Multimodal Retrieval-Augmented Generation via Semantic-Aware Deep Hashing and Approximate Nearest Neighbor Search

Guohao Duan
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
guohaowork@buffalo.edu

Abstract

Retrieval-augmented generation (RAG) architectures have transformed the landscape of large-scale natural language processing by grounding generative outputs in external knowledge repositories, thereby reducing factual hallucination and improving response quality. As real-world applications increasingly demand the integration of textual, visual, and other modalities, extending RAG to multimodal settings introduces profound systems challenges related to indexing latency, storage efficiency, semantic alignment, and retrieval quality. This paper presents a comprehensive system-level investigation of multimodal RAG frameworks underpinned by semantic-aware deep hashing and approximate nearest neighbor (ANN) search. We examine the design space where high-dimensional multimodal embeddings are compressed into compact binary hash codes that preserve cross-modal semantic similarity while enabling sub-linear retrieval over billion-scale repositories. The discussion encompasses architectural trade-offs in joint and modality-specific hashing, the interplay between hashing objectives and ANN index structures, and the systemic implications for deployment, scalability, energy consumption, and fairness. We analyze governance and policy considerations arising from large-scale multimodal retrieval, including provenance attribution, bias amplification across modalities, and the sustainability of indexing infrastructure. By synthesizing cross-domain perspectives, the paper provides forward-looking insights into building robust, efficient, and ethically grounded multimodal RAG systems that can serve as knowledge-intensive backbones in high-stakes environments.

Keywords

retrieval-augmented generation, multimodal systems, deep hashing, approximate nearest neighbor search, semantic indexing, system architecture, sustainability, fairness.

1. Introduction

The emergence of large language models has reshaped the frontier of artificial intelligence, yet their intrinsic tendency to generate factually incorrect or outdated information has spurred the development of retrieval-augmented generation (RAG) as a corrective paradigm [1,2,3]. By decoupling parametric memory from non-parametric access to external document collections, RAG systems enable language models to condition their outputs on retrieved evidence, significantly improving factual accuracy and adaptability without exhaustive retraining. As the scope of these systems expands beyond text, practitioners are increasingly resorting to multimodal RAG architectures that incorporate visual, auditory, and structured data sources, thereby aligning with the inherently multimodal nature of human knowledge work. The challenge, however, is that retrieval over multimodal corpora demands indexing and search mechanisms capable of handling heterogeneous representations at enormous scale

while maintaining semantic fidelity across modalities. This paper examines a systems-centric solution pathway that integrates semantic-aware deep hashing with approximate nearest neighbor search to realize efficient, large-scale multimodal RAG.

The core motivation stems from the tension between retrieval quality and computational feasibility. Dense vector representations derived from transformer-based multimodal encoders, such as those exemplified by contrastive language-image pre-training [4], now routinely span hundreds to thousands of dimensions. Performing exhaustive nearest neighbor search in such spaces becomes prohibitively expensive as corpora grow to billions of items. In the multimodal setting, this complexity is compounded by the need to reconcile multiple embedding spaces, cross-modal relevance scoring, and dynamic update demands in production environments. Deep hashing techniques offer a principled compression pathway by learning to map high-dimensional representations to compact binary codes while preserving the underlying semantic neighborhood structure. When coupled with ANN indices, these binary codes unlock logarithmic or constant-time lookup, dramatically reducing latency and memory footprint. The research challenge is to design hashing frameworks that are not only efficient but also semantically adaptive to multimodal distributions, resilient to drift, and robust against fairness-related degradation that can emerge when compressing diverse cultural or demographic signals.

This paper adopts a system-level lens, deliberately eschewing formulaic detail to focus on architectural reasoning, infrastructure trade-offs, and societal implications. We trace the evolution from unimodal RAG to multimodal deployments, dissect the interplay between deep hashing design and ANN backbone selection, and interrogate governance dimensions including auditability, bias migration, and environmental sustainability. Through this integrative treatment, the paper aims to provide a conceptual foundation for engineers, policymakers, and interdisciplinary researchers tasked with building the next generation of large-scale retrieval-augmented systems.

2. Background and Motivation

Contemporary retrieval-augmented generation architectures operate in a two-stage pipeline: a retriever identifies relevant context passages from an external knowledge base, and a generator synthesizes a response conditioned on both the input query and the retrieved evidence. Foundational systems such as dense passage retrieval combined with sequence-to-sequence models [2] and retrieval-augmented pre-training [3] established that learned dense representations outperform traditional sparse term-matching baselines. The subsequent scaling of these architectures to web-scale corpora, as illustrated by models that retrieve from trillions of tokens during generation [5], demonstrated that retrieval quality and latency are the central bottlenecks rather than generator capacity per se. In parallel, work on fusion-in-decoder [6] and self-reflective retrieval [7] has shown that architectural integration between retriever and generator can further improve answer fidelity, but these advances amplify the need for indexing systems that sustain high recall under stringent latency budgets.

The transition to multimodal contexts is a natural progression driven by applications in visual question answering, medical diagnostics, e-commerce search, and multimedia content moderation. Unlike text-only scenarios, multimodal RAG must bridge disparate similarity notions, such as semantic text-to-image relevance, visual co-occurrence patterns, and cross-modal analogical mapping. Vision-language models trained with contrastive objectives [4] produce aligned embedding spaces in which semantically similar concepts are placed in proximity, yet the high dimensionality and continuous nature of these embeddings challenge

conventional indexing pipelines. Furthermore, real-world multimodal corpora are often constructed incrementally from heterogeneous providers, each with its own content distributions, quality standards, and metadata conventions, thereby introducing distributional shifts that can degrade static index accuracy.

The scalability imperative is underscored by the explosion of multimodal data on the internet and in enterprise knowledge bases. A video platform's search infrastructure, for example, might need to retrieve relevant thumbnails, audio segments, and subtitles simultaneously for a user's natural language query, all within a few hundred milliseconds. Achieving this requires compressing multimodal embeddings into representations that permit efficient comparison via hardware-friendly Hamming distance computations rather than floating-point cosine similarity. This very compression, however, risks discarding the fine-grained semantic distinctions that distinguish, for instance, a medicinal product from a lookalike candy wrapper in a visual scan. The design of hashing functions that preserve such subtle yet critical boundaries is therefore not merely an optimization exercise; it is a prerequisite for safe deployment in domains where retrieval errors bear safety, legal, or ethical consequences. These concerns motivate a deep inquiry into semantic-aware hashing as the linchpin of scalable multimodal RAG.

3. System Architecture of Multimodal Retrieval-Augmented Generation

A well-architected multimodal RAG system comprises four principal subsystems: modality-aware encoding, semantic hashing, approximate nearest neighbor indexing, and generation conditioned on retrieved multimodal context. The encoder layer ingests heterogeneous inputs and projects them into a shared semantic space or into modality-specific spaces equipped with cross-modal alignment heads. In many deployments, separate encoders for text, image, and audio are trained with contrastive objectives so that the resulting embeddings can be compared directly; alternatively, a late-fusion approach maintains separate indices per modality and employs cross-modal re-ranking at query time. The choice between early fusion into a unified index and late fusion with multiple coordinated indices entails trade-offs in storage footprint, update granularity, and recall.

Once embeddings are produced, the hashing subsystem converts continuous vectors into binary codes. The encoder and hashing function may be trained jointly or sequentially. Joint training typically yields better semantic preservation because the hashing mechanism can influence the representation geometry, whereas decoupled training allows the reuse of powerful pre-trained encoders without end-to-end re-optimization, which is often preferable in industrial settings with frozen external models. The design of the hashing function itself can range from locality-sensitive random projections to learned deep neural networks that exploit label information, reconstruction objectives, or adversarial regularization. The critical requirement in multimodal RAG is cross-modal semantic preservation: a text query about a red dress should produce binary codes that are close to the codes of relevant dress images even if they occupy different modality manifolds.

The binary codes populate an ANN index that accelerates retrieval. Modern indexing libraries such as Faiss [8] and ScaNN [9] provide a suite of algorithms, including inverted file systems with product quantization and graph-based structures such as Hierarchical Navigable Small Worlds (HNSW) [10]. The index structure must be carefully matched to the properties of the hashing output. Binary codes enable exact Hamming distance computation in a handful of CPU cycles, which allows exhaustive search over millions of codes per second, but at billion-scale such linear scans remain insufficient. Inverted multi-index structures that subdivide the

code space can reduce the search scope dramatically, yet they impose constraints on code length and distributional uniformity that the hashing process must satisfy. Furthermore, graph-based indices like HNSW rely on a distance metric that is ideally a metric space; Hamming distance can be adapted but may not perfectly capture semantic granularity, leading engineers to explore hybrid schemes where binary codes act as a filtering step before re-ranking with full-precision embeddings.

The generator receives the top-k retrieved multimodal items and constructs a contextual representation that can be consumed by a language model or a multimodal generative model. Architecturally, the generation stage may use cross-attention over retrieved token sequences, late concatenation of modality-specific representations, or learned memory augmentations. As retrieval quality varies, the system must also implement confidence calibration and fallback mechanisms, such as abstaining from generation when retrieval scores fall below a threshold or signaling provenance uncertainty to users. The overall architecture thus constitutes a socio-technical pipeline where indexing decisions ultimately shape both user-facing outcomes and operational cost profiles.

4. Semantic-Aware Deep Hashing: Design Principles and Trade-offs

Deep hashing for multimodal retrieval must navigate a tension between compactness and semantic fidelity. The fundamental design principle is to learn a mapping from high-dimensional input spaces to a low-dimensional Hamming space such that semantically similar items have low Hamming distance and dissimilar items have large distance. Early supervised deep hashing formulations [11,12] established that convolutional neural networks could be trained with pairwise similarity labels to produce discriminative codes for image retrieval. These methods typically employed continuous relaxations of the sign function during training and enforced quantization through regularization, yielding codes that significantly outperformed data-independent hashing baselines. The transition to large-scale multimodal settings introduces additional complexity because the similarity signal must account for cross-modal relationships derived from co-occurrence statistics, user feedback, or explicit cross-modal annotations.

Semantic awareness in this context implies that the hashing function should align with high-level conceptual taxonomies rather than merely replicating low-level modality-specific correlations. For example, an image of a stethoscope and a textual description of a cardiologist should map to nearby codes, reflecting their conceptual relatedness within the medical domain, even though their low-level features are starkly different. Achieving this alignment demands training objectives that integrate multiple supervision modalities: classification labels provide categorical boundaries, triplet loss enforces relative semantic ordering, and cross-modal contrastive loss ensures that paired text and image codes are mutually retrievable. Adversarial training strategies can further disentangle semantic content from modality-specific style, leading to codes that generalize more robustly to unseen modality combinations.

Recent advances in self-supervised asymmetric semantic excavation have addressed a critical bottleneck in prior deep hashing work, namely the difficulty of mining informative negative pairs without exhaustive label annotation [15]. By employing asymmetric network branches that jointly excavate hard semantic negatives and enforce a margin-scalable constraint, such methods improve the discriminability of compact codes under limited supervision. This line of work illustrates a broader trend in hashing research: moving from symmetric architectures that treat all modalities identically toward asymmetric, modality-sensitive designs where the

hashing function for images might exploit spatial attention patterns while the text branch leverages token-level contextual embeddings. The resulting codes can be more semantically precise, but the asymmetry introduces engineering complexity in synchronizing index segments and handling modality-specific updates. Moreover, the margin-scalable property enables system operators to tune the acceptable retrieval grain according to application risk: in high-recall scenarios where missing a relevant item carries safety consequences, a tighter margin can be enforced to ensure near-exhaustive candidate inclusion, at the cost of increased index size and computational load.

The choice of code length is a pivotal architectural decision. Shorter codes yield faster Hamming distance computation and smaller index footprints, but they severely limit the vocabulary of distinct representations and can amplify collisions among semantically distinct items, particularly in multimodal corpora where the diversity of concepts is vast. Longer codes improve precision but impose heavier storage and scanning overhead. In deployment, system architects often select a code length between 64 and 256 bits for medium-scale applications and may extend to 512 bits for web-scale indices. Empirical studies across different hashing methods suggest that code length interacts with the complexity of the embedding manifold: simpler manifold structures arising from pre-aligned vision-language models tolerate shorter codes, while heterogeneous collections with incomplete alignment benefit from longer codes and iterative re-ranking after initial bit-level retrieval [13,14,16]. These trade-offs underscore the importance of treating hashing not as an isolated module but as an integral component of an end-to-end system that jointly considers encoding, indexing, and retrieval evaluation.

5. Approximate Nearest Neighbor Search and Infrastructure Considerations

The integration of deep hashing with ANN search forms the retrieval backbone of multimodal RAG. ANN methods fall broadly into partitioning-based, graph-based, and quantization-based categories, each imposing distinct compatibility requirements on binary hash codes. Partitioning approaches such as inverted file systems divide the vector space into Voronoi cells according to a coarse quantizer and limit exhaustive search to a few candidate cells. When binary codes are used, the coarse quantizer can be constructed directly in Hamming space, and the product quantization of residual codes [17] allows efficient compression of the original embeddings for re-ranking. However, the uniformity of the cell distribution in Hamming space depends heavily on the training of the hashing model; hashing functions that produce correlated bits or highly imbalanced code distributions degrade partitioning efficiency, leading to load imbalance and hotspots in distributed retrieval clusters.

Graph-based indices, particularly HNSW, have emerged as leading solutions for high-recall approximate search in continuous spaces [10]. Adapting HNSW to binary codes is feasible by substituting the inner distance function with Hamming distance, yet the graph construction algorithm relies on the continuity of the metric to select effective long-range links. In Hamming space, the granularity is discrete, and the number of distinct distances is limited by code length, which can result in spurious equalities that confuse the graph navigation. Nevertheless, hybrid strategies that use a fast graph search over binary codes to produce a candidate set and then apply exact distance comparison on full-embedding residuals for a small fraction of items have demonstrated state-of-the-art latency-recall trade-offs. Such configurations demand careful co-design of the hashing output distribution and the graph construction hyperparameters.

Infrastructure-level decisions around storage and memory hierarchy also shape retrieval pipeline performance. Binary codes are ideally suited to in-memory indexing because a billion 128-bit codes occupy only 16 gigabytes, a fraction of the storage required for 768-dimensional float embeddings. This compactness allows the entire index to reside in RAM across a modest number of nodes, minimizing costly disk I/O. However, in multimodal systems where the index must support frequent insertion of new content, the compactness advantage must be weighed against the amortized cost of re-hashing and index rebuilding. Incremental hashing strategies that assign binary codes via locality-sensitive projections or through incremental learning with rehearsal buffers can mitigate this, but they risk semantic drift as the data distribution evolves. System monitoring must track drifts in retrieval recall and code entropy to trigger index rebuilding before serving accuracy degrades beyond operational thresholds.

The ecosystem of production-grade ANN libraries now includes GPU-accelerated search for dense embeddings [8] and highly optimized CPU-based implementations for binary codes. For multimodal RAG, the trend toward heterogeneous compute clusters that combine GPU-based re-ranking with CPU-based binary search is becoming mainstream. This architectural pattern enables the system to handle bursts of queries while maintaining energy proportionality, a consideration that is increasingly relevant as regulatory frameworks begin to scrutinize the carbon footprint of large-scale AI infrastructure. By keeping the primary index compact and efficient, deep hashing directly contributes to the sustainability posture of the overall system, a theme we expand upon in the next section.

6. Deployment, Scalability, and Sustainability

Operationalizing a multimodal RAG system at scale demands end-to-end lifecycle management encompassing data ingestion, continuous indexing, retrieval, and generation serving. The deployment architecture typically follows a microservices paradigm in which the encoder, hasher, index, and generator are independently scalable components coordinated via a message queue. The hashing service can be maintained as a stateless layer that receives batches of new multimodal documents, computes binary codes, and pushes them to index shards. Index sharding based on code prefixes or on modality identifiers distributes the query load, but it also requires a query planner that merges partial results from multiple shards while respecting global ranking constraints. This distributed design surfaces classic trade-offs around consistency, availability, and partition tolerance. Eventual consistency models are often acceptable for web search scenarios where minor staleness is tolerable, but for high-stakes applications such as legal discovery or clinical decision support, stronger consistency guarantees must be enforced, potentially reducing throughput and increasing cost.

Scalability is tested when the system serves multinational user bases with locality constraints. Data sovereignty regulations may require that certain corpora be indexed and served within specific jurisdictional boundaries, leading to geo-federated index deployments. Deep hashing offers a distinct advantage here because compact binary codes can be replicated across regions at minimal bandwidth cost, enabling a logically global index with region-local physical instances. Nevertheless, cross-modal semantic drift between region-specific content and global encoder training can lead to inconsistent retrieval experiences, motivating federated fine-tuning loops that adjust hashing parameters without exporting raw user data. The governance of such federated pipelines must balance utility against data privacy, a challenge that is actively explored in privacy-preserving deep hashing research.

Sustainability considerations have become central to the design of large retrieval systems. Training state-of-the-art encoders and hashing models consumes substantial computational energy [18], and the operational energy of serving billions of queries per day can rival the training footprint over the system's lifetime. Compact hashing reduces both the memory footprint and the number of machine cycles per query, leading to measurable reductions in power consumption. For instance, moving from dense index search to binary pre-filtering in a large-scale multimodal engine can lower query-time energy costs by an order of magnitude, particularly when search is performed on low-power edge devices. Lifecycle assessments of RAG pipelines must account for the embodied carbon of specialized hardware, such as tensor processing units and flash storage, and for the electricity mix of hosting data centers. Architectural decisions that prioritize code compactness, index freshness, and retrieval efficiency are therefore aligned with broader environmental goals, creating a synergy between technical performance and responsible innovation.

7. Fairness, Governance, and Policy Implications

The compression inherent in deep hashing can amplify societal biases present in multimodal training data. When high-dimensional semantic spaces are projected into compact binary codes, minority attributes and underrepresented modalities risk losing the critical representational granularity that would otherwise surface them in retrieval results. This phenomenon, which we term representational compression bias, is especially concerning in multimodal search applications where queries involving intersectional identities may retrieve less relevant or stereotypical content [19,20]. For example, a cross-modal query for "professional attire" might map to codes that over-index on Western business suits due to training set imbalances, marginalizing culturally specific professional dress. Addressing such biases requires fairness-aware hashing objectives that explicitly enforce demographic parity in code assignments or minimize variance in retrieval recall across protected groups, a design space that intersects with broader algorithmic fairness frameworks [21,22].

Governance structures for multimodal RAG systems must account for the opacity introduced by hashing. Unlike a nearest neighbor search over raw embeddings, where retrieved items can be attributed to a continuous similarity score with clear geometric interpretation, the Hamming distance in binary space is a discrete measure that may obscure the rationale behind retrieval rankings. This opacity complicates external audits and makes it difficult to explain to users why a particular piece of evidence was selected. Policy interventions such as mandatory model cards [23] for retrieval components and documented disclosure of hashing hyperparameters are emerging as best practices to improve transparency. Furthermore, the provenance of multimodal content indexed in RAG systems raises intellectual property and attribution challenges. Since binary codes are often stored without reference to the original data, tracing a retrieved code back to its source for licensing verification or content moderation purposes demands robust provenance logging that can withstand legal scrutiny, a significant engineering overhead that many current deployments overlook.

The cross-modal nature of these systems also introduces unique regulatory tensions. A retrieval index that combines news images, user-generated videos, and textual encyclopedia entries may fall under divergent content liability regimes in different jurisdictions. Governance frameworks must delineate the responsibilities of platform operators when retrieved multimodal content informs generated text that could be deemed defamatory, privacy-violating, or dangerous. The European Union's Artificial Intelligence Act, among other emerging regulations, classifies certain RAG applications as high-risk when used in

critical infrastructure, education, or law enforcement, imposing requirements for accuracy, robustness, and human oversight. Deep hashing modules integrated into such high-risk systems must undergo rigorous conformity assessments, including evaluation of worst-case retrieval failures and adversarial robustness against code-space perturbations. Proactive engagement between system architects and policy bodies is essential to harmonize technical design choices with evolving legal standards and to ensure that multimodal RAG technologies serve the public interest equitably.

8. Robustness and Reliability

The reliability of a multimodal RAG pipeline is contingent on its resilience to distributional shift, adversarial manipulation, and component failure. Distributional shift in multimodal data can arise from sudden changes in content popularity, seasonal events, or platform policy changes that alter the mix of modalities. When the hashing function is not regularly updated, the semantic alignment between query codes and index codes degrades, leading to silent recall loss that may go undetected if retrieval health metrics are not modality-stratified. Continuous monitoring dashboards that track per-modality code entropy, query-code collision rates, and retrieval precision on a holdout set of known relevant multimodal pairs are critical operational safeguards. When drift is detected, automated fine-tuning pipelines can trigger re-hashing and partial index rebuilds while the system remains online, an architectural requirement known as hot-swappable index partitioning.

Adversarial robustness in the context of binary codes presents distinctive challenges. Attackers can craft multimodal inputs designed to produce binary codes that collide with sensitive or harmful content, effectively poisoning the retrieval pool even if the generated output is later filtered. Defense mechanisms include randomized hashing ensembles, where multiple independent hashing functions are applied and the consensus among retrieved sets is required, and query-side bit flip detection that monitors for anomalous code patterns indicative of adversarial construction. These defenses, however, increase query latency and introduce additional hyperparameters that must be calibrated against baseline performance. The system-level decision to employ adversarial defenses is thus a risk management choice that weighs the threat model against operational costs.

Reliability engineering for multimodal RAG must consider the cascade effect of retrieval failure on generation quality. If the retriever fails to surface relevant evidence, the generator may produce fabrications that are then presented to users with high confidence. Mitigations such as confidence-aware generation, where the generator outputs a calibrated uncertainty estimate derived from retrieval scores, and defined fallback responses when no high-confidence retrieval is available, are becoming standard in production systems. The robustness of the overall pipeline can be enhanced through redundancy in retrieval pathways: a primary hashing-based ANN index can be complemented by a slower but exact secondary retriever over a window of recent content, serving as a safety net for high-stakes queries. This layered retrieval architecture exemplifies the principle of defense in depth, adapted to the retrieval setting.

Finally, the long-term reliability of multimodal RAG depends on the sustainability of the indexing ecosystem itself. As language evolves, visual styles change, and new modalities such as haptic or olfactory data emerge, the encoder and hashing models must be periodically retrained. The archival of historical indices and the ability to replay queries against past states become essential for longitudinal studies, reproducibility audits, and legal discovery. Designing for such archival capabilities from the outset, rather than retrofitting them, is a

systems engineering discipline that is increasingly recognized as part of responsible AI practice [24]. The interplay between storage cost, retrieval fidelity, and archival depth represents yet another multi-objective optimization that system architects must navigate, guided by the intended use domain and the expected duration of system operation.

9. Conclusion

Multimodal retrieval-augmented generation represents a paradigm shift in how AI systems access and reason over the world's knowledge, moving from monolithic model memory to dynamic, heterogeneous evidence integration. This paper has argued that the combination of semantic-aware deep hashing and approximate nearest neighbor search provides a foundational systems architecture capable of meeting the scalability, efficiency, and semantic fidelity demands of multimodal RAG. Through a thorough examination of design choices, we have shown that effective hashing is not merely about compression ratios but about the delicate preservation of cross-modal semantics under tight computational budgets. The architectural decisions surrounding code length, index structure, training asymmetry, and update strategy all have cascading effects on retrieval latency, energy consumption, fairness, and auditability.

We have further illuminated the governance and sustainability dimensions that elevate multimodal RAG from a technical subsystem to a socio-technical infrastructure. The deployment of these systems at scale requires a holistic view that encompasses federated index management, carbon-aware engineering, fairness-aware hashing objectives, and adversarial robustness. As regulation tightens and public expectations for transparency rise, system architects must engage proactively with policy communities and embed ethical considerations into the very fabric of indexing and retrieval pipelines. The future evolution of multimodal RAG will likely involve dynamic hashing functions that adapt to usage patterns, retrieval modalities yet to be invented, and continual alignment with societal values. It is our hope that the system-level perspective advanced in this paper will serve as a conceptual map for researchers and practitioners committed to building multimodal knowledge systems that are not only powerful but also trustworthy and sustainable.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

2. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769–6781).

3. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. In International Conference on Machine Learning (pp. 3929–3938). PMLR.

4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

5. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In International Conference on Machine Learning (pp. 2206–2240). PMLR.

6. Izacard, G., Grave, E., Joulin, A., & Usunier, N. (2022). Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 23(251), 1–42.

7. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511.

8. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.

9. Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., & Kumar, S. (2020). Accelerating large-scale inference with anisotropic vector quantization. In International Conference on Machine Learning (pp. 3887–3896). PMLR.

10. 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.

11. Liu, H., Wang, R., Shan, S., & Chen, X. (2016). Deep supervised hashing for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2064–2072).

12. Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). HashNet: Deep learning to hash by continuation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 5608–5617).

13. Su, S., Zhang, C., Han, K., & Tian, Y. (2018). Greedy hash: Towards fast optimization for accurate hash coding in CNN. Advances in Neural Information Processing Systems, 31.

14. Shen, F., Shen, C., Liu, W., & Tao, D. (2018). Deep semantic hashing with generative adversarial networks. In Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 225–234).

15. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

16. Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1), 117–128.

17. 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).

18. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

19. 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).

20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

21. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29.

22. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

23. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

24. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.