Retrieve in Any Modality: A Survey on Modality Retrieval-Augmented Generation for Large Language Models

Maojun SUN
Hong Kong Polytechnic University
mj.sun@connect.polyu.hk

Abstract

Large Language Models (LLMs) often suffer from outdated knowledge and factual hallucinations due to their reliance on static training data. Traditional Retrieval-Augmented Generation (RAG) addresses these issues by incorporating external dynamic knowledge during inference, enhancing factual accuracy and up-to-date grounding. With the advancement of multimodal models, Multimodal RAG enabling retrieval and generation across diverse modalities such as text, images, audio, and video. This survey systematically reviews the methodologies, datasets and evaluation metrics of Multimodal RAG. We further discuss some open research challenges, providing insights into future directions for developing more reliable and capable multimodal generation systems.

1 Introduction

The rapid advancement of Large Language Models (LLMs) stems from breakthroughs in transformer architectures (Vaswani et al., 2017), improved computational resources, and the availability of massive-scale training corpora (Naveed et al., 2024). These factors have enabled LLMs to excel across diverse natural language processing (NLP) tasks, including instruction following (Qin et al., 2024), complex reasoning (Wei et al., 2024b), in-context learning (Brown et al., 2020), and multilingual translation (Zhu et al., 2024a). However, despite these achievements, LLMs still face inherent weaknesses, particularly in terms of factual accuracy, outdated knowledge, and unverifiable reasoning paths (Huang et al., 2024; Xu et al., 2024). This is largely due to their reliance on parametric memory, which limits their ability to access evolving external knowledge directly. Retrieval-Augmented Generation (RAG) offers a practical solution to these issues by combining retrieval from external knowledge sources with generative capabilities (Lewis et al., 2020). In a typical RAG pipeline, a retriever identifies relevant content using embedding-based similarity search (Chen et al., 2024; Rau et al., 2024), followed by re-ranking to refine relevance (Dong et al., 2024a). The retrieved knowledge is then fed into the generator as additional context, enabling the model to produce more accurate and verifiable responses. While this approach improves factual grounding, conventional RAG systems are almost exclusively designed for text retrieval, limiting their applicability in scenarios requiring multimodal information.

Multimodal RAG The evolution toward Multimodal RAG integrates these two trends—retrieval-augmented generation and multimodal learning—enabling retrieval and generation across modalities such as text, images, and audio (Liu et al., 2023; Team et al., 2024; Li et al., 2023). Systems like GPT-4, which can process both images and text (OpenAI et al., 2024), exemplify this direction. In multimodal RAG, external knowledge is no longer limited to textual documents but can include visual content, speech transcripts, or even video clips, enriching the context provided to the generator (Hu et al., 2023; Chen et al., 2022a). This unlocks new capabilities but also introduces unique challenges such as modality selection, cross-modal fusion, and retrieval precision across heterogeneous data types (Zhao et al., 2023).

The objective of Multimodal RAG is generating a multimodal response, denoted as $\mathbf{y}$, given a multimodal query $\mathbf{x}$.

Let $\mathcal{C} = \{\sqcap_1, \sqcap_2, \ldots, \sqcap_N\}$ denote a multimodal knowledge corpus, where each knowledge element $\sqcap_j$ is associated with a specific modality $\mu(\sqcap_j)$. Each knowledge element is processed using an encoder tailored for its modality, yielding:

$$\mathbf{v}_j = \mathcal{E}_{\mu(\sqcap_j)}(\sqcap_j) \tag{1}$$

The complete set of encoded representations is:

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\} \tag{2}$$

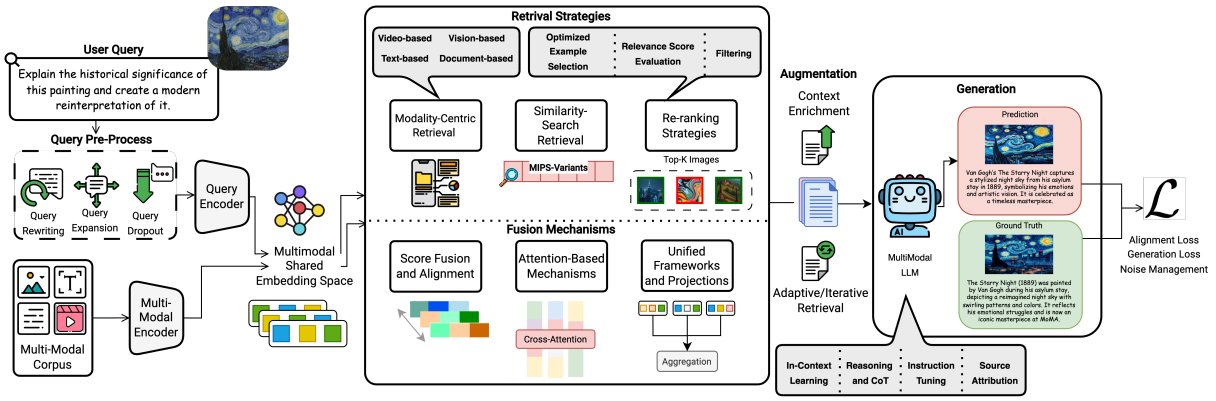

Figure 1: This figure illustrates the overall process of multimodal RAG, showing how queries are processed, relevant multimodal data is retrieved and fused, and external knowledge is incorporated to guide generation. Each stage applies tailored techniques, from retrieval refinement to fusion and generation optimization, enabling the system to handle diverse modalities effectively (Abootorabi et al., 2025).

These modality-specific encoders project diverse content into a shared representation space to enable cross-modal relevance estimation. A retriever module $\mathcal{R}$ computes the relevance score between the encoded query representation $\mathbf{x}_{\text{enc}}$ and each candidate knowledge unit $\mathbf{v}_j$ using a scoring function ϕ, as follows:

$$\phi(\mathbf{x}_{\text{enc}}, \mathbf{v}_j) \tag{3}$$

Knowledge items are retrieved if their scores exceed a modality-dependent threshold $\theta_{\mu(\sqcap_j)}$ (Sun et al., 2024a), resulting in the retrieval set:

$$\mathcal{S} = \{\sqcap_j \mid \phi(\mathbf{x}_{\text{enc}}, \mathbf{v}_j) \geq \theta_{\mu(\sqcap_j)}\} \tag{4}$$

Finally, a multimodal generator $\mathcal{G}$ produces the response $\mathbf{y}$, conditioned on the input query $\mathbf{x}$ and the retrieved set $\mathcal{S}$:

$$\mathbf{y} = \mathcal{G}(\mathbf{x}, \mathcal{S}) \tag{5}$$

This formalism characterizes the end-to-end workflow of multimodal RAG, combining retrieval relevance with effective cross-modal content fusion.

2 Methodologies

2.1 Retrieval

Multimodal RAG systems map inputs from various modalities into a shared embedding space for cross-modal retrieval. Recent innovations in CLIP-based (Radford et al., 2021) and BLIP-inspired (Li et al., 2022) models have advanced contrastive learning, enhancing multimodal retrieval through new architectures and training approaches (Zhou et al., 2024b; Wei et al., 2024a; Zhang et al., 2024c). These multi-encoder models rely on efficient search strategies to retrieve relevant knowledge from large datasets.

Variants of Maximum Inner Product Search (MIPS) (Tiwari et al., 2024; Wang et al., 2024) are widely adopted for rapid similarity comparisons. Systems like RA-CM3 (Yasunaga et al., 2023) and MuRAG (Chen et al., 2022a) use approximate MIPS to retrieve top-k candidates by maximizing the inner product between the query and image-text embeddings. Distributed MIPS approaches, such as TPU-KNN (Chern et al., 2022), are used for large-scale, high-speed retrieval. Other efficient methods include ScaNN (Guo et al., 2020) and approximate KNN (Caffagni et al., 2024). Besides,MIPS optimization focuses on improving retrieval speed and accuracy, utilizing techniques like adaptive quantization (Zhang et al., 2023) and hybrid sparse-dense representations (Zhang et al., 2024a) to balance performance and precision.

Text Modality This remains crucial in multimodal systems, with both traditional methods (e.g., BM25 (Robertson and Zaragoza, 2009)) and newer dense retrieval models like MiniLM (Wang et al., 2020). Approaches such as ColBERT (Khattab and Zaharia, 2020) and PreFLMR (Lin et al., 2024) focus on precise semantic matching and domain-specific queries, enhancing retrieval accuracy.

Vision Modality Vision-based methods focus on extracting knowledge from images. Systems like EchoSight (Yan and Xie, 2024) and ImgRet (Shohan et al., 2024) use image queries for retrieval. Compositional Image Retrieval (CMI) (Feng et al., 2023) enhances this by merging multiple image features into one representation. Models like Pic2word (Saito et al., 2023) map visual con-

tent to text, enabling zero-shot retrieval.

Video Modality Extending vision-based methods, video-centric retrieval incorporates temporal dynamics. Techniques like iRAG (Arefeen et al., 2024) introduce incremental retrieval for sequential videos, while MV-Adapter (Jin et al., 2024) optimizes video-text transfer learning. Long-context video retrieval models, like Video-RAG (Luo et al., 2024) and VideoRAG (Ren et al., 2025), handle long-duration content effectively.

2.2 Representation Fusion

Score-based Methods Fusion strategies for multimodal representations include converting various modalities (text, tables, images) into unified formats. Zhi Lim et al. (2024) use cross-encoders for relevance scoring, while Sharifymoghaddam et al. (2024) and Li et al. (2022) apply CLIP and BLIP fusion for image-text alignment. Systems like Wiki-LLaVA (Caffagni et al., 2024) and MegaPairs (Zhou et al., 2024a) employ shared embedding spaces for images and queries. Other models, such as VISA (Ma et al., 2024a) and REVEAL (Hu et al., 2023), align text and visual data into common embedding spaces, while RA-BLIP (Ding et al., 2024) uses a multi-layer fusion module for visual-text unification. Re-IMAGEN (Chen et al., 2022b) balances creativity in image generation with semantic fidelity.

Attention-based Methods Attention mechanisms facilitate task-specific cross-modal interaction. EMERGE (Zhu et al., 2024b) and MORE (Cui et al., 2024) utilize cross-attention for data integration. RAMM (Yuan et al., 2023) uses a dual-stream co-attention transformer to merge biomedical data, while RAGTrans (Cheng et al., 2024) applies user-aware attention for social media. MV-Adapter (Jin et al., 2024) and M2-RAAP (Dong et al., 2024b) employ cross-modal techniques for video-text alignment, using auxiliary strategies and similarity-based re-weighting for fusion. Mu-RAG (Chen et al., 2022a) and Kim et al. (2024) focus on video-text alignment through attention-based methods, optimizing performance through cross-modal memory retrieval and feature refinement.

2.3 Augmentation

Context Enhancement enriches retrieved content by introducing complementary information, such as additional textual segments or visual descriptors, to provide a more comprehensive foundation for downstream generation (Caffagni et al., 2024; Xue et al., 2024). Systems like EMERGE (Zhu et al., 2024b) incorporate entity-level associations to strengthen contextual relevance, while MiRAG (Adjali et al., 2024) augments initial queries by retrieving semantically related entities. Video-RAG (Luo et al., 2024) enhances video content understanding through query rewriting strategies tailored to structured retrieval processes. Img2Loc (Zhou et al., 2024c) further improves location identification by blending both similar and contrastive retrieval results into the augmented context.

Dynamic Retrieval Rather than using static retrieval pipelines, adaptive strategies modulate the retrieval process based on the complexity and ambiguity of the input query. For example, SKURG (Yang et al., 2023) dynamically adjusts retrieval depth according to contextual demands, while SAM-RAG (Zhai, 2024) and mR^{2}AG (Zhang et al., 2024b) actively evaluate and remove irrelevant or redundant content. MMed-RAG (Xia et al., 2024) prioritizes content filtering to improve the informativeness of retrieved evidence, and OmniSearch (Li et al., 2024) breaks down complex queries into more manageable sub-questions to guide retrieval in a stepwise manner.

In addition, iterative refinement techniques continuously enhance retrieval quality by adjusting queries based on the analysis of prior retrieval cycles. IRAMIG (Liu et al., 2024) applies query reformulation across multiple turns, while OMG-QA (Nan et al., 2024) leverages episodic memory to iteratively reshape retrieval behavior. RAGAR (Khaliq et al., 2024) reinforces the reliability of evidence by incorporating successive rounds of query adjustment and validation.

2.4 Generation

Recent generation techniques on multimodal RAG focus on improving cross-modal coherence, reasoning depth, and adaptability to task-specific contexts through techniques such as in-context learning, instruction tuning, and enhanced source attribution (Tan et al., 2024; Ding et al., 2024; Ma et al., 2024a). In-context learning methods retrieve examples to guide generation without retraining, while reasoning techniques—like chain-of-thought extensions—promote stepwise cross-modal logic to enhance answer quality (Khaliq et al., 2024; Suri et al., 2024). Instruction tuning further refines generation, with approaches like RA-BLIP tailoring feature extraction to specific queries and mR^{2}AG training models to dynamically invoke retrieval

based on real-time evidence needs (Zhang et al., 2024b; Dai et al., 2023). Transparency and attribution are also prioritized, with models explicitly citing multimodal evidence to strengthen answer credibility (Zhu et al., 2025; Nan et al., 2024). Training strategies increasingly emphasize multi-stage processes that combine large-scale cross-modal pretraining with targeted fine-tuning to align retrieval, fusion, and generation components, with techniques like contrastive loss and alignment regularization helping models balance internal reasoning with retrieved knowledge (Hu et al., 2023; Chen et al., 2022a).

3 Datasets and Benchmarks

Multimodal RAG research utilizes a broad range of datasets and benchmarks, covering image-text retrieval, captioning, visual question answering, video understanding, and medical applications. Representative datasets include MS-COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), OK-VQA (Marino et al., 2019), ActivityNet (Caba Heilbron et al., 2015), and MIMIC-CXR (Johnson et al., 2019), among others. These datasets are often combined to support cross-modal retrieval and generation.

Benchmarks evaluate both retrieval relevance and generation quality. M^2RAG (Ma et al., 2024b) and MRAG-Bench (Hu et al., 2024) offer unified metrics for multimodal tasks. Other benchmarks, such as Dyn-VQA (Li et al., 2024) and ScienceQA (Saikh et al., 2022), assess dynamic retrieval and multi-hop reasoning, while RAG-Check (Mortaheb et al., 2025) focuses on retrieval reliability.

4 Evaluation Metrics

Retrieval Evaluation Retrieval effectiveness is commonly measured by Recall@K, Mean Reciprocal Rank (MRR), and F1 score, following practices from (Adjali et al., 2024; Nguyen et al., 2024).

Modality Evaluation For text outputs, evaluation relies on metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Image captions are assessed using CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and SPIDEr (Liu et al., 2017). Semantic alignment is quantified by BERTScore (Zhang et al., 2020) and CLIP Score (Hessel et al., 2021). Image quality can be assessed using FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), and Inception Score. Audio evaluation employs Fréchet Audio Distance (FAD) (Kilgour et al., 2019).

System Efficiency Efficiency measures, including FLOPs, inference time, and retrieval latency, are also reported to assess practical feasibility (Nguyen et al., 2024; Strand et al., 2024).

5 Challenges and Future Works

Despite recent advancements, multimodal RAG systems still face critical challenges in generalization, robustness, reasoning, retrieval, and scalability (Zhang et al., 2024b; Ma et al., 2024a). These systems often exhibit modality biases, over-relying on text while struggling to seamlessly integrate diverse modalities like images and speech (Winterbottom et al., 2020). Explainability also remains limited, as existing approaches frequently cite entire documents or large image segments rather than pinpointing the exact evidence (Hu et al., 2023). Effective compositional reasoning across modalities is underdeveloped, and the integration of external knowledge graphs for cross-modal reasoning remains rare (Procko and Ochoa, 2024). Retrieval processes face biases from positional sensitivity and redundant retrieval, and current embedding spaces struggle to unify different modalities into a cohesive representation (Hu et al., 2024). Future research could also explore agent-based multimodal RAGs capable of iterative self-guided refinement through dynamic feedback and real-time reasoning across modalities (Dong et al., 2024b; Sun et al., 2024b). Additionally, addressing long-context challenges, improving computational efficiency, enabling user-specific personalization while safeguarding privacy, and curating benchmarks with complex multimodal reasoning tasks will be essential for advancing this field (Kandhare and Gisselbrecht, 2024).

6 Conclusion

This survey presents a systematic analysis and taxonomy of research progress in multimodal RAG. Covering retrieval processes, development of each multimodal, augmentation techniques, and generation strategies. Furthermore, we summarize the relevant datasets, and evaluation metrics, pointing out features of them. Finally, we highlight existing technical challenges like modality biases of retrieval processes, while outlining future research directions like effective reasoning and improving retrieval relevance across modalities. We hope this survey provide insights to future researchs.

References

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

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 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.

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.

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.

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.

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.

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.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. 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.

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.

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.

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.

Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, and Liqiang Nie. 2024. Ra-blip: Multimodal adaptive retrieval-augmented boot-

strapping language-image pre-training. *Preprint*, arXiv:2410.14154.

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.

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.

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.

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.

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

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.

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*.

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

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

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

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

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.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *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.

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

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.

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. 2024. 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.

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

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

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

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

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

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

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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

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

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

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

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.

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

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. 2024a. Lambda: A large model based data agent. *arXiv preprint arXiv:2407.17535*.

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. 2024b. A survey on large language model-based agents for statistics and data science. *arXiv preprint arXiv:2412.14222*.

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*.

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.

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*.

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.

Mengzhao Wang, Xiangyu Ke, Xiaoliang Xu, Lu Chen, Yunjun Gao, Pinpin Huang, and Runkai Zhu. 2024. 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.

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

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 *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*, page 387–404, Berlin, Heidelberg. Springer-Verlag.

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

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

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

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

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

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

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

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.

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.

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.

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. 2023. Query-aware quantization for maximum inner product search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4875–4883.

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. 2024b. mr2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *ArXiv*, abs/2411.15041.

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. 2024c. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

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

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.

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.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024b. 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.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024c. 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.

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.