

Multimodal Retrieval-Augmented Generation for Knowledge-Grounded Visual Question Answering

Edu Royo

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
helloedu@unh.edu

Abstract

Knowledge-grounded visual question answering asks a system to answer natural-language questions about images by combining visual perception, linguistic reasoning, and external evidence that is not contained in the image alone. Conventional vision-language models often rely on parametric memory and therefore struggle with rare entities, time-sensitive facts, specialized domains, and questions that require verifiable provenance. Multimodal retrieval-augmented generation addresses this limitation by coupling cross-modal retrieval with generative reasoning, allowing the system to fetch relevant text, images, tables, diagrams, or structured records before producing an answer. This paper examines the architectural, infrastructural, and governance implications of multimodal retrieval-augmented generation for knowledge-grounded visual question answering. It analyzes retrieval architectures, unified semantic spaces, fusion mechanisms, and deployment trade-offs involving latency, recall, freshness, robustness, fairness, privacy, and sustainability. The discussion emphasizes that multimodal retrieval-augmented systems are not merely larger models; they are socio-technical infrastructures whose behavior depends on index construction, retrieval policies, content provenance, human oversight, and jurisdictional constraints. The paper compares deployment scenarios across medicine, education, industry, and public services, showing that different domains impose distinct requirements for auditability, abstention, and update cycles. It concludes that future progress depends on evaluation regimes that measure groundedness, citation quality, abstention, and long-term operational resilience rather than only answer accuracy.

Keywords

multimodal retrieval-augmented generation; visual question answering; knowledge grounding; vision-language models; system architecture; governance; robustness; fairness; sustainability.

1. Introduction

Visual question answering has become a central test of whether artificial intelligence systems can connect language, perception, and reasoning. Early datasets and models established the basic task of answering questions about images, but they also revealed that success on benchmark accuracy can coexist with shallow reasoning, linguistic priors, and reliance on dataset regularities [1]. Subsequent work improved visual grounding through attention mechanisms, region-level representations, and multimodal pretraining, yet many systems still answer from internal parameters rather than from verifiable external evidence [2]. This limitation becomes acute when questions involve recent events, niche entities, institutional policies, medical findings, or product specifications that were absent or underrepresented during model training. In such settings, a system that cannot retrieve and cite external knowledge may produce fluent but unsupported answers. Retrieval-augmented generation offers a structural response by separating knowledge storage from language generation and by

introducing a retrieval step before answer synthesis [4]. When extended to images, this separation becomes multimodal retrieval-augmented generation, a system class that must coordinate visual encoders, text encoders, cross-modal indices, ranking policies, fusion modules, and generation controls.

The importance of this architecture extends beyond model accuracy. Knowledge-grounded visual question answering is increasingly proposed for clinical decision support, educational tutoring, industrial inspection, accessibility tools, and public information services. In these domains, the answer alone is insufficient; users and auditors need to know which evidence was retrieved, why it was considered relevant, whether it remains current, and how uncertainty was communicated. A multimodal retrieval-augmented system therefore operates as an information infrastructure. It depends on source curation, indexing schedules, embedding models, access controls, logging, and fallback procedures. Its failure modes include stale indices, biased retrieval corpora, adversarial poisoning, privacy leakage through retrieved content, and opaque fusion that hides whether the answer came from the image, the question, or an external document. A purely model-centric account misses these structural concerns. This paper argues that knowledge-grounded visual question answering should be studied as a layered socio-technical system in which retrieval, fusion, governance, and deployment constraints jointly determine trustworthiness.

The discussion proceeds from foundations to architecture and then to governance and deployment. It first reviews the problem of knowledge grounding in visual question answering. It then examines multimodal retrieval architectures, including dense retrieval, cross-modal embedding alignment, and heterogeneous source fusion. It analyzes generation and grounding mechanisms, with attention to how retrieved evidence is integrated and cited. The paper then considers system-level trade-offs, infrastructure requirements, fairness and privacy, robustness evaluation, and sustainability. Case illustrations from medicine, education, industry, and public services show that no single architecture satisfies all domains. The conclusion emphasizes that multimodal retrieval-augmented generation for visual question answering requires evaluation and governance frameworks that treat evidence quality, abstention, provenance, and operational resilience as first-class outcomes.

2. Foundations and Problem Setting

Knowledge-grounded visual question answering differs from conventional visual question answering because the answer may require information that is not visually present or inferable from common sense. A question about a patient’s medication interaction based on a photographed label, for example, requires both optical character recognition and authoritative pharmacological knowledge. A question about a historical artifact in a museum may require provenance records, conservation notes, and scholarly interpretations. A question about a machine part on a factory floor may require maintenance manuals, sensor histories, and safety regulations. In each case, the image supplies part of the context, while external knowledge supplies the rest. Conventional vision-language pretraining has advanced the ability to align images and text, but parametric models compress knowledge in ways that are difficult to inspect, update, or attribute [6]. Retrieval augmentation introduces an explicit evidence layer that can be updated independently and audited after deployment [4]. This separation is valuable for knowledge-intensive tasks, but it also creates new interface problems between visual perception, query formulation, retrieval ranking, and answer generation.

The first generation of visual question answering models established benchmark datasets and evaluation practices that shaped the field [1]. Later models introduced attention over image

regions and question words, improving the ability to focus on relevant visual content [2]. Transformer architectures then enabled deeper cross-modal interaction through self-attention and co-attention [3]. Models such as ViLBERT, LXMERT, and UNITER demonstrated that large-scale pretraining on image-text pairs could produce representations useful for many downstream tasks [7][8][9]. More recent vision-language systems, including BLIP, BLIP-2, and Flamingo, have expanded the capacity to connect visual inputs with language generation [10][11][12]. These systems are powerful, but their knowledge remains largely implicit. They may answer a question correctly because the answer appeared in pretraining data, not because the system located a current authoritative source. For knowledge-grounded visual question answering, implicit memory is insufficient when the domain demands traceability, timeliness, or jurisdiction-specific accuracy.

The problem setting therefore involves three interacting uncertainties. The visual uncertainty concerns what is actually present in the image, including occlusion, resolution, lighting, perspective, and ambiguity. The linguistic uncertainty concerns what the question means, including reference resolution, domain terminology, and implicit assumptions. The knowledge uncertainty concerns whether the retrieved evidence is relevant, current, authoritative, and sufficient. A system that optimizes only answer likelihood may collapse these uncertainties into a confident response. A knowledge-grounded system must instead decide when to retrieve, what to retrieve, how much evidence to use, and when to abstain. This decision process is not merely technical. In medical or legal settings, an unsupported answer can cause harm, while in educational settings an overly cautious system may frustrate learners. The architecture must therefore expose enough internal state to support calibration, explanation, and oversight.

3. Multimodal Retrieval Architectures

Multimodal retrieval architectures determine what evidence is available to the generator. Dense passage retrieval demonstrated that learned embeddings can outperform sparse lexical matching for open-domain question answering by encoding semantic similarity between questions and passages [5]. In the multimodal setting, retrieval must operate across images, text, tables, diagrams, audio transcripts, and structured records. Cross-modal encoders such as CLIP align images and text in a shared embedding space, enabling image-to-text and text-to-image search [6]. However, a single shared space may not capture fine-grained domain relations. A medical image and a clinical guideline may be semantically related at a high level but require specialized features for lesion type, anatomical location, and patient history. A unified semantic space and attention-based fusion framework for multi-source heterogeneous retrieval-augmented generation illustrates the need to reconcile sources that differ in modality, granularity, and reliability [16]. The architectural challenge is not only to embed heterogeneous content but also to preserve source-specific metadata so that ranking and fusion can account for authority, recency, and jurisdiction.

A practical multimodal retrieval stack typically includes several layers. The ingestion layer normalizes documents, images, tables, and metadata; extracts text through optical character recognition or captioning; and records provenance. The indexing layer creates dense, sparse, or hybrid representations and stores them in vector databases or search engines. The retrieval layer receives a query derived from the image, question, or both, and returns candidate evidence. The ranking layer applies cross-encoders, learned re-rankers, or policy rules to order candidates. The fusion layer combines retrieved items with visual and textual representations. The generation layer produces an answer with citations, confidence signals,

or abstention. Each layer introduces trade-offs. Dense retrieval improves semantic recall but can be computationally expensive [5]. Approximate nearest neighbor methods such as hierarchical navigable small world graphs and GPU-accelerated similarity search make large-scale retrieval feasible, but they introduce recall loss and operational complexity [18][19]. Hybrid retrieval may improve robustness by combining lexical precision with semantic generalization, yet it complicates index maintenance and scoring.

Heterogeneous sources create additional design tensions. A knowledge-grounded visual question answering system may retrieve from curated encyclopedias, internal manuals, patient records, product catalogs, legal codes, or real-time sensor streams. These sources differ in update frequency, licensing, access control, and error characteristics. A hospital may require that patient data never leave a secure environment, while a public educational system may prioritize open educational resources. A factory may need sub-second retrieval from local manuals because network connectivity is unreliable. A government service may need multilingual retrieval with jurisdiction-specific legal validity. The architecture must therefore support source-aware retrieval, where the query planner chooses collections based on domain, sensitivity, and latency. It must also support provenance-aware ranking, where a recent official document outranks an older forum post even if the latter is semantically closer. Retrieval-augmented generation surveys emphasize that retrieval quality and integration strategy strongly influence downstream performance [14]. In multimodal settings, this principle expands to include visual grounding and cross-modal alignment.

4. Fusion, Generation, and Grounding

Fusion is the process by which retrieved evidence is integrated with visual and textual representations. Early fusion combines modalities before reasoning, while late fusion combines outputs after independent processing. Attention-based fusion allows the model to weigh visual regions, question tokens, and retrieved passages dynamically [3]. Cross-modal transformers can attend jointly over image patches, text tokens, and retrieved documents, but the computational cost grows with the number and length of evidence items. Systems such as BLIP and BLIP-2 use querying transformers and frozen encoders to connect visual representations with language models [10][11]. Flamingo demonstrates few-shot multimodal learning through interleaved visual and textual inputs [12]. These architectures provide strong generative capacity, but they do not automatically guarantee that generated answers are grounded in retrieved evidence. A model may attend to a retrieved passage while still relying on parametric memory, producing citations that appear plausible but are not causally responsible for the answer.

Knowledge-grounded generation therefore requires mechanisms that encourage faithfulness and attribution. One approach is to train or prompt the generator to quote or summarize retrieved evidence before answering. Another is to use verifier models that check whether the answer is entailed by the retrieved context. A third is to constrain decoding so that named entities, numbers, and dates must appear in the evidence. Retrieval-augmented generation originally demonstrated that combining a dense retriever with a sequence-to-sequence generator can improve knowledge-intensive tasks [4]. Subsequent work showed that retrieval can be integrated into language modeling at scale [17]. In open-domain question answering, generative models can condition on retrieved passages to produce more accurate answers [15]. For visual question answering, however, the generator must also justify why the visual evidence supports the retrieved knowledge. A question about a damaged component in an image may retrieve a repair manual, but the answer depends on correctly identifying the

component and matching it to the manual’s diagram. Fusion must therefore be both cross-modal and cross-source.

Grounding also has an interface dimension. Users need to see which image regions and which retrieved passages support the answer. This requires the system to preserve alignment between generated claims and evidence units. If the fusion module is opaque, explanations may be post hoc and unreliable. If the retrieval layer returns long documents, the generator may cite the document without identifying the relevant paragraph. If the visual encoder produces only global embeddings, the system may not be able to point to the specific object or region that triggered the answer. A robust architecture should support region-level visual grounding, passage-level textual grounding, and source-level provenance. It should also support abstention when evidence is missing or contradictory. In high-stakes domains, an answer that says the evidence is insufficient is more valuable than a confident hallucination. This makes calibration and uncertainty communication integral to fusion and generation, not optional post-processing.

5. System-Level Trade-Offs and Infrastructure

Multimodal retrieval-augmented generation is an infrastructure-intensive system. It requires storage for multimodal indices, compute for encoding and ranking, network capacity for distributed retrieval, and orchestration for coordinating model calls. The most visible trade-off is latency versus accuracy. Retrieving more evidence, re-ranking with cross-encoders, and running verification can improve answer quality but increase response time. Interactive applications such as accessibility assistants or surgical support may require sub-second responses, while asynchronous applications such as archival research may tolerate longer waits. A tiered architecture can balance these needs by using fast approximate retrieval for initial candidates and slower verification only when confidence is low or risk is high [19]. Caching can reduce repeated retrieval costs, but stale caches may return outdated evidence. Index freshness is another critical dimension. A system that indexes daily may miss breaking information, while continuous indexing increases cost and raises consistency challenges.

Infrastructure choices also affect resilience and sovereignty. Cloud-based retrieval offers elasticity and managed services, but it may conflict with data residency requirements or create dependency on external providers. On-premises deployment offers control and privacy but requires expertise and capital investment. Hybrid deployment can keep sensitive data local while using cloud services for non-sensitive retrieval. Edge deployment can reduce latency and preserve privacy in industrial or medical settings, but edge devices have limited memory and compute. The architecture must therefore specify which components can be distributed and which must remain centralized. Vector database sharding, replication, and backup policies are not secondary details; they determine availability and recovery. Large-scale retrieval systems such as Faiss show that efficient similarity search is achievable at scale, but they also require careful tuning of index parameters and hardware acceleration [18][19]. A knowledge-grounded visual question answering service may fail not because the generator is weak but because the retrieval index is unavailable, inconsistent, or poorly partitioned.

Operational governance is equally important. Retrieval-augmented systems depend on source agreements, licensing, access control, and update workflows. A source may be authoritative but unavailable for commercial use; another may be open but unreliable. The system needs metadata standards for provenance, versioning, and jurisdiction. It also needs monitoring for retrieval drift, index staleness, and query distribution shift. When the underlying corpus changes, evaluation results may no longer hold. When a new visual domain is added, the

embedding model may not transfer. When a language is added, cross-modal alignment may degrade. These issues require cross-functional teams including domain experts, security engineers, legal counsel, and user experience researchers. The infrastructure is therefore not just hardware and software; it is an organizational capability. Systems that treat retrieval as a static preprocessing step will struggle in dynamic environments where knowledge changes and accountability matters.

6. Governance, Fairness, Privacy, and Policy

Governance begins with data provenance and consent. Multimodal retrieval-augmented generation may ingest images, documents, tables, and personal records. Each source carries rights, restrictions, and potential harms. Datasheets for datasets provide a useful model for documenting motivation, composition, collection process, and intended uses, and similar discipline is needed for retrieval corpora [21]. Without such documentation, downstream users cannot assess whether the evidence base is representative or appropriate for their context. Privacy is especially complex because images can contain faces, locations, medical conditions, or identifying objects, while retrieved documents may contain personal data. A system that retrieves a patient record in response to a clinical image question must enforce access controls, purpose limitation, and audit logging. A public-facing system that retrieves from the open web may inadvertently expose copyrighted or sensitive content. The architecture must therefore treat privacy as a retrieval constraint, not only as a model training issue.

Fairness concerns arise at multiple layers. If the retrieval corpus underrepresents certain skin tones, languages, regions, or cultural artifacts, the system will provide weaker answers for affected users. If the visual encoder performs poorly on low-resolution images or uncommon perspectives, retrieval queries may be malformed. If the ranking policy favors popular sources, minority perspectives may be excluded. If the generator defaults to majority-language answers, multilingual users may receive lower-quality grounding. These disparities can compound. A knowledge-grounded visual question answering system used in education, for example, may consistently retrieve examples from dominant cultures, thereby reinforcing narrow perspectives. In medicine, underrepresentation in training and retrieval data can lead to diagnostic disparities. Governance must therefore include bias auditing, source diversity requirements, and mechanisms for affected communities to contest harmful outputs. The risks of large-scale language models without such safeguards have been widely discussed, including environmental costs, hegemonic bias, and misleading fluency [22]. Retrieval augmentation can mitigate some risks by grounding answers, but it can also amplify whatever biases exist in the index.

Policy implications extend to transparency, liability, and accountability. Users should know when an answer is generated from retrieved evidence, when it is based on model memory, and when the system is uncertain. Regulators may require explanations for automated decisions in domains such as credit, employment, healthcare, and public benefits. A multimodal retrieval-augmented system can support such explanations by logging retrieved evidence and fusion decisions, but logging also creates privacy and security risks. Records may contain sensitive images or queries, and logs may be subpoenaed or breached. The system must therefore use privacy-preserving audit methods, access controls, and retention limits. Liability is another unresolved issue. If a retrieved document is outdated and the generated answer causes harm, responsibility may be shared among the model provider, the data source, the deployer, and the user. Clear contracts, provenance metadata, and human oversight procedures are necessary to

allocate responsibility. Policy frameworks should encourage evidence-based deployment without creating impossible compliance burdens for smaller organizations.

7. Robustness, Evaluation, and Sustainability

Robustness in multimodal retrieval-augmented generation includes resistance to distribution shift, adversarial inputs, retrieval poisoning, and index corruption. A visual question answering system may encounter images from a new camera, a new language, or a new domain. The retrieval query may then fail to retrieve relevant evidence even if the generator is capable. Adversarial images can manipulate visual encoders, causing the system to retrieve misleading documents. A malicious actor could poison the index by injecting false documents that rank highly for common queries. These risks require red-teaming, anomaly detection, source authentication, and retrieval diversity. Evaluation should therefore go beyond answer accuracy. It should measure retrieval recall, evidence precision, citation correctness, faithfulness, abstention calibration, and robustness under perturbation. A system that answers ninety percent of questions correctly but cites irrelevant evidence may be less trustworthy than a system that answers fewer questions but abstains when evidence is weak.

Evaluation is complicated by the lack of benchmarks that combine visual questions, external evidence, and provenance requirements. Existing visual question answering datasets often assume that all necessary information is in the image or in common knowledge. Retrieval-augmented question answering datasets are often text-only. Multimodal knowledge-grounded benchmarks are emerging, but they may not reflect domain-specific authority structures. In medicine, an answer may be correct according to a general reference but wrong for a particular patient population. In law, a correct answer depends on jurisdiction and date. In industry, a correct answer depends on the exact model of a machine. Evaluation must therefore be context-sensitive. It should include temporal splits, geographic splits, and source-authority checks. It should also include human evaluation by domain experts, because automatic metrics may reward fluent but unsupported answers. Retrieval-augmented generation surveys emphasize that evaluation must consider retrieval quality and generation faithfulness together [14]. In multimodal settings, this requires joint assessment of visual grounding, textual evidence, and cross-modal consistency.

Sustainability is an often overlooked dimension. Large vision-language models and retrieval infrastructure consume significant energy for training, indexing, and inference. Energy and policy considerations for deep learning in natural language processing show that model size and training choices have environmental consequences [20]. Retrieval augmentation can reduce the need for ever-larger parametric models because knowledge is externalized, but it also adds inference-time retrieval and ranking costs. The net environmental impact depends on architecture. A system that retrieves from a small, well-curated index may be more efficient than a massive model that memorizes everything. A system that performs multi-stage re-ranking with large cross-encoders may be less efficient. Sustainable design includes caching, quantization, distillation, index compression, and carbon-aware scheduling. It also includes lifecycle planning for hardware and data storage. Organizations should measure energy use per query and consider whether the marginal accuracy gain justifies the additional cost. Sustainability is not only an environmental issue; it affects affordability and access, particularly for smaller institutions and lower-resource regions.

8. Deployment Scenarios and Future Directions

Deployment scenarios reveal how requirements diverge. In clinical settings, a multimodal retrieval-augmented visual question answering system might help clinicians interpret images or medication labels. It must respect patient privacy, provide citations to clinical guidelines, handle uncertainty, and integrate with electronic health records. Latency matters, but accuracy and auditability matter more. In education, a system might help learners understand diagrams, historical photographs, or scientific figures. It must be age-appropriate, multilingual, and transparent about sources. It must also avoid reinforcing stereotypes and should encourage learning rather than simply providing answers. In industrial inspection, a system might guide technicians through repairs using images and manuals. It must operate in low-connectivity environments, tolerate noisy images, and prioritize safety-critical information. In public services, a system might answer questions about permits, benefits, or civic documents. It must be accessible, jurisdiction-aware, and resistant to misinformation. These scenarios share a need for knowledge grounding, but they differ in risk tolerance, update frequency, and governance.

Future research should address adaptive retrieval, where the system decides how much evidence to retrieve based on question difficulty, risk, and uncertainty. It should develop federated retrieval architectures that allow organizations to share knowledge without centralizing sensitive data. It should explore compositional reasoning over multiple images, documents, and tables, because real questions often require combining evidence across sources. It should improve multimodal citation, so that generated answers point to specific image regions, passages, and table cells. It should develop standards for provenance and model cards for retrieval corpora. It should also investigate human-in-the-loop workflows where users can correct retrievals, flag stale sources, and provide feedback. Systems such as visual instruction-tuned models show the potential of general-purpose multimodal assistants [13], but domain deployment requires more than general capability. It requires bounded behavior, source-aware retrieval, and governance. The required fusion of heterogeneous sources [16] points toward architectures that can reconcile different modalities and authorities without collapsing them into a single opaque representation.

The policy environment will shape these developments. Data protection laws, copyright rules, accessibility requirements, and sector-specific regulations will influence what can be retrieved and how it can be used. Standards for transparency and auditability may become prerequisites for public procurement. Organizations will need to document source provenance, update schedules, and failure modes. They will also need to train users to interpret evidence and uncertainty. The future of knowledge-grounded visual question answering is therefore not only a matter of better models. It is a matter of building institutions and infrastructures that can sustain trustworthy retrieval over time. That includes funding for curated knowledge bases, support for domain expertise, and evaluation practices that reward honesty over fluency.

9. Conclusion

Multimodal retrieval-augmented generation offers a promising path toward knowledge-grounded visual question answering because it separates knowledge storage from generation and makes evidence available for inspection. This separation improves updateability, provenance, and domain adaptability, but it also introduces architectural complexity. Systems must coordinate visual and textual encoders, heterogeneous retrieval sources, ranking policies, fusion mechanisms, and generation controls. They must balance latency, accuracy, recall, freshness, privacy, and cost. They must be governed by policies that address bias, consent, copyright, liability, and sustainability. They must be evaluated not only by answer accuracy

but also by groundedness, citation quality, abstention, robustness, and long-term operational resilience. The paper has argued that multimodal retrieval-augmented generation should be understood as a socio-technical infrastructure rather than a single model. Deployment in medicine, education, industry, and public services shows that different domains require different trade-offs, but all require evidence that is current, relevant, and accountable. Future progress will depend on adaptive retrieval, federated architectures, multimodal citation, and governance frameworks that make trust measurable. If these elements are developed together, knowledge-grounded visual question answering can move from impressive demonstration to dependable infrastructure.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425-2433). IEEE.
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6077-6086). IEEE.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
5. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769-6781). Association for Computational Linguistics.
6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763). PMLR.
7. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 13-23.
8. Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 5100-5111). Association for Computational Linguistics.
9. Chen, Y. C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: Universal image-text representation learning. In *European Conference on Computer Vision* (pp. 104-120). Springer.

10. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (pp. 12888-12900). PMLR.
11. Li, J., Li, D., Savarese, S., & Hoi, S. (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 (pp. 19730-19742). PMLR.
12. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736.
13. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36, 34892-34916.
14. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
15. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 874-880). Association for Computational Linguistics.
16. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In 2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 591-597). IEEE.
17. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J. B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (pp. 2206-2240). PMLR.
18. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824-836.
19. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547.
20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics.
21. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
22. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM.