

Cross-Modal Mamba Networks for Multimodal Brain Tumor Diagnosis Using MRI and Clinical Data

Shanrui Deng

School of Computing, Clemson University, Clemson, SC, USA.
shanruiwork@clemson.edu

Hudson Peters

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.

hpeters@uab.edu

Jordan J. Wilson

Department of Computer Science, University of Houston, Houston, TX, USA.
jordanmail@uh.edu

Abstract

Multimodal brain tumor diagnosis requires the joint interpretation of heterogeneous magnetic resonance imaging sequences and non-imaging clinical variables, yet most diagnostic systems remain constrained by unimodal processing or by attention mechanisms whose computational costs grow rapidly with sequence length. This paper examines cross-modal Mamba networks as a system-level alternative for multimodal brain tumor diagnosis, focusing on architectural design, data infrastructure, clinical robustness, governance, fairness, deployment, and sustainability rather than isolated accuracy metrics. Selective state space sequence modeling enables linear-time processing of long multimodal token sequences, making it possible to align MRI-derived visual tokens with variable-length clinical text without the quadratic memory burden of conventional attention. The paper analyzes structural trade-offs in cross-modal fusion, including the ordering of modality encoders, fusion depth, state expansion, and the management of missing clinical data. It further discusses the institutional and regulatory conditions needed for safe translation, including federated learning, demographic fairness auditing, post hoc interpretability, prospective validation, and energy-conscious deployment. By treating cross-modal Mamba architectures as socio-technical infrastructures, the discussion connects computational design choices to broader concerns of clinical accountability, data governance, and health system sustainability. The paper argues that the value of such architectures lies not only in their efficiency but also in their capacity to support auditable, robust, and equitable multimodal diagnosis across heterogeneous hospital environments.

Keywords

cross-modal learning; selective state space models; brain tumor diagnosis; MRI; clinical data; system governance.

1. Introduction

Brain tumor diagnosis is an intrinsically multimodal clinical task that requires the integration of structural and functional information from multiparametric magnetic resonance imaging, including T1-weighted, T1-contrast-enhanced, T2-weighted, and fluid-attenuated inversion

recovery sequences, together with clinical variables such as age, performance status, molecular markers, and treatment history. Despite substantial progress in medical image analysis, many deployed diagnostic systems remain unimodal or rely on separate pipelines that do not learn joint representations from imaging and clinical text. This separation limits the ability to capture correlations between radiological appearance and patient-level clinical context, which are often essential for distinguishing tumor grade, predicting progression, and guiding intervention. Recent developments in selective state space sequence modeling have introduced linear-time sequence mixing [1], and their visual counterparts have demonstrated competitive representation learning with bidirectional scanning across image patches [2]. These methods offer a compelling structural alternative to conventional attention mechanisms [3] and vision transformers [4], particularly when processing high-resolution volumetric MRI and long variable-length clinical strings.

Public benchmark collections such as the Multimodal Brain Tumor Image Segmentation Benchmark [5] and The Cancer Genome Atlas glioma collections [6] have driven much of the progress in automated tumor segmentation and radiogenomic analysis. These resources provide expert annotations and molecular labels that support supervised learning, but they also reflect curated conditions that are often distant from routine clinical workflows. Real-world diagnosis must contend with heterogeneous scanners, missing sequences, incomplete clinical records, institutional documentation differences, and variable acquisition protocols. Under these conditions, the central challenge is not merely to maximize segmentation overlap but to build a cross-modal system that can operate reliably within constrained computational resources, protect patient privacy, and produce decisions that can be explained and audited. This paper therefore adopts a systems perspective, analyzing cross-modal Mamba networks for multimodal brain tumor diagnosis as an integrated computational and organizational infrastructure. The discussion emphasizes structural trade-offs, data governance, fairness, robustness, and deployment feasibility rather than isolated benchmark performance.

2. Related Work and Technical Context

Medical image computing has long been shaped by convolutional architectures that exploit local spatial regularities. Early fully convolutional segmentation pipelines were substantially improved by self-configuring frameworks such as nnU-Net [7], which automate many preprocessing, architecture, and postprocessing decisions that previously required expert tuning. Nested architectures with deep supervision and multiscale skip pathways [8] further improved feature reuse across resolutions. These systems build on foundational residual learning [9] and dense connectivity [10], which remain influential because they stabilize gradient flow and encourage feature reuse in deep networks. A broad survey of deep learning in medical image analysis [11] identifies multimodal fusion, data heterogeneity, and interpretability as persistent obstacles across applications. In brain tumor imaging, these obstacles are especially acute because the relevant diagnostic signal is distributed across multiple MRI contrasts and non-imaging clinical variables, and because the tumor microenvironment cannot be summarized by a single imaging feature.

A cross-fusion Mamba formulation has recently been proposed for multi-view medical image classification [12], demonstrating that selective state space models can be used to align heterogeneous views without relying on quadratic attention. This approach illustrates the broader potential of state space methods for medical image understanding, particularly when the input consists of multiple related but non-identical views. In the context of brain tumor diagnosis, the cross-modal problem is more demanding than multi-view fusion because the

imaging and clinical modalities differ in representational form, temporal resolution, missingness patterns, and semantic granularity. Nevertheless, the core insight is transferable: state space sequence mixing can integrate heterogeneous information in a memory-efficient manner. The present paper extends this conceptual direction by examining how cross-modal Mamba architectures can be embedded within a full diagnostic infrastructure.

3. Cross-Modal Mamba Architecture and System Design

A cross-modal Mamba network for brain tumor diagnosis can be understood as a sequence-to-sequence system in which MRI volumes and clinical records are converted into token sequences and processed by shared or modality-specific selective state space layers. The MRI branch typically divides each volume into patches or slabs, embeds them through a convolutional or linear projection, and then applies bidirectional Mamba blocks to capture local and long-range spatial dependencies. The clinical branch tokenizes demographic, laboratory, molecular, and symptom information into a compact sequence that may vary in length across patients. The two branches are then aligned through cross-fusion layers that allow imaging tokens to attend to clinical summaries and clinical tokens to attend to imaging features. Unlike conventional cross-attention, which incurs quadratic memory costs with respect to sequence length, cross-modal Mamba layers preserve linear complexity by maintaining and updating a compact hidden state as they scan the fused sequence.

This architectural choice creates several structural trade-offs. First, the order in which modalities are scanned affects how information propagates through the hidden state. Scanning imaging tokens before clinical tokens may encourage the clinical branch to refine radiological features, while the reverse order may allow imaging features to be interpreted in the context of patient-level variables. In practice, multiple scan directions and repeated interleaving can mitigate order sensitivity, but this increases the number of parameters and the complexity of gradient flow. Second, the depth at which fusion occurs matters. Early fusion allows low-level imaging features and raw clinical variables to interact directly, which may be useful when molecular markers correlate with subtle radiological patterns, but it also risks propagating noisy or missing clinical values into the imaging stream. Late fusion preserves modality-specific feature extraction but may fail to learn fine-grained cross-modal dependencies. A hybrid design with intermediate fusion at multiple levels offers a compromise, though it requires careful regularization and validation.

A further structural consideration is the management of missing data. Clinical records are frequently incomplete because laboratory tests are ordered selectively, molecular profiling is not always available, and follow-up documentation varies across institutions. A sequence model can represent missing values as learned embeddings or mask tokens, but the system designer must decide whether missingness should be treated as informative or neutral. Treating missingness as informative may improve predictions when missing values reflect clinical decision-making, but it can also encode institutional differences and exacerbate bias. Cross-modal Mamba architectures can support missingness-aware fusion by allowing the clinical branch to condition its state updates on observed variables, yet this increases the complexity of the system and demands careful auditing. These design choices illustrate that architectural decisions are not purely technical; they shape how the system will behave under real-world data incompleteness.

4. Data Infrastructure and Multimodal Preprocessing

The successful deployment of cross-modal Mamba networks depends on data infrastructure that can harmonize MRI acquisitions and clinical records while preserving institutional privacy and regulatory compliance. MRI preprocessing typically includes skull stripping, intensity normalization, resampling to a common anatomical space, and tumor region cropping. Clinical data preprocessing involves entity normalization, missing value handling, and conversion of unstructured text into structured variables. These steps introduce a tension between standardization and information loss. Aggressive normalization can remove scanner-specific artifacts but may also erase subtle signal related to acquisition differences that are predictive of site or population. Insufficient normalization can leave the model vulnerable to domain shift when it encounters a new hospital or scanner vendor. A robust infrastructure must therefore support provenance tracking, preprocessing versioning, and continuous evaluation of data drift.

Privacy-preserving infrastructure is particularly important because brain tumor datasets are relatively small and often distributed across institutions. Federated learning allows multiple hospitals to jointly train a model without consolidating raw patient data [13]. In a cross-modal Mamba setting, federated training introduces additional complexity because the imaging and clinical branches may be updated at different frequencies, and local clinical vocabularies may differ across sites. Secure aggregation, differential privacy, and model-sharing agreements are necessary governance mechanisms, but they also influence architectural decisions. For example, if clinical text cannot leave a local institution due to privacy constraints, the clinical encoder may need to be trained locally while only aggregated gradients or image-derived representations are shared. This constraint may favor modular cross-modal architectures in which modality-specific encoders can be trained under heterogeneous privacy regimes.

5. Clinical Validation, Robustness, and Explainability

Translating cross-modal Mamba networks into clinical workflows requires evidence beyond retrospective benchmark performance. Recent analyses of artificial intelligence in health and medicine [14] emphasize prospective validation, workflow integration, and continuous monitoring as essential components of responsible clinical deployment. Retrospective studies on public benchmarks are valuable for initial development, but they often overestimate real-world utility because curated datasets exclude ambiguous cases, low-quality scans, and incomplete records. Prospective evaluation in a clinical setting reveals how the system interacts with radiologists, whether it changes decision confidence, and whether it introduces automation bias. Cross-modal systems are especially difficult to validate because clinical variables may affect predictions in ways that are not visible in the imaging data alone. This necessitates evaluation protocols that measure not only overall accuracy but also performance conditional on clinical subgroups and missing data patterns.

Robustness and explainability are tightly linked in clinical adoption. Post hoc visual explanation methods such as gradient-weighted class activation mapping [15] can highlight suspicious regions in MRI scans, while additive feature attribution methods [16] can estimate the contribution of individual clinical variables to the final prediction. These tools provide partial transparency, but they do not fully explain the dynamics of a cross-modal state space model. Because Mamba layers maintain a hidden state that evolves over the sequence, the contribution of a particular imaging region may depend on the preceding clinical tokens. This makes static attribution incomplete. Clinicians may require counterfactual explanations that show how a prediction would change if a clinical variable were altered or if a particular MRI sequence were removed. Designing such explanations demands additional infrastructure and

may influence model architecture, for example by encouraging modular fusion paths that can be ablated during explanation generation.

6. Governance, Fairness, and Policy Implications

Fairness in multimodal brain tumor diagnosis requires systematic auditing across demographic subgroups, scanner types, and contributing institutions. Bias may enter through unbalanced training data, differences in MRI acquisition, variable clinical documentation, or the use of proxies such as insurance codes and hospital location. Fairness surveys in machine learning [17] distinguish between different definitions of fairness and warn that optimizing one fairness criterion can violate another. In clinical settings, the relevant fairness criteria should be negotiated with clinicians, patients, and regulators rather than imposed as a single mathematical constraint. Cross-modal models may inadvertently learn institutional identity from clinical text or scanner artifacts, and these shortcuts can produce spurious correlations that fail when the model is deployed in a new population. Regular auditing, subgroup analysis, and external validation across geographically distinct sites are therefore essential governance practices.

Ethical oversight must also address accountability, consent, and the distribution of responsibility among developers, hospitals, and regulatory bodies. Machine learning in medicine raises ethical challenges related to transparency, liability, and the protection of patient autonomy [18]. When a cross-modal model contributes to diagnosis, it may influence treatment decisions even if the model is formally advisory. This makes it necessary to define the role of human clinicians, the conditions under which model outputs can be overridden, and the processes for reviewing adverse events. Big data approaches in health care [19] have demonstrated the benefits of linking large clinical repositories, but they also create risks related to re-identification, secondary use, and algorithmic discrimination. Governance frameworks should require clear documentation of training data, preprocessing choices, model limitations, and postmarket surveillance plans before cross-modal diagnostic systems are integrated into routine care.

7. Deployment, Sustainability, and Future Directions

Deployment of cross-modal Mamba networks in hospital environments involves constraints related to computational hardware, interoperability with electronic health record systems, and long-term maintenance. Although selective state space models reduce memory requirements compared with attention-based systems, they still demand specialized processing for three-dimensional MRI volumes, particularly when multiple sequences are processed jointly. Model compression, knowledge distillation, and selective sequence pruning can reduce inference costs, but these strategies must be validated to ensure that they do not disproportionately degrade performance for rare tumor subtypes or underrepresented patient groups. Sustainability also includes energy consumption, hardware refresh cycles, and the organizational capacity to retrain or fine-tune models as clinical guidelines evolve. A model that is accurate at initial deployment but cannot be updated without costly infrastructure changes may become obsolete or unsafe.

Clinical data infrastructures such as MIMIC-III [20] illustrate the value of linked multimodal repositories for secondary analysis, but brain tumor diagnosis requires tighter integration between imaging archives, pathology systems, and longitudinal outcome registries. Future cross-modal systems may need to incorporate molecular pathology, surgical reports, and serial imaging to support treatment response assessment. This expansion increases the

complexity of sequence modeling and data governance, but it also creates opportunities for state space models that can process long patient trajectories efficiently. The long-term value of cross-modal Mamba networks will depend less on incremental benchmark improvements than on their compatibility with federated learning, auditability, modular updating, and resource-limited clinical environments. Research should therefore prioritize open evaluation protocols, standardized reporting of data processing, and prospective studies that measure clinical utility rather than only technical performance.

8. Conclusion

Cross-modal Mamba networks represent a promising architectural direction for multimodal brain tumor diagnosis, offering a memory-efficient mechanism for integrating MRI-derived visual features with clinical variables. However, their clinical value cannot be reduced to sequence-model efficiency or segmentation accuracy. The design of these systems involves structural choices about fusion order, missing data representation, privacy-preserving training, and interpretability that have direct consequences for robustness, fairness, and accountability. When embedded within appropriate data infrastructure and governance frameworks, such models can support more integrated and auditable diagnostic workflows. When deployed without attention to institutional heterogeneity and ethical oversight, they may reproduce or amplify existing disparities. A systems-level perspective is therefore essential. The future development of cross-modal Mamba networks should be guided by prospective clinical validation, federated governance, subgroup fairness auditing, energy-conscious deployment, and transparent documentation of limitations. Under these conditions, selective state space architectures can contribute meaningfully to the long-term goal of safe, equitable, and sustainable multimodal diagnosis in neuro-oncology.

References

1. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.
2. 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.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
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. *International Conference on Learning Representations*.
5. Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M. A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993-2024.
6. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma

MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 170117.

7. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203-211.
8. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3-11). Springer.
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.
10. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708.
11. 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.
12. 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.
13. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 119.
14. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38.
15. 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. *Proceedings of the IEEE International Conference on Computer Vision*, 618-626.
16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
17. 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.
18. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), e1002689.
19. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318.
20. 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.