Vision-Language Mamba Models for Multimodal Medical Image and Clinical Report Understanding

Dennis Warner
School of Computing, Clemson University, Clemson, SC, USA.
dennis.work@clemson.edu

Aarohi Goel
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
goel-aarohi@unr.edu

Abstract

The convergence of medical imaging and clinical narrative text holds substantial promise for diagnostic decision support, yet it also exposes deep technical and organizational challenges. Transformer-based vision-language models have advanced multimodal representation learning but are constrained by quadratic scaling when applied to high-resolution volumetric images and lengthy clinical documents. Selective state space models, particularly the Mamba family, provide a compelling alternative through linear-time sequence modeling and input-dependent state transitions. This paper presents a system-level examination of vision-language Mamba architectures for multimodal medical image and clinical report understanding. It analyzes architectural choices involving visual tokenization, bidirectional scanning, cross-modal fusion, and report generation. Beyond model design, the paper considers data governance, infrastructure requirements, deployment constraints, robustness, fairness, and regulatory oversight. It argues that efficiency gains must be accompanied by rigorous evidence of subgroup performance, auditability, and resilience to distribution shift. The discussion connects clinical safety requirements with systems engineering trade-offs, emphasizing that multimodal medical models are best understood as components of broader sociotechnical infrastructures rather than isolated predictors. A forward-looking agenda is proposed for integrating state space vision-language systems into clinical workflows in a manner that is sustainable, equitable, and accountable.

Keywords

vision-language models; state space models; medical imaging; clinical reports; multimodal learning; health AI governance.

1. Introduction

The clinical interpretation of medical images rarely occurs in isolation. Radiologists, pathologists, and other specialists integrate visual findings with patient history, laboratory results, prior reports, and clinical guidelines. This integration is central to diagnosis, but it has proven difficult to replicate in computational systems because visual and linguistic modalities differ markedly in structure, scale, and meaning. Transformer architectures have become the dominant approach to multimodal learning because self-attention can model long-range dependencies across tokens [1]. Nevertheless, self-attention incurs quadratic computational cost with respect to sequence length, which creates practical barriers when inputs include three-dimensional imaging volumes, whole-slide pathology images, or long longitudinal

clinical documents. The need for more efficient sequence modeling is especially pressing in medicine, where acceptable latency and resource consumption are tightly constrained.

Selective state space models have recently emerged as an important alternative to attention. The Mamba architecture replaces dense attention with input-dependent state transitions, enabling linear-time sequence processing while retaining a selective focus on relevant information [2]. The relationship between attention and structured state space models has been further clarified through generalized algorithms that connect the two families under a common computational framework [3]. Visual adaptations of Mamba have extended state space modeling to image classification and dense visual tasks by scanning image patches across multiple spatial directions [4][5]. These developments indicate that state space models may reduce the computational burden of medical vision-language processing without sacrificing the capacity to reason over long contexts.

Natural image vision-language systems such as contrastive language-image pretraining and bootstrapped image-text models have shown that large-scale alignment between visual and textual representations can support zero-shot transfer and multimodal reasoning [6][7]. In medical imaging, cross-fusion state space methods have begun to address multi-view image classification by aggregating complementary views through selective cross-modal updates [8]. These approaches suggest that Mamba-based vision-language models may be well suited to multimodal medical tasks that require simultaneous interpretation of images and reports.

This paper does not present a new benchmark or a single architecture. Instead, it offers a system-level analysis of vision-language Mamba models for multimodal medical image and clinical report understanding. The discussion spans architectural trade-offs, data governance, infrastructure, deployment, sustainability, robustness, fairness, and policy. The central argument is that efficiency alone is insufficient. Medical multimodal systems must be designed with attention to institutional context, clinical evidence standards, and the distributed responsibilities of developers, deployers, and regulators.

2. Background and Related Work

The shift from narrow medical image classifiers to foundation models reflects a broader ambition to build generalist systems that can perform multiple tasks across modalities [9]. Large language models have demonstrated substantial clinical knowledge in question-answering and reasoning tasks, although their deployment requires careful alignment with professional norms and safety requirements [10]. Domain-specific clinical language models trained on electronic health record text have improved performance on report classification, named entity recognition, and similar tasks by capturing the lexical patterns of clinical documentation [11]. These text-only advances, however, do not fully address the coupled challenge of interpreting images and reports together.

Public datasets have played a decisive role in shaping medical imaging research. MIMIC-CXR provides chest radiographs paired with clinical reports, enabling studies of image-text alignment and report generation [12]. CheXpert offers large-scale expert-labeled chest radiographs with uncertainty labels, supporting robust classification research [13]. Evaluations on such datasets have revealed important fairness problems. Models trained on chest radiographs can systematically underdiagnose disease in underserved populations, partly because of label noise and population-specific prevalence differences [14]. Early high-profile systems such as CheXNet demonstrated that deep networks can achieve radiologist-level pneumonia detection on chest radiographs [15]. However, single-task systems of this kind

offer limited support for open-ended report interpretation and can produce hidden failure modes that are difficult to detect in aggregate performance metrics.

Concerns about fairness and safety extend beyond imaging. An influential analysis of a commercial risk prediction algorithm showed how biased labels can reproduce and amplify racial disparities in health care [16]. Foundation models introduce additional risks related to homogenization, opacity, emergent behavior, and the concentration of power in a small number of large-scale systems [17]. These considerations inform the design of vision-language Mamba models because architectural efficiency does not automatically mitigate the social and clinical risks associated with large-scale multimodal learning. Instead, systems must be constructed with explicit mechanisms for auditing data, monitoring subgroup performance, and documenting deployment assumptions.

3. Architecture and System Design

A vision-language Mamba system for medical understanding typically consists of three major components: a visual encoder that maps images or volumes into sequence representations, a language encoder or decoder that processes clinical text, and a fusion module that aligns and integrates the two modalities. Visual encoders based on state space models can process image patches by scanning along multiple axes, thereby preserving spatial structure while maintaining linear-time complexity. Language components may be implemented with efficient transformer decoders or with state space language models. Open language models such as LLaMA illustrate the value of efficient parameterization for domain adaptation and local deployment [18]. At the same time, promptable segmentation models such as Segment Anything show that flexible visual encoders can support interactive and task-agnostic image understanding, which may be useful for clinical annotation and region-of-interest analysis [19]. In a medical multimodal system, these components must be integrated through a fusion strategy that balances expressiveness, latency, and interpretability.

Fusion placement is a central architectural choice. Early fusion combines raw visual and textual features before deep processing, which can capture low-level correspondences but may obscure modality-specific structure. Late fusion processes each modality independently and combines representations only at the decision stage, which is simpler but often misses fine-grained interactions. Intermediate fusion, including cross-attention and cross-state mechanisms, allows selective exchange of information between modalities at multiple depths. Mamba-based intermediate fusion can be implemented by conditioning state updates on cross-modal features, thereby enabling a report token to influence the representation of an image region or an image region to guide text encoding. This type of selective cross-modal update is especially relevant for multi-view andmultimodal clinical scenarios, where a single representation must integrate heterogeneous sources without incurring quadratic attention costs. A cross-fusion Mamba module may maintain separate state vectors for visual and textual streams while allowing one stream to read the other's hidden state at each scanning step. This design is notable because it preserves modality-specific state dynamics but also permits coordinated updates when visual evidence conflicts with narrative findings. The trade-off is primarily architectural: stronger coupling can improve alignment but may increase optimization difficulty and reduce the interpretability of individual streams. Systems designers must therefore decide whether cross-modal state exchange should occur at every layer or only at selected fusion layers. Deeper coupling can increase representational capacity but also raises the risk of spurious correlations when training data are limited or when reports contain incidental references to unrelated conditions.

In many medical vision-language systems, the decoder is responsible for generating clinical reports from visual and textual inputs. The decoder may be autoregressive, producing tokens sequentially conditioned on prior tokens and fused visual features. State space language decoders offer lower memory footprints than transformer decoders, which is valuable when generating long reports that include multiple sections such as findings, impression, and recommendations. However, autoregressive generation remains sequential. The efficiency advantage of Mamba is most pronounced during encoding and fusion rather than during the token-by-token decoding process. Nevertheless, the reduction in peak memory during training can enable larger batch sizes and longer sequences, contributing to more stable optimization across high-resolution imaging inputs. Visual tokenization also deserves attention. A common approach is to divide an image into non-overlapping patches and process them through a visual state space encoder. For volumetric modalities such as computed tomography or magnetic resonance imaging, the image can be decomposed into slices or three-dimensional patches. Bidirectional or multi-axis scanning is then used to capture spatial dependencies. This scanning strategy introduces a structural trade-off: more scanning directions improve spatial coverage but increase sequence length and may dilute selective state updates. Designers must balance the need for contextual richness with the constraints of available memory and the risk of overfitting in small clinical cohorts.

4. Data Governance and Infrastructure

Multimodal medical data combine visual and textual records that are highly sensitive and often governed by different legal and institutional regimes. Radiological images are protected health information under many jurisdictions, while clinical reports include unstructured narratives that may reveal family history, social circumstances, or behavioral details. The construction of a vision-language training corpus therefore requires careful attention to de-identification, consent, and data use agreements. Public datasets have accelerated research by providing accessible paired image-text examples [12][13], but their geographic and demographic coverage remains uneven. A system trained predominantly on data from a small number of institutions may perform inconsistently when transferred to other clinical settings. Data governance must thus be understood not merely as a compliance activity but as a determinant of model behavior across subpopulations and clinical presentations. Institutions that contribute data need mechanisms for reviewing how their records are used and for withdrawing consent where applicable. Federated learning offers one path toward reducing centralization while preserving privacy, though it introduces additional coordination costs and heterogeneous local data formats.

Infrastructure for vision-language Mamba training must support high-throughput access to large imaging archives and corresponding reports. Although linear-time state space models reduce the asymptotic cost of sequence processing, the practical requirements of training foundation-scale systems remain substantial. Image preprocessing pipelines may generate thousands of patches per study, and report tokenization across large corpora creates long sequence streams. The efficiency of Mamba can lower the barrier to training on longer visual contexts, but it does not remove the need for robust data pipelines, versioned datasets, checkpoint management, and fault-tolerant distributed training. In health systems, compute resources are often distributed across on-premise servers and regulated cloud environments. The choice between centralized training and federated fine-tuning is not purely technical; it interacts with institutional trust, data sovereignty, and the cost of maintaining local infrastructure. Open language models have demonstrated that efficient parameterization can

enable domain adaptation outside a single central provider [18], and a similar logic applies to medical vision-language models built on state space backbones. Local fine-tuning may improve alignment with local documentation practices, but it also requires institutions to maintain evaluation pipelines and to monitor drift after deployment.

Evaluation infrastructure is a further concern. Model performance in a research setting often differs from performance in clinical practice. Independent benchmarks may reflect narrow distributions of disease prevalence and imaging protocols. External validation requires curated test sets that span multiple vendors, scanners, and patient groups. For vision-language systems, evaluation must include both image-only metrics and text generation quality. The latter is especially difficult because clinical language can be highly variable, and surface-level similarity to a reference report may not capture factual correctness. Automated metrics may reward fluent but clinically inaccurate text. Infrastructure for human evaluation, including structured expert review and disagreement resolution, is therefore essential. Without such infrastructure, efficiency gains cannot be interpreted as evidence of clinical utility. The governance of foundation models in medicine must establish expectations for documentation, audit trails, and version control that go beyond typical research reporting [9][17]. A multimodal foundation model is not a static artifact but a continuously updated system whose changed behavior may affect downstream decisions.

5. Deployment and Clinical Integration

Deployment of vision-language Mamba models in clinical settings raises questions about how these systems fit into existing workflows. Radiologists commonly read images while dictating or reviewing reports, and their attention is already divided among multiple information sources. A multimodal assistant might suggest differential diagnoses, flag discrepancies between images and prior reports, or draft preliminary findings for review. However, the integration of such tools must avoid creating automation bias or alert fatigue. Systems that generate plausible but incorrect language can be especially hazardous because textual output may carry an unwarranted sense of authority. The interaction design should therefore emphasize uncertainty, provenance, and the need for clinician verification. A generated report should be presented as a hypothesis rather than a conclusion, and the visual evidence supporting each statement should be retrievable.

The choice of deployment architecture also affects safety and maintainability. On-premise deployment may reduce concerns about patient data leaving the institution, but it requires local compute capacity and ongoing software maintenance. Cloud-based inference offers scalability but necessitates robust access controls and data residency guarantees. Hybrid models may allow preprocessing and encryption on site while offloading heavy inference to authorized cloud services. Latency requirements vary across clinical contexts. An emergency department may require rapid preliminary interpretation, while a subspecialty oncology workflow may tolerate longer analysis in exchange for more detailed multimodal reasoning. Mamba-based systems can potentially reduce inference latency relative to deep transformer stacks, enabling more responsive deployments even on modest hardware. The linear-time property is particularly relevant for three-dimensional imaging and long report sequences, where transformer attention might require aggressive patch reduction or truncation. Nonetheless, latency alone does not determine usability. Clinicians must be trained to interpret model outputs, and institutional protocols must define when a model may be used and how disagreements between the model and the clinician are resolved.

Sustainability is an often underestimated dimension of deployment. Medical AI systems consume energy during training, inference, and repeated retraining. Although state space models are more parameter efficient in some regimes, the aggregate environmental and financial cost depends on the frequency of model updates, the size of the serving infrastructure, and the useful lifespan of a deployed model. Reusing pretrained backbones and fine-tuning on local data can reduce the need for repeated large-scale training. Model distillation, quantization, and selective activation can further lower operational energy demand. Yet these techniques may interact with fairness and robustness in complex ways. For example, aggressive quantization might preserve average accuracy while degrading performance for rare findings or underrepresented groups. Life-cycle assessment should therefore accompany any sustainability claim. A system that is efficient on a single benchmark but requires constant recalibration may be less sustainable than a somewhat larger system that remains stable across distribution shifts.

6. Robustness, Fairness, and Policy Implications

Robustness in multimodal medical systems is closely linked to the alignment between visual and textual evidence. A vision-language model may correctly identify a finding in one image but fail when the imaging protocol changes, when contrast administration differs, or when a report uses institution-specific abbreviations. Distribution shift is pervasive in medical imaging, and models trained on research datasets may exhibit brittle behavior in community hospitals or international settings. In addition, medical text often carries contextual implications that are not directly visible in the image. A report may mention a finding that is absent, uncertain, or represented only in a prior study. The model must learn to distinguish current visual evidence from historical narrative context. Without explicit handling of temporal and causal structure, multimodal models may blend information across time in ways that undermine diagnostic reasoning. Evaluation protocols should include stress tests with corrupted images, varied scanners, and synthetic distribution shifts. The linear-time processing of Mamba does not inherently confer robustness, but its reduced memory footprint can make it feasible to train on larger and more diverse sequences, potentially improving generalization when paired with deliberate data curation.

Fairness concerns in vision-language medical models arise from several interacting sources. Imbalanced training data can result in underdiagnosis for minority populations, as demonstrated in studies of chest radiograph classifiers [14]. Label noise may be more prevalent in reports from underrepresented groups or from settings with less standardized documentation. The textual component can introduce additional bias because clinical language may vary by patient demographics, clinician style, and institutional norms. Even if an image encoder performs uniformly, the language decoder may generate reports that differ in tone or completeness across groups. Fairness evaluation should therefore move beyond aggregate accuracy to include subgroup-specific metrics for both classification and text generation. The choice of reference standard matters; a model may appear fair when evaluated against biased labels but unfair when evaluated against outcomes or expert consensus. The broader algorithmic fairness literature has shown how label bias can perpetuate disparities in health systems [16]. Vision-language Mamba models are not exempt from this dynamic, and their selective state updates do not automatically neutralize social bias encoded in training text.

Policy and regulatory implications follow from these technical and social constraints. Medical AI systems are increasingly subject to risk-based oversight, including requirements for

validation, documentation, and post-market monitoring. A vision-language model that generates clinical reports may be treated as a medical device in some jurisdictions, especially if its output is intended to support diagnostic decisions. Regulatory approval generally requires evidence of safety and effectiveness, but multimodal generative models pose distinct challenges because their output space is open-ended and context-dependent. Traditional performance metrics may fail to capture rare but serious errors. Policy frameworks must therefore encourage prospective evaluation in representative settings, continuous monitoring after deployment, and clear assignment of responsibility among developers, integrators, and clinical users. The concentration of foundation models in a small number of organizations raises additional governance concerns, including data access, pricing, and the ability of public institutions to audit model behavior. Open models and transparent evaluation resources can mitigate some risks but require institutional capacity to deploy and maintain.

Transparency and auditability are particularly important for state space models because their internal dynamics are less familiar to many clinicians and regulators than attention-based architectures. The selective state transition mechanism may be difficult to interpret directly. Explainability should therefore rely on multiple complementary approaches, including input attribution, counterfactual analysis, and retrieval of supporting evidence from the source image or report. Audit trails should record model version, input preprocessing, generated output, and any clinician modifications. Such records support retrospective review and enable institutions to detect systematic failures. In the broader policy landscape, the requirement for model cards, data sheets, and post-market surveillance aligns with a shift toward safety cases that articulate the intended use, residual risk, and monitoring plan. A vision-language Mamba system should not be deployed as a standalone diagnostic authority. Its role is better understood as one element of a distributed clinical reasoning process that includes humans, organizational protocols, and electronic health record systems.

7. Conclusion

Vision-language Mamba models offer a promising direction for multimodal medical image and clinical report understanding. Their linear-time state space processing can reduce the computational burden of handling long visual sequences and extended clinical narratives, making it feasible to train and deploy systems on high-resolution imaging data. However, the architectural advantages do not automatically address the deeper challenges of medical AI. Data governance, evaluation infrastructure, clinical workflow integration, robustness across populations, fairness, and regulatory accountability remain central to the success of these systems. This paper has argued for a system-level perspective in which architecture is only one component of a broader sociotechnical infrastructure. Efficient fusion mechanisms must be paired with careful dataset construction, transparent subgroup evaluation, and enduring institutional oversight. Cross-fusion state space methods for multi-view image classification [8] illustrate the potential of selective multimodal aggregation, but such technical developments must be connected to clinical evidence and governance structures to have meaningful impact. Future research should examine longitudinal multimodal reasoning, prospective validation in diverse healthcare settings, and standardized methods for monitoring model drift over time. The ultimate goal is not merely to generate fluent reports from images but to build sustainable, fair, and auditable systems that support clinicians in the complex work of patient care.

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. Advances in Neural Information Processing Systems, 30, 5998–6008.

2. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

3. Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In Proceedings of the 41st International Conference on Machine Learning. PMLR.

4. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In Proceedings of the 41st International Conference on Machine Learning. PMLR.

5. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model. In Advances in Neural Information Processing Systems, 37.

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, 139, 8748–8763. PMLR.

7. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In European Conference on Computer Vision (pp. 88–104). Springer.

8. 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.

9. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265.

10. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

11. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). Association for Computational Linguistics.

12. Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1), 317.

13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, 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(1), 590–597.

14. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182.

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. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

18. 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.

19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015–4026).