Time is continuous, why chunk it? Token-Free State-Space Models for ECG

Sukwon Kim [1] **William Jongwon Han** [1] **Ding Zhao** [1]

Abstract

Recent successes in foundation models for electrocardiograms (ECGs) are largely constrained by tokenization and fixed-length input paradigms. These fixed-length paradigms restrict models through heuristics like padding or sliding windows, which can disrupt the continuous temporal dynamics of the signal. To challenge these paradigms, we investigate State-Space Models (SSMs) as a foundational architecture, demonstrating their capability to natively process raw signals without artificial chunking. Using a simple SSM-based encoder, experiments show competitive classification performance compared to state-of-the-art baselines on standard fixed-length benchmarks. Crucially, further evaluations on truncated and long signals reveal significantly less performance degradation, confirming the inherent robustness of SSM architecture over fixed-length approaches.

1. Introduction

Recent advances in deep learning have driven significant success in multi-lead electrocardiogram (ECG) analysis, mostly with complex Convolutional Neural Networks (CNNs) and Transformer-based architectures(Vaswani et al., 2017). Despite their remarkable success, the vast majority of these state-of-the-art models are structurally constrained by fixed-length input paradigms, which are typically optimized for standard 10-second segments that are prevalent in widely used benchmark datasets (Wagner et al., 2020).

This structural rigidity fundamentally misaligns with the clinical reality of ECG monitoring, where real-world data inherently manifests in highly variable lengths—ranging from few-second-long signals generated by wearable smart devices to continuous 24-hour Holter monitor recordings. The inability to natively process variable-length inputs exposes severe structural vulnerabilities in prevailing architectures. Standard CNNs are conventionally bounded by fixed receptive fields, making it architecturally challenging to capture long-range temporal dependencies (Ansari et al., 2023). While recent innovations demonstrate that CNNs can model global context—for instance, by employing continuous kernel parameterizations (Romero et al., 2022)—such adaptations require replacing traditional discrete convolutions entirely with implicit neural network generators.

Conversely, Transformer-based architectures capture global context comprehensively but suffer from a quadratic computational complexity as the sequence length grows. Consequently, processing long-term continuous physiological signals in a single turn becomes restricted. To circumvent these limitations, current methodologies rely on destructive artificial heuristics: zero-padding for severely truncated signals, which disrupts temporal distribution and calibration, or sliding-window chunking for long sequences, which fragments the global context.

Recently, Mamba (Gu & Dao, 2023) has emerged as a powerful alternative, offering linear complexity with a selective state-update mechanism. This architecture has demonstrated remarkable efficacy in domains requiring long sequence modeling. For example, the original Mamba architecture excelled in modeling raw DNA sequences at the single-nucleotide level, and subsequent works like MambaByte (Wang et al., 2024) proved that Mamba can effectively process untokenized raw byte streams without performance degradation.

Despite these breakthroughs in token-free modeling, recent studies attempting to apply Mamba to cardiology predominantly utilize CNN-based front-ends or patch tokenizers before feeding the embeddings into the Mamba blocks (Qiang et al., 2024; Jiang et al., 2025). By artificially chunking the signal into fixed token spaces, these approaches inadvertently inherit the limitations of fixed-length models, failing to leverage Mamba's intrinsic capability for token-free, raw continuous signal processing.

To address these limitations, we propose a pure Mamba-

[1]Carnegie Mellon University, USA. Correspondence to: William Jongwon Han <wjhan@andrew.cmu.edu>, Ding Zhao <dingzhao@andrew.cmu.edu>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

based encoder to confirm that the continuous state-space formulation is uniquely suited for modeling physiological time-series data. By treating the raw ECG signal as a continuous byte-stream optimized via self-supervised Masked Signal Modeling (MSM), we systematically investigate its modeling capabilities.

Our main contributions are summarized as follows:

- **Raw Physiological Modeling:** We introduce a pure, continuous Mamba-based encoder that natively processes raw ECG signals. This establishes a novel paradigm that removes the need for patch tokenization or CNN-based feature extraction, thereby preserving the fine-grained temporal resolution of the physiological signal.
- **Robustness Across Variable-Length Inputs:** We empirically demonstrate that our continuous state-space formulation remains highly resilient across widely varying input lengths. By natively processing truncated and extended signals, it effectively circumvents the severe performance degradation and calibration instability caused by artificial zero-padding in traditional fixed-length architectures.
- **Heuristic-Free Linear Processing and Efficiency:** We validate the $\mathcal{O}(N)$ computational scalability of our token-free approach. By avoiding heuristics like sliding-window chunking, our model achieves superior inference throughput and processes entire recordings in a single pass, seamlessly preserving global temporal context.

2. Method

2.1. Architecture Overview

Unlike prevailing Vision Transformers (ViTs) (Dosovitskiy et al., 2021) or CNN models that require explicit patch tokenization, our architecture is designed to process the raw continuous sequence directly. The model ingests the raw 12-lead ECG signal and maps it into a high-dimensional hidden space ($d_{model} = 512$) utilizing only a simple linear projection layer.

The core encoder consists of four sequentially stacked Mamba blocks (configured with a state dimension $d_{state} = 16$ and an expansion factor of 2). For decoder, a single Mamba block with linear layer was used to maintain comparatively lighter architecture than encoder. Each block is followed by Layer Normalization to stabilize the continuous state transitions. The Mamba layers utilize a data-dependent selective mechanism to dynamically filter out irrelevant noise while retaining critical morphological information over extended contexts.

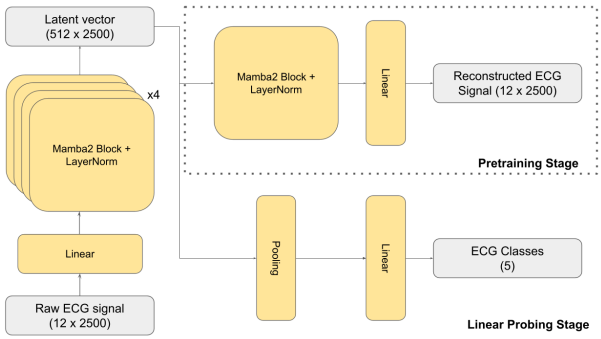

Figure 1. Overall architecture of the token-free Mamba encoder.

During the self-supervised pretraining stage, a lightweight Mamba decoder is attached to reconstruct the masked signals. For the downstream linear probing stage, the encoder weights are frozen, and the latent representations are aggregated via global average pooling before being passed to a linear classifier.

2.2. Pretraining Strategy

To acquire robust representations of cardiac morphology, we pretrain the token-free encoder using a self-supervised Masked Signal Modeling (MSM) objective, heavily inspired by the Masked Autoencoder (MAE) framework (He et al., 2022). The model is pretrained on standard 12-lead 10-second ECG segments from the MIMIC-IV-ECG dataset (Johnson et al., 2023). Crucially, the masking operation is applied directly to the raw continuous signals prior to the linear projection. We utilize a block-wise masking strategy with a 50% masking ratio to prevent the model from exploiting trivial local temporal interpolations. Each blocks are 100 ms long and are randomly masked.

Following this MAE-style asymmetric encoder-decoder design philosophy for computational efficiency, the 4-block Mamba encoder operates solely on the unmasked visible segments of the signal. Trainable mask tokens are subsequently inserted to restore the original temporal sequence, which is then processed by a highly lightweight decoder consisting of a single Mamba block. A linear projection head maps the decoder's output back to the original 12-lead signal space. The model optimizes the Mean Squared Error (MSE) strictly over the masked timesteps, forcing the architecture to leverage its continuous state-updates to infer missing rhythmic patterns and granular local morphologies.

3. Experiments

3.1. Datasets and Downstream Configuration

All evaluated models process ECG signals sampled at 250 Hz, resulting in 2,500 timesteps for standard 10-second

inputs. Following the pretraining on the MIMIC-IV-ECG dataset, we evaluate the representation learning quality on a downstream arrhythmia classification task via linear probing on the PTB-XL dataset (Wagner et al., 2020). During this phase, the pretrained encoders are strictly frozen, and a linear classifier is trained. To systematically evaluate representation quality and ensure a fair, rigorous comparison, the training dataset size was strictly controlled (4,310 instances).

Furthermore, to assess the model's ability to generalize to longer, unseen data without additional fine-tuning, we evaluate our approach on the INCART dataset (Yakushenko, 2008), which features continuous 30-minute recordings. To ensure consistency of evaluation, the diagnostic labels in INCART were assigned to the five PTB-XL superclasses (NORM, MI, STTC, CD, HYP) by extracting clinical keywords and applying a hierarchy-based single-label assignment (MI $>$ STTC $>$ HYP $>$ CD).

3.2. Baselines and Architectural Adaptations

To establish a rigorous comparison, we benchmark our proposed architecture against two recent state-of-the-art representation learning models optimized for standard ECGs: MERL (Liu et al., 2024), which utilizes a CNN backbone, and ST-MEM (Na et al., 2024), which employs a Transformer-based masked autoencoder.

A critical aspect of our experimental design involves evaluating these models across widely varying temporal contexts. However, because both baselines are strictly bounded by fixed-length input requirements (10 seconds), evaluating them on shortened inputs (e.g., 2-second and 5-second segments) necessitated zero-padding. To mitigate the disruptive effects of this artificial padding and ensure the fairest possible baseline performance, we implemented targeted masking strategies during inference. Specifically, we explicitly excluded the zero-padded regions from the global average pooling calculation for MERL, and supplied explicit attention masks for ST-MEM to prevent computing on padded areas. For the 30-minute continuous sequences from the INCART dataset, we employed a sliding-window chunking strategy for ST-MEM and MERL, uniformly segmenting the signals into independent 10-second chunks and aggregating the final prediction via mean-pooling.

In contrast, our proposed Mamba encoder natively processes both truncated and extended variable-length sequences through continuous state transitions, without requiring any padding, masking, or chunking heuristics.

3.3. Evaluation Protocol and Statistical Analysis

We benchmark the computational efficiency of our approach against baseline architectures by measuring inference throughput and peak VRAM consumption.

Table 1. Baseline performance on 10s fixed-length ECG signals (PTB-XL). Mamba achieves competitive diagnostic accuracy compared to SOTA baselines, establishing a foundation for variable-length evaluation.

Method	Accuracy (%)	Top-2 Recall (%)	AUROC	Throughput	Peak VRAM
MERL	63.33	82.86	0.815	717.5	292.3
ST-MEM	60.12	77.26	0.756	223.5	863.4
Mamba	58.26	76.73	0.763	351.4	1968.8

Crucially, to rigorously evaluate the statistical significance of the performance differences on variable-length inputs, we employ a paired bootstrap resampling approach ($N = 2,000$ iterations). Rather than merely comparing absolute performance scores, we calculate the empirical p-value based on the difference-in-differences (DiD) of the performance drop (from standard 10s to truncated 2s) between the Mamba encoder and the baseline models.

4. Results

4.1. Baseline Performance on Fixed-Length Inputs

We first evaluate the models on standard 10-second sequences (PTB-XL) (Table 1). In this constrained setting, the highly optimized MERL architecture achieves the highest diagnostic performance (AUROC 0.815). This likely stems from MERL's use of a contrastive learning objective during pre-training, which explicitly includes diagnostic statements. Our pure Mamba encoder establishes a competitive foundation with an AUROC of 0.763, performing on par with the ST-MEM baseline (0.756). This predefined 10-second environment represents the optimal operational condition for traditional fixed-window architectures. Rather than demonstrating Mamba's full potential, this baseline serves to expose the structural vulnerabilities of traditional models when confronted with real-world, variable-length signals.

4.2. Robustness to Short-Term Truncated Signals

The primary structural vulnerability of traditional architectures is dramatically exposed when processing truncated inputs (Fig 2). When evaluating on 2-second partial signals—which forces baselines to rely on heavy zero-padding—MERL and ST-MEM exhibit severe performance degradation, with Top-2 recall dropping by 7.5% and 4.7%, respectively.

In sharp contrast, the Mamba encoder demonstrates exceptional stability, even achieving a marginal improvement (+0.8%) on the same 2-second segments. Difference-in-differences (DiD) analysis confirms that Mamba's superior robustness over the baselines is highly statistically significant ($p < 0.001$).

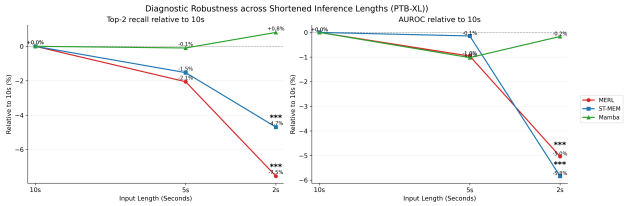

Figure 2. Diagnostic Robustness across Shortened Inference Lengths (PTB-XL)

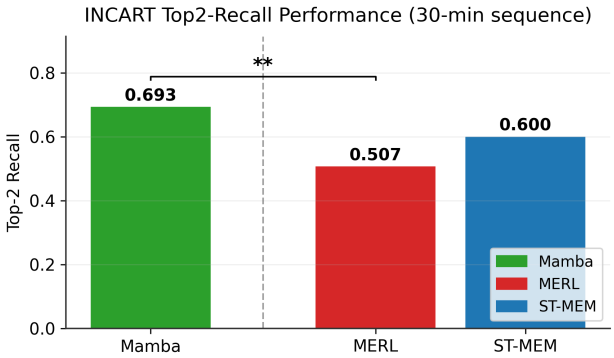

Figure 3. Diagnostic Extrapolation across 30-Minute Sequences (INCART)

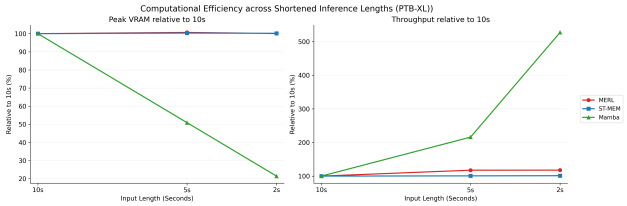

Figure 4. Computational Efficiency across Shortened Inference Lengths (PTB-XL)

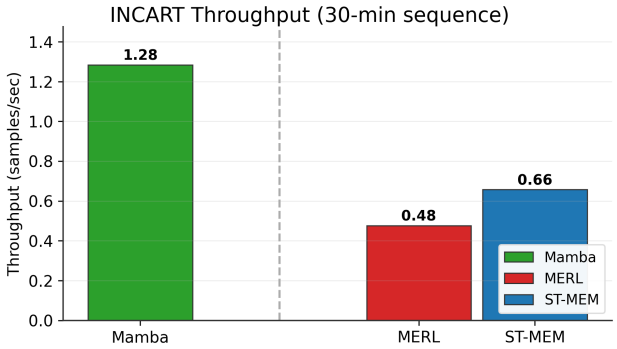

Figure 5. Computational Scalability across 30-Minute Sequences (INCART)

Extensive evaluations across multiple data distribution splits confirmed that the calibration instability of the baselines is fundamentally caused by the zero-padding artifacts (Table 2). Conversely, Mamba's native continuous state transitions allow it to extract meaningful morphological features from partial signals flawlessly.

4.3. Generalization to Longer ECG Sequences

The ultimate utility of the Mamba encoder lies in its scalability to much longer input sequences. In evaluating generalization on 30-minute INCART recordings, 180 times longer than the training context, Mamba demonstrated superior diagnostic stability (Fig 3). It achieved a Top-2 recall of 0.693, significantly outperforming MERL (0.507, $p < 0.01$) and ST-MEM (0.600). While baseline models suffer from fragmented global context due to their reliance on sliding-window chunking, Mamba's continuous state-space mechanism seamlessly preserves and updates the cardiac state across the entire 30-minute duration.

4.4. Computational Efficiency and Linear Scaling

The structural advantages of Mamba are equally pronounced in computational scalability. As input length decreases to 2 seconds, Mamba exhibits linear scaling: peak VRAM usage drops by 80%, and throughput surges over 5× compared to its 10-second baseline (Fig 4). Baseline models, forced to process padded zeros, suffer from massive computational waste and maintain fixed latency. This efficiency gap widens significantly on long-term sequences. Whereas MERL and ST-MEM are forced to employ sliding-window chunking heuristics to process 30-minute signals due to their inherent context limitations, Mamba achieves an approximate 2× inference speedup over the nearest baseline (1.28 vs. 0.66 samples/sec) (Fig 5). Even compared to these optimized heuristics, Mamba's linear time complexity allows it to process the entire long-term sequence in a single, efficient pass, validating its role as a uniquely scalable solution for real-time clinical monitoring.

5. Conclusion

In this study, we demonstrated that a pure Mamba encoder can serve as a foundational architecture to directly process raw ECG sequences regardless of their length.

In contrast to traditional fixed-length models, which exhibit a 5–10% performance drop, our Mamba-based encoder demonstrated virtually no performance degradation on 2-second signals. Simultaneously, the Mamba encoder showed significantly superior performance on an unseen dataset with 30-minute recordings compared to that of MERL. This dual capability—scaling efficiently to long sequences while remaining resilient to severe truncation—confirms that SSMs provide the flexibility necessary for dynamic clinical environments.

Furthermore, our findings have validated the computational

advantage of Mamba's linear time complexity over the quadratic overhead inherent to Transformer-based models. This architectural efficiency enables the direct, single-pass processing of 30-minute continuous recordings, preserving global temporal context without heuristic chunking.

Limitations and Future Work

While our straightforward implementation serves as a structural proof-of-concept for raw physiological modeling, we acknowledge several limitations that map out critical directions for future research.

First, a performance gap remains on standard 10-second sequences between our pure Mamba encoder (AUROC 0.763) and highly optimized CNN-based frameworks like MERL (AUROC 0.815). Further investigation is warranted to analyze the specific inductive biases of convolutional layers that excel in capturing localized, fixed-window features.

Second, our self-supervised pre-training utilized a fixed 100ms block masking strategy with a 50% random ratio. Given that a typical high-frequency QRS complex spans approximately 80–120ms, a rigid 100ms mask risks obliterating entire R-peaks, making it challenging for the model to genuinely learn cardiac electrophysiology rather than merely interpolating from low-frequency surroundings. Future work will include extensive ablation studies on mask lengths and explore ECG-aware, domain-specific masking strategies to ensure robust physiological representations.

Finally, although our model scales linearly for long sequences, it exhibits a higher base memory footprint on standard short-duration clips compared to traditional baselines that operates based on sliding window. This optimization overhead presents challenges for real-world deployment in computation-constrained clinical edge environments. To address this, future efforts will investigate introducing a lightweight CNN-based front-end to downsample the raw signal and reduce the token sequence length before it enters the Mamba blocks, thereby optimizing the memory footprint for strict clinical hardware settings while fully leveraging Mamba's modeling capabilities.

Ultimately, this native variable-length processing framework opens up new possibilities for more adaptable and deployable foundational models in real-world healthcare settings.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically in the domain of physiological time-series analysis. The potential societal consequences of our work include enabling more accessible, efficient, and robust continuous cardiac monitoring, which could significantly improve the early diagnosis and management of cardiovascular diseases. However, as with all healthcare AI models, ensuring equitable performance across diverse patient demographics and mitigating potential algorithmic biases in training data remain critical ethical considerations for future clinical deployment.

References

Ansari, Y., Mourad, O., Qaraqe, K., and Serpedin, E. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, 14:1246746, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Jiang, H., Mutahira, H., Wei, S., and Muhammad, M. S. ECG-Mamba: Cardiac abnormality classification with non-uniform-mix augmentation on 12-lead ECGs. *IEEE Journal of Translational Engineering in Health and Medicine*, 2025.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Moody, B., Gow, B., wei H. Lehman, L., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 2023.

Liu, C., Wan, Z., Ouyang, C., Shah, A., Bai, W., and Arcucci, R. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

Na, Y., Park, M., Tae, Y., and Joo, S. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

Qiang, Y., Dong, X., Liu, X., Yang, Y., Fang, Y., and Dou, J. ECGMamba: Towards efficient ECG classification with BiSSM. *arXiv preprint arXiv:2406.10098*, 2024.

Romero, D. W., Knigge, D. M., Gu, A., Bekkers, E. J., Gavves, E., Tomczak, J. M., and Hoogendoorn, M. Towards a general purpose cnn for long range dependencies in nd. 2022.

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

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

Wang, J., Gangavarapu, T., Yan, J. N., and Rush, A. M. MambaByte: Token-free selective state space model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Yakushenko, E. St petersburg INCART 12-lead arrhythmia database v1.0.0, 2008. URL https://physionet.org/content/incartdb/1.0.0/.

Table 2. Variable-length robustness across different Train/Test data distributions. Baseline models (MERL, ST-MEM) consistently exhibit severe performance drops (Δ) on 2-second padded inputs regardless of the distribution, while Mamba maintains stability.

Train / Test Split	Method	Accuracy (%)		Top-2 Recall		AUROC	
		10s	2s (Δ)	10s	2s (Δ)	10s	2s (Δ)
Unbalanced / Unbalanced	MERL	63.33	55.13 (-8.20)	0.829	0.766 (-0.063)	0.815	0.774 (-0.041)
	ST-MEM	60.12	55.11 (-5.01)	0.773	0.736 (-0.037)	0.756	0.712 (-0.044)
	Mamba	58.26	59.12 (+0.86)	0.767	0.773 (+0.006)	0.763	0.762 (-0.001)
Unbalanced / Balanced	MERL	50.73	46.17 (-4.56)	0.710	0.677 (-0.033)	0.800	0.761 (-0.039)
	ST-MEM	39.27	36.47 (-2.80)	0.588	0.545 (-0.043)	0.732	0.686 (-0.046)
	Mamba	36.83	38.96 (+2.13)	0.593	0.617 (+0.024)	0.714	0.725 (+0.011)
Balanced / Balanced	MERL	50.00	45.54 (-4.46)	0.751	0.710 (-0.041)	0.798	0.762 (-0.036)
	ST-MEM	46.34	36.32 (-10.02)	0.722	0.633 (-0.089)	0.760	0.697 (-0.063)
	Mamba	38.29	36.13 (-2.16)	0.622	0.600 (-0.022)	0.696	0.687 (-0.009)
Balanced / Unbalanced	MERL	59.61	49.80 (-9.81)	0.778	0.709 (-0.069)	0.809	0.770 (-0.039)
	ST-MEM	58.26	46.59 (-11.67)	0.762	0.666 (-0.096)	0.766	0.711 (-0.055)
	Mamba	50.90	41.94 (-8.96)	0.688	0.613 (-0.075)	0.737	0.704 (-0.033)

A. Impact of Data Distribution on Variable-Length Robustness

To ensure that the observed performance degradation on shortened signals is inherently a structural limitation of the baseline models rather than an artifact of class imbalance, we expanded our evaluation across four distinct data distribution splits (Train/Test): Balanced/Balanced, Balanced/Unbalanced, Unbalanced/Balanced, and Unbalanced/Unbalanced. We recorded the absolute performance at the standard 10-second length and measured the relative degradation when the input was truncated to 2 seconds.

As detailed in Table 2, while the absolute baseline performance fluctuates depending on the training and testing distributions, a consistent pattern emerges regarding variable-length robustness. Regardless of the data distribution, baseline architectures strictly bound by fixed-length inputs experience severe and consistent performance drops when processing 2-second padded signals. This consistency confirms that their calibration instability is fundamentally caused by the zero-padding methodology disrupting temporal representations, independent of class distribution.

In contrast, the Mamba encoder, which natively processes variable-length sequences through continuous state transitions, demonstrates extreme resilience across the board. In the natural Unbalanced/Unbalanced setting, Mamba's AUROC remains remarkably stable, exhibiting a negligible change from 0.763 to 0.762, while its accuracy even slightly increases. These findings conclusively demonstrate that Mamba's structural robustness to temporal length variation is an intrinsic architectural advantage that persists robustly across varying data distributions.