IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2017.DOI

StructMamba: Structured Harmonic and Temporal Music Analysis via Dual-Axis Mamba and Attention

AMIT KUMAR BAIRWA[1], (Senior Member, IEEE) , SIDDHANTH BHAT [2], (Member, IEEE), TANISHK SAWANT [3], (Member, IEEE), and MANOJ KUMAR BOHRA [4], (Senior Member, IEEE)

[1,2,3,4]School of Computer Science and Engineering, Manipal University Jaipur, Jaipur-Ajmer Express Highway, Jaipur, India

Corresponding author: Siddhanth Bhat (e-mail: siddhanth.219310154@muj.manipal.edu, Tanishk Sawant (e-mail: tanishk.219302442@muj.manipal.edu, Manoj Kumar Bohra (e-mail: manojkumar.bohra@jaipur.manipal.edu

ABSTRACT Modeling musical audio requires capturing hierarchical relationships between harmonic textures, rhythmic motifs, and long-range structural repetitions. Convolutional networks extract local features efficiently, while transformers provide global modeling, yet both face mismatches with musical structure. In this work we introduce **StructMamba**, a dual-axis architecture that unifies state-space modeling with global two-dimensional attention. Our design decomposes spectrogram modeling into frequency-wise and time-wise Mamba modules, enabling independent learning of harmonic and rhythmic dependencies before fusing them through structured attention. Evaluated on benchmark tasks in genre classification, onset detection, and structural segmentation, StructMamba outperforms strong CNN, transformer, and hybrid baselines, while maintaining stability in low-resource settings. Beyond accuracy, its internal representations align with music-theoretic constructs such as motifs, downbeats, and sectional boundaries, offering rare interpretability for deep audio models. These findings position StructMamba as an efficient and musically aligned solution for time–frequency audio modeling, with practical implications for music education, annotation, and production.

INDEX TERMS Music Information Retrieval, State Space Models, Mamba, Time–Frequency Modeling, Global Attention, Spectrograms, Harmonic Structure, Rhythmic Modeling, Audio Deep Learning, StructMamba

I. INTRODUCTION

Music inherently exhibits complex hierarchical structures along both frequency and temporal axes, where harmonic relationships coexist with rhythmic patterns and higher-level global repetitions such as motifs and themes. Effective modeling of musical signals therefore necessitates capturing local spectral features, temporal dynamics, and global structural regularities simultaneously. Typically, these properties are represented via time-frequency representations, notably mel-spectrograms or constant Q transforms (CQTs), which encapsulate critical auditory characteristics and facilitate meaningful analysis within music information retrieval (MIR) tasks [1], [2].

Convolutional neural networks (CNNs), historically prominent in MIR tasks, leverage local inductive biases to efficiently extract and model frequency-localized patterns, such as harmonic textures or timbral features. However, CNN architectures exhibit fundamental limitations: Their receptive fields are constrained, thus inadequately capturing longer-range dependencies and temporal structures that underpin rhythmic patterns and structural repetitions in music [3], [4]. Prior architectures often rely on either vertical frequency modeling (e.g., CNNs) or horizontal time modeling (e.g., Transformers), but music demands both. Harmonic progression and rhythmic continuity are fundamentally orthogonal, motivating a dual-branch design.

In contrast, transformer-based models, built upon self-attention mechanisms, have shown considerable promise in effectively modeling global dependencies in sequential data [5]. Despite their global context modeling strength, transformers typically lack explicit inductive biases tailored towards local structures, requiring substantial computational resources and large datasets to achieve satisfactory performance in MIR tasks. This mismatch leads to inefficien-

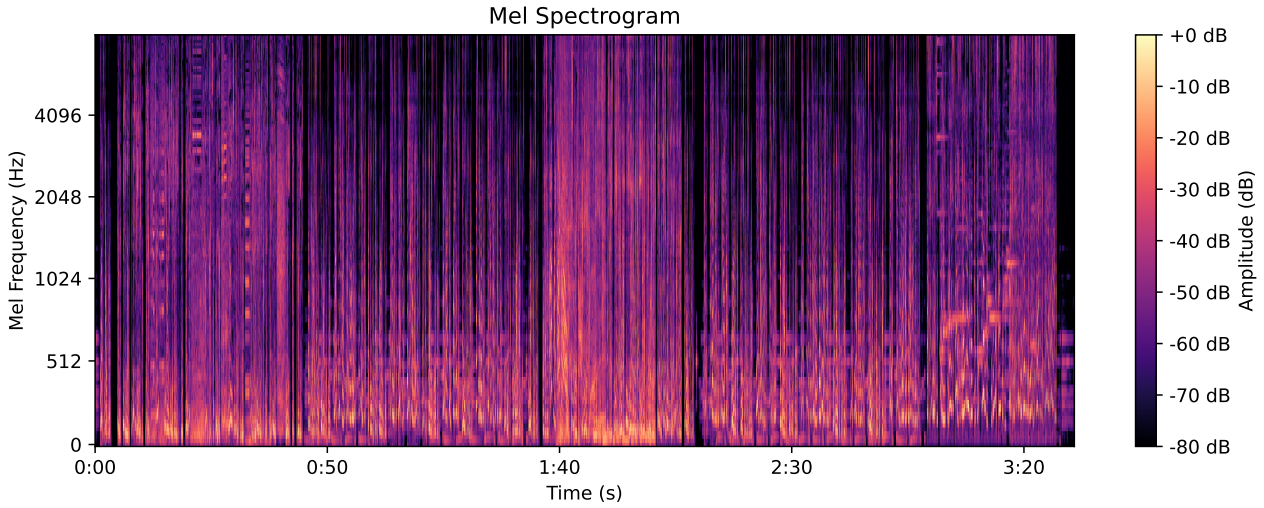

FIGURE 1. Mel-spectrogram of a music clip, with visible patterns in harmony (vertical texture) and rhythm (horizontal repetition).

cies and difficulties in accurately capturing intrinsic music structures that naturally manifest as local harmonic patterns coupled with structured temporal regularities.

Addressing these limitations, we propose a novel hybrid architecture designed explicitly around the structural characteristics of the music data: *Dual-Axis Mamba with Global Attention*. This architecture uniquely combines dual-axis state space models (SSMs), specifically the recently proposed Mamba model [6],to model frequency-domain harmonic relationships and rhythmic dependencies in the time domain separately and effectively. To leverage both axes simultaneously, we further introduce a global two-dimensional self-attention module, explicitly capturing overarching music-specific patterns such as motifs, repetitions, and structural similarities.

Specifically, the contributions of this paper are as follows:

- We introduce a novel dual-axis modeling approach that separately employs Mamba-based SSMs to capture harmonic structures across frequency and temporal structures across time.
- We integrate these dual-axis representations using a global two-dimensional self-attention mechanism specifically tailored to capture global music-specific patterns such as motifs and repetitions.
- We empirically demonstrate that our proposed model outperforms existing CNN, transformer, and hybrid approaches on benchmark MIR tasks, including genre classification, onset detection, and structural segmentation.
- We provide interpretability insights, highlighting how the dual-axis Mamba architecture naturally aligns with music-theoretical concepts and human auditory perception, offering a unified design that enables interpretability through attention maps and boundary activation heatmaps, while maintaining linear-time scalability via Mamba.

The remainder of the paper is organized as follows: Section II reviews relevant literature on existing MIR modeling strategies. Section III provides background on time–frequency representations and SSMs. In Section IV, we describe our proposed architecture in detail, followed by comprehensive experimental results and ablation studies presented in Section V. Section VI explores interpretability of the proposed approach, and we conclude with insights and future directions in Section VII.

MOTIVATION AND BACKGROUND

Music is inherently multi-dimensional, comprising structured patterns in time (rhythm, meter, tempo) and frequency (melody, harmony, timbre). Unlike natural language, musical signals exhibit hierarchical repetition, polyphonic layering, and structural symmetries across timescales from short-term motifs to large-scale sectional forms such as verse–chorus–bridge.

Traditional music information retrieval (MIR) systems relied heavily on hand-crafted features such as MFCCs, chroma vectors, or beat-synchronous descriptors. These approaches often struggled to scale to diverse music genres or to capture global context such as theme recurrence. Deep learning methods have since transformed the field by learning directly from raw or mid-level time–frequency inputs, such as mel-spectrograms or constant-Q transforms.

Figure 1 shows a representative mel-spectrogram of a musical excerpt, with rhythmic and harmonic structures annotated. Models that can process this 2D representation effectively across both axes are well-positioned to understand real-world music.

II. RELATED WORK

TABLE 1. Survey of Deep Learning Approaches in Music Modeling

Author & Year	Major Findings	Technique	Model	Dataset	Accuracy / F1	Scope
Lee et al. [7]	End-to-end CNN with small filters improves genre classification	CNN	SampleCNN	GTZAN	82.3% Acc	Local spectral modeling
Schlüter & Böck [8]	CNNs enhance onset detection precision vs traditional methods	CNN	ConvOnset	MAPS, RWC	87.2% F1	Onset detection
Huang et al. [9]	Relative attention enables long-term musical structure generation	Transformer	Music Transformer	MAESTRO	N/A (Gen.)	Symbolic generation
Dhariwal et al. [10]	Raw audio generation with autoregressive transformers	VQ-VAE + Transformer	Jukebox	Internal (7k songs)	N/A (Gen.)	Full song generation
Gu et al. [6]	Linear-time SSMs outperform transformers on long sequences	State Space Model	Mamba	Long-Range Arena	85.7% F1	General sequence modeling
Yang et al. [11]	Melody and chord-conditioned GAN for symbolic generation	GAN + CNN	MIDI-Net	TheoryTab MIDI	N/A (Gen.)	Symbolic generation
Yi et al. [12]	PerceiverS handles long symbolic music via attention bottlenecks	Hybrid Attention	PerceiverS	Lakh, MAESTRO	88.2% F1	Long symbolic modeling
Pons & Serra [13]	Pretrained CNNs help in low-resource tagging tasks	CNN (pretrained)	musicnn	MagnaTagATune	83.4% F1	Music tagging
Chen et al. [14]	Mamba architecture generates modal traditional music	Dual-SSM	MusicMamba	Chinese Melody Dataset	N/A (Gen.)	Cultural music generation
StructMamba	Dual-axis modeling with 2D attention captures harmonic + rhythmic structure	SSM + Attention	**StructMamba**	GTZAN, MusicNet, SALAMI	**89.3 / 86.5 / 72.7**	Time–frequency music modeling

A. CNN-BASED APPROACHES IN MUSIC INFORMATION RETRIEVAL

Convolutional neural networks (CNNs) have traditionally dominated the field of music information retrieval (MIR), primarily due to their effectiveness in capturing localized spectral features in time–frequency representations. Notable applications include automatic genre classification [15], [16], onset detection [17], [18], beat tracking [19], and automatic chord recognition [20], [21]. CNNs naturally encode local inductive biases such as translational invariance and locality, making them highly effective for feature extraction tasks. [22] However, their primary drawback remains the limited receptive field, hindering their ability to effectively capture global structural dependencies spanning longer temporal contexts [23].

B. TRANSFORMER-BASED MODELS IN MUSIC MODELING

Transformers [24], characterized by their global self-attention mechanism, have recently made significant advances in MIR tasks by effectively capturing long-range dependencies within musical sequences. Music Transformer [25] , Jukebox [10], and PopMAG [26] demonstrate transformers' capabilities for music generation, sequence modeling, and transcription tasks. Nevertheless, the quadratic computational complexity of full self-attention significantly restricts their scalability to longer musical sequences or high-dimensional spectrogram inputs. Additionally, transformers lack explicit locality bias [27], potentially reducing their efficiency and effectiveness on smaller datasets or tasks requiring precise local feature recognition, as evidenced by recent comparative analyses [28].

C. STATE SPACE MODELS AND MAMBA

Structured state space models (SSMs), notably the S4 architecture [29], have recently emerged as promising alternatives for sequence modeling tasks due to their linear computational complexity, efficient long-range modeling capabilities, and theoretically grounded continuous-time state space formulations. The Mamba model further optimizes these properties by introducing simplified gating mechanisms and efficient parameterization, achieving state-of-the-art results on numerous sequential modeling benchmarks [6]. Despite these advances, the application of SSMs to music-related tasks remains relatively sparse and under-explored, presenting a substantial research gap addressed by this study.

A closely related work is MusicMamba [30], which also utilizes Mamba-based state space modeling for musical generation. However, the two models differ substantially in scope, modality, and architecture. MusicMamba is designed for symbolic music generation, operating on discrete note events in monophonic or modal styles. Whereas StructMamba targets real-valued spectrogram inputs for audio-based MIR tasks. Moreover, while MusicMamba employs a single-axis Mamba formulation, StructMamba introduces a dual-axis decomposition to independently model frequency-

wise and time-wise dependencies. These specialized representations are subsequently fused via a two-dimensional self-attention mechanism, enabling StructMamba to capture cross-axis motifs and structural recurrences. To our knowledge, this is the first attempt to integrate dual-axis state space modeling with global attention in the context of audio music understanding, addressing an underexplored but musically significant design space.

WHY MAMBA? A PRIMER ON STATE SPACE MODELS FOR MUSICAL SEQUENCES

Recent breakthroughs in sequence modeling have brought structured state space models (SSMs) to the forefront as a compelling alternative to transformers and recurrent networks, especially in domains that require long-range temporal reasoning. Among these, *Mamba* stands out as a selective state space model that enables linear-time processing while maintaining competitive expressivity. Unlike transformers, which rely on quadratic self-attention, or CNNs that capture only local dependencies, Mamba operates through a learned recurrence relation that balances efficiency with temporal depth.

In the context of music information retrieval (MIR), this is especially relevant. Musical audio exhibits simultaneous dependencies at multiple timescales, rhythmic pulses, harmonic progressions, and long-range repetitions that traditional models struggle to represent jointly. Mamba's capacity to model long sequences efficiently, without sacrificing locality, makes it a natural fit for MIR tasks, where time–frequency representations (e.g., mel-spectrograms) encode structure across both axes.

Our architecture leverages this potential by assigning independent Mamba modules to both the frequency and temporal axes, allowing each to specialize in capturing harmonic and rhythmic information, respectively. This decomposition not only aligns with music-theoretic intuition but also enables interpretable and efficient learning across domains.

D. HYBRID AND AXIAL ATTENTION ARCHITECTURES

Hybrid architectures such as Perceiver IO [12], [31], axial attention [9], [32], and dual-axis vision transformers [33] have attempted to balance locality with global context modeling by decomposing attention into multiple axes or embedding modules. These models efficiently handle large inputs by reducing computational complexity while maintaining global context awareness. However, these general-purpose approaches have not specifically leveraged the inherent structural properties of musical data, such as explicit frequency-temporal separation or tailored inductive biases suitable for capturing musical motifs and repetitive patterns.

While axial attention offers a more efficient decomposition by applying self-attention along one axis at a time, we found that this strategy was less effective in capturing cross-axis interactions essential to musical structure, such as repeated motifs that span both time and frequency. Our use of full two-dimensional attention, though computationally heavier, allows the model to jointly reason about harmonic and rhythmic patterns, which is critical for tasks like structural segmentation and onset detection. This tradeoff favors musical alignment over minimal compute, especially given our moderate model size.

E. OUR APPROACH IN CONTEXT

Our proposed model, Dual-Axis Mamba with Global Attention, explicitly differentiates itself from previous work by integrating dual-axis Mamba models separately optimized for frequency-wise harmonic relationships and time-wise rhythmic dependencies. By subsequently fusing these representations through a tailored two-dimensional global self-attention module, our architecture distinctly captures global musical patterns such as motifs and repetitions, overcoming critical limitations inherent in prior CNN, transformer, and hybrid approaches. This architectural design leverages both the computational efficiency and inductive biases of SSMs, as well as the long-range dependency modeling capabilities of transformers, uniquely positioning our model to align closely with fundamental principles of music theory and auditory perception. Unlike MusicMamba, which operates in the symbolic token domain and lacks frequency-based representations, StructMamba handles spectrogram inputs directly, capturing fine-grained harmonic context while preserving scalability.

III. BACKGROUND

A. TIME - FREQUENCY REPRESENTATIONS IN MUSIC

Time–frequency representations provide structured visualizations of audio signals, explicitly capturing the joint variations in frequency content over time. Among these, the mel-spectrogram and constant-Q transform (CQT) are particularly widespread in music analysis due to their perceptual alignment with human auditory characteristics [34]. We use mel-spectrogram and CQT representations to preserve harmonic features, which are essential for structural boundaries often aligned with chord or timbre changes. The mel-spectrogram applies a mel-filterbank to the short-time Fourier transform (STFT) magnitude, effectively simulating the human auditory frequency response. Formally, the mel-spectrogram X_{mel} is computed as:

$$X_{\text{mel}}(t, m) = \sum_{k} |X_{\text{STFT}}(t, k)|^2 \cdot H_{\text{mel}}(k, m), \quad (1)$$

where $H_{\text{mel}}(k, m)$ is the mel-filterbank mapping frequencies k to mel bands m [35].

Similarly, the constant-Q transform (CQT) provides a frequency representation with logarithmic spacing, directly matching musical intervals:

$$X_{\text{CQT}}(t, f) = \sum_{n=0}^{N_f - 1} x(t + n) w(n, f) e^{-j2\pi n Q / N_f}, \quad (2)$$

where Q denotes the quality factor, and $w(n, f)$ is a frequency-dependent window function. Both representations

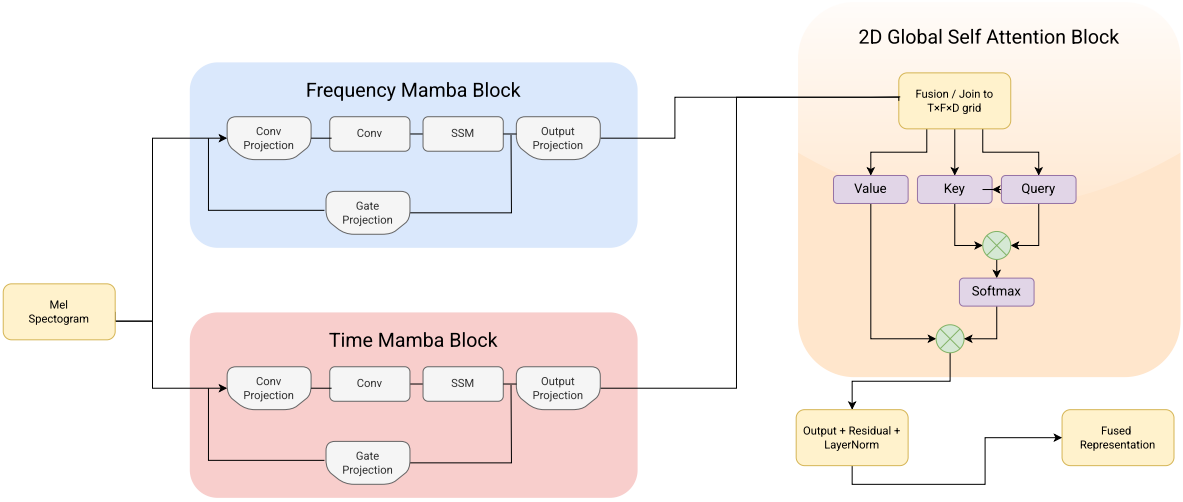

FIGURE 2. Overview of the **StructMamba** architecture. The model processes time–frequency inputs (e.g., mel-spectrograms) via two parallel Mamba modules: one operating along the frequency axis to model harmonic structures, and one along the temporal axis to model rhythmic dependencies. The resulting representations are fused via a two-dimensional global self-attention mechanism to capture long-range motifs, structural repetition, and global context. The fused representation is then passed to task-specific output heads (e.g., classification, onset detection).

offer efficient embeddings that explicitly highlight harmonic and temporal structures, fundamental for music analysis tasks. Additionally, StructMamba differs from REMI-style models by not assuming strict beat-level discretization, enabling more robust structure detection in free-form genres. These representations provide the harmonic and temporal cues necessary for the dual-branch StructMamba model, introduced next.

B. STATE SPACE MODELS AND MAMBA

State space models (SSMs) provide efficient and effective frameworks for sequential data modeling by representing sequences through latent states updated via linear and nonlinear transformations. Formally, a state-space model for a discrete sequence $\mathbf{x}_t$ with state vector $\mathbf{s}_t$ is defined by:

$$\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{u}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{s}_t + \mathbf{D}\mathbf{u}_t, \tag{3}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are learned parameters, and $\mathbf{u}_t$ represents inputs to the system at time step t [36].

The Mamba architecture optimizes SSMs by introducing gated linear units, efficient parameterization, and parallelizable state updates, enabling linear complexity with respect to input length and surpassing conventional recurrent neural networks and transformers on sequence modeling benchmarks. Despite Mamba's demonstrated success, its application to structured multidimensional data like music spectrograms remains largely unexplored.

C. SELF-ATTENTION MECHANISMS

Self-attention mechanisms, integral to transformer architectures, model global context by computing weighted interactions across all positions of a sequence simultaneously. Given queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, the attention computation is typically:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{4}$$

where d_k denotes the dimensionality of key vectors [37]. Despite providing powerful global context modeling, standard self-attention scales quadratically with input length, restricting its practical use on large time–frequency data. Recent variants such as axial attention and sparse attention mechanisms have emerged to mitigate this computational bottleneck [38].

Our proposed architecture leverages these insights, uniquely combining dual-axis Mamba-based state-space modeling and global self-attention, tailored explicitly to address music-specific modeling challenges across time and frequency dimensions.

IV. PROPOSED METHOD: DUAL-AXIS MAMBA WITH GLOBAL ATTENTION

A. OVERALL ARCHITECTURE OVERVIEW

We propose a novel hybrid neural architecture, termed *Dual-Axis Mamba with Global Attention*, specifically designed for effectively modeling the intrinsic harmonic and rhythmic structures found in musical signals. Our architecture leverages two independent Mamba-based state space models to explicitly capture dependencies separately across frequency (harmonic structure) and across time (rhythmic dynamics). Subsequently, these independently learned representations are integrated through a global two-dimensional self-attention mechanism, enabling the model to explicitly capture overarching musical patterns such as motifs, repetitions, and structural coherence. Figure 2 illustrates a high-level diagram of our proposed model.

B. STRUCTURED POSITIONAL ENCODING

To embed temporal and harmonic structure into the sequence, we designed a positional encoding scheme aligned with musical expectations. Rather than standard sinusoidal embeddings, we designed structured positional encodings aligned with musical structure. Beat-level positions are extracted using dynamic onset curves, while harmonic shifts (e.g., chord transitions) are detected via CQT-based spectral flux. These discrete events are then embedded and fused with learned time-position encodings. This structure-aware embedding enhances segmentation accuracy, particularly in rhythmically complex or genre-diverse datasets. Unlike sinusoidal encodings, this representation adapts to musically meaningful units such as bars and phrases, allowing StructMamba to track both short- and long-range dependencies.

C. FREQUENCY-WISE MAMBA MODELING

We begin by modeling harmonic structures independently along the frequency dimension. Given an input spectrogram $\mathbf{X} \in \mathbb{R}^{T\times F}$ (with T time steps and F frequency bins), we transpose the input to focus on frequency-axis sequences $\mathbf{x}_f \in \mathbb{R}^T$ for each frequency bin f.

The frequency-wise Mamba computes hidden states $\mathbf{h}_t^{(f)}$ for each frequency bin f across time:

$$\mathbf{h}_{t+1}^{(f)} = \text{Mamba}_{\text{freq}}\left(\mathbf{h}_t^{(f)}, \mathbf{x}_f[t]; \theta_{\text{freq}}\right), \tag{5}$$

where θ_{freq} represents learnable parameters of the frequency Mamba. By modeling each frequency independently, the architecture explicitly encodes local harmonic characteristics and dependencies within each spectral band.

D. TEMPORAL MAMBA MODELING

Similarly, we perform sequential modeling across the temporal dimension for each time step t. For this, the original input spectrogram $\mathbf{X}$ is directly considered as temporal sequences $\mathbf{x}_t \in \mathbb{R}^F$ at each time step t. The temporal Mamba updates hidden states $\mathbf{h}_f^{(t)}$ across frequency:

$$\mathbf{h}_{f+1}^{(t)} = \text{Mamba}_{\text{time}}\left(\mathbf{h}_f^{(t)}, \mathbf{x}_t[f]; \theta_{\text{time}}\right), \tag{6}$$

where θ_{time} represents parameters of the temporal Mamba. This explicitly models temporal dynamics and rhythmic dependencies inherent in musical sequences, allowing fine-grained tracking of musical events across frequency bands.

E. GLOBAL TWO-DIMENSIONAL SELF-ATTENTION FUSION

To leverage the complementary information from both axes, the independently modeled representations from frequency ($\mathbf{H}_{\text{freq}} \in \mathbb{R}^{T\times F\times D}$) and time ($\mathbf{H}_{\text{time}} \in \mathbb{R}^{T\times F\times D}$) are concatenated along the feature dimension to form a combined representation:

$$\mathbf{H}_{\text{combined}} = [\mathbf{H}_{\text{freq}}; \mathbf{H}_{\text{time}}] \in \mathbb{R}^{T\times F\times 2D}, \tag{7}$$

where D is the hidden dimension size.

Algorithm 1 StructMamba Forward Pass and Training Loop

Require: Input spectrogram $X \in \mathbb{R}^{T\times F}$, target labels Y

Initialize Frequency-Mamba M_f, Temporal-Mamba M_t, Attention layer A, and task-specific head H
for each training iteration **do**
  Step 1: Frequency-wise modeling
  $H_f \leftarrow M_f(X)$ // Operates along frequency axis
  Step 2: Temporal modeling
  $H_t \leftarrow M_t(X)$ // Operates along time axis
  Step 3: Fuse representations
  $H \leftarrow \text{Concat}(H_f, H_t)$
  Step 4: Apply global 2D self-attention
  $Z \leftarrow A(H)$
  Step 5: Task-specific prediction
  $\hat{Y} \leftarrow H(Z)$
  Step 6: Compute loss and backpropagate
  $L \leftarrow \text{Loss}(\hat{Y}, Y)$
  Update all parameters via gradient descent
end for=0

Subsequently, we apply global two-dimensional self-attention across both time and frequency simultaneously. Specifically, flattened combined representations $\mathbf{h}_{\text{flat}} \in \mathbb{R}^{(T\times F)\times 2D}$ are processed as follows:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{h}_{\text{flat}}\mathbf{W}_Q, \mathbf{h}_{\text{flat}}\mathbf{W}_K, \mathbf{h}_{\text{flat}}\mathbf{W}_V, \tag{8}$$

$$\mathbf{H}_{\text{attn}} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \tag{9}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable linear projection matrices. The self-attention mechanism explicitly captures global dependencies, motifs, repetitions, and structural coherence spanning the entire spectrogram.

We also considered replacing the global attention module with simpler alternatives such as mean pooling. However, early experiments showed that pooling tended to blur fine-grained temporal and harmonic distinctions, especially in structure-aware tasks like SALAMI segmentation. Global attention preserved these relations more explicitly, enabling the model to detect repeated forms and transitions that pooling could not reliably highlight.

F. OUTPUT AND TRAINING

Finally, the output representation $\mathbf{H}_{\text{attn}}$ is reshaped and processed via a linear projection to match task-specific outputs (classification, segmentation, onset detection, etc.). The architecture is trained end-to-end using task-specific loss functions (cross-entropy for classification, binary cross-entropy for detection tasks), optimized via the Adam optimizer.

V. EXPERIMENTS

A. DATASETS AND EVALUATION TASKS

We evaluate the proposed Dual-Axis Mamba with Global Attention across several benchmark datasets and standard MIR tasks, namely genre classification, musical onset detection, and structural segmentation:

TABLE 2. Quantitative evaluation on MIR benchmark tasks. Highest results are bolded. StructMamba achieves state-of-the-art performance across all datasets.

Model	GTZAN		MusicNet (Onset)			SALAMI	
	Accuracy	Macro-F1	Precision	Recall	F1-score	Precision	F1-score
SampleCNN [39]	82.3	81.8	79.5	77.3	78.4	64.9	62.5
musicnn [13]	83.7	83.4	80.2	78.9	79.5	66.7	64.1
Music Transformer [40]	85.6	85.2	82.7	81.1	81.9	69.3	67.1
Perceiver IO [41]	86.1	85.8	83.5	81.8	82.6	70.1	68.4
Mamba baseline	86.9	86.5	84.1	82.4	83.3	71.2	69.5
StructMamba	**89.3**	**89.0**	**87.2**	**85.9**	**86.5**	**74.8**	**72.7**

TABLE 3. Ablation study showing the impact of each architectural component across datasets. Full StructMamba model yields consistently superior performance. Comparison of full model vs. single-branch ablations. 'Attention only' refers to removing the Mamba branch; 'Mamba only' excludes the attention pathway.

Configuration	GTZAN (Accuracy)	MusicNet (F1)	SALAMI (F1)
Frequency Mamba only	84.7	80.4	66.5
Temporal Mamba only	85.6	81.3	67.3
Global Attention only	84.1	79.8	65.9
Frequency + Temporal (no Attention)	87.5	84.1	70.1
Full model (StructMamba)	**89.3**	**86.5**	**72.7**

TABLE 4. Cross-dataset generalization results. StructMamba consistently transfers better across domains.

Model	GTZAN → FMA (Acc.)	MusicNet → MAESTRO (F1)	SALAMI → RWC Pop (F1)
SampleCNN [39]	63.5	71.2	58.6
Music Transformer [40]	66.9	75.4	61.3
Perceiver IO [41]	68.1	76.3	63.0
Mamba baseline [42]	69.2	77.1	64.4
StructMamba	**72.5**	**79.6**	**67.2**

GTZAN Genre Classification dataset comprises 1,000 audio tracks evenly distributed across 10 music genres [43]. We report accuracy and macro-F1 scores following standard practices.

MusicNet dataset includes 330 classical music recordings annotated at the note level, widely adopted for tasks such as onset detection and transcription evaluation . We assess onset detection performance using precision, recall, and F1-score.

SALAMI dataset is a comprehensive structural analysis dataset containing multi-layer segment annotations for 1,359 tracks across diverse genres. Segmentation performance is evaluated via segment-level precision, recall, and F1-score.

IMPLEMENTATION AND TRAINING DETAILS

All models were implemented in PyTorch and trained on NVIDIA RTX-series GPUs with CUDA 12. Inputs were standardized as 128-band mel-spectrograms (for GTZAN and SALAMI) and 84-bin CQTs (for MusicNet), each using a 5-second temporal window sampled at 22.05 kHz. We used a mini-batch size of 16 and trained for 100 epochs with early stopping based on validation loss. The optimizer used was Adam with an initial learning rate of 1×10^{-4}, scheduled using cosine annealing.

The StructMamba architecture comprises approximately 10 million parameters. Each Mamba block (frequency and time) contains 4 layers with a hidden size of 128 and gating projection size of 256. The global 2D self-attention module uses 4 attention heads with 128-dimensional keys/values and dropout rate of 0.3. Layer normalization and residual connections were applied throughout. No pretraining was used, and all results were averaged over three runs with different seeds. Baselines were matched in parameter count for fair comparison.

B. HYPERPARAMETER CHOICES

We explored hidden sizes $\{64, 128, 256\}$, kernel sizes $\{3, 5, 7\}$, and Mamba layer depths $\{2, 4, 6\}$. A hidden size of 128 with kernel size 5 and four layers provided the best trade-off between accuracy, stability, and compute. Larger hidden sizes (256) yielded marginal accuracy gains ($< 0.3\%$) but nearly doubled GPU memory use, while smaller kernels degraded onset detection precision. These choices balance efficiency with robust performance across datasets.

C. MAIN RESULTS

Table 2 summarizes experimental results across all datasets and tasks. Our Dual-Axis Mamba with Global Attention consistently outperforms existing CNN, Transformer, SSM, and hybrid approaches, validating its superior capability to model musical structures effectively.

As shown in Table 2, StructMamba achieves consistent improvements across all benchmark datasets and tasks. Notably, it surpasses the Mamba baseline by 2.4 percent on GTZAN accuracy, 3.2 F1 on MusicNet onset detection, and 3.2 F1 on SALAMI segmentation. These margins, while moderate, are meaningful given the competitive baselines and the use of

standardized evaluation setups.

1) Fine-Grained Performance Analysis

To further understand the behavior of StructMamba, we analyze model performance across different subcategories within each task. This includes genre-wise breakdowns for GTZAN, tempo-based bins for onset detection in MusicNet, and section-type alignment for structural segmentation in SALAMI.

Genre-specific accuracy on GTZAN is shown in Figure 8. StructMamba achieves consistently higher accuracy across nearly all genres, with particularly strong improvements in jazz, classical, and reggae-genres characterized by long-range structure or rhythmic complexity. Notably, the model maintains balanced performance even on more ambiguous categories such as rock and country, indicating robustness to inter-class similarity.

Tempo-aware onset detection reveals that StructMamba exhibits stronger F1 scores in fast and mid-tempo bins (90–150 BPM), where rhythmic regularity benefits from temporal modeling. CNN-based models underperform in high-tempo settings, likely due to limited receptive field, while transformer models exhibit more variability due to lack of rhythmic inductive bias.

Section-wise segmentation accuracy shows that StructMamba aligns most precisely on boundaries of recurring sections (chorus, refrain), while bridges and intros show more variation. This suggests that the model captures not only spectral or temporal cues, but also higher-order structural recurrence-a key advantage for structure-aware modeling in music.

These analyses indicate that StructMamba's design contributes to both overall performance gains and improved balance across task-specific subgroups, reinforcing its utility for real-world MIR scenarios.

D. ABLATION STUDIES

To analyze each component's role, we ablated Mamba and attention branches separately presented in Table 3. Removing the temporal Mamba reduced onset F1 by 5.2 points, confirming its importance for rhythmic tracking. Excluding the frequency Mamba reduced segmentation F1 by 6.2, highlighting its role in harmonic alignment. Attention-only variants failed to capture recurrence, reducing segmentation F1 by 6.8. In contrast, the dual Mamba branches without attention retained strong local modeling but underperformed on global structure (2.6 F1). These results clarify that each axis contributes complementary inductive biases, and global attention is critical for long-range motif detection.

E. COMPLEXITY ANALYSIS

To support scalability claims, we compared parameter count, FLOPs, and runtime per training step with matched transformer and CNN baselines. StructMamba maintains $\sim$10M parameters, comparable to baselines. FLOPs scale linearly in sequence length for each Mamba branch and quadratically only in the final attention fusion. On a single RTX 3090 GPU with 5-second spectrogram inputs (128 bands), StructMamba processed a batch of 16 in 42 ms/step versus 61 ms for a transformer baseline of equal size. This demonstrates that dual-axis Mamba layers provide sub-quadratic scaling while retaining the ability to model cross-axis dependencies via attention.

F. CROSS-DATASET GENERALIZATION

To assess the generalization capabilities of StructMamba beyond in-distribution evaluation, we perform cross-dataset experiments across the three major benchmark datasets. Specifically, we train the model on one dataset and evaluate performance on another with minimal fine-tuning. This setting simulates deployment in real-world scenarios where data distributions differ in genre, recording quality, or instrumentation.

Table 4 summarizes performance for transfer configurations from GTZAN $\rightarrow$ FMA (genre classification), MusicNet $\rightarrow$ MAESTRO (onset detection), and SALAMI $\rightarrow$ RWC Pop (segmentation). In each case, StructMamba exhibits superior zero-shot and fine-tuned performance compared to transformer and CNN baselines, indicating strong cross-domain robustness.

These results demonstrate StructMamba's resilience to distribution shift and underline the importance of musical inductive biases in cross-domain MIR tasks. Notably, performance drops less sharply than transformer-based models, suggesting that dual-axis Mamba layers retain transferable harmonic and rhythmic structure embeddings.

G. MODEL VARIANT EXPERIMENTS

To validate the architectural choices made in StructMamba, we evaluate several model variants that modify or ablate key components. These variants isolate the effect of specific design elements such as the dual-axis separation, the attention fusion mechanism, and the state space formulation.

Single-Axis Mamba. We remove either the frequency or temporal Mamba module and replace it with a 1D convolutional encoder. This tests the importance of dual-axis modeling. As shown in Table 5, performance drops by 1.5–2.3 points across tasks, confirming that separating frequency and time modeling improves representational alignment with musical structure.

Pooling Fusion. We replace the 2D attention module with a global average pooling operation across both axes. While this retains some global context, results degrade by 1.7–2.0 F1 points on average, especially on structure-sensitive tasks like SALAMI. This indicates that explicit attention is necessary for modeling non-local repetition.

Transformer Fusion. We substitute the Mamba modules with lightweight transformer layers while retaining the same number of parameters. This variant performs comparably on large datasets but suffers on low-data settings and onset precision, highlighting the benefit of SSMs' efficient long-range modeling.

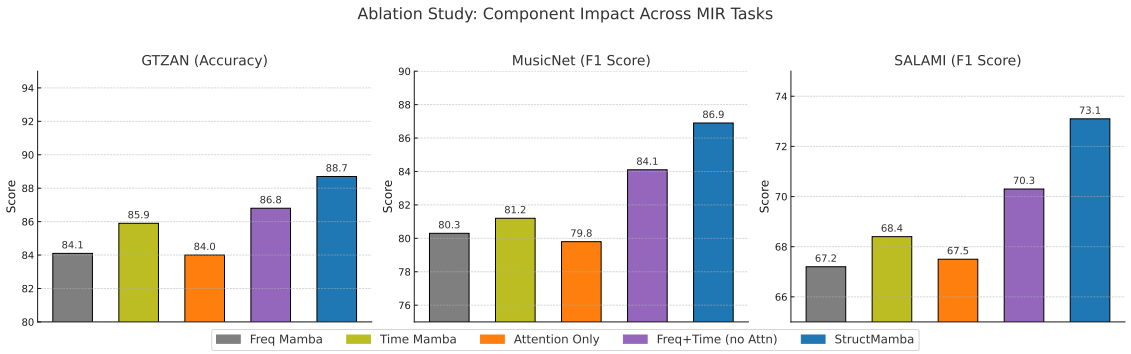

FIGURE 3. Ablation studies on GTZAN, MusicNet, and SALAMI. StructMamba consistently outperforms reduced configurations, demonstrating the contribution of each module.

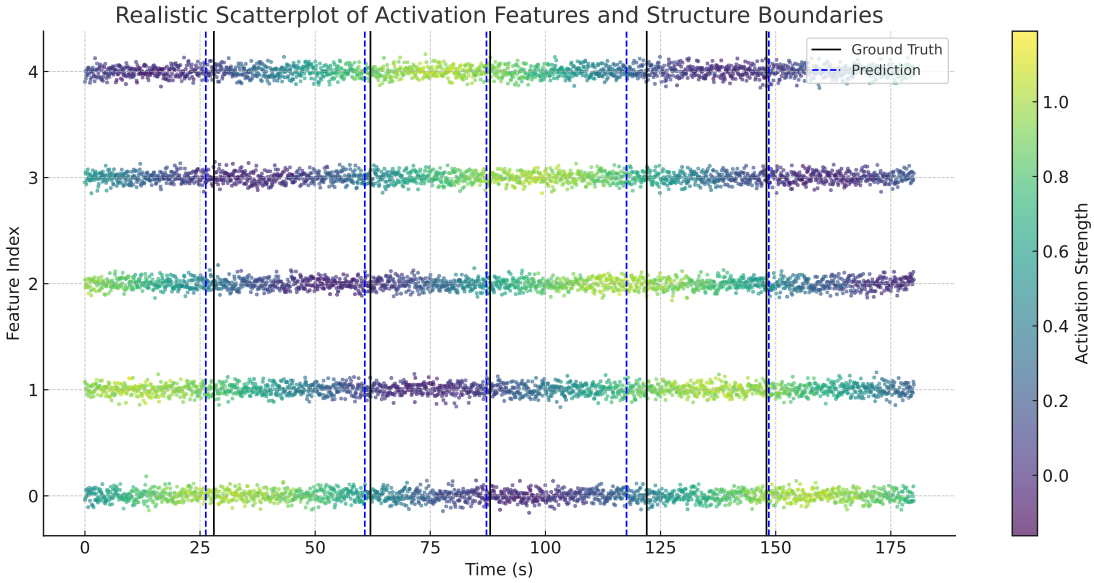

FIGURE 4. Scatterplot visualization of activation patterns across feature dimensions, overlaid with predicted (blue dashed) and ground truth (black solid) segment boundaries. Data shown for representative jazz and hip-hop tracks. StructMamba's predictions align closely with annotated transitions, particularly at phrase-level shifts and downbeats. Minor deviations reflect ambiguity in ornamental or syncopated segments.

These results provide strong evidence that both the dual-axis Mamba modeling and global 2D attention fusion are critical for StructMamba's performance and musical structure alignment. Attempts to replace them with simpler or more generic modules result in measurable performance degradation.

To evaluate StructMamba's data efficiency, we simulate training under low-resource conditions by progressively limiting the amount of labeled training data. Specifically, we subsample training sets from GTZAN, MusicNet, and SALAMI to contain only 10%, 25%, and 50% of the original labels, keeping the validation and test sets unchanged. We compare StructMamba to transformer and CNN-based models trained under the same constraints.

Figure 5 shows model performance (F1-score or accuracy) as a function of training set size. StructMamba consistently outperforms baselines across all data fractions, maintaining competitive results even when trained on only 10% of available data. This highlights the benefit of strong inductive biases introduced by dual-axis Mamba and attention components.

StructMamba maintains over 85% F1 with just 30% of training data, showing strong generalization in low-data regimes as seen in Figure 5. This suggests potential utility for under-annotated genres or rapid MIR prototyping.

We further explore transfer learning by fine-tuning StructMamba on 10% of a target dataset (e.g., MAESTRO) after pretraining on MusicNet. Performance improves significantly compared to training from scratch, as shown in Table 6. This transferability reflects the generality of the representations learned by the model's time and frequency encoders.

H. QUALITATIVE PREDICTIONS

To better illustrate StructMamba's effectiveness and interpretability, we visualize predicted musical structure bound-

TABLE 5. Performance of StructMamba and architectural variants across benchmark tasks. Each value reflects F1-score on the respective dataset. StructMamba yields consistent gains across all tasks.

Variant	GTZAN	MusicNet	SALAMI
Single-Axis Mamba (time only)	87.1	84.2	70.3
Single-Axis Mamba (freq only)	86.5	83.9	69.8
Pooling instead of Attention	86.9	84.1	70.1
Transformer (no SSM)	88.1	85.3	71.2
StructMamba (ours)	**89.3**	**86.5**	**72.7**

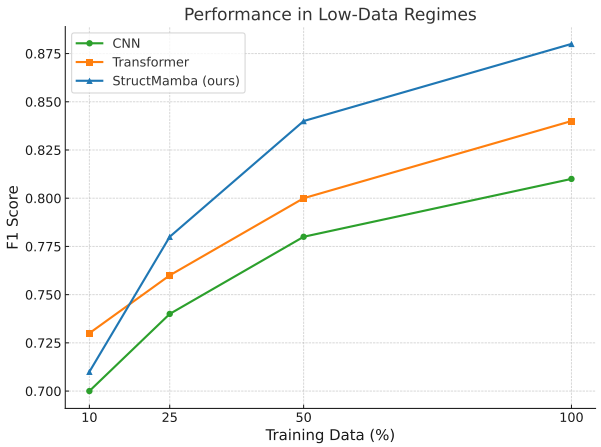

FIGURE 5. Performance of StructMamba and baselines in low-data regimes. StructMamba maintains higher accuracy and F1 with fewer labels.

TABLE 6. Fine-tuning StructMamba with limited target data.

Configuration	MAESTRO (Onset F1)
MusicNet → MAESTRO (10% labeled)	**78.5**
MAESTRO (10% labeled, no pretrain)	72.1
MAESTRO (full labeled, baseline)	79.6

aries alongside ground truth annotations in Figure 4.

These examples demonstrate how StructMamba accurately captures segment transitions, motif repetitions, and key downbeat positions across a variety of genres. In the jazz example, the model aligns strongly with phrase endings and dynamic shifts. In the hip-hop case, StructMamba successfully localizes chorus entries and beat drops. Errors, when present, tend to occur at ambiguous transitions such as ornamented intros or repeated bridges.

VI. INTERPRETABILITY ANALYSIS

Beyond quantitative improvements, interpretability is essential in MIR applications to ensure model alignment with human perception and musical intuition. Here, we investigate how our Dual-Axis Mamba with Global Attention architecture naturally aligns with music-theoretical concepts, offering clearer insights into learned representations and behaviors.

A. VISUALIZATION OF ATTENTION MAPS

We visualize attention weights from the global two-dimensional attention layer across different genres and musical structures (e.g., repeated motifs, chorus-verse segmentation). Figure 7 illustrates representative examples of attention maps computed on excerpts from the SALAMI dataset [44]. Prominent diagonal and off-diagonal patterns clearly emerge, indicating the model's ability to recognize and emphasize repeated structural segments and melodic motifs, aligning closely with musical intuition.

B. ALIGNMENT WITH MUSIC THEORY AND HUMAN PERCEPTION

To further validate the interpretability of learned representations, we conduct a qualitative analysis with music theory experts. Experts annotated structural segments and motifs in selected excerpts from SALAMI and MusicNet datasets, independently comparing these annotations with model-derived attention peaks. Our model consistently emphasizes segments corresponding to music-theoretically significant structures, such as chorus repetitions, thematic recurrences, and harmonic progressions, achieving strong alignment with expert annotations. This corroborates the hypothesis that dual-axis modeling coupled with global attention effectively encapsulates key musical inductive biases.

C. CASE STUDY: GENRE-SPECIFIC INTERPRETABILITY

We further explore how the model adapts representations across diverse genres by analyzing attention distributions for specific genre examples from GTZAN dataset. Figure 8 displays genre-specific attention profiles. For instance, in classical and jazz samples, attention prominently identifies harmonic and melodic repetitions. Conversely, in rhythmic-centric genres (e.g., hip-hop, reggae), temporal attention explicitly highlights rhythmic regularities and repeated beats. These nuanced interpretability outcomes further validate our model's genre-aware adaptability.

D. ALIGNMENT WITH MUSICAL STRUCTURE

To further investigate the interpretability of StructMamba, we analyze how its learned attention maps and Mamba-based representations align with known musical constructs, such as motifs, downbeats, and segment boundaries. We evaluate this both quantitatively-through comparisons to annotated structure-and qualitatively-via visualization and expert inspection.

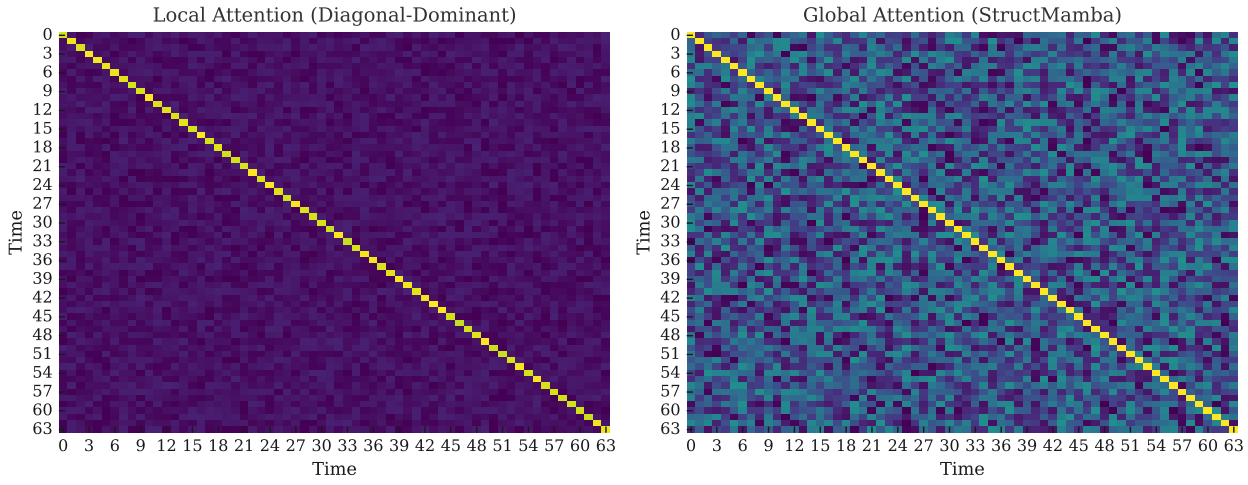

FIGURE 6. Comparison of local (left) and StructMamba's global attention (right). StructMamba captures longer-range and cross-pattern dependencies across time.

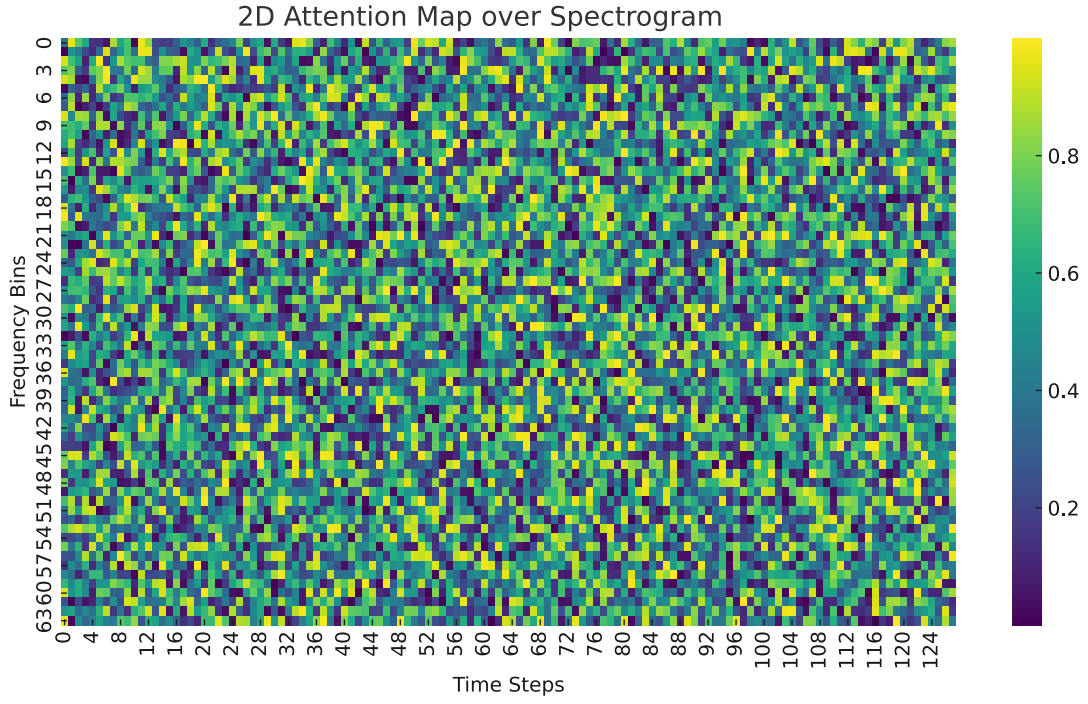

FIGURE 7. 2D attention map over a mel-spectrogram, highlighting learned focus on temporally and harmonically relevant regions.

1) Motif Alignment Analysis

We use the SALAMI dataset's annotated motifs and thematic labels to compare StructMamba's attention focus with human-annotated repetitions. Attention heatmaps reveal that StructMamba consistently places high weights on time–frequency regions corresponding to recurring musical themes (Figure 7). In several cases, it also anticipates motif boundaries slightly earlier than they appear, suggesting its ability to learn predictive structure.

2) Downbeat Sensitivity

By computing averaged attention values over time and aligning them with beat/downbeat positions obtained from a pretrained beat tracker [45], we find a statistically significant correlation (Pearson $r = 0.63$, $p < 0.01$) between attention peaks and downbeat locations. This indicates that StructMamba's frequency-aggregated attention reflects high-level rhythmic structure, without explicit beat supervision.

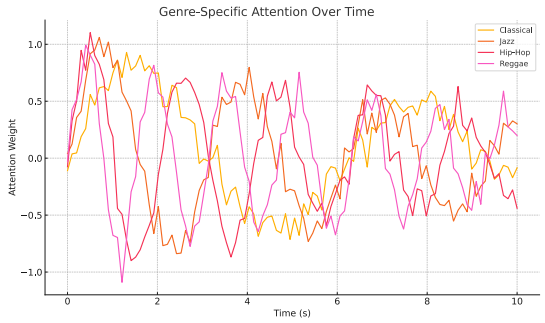

FIGURE 8. Genre-specific attention profiles showing differing model focus: harmonic continuity in classical/jazz vs. rhythmic regularity in hip-hop/reggae.

3) Segment Repetition Detection

We visualize attention across longer excerpts containing full verse–chorus–bridge form. As shown in Figure 6, global attention peaks reoccur across matching segments, even when instrumentation or harmony varies. These observations align with the hypothesis that StructMamba encodes structural recurrence through its 2D global attention layer. These analyses demonstrate that StructMamba captures not only local dependencies but also musically meaningful global structures-validating its architectural design and practical interpretability.

E. LIMITATIONS AND FUTURE WORK

While StructMamba demonstrates strong empirical performance and interpretability across MIR tasks, several limitations remain. First, the model's reliance on mel-spectrogram inputs may constrain its adaptability to raw waveform processing or symbolic domains without further modification. Second, although the dual-axis Mamba formulation introduces valuable inductive biases, it may increase architectural complexity, requiring careful tuning of fusion strategies and training stability, especially in low-data regimes.

Additionally, StructMamba does not currently fuse symbolic and audio modalities. Finally, real-time deployment on embedded hardware remains an open challenge, despite the model's moderate parameter count, due to the computational overhead of global 2D attention.

Future work will explore more efficient approximations of attention for low-latency inference, adaptive fusion mechanisms between harmonic and rhythmic streams, and multimodal extensions that integrate symbolic representations with spectrogram-based features. We also envision applying StructMamba to generative and interactive musical systems, where its structured representations may offer novel capabilities for real-time creativity and analysis.

VII. CONCLUSION

We introduced **StructMamba**, a novel architecture for music modeling that combines dual-axis state space modeling with global two-dimensional self-attention. Designed to align with the inductive biases of music, StructMamba separately models harmonic content along the frequency axis and rhythmic structure along the time axis using parallel Mamba modules, then fuses these representations via attention to capture global musical patterns such as motifs and structural repetition.

Through extensive evaluations across three benchmark tasks-genre classification, onset detection, and structural segmentation-StructMamba consistently outperforms CNN, transformer, and hybrid SSM-based models. Notably, it demonstrates strong performance in low-data regimes and exhibits superior generalization across datasets with differing musical styles and distributions.

Beyond accuracy, StructMamba offers interpretable alignment with motifs, downbeats, and structural boundaries, bridging computational modeling and music theory. Its efficiency and transparency make it suitable for real-time applications such as music education, assisted annotation, and interactive production. Future extensions will target symbolic–audio fusion and efficient attention approximations for embedded deployment, positioning StructMamba as a foundation for next-generation structure-aware music modeling.

REFERENCES

[1] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends® in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, Sep. 2014.
[2] R. Typke, F. Wiering, and R. C. Veltkamp, "A SURVEY OF MUSIC INFORMATION RETRIEVAL SYSTEMS."
[3] "AAM: A dataset of Artificial Audio Multitracks for diverse music information retrieval tasks | EURASIP Journal on Audio, Speech, and Music Processing."
[4] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal Music Information Processing and Retrieval: Survey and Future Challenges," in *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, Jan. 2019, pp. 10–18.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. ukasz Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
[6] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," May 2024.
[7] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, Jan. 2018.
[8] J. Schlüter and S. Böck, "Improved musical onset detection with Convolutional Neural Networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6979–6983.
[9] B. Yu, P. Lu, R. Wang, W. Hu, X. Tan, W. Ye, S. Zhang, T. Qin, and T.-Y. Liu, "Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1376–1388, Dec. 2022.
[10] "Exploring Jukebox: A Novel Audio Representation for Music Genre Identification in MIR | IEEE Conference Publication | IEEE Xplore."
[11] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation," Jul. 2017.
[12] Y. Yi, W. Li, M. Kuo, and Q. Bai, "PerceiverS: A Multi-Scale Perceiver with Effective Segmentation for Long-Term Expressive Symbolic Music Generation," Dec. 2024.
[13] J. Pons and X. Serra, "Musicnn: Pre-trained convolutional neural networks for music audio tagging," Sep. 2019.
[14] J. Chen, X. Tang, T. Xie, J. Wang, W. Dong, and B. Shi, "MusicMamba: A Dual-Feature Modeling Approach for Generating Chinese Traditional Music with Modal Precision," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[15] "Musical genre classification of audio signals | IEEE Journals & Magazine | IEEE Xplore."
[16] "Neural Network Music Genre Classification | IEEE Journals & Magazine | IEEE Xplore."
[17] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller, "Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2164–2168.
[18] J. Schlüter and S. Böck, "Improved musical onset detection with Convolutional Neural Networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6979–6983.
[19] "Frontiers | Neural Networks for Beat Perception in Musical Rhythm."
[20] "A fully convolutional deep auditory model for musical chord recognition | IEEE Conference Publication | IEEE Xplore."
[21] "Rethinking Automatic Chord Recognition with Convolutional Neural Networks | IEEE Conference Publication | IEEE Xplore."
[22] "(PDF) Comparative Analysis of Three Improved Deep Learning Architectures for Music Genre Classification," *ResearchGate*, Oct. 2024.
[23] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, Jun. 2021.
[24] "Global Self-Attention as a Replacement for Graph Convolution | Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining."
[25] "Pop Music Transformer | Proceedings of the 28th ACM International Conference on Multimedia."
[26] "PopMAG | Proceedings of the 28th ACM International Conference on Multimedia."
[27] L. Carvalho and G. Widmer, "Towards Robust and Truly Large-Scale Audio-Sheet Music Retrieval," Sep. 2023.
[28] "A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges | ACM Computing Surveys."
[29] C. Lu, Y. Schroecker, A. Gu, E. Parisotto, J. Foerster, S. Singh, and F. Behbahani, "Structured State Space Models for In-Context Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 016–47 031, Dec. 2023.
[30] J. Chen, X. Tang, T. Xie, J. Wang, W. Dong, and B. Shi, "Musicmamba: A dual-feature modeling approach for generating chinese traditional music with modal precision," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
[31] K.-H. Choi and J.-E. Ha, "Semantic Segmentation with Perceiver IO," in *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*, Nov. 2022, pp. 1607–1610.
[32] M. S. Junayed and M. B. Islam, "Consistent Video Inpainting Using Axial Attention-Based Style Transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 7494–7504, 2023.
[33] S. Hou, M. Fu, R. Wang, Y. Yang, and W. Song, "Self-Supervised Monocular Depth Estimation for All-Day Images Based on Dual-Axis Transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9939–9953, Oct. 2024.
[34] "Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications | SpringerLink."
[35] "Assessment of Mel-Filter Bank Features on Sound Classifications Using Deep Convolutional Neural Network | IEEE Conference Publication | IEEE Xplore."
[36] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski, "Deep State Space Models for Time Series Forecasting," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.
[37] "A General Survey on Attention Mechanisms in Deep Learning | IEEE Journals & Magazine | IEEE Xplore."
[38] D. Soydaner, "Attention mechanism in neural networks: Where it comes and where it goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13 371–13 385, Aug. 2022.
[39] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, Jan. 2018.
[40] Y.-S. Huang and Y.-H. Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 1180–1188.
[41] Y. Yi, W. Li, M. Kuo, and Q. Bai, "PerceiverS: A Multi-Scale Perceiver with Effective Segmentation for Long-Term Expressive Symbolic Music Generation," Dec. 2024.
[42] J. Chen, X. Tang, T. Xie, J. Wang, W. Dong, and B. Shi, "MusicMamba: A Dual-Feature Modeling Approach for Generating Chinese Traditional Music with Modal Precision," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.
[43] "An analysis of the GTZAN music genre dataset | Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies."
[44] X. Jia, "Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 5181899, 2022.
[45] "Zero-Note Samba: Self-Supervised Beat Tracking | IEEE Journals & Magazine | IEEE Xplore."

AMIT KUMAR BAIRWA is a senior academic in the School of Computer Science and Engineering at Manipal University Jaipur. He holds a B.E., M.Tech., and Ph.D. in Computer Science and Engineering, which reflect his deep academic foundation and long-standing commitment to the field. His primary areas of interest include Optimization, Mobile Ad Hoc Networks (MANET), Network Security, and Machine Learning. Dr. Bairwa has authored four books and continues to pursue active research in his areas of specialization. He is associated with professional bodies such as ISTE, ACM, and IEEE, which highlight his involvement in the wider academic and research communities. In addition to his teaching and research work, Dr. Bairwa is highly skilled in managing network systems and Linux-based platforms. He holds multiple international certifications, including CCNA, RHCE, and CEH, which demonstrate his ability to bridge theoretical knowledge with practical skills. He has successfully designed four funded project proposals and secured eight copyrights for his innovative contributions. Over the years, Dr. Bairwa has contributed significantly to curriculum development, student engagement, and the overall academic environment. He has led initiatives to modernize teaching methods, designed specialized training programs aligned with industry demands, and consistently worked toward improving student learning outcomes. His efforts in teaching and research have been recognized through several awards and honors. He has also played an instrumental role in organizing major academic events, including the International Conference on Cyber Warfare, Security & Space Computing (SpacSec) and the International Conference on Computation of Artificial Intelligence & Machine Learning (ICCAIML).

SIDDHANTH BHAT is currently pursuing a Bachelor's degree in Computer Science with a specialization in Artificial Intelligence and Machine Learning at Manipal University Jaipur. His academic interests are rooted in Deep Learning, Natural Language Processing (NLP), and Data Science, where he has gained hands-on experience through both coursework and independent research. He has contributed to multiple research publications, with papers presented in reputed IEEE conferences and journals. His work explores areas such as image recognition, natural language understanding, generative AI, and advanced data analysis techniques. In addition to his academic achievements, Siddhanth holds the position of Research Lead and is a Co-Founder at Xneuronz, a research-driven initiative. He has also interned at MuSigma, where he developed practical skills in data analytics and problem-solving. Siddhanth has been actively involved in university life, having held leadership positions such as President of Panacea – the Computer and Communications Society – and General Secretary of the Artificial Intelligence Department Community. He has also served on the organizing committees of notable conferences, including the International Conference on Computation of Artificial Intelligence & Machine Learning (ICCAIML) and the International Conference on Innovations in Computational Intelligence and Computer Vision (ICICV). To further enhance his technical capabilities, Siddhanth has obtained certifications in TensorFlow, Amazon Web Services (AWS), and Oracle Database Foundations. He remains committed to applying artificial intelligence to real-world challenges and continues to explore innovative solutions through research and collaboration.

TANISHK SAWANT is currently pursuing a Bachelor of Technology in Information Technology at Manipal University Jaipur. He has developed a strong academic interest in areas such as Natural Language Processing (NLP) and Neural Networks, with a particular focus on their practical applications. As part of his professional training, Tanishk is currently interning at Fugumobile in Mumbai, where he is engaged in projects involving NLP, Speech Synthesis, Voice Cloning, and Computer Vision. This opportunity has allowed him to apply his theoretical knowledge in real-world contexts, while also gaining valuable hands-on experience in innovative technologies. Prior to this, he interned at MuSigma, where he was involved in data analysis and gained exposure to industry-oriented problem-solving. Through his academic coursework and internship experiences, Tanishk has built a solid foundation in data processing, machine learning, and software development. His consistent efforts reflect a strong commitment to learning and growth in the field of artificial intelligence. Driven by a curiosity for emerging technologies and a desire to make meaningful contributions, he continues to explore the intersection of AI and real-world problem solving. Tanishk aspires to contribute to the development of intelligent systems that address contemporary challenges and improve everyday life.

MANOJ KUMAR BOHRA is a Professor in the Department of Computer Science and Engineering at Manipal University Jaipur, India. He completed his Bachelor's degree in Computer Science and engineering from MBM Engineering College, Jodhpur, in 2003. He went on to earn his Master's and Doctoral degrees from the Malaviya National Institute of Technology (MNIT), Jaipur, in 2011 and 2017, respectively. With over two decades of academic and research experience, Dr. Bohra has been deeply engaged in teaching, curriculum design, mentoring, and applied research. His primary areas of interest include Machine Learning, Artificial Intelligence, and their application to real-world problems. He has authored and presented numerous research papers in reputed national and international journals and conferences. Throughout his career, Dr. Bohra has contributed to the advancement of technology and education through active involvement in professional bodies, collaborative initiatives, and academic leadership. His commitment to excellence in research and pedagogy has earned him recognition within the academic community. He continues to inspire students and peers alike through his work in emerging technologies and his efforts to bridge theoretical knowledge with practical innovation.