Mamba-Based Multimodal Fusion for Robust Medical Image Segmentation and Lesion Delineation

Casper Rhodes
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
casperr@uc.edu

Brooks Vega
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
vega440@ucf.edu

Abstract

Medical image segmentation and lesion delineation are essential tasks in diagnostic imaging, oncology, and image-guided intervention, but their clinical utility depends on robustness across imaging modalities, acquisition sites, patient populations, and lesion types. Recent advances in state space modeling, particularly the Mamba selective scan mechanism, offer a promising alternative to convolutional and transformer-based segmentation pipelines by combining long-range dependency modeling with improved computational scaling. When embedded within multimodal fusion architectures, Mamba-based models can integrate complementary imaging sequences while preserving boundary detail and reducing memory demands. However, the design of such systems extends well beyond model architecture. This paper presents a system-level examination of Mamba-based multimodal fusion for segmentation and lesion delineation, focusing on structural trade-offs, robustness, uncertainty, data governance, fairness, infrastructure, deployment, sustainability, and policy implications. We argue that clinical value emerges not from segmentation accuracy alone but from the alignment of model design with institutional workflows, regulatory requirements, and sociotechnical safeguards. The discussion integrates architectural analysis with cross-domain comparisons and forward-looking perspectives on federated evaluation, resource-aware deployment, and accountable clinical integration. The result is a comprehensive framework for developing and deploying multimodal segmentation systems that are technically capable, operationally sustainable, and ethically defensible.

Keywords

medical image segmentation; multimodal fusion; state space models; lesion delineation; robustness; clinical AI governance; sustainable deployment.

1. Introduction

Medical image segmentation and lesion delineation are foundational tasks in diagnostic imaging, radiation oncology, and image-guided intervention. The field has shifted from handcrafted feature extractors to deeply supervised architectures, as evidenced by broad surveys of deep learning in medical image analysis [1]. The U-shaped convolutional architecture introduced in earlier work established a durable design pattern based on encoder-decoder symmetry and skip connections [2]. Self-configuring pipelines such as nnU-Net further demonstrated that careful preprocessing, data augmentation, and postprocessing are often as important as architectural novelty [3]. These systems produce high-quality segmentations when training and deployment distributions are closely aligned, yet their

behavior can degrade under acquisition heterogeneity, rare lesion morphologies, and ambiguous boundaries.

The introduction of transformer mechanisms [4] and vision transformers [5] changed the design space by enabling global context aggregation. Architectures such as UNETR [6] applied transformer-based representations to volumetric medical data, improving long-range dependency modeling at the cost of increased memory and computational demand. More recently, state space models and the selective scan mechanism of Mamba [7] have offered an alternative that scales more gracefully with sequence length. In multimodal medical image analysis, cross-fusion designs have explored how state space representations can integrate complementary views [8]. Although that work targets classification rather than segmentation, it illustrates a broader shift toward selective state space models for heterogeneous medical inputs.

This paper examines the system-level implications of Mamba-based multimodal fusion for segmentation and lesion delineation. Rather than presenting a single benchmark result, we analyze architectural trade-offs, robustness behavior, data governance, infrastructure requirements, sustainability, fairness, and policy integration. The objective is to provide a coherent design perspective that can guide translational research and hospital deployment while acknowledging that segmentation quality alone is insufficient for clinical value.

2. Background and Related Work

Early segmentation systems relied on convolutional inductive biases and localized filters, which made them efficient but limited in capturing distant anatomical relationships. The U-Net architecture [2] addressed localization through skip connections, and nnU-Net [3] demonstrated that adaptive configuration of preprocessing and topology often outperforms bespoke designs. These methods remain important baselines because they are well understood, relatively easy to validate, and compatible with hospital compute environments. However, they can struggle with subtle boundaries, heterogeneous contrast, and integration of multiple imaging sequences.

Transformer-based segmentation methods use self-attention to model global context. The original attention mechanism [4] and vision transformer [5] established the conceptual foundation, while UNETR [6] adapted these ideas to three-dimensional medical volumes. FlashAttention [9] later reduced the memory footprint of exact attention through input-output-aware scheduling, but the quadratic complexity of attention remains a structural concern for very large volumes or high-resolution multimodal inputs. These methods often require extensive pretraining or large datasets to fully exploit their representational capacity.

State space models provide a complementary approach. Mamba [7] introduced a selective state space mechanism that can capture long-range dependencies without the quadratic growth associated with full self-attention. Visual adaptations such as VMamba [10] and segmentation-oriented models such as SegMamba [11] have begun to translate these benefits to imaging tasks. Foundation models for image segmentation, including Segment Anything [12], have also changed expectations about generalization and prompt-based interaction. Mamba-based multimodal fusion can potentially combine the global receptive field of state space models with the boundary precision of U-shaped convolutional decoders.

3. Mamba-Based Multimodal Fusion Architecture

A Mamba-based multimodal fusion architecture for segmentation can be organized into modality-specific tokenization, selective state space encoding, cross-modal interaction, and decoder reconstruction. Tokenization can convert two-dimensional slices or volumetric patches into sequences while preserving positional information. Modality-specific encoders then apply selective scanning along biologically meaningful directions, allowing the model to emphasize salient regions rather than treating all states uniformly. This selective behavior is central to Mamba [7] and distinguishes it from earlier state space models that used fixed transition rules.

Multimodal fusion requires careful alignment between imaging sequences such as T1-weighted, T2-weighted, fluid-attenuated inversion recovery, computed tomography, or positron emission tomography. A cross-fusion state space design can share or exchange hidden state representations after each block, enabling one modality to modulate the scanning behavior of another. The cross-fusion perspective described in [8] is relevant because it shows how separate state space streams can be combined without simply concatenating feature maps at the input. For segmentation, cross-modal gating can help resolve ambiguity in lesions that are conspicuous in one modality but poorly defined in another.

The architecture must also balance sequence length, scan direction, and positional encoding. Visual Mamba models [10] have explored bidirectional scanning to capture spatial context, while SegMamba [11] has applied long-range sequential modeling to volumetric inputs. In multimodal systems, the number of scans grows with the number of modalities, so memory consumption and state propagation become structural design parameters. Unlike conventional attention, which computes pairwise interactions among tokens, Mamba-style selective scanning maintains a recurrent state that can be updated efficiently. This makes it possible to process multimodal token streams with lower asymptotic complexity, although implementation efficiency depends on hardware-aware kernels such as those developed for efficient attention [9].

Decoder design remains important for lesion delineation. A U-shaped decoder with skip connections from modality-specific and fused encoders can recover fine spatial detail. Boundary refinement can be supported by auxiliary supervision on edge maps or distance transforms. The system-level trade-off is that deeper fusion increases representational capacity but complicates validation, interpretability, and failure analysis. A modular architecture with clearly separated modality encoders, a fusion bottleneck, and a shared decoder offers a pragmatic balance between model flexibility and clinical auditability.

4. Robustness and Uncertainty in Lesion Delineation

Robustness in medical image segmentation is not reducible to average overlap metrics. Acquisition protocols, scanner vendors, reconstruction kernels, patient motion, contrast timing, and lesion rarity all introduce distributional variation. A Mamba-based multimodal system can improve robustness by learning selective representations that ignore irrelevant background and focus on lesion-relevant context. However, this selectivity can also become a vulnerability if the model learns spurious correlations with scanner markers or acquisition artifacts. Robustness must therefore be evaluated through external validation, stratified subgroup analysis, and stress testing under realistic perturbations.

Uncertainty estimation is essential because lesion boundaries are often inherently ambiguous. Inter-reader variability can be large for infiltrative tumors, ground-glass opacities, or inflammatory lesions. Segmentation models that output deterministic boundaries may create a

false impression of certainty. A system-level design should support multiple samples, test-time augmentation, and confidence surfaces that highlight areas where the model and multiple human readers are likely to disagree. Mamba-based architectures can be integrated with Monte Carlo dropout or deep ensembles, but these strategies increase computational cost and must be weighed against latency requirements.

Explainability tools provide another layer of robustness assessment. Gradient-based localization methods such as Grad-CAM [13] and additive feature attribution methods such as SHAP [14] can reveal whether the fusion system attends to clinically meaningful regions. These tools are not sufficient for regulatory acceptance, but they can help detect failures related to out-of-distribution inputs or biased attention. For multimodal fusion, explanations should be reported per modality and for the fused representation, because a model may appear reasonable in aggregate while relying on an incorrect modality for specific patient subgroups.

Calibration under label noise and referral bias is a persistent challenge. Segmentation datasets often contain labels from a single institution or annotator, and lesion boundaries may be systematically narrower or wider than clinical practice. A robust deployment pipeline should include label quality audits, reader studies, and periodic recalibration. Multimodal fusion can reduce ambiguity by leveraging complementary modalities, but only if the fusion process is conditioned on data quality and not merely on the presence of an input channel.

5. Data Governance, Fairness, and Bias Mitigation

Medical imaging data governance must address privacy, provenance, consent, and access control. Machine learning systems inherit biases from population selection, annotation practices, and equipment availability. A landmark study [15] demonstrated that algorithmic bias can arise from using proxy outcomes that encode structural disadvantage rather than biological need. In segmentation, similar mechanisms can produce models that perform differently across demographic groups, imaging sites, or disease subtypes. Governance frameworks should treat training data as a sociotechnical artifact rather than a neutral input.

Data preparation and curation are critical infrastructure. The medical imaging community has developed recommendations for organizing, cleaning, and sharing imaging data for machine learning [16]. These practices include standardized metadata, scanner harmonization, de-identification, and clear data use agreements. Without such practices, multimodal fusion models may learn site-specific shortcuts that fail when deployed in new hospitals. Federated learning offers a governance-compatible framework for training across institutions without centralizing raw images [17]. Mamba-based segmentation models can be adapted to federated settings, but communication costs and state space parameter synchronization must be carefully designed.

Fairness in lesion delineation requires more than equal overall accuracy. Clinically meaningful subgroups defined by age, sex, organ size, comorbidity, and imaging characteristics can exhibit different error patterns. Evaluation should include subgroup-specific boundary errors, false positive and false negative rates, and downstream effects on volume estimation or treatment planning. A multimodal model may improve sensitivity for small lesions but also increase false positives in benign findings. Bias mitigation can include stratified sampling, adversarial domain alignment, reweighting, and prospective validation on underrepresented populations.

Transparency and accountability are central to responsible deployment. Reviews of machine learning in health [18] and analyses of algorithmic challenges [19] emphasize that model

developers must document intended use, data limitations, and performance conditions. Ethical frameworks for machine learning in medicine [20] similarly call for institutional oversight, patient engagement, and mechanisms for redress. Segmentation systems that inform radiotherapy or surgical planning are particularly high stakes, because boundary errors can translate directly into dose distribution changes or incomplete resection.

6. Infrastructure and Deployment Considerations

Deployment of Mamba-based multimodal fusion systems requires a robust computational infrastructure that spans data ingestion, model inference, visualization, and electronic health record integration. Unlike research environments, clinical systems must operate continuously, handle incomplete or corrupted inputs, and produce results within clinically acceptable latencies. Containerization, versioned model registries, and audit logs are necessary to ensure reproducibility. The infrastructure should also support graceful fallback when one modality is unavailable, because patients may not always receive a full imaging protocol.

Interoperability with picture archiving and communication systems, radiology information systems, and treatment planning software is a major engineering challenge. Segmentation outputs must be converted into standard formats and displayed in a way that clinicians can review and correct. The user interface should permit boundary editing rather than presenting the model as an unalterable authority. Governance operations should track model version, input quality flags, and clinician corrections to feed back into model maintenance.

Continuous monitoring is essential after deployment. Model performance can drift as scanner software updates, acquisition protocols change, or patient populations shift. A deployment architecture should include statistical process control on input distributions and segmentation quality indicators. Mamba-based models, like other deep learning systems, require periodic retraining or recalibration, which must be coordinated with institutional data governance and regulatory requirements. Federated learning [17] can facilitate collaborative monitoring without moving protected health information.

7. Sustainability and Resource-Aware System Design

Sustainability in medical AI includes energy consumption, hardware lifecycles, dataset reuse, and long-term maintenance burden. Transformer-based segmentation models can be computationally demanding, especially when applied to high-resolution volumetric images. Mamba-based state space models [7] offer a structural advantage in asymptotic complexity, but realized energy savings depend on hardware-aware implementation and batch scheduling. The selective scanning mechanism can reduce unnecessary computation by focusing on salient regions, yet careful ablation is needed to confirm that these savings are not offset by specialized kernel requirements.

Resource-aware design should consider the entire lifecycle from training to inference. Large multimodal models may require distributed training, but a more compact deployment model can be distilled or pruned once the fusion logic has stabilized. Efficiency innovations such as FlashAttention [9] demonstrate that hardware-aware redesign can reduce memory traffic and improve throughput. Similar engineering effort for state space models can make them practical for hospital clusters and edge devices. Sustainability also includes data storage, because multimodal imaging datasets are large and often duplicated across institutions and research groups.

A mature sustainability strategy should prioritize modularity and reuse. Shared encoders, standardized preprocessing pipelines, and openly documented interfaces reduce duplicated effort. Dataset documentation and model cards can help future teams understand whether retraining is necessary. Additionally, model compression and quantization can lower the carbon footprint of inference, but only if they do not disproportionately degrade performance on rare lesion subtypes or underrepresented populations. This is a structural trade-off that must be explicitly governed.

8. Clinical Integration and Policy Implications

Clinical integration of automated lesion delineation requires alignment with workflow, professional roles, and evidentiary standards. A segmentation model that produces high overlap scores but requires excessive manual correction may reduce rather than improve radiologist productivity. The system should be evaluated through human-computer interaction studies that measure correction time, inter-reader agreement, and clinical decision impact. Regulatory pathways increasingly require prospective validation and post-market surveillance for artificial intelligence systems [18].

Policy implications extend to reimbursement, liability, and data sovereignty. If a segmentation error contributes to a missed diagnosis or inappropriate radiation plan, responsibility may be distributed across the vendor, integrating institution, and supervising clinician. Clear documentation of model limitations and clinician override authority is necessary. Ethical guidance [20] suggests that algorithmic tools should augment rather than replace professional judgment. Institutions should establish multidisciplinary governance committees that include radiologists, pathologists, oncologists, data scientists, and patient representatives.

The fairness and robustness requirements discussed earlier should be formalized as policy instruments. Procurement criteria for segmentation software can include external validation across multiple sites, subgroup performance reports, uncertainty outputs, and interoperability standards. Regulators may require audit trails that capture model version, input quality, and clinician modifications. These mechanisms are especially important when segmentation outputs influence radiation dose volumes, surgical margins, or drug response assessment. Prospective registries and federated evaluation networks [17] can provide ongoing evidence about performance in diverse populations.

9. Conclusion

Mamba-based multimodal fusion offers a promising architectural direction for robust medical image segmentation and lesion delineation. It combines the scalability of selective state space models with the complementary information available from multiple imaging modalities. This paper has argued that the technical merits of such systems cannot be separated from their organizational, ethical, and policy contexts. Architecture choices interact with data governance, fairness, infrastructure, sustainability, and clinical workflow in ways that determine whether performance improvements produce real patient benefit.

Future work should develop standardized robustness benchmarks, uncertainty-aware evaluation protocols, and federated training frameworks for Mamba-based segmentation systems. Research should also compare state space multimodal fusion against transformer and convolutional baselines across diverse clinical tasks, including tumor volumetry, stroke lesion quantification, and radiotherapy target delineation. By treating segmentation as a

sociotechnical system, the field can move beyond incremental metric gains toward safer, more equitable, and more sustainable clinical artificial intelligence.

References

1. 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.

2. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234–241). Springer.

3. 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.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

5. 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.

6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. arXiv preprint arXiv:2103.10504.

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

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. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Advances in Neural Information Processing Systems, 35, 16344–16359.

10. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model. arXiv preprint arXiv:2401.10166.

11. Xing, Z., Ye, T., Yang, Y., Liu, G., & Zhu, L. (2024). SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. arXiv preprint arXiv:2401.13560.

12. 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. arXiv preprint arXiv:2304.02643.

13. 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).

14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

15. 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.

16. Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. Radiology, 295(1), 4–15.

17. 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.

18. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

19. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689.

20. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.