

Computationally Efficient Mamba Models for Large-Scale Medical Image Retrieval

Claude R. Norris

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

Brent Ortiz

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
brentortiz@ku.edu

Abstract

Large-scale medical image retrieval systems must reconcile high diagnostic relevance, low inference latency, memory efficiency, and robust governance across heterogeneous clinical settings. This article analyzes the integration of Mamba, a selective state space sequence model, into retrieval pipelines for medical image repositories. The discussion adopts a system-level orientation rather than focusing narrowly on benchmark accuracy. It examines architectural decisions, computational trade-offs, deployment infrastructure, fairness, sustainability, and clinical accountability. Mamba architectures replace quadratic attention mechanisms with selective scan operations that scale approximately linearly with sequence length, making them attractive for high-resolution radiology, pathology, and multi-view imaging. However, moving from research prototypes to production retrieval systems raises structural issues. Patch tokenization order, bidirectional state propagation, hybrid attention-state space designs, and embedding normalization directly influence retrieval precision and failure modes. At the infrastructure layer, the paper considers vector database integration, sharding, query planning, caching, and fallback strategies in hospital networks. Governance and fairness concerns arise from dataset composition, minority anatomical presentation, disease prevalence imbalances, and the difficulty of explaining similarity decisions. The article further discusses energy use, model refresh cycles, and procurement policies. By synthesizing architecture, infrastructure, and policy perspectives, the paper offers a cross-disciplinary framework for building computationally efficient Mamba-based medical image retrieval systems that are not only accurate but also operationally sustainable and ethically defensible. Efficiency gains must be coupled with transparent evaluation, resource-aware deployment, and inclusive governance mechanisms to avoid amplifying existing clinical inequities.

Keywords

Medical image retrieval, Mamba, selective state space models, computational efficiency, large-scale infrastructure, fairness, governance.

1. Introduction

The retrieval of clinically relevant images from large-scale repositories has become a foundational problem in contemporary health systems. Picture archiving and communication systems now contain millions of radiological, pathological, and dermatological images, yet traditional search mechanisms remain dominated by metadata fields such as modality, body

part, and acquisition date. These metadata fields cannot fully express visual similarity, subtle disease progression, or cross-sectional anatomical relationships. Attention-based architectures have significantly improved representation learning for visual and textual data [1], but their computational behavior creates challenges for very large retrieval environments. Selective state space models, particularly Mamba [2], have emerged as a compelling alternative because they replace dense attention operations with recurrent state propagation that scales efficiently with sequence length. This property is important for medical imaging because high-resolution scans and three-dimensional volumes produce long token sequences that stress memory and processing capacity.

Before the emergence of selective state space models, deep convolutional networks established the standard for medical image feature extraction due to their strong local receptive fields and translational inductive biases [3]. Vision transformers later demonstrated that patch-based tokenization and self-attention can model long-range dependencies [4]. However, these architectures often require substantial computational resources when applied to large image collections. A recent cross-fusion Mamba approach for multi-view medical image classification illustrates the growing interest in applying state space models to clinically structured image sets [5]. Closely related work on vision Mamba variants has extended selective scanning to two-dimensional visual tasks [6], while input-output-aware attention kernels have shown that efficiency gains can be achieved through hardware-conscious design [7]. Content-based medical image retrieval systems have a long research history [8], and deep learning has expanded their representational capacity considerably [9].

This article departs from algorithm-centered comparisons by treating medical image retrieval as a sociotechnical system. The core argument is that Mamba-based architectures must be evaluated not only by retrieval metrics but also by their interaction with storage infrastructure, inference workloads, clinical governance, fairness obligations, sustainability targets, and institutional procurement practices. Such an analysis is necessary because medical retrieval systems are embedded in hospitals, clinics, and public health agencies where resource constraints, privacy rules, and clinical workflows shape technical viability. The following sections examine architectural foundations, system-level trade-offs, deployment infrastructure, robustness, fairness, governance, and policy implications. By integrating these perspectives, the paper provides a structured account of the challenges and opportunities associated with computationally efficient Mamba models in large-scale medical image retrieval.

2. Background and Related Work

Early medical image retrieval research concentrated on handcrafted visual features and query-by-example interfaces. These systems were often limited by the semantic gap between low-level image statistics and high-level clinical meaning. The field has since shifted toward learned representations that can capture tissue patterns, lesion boundaries, and subtle textural variations. Content-based retrieval pipelines generally consist of feature extraction, embedding construction, similarity computation, and result presentation. Deep learning has improved every stage, but it has also introduced new operational burdens related to model size, inference latency, and annotation requirements. The medical imaging literature has documented these transformations and emphasized the need for clinically meaningful evaluation beyond precision and recall [8]. The introduction of deep learning into medical image analysis broadened the scope of possible retrieval tasks while also raising questions about generalization, reproducibility, and data quality [9].

Contrastive learning has been especially influential for retrieval because it directly optimizes the similarity structure of the embedding space [10]. By pulling representations of semantically similar images together and pushing dissimilar images apart, contrastive objectives support nearest-neighbor retrieval without relying on manually assigned disease labels. Vision-language pretraining has further enriched retrieval by aligning images with textual descriptions, enabling queries that combine visual findings and clinical language [11]. These representation learning advances depend on efficient approximate nearest-neighbor search when the gallery contains millions of items. GPU-based similarity search frameworks have made billion-scale retrieval feasible in consumer and industrial settings [12]. In medical contexts, large public datasets such as CheXpert have facilitated the development of deep models for chest radiograph interpretation, while also exposing issues of label uncertainty and class imbalance [13]. Together, these developments create a foundation for studying selective state space models in retrieval systems.

3. Architectural Foundations of Mamba for Retrieval

Mamba extends the state space modeling tradition by introducing selective gating over state transitions. Unlike fixed state space models, Mamba allows state coefficients to depend on the input, which enables the model to retain or discard information based on the relevance of each token. This selective behavior is particularly valuable for medical images because not all regions carry diagnostic importance. A chest radiograph may contain large areas of unremarkable lung tissue, while a small opacity near the pleural boundary is clinically decisive. Selective state space models can in principle allocate representational capacity unevenly across tokens without computing pairwise attention between every pair of positions. This structural property supports memory-efficient inference and makes it possible to process long sequences that would be costly under attention-based designs.

Medical images must be converted into sequences of patch tokens before being processed by a Mamba encoder. The chosen scanning order profoundly affects the model because selective state propagation has a directional bias. Naive raster scanning may separate anatomically connected structures that lie above and below one another. Bidirectional scanning, as used in several vision Mamba variants [6], mitigates this limitation by allowing information to flow in both directions across the image plane. Multi-view medical classification further depends on cross-view fusion, where representations from different orientations must be integrated without losing view-specific diagnostic details [5]. For retrieval, the fusion strategy determines whether the resulting embedding is stable across repeated scans, subtle rotations, and acquisition differences. Public critical care databases such as MIMIC-III illustrate the value of aligning imaging findings with broader clinical signals [14], and chest radiograph baselines such as CheXNet demonstrate the clinical utility of deep visual representations [15]. These resources provide useful benchmarks for evaluating whether Mamba-based retrieval embeddings retain clinically relevant variation while suppressing nuisance variation.

4. System-Level Design and Computational Trade-offs

The adoption of Mamba models for large-scale medical image retrieval introduces a series of system-level trade-offs that cannot be captured by accuracy alone. The first trade-off concerns the balance between selective state propagation and attention mechanisms. Pure Mamba encoders reduce memory consumption for long sequences, but they may be less effective at capturing some forms of long-range consistency that attention handles naturally. Hybrid designs that interleave selective scan layers with limited attention blocks can improve retrieval quality at the cost of higher latency and more complex computation graphs. System

designers must decide whether the retrieval task demands global context across an entire image or whether local anatomical structure is sufficient. For mammography, the relation between distant image regions is often clinically relevant, whereas for dermatology the local texture may dominate. These task-specific differences shape the optimal trade-off between efficiency and representational richness.

A second trade-off involves embedding dimensionality and gallery size. Large embedding vectors can encode finer distinctions, but they increase storage requirements, index construction time, and network transfer costs. Small embedding vectors reduce retrieval latency but may collapse clinically distinct findings into nearby points. The selective scan mechanism itself can be configured with different state dimensions, expansion factors, and block counts. These choices affect both the discriminative power of the embedding and the memory footprint during inference. Fairness audits have shown that such design choices can disproportionately affect underrepresented groups when the model relies on spurious correlates [16]. Mamba-based retrieval systems must therefore be evaluated on subgroup-level retrieval consistency, not only on average accuracy. This is especially important in medical imaging because disease presentation varies by age, sex, genetic background, and comorbidity profile.

A third trade-off concerns training energy and inference sustainability. State space models are often praised for inference efficiency, but large-scale pretraining remains energy-intensive. The environmental cost of model development must be considered alongside clinical benefit [17]. In addition, interpretability remains a concern. Clinicians often request explanations for why a particular image was retrieved, and gradient-based visualization methods offer one route for generating spatial explanations [18]. Selective state space models can be paired with such methods, but the explanations must be validated against clinical reasoning rather than assumed to be inherently trustworthy. These system-level considerations indicate that the choice of Mamba architecture is not merely a modeling decision; it is a governance decision with consequences for storage, energy, fairness, and clinical communication.

5. Large-Scale Retrieval Infrastructure and Deployment

Deploying Mamba-based medical retrieval systems in production requires careful integration with vector databases, caching layers, and hospital network infrastructure. Medical image repositories are typically distributed across departments, with radiology, pathology, and cardiology maintaining separate storage systems. A retrieval service must therefore coordinate multiple data sources without creating unmanageable synchronization bottlenecks. Embedding extraction can be performed offline for large batches, but new studies must be indexed promptly to support clinical workflow. Online extraction requires reserved GPU capacity, while offline extraction allows workload smoothing but introduces staleness. Vector databases support approximate nearest-neighbor search, but they require tuning for index type, compression, and refresh frequency. GPU-based similarity search frameworks have demonstrated that high-performance retrieval can be achieved on large collections [12], yet medical deployments must additionally satisfy access control, audit logging, and data residency requirements.

Institutional deployment also requires fallback strategies during network outages, hardware failures, or model updates. A retrieval system that fails silently may delay clinical decisions, so the infrastructure must include explicit degradation modes. For example, if the learned embedding service is unavailable, the system may revert to metadata-based search while alerting the clinician that image similarity results are incomplete. Open foundation model

families such as LLaMA illustrate the value of accessible architectures that can be adapted for domain-specific tasks without complete retraining [19]. A similar logic applies to medical retrieval: Mamba encoders may be pretrained on broad visual data and then fine-tuned on clinical collections. However, distribution shift remains a serious risk. Models trained on curated public datasets may not generalize to older imaging equipment, different acquisition protocols, or underrepresented patient populations. Robustness studies in computer vision have shown that models can fail under distribution shifts even when benchmark accuracy is high [20]. Medical retrieval systems must therefore include continuous monitoring and recalibration procedures.

6. Robustness, Fairness, and Governance Considerations

Medical image retrieval systems inherit the fairness challenges of the datasets and representations they use. If the dataset contains mostly adult chest radiographs, the retrieval system may be less reliable for pediatric patients. If the training collection overrepresents one disease prevalence, rare conditions may be systematically overlooked. These biases can be amplified by contrastive objectives that emphasize dominant visual patterns. A fairness-oriented evaluation must measure retrieval performance across demographic subgroups, disease categories, and imaging devices. Bias surveys in machine learning emphasize that fairness is not a single metric but a multidimensional construct requiring institutional commitment [16]. For medical retrieval, fairness also includes ensuring that patients with rare diseases are not excluded from relevant clinical comparisons simply because their image features lie in sparse regions of the embedding space.

Governance mechanisms must address the opacity of similarity decisions. Clinicians may accept a retrieval result if it is accompanied by a plausible explanation, but they may distrust a black-box similarity score in high-stakes settings. Regulatory frameworks such as the European Union General Data Protection Regulation have prompted debates about the scope of the right to explanation for automated decision-making [21]. Although the legal interpretation of this right remains contested, its presence reflects broader public expectations of accountability. Mamba-based retrieval systems should therefore include audit trails that record which model version, preprocessing scheme, and index configuration produced each result. Retrieval transparency also requires communicating the limitations of the system. If the underlying model has not been validated for a particular modality, the interface should disclose this uncertainty rather than presenting similarity results as definitive diagnostic evidence.

7. Sustainability and Operational Policy Implications

Sustainability considerations are becoming more prominent in medical artificial intelligence. Although Mamba models may reduce inference energy relative to attention-based models, the full lifecycle includes pretraining, fine-tuning, embedding extraction, vector index construction, and model refresh. Each of these stages consumes energy and hardware resources. Energy and policy research has shown that deep learning development can have substantial environmental costs, especially when large model sweeps are performed without careful resource planning [17]. Medical institutions should therefore adopt procurement criteria that include energy efficiency, hardware longevity, and the carbon intensity of the local electricity supply. Efficiency should not be treated as a one-time architectural property; it must be maintained through monitoring, workload scheduling, and periodic model consolidation.

Operational policy must also address model governance. A Mamba-based retrieval system embedded in a hospital requires clear ownership of model updates, data access, and error correction. Without assigned responsibility, technical drift may go unnoticed until retrieval quality degrades. Institutional review boards and clinical informatics committees should periodically evaluate whether the retrieval system continues to serve its intended purpose. Procurement policies should favor vendors that provide reproducible training configurations, documented failure modes, and auditable evaluation scripts. Open model architectures can support reproducibility, but they also require institutional capacity for ongoing maintenance. The long-term success of Mamba-based medical retrieval depends less on a single architectural breakthrough than on the development of reliable institutional practices for validation, deployment, and review.

8. Conclusion

Computationally efficient Mamba models offer a promising direction for large-scale medical image retrieval. Their selective state space mechanism supports linear-time sequence processing and memory-efficient inference, which is highly relevant for high-resolution and multi-view medical images. However, the transition from research prototypes to clinically embedded retrieval systems requires attention to architectural trade-offs, infrastructure design, fairness, robustness, governance, and sustainability. Hybrid attention-state space configurations, careful patch traversal, embedding dimension selection, and vector database integration all shape the behavior of the deployed system. Retrieval quality must be measured not only by average accuracy but also by subgroup consistency, robustness to distribution shift, and explainability. Medical institutions should treat Mamba-based retrieval as a sociotechnical intervention rather than a purely computational upgrade. Procurement, validation, monitoring, and error correction must be built into the deployment lifecycle. By integrating these perspectives, researchers and practitioners can develop retrieval systems that are efficient, clinically useful, and ethically defensible in large healthcare ecosystems.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates.
2. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
5. Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 672–682). Cham: Springer Nature Switzerland.

6. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417.
7. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems* (pp. 16344–16359). Curran Associates.
8. Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1), 1–23.
9. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
10. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). PMLR.
11. 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.
12. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.
13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590–597.
14. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.
15. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
16. 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.
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). Association for Computational Linguistics.
18. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE.

19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
20. Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In Proceedings of the 36th International Conference on Machine Learning (pp. 5389–5400). PMLR.
21. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99.