ denote a multimodal corpus. For clarity, we assume each document $d_i \in D$ is associated with a single modality M_{d_i}. In practice, however, documents often span multiple modalities—for example, a scientific article containing both text and images. Such cases are typically addressed by either decomposing the document into modality-specific sub-documents or employing universal encoders capable of jointly processing multiple modalities.

Each document d_i is encoded using its corresponding modality-specific encoder, yielding $z_i = \text{Enc}_{M_{d_i}}(d_i)$. The collection of all encoded representations is denoted as $Z = \{z_1, z_2, \ldots, z_n\}$. These modality-specific encoders project diverse input modalities into a shared semantic space, enabling cross-modal alignment.

A retrieval model R computes a relevance score $s(e_q, z_i)$ between the encoded query representation e_q (obtained by encoding q using the appropriate encoders) and each document representation z_i. The retrieval-augmented multimodal context X is constructed by selecting documents whose relevance scores exceed a modality-specific threshold:

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

where $\tau_{M_{d_i}}$ is the relevance threshold for the modality M_{d_i}, and s is the scoring function that measures semantic relevance. Finally, the generative model G produces the response conditioned on the original query q and the retrieved context X, formally defined as $r = G(q, X)$.

Related Works Multimodal RAG is a rapidly emerging field, yet a comprehensive survey dedicated to its recent advancements remains lacking. While over ten surveys discuss RAG-related topics such as Agentic RAG (Singh et al., 2025), none specifically focus on the multimodal setting. To our knowledge, the only relevant work (Zhao et al., 2023a) categorizes multimodal RAGs by application and modality. In contrast, our survey adopts a more innovation-driven perspective, offering a detailed taxonomy and addressing recent trends and open challenges. We review over 100 recent papers, primarily from the ACL Anthology, reflecting the growing interest and progress in this domain.

Contributions In this work, **(i)** we present a comprehensive review of multimodal RAG, covering task formulation, datasets, benchmarks, applications, and key innovations across retrieval, fusion, augmentation, generation, training strategies, loss functions, and agent frameworks. **(ii)** We propose a structured taxonomy (Figure 2) that categorizes state-of-the-art models by their core contributions, highlighting methodological advances and emerging trends. **(iii)** We provide open-access resources, including datasets, benchmarks, and implementation details, to facilitate future research. **(iv)** Finally, we identify research gaps and offer insights to guide future directions in this rapidly evolving field.

2 Datasets, Evaluation, and Applications

We review diverse datasets and benchmarks supporting tasks such as multimodal summarization, visual QA, video understanding, and more. For full details, see Appendix (§B) and Tables 1 and 2. Multimodal RAG has been applied across various domains, including healthcare, software engineering, fashion, entertainment, and emerging fields. An overview of tasks and applications are detailed in Appendix (§E) and Figure 3. Evaluating these systems requires multiple metrics, covering retrieval performance, generation quality, and modality alignment. The complete evaluation methods, metrics, and their definitions and formulations are in Appendix (§C).

3 Key Innovations and Methodologies

3.1 Retrieval Strategy

Efficient Search and Similarity Retrieval Modern multimodal RAG systems encode diverse input modalities into a unified embedding space to enable direct cross-modal retrieval. Early CLIP-based (Radford et al., 2021) methods often struggled to balance retrieval precision and computational cost. BLIP-inspired (Li et al., 2022) approaches addressed some of these trade-offs by integrating cross-modal attention during training, yielding richer alignments between visual and textual features. To reconcile high accuracy with efficiency, contrastive retrieval frameworks such as MARVEL (Zhou et al., 2024c) and Uni-IR (Wei et al., 2024a) improved inter-modal discrimination through hard-negative mining and balanced sampling strategies (Zhang et al., 2024i; Lan et al., 2025). Despite these representational gains, direct search over millions of embeddings demands fast similarity computation. Maximum inner product search (MIPS) variants offer sublinear lookup by approximating top-k inner products (Tiwari et al., 2024; Wang et al., 2024c; Zhao et al., 2023b). However, coarse quantization can degrade recall. To mitigate this, adaptive quantization methods (Zhang et al., 2023a; Li et al., 2024a) dynamically allocate bits where the embedding distribution is dense, resulting in recall improvements over static schemes. Hybrid sparse–dense retrieval (Nguyen et al., 2024; Zhang et al., 2024a) further complements dense embeddings with sparse lexical signals. Systems such as MuRAG (Chen et al., 2022a) and RA-CM3 (Yasunaga et al., 2023) employ approximate MIPS for efficient top-k candidate retrieval from large collections of image–text embeddings. Large-scale implementations leverage distributed MIPS techniques, such as TPU-KNN (Chern et al., 2022), for high-speed retrieval. Other efficient similarity computation methods include 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). Emerging approaches explore learned index structures (Zhai et al., 2023; Basnet et al., 2024), which embed the search tree itself in neural parameters, aiming to adapt retrieval paths to the data distribution and reduce both latency and storage overhead.

Modality-Based Retrieval Modality-aware retrieval techniques optimize efficiency by leveraging the unique characteristics of each modality. ***(i) Text-centric retrieval*** 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) dominating text-based evidence retrieval (Chen et al., 2022b; Suri et al., 2024; Nan et al., 2024b). 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 by ensuring accurate citation of text spans. ***(ii) Vision-centric retrieval*** leverages image representations for knowledge extraction (Kumar and Marttinen, 2024; Yuan et al., 2023). Systems such as EchoSight (Yan and

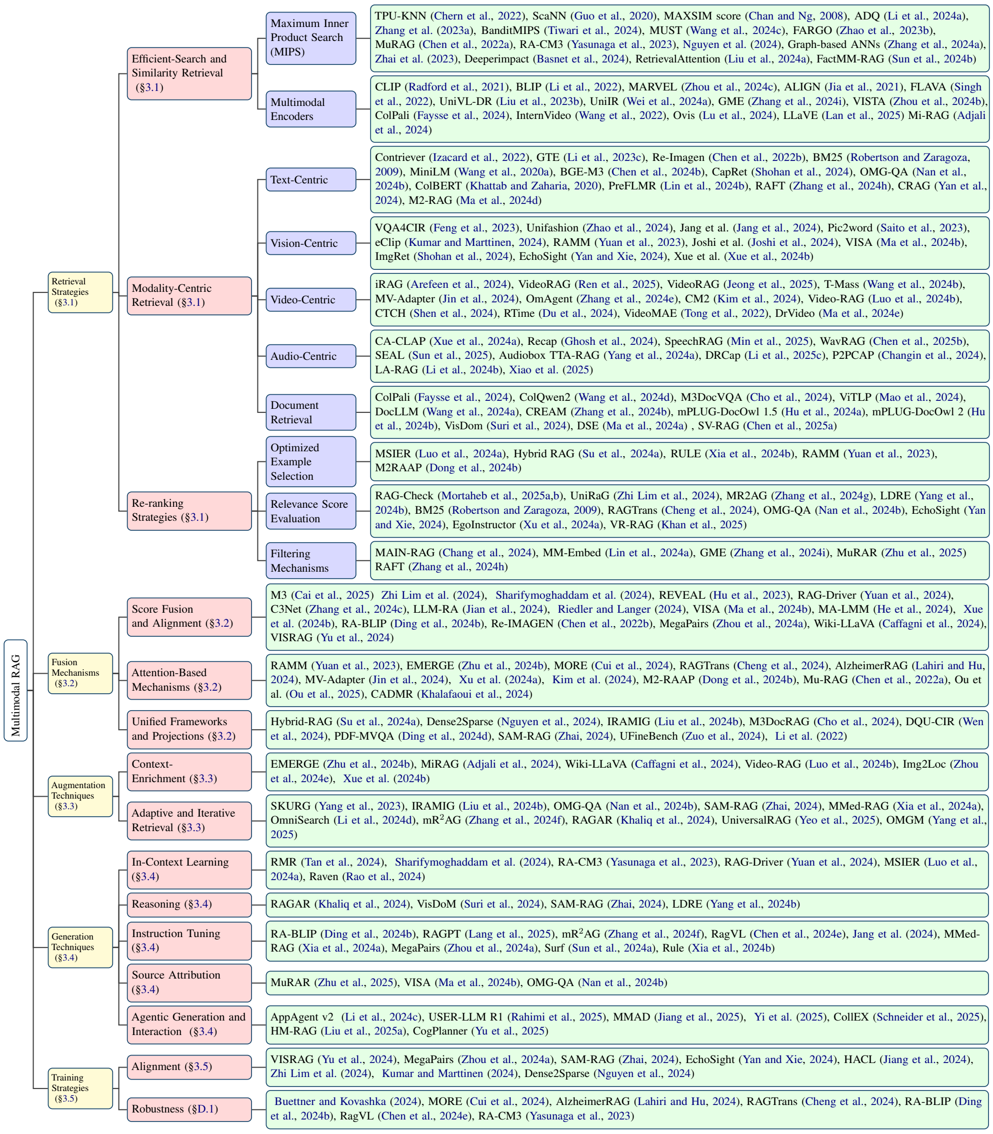

Figure 2: Taxonomy of recent advances in Multimodal RAG. Refer to Appendix (§A) for further details.

Xie, 2024) and ImgRet (Shohan et al., 2024) retrieve visually similar content by using reference images as queries. In addition, composed image retrieval methods (Feng et al., 2023; Zhao et al., 2024; Jang et al., 2024; Saito et al., 2023) integrate multiple image features into unified query representations, enabling zero-shot image retrieval. ***(iii) Video-centric retrieval*** extends vision-based techniques by incorporating temporal dynamics and large video-language models. For instance, iRAG (Arefeen et al., 2024) enables incremental retrieval for sequential video understanding, addressing the need for temporal coherence, while T-Mass (Wang et al., 2024b) uses stochastic text embeddings to improve robustness in text-video alignment. Tackling long-context processing, Video-RAG (Luo et al., 2024b) avoids reliance on proprietary models by using auxiliary OCR/ASR texts, whereas VideoRAG (Ren et al., 2025) employs dual-channel architectures and graph-based knowledge grounding for extreme-length videos. To capture temporal reasoning, CTCH (Shen et al., 2024) applies contrastive transformer hashing for long-term dependencies, which RTime (Du et al., 2024) further refines by introducing reversed-video hard negatives for more robust causality benchmarking. Finally, OmAgent (Zhang et al., 2024e) addresses the challenge of complex video understanding with a divide-and-conquer framework, while DRVideo (Ma et al., 2024e) takes a complementary document-centric approach to enhance narrative preservation. ***(iv) Audio-centric retrieval*** aims to bypass traditional ASR pipelines while improving contextual alignment and real-time processing (Xue et al., 2024a; Ghosh et al., 2024; Min et al., 2025). Pioneering frameworks like WavRAG (Chen et al., 2025b) and

SEAL (Sun et al., 2025) introduce unified embedding architectures, directly mapping raw audio into a shared latent space to enable retrieval from hybrid knowledge bases. Audiobox TTA-RAG (Yang et al., 2024a) conditions text-to-audio synthesis on retrieved acoustic samples, thereby enhancing zero-shot performance by enriching prompts with unlabeled audio context. For audio captioning, DRCap (Li et al., 2025c) bridges the audio-text latent space of CLAP (Wu et al., 2023) via text-only training for domain-adaptable descriptions without paired data. In parallel, P2PCAP (Changin et al., 2024) improves retrieval precision by regenerating captions as dynamic queries. Further innovations address error correction and efficiency. LA-RAG (Li et al., 2024b) utilizes fine-grained speech-to-speech retrieval and forced alignment to enhance ASR accuracy through LLM in-context learning. Meanwhile, hybrid systems, such as Xiao et al. (2025), integrate LLMs to correct errors in noisy environments using retrieved text/audio context.

Document Retrieval and Layout Understanding Recent research has moved beyond traditional unimodal retrieval, developing models that process entire documents by integrating textual, visual, and layout information. 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; Faysse et al., 2024) 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 computational 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, while SV-RAG (Chen et al., 2025a) leverages MLLMs' intrinsic retrieval capabilities via dual LoRA adapters: one for evidence page retrieval and the other for question answering.

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 improve retrieval quality through optimized example selection, refined relevance scoring, and filtering mechanisms. ***(i) Optimized example selection*** techniques often employ multi-step retrieval, integrating both supervised and unsupervised selection approaches (Luo et al., 2024a; Yuan et al., 2023). Supervised methods like Su et al. (2024a) enhance multimodal inputs using probabilistic control keywords, whereas RULE (Xia et al., 2024b) calibrates retrieved context via statistical methods like the Bonferroni correction (Haynes, 2013) to mitigate factuality risks. Clustering-based key-frame selection ensures diversity in video-based retrieval (Dong et al., 2024b). Advanced ***(ii) scoring mechanisms*** are employed by several methods 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 documents. Some frameworks combine similarity scores derived from various modalities for more robust re-ranking. For example, VR-RAG (Khan et al., 2025) proposes a visual re-ranking framework that combines cross-modal text-image similarity with intra-modal visual similarity using DINOv2 (Oquab et al., 2023), demonstrating significant improvements in open-vocabulary recognition tasks. Hierarchical post-processing integrates 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., 2024b) employs semantic ensemble methods to adaptively weigh multiple caption features, while RAGTrans (Cheng et al., 2024) and OMG-QA (Nan et al., 2024b) incorporate traditional ranking functions like BM25 (Robertson and Zaragoza, 2009). ***(iii) Filtering methods*** ensure 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, 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. Dynamic modality filtering methods, such as RAFT (Zhang et al., 2024h) and MAIN-RAG (Chang et al., 2024), train retrievers to disregard confusing data, improving multimodal retrieval robustness.

3.2 Fusion Mechanisms

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., 2022). Wiki-LLaVA (Caffagni et al., 2024), C3Net (Zhang et al., 2024c), Riedler and Langer (2024), and MegaPairs (Zhou et al., 2024a) embed images and queries into a shared CLIP space. In particular, MegaPairs (Zhou et al., 2024a) scales this approach by integrating both CLIP-based and MLLM-based retrieval, fusing their scores to leverage complementary strengths, but at the cost of increased inference complexity. 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., 2022). 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. (2024b) 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, VIS-RAG (Yu et al., 2024) applies position-weighted mean pooling over VLM hidden states, giving higher relevance to later tokens. In contrast, RAG-Driver (Yuan et al., 2024) aligns visual and language embeddings through visual instruction tuning and an MLP projector.

Attention-Based Mechanisms Attention-based methods dynamically modulate cross-modal interactions to enable fine-tuned reasoning across tasks, balancing specificity and interpretability. Cross-attention is frequently used to integrate heterogeneous modalities, as in EMERGE (Zhu et al., 2024b), MORE (Cui et al., 2024), and Alzheimer-RAG (Lahiri and Hu, 2024), though often requiring task-specific attention heads. 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. RAGTrans (Cheng et al., 2024) applies user-aware attention to social media features. MV-Adapter (Jin et al., 2024) introduces Cross-Modality Tying to align video-text embeddings by sharing latent factors, improving robustness but reducing granularity of modality-specific features. 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, then uses a mutual-guided alignment head to filter misaligned features via dot-product similarity and frame-to-token attention; however, this method is computationally intensive. Xu et al. (2024a) condition text generation on visual features using gated cross-attention, optimizing controllability but requiring aligned supervision, 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.

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, enhancing consistency but requiring multiple reasoning passes. M3DocRAG (Cho et al., 2024) flattens multi-page documents into a single embedding tensor, and PDF-MVQA (Ding et al., 2024d) proposes a joint-grained retriever that fuses coarse-grained semantic entity representations with their fine-grained token-level textual content, creating a richer, unified representation. 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 to unimodal text for subsequent processing. UFineBench (Zuo et al., 2024) uses 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., 2022) into sparse lexical vectors using layer normalization and probabilistic expansion control to optimize storage and interpretability.

3.3 Augmentation Techniques

Basic RAG systems typically retrieve content in a single step, directly passing it to generation, often leading to inefficiencies and suboptimal outputs. Augmentation techniques refine retrieved data beforehand, improving multimodal interpretation, structuring, and integration (Gao et al., 2023).

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., 2024b). 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) boosts accuracy by including both similar and dissimilar points in prompts, helping rule out implausible locations.

Adaptive and Iterative Retrieval For more complex queries, dynamic retrieval mechanisms have proven effective. Adaptive retrieval approaches optimize relevance by adjusting retrieval dynamically. For instance, UniversalRAG (Yeo et al., 2025) introduces a framework that adapts retrieval by dynamically routing queries to the most suitable corpus based on both the required modality and granularity (e.g., paragraph vs. document, clip vs. full video), thereby addressing the specific knowledge type and scope demanded by the query. 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., 2024d) decomposes multimodal queries into structured sub-questions, planning retrieval actions in real time. Iterative approaches refine results over multiple steps by incorporating feedback from prior iterations. For example, OMGM (Yang et al., 2025) orchestrates a multi-step, coarse-to-fine retrieval process for knowledge-based visual question answering, starting with a broad entity search and progressively refining the selection through multimodal reranking and fine-grained textual filtering to pinpoint the most relevant knowledge, achieving superior retrieval performance in comparison to prior methods. IRAMIG (Liu et al., 2024b) improves multimodal retrieval by dynamically updating queries based on retrieved content. OMG-QA (Nan et al., 2024b) integrates episodic memory to refine retrieval across multiple rounds, ensuring continuity in reasoning. RAGAR (Khaliq et al., 2024) further enhances contextual consistency by iteratively adjusting retrieval based on prior responses and multimodal analysis.

3.4 Generation Techniques

In-Context Learning (ICL) 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. MSIER (Luo et al., 2024a) improves example selection with a multimodal supervised in-context examples retrieval framework, using an MLLM scorer to assess textual and visual relevance. Raven (Rao et al., 2024) introduces Fusion-in-Context Learning, integrating diverse in-context examples for superior performance over standard ICL.

Reasoning Reasoning methods, like chain of thought (CoT), decompose complex reasoning into sequential steps, improving coherence and robustness in multimodal RAG. RAGAR (Khaliq et al., 2024) refines fact-checking queries and explores branching reasoning paths by introducing Chain of RAG and Tree of RAG, while VisDoM (Suri et al., 2024) and SAM-RAG (Zhai, 2024) integrate CoT with evidence curation and multi-stage verification to enhance accuracy and support. Notably, VisDoM performs well in scenarios where key information is distributed across modalities. LDRE (Yang et al., 2024b) applies LLMs for divergent compositional reasoning by refining captions using dense descriptions and textual modifications, achieving superior zero-shot composed image retrieval results.

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) and RagVL (Chen et al., 2024e) train MLLMs to invoke retrieval adaptively, identify relevant evidence, and enhance ranking capabilities for improved response accuracy. 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 a medical large vision language model through direct preference optimization to mitigate overreliance on retrieved contexts.

Source Attribution and Evidence Transparency Ensuring source attribution in multimodal RAG systems is a significant research focus. OMG-QA (Nan et al., 2024b) prompts LLMs for explicit evidence citation in generated responses. MuRAR (Zhu et al., 2025) refines an LLM's initial response by integrating multimodal information from a source-based retriever to improve informativeness. However, its recall is constrained, as the retriever may miss evidence spanning different sections or web documents. Similarly, VISA (Ma et al., 2024b) employs vision-language models to generate answers with visual source attribution by highlighting evidence in retrieved screenshots. Nevertheless, its attribution accuracy degrades when evidence spans multiple sections or requires cross-modal integration.

Agentic Generation and Interaction Agent-driven multimodal RAG uses versatile autonomous/semi-autonomous systems across diverse interaction paradigms and specialized domains, often generating complex outputs. For user interaction, AppAgent v2 (Li et al., 2024c) enables mobile GUI navigation while USER-LLM R1 (Rahimi et al., 2025) creates personalized conversational agents via dynamic profiling, particularly for elderly users. In specialized applications, MMAD (Jiang et al., 2025) addresses industrial anomaly detection with training-free enhancement strategies, Yi et al. (2025) improve clinical report generation while reducing hallucination, and CollEX (Schneider et al., 2025) facilitates scientific collection exploration for researchers and learners. For complex reasoning, HM-RAG (Liu et al., 2025a) coordinates hierarchical multi-agent collaboration across multimodal data streams, while CogPlanner (Yu et al., 2025) introduces a cognitively inspired planning framework that iteratively refines queries and selects retrieval strategies adaptively.

3.5 Training Strategies

Training multimodal RAG models follows a multi-stage process to effectively capture cross-modal interactions (Chen et al., 2022a). Pretraining on large paired datasets establishes cross-modal relationships, while fine-tuning adapts models to task-specific objectives by aligning outputs with task requirements (Ye et al., 2019). For example, REVEAL (Hu et al., 2023) integrates multiple training objectives. Its pretraining phase optimizes Prefix Language Modeling Loss (L_{PrefixLM}), where text is predicted from a given prefix and an associated image. Supporting losses include Contrastive Loss (L_{contra}) which aligns queries with pseudo-ground-truth knowledge, Disentangled Regularization Loss (L_{decor}) to enhance embedding expressiveness, and Alignment Regularization Loss (L_{align}) to refine query-knowledge alignment. Fine-tuning employs a cross-entropy objective for downstream tasks like visual question answering or image captioning. Details on robustness advancements and loss formulations are in Appendix (§D).

Alignment Contrastive learning improves representation quality by pulling positive pairs closer and pushing negative pairs apart in the embedding space. The InfoNCE loss (van den Oord et al., 2019) is widely employed in multimodal RAG models, including VISRAG (Yu et al., 2024), MegaPairs (Zhou et al., 2024a), and SAM-RAG (Zhai, 2024), to improve retrieval-augmented generation. Several models introduce refinements to contrastive training. EchoSight (Yan and Xie, 2024) enhances retrieval accuracy by selecting visually similar yet semantically distinct negatives, while HACL (Jiang et al., 2024) mitigates hallucinations by incorporating adversarial captions as distractors. Similarly, UniRaG (Zhi Lim et al., 2024) improves retrieval robustness by leveraging hard negative documents to help the model discriminate between relevant and irrelevant contexts. The eCLIP loss (Kumar and Marttinen, 2024) extends contrastive learning by integrating expert-annotated data and an auxiliary MSE loss to refine embedding quality. Mixup strategies further improve generalization by generating synthetic positive pairs (Kumar and Marttinen, 2024). Dense2Sparse (Nguyen et al., 2024) employs image-to-caption $\ell(I \rightarrow C)$ and caption-to-image $\ell(C \rightarrow I)$ losses, while enforcing sparsity through $\ell 1$ regularization, optimizing retrieval precision by balancing dense and sparse representations.

4 Open Problems and Future Directions

Additional challenges and future directions about long-context processing, scalability, efficiency, and personalization are discussed in Appendix (§F).

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 often attribute responses to broad sources, citing entire documents or large visual regions instead of pinpointing exact contributing elements across modalities (Ma et al., 2024b; Hu et al., 2023). Moreover, the interplay between modalities affects the outcome quality; 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 degrades when relying on low-quality or outdated sources (Chen et al., 2022b). MM-PoisonRAG (Ha et al., 2025) and Poisoned-MRAG (Liu et al., 2025b) demonstrate that even a few adversarial knowledge injections can hijack cross-modal retrieval and derail generation, underscoring the imperative for robust defense mechanisms against knowledge poisoning in multimodal RAG systems. While the trustworthiness of unimodal RAGs has been studied (Zhou et al., 2024d), ensuring robustness in multimodal RAGs remains an open challenge and a crucial research direction.

Reasoning, Alignment, and Retrieval Enhancement Multimodal RAGs struggle with compositional reasoning, requiring logical integration of information across modalities for coherent, context-rich outputs. While cross-modal techniques like Multimodal-CoT (Zhang et al., 2023b) have emerged, further advancements are needed to enhance coherence and contextual relevance. Improving modality alignment and entity-aware retrieval is crucial. Moreover, despite the potential of knowledge graphs to enrich cross-modal reasoning, they remain 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), redundancy (Nan et al., 2024b), and biases from training data or retrieved content (Zhai, 2024), pose significant challenges. A promising direction is a unified embedding space for all modalities, enabling direct multimodal search without intermediary models (e.g., ASRs). Despite progress, mapping multimodal knowledge into a unified space remains an open challenge with substantial potential.

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 RAGs, future systems should incorporate interactive feedback and self-guided decision-making to iteratively refine outputs. Existing feedback mechanisms often fail to 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 remains largely overlooked but holds significant potential for assessing whether retrieval is necessary, evaluating the relevance of retrieved content, and dynamically determining the most suitable modalities for response generation. Robust support for any-to-any modality is crucial for 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 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.

5 Conclusion

This study provides a comprehensive review of multimodal RAG, categorizing key advancements in retrieval, multimodal fusion, augmentation, generation, training strategies, and agents. We also examine task-specific applications, datasets, benchmarks, and evaluation methods while highlighting open challenges and 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.

6 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.

7 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.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

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*.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. 2024. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828.

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.

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*.

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.

Choi Changin, Lim Sungjun, and Rhee Wonjong. 2024. Audio captioning rag via generative pair-to-pair retrieval with refined knowledge base. *Preprint*, arXiv:2410.10913.

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.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, and Tong Sun. 2025a. Sv-rag: Lora-contextualizing adaptation of mllms for long document understanding. *Preprint*, arXiv:2411.01106.

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.

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

Ran Chen, Xueqi Yao, and Xuhui Jiang. 2024d. 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.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025b. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. *Preprint*, arXiv:2502.14727.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024e. 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.

Kyoyun Choi, Byungmu Yoon, Soobum Kim, and Jonggwon Park. 2025. Leveraging llms for multimodal retrieval-augmented radiology report generation via key phrase extraction. *arXiv preprint arXiv:2504.07415*.

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.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *Int. J. Comput. Vision*, 130(1):33–55.

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).

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. Mmvqa: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI*, pages 3–9.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024d. 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.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

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.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha. 2024. Recap: Retrieval-augmented audio captioning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1161–1165.

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.

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, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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.

Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. 2025. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. *Preprint*, arXiv:2502.17832.

Winston Haynes. 2013. *Bonferroni Correction*, pages 154–154. Springer New York, New York, NY.

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.

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.

Liting Huang, Zhihao Zhang, Yiran Zhang, Xiyue Zhou, and Shoujin Wang. 2025. Ru-ai: A large multimodal dataset for machine-generated content detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 733–736.

Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*.

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.

Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. 2020. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. *Preprint*, arXiv:2004.12276.

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.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. 2025. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. *Preprint*, arXiv:2410.09453.

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 Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

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.

Faizan Farooq Khan, Jun Chen, Youssef Mohamed, Chun-Mei Feng, and Mohamed Elhoseiny. 2025. Vr-rag: Open-vocabulary species recognition with rag-assisted large multi-modal models. *arXiv preprint arXiv:2505.05635*.

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.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

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)*.

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.

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

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.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'22, New York, NY, USA. Association for Computing Machinery.

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.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

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

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. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. 2025b. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. In *ICLR*.

Shaojun Li, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Zongyao Li, Xianghui He, Min Zhang, and Hao Yang. 2024b. La-rag:enhancing llm-based asr accuracy with retrieval-augmented generation. *Preprint*, arXiv:2409.08597.

Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen. 2025c. Drcap: Decoding clap latents with retrieval-augmented generation for zero-shot audio captioning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning and retrieval.

Yanda Li, Chi Zhang, Wanqi Yang, Bin Fu, Pei Cheng, Xin Chen, Ling Chen, and Yunchao Wei. 2024c. Appagent v2: Advanced agent for flexible mobile interactions. *Preprint*, arXiv:2408.11824.

Yangning Li, Yinghui Li, Xingyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, et al. 2024d. 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.

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

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.

Yinuo Liu, Zenghui Yuan, Guiyao Tie, Jiawen Shi, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025b. Poisoned-mrag: Knowledge poisoning attacks to multimodal retrieval augmented generation. *arXiv preprint arXiv:2503.06254*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025c. 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*.

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

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)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024c. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024d. 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. 2024e. 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.

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

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

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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)*.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. 2024a. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*.

Linyong Nan, Weining Fang, Aylin Rasteh, Pouya Lahabi, Weijin Zou, Yilun Zhao, and Arman Cohan. 2024b. OMG-QA: Building open-domain multi-modal 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 multi-modal 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.

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.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision.

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.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362, New York, NY, USA. Association for Computing Machinery.

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.

Abhirama Subramanyam Penamakuri, Manish Gupta, Mithun Das Gupta, and Anand Mishra. 2023. Answer mining from a pool of images: Towards retrieval-based visual question answering. In *IJCAI*. ijcai.org.

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.

Hamed Rahimi, Jeanne Cattoni, Meriem Beghili, Mouad Abrini, Mahdi Khoramshahi, Maribel Pino, and Mohamed Chetouani. 2025. Reasoning llms for user-aware multimodal conversational agents. *Preprint*, arXiv:2504.01700.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition. *Preprint*, arXiv:2301.02560.

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, Nihar 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.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*.

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.

Fulvio Sanguigni, Davide Morelli, Marcella Cornia, and Rita Cucchiara. 2025. Fashion-rag: Multimodal fashion image editing via retrieval-augmented generation. *arXiv preprint arXiv:2504.14011*.

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

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. 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*.

Chunyu Sun, Bingyu Liu, Zhichao Cui, Anbin Qi, Tian hao Zhang, Dinghao Zhou, and Lewei Lu. 2025. Seal: Speech embedding alignment learning for speech large language model with retrieval-augmented generation. *Preprint*, arXiv:2502.02603.

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*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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. 2023a. 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, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024b. 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.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*.

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.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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.

Cihan Xiao, Zejiang Hou, Daniel Garcia-Romero, and Kyu J Han. 2025. Contextual asr with retrieval augmented large language model. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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.

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.

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

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. 2024a. Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining. *Preprint*, arXiv:2406.03714.

Junxiao Xue, Quan Deng, Fei Yu, Yanhao Wang, Jun Wang, and Yuehua Li. 2024b. 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.

Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. 2024a. Audiobox tta-rag: Improving zero-shot and few-shot text-to-audio with retrieval-augmented generation. *Preprint*, arXiv:2411.05141.

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.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. 2025. Omgm: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. *arXiv preprint arXiv:2505.07879*.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024b. 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.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. 2025. Universalrag: Retrieval-augmented generation over multiple corpora with diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*.

Ziruo Yi, Ting Xiao, and Mark V. Albert. 2025. A multimodal multi-agent framework for radiology report generation. *Preprint*, arXiv:2505.09787.

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.

Xiaohan Yu, Zhihan Yang, and Chong Chen. 2025. Unveiling the potential of multimodal retrieval augmented generation with planning. *Preprint*, arXiv:2501.15470.

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 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 multimodal encoders) and modality-centric techniques that distinguish between text-, vision-, audio-, and video-centric as well as document retrieval models. Re-ranking strategies further refine these methods via optimized example selection, relevance scoring, and filtering.

- **Fusion mechanisms** cover score fusion and alignment techniques, including CLIP score fusion and prototype-based embeddings that unify multimodal representations, attention-based methods such as cross-attention and co-attention transformers that dynamically weight cross-modal interactions, and unified frameworks and projections like hierarchical fusion and dense-to-sparse projections that consolidate multimodal inputs.

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

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

- **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

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 (QA) 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 should be noted 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., 2024d) 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. In addition, (Liu et al., 2025d) introduce another specialized benchmark for multimodal RAG that evaluates performance across image captioning, multi-modal question answering, fact verification, and image reranking in an open-domain retrieval setting. 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., 2024d), MMBench (Liu et al., 2025c), and ScienceQA (Lu 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 like 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.

The choice of dataset significantly influences the evaluation focus, ranging from foundational pretraining on large-scale image-text corpora like LAION-5B (Schuhmann et al., 2022) (5.85 billion pairs) or MINT-1T (Awadalla et al., 2024) (3.4 billion images with 1 trillion text tokens), to more specialized tasks such as video understanding with HowTo100M (Miech et al., 2019) (136 million video clips) or medical report generation using MIMIC-CXR (Johnson et al., 2019) (125,417 image-report pairs).

Datasets are often tailored for specific downstream tasks. For visual question answering, VQA (Antol et al., 2015) and A-OKVQA (Schwenk et al., 2022) specifically require external knowledge, making them suitable for evaluating RAG systems' ability to retrieve and reason over such knowledge. For document understanding, datasets such as DocVQA (Mathew et al., 2021) and M3DocVQA (Cho et al., 2024) are essential. As discussed in the benchmarks overview above, unified evaluation frameworks like M^2RAG (Ma et al., 2024d) provide a comprehensive assessment across multiple tasks, including image captioning, visual question answering, and fact verification.

Evaluating complex reasoning capabilities in multimodal RAG systems has become increasingly important. Datasets such as MultimodalQA (Talmor et al., 2021), WebQA (Chang et al., 2022), and ScienceQA (Lu et al., 2022) are specifically designed to benchmark multi-hop reasoning abilities crucial for advanced RAG systems, with Dyn-VQA (Li et al., 2024d) additionally focusing on robustness to changing information.

Comparative Analysis of Datasets Understanding the strategic trade-offs in dataset design is crucial for multimodal RAG development, as different dataset characteristics serve distinct purposes across the model development pipeline.

(i) Scale and Diversity vs. Curation: Large-scale datasets such as LAION-5B (Schuhmann et al., 2022) and Conceptual Captions (Sharma et al., 2018)

provide substantial scale essential for pre-training, enabling models to learn generalizable representations across diverse domains. However, their reliance on web-crawled data introduces inherent noise that can compromise training quality. Conversely, smaller, meticulously curated datasets like Flickr30K (Young et al., 2014) (31,000 images with human annotations) and domain-specific collections such as Fashionpedia (Jia et al., 2020) (48,000 images with segmentation masks) prioritize annotation quality over scale, making them essential for fine-tuning models and assessing specialized performance.
(ii) Modality Focus and Combination: While many systems aggregate unimodal datasets to construct multimodal contexts, datasets explicitly designed for multimodal tasks demonstrate superior alignment between modalities. Foundational datasets like MS-COCO (Lin et al., 2014) and VQA (Antol et al., 2015) establish benchmarks for image-text understanding, while specialized collections such as AudioSet (Gemmeke et al., 2017) (2 million audio clips) and AudioCaps (Kim et al., 2019) (46,000 audio clips with captions) address audio-language integration. Emerging modalities like 3D (e.g., ShapeNet (Chang et al., 2015)) remain underrepresented, yet are essential for expanding RAG applications into spatial reasoning domains.
Table 1 and Table 2 present a comprehensive overview of datasets and benchmarks commonly 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.

Limitations of Existing Datasets and Benchmarks
While the datasets and benchmarks discussed above have significantly advanced multimodal RAG research, several limitations persist that offer important avenues for future work:
(i) Bias and Fairness: Large datasets, especially those scraped from the web, can inherit societal biases related to gender, race, or culture. This can lead to skewed model behavior and unfair outcomes. Efforts to create more balanced datasets are crucial, but comprehensive bias auditing across modalities remains a challenge.
(ii) Annotation Quality and Noise: The trade-off between dataset scale and annotation quality remains a persistent challenge. While large datasets facilitate broad learning, their often noisy or weakly supervised labels (e.g., alt-text for images) can hinder precise model training. As demonstrated by OHRBench (Zhang et al., 2024d), OCR errors exemplify how noise in one modality can cascade and affect overall RAG system performance.
(iii) Coverage and Generalization Gaps: Many datasets are domain-specific, which can limit the generalization of models to out-of-domain scenarios. There is a need for more datasets covering a wider array of real-world contexts and less common modalities.
(iv) Real-World Complexity and Long-Context Understanding: Current datasets inadequately capture real-world multimodal information complexity. Challenges include efficient sampling of relevant video frames, handling multi-page documents with numerous images, and processing dynamic information environments; benchmarks like Dyn-VQA

(Li et al., 2024d) are, however, beginning to address this latter challenge.

(v) Lack of Adversarial and Robustness Testing: While benchmarks like Counterfactual VQA (Niu et al., 2021) specifically test robustness against certain perturbations, there is a general scarcity of datasets containing multimodal adversarial examples or structured negative instances. Such datasets are vital for developing more robust and reliable RAG systems that can handle out-of-distribution inputs or misleading information.

(vi) Retrieval-Generation Integration: Many benchmarks evaluate retrieval and generation components separately rather than assessing their synergistic interplay. More holistic evaluation frameworks are needed that jointly measure retrieval accuracy, relevance of retrieved multimodal context, and final output quality, as aimed by benchmarks like MRAG-Bench (Hu et al., 2024c) for visual integration and RAG-Check (Mortaheb et al., 2025a) for retrieval relevance.

(vii) Limited Support for "Any-to-Any" Modalities: While current research primarily focuses on text, image, video, and audio, future RAG systems are envisioned to support any-to-any modality interactions. Existing datasets offer limited support for such comprehensive multimodality.

C Evaluation and Metrics

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. 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. 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:

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

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). MRR measures the rank position of the first relevant result in the returned list. The formula for calculating MRR is:

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q} \tag{2}$$

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

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). 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:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}(\text{gram}_N)} \tag{3}$$

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

$$\text{ROUGE-L} = \frac{LCS(X, Y)}{|Y|} \tag{4}$$

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

$$\text{BLEU}(p_n, \text{BP}) = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5}$$

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:

$$\text{BP} = \begin{cases} 1 & \text{length} > rl \\ \exp\left(1 - \frac{rl}{cl}\right) & \text{length} \leq rl \end{cases} \tag{6}$$

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

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

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., 2024d).

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). The formula for calculating CLIPScore is:

$$\text{CLIPScore} = \frac{\mathbf{t}.\mathbf{i}}{||\mathbf{t}||\,||\mathbf{i}||} \tag{7}$$

where t and i are text and image embedding, respectively.

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. The formula for FID is:

$$\text{FID} = ||\mu_r - \mu_g||^2 + tr(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \tag{8}$$

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 (Szegedy et al., 2016) is typically used.

Inception Score (IS) evaluates image diversity and quality through classification probabilities (Zhi Lim et al., 2024). For audio evaluation, Zhang et al. (2024c) use 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).

D Robustness Advancements and Loss Functions

D.1 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. RAGTrans (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., 2024e) 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.

D.2 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:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (9)$$

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:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (10)$$

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:

$$\mathcal{L}_G = \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (11)$$

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:

$$\mathcal{L} = \sum_{i=1}^{N} \max(0, \|f(x_a^i) - f(x_p^i)\|^2 - \|f(x_a^i) - f(x_n^i)\|^2 + \alpha) \quad (12)$$

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.

E Applications and Relevant Tasks

Multimodal RAG extends traditional RAG beyond unimodal settings to cross-modal tasks. In content generation, it enhances 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. It also improves coherence in visual storytelling and factual alignment in multimodal summarization (Tonmoy et al., 2024). In knowledge-intensive applications, multimodal RAG supports open-domain QA (Chen et al., 2024e; Ding et al., 2024b; Yuan et al., 2023), video-based QA (Luo et al., 2024b), fact verification (Khaliq et al., 2024), and zero-shot image–text retrieval (Yang et al., 2024b), grounding responses in retrieved knowledge and thereby mitigating hallucinations.
Additionally, the incorporation of chain-of-thought reasoning (Zhai, 2024; Khaliq et al., 2024) further enhances complex problem-solving and inference. Finally, their integration into AI assistants such as Gemini (Team et al., 2024) enables natural language-driven visual search, document understanding, and multimodal reasoning.
Multimodal RAGs are increasingly applied across diverse domains, including healthcare, software engineering, and creative industries (e.g., fashion and design automation). 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. In automated report generation, RULE (Xia et al., 2024b) mitigates hallucinations through dynamic retrieval of clinically similar cases. Similarly, RA-RRG (Choi et al., 2025) first leverages an LLM to extract key textual phrases from a report corpus, then employs a multimodal retriever to link the visual features to these relevant phrases. The coherent report is generated after being retrieved by another LLM without fine-tuning, thereby reducing hallucinations. FactMM-RAG (Sun et al., 2024b) further automates radiology report drafting by retrieving biomarker correlations from medical ontologies, exemplifying the paradigm's capacity to operationalize expert knowledge at scale. 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 Su et al. (2024a) advance privacy-preserving architectures for federated clinical data integration.

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. For fashion image editing, Fashion-RAG (Sanguigni et al., 2025) employs a retrieval-augmented approach, retrieving garments that match textual descriptions and integrating their attributes into image generation via textual inversion techniques within diffusion models, ensuring personalized and contextually relevant outputs. LLM4DESIGN (Chen et al., 2024d) 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.

F Additional Future Directions

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, like user-specific retrieval (e.g., adapting to medical history), remain underexplored. As these systems evolve, ensuring privacy and preventing sensitive data leakage in multimodal outputs is critical. Lastly, the lack of datasets with complex reasoning tasks and multimodal adversarial examples limits robust evaluation.

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)	400M image–text pairs; used for pre-training multimodal models.	Image, Text	LAION-400M
	Conceptual-Captions (CC) (Sharma et al., 2018)	More than 3M 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)	5.85B image–text pairs used as external memory for retrieval.	Image, Text	LAION-5B
	COCO-CN (Li et al., 2019)	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
	MINT-1T (Awadalla et al., 2024)	1T text tokens and 3.4B images; 10x larger than existing open-source datasets.	Image, Text	MINT-1T
	ShareGPT4V (Chen et al., 2024c)	1.2M images with GPT-4-generated captions, including spatial and factual details.	Image, Text	ShareGPT4V
	OmniCorpus (Li et al., 2025b)	8.6B images and 1.7T tokens across 2.2B web documents; interleaved text-image layout.	Image, Text	OmniCorpus
Video-Text	BDD-X (Kim 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
	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
	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
	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 (Damen et al., 2022)	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
	WebVid (Bain et al., 2021)	10M video–text pairs (refined to WebVid-Refined-1M).	Video, Text	WebVid
	InternVid (Wang et al., 2023b)	7M YouTube videos (760K hours), 234M clips, 4.1B words; used for video-text pretraining and representation learning.	Video, Text	InternVid
	OpenVid-1M (Nan et al., 2024a)	1 million video-text pairs for multimodal learning.	Video, Text	OpenVid-1M
	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
	AudioCaps (Kim et al., 2019)	46K audio clips paired with human-written text captions.	Audio, Text	AudioCaps
	MusicCaps (Agostinelli et al., 2023)	It is composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.	Audio, Text	MusicCaps
	Clotho (Drossos et al., 2020)	Audio captioning dataset with diverse soundscapes.	Audio, Text	Clotho
	WavCaps (Mei et al., 2024)	Large-scale weakly-labeled audio-text dataset, comprising approximately 400k audio clips with paired captions.	Audio, Text	WavCaps
	Spoken SQuAD (Li et al., 2018)	Audio version of the SQuAD dataset for spoken question answering, focusing on the listening comprehension task.	Audio, Text	Spoken SQuAD
	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
	Quilt-1M (Ikezogwo et al., 2023)	438K medical images with 768K text pairs; includes microscopic images and UMLS entities.	Image, Text	Quilt-1M
Fashion	Fashion-IQ (Wu et al., 2021)	77,684 images across three categories; evaluated with Recall@10 and Recall@50.	Image, Text	Fashion IQ
	FashionGen (Rostamzadeh 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 (Jia et al., 2020)	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)	Covering 55 common object categories with 51,300 unique 3D 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
	MultimodalQA (Talmor et al., 2021)	29,918 questions requiring multi-modal multi-hop reasoning over text, tables, and images.	Image, Text, Table	MultimodalQA
	ViQuAE (Lerner 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 (Chang et al., 2022)	46K queries that require reasoning across text and images.	Text, Image	WebQA
	Infoseek (Chen et al., 2023)	Fine-grained visual knowledge retrieval using a Wikipedia–based knowledge base (6M passages).	Image, Text	Infoseek
	ClueWeb22 (Overwijk 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
	DocVQA (Mathew et al., 2021)	12,767 diverse document images with 50K QA pairs, categorized by reasoning type to evaluate DocVQA methods.	Image, Text	DocVQA
	ChartQA (Masry et al., 2022)	9.6K human-written QA pairs + 23.1K generated from chart summaries.	Image, Text	ChartQA
	DVQA (Kafle et al., 2018)	3.5M QA pairs on 300K diagrams, evaluating structure, data retrieval, and reasoning.	Image, Text	DVQA
	RETVQA (Penamakuri et al., 2023)	416,000 QA samples where retrieval from a large image set is needed to answer questions; emphasizes RAG pipeline.	Image, Text	RETVQA
	SEED-Bench (Li et al., 2023a)	19K multiple–choice questions with accurate human annotations across 12 evaluation dimensions.	Text	SEED-Bench
	M3DocVQA (Cho et al., 2024)	2,441 multi-hop questions across 3,368 PDF documents; evaluates open-domain DocVQA.	Image, Text	M3DocVQA
	MMLongBench-Doc (Ma et al., 2024c)	135 lengthy PDFs with 1,091 questions; focuses on multi-hop reasoning in single documents.	Image, Text	MMLongBench-Doc
Other	GeoDE (Ramaswamy et al., 2023)	61,940 images from 40 classes across 6 world regions; emphasizes geographic diversity in object recognition.	Image	GeoDE
	RU-AI (Huang et al., 2025)	1.47M samples of real vs AI-generated content for fake detection robustness.	Image, Text, Audio	RU-AI
	MIMIC-IT (Li et al., 2025a)	2.8M multimodal instruction-response pairs for model alignment.	Image, Video, Text	MIMIC-IT
	MMVQA (Ding et al., 2024c)	262K question-answer pairs across 3,146 multipage research PDFs for robust multimodal information retrieval.	Image, Text	MMVQA

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
	M^2RAG (Ma et al., 2024d)	Benchmarks multimodal RAG; evaluates retrieval, multi-hop reasoning, and integration.	Images + Text	M^2RAG
	Dyn-VQA (Li et al., 2024d)	Focuses on dynamic retrieval, multi-hop reasoning, and robustness to changing information.	Images + Text	Dyn-VQA
	MMBench (Liu et al., 2025c)	Covers VQA, captioning, retrieval; evaluates cross-modal understanding across vision, text, and audio.	Images + Text + Audio	MMBench
	ScienceQA (Lu 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