Spectral Mamba: Mechanistic Auditing of Spectral Dynamics in Mamba-Based Medical Image Segmentation

Anonymous Author(s)

Anonymous Organization
@.***

Abstract. State Space Model (SSM) architectures such as Mamba have recently emerged as computationally efficient alternatives to Transformer-based medical image segmentation systems. Despite strong downstream segmentation performance, the internal representational dynamics induced by selective scan mechanisms remain insufficiently understood. In this work, we present a mechanistic audit of spectral dynamics in VM-UNet relative to Swin-Tiny and UNet-ResNet50 using a DC-corrected frequency-domain analysis framework centered around Alias Violation Ratio (AVR), a stage-wise metric computed from intermediate encoder activations.

Across architectures, VM-UNet exhibited elevated early-stage high-frequency concentration followed by pronounced late-stage spectral suppression. However, after intensity-bias correction through mean-centering, the observed spectral signatures exhibited near-zero pooled correlation with Boundary F1 (BF1) ($r = 0.0108$, $p = 0.6704$), indicating the absence of a strong global relationship between measured spectral organization and downstream boundary precision.

These findings suggest that the observed spectral behavior primarily reflects architecture-specific representational dynamics associated with selective scan processing rather than a universal predictor of segmentation boundary degradation. More broadly, this study demonstrates that frequency-domain auditing can provide useful mechanistic insight into emerging SSM-based medical image segmentation architectures while highlighting the importance of confound-aware interpretation in representation-level medical AI analysis.

Keywords: Mechanistic Interpretability · Medical Image Segmentation · Mamba · State Space Models · Spectral Analysis

1 Introduction

State Space Model (SSM) vision architectures have recently emerged as computationally efficient alternatives to Transformer-based segmentation systems [3]. In particular, Mamba-based architectures achieve linear-complexity sequential processing while maintaining competitive downstream segmentation performance [3].

As SSM-based segmentation systems continue to gain traction in computational medical imaging, understanding their internal representational behavior becomes increasingly important.

Unlike convolutional operators and self-attention mechanisms, selective scan operations introduce directional and ordering-dependent representation dynamics [8]. These sequential dynamics may induce distinctive frequency-domain organization within intermediate feature representations. In medical image segmentation, where accurate delineation of fine structural boundaries is particularly important, understanding these spectral dynamics may provide useful insight into emerging segmentation architectures [1].

This work investigates the following question: what characteristic spectral behaviors emerge within Mamba-based segmentation architectures, and do these spectral signatures meaningfully correspond to downstream boundary precision?

To address this question, we introduce a DC-corrected spectral auditing framework centered around Alias Violation Ratio (AVR), a stage-wise frequency-domain metric applied directly to intermediate encoder activations after mean-centering to suppress intensity-driven spectral bias.

Across VM-UNet, Swin-Tiny, and UNet-ResNet50, our analysis reveals a distinct stage-dependent spectral pattern within VM-UNet consisting of elevated early-stage AVR followed by pronounced late-stage spectral suppression. However, after DC correction, the observed spectral signatures exhibit near-zero pooled correlation with Boundary F1 (BF1), suggesting that the measured spectral behavior primarily reflects architecture-specific representational organization rather than a universal predictor of downstream boundary degradation.

1.1 Contributions

- We introduce a DC-corrected spectral auditing framework for analyzing intermediate encoder representations in Mamba-based medical image segmentation architectures.
- We operationalize Alias Violation Ratio (AVR), a stage-wise frequency-domain metric for characterizing relative spectral organization across encoder depth.
- We identify a characteristic dual-stage spectral trajectory in VM-UNet consisting of elevated early-stage AVR followed by substantial late-stage spectral suppression.
- We demonstrate that apparent AVR-BF1 relationships substantially collapse after DC correction, providing evidence against the strongest formulation of the proposed spectral degradation hypothesis.
- We release the complete analysis framework and experimental pipeline for reproducible mechanistic auditing of medical segmentation architectures.[1]

[1] The complete codebase, including training scripts and the DC-corrected spectral auditing framework, will be made publicly available upon acceptance.

2 Related Work

Frequency-domain analysis has increasingly been used to study internal representational behavior in deep vision architectures [1,7,8,12]. Prior work has shown that spectral statistics can reveal architectural biases, robustness characteristics, texture sensitivity, and implicit frequency preferences not directly observable through downstream accuracy metrics alone. In medical imaging specifically, frequency-domain analysis has been used to investigate representation collapse, robustness behavior, and segmentation stability under distributional perturbation [1, 7].

Recent work has additionally demonstrated that modern deep architectures exhibit characteristic spectral biases during optimization and representation formation [8]. Convolutional architectures often display localized frequency selectivity driven by spatial inductive bias, while Transformer-based systems exhibit different long-range representational organization due to global self-attention dynamics.

More recently, State Space Models (SSMs) such as Mamba [3] have emerged as efficient alternatives to self-attention for dense vision tasks. VM-UNet [10] introduced selective scan mechanisms into medical image segmentation and demonstrated competitive segmentation performance with improved computational efficiency. However, the internal spectral behavior induced specifically by selective scan dynamics remains insufficiently characterized.

Mechanistic interpretability approaches in medical imaging have increasingly emphasized representation-level analysis beyond conventional segmentation metrics. Prior work has explored anti-aliasing behavior, spectral suppression, and frequency-domain auditing in medical segmentation systems [5]. In this work, we extend this direction through DC-corrected spectral auditing of intermediate feature representations across convolutional, Transformer-based, and SSM-based segmentation architectures.

3 Methodology

3.1 Alias Violation Ratio (AVR)

To quantify architecture-specific spectral organization within intermediate feature representations, we introduce Alias Violation Ratio (AVR), a stage-wise frequency-domain metric computed directly from encoder activations.

Given an intermediate feature tensor $F \in \mathbb{R}^{C\times H\times W}$, channel-wise mean-centering is first applied:

$$\hat{F}_c = F_c - \mu(F_c) \tag{1}$$

A two-dimensional FFT is then computed:

$$S_c(u,v) = |\mathcal{F}(\hat{F}_c(u,v))|^2 \tag{2}$$

AVR is defined as:

$$\mathrm{AVR}_c = \frac{\sum_{(u,v)\in\Omega_H} S_c(u,v)}{\sum_{(u,v)\in\Omega_L} S_c(u,v)} \tag{3}$$

where Ω_H and Ω_L denote radially defined high-frequency and low-frequency spectral regions respectively.

Higher AVR values indicate greater relative concentration of spectral energy within predefined high-frequency regions of the representation space. AVR is intended as an operational measure of relative spectral organization rather than a formal estimator of Nyquist aliasing behavior [11].

3.2 DC-Corrected Spectral Analysis

Raw Fourier statistics are strongly influenced by low-frequency intensity structure and DC activation bias. To suppress this confound, all feature maps were mean-centered prior to FFT computation. All reported AVR measurements therefore reflect DC-corrected spectral organization rather than global activation magnitude differences across encoder stages or architectures.

4 Experimental Setup

Experiments were conducted on the ISIC 2018 Skin Lesion Segmentation Challenge dataset [2] using an 80/20 train-validation split with fixed seed initialization.

We evaluate spectral behavior across VM-UNet [10], Swin-Tiny [6], and UNet-ResNet50 [4,9], representing SSM-based, Transformer-based, and convolutional segmentation paradigms respectively. VM-UNet and UNet-ResNet50 operated on 256×256 inputs, while Swin-Tiny used 224×224 inputs consistent with its pretrained backbone configuration.

Intermediate encoder activations were extracted using forward-hook instrumentation at semantically corresponding encoder stages. All models were trained for 100 epochs using pretrained backbone initialization where supported. Experiments were conducted on an NVIDIA RTX 4050 Laptop GPU with 6 GB VRAM.

5 Results

5.1 Performance and Spectral Dynamics

VM-UNet achieved the highest Dice score while Swin-Tiny produced the strongest BF1 performance (Table 1). Despite similar downstream segmentation quality, stage-wise AVR analysis revealed substantially different spectral trajectories across architectures. Most notably, VM-UNet exhibited elevated early-stage AVR followed by substantial late-stage spectral suppression (Fig. 1), producing a characteristic dual-stage spectral trajectory not observed in either comparison baseline.

Table 1: Segmentation performance and stage-wise spectral statistics.

Model	Dice	BF1	S1	S2	S3	S4	Mean AVR
UNet-ResNet50	0.9000	0.4470	0.3427	0.3613	0.2974	0.1802	0.2954
Swin-Tiny	0.9023	**0.5259**	0.3276	0.3744	0.2519	0.3623	**0.3291**
VM-UNet	**0.9027**	0.4939	**0.4600**	**0.3840**	0.1408	0.1346	0.2799

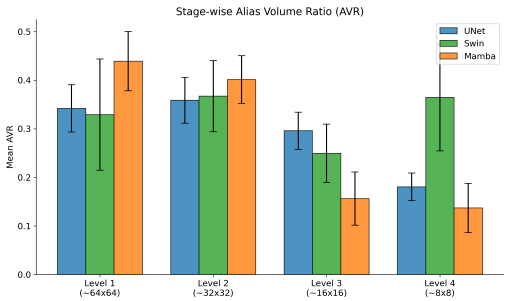

Fig. 1: Stage-wise AVR dynamics across encoder levels.

5.2 Correlation Analysis

After mean-centering, pooled AVR-BF1 correlation collapsed to near-zero ($r = 0.0108$, $p = 0.6704$), indicating the absence of a strong global relationship between spectral organization and downstream boundary precision.

Importantly, within-architecture correlation behavior also differed substantially across models. VM-UNet exhibited a weak positive correlation between AVR and BF1, while UNet-ResNet50 demonstrated a weak negative correlation and Swin-Tiny exhibited no statistically significant relationship. This absence of a consistent directional trend suggests that any AVR-BF1 relationship may be architecture-dependent rather than globally generalizable.

5.3 Spectral Fingerprints

Spectral visualizations further revealed characteristic frequency-domain organization within VM-UNet relative to both Swin-Tiny and UNet-ResNet50.

Table 2: Correlation analysis between AVR and BF1 after DC correction.

Population	Pearson r	p-value	Interpretation
VM-UNet	+0.0998	0.0229	Weak positive association
Swin-Tiny	+0.0188	0.6686	No significant correlation
UNet-ResNet50	-0.1880	< 0.0001	Weak negative association
Pooled	+0.0108	0.6704	Correlation collapse
Partial	-0.0001	0.9956	No structural link

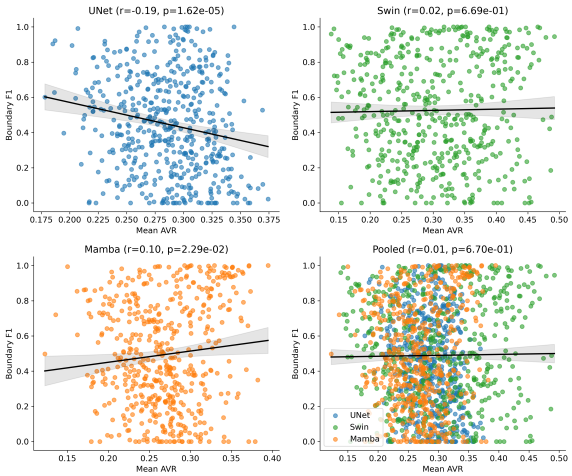

Fig. 2: Relationship between AVR and Boundary F1 (BF1) across architectures.

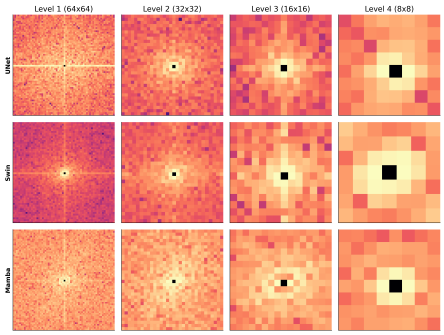

Fig. 3: Representative spectral heatmaps across encoder stages.

As shown in Fig. 3, early VM-UNet encoder stages exhibited broader peripheral spectral activation patterns compared to both comparison architectures, indicating relatively elevated high-frequency concentration during earlier stages of representation formation. However, deeper VM-UNet stages demonstrated progressively centralized low-frequency spectral concentration accompanied by attenuation of peripheral spectral energy.

This stage-dependent transition was further supported by frequency-band decomposition analysis (Fig. 4). VM-UNet demonstrated substantially higher early-stage high-frequency energy proportions relative to later encoder stages, where low-frequency dominance became increasingly pronounced.

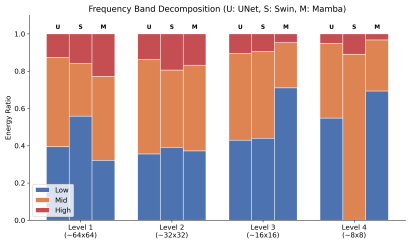

Fig. 4: Frequency-band energy decomposition across encoder stages.

6 Discussion and Limitations

6.1 Interpretation of Spectral Dynamics

Across analyses, VM-UNet consistently exhibited characteristic frequency-domain organization relative to both Transformer-based and convolutional segmentation baselines. Most notably, stage-wise AVR analysis revealed elevated early-stage high-frequency concentration followed by pronounced late-stage spectral suppression.

One possible interpretation is that selective scan dynamics may induce distinctive representational organization across encoder depth relative to convolutional and attention-based processing. Under this interpretation, earlier encoder stages may preserve broader high-frequency activation patterns, while deeper stages progressively consolidate representations into lower-frequency structure. This interpretation remains speculative rather than causal.

The spectral heatmaps and frequency-band decomposition analyses qualitatively support this interpretation, with VM-UNet exhibiting broader peripheral spectral activation in earlier encoder stages followed by progressively centralized low-frequency concentration in deeper stages.

These observations should not be interpreted as pathological aliasing in a strict signal-processing sense. Rather, the observed spectral trajectories appear to reflect architecture-specific representational dynamics associated with selective scan processing.

Interpretation of the Null Correlation Result. A central motivation of this work was the hypothesis that elevated high-frequency spectral concentration may correspond to degraded downstream boundary precision. However, after mean-centering and suppression of low-frequency intensity bias, pooled AVR-BF1 correlation collapsed to near-zero.

This negative result is itself informative. Rather than supporting a universal relationship between spectral organization and downstream boundary precision, the findings suggest that much of the initially observed correspondence was driven by intensity-related confounds rather than structural representational behavior alone.

Consequently, the strongest formulation of the proposed spectral degradation hypothesis was not supported after confound correction.

6.2 Limitations

Several limitations should be considered when interpreting these findings. First, the analysis was conducted on a single dermoscopic segmentation dataset using a limited set of comparison architectures and a single random seed. Second, VM-UNet was trained from scratch while Swin-Tiny and UNet-ResNet50 used pretrained ImageNet encoders, introducing a potentially important representational asymmetry across models.

Additionally, AVR is intended as an operational measure of relative spectral organization rather than a formal estimator of theoretical aliasing behavior. Finally, the present study remains observational and does not establish causal relationships between spectral organization and downstream segmentation behavior.

7 Conclusion

We presented a mechanistic audit of spectral dynamics in VM-UNet relative to Transformer-based and convolutional medical image segmentation baselines. Using a DC-corrected spectral auditing framework centered around AVR, we identified a stage-dependent spectral transition within VM-UNet characterized by early high-frequency concentration followed by progressive late-stage suppression.

However, after intensity-bias correction, the observed spectral signatures exhibited near-zero pooled correlation with downstream boundary precision, suggesting that the measured spectral behavior primarily reflects architecture-specific representational organization rather than a universal predictor of segmentation failure.

Overall, this study demonstrates that frequency-domain auditing can provide useful mechanistic insight into emerging SSM-based medical image segmentation architectures while highlighting the importance of confound-aware interpretation in representation-level medical AI analysis.

References

1. Chen, M., et al.: Frequency domain analysis of deep learning models in medical imaging. Medical Image Analysis (2023)
2. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). In: 2018 IEEE International Symposium on Biomedical Imaging (2018)
3. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
4. He, K., et al.: Deep residual learning for image recognition. In: CVPR (2016)

5. Kashyap, S.: The structural cost of anti-aliasing in 3d volumetric segmentation. In: Medical Imaging with Deep Learning - Short Papers (2026), https://openreview.net/forum?id=F7RaWxyKXD
6. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
7. Ortega, P., et al.: Spectral analysis of deep representations in medical imaging. In: Medical Imaging with Deep Learning (2024)
8. Rahaman, N., et al.: Spectral bias and neural networks. ICML (2019)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
10. Ruan, J., et al.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
11. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal (1949)
12. Wang, H., et al.: High-frequency component helps explain the generalization of convolutional neural networks. CVPR (2020)