

Self-Supervised Mamba Representation Learning for Generalizable Medical Image Analysis

Stanley Carlson

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
stanley1998@colostate.edu

Aamir Murthy

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

Ralph Zimmerman

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
contactralph@uc.edu

Abstract

The increasing availability of large-scale medical imaging archives has created a pressing need for representation learning methods that remain clinically meaningful when supervision is sparse, domain shifts emerge, and deployment settings diverge from training conditions. This paper examines self-supervised Mamba representation learning as a system-level framework for generalizable medical image analysis. Rather than focusing narrowly on architectural novelty, the discussion addresses how selective state-space models, self-supervised pretraining objectives, and medical imaging workflows interact across compute infrastructure, data governance, robustness assessment, and clinical integration. The paper analyzes the structural trade-offs between transformer-based attention mechanisms and linear-time state-space sequence modeling, emphasizing the implications for high-resolution two-dimensional images, volumetric scans, and multi-view diagnostic protocols. It further considers how pretraining datasets, augmentation policies, and institutional data-sharing agreements shape representation generality and fairness. The analysis extends to deployment concerns, including model updating, energy consumption, interpretability, regulatory alignment, and the socio-technical conditions under which representation learners become durable clinical infrastructure. By approaching self-supervised Mamba models as embedded components of larger learning systems, the paper provides a cross-disciplinary perspective on their potential to improve generalization, reduce annotation burdens, and support equitable medical image analysis while exposing governance and sustainability requirements that remain underexamined in the literature.

Keywords

self-supervised learning; state-space models; medical image analysis; representation learning; clinical deployment; algorithmic fairness; infrastructure governance.

1. Introduction

Medical image analysis has become one of the most consequential application areas for representation learning because diagnostic decisions increasingly depend on the ability to derive stable features from complex anatomical and pathological patterns. Supervised deep

learning has produced impressive results for narrow tasks, yet its dependence on large quantities of expert-annotated data remains a fundamental constraint. Self-supervised contrastive methods such as momentum contrast [1] and simple contrastive learning frameworks [2] demonstrated that strong visual representations can emerge without task-specific labels. These approaches learn by pulling augmented views of the same instance together while separating distinct instances, thereby encoding invariance to nuisance transformations that often destabilize medical imaging models. However, the organizational requirements of contrastive learning at scale, including negative sample management, augmentation design, and large-batch optimization, have proven difficult to transfer directly into hospital and research imaging environments.

Transformer architectures extended representation learning by introducing global attention over image patches [3], building on the broader attention mechanism [4]. Transformers are highly expressive but impose substantial memory and computation costs as sequence length grows. This is particularly problematic in medical imaging because volumetric scans and whole-slide pathology images can produce extremely long token sequences. Self-supervised transformer pretraining therefore often requires careful patch sampling, resolution reduction, or distributed training strategies that may sacrifice diagnostically relevant detail. The system-level consequence is that high-capacity attention models can be technically excellent in controlled benchmarks while remaining difficult to operate in settings with limited computational resources or strict latency requirements.

State-space models offer an alternative sequence modeling path that maintains global context aggregation while avoiding quadratic attention complexity. The selective state-space formulation introduced by Mamba [5] enables input-dependent gating of information flow across long sequences, and visual extensions such as VMamba [6] have adapted this principle to two-dimensional image encoding through scanning strategies. These models introduce different system trade-offs: they may require careful token ordering, bidirectional scanning, and multi-scale aggregation to approximate the spatial reasoning that attention provides naturally. This paper argues that self-supervised Mamba representation learning should be understood not as a direct replacement for convolutional or attention-based systems, but as a candidate component within a broader infrastructure that includes data curation, training orchestration, governance mechanisms, deployment monitoring, and fairness auditing. Such a framing is necessary because generalization in medical image analysis is not solely an architectural property; it is produced by an alignment among modeling assumptions, training data, clinical workflows, and institutional policies.

2. Background and Related Work

Recent adaptations of Mamba to medical image segmentation illustrate the promise and complexity of state-space modeling in diagnostic contexts. SegMamba [7] applies long-range sequential modeling to three-dimensional medical images, demonstrating that selective state propagation can capture contextual dependencies across volumetric regions with reduced memory requirements. U-Mamba [8] similarly integrates state-space layers into a biomedical segmentation framework, aiming to balance local precision with global coherence. These architectures tend to be evaluated under supervised regimes where segmentation masks define the optimization target. Their success indicates that state-space models can process large medical inputs efficiently, but it leaves open the question of how such models should be pretrained when annotated masks are unavailable or sparse. The present work extends this line

of inquiry by considering self-supervised objectives as a precondition for downstream generalization.

Before the rise of state-space models, medical image analysis was dominated by convolutional architectures and self-configuring pipelines. The nnU-Net framework [9] systematized preprocessing, topology selection, and postprocessing for biomedical segmentation, and its organizing logic can be traced to the original U-Net [10]. These approaches encode strong spatial priors and remain highly effective for many tasks, yet their representations are typically trained from scratch and do not exploit unlabeled images. Self-supervised medical imaging research subsequently borrowed from natural image pretraining, with large-scale contrastive learning on chest radiographs showing that representations learned without diagnostic labels can transfer to classification tasks [11]. Restorative pretraining in volumetric medical imaging further demonstrated that models can learn anatomical structure by reconstructing corrupted regions [12]. Public benchmark systems such as CheXNet [13] and CheXpert [14] provided important evaluation substrates for chest radiograph interpretation, but they also exposed persistent challenges related to label noise, domain shift, and the limited representation of rare conditions.

These developments reveal a recurring pattern in medical image analysis: architectural progress usually precedes systematic attention to data governance, representation evaluation, and deployment reliability. Self-supervised Mamba models inherit this pattern while also introducing state-space-specific concerns. Token scanning order, patch selection, and bidirectional fusion are not merely implementation details; they shape the inductive biases that determine which anatomical relationships are preserved during pretraining. Therefore, a rigorous system-level account must connect architectural choices to the institutional and clinical conditions under which models are trained, audited, and maintained.

3. System Architecture and Representation Learning Design

A self-supervised Mamba representation learning system can be decomposed into several interacting components: an input tokenization layer, a selective state-space encoder, a projection head for pretraining objectives, and a downstream task interface. The tokenization layer determines how raw medical images are converted into sequences, with options including fixed patch grids, region-of-interest crops, and multi-scale feature pyramids. In volumetric imaging, patch ordering across axial, coronal, and sagittal planes has a strong influence on representation quality because state-space models process tokens sequentially. Unlike attention, which can compare any pair of patches directly, selective state-space propagation relies on ordered recurrence. Bidirectional scanning and cross-scan fusion can mitigate this limitation but increase architectural complexity and require additional memory bandwidth during training.

The pretraining objective is equally important. Contrastive objectives encourage invariance to augmentations such as rotation, cropping, color perturbation, and noise injection, but these augmentations may not align with clinically meaningful variations. Restorative objectives, including masked patch reconstruction or inpainting, may be better suited to anatomical data because they force models to recover structural information. Hybrid objectives can combine contrastive and restorative signals, but they introduce optimization trade-offs related to gradient balancing, batch composition, and computational cost. In state-space models, the latent state must preserve enough local detail for reconstruction while also forming globally coherent representations for downstream tasks. The selective state mechanism supports this by allowing input-dependent forgetting and retention, but its behavior depends heavily on

how token sequences are constructed and how augmentation policies interact with scanning order.

Transformer-based segmentation models such as UNETR [15] illustrate how global token interactions can improve volumetric medical image analysis, yet they also show the computational limits of attention when patch counts grow. Validation studies for clinically deployed systems, such as diabetic retinopathy screening [16], highlight the need for evidence beyond offline benchmark performance, including prospective evaluation, workflow compatibility, and failure analysis in underrepresented populations. Fairness investigations in clinical risk prediction [17] further demonstrate that apparently effective models can encode structural inequities if training data and labels reflect biased care patterns. Cross-fusion Mamba designs for multi-view medical image classification [18] indicate that selective state propagation across complementary diagnostic views can improve representation consistency, but they also show that view alignment must be treated as a governance problem rather than a purely technical one. These perspectives collectively suggest that architecture selection cannot be separated from data provenance, evaluation policy, and institutional oversight.

System designers should therefore adopt a layered architecture in which representation learning is encapsulated as a shared service that can be versioned, monitored, and updated independently of downstream task heads. Such layering supports incremental retraining, audit logging, and controlled deployment across different clinical units. It also allows governance bodies to evaluate pretraining data influence, augmentation policies, and state-space parameterization without requiring full model retraining for every investigation. This is particularly important for generalizable medical image analysis because a representation model deployed across multiple hospitals must be treated as shared infrastructure rather than as a static artifact.

4. Training Infrastructure and Data Governance

Training self-supervised Mamba models on medical images places substantial demands on data infrastructure and institutional coordination. Medical image archives are often distributed across hospitals that operate under different privacy regulations, consent workflows, and data quality standards. The development of a generalizable representation learner therefore depends less on raw data volume than on the governance mechanisms that determine how data are collected, deidentified, licensed, and monitored over time. Federated training can reduce the need for centralized data aggregation, but it introduces additional system-level concerns such as communication efficiency, local update divergence, and the auditability of contributions from different sites. State-space models may have advantages in federated settings because their linear-time sequence processing can reduce computational cost at each client, yet model aggregation must still account for heterogeneity in scanner types, imaging protocols, and population distributions.

Foundation model development has increasingly shown that data curation is a first-order design decision rather than a preprocessing formality. Large-scale segmentation models such as Segment Anything [19] rely on carefully assembled datasets and iterative annotation workflows, and foundation model research has drawn attention to the risks of undocumented data provenance, licensing ambiguity, and unintended deployment [20]. In medical imaging, these risks are amplified because training data may encode differential access to care, referral bias, and historical disparities. Self-supervised pretraining can absorb such biases without explicit labels, making them harder to detect in downstream performance summaries. A robust training infrastructure must therefore include provenance tracking, dataset versioning,

and periodic audits that connect representation behavior to the demographic and clinical composition of pretraining corpora.

From an operational perspective, training large state-space models requires close integration among storage systems, accelerators, scheduling software, and experiment tracking platforms. Efficient scanning implementations, mixed-precision training, and gradient checkpointing can reduce resource usage, but these techniques must be governed by sustainability policies that account for energy consumption and carbon intensity. Academic and clinical institutions increasingly face pressure to report computational costs alongside accuracy metrics, because large-scale pretraining can otherwise consume resources disproportionate to marginal clinical benefit. A system-level approach therefore evaluates self-supervised Mamba models not only by downstream accuracy but also by the energy required to produce, update, and serve representations across heterogeneous clinical environments.

5. Robustness, Fairness, and Generalization Trade-offs

Generalization in medical image analysis is frequently framed as the ability to maintain performance across scanners, patient populations, and clinical sites. However, robustness and fairness are distinct but related properties that require separate governance attention. Robustness refers to stability under distribution shift, including changes in acquisition parameters, patient positioning, pathology prevalence, and image noise. Fairness refers to the distribution of model benefits and harms across social groups, particularly groups defined by race, ethnicity, sex, age, disability, and geography. A model may be robust in aggregate while remaining unfair to a minority population, and fairness audits [21] provide structured methods for interrogating how algorithmic systems allocate benefits and burdens. Self-supervised Mamba representations inherit these concerns because pretraining objectives can amplify dominant patterns and suppress clinically relevant rare features.

Empirical studies of self-supervised vision transformers have shown that transfer quality depends on optimization duration, augmentation intensity, and architectural capacity [22]. Similar trade-offs are likely to appear in state-space models, but the interaction between scanning order and data diversity remains underexamined. For example, a model pretrained predominantly on supine chest radiographs may learn strong representations for common pulmonary findings while failing to capture findings that manifest differently in upright portable radiographs. Fairness is not automatically improved by removing labels or by increasing dataset size; it requires deliberate sampling strategies, subgroup evaluation protocols, and mechanisms for contesting model behavior in clinical settings.

Generalization should therefore be treated as an institutional property rather than an isolated model property. A generalizable representation learning system includes not only the learned encoder but also the evaluation infrastructure, demographic reporting standards, and remediation processes that govern its use. Cross-domain comparisons with transformer and convolutional systems can provide useful diagnostic information, but they cannot substitute for prospective evaluation under realistic workflow constraints. The organizational capacity to detect performance drift, investigate root causes, and redistribute computational resources toward underrepresented sites is as central to generalization as the choice of self-supervised objective or state-space architecture.

6. Deployment, Sustainability, and Clinical Integration

Deploying self-supervised Mamba representations in clinical environments requires careful consideration of latency, interoperability, interpretability, and lifecycle management.

Representation models are often embedded within larger diagnostic pipelines that include image retrieval, report generation, decision support, and quality assurance. The value of a representation learner depends on whether it can be integrated into these workflows without creating new bottlenecks or obscuring clinical reasoning. State-space models may offer favorable inference latency for long sequences, but their sequential scanning behavior can complicate batch processing and require specialized kernels on hospital hardware. Deployment teams must therefore evaluate whether the architectural benefits observed in research benchmarks persist under real-world constraints related to network bandwidth, electronic health record integration, and radiologist workflow.

Sustainability is another critical concern. Continuous pretraining and periodic fine-tuning are often proposed as ways to maintain performance over time, yet these processes consume energy and can introduce instability if not governed carefully. A representation model that requires frequent retraining may be less sustainable than a simpler model that degrades more predictably. Clinical institutions must balance the marginal performance gains of updated representations against the costs of data revalidation, model recertification, and infrastructure maintenance. Regulatory frameworks in many jurisdictions require evidence of safety and effectiveness for software that influences medical decisions, and self-supervised representations complicate this requirement because they are not directly interpretable as diagnostic rules. Governance mechanisms must therefore include model cards, data sheets, and audit trails that connect pretraining data, architectural choices, and downstream performance to accountable institutional actors.

The socio-technical integration of representation learners also raises questions about professional autonomy and clinical trust. Radiologists and pathologists are more likely to accept model outputs when they understand the conditions under which representations were trained and when they can contest predictions that conflict with local clinical experience. Self-supervised Mamba models may support such trust by enabling rapid adaptation to local data distributions, but only if adaptation workflows are designed with clinical participation. Ultimately, sustainable deployment requires institutional structures that treat representation learning as an ongoing sociotechnical practice rather than a one-time technical intervention.

7. Conclusion

Self-supervised Mamba representation learning offers a promising direction for generalizable medical image analysis by combining label-efficient pretraining with linear-time sequence modeling. This paper has argued that the clinical and organizational value of such models depends on more than architectural performance. System-level factors including tokenization and scanning strategies, training infrastructure, data governance, fairness auditing, energy consumption, and deployment integration jointly determine whether representations generalize beyond narrow benchmark settings. Selective state-space models may reduce computational barriers to processing large medical volumes and multi-view data, but they also introduce new governance requirements related to sequence ordering, distributed training, and the interpretation of latent states. By treating self-supervised Mamba models as components of larger learning infrastructures, researchers and clinical institutions can better align technical innovation with the demands of equitable, robust, and sustainable medical image analysis.

References

1. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729–9738). <https://doi.org/10.1109/CVPR42600.2020.00975>
2. 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.
3. 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*. <https://arxiv.org/abs/2010.11929>
4. 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 30* (pp. 5998–6008).
5. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv*. <https://arxiv.org/abs/2312.00752>
6. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., & Liu, Y. (2024). VMamba: Visual state space model. *arXiv*. <https://arxiv.org/abs/2401.10166>
7. Xing, Z., Ye, T., Yang, Y., Liu, G., & Zhu, L. (2024). SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. *arXiv*. <https://arxiv.org/abs/2403.08460>
8. Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv*. <https://arxiv.org/abs/2401.04722>
9. 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. <https://doi.org/10.1038/s41592-020-01008-z>
10. 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. https://doi.org/10.1007/978-3-319-24574-4_28
11. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3478–3488). <https://doi.org/10.1109/ICCV48922.2021.00347>
12. Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., & Liang, J. (2021). Models Genesis. *Medical Image Analysis*, 67, 101840. <https://doi.org/10.1016/j.media.2020.101840>
13. 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*. <https://arxiv.org/abs/1711.05225>

14. 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., Langlotz, C., 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. <https://doi.org/10.1609/aaai.v33i01.3301590>
15. 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*. <https://arxiv.org/abs/2103.10504>
16. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(23), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
17. 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. <https://doi.org/10.1126/science.aax2342>
18. 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.
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). <https://doi.org/10.1109/ICCV51070.2023.00371>
20. 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*. <https://arxiv.org/abs/2108.07258>
21. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145–151). <https://doi.org/10.1145/3375627.3375820>
22. Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9640–9649). <https://doi.org/10.1109/ICCV48922.2021.00950>