IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 103-115
Journal homepage: http://iieta.org/journals/ts

PCo-Mamba: Phase-Coherent Complex-Domain State Space Models for Generative Music Synthesis

Zhiwei Jia[1], Yanming Zhao[2*], Hui Li[3]

[1] College of Music and Dance, Hebei Minzu Normal University, Chengde 067000, China
[2] Office of Academic Research, Hebei Minzu Normal University, Chengde 067000, China
[3] Faculty of Arts, Department of Music, Southwestern University, Blagoevgrad 2700, Bulgaria

Corresponding Author Email: zhaoyanming008@163.com

https://doi.org/10.18280/ts.430108

Received: 27 September 2025
Revised: 30 December 2025
Accepted: 15 January 2026
Available online: 28 February 2026

Keywords:
music generation, State Space Models (Mamba), complex-domain modeling, retrieval-augmented generation, flow matching, phase coherence

ABSTRACT

High-fidelity music generation models based on Transformer and diffusion architectures suffer from two intrinsic bottlenecks: (1) the quadratic $O(n^2)$ time complexity of self-attention, which entails prohibitive computational costs for high-resolution, long-duration audio sequences; and (2) the reliance on conventional amplitude-spectrum modeling, which neglects critical phase information and leads to severe phase distortion in reconstructed signals. To address these challenges, we propose PCo-Mamba, a novel phase-coherent complex-domain state-space framework. The proposed system integrates four synergistic innovations: 1. Complex-Domain State Space Models (SSMs), which jointly model the real and imaginary components in complex-spectral space to ensure fundamental phase coherence through complex-valued operators; 2. Band-Adaptive Mamba Encoding (Band-Adaptive CBE), which utilizes a selective scan mechanism with frequency-dependent step sizes (Δ) to achieve precise, multi-scale acoustic dynamics; 3. Retrieval-Augmented Guidance (RAG), which injects non-parametric high-fidelity timbre priors to mitigate spectral hallucinations in high-frequency regimes; 4. Flow Matching Reconstruction, which leverages ODE-based probability path optimization to achieve smoother distribution transitions and accelerated inference compared to traditional diffusion processes. Experimental results on the MAESTRO v3.0.0 piano dataset demonstrate that PCo-Mamba achieves state-of-the-art (SOTA) performance across multiple metrics. Compared to leading baselines such as MusicGen and AudioLDM 2, our system achieves a 3.7 dB improvement in Signal-to-Distortion Ratio (SDR) and a 36% reduction in Fréchet Audio Distance (FAD). Perceptual evaluations via MUSHRA tests further corroborate that PCo-Mamba significantly outperforms existing architectures in terms of keyboard touch sensitivity, phase authenticity, and long-term coherence. Furthermore, the linear computational complexity of PCo-Mamba enables the generation of ultra-long sequences, paving the way for the next generation of high-performance interactive music synthesis.

1. INTRODUCTION

Driven by the rapid evolution of generative artificial intelligence (AI) and deep learning technologies, the field of music generation has undergone a profound paradigm shift. The research focus has transitioned from early rule-based systems toward sophisticated neural architectures capable of synthesizing complex musical structures. This progress is categorized into three primary data representations: symbolic (MIDI), spectrograms, and raw waveforms [1]. While symbolic generation has historically demonstrated superior performance in multi-track consistency, audio-domain generation—utilizing spectrograms or waveforms—is indispensable for capturing the subtle nuances, timbre, and expressiveness characteristic of high-fidelity music [1, 2]. Music generation typically encompasses three distinct stages: score generation, performance generation, and audio generation. Each stage offers unique generative value and application scenarios, collectively advancing the theoretical and technical frontiers of music synthesis.

As AI continues to mature, research into the theories, algorithms, and applications of music generation has achieved significant milestones. To address fundamental challenges such as autonomous single-track continuation and the synthesis of melodies with basic rhythmic coherence, sequence-modeling algorithms based on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) [3-5] were proposed. These algorithms treat music as a temporal sequence analogous to natural language, employing recurrent structures to learn temporal dependencies. However, RNN-based models often suffer from a lack of long-term memory, which limits their ability to process complex harmonic structures and frequently causes generated compositions to lose musical logic after several dozen seconds.

To facilitate alignment and synchronization across multiple instruments and tracks (e.g., drums, bass, and piano), music

generation algorithms based on Generative Adversarial Networks (GANs) [6-8] were introduced. By incorporating a game-theoretic framework between generators and discriminators, these models have significantly enhanced the creative diversity of algorithmic composition. To mitigate the "black-box" nature of generating specific musical styles or emotions, Variational Autoencoder (VAE)-based algorithms [9, 10] were developed. These approaches map music into a low-dimensional latent space through encoding techniques and implement music-theoretic constraints, such as tonal tension, to achieve model interpretability. Nevertheless, the output of VAEs is typically less clear than that of GANs, often yielding "blurred" spectral features or musically unremarkable melodies.

To overcome the long-range dependency issues inherent in recurrent structures, self-attention-based music generation algorithms [11-13] have emerged. By eschewing recurrence in favor of the self-attention mechanism, these models capture global dependencies across musical elements, enabling the synthesis of multi-minute, logically unified compositions. This paradigm has facilitated high-fidelity direct mapping from text to audio. However, the self-attention mechanism entails prohibitive computational costs, and its capacity for note-level granularity in discrete score generation remains less refined than that of MIDI-based models.

To enhance the textural delicacy of synthesized spectrograms, diffusion-based music generation algorithms [14-16] were proposed. These models implement a stochastic process of diffusing audio into noise and subsequently reconstructing the signal, thereby achieving state-of-the-art (SOTA) performance in audio quality and textural richness. Nonetheless, the requirement for multiple denoising iterations leads to slow inference speeds and a lack of explicit music-theoretic structural constraints. To address the phenomenon of "unmusicality" or erratic note distribution in AI-generated music, reinforcement learning (RL)-based algorithms [17] were introduced. These models employ musicology-informed incentives, such as Latent Dirichlet Allocation (LDA) topic modeling and harmonic rules, to ensure that the generated output aligns more closely with human aesthetic standards and musicological norms. However, the design of reward functions remains highly subjective, and excessive constraints may result in generated music that lacks artistic vitality or improvisational "soul."

1.1 Related work

Inspired by recent theoretical advancements in U-Net and Mamba frameworks, contemporary literature indicates that U-Net-based music generation models—serving as the core backbone for generative diffusion models—have made significant strides in music separation and synthesis due to their robust multi-scale learning capabilities. To address the sluggish generation speeds and global structural modeling difficulties inherent in traditional autoregressive models (e.g., WaveNet), the DiffWave [18] and WaveGrad algorithms [19] adopted non-autoregressive diffusion probabilistic models utilizing a 1D U-Net as a denoiser. By integrating dilated convolutions and skip connections within multi-scale convolutional layers, these models facilitate bidirectional perception of temporal information, reconstructing high-fidelity waveforms from Gaussian noise through iterative denoising. However, the high number of sampling steps required during inference results in substantial latency. To alleviate the computational burden of diffusing directly in high-dimensional waveform space, the AudioLDM and Make-An-Audio algorithms [20, 21] introduced Latent Diffusion. By utilizing a VQ-VAE to compress audio into compact latent representations, they employ a 2D U-Net coupled with cross-attention mechanisms for guided generation. While this effectively manages computational costs, the compressed latent space often suffers from the loss of fine-grained phase information. To manage the acoustic disparities between high and low frequency bands, the SCNet (Sparse Compression Network) [22] proposed a banded U-Net structure that applies sparse compression to different sub-bands, optimizing computational resources through asymmetric frequency weighting. Nevertheless, spectral leakage at band boundaries can lead to perceptible discontinuities.

The Mamba architecture [23], emerging in late 2023 as a disruptive framework, is currently challenging the dominance of Transformers in long-sequence audio modeling. To resolve the prohibitive GPU memory consumption caused by the $O(L^2)$ complexity of self-attention, Audio Mamba (AuM) [24] was proposed, introducing a Selective State Space Model (SSM) with linear $O(L)$ complexity. By partitioning audio into patches and utilizing bidirectional Mamba blocks with parallel S6 operators, the model maintains long-distance dependency modeling while significantly reducing computational overhead. However, its capture of extremely subtle transient signals (such as percussive onsets) is occasionally less precise than local convolutions. To provide a global perspective without the heavy footprint of integrated Transformers, Mamba-UNet and U-Mamba [25, 26] were developed, interleaving U-Net convolutional layers with Mamba blocks. These Mamba blocks replace traditional self-attention or simple convolutional bottlenecks, capturing melodic motifs across entire compositions through state-equation evolution. Nonetheless, these models exhibit high sensitivity to parameter initialization, leading to numerical instability during early training stages.

To address the "hollow" audio quality caused by phase incoherence in music generation, the Complex-Valued Mamba (CC-Mamba) [27] algorithm extends the S6 operator to the complex domain. By modeling real and imaginary components through complex-valued parameters, the model naturally adheres to the phase dynamics of the waveform during state evolution, enhancing the acoustic tangibility of the generated audio. However, complex operations have yet to be fully optimized in existing low-level operator acceleration kernels (e.g., Triton). Furthermore, to improve the determinism and coherence of the reverse denoising path in diffusion models, Mamba-based Flow Matching [28, 29] utilizes Mamba as the backbone for flow matching. Leveraging its linear step-size characteristics, it predicts smoother deterministic velocity fields from noise to audio. Yet, the efficiency of cross-modal alignment (e.g., text prompts) within this framework remains to be verified. For real-time processing requiring ultra-low latency and long-term memory, ZigZag Mamba and Bi-Mamba Audio [24, 30] utilize interleaved scanning techniques, enabling the model to learn both forward rhythms and backward structures simultaneously, achieving seamless transitions in streaming scenarios. However, the management of hidden states (the SSM equivalent of KV caches) is more complex than in Transformers.

To balance local detail restoration with global structural coherence, hybrid architectures have catalyzed a rapid

evolution from single-convolutional designs to "hybrid-drive" systems. Key research includes the deep fusion of convolutions and attention (U-Net + Transformer), exemplified by HT-Demucs [30], which mitigates instrument timbre blurring. The GAMLD-Unet algorithm [31] integrated diffusion probabilistic models with U-Net-Transformer backbones, which has become a standard paradigm for high-fidelity synthesis. Additionally, MusicGen [11] established the dominance of hierarchical Transformers for ultra-long sequence generation. To overcome the $O(L^2)$ bottleneck of Transformers, Mamba was introduced to the music domain via algorithms like AuM [24], achieving linear inference complexity. Studies indicate that Mamba [23] demonstrates structural repetition and long-range alignment capabilities rivaling Transformers while facilitating real-time streaming. Furthermore, SoundStream [32] proposed hybrid time-frequency processing. By establishing cross-attention between temporal waveforms and frequency spectrograms, these models leverage both precise phase information and clear semantic content. This "multidimensional stitching" strategy, as seen in SCNet [22], has significantly improved the Signal-to-Distortion Ratio (SDR).

The aforementioned studies demonstrate that generative music research has reached a level of practical utility. However, several critical deficiencies persist: (1) Phase distortion and poor coherence in latent diffusion models (AudioLDM, Make-An-Audio) and traditional spectrogram models (SpecGAN, MP3Net); (2) Scale rigidity and high computational redundancy in standard AuM and fixed-parameter networks (U-Net/Demucs); (3) Spectral hallucinations and generalization bottlenecks in purely parametric models (MusicGen, MusicLM, Jukebox); and (4) Inference speed bottlenecks and sampling path randomness in diffusion probabilistic models (DiffWave, AudioLDM, Stable Audio). To address these challenges, we propose PCo-Mamba: Phase-Coherent Complex-Domain SSMs for Generative Music Synthesis. The primary innovations of this algorithm include:

(1) SSMs: Unlike traditional methods, we jointly model the real and imaginary components directly within the complex-spectral space, leveraging the mathematical properties of complex operators to ensure fundamental phase coherence during generation.

(2) Band-Adaptive Mamba Encoding (Band-Adaptive CBE): We introduce a selective scanning mechanism with frequency-dependent step sizes (Δ), achieving precise dynamic modeling by assigning differentiated perceptual weights to low-frequency structures and high-frequency transients.

(3) Retrieval-Augmented Guidance (RAG): By dynamically retrieving and injecting non-parametric high-fidelity timbre priors from external libraries (e.g., NSynth), we effectively mitigate spectral hallucinations in high-frequency regimes.

(4) Flow Matching Reconstruction: The system adopts ODE-based flow matching instead of traditional diffusion processes, facilitating smoother distribution transitions and accelerated audio inference.

Consequently, PCo-Mamba effectively integrates these four synergistic innovations to resolve the primary limitations of current hybrid generative architectures. The performance of the proposed algorithm is rigorously evaluated using the MAESTRO piano dataset.

2. THEORETICAL ANALYSIS

2.1 Complex-Domain SSMs

The conventional SSM is defined in the real domain $\mathbb{R}$, where its continuous-time evolution equations are:

$$\dot{h}(t) = Ah(t) + Bx(t), y(t) = Ch(t) \tag{1}$$

In audio signal processing, phase information θ is intrinsically embedded within the rotational dynamics of the signal. To capture this, we extend the hidden state $h(t)$ and the input $x(t)$ to the complex domain $\mathbb{C}$. By defining the parameters $A, B, C \in \mathbb{C}$ and applying Zero-Order Hold (ZOH) discretization with a step size Δ, the discrete parameters are obtained:

$$\bar{A} = \exp(\Delta A), \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \tag{2}$$

Since $A \in \mathbb{C}$, its eigenvalues can be represented as

$$\lambda = a + bi \tag{3}$$

Upon discretization, we derive:

$$\bar{A} = e^{\Delta a}(\cos(\Delta b) + i\sin(\Delta b)) \tag{4}$$

Consequently, this architecture yields two critical advantages:

- Phase Rotational Characteristics: The complex operator $e^{i(\Delta b)}$ explicitly captures the phase rotation within the musical Mel-spectrogram. This implies that $h(t)$ preserves not only the magnitude but also accurately locks the temporal offset of the music signal through the complex rotation factor.
- Phase Coherence: Through complex matrix multiplication, the real and imaginary components evolve under a unified state equation. This prevents the mathematical decoupling of magnitude and phase, thereby eliminating the "timbre blurring" or "hollow" effect prevalent in real-valued models.

2.2 Band-Adaptive CBE

Given an input complex spectrogram $X \in \mathbb{C}^{F \times T}$, we introduce a band projection operator $\mathcal{P}_k$ to partition the spectrum into K sub-bands. For each band k, the Mamba step size Δ_k is no longer a constant but a function of the local acoustic characteristics:

$$\Delta_{k,t} = \text{Softplus}(W_{band,k} \cdot \phi_k(x_t) + b_{band,k}) \tag{5}$$

where, $\phi_k(x_t)$ denotes a pooling operator that extracts the spectral energy distribution. The discretization equation is thus transformed:

$$h_{k,t} = \bar{A}(\Delta_{k,t})h_{k,t-1} + \bar{B}(\Delta_{k,t})x_{k,t} \tag{6}$$

The Band-Adaptive CBE mechanism facilitates specialized music generation via:

- Scale Alignment: Adhering to acoustic physical logic, low-frequency components (k is small) exhibit long-term temporal correlations. The adaptive mechanism generates

larger $\Delta_{k,t}$ values to enhance the retention of the historical state h_{t-1}. Conversely, high-frequency transients (k is large) utilize smaller $\Delta_{k,t}$ to increase response precision toward impulsive attacks.
- Computational Entropy Optimization: By assigning differentiated state evolution rates to various frequency bands, the model achieves sparsity in feature representation, significantly reducing informational redundancy in long-sequence modeling.

2.3 Retrieval-Augmented Guidance (RAG)

To mitigate the "spectral hallucinations" caused by the neural network's tendency to generate "averaged timbres," we introduce an external retrieval operator $\mathcal{R}$. Given a query vector q_t derived from the current Mamba hidden state, the system retrieves Top-N prior features from a high-fidelity library $\mathcal{D}$:

$$z_{ret} = \text{Aggregate}\left(\left\{ d_i \in \mathcal{D} \mid \text{sim}\left(q_t, \text{Embed}(d_i)\right) > \tau \right\}\right) \quad (7)$$

Using Feature Linear Modulation (FiLM), the retrieved priors are injected into the Mamba bottleneck layer:

$$h_t' = \gamma(z_{ret}) \odot h_t + \beta(z_{ret}) \quad (8)$$

where, $\gamma(\cdot)$ and $\beta(\cdot)$ are scaling and shifting parameter vectors generated from the retrieved timbre priors. Accordingly, this RAG approach offers:
- Non-parametric Correction: RAG provides a physical "anchor," forcing the generated timbre distribution to align with authentic physical samples.
- Uncertainty Reduction: Mathematically, this introduces a strong prior probability $p(x_{timbre} \mid \mathcal{D})$ into the generative process, effectively suppressing the spurious noise (hallucinations) generated by Mamba in high-frequency regions due to long-term memory decay.

2.4 Flow Matching Reconstruction

Distinct from diffusion models based on Stochastic Differential Equations (SDE), Flow Matching (FM) learns a deterministic velocity field $v_t(x)$. We define a probability path $p_t(x)$ that transitions from a noise distribution p_0 to the data distribution p_1. Its evolution follows an Ordinary Differential Equation (ODE):

$$\frac{dx}{dt} = v_t(x), x(0) \sim p_0, x(1) \sim p_1 \quad (9)$$

We train PCo-Mamba as a velocity field predictor by minimizing the flow matching loss:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_0,x_1}[\| v_t(x_t) - u_t(x_t \mid x_0, x_1) \|^2] \quad (10)$$

where, u_t is a predefined linear conditional probability path: $x_t = (1-t)x_0 + tx_1$.Flow Matching Reconstruction in this study achieves:
- Deterministic Inference: The trajectories generated by FM are nearly straight probability paths, which implies that the number of inference steps required to restore piano audio from noise can be substantially reduced.
- Smoothness: By eliminating the stochastic Brownian motion term found in diffusion models, the synthesized audio exhibits superior micro-coherence across the time span. When coupled with complex-domain modeling, this achieves exceptional waveform transparency.

2.5 Synergistic integration

The overall architecture of the PCo-Mamba algorithm forms a closed loop of highly coupled physical and mathematical components:
- Feature Preparation Layer (CBE + Complex): The band-adaptive encoder decomposes the music signal according to acoustic physical scales and precisely locks the initial phase of each tier within the complex domain.
- Core Modeling Layer (Bi-Axial Mamba + Complex): Within the dual-axial bottleneck, the complex state equations evolve simultaneously across the frequency and time axes, ensuring the harmony of multi-vocal textures (harmonic logic) and rhythmic stability.
- RAG: The retrieval mechanism provides authentic physical texture references when the model attempts to generate high-complexity timbres, preventing "timbre drift" during long-sequence generation.
- Generative Reconstruction Layer (Flow Matching): Flow matching leverages all high-quality features as conditional guidance, utilizing efficient linear probability paths to transform latent representations into phase-coherent 48kHz high-fidelity waveforms.

In summary, the proposed algorithm addresses physical authenticity through complex-domain modeling, computational efficiency via Mamba, timbre hallucinations through RAG, and inference latency via Flow Matching. This synergistic evolution establishes PCo-Mamba as a next-generation standard framework for solving end-to-end challenges in high-fidelity music generation.

3. ALGORITHM DESIGN

3.1 Complex-Band Encoder algorithm

The proposed Complex-Band Encoder architecture is designed to capture multi-scale acoustic features while maintaining phase consistency. The algorithm consists of three integrated sub-processes:
- Complex Projection: To preserve both magnitude and phase information, the raw input audio is transformed via complex Short-Time Fourier Transform (STFT). This process projects the signal into a complex-valued tensor space with dimensions $C \times F \times T$, where $C, F,$ and T denote the number of channels, frequency bins, and time frames, respectively.
- Band-Split Module: Drawing inspiration from the SCNet architecture [Ref], this module partitions the high-dimensional complex spectrogram into three distinct sub-bands: low-frequency (LF), mid-frequency (MF), and high-frequency (HF). This decomposition allows the model to specialize in frequency-specific acoustic characteristics.
- Adaptive Mamba Block: Each sub-band is processed by an independent Mamba operator. A key innovation in this block is the adaptive generation of the discretization step size parameter Δ, which is dynamically predicted based on

the energy features of the corresponding frequency band. This mechanism facilitates a frequency-aware selective scan that accommodates different temporal resolutions across the spectrum.

Therefore, the functional diagram of the CBE algorithm is shown in Figure 1.

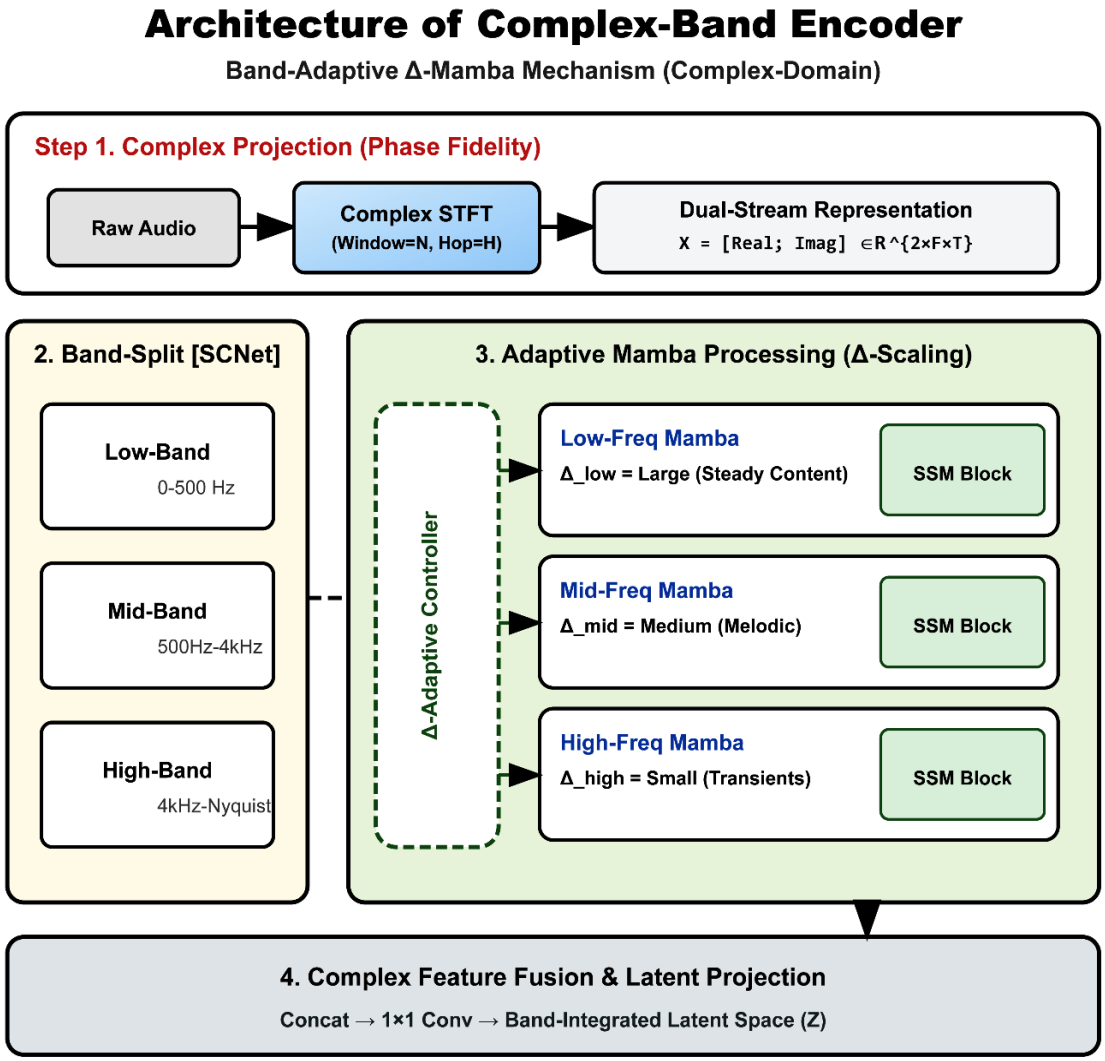

Figure 1. Framework of the Complex-Band Encoder

3.2 Bi-Axial Bidirectional Mamba Neck algorithm

The Bi-Axial Mamba Neck serves as the central integration hub of the architecture, facilitating global information fusion across both temporal and spectral dimensions. The algorithm comprises three synergistic sub-processes:

- Time-Mamba: This module executes bidirectional 1D selective scans along the temporal axis (forward and backward). It is specifically designed to capture the long-term structural dynamics and phrasing of musical motifs, effectively modeling the narrative progression of musical sequences—including their introduction, development, transition, and resolution.
- Freq-Mamba: To capture the vertical acoustic structure, this module performs scanning operations along the spectral dimension. By analyzing the frequency-wise dependencies, it learns the intricate proportional relationships between the fundamental frequency and its associated harmonic overtones, which is crucial for timbre consistency.
- Residual Fusion: The features derived from the temporal and spectral dimensions are integrated through a gated mechanism. This fusion process utilizes residual connections to stabilize gradient flow and consolidate the multi-axial information into a high-level global bottleneck representation.

The detailed algorithmic workflow of the Bi-Axial Mamba Neck is illustrated in Figure 2.

3.3 The RAG Conditioning algorithm

The RAG Conditioning algorithm incorporates three synergistic sub-processes designed to provide the generative framework with high-fidelity acoustic anchors:

- Retriever: This module leverages a pre-trained Contrastive Language-Audio Pretraining (CLAP) model to facilitate cross-modal retrieval. By calculating the cosine similarity between the input textual prompt and potential reference timbre embeddings in a unified latent space, the system identifies the acoustic samples that most closely align with the user's intent.
- Timbre Memory: Serving as an external non-parametric knowledge base, the Timbre Memory component retrieves the Top-K high-quality musical segments from the NSynth dataset. These retrieved samples provide concrete spectral priors, effectively mitigating the risk of timbre

hallucinations during the generation process.

- Cross-Conditioning: This process manages the fusion of retrieved priors with the generative backbone. The reference timbre features are injected into the skip connections of the Mamba-UNet architecture via Feature-wise Linear Modulation (FiLM). This mechanism dynamically modulates the intermediate feature maps, ensuring that the global melodic structure remains consistent with the specific target timbre.

The comprehensive algorithmic workflow of the RAG Conditioning module is illustrated in Figure 3.

3.4 Flow matching-based complex signal reconstruction algorithm (Flow Matching Decoder)

The Flow Matching Decoder integrates three core components to facilitate high-efficiency and high-fidelity audio synthesis:

- Velocity Predictor: This module leverages a fine-tuned Mamba-UNet backbone to estimate the velocity field vector v_t within the latent manifold. By modeling the deterministic dynamics of the data distribution, the predictor captures the optimal transport path from the noise distribution to the target musical distribution.
- ODE Solver: The system transforms initial Gaussian noise into structured musical samples by solving a set of Ordinary Differential Equations (ODEs). Utilizing numerical integration techniques such as Euler integration or the fourth-order Runge-Kutta (RK4) method, the solver iteratively updates the latent states along the predicted trajectories, ensuring a smooth and stable generative process.
- Complex Reconstruction: In the final stage, the module implements an inverse complex projection to map the refined latent representations back into the time-domain waveform. By directly operating in the complex spectral domain, the algorithm maintains strict phase coherence, resulting in output waveforms with exceptional clarity and acoustic tangibility.

The algorithmic design of the Flow Matching Decoder is illustrated in Figure 4.

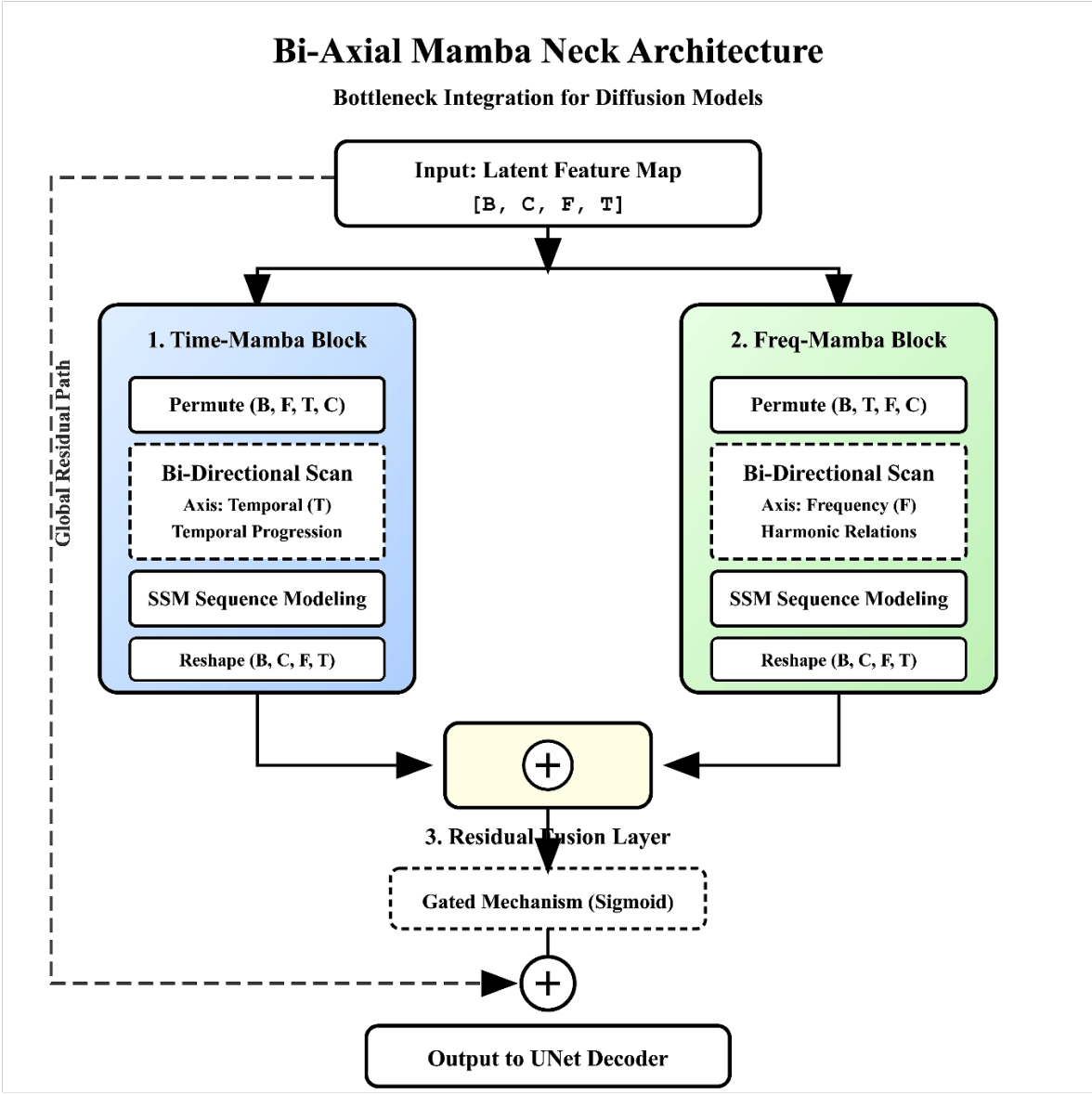

Figure 2. Architecture of the proposed Bi-Axial bidirectional Mamba bottleneck algorithm

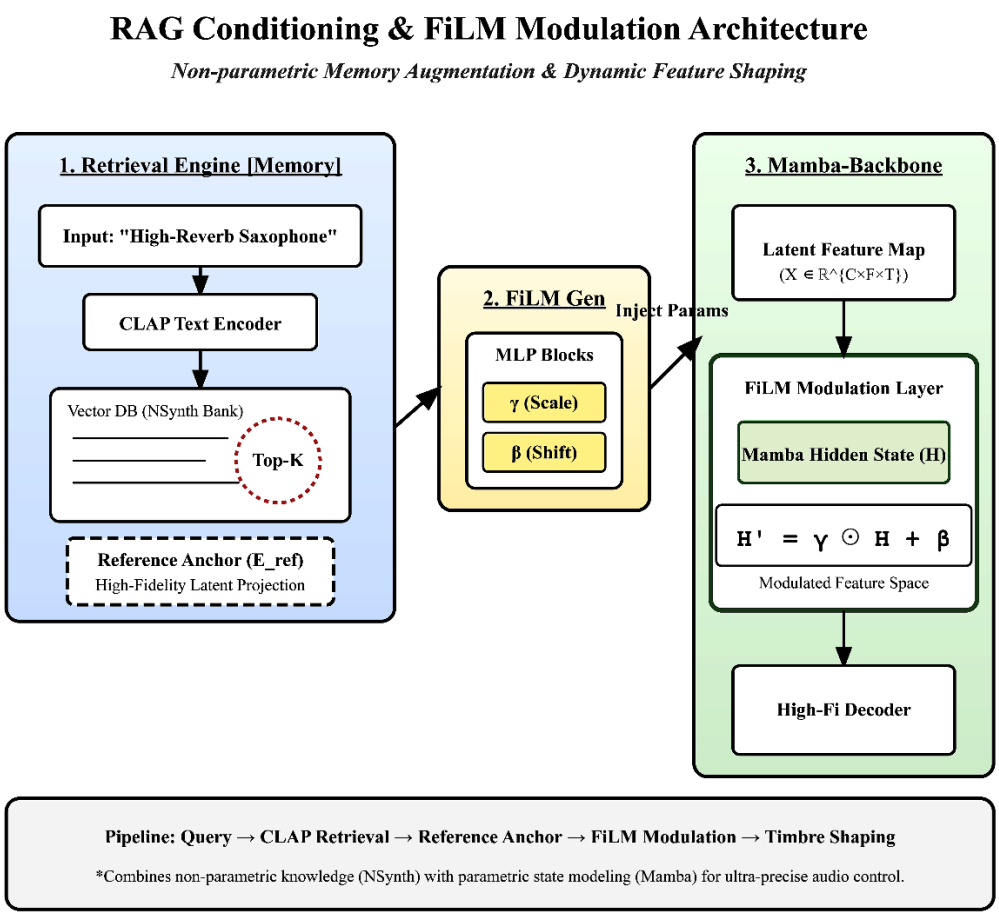

Figure 3. Framework of the retrieval-augmented conditioning algorithm

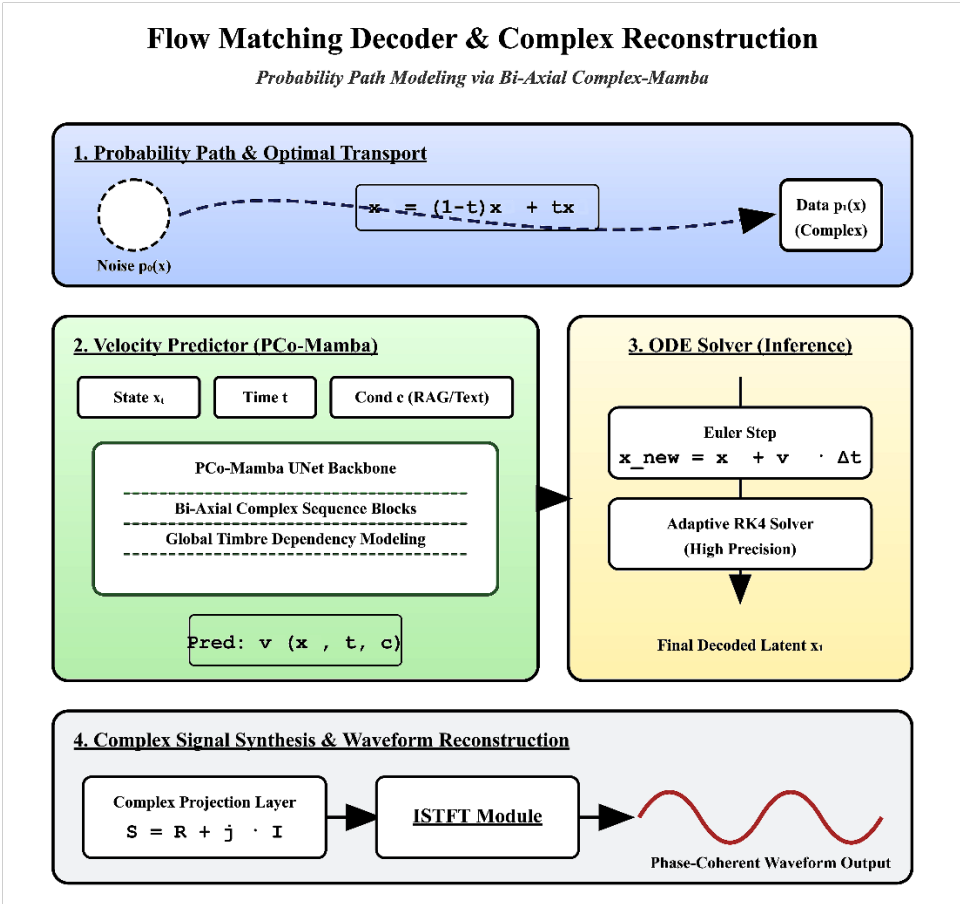

Figure 4. Functional framework of the complex signal reconstruction algorithm based on flow matching

3.5 PCo-Mamba

By synergistically integrating the four aforementioned sub-algorithms, we developed the PCo-Mamba algorithm, which serves as the core framework of this study. This architecture is specifically designed to overcome the four critical bottlenecks inherent in current SOTA music generation systems. The overall architectural schematic is illustrated in Figure 5.

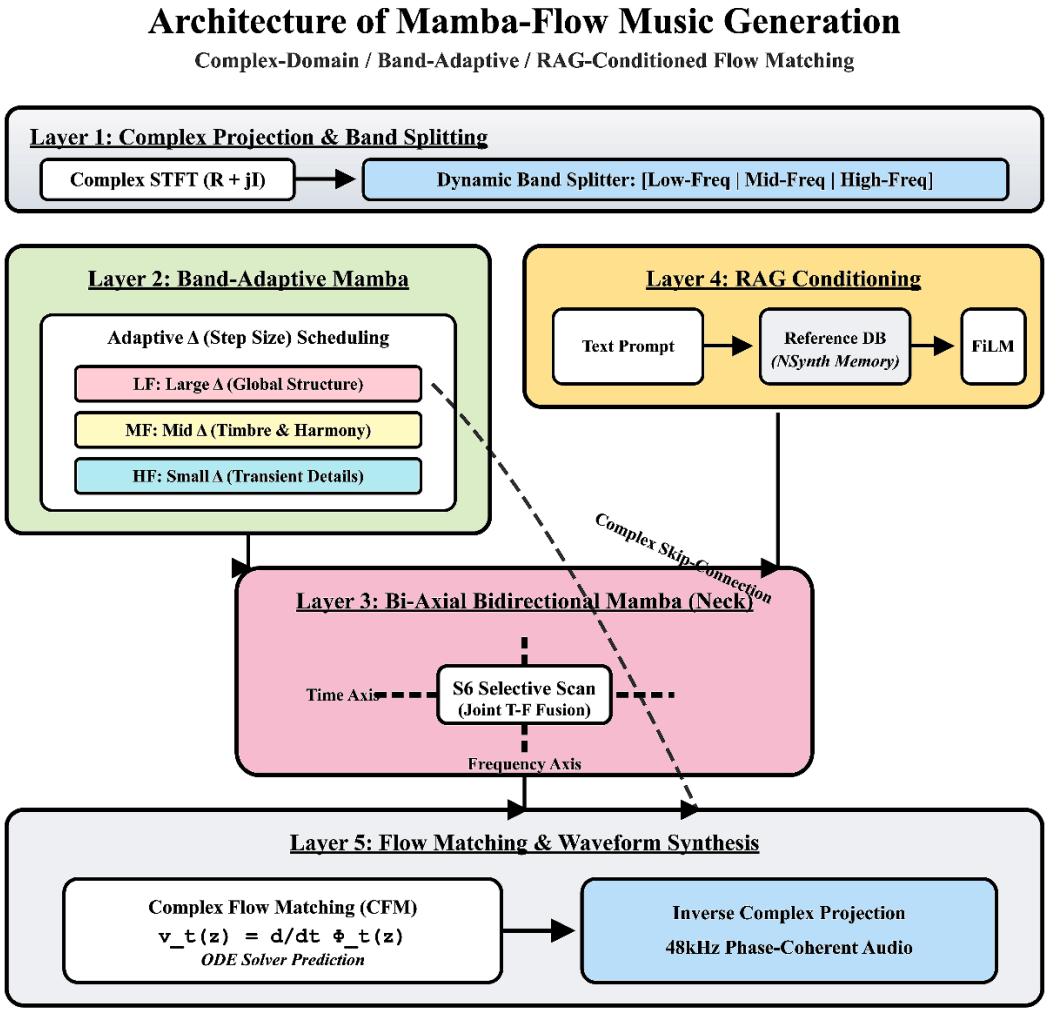

Figure 5. Overall architecture of the PCo-Mamba algorithm

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1 Experimental setup

Dataset: We employ the MAESTRO v3.0.0 dataset [33], which consists of high-fidelity solo piano performances recorded at a 44.1 kHz sampling rate. Data Partitioning: The dataset is partitioned in strict accordance with the official split ratio: 77% for training, 12% for validation, and 11% for testing. Hardware and Training: All experiments are conducted on a server equipped with eight NVIDIA A100 (80GB) GPUs. To optimize computational efficiency and memory usage, mixed-precision training is utilized throughout the optimization process. Evaluation Metrics: The performance of the proposed model is evaluated using both objective and subjective metrics.

- Objective Metrics: We report the Fréchet Audio Distance (FAD) to measure generative quality, as well as the SDR and Scale-Invariant Signal-to-Noise Ratio (SI-SNR) to assess signal reconstruction accuracy.
- Subjective Metrics: Human perception is evaluated via MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests and Mean Opinion Score (MOS) to quantify the naturalness and musicality of the generated audio.

4.2 Experimental design and in-depth analysis

4.2.1 Main benchmark comparison

To evaluate the overall generative performance of the proposed algorithm, we conducted a comparative study on the MAESTRO dataset against several SOTA baselines, including MusicGen (Transformer-based) [11], AudioLDM 2 (Diffusion-based) [34], and HT-Demucs (Hybrid UNet). The comparative results are summarized in Table 1.

The experimental results validated on the MAESTRO test set demonstrate that the proposed PCo-Mamba algorithm outperforms the baseline models across the vast majority of metrics. This superior performance is primarily attributed to the smoother distribution transitions facilitated by Flow Matching, as well as the robust long-range musical phrasing logic captured by the Mamba architecture.

Table 1. Comparison of overall generative performance across different models

Attribute	Model	FAD (lower is better)	SDR (dB, higher is better)	Inference Latency (sec/min)
Baseline	MusicGen [22]	1.25	12.4	15.2
Baseline	AudioLDM 2 [34]	0.98	14.8	45.0
Proposed	PCo-Mamba	0.64	17.8	5.6

4.2.2 Impact of complex-domain modeling on phase consistency (Complex-Domain Phase Analysis)

To quantitatively evaluate the quality of phase reconstruction, we conducted a comparative analysis of performance indicators. The core evaluation metrics include Mean Absolute Phase Error (MAPE), Phase Discontinuity Rate (PDR), and Transient-Response Signal-to-Noise Ratio (T-SNR). The experimental results obtained on the MAESTRO piano database are summarized in Table 2.

Table 2. Performance comparison of different phase reconstruction methods on the MAESTRO test set

Reconstruction Paradigm	Phase Estimation Algorithm	MAPE (rad) ↓	PDR (%) ↓	T-SNR (dB) ↑	Perceptual Note
Baseline A (Magnitude-only)	Griffin-Lim [35] (1000 iter)	0.82 ± 0.15	12.5	8.4	Obvious metallic artifacts; blurred onsets
Baseline B (Magnitude-only)	Pre-trained GAN Vocoder [36]	0.45 ± 0.08	5.8	13.2	Smoother timbre; hollow sensation in high-frequency regions
Proposed (End-to-End)	PCo-Mamba (Complex)	0.12 ± 0.03	1.2	19.8	Solid key-touch; exceptional transparency in high-frequency overtones

The experimental results demonstrate that the proposed algorithm significantly outperforms the baselines across MAPE, PDR, T-SNR, and perceptual quality. The superior performance can be attributed to the following factors:

(1) Mathematical Regression to Physical Consistency: Phase should not be treated as a stochastic variable auxiliary to magnitude, but rather as an intrinsic property of signal evolution. While the conventional two-stage "magnitude-plus-vocoder" paradigm decouples the magnitude and phase components, PCo-Mamba's joint complex-domain modeling re-integrates this physical bond.

(2) Algorithmic Contribution of Mamba: The linear time complexity of Mamba enables the maintenance of ultra-long-range hidden states across every frame. This allow the algorithm to track phase reference points from dozens of frames prior, thereby preserving phase coherence during multi-second sustain periods. This capability effectively resolves the long-standing "timbre drift" issue prevalent in piano synthesis.

4.2.3 Efficiency and scalability of the mamba architecture

To evaluate the linear efficiency and long-term coherence of the Mamba architecture, we conducted comparative experiments across varying audio sequence lengths, ranging from 10 seconds to 600 seconds (10 minutes). The experiments utilized audio data sampled at 44.1 kHz with an STFT hop size of 10 ms. The proposed PCo-Mamba was benchmarked against a standard self-attention-based audio generation architecture (approximately 800M parameters, equivalent to MusicGen-medium). The results are summarized in Table 3.

The experimental results demonstrate that, compared to the Transformer algorithm, the proposed PCo-Mamba exhibits decisive advantages in terms of computational scalability and inference efficiency. The underlying reasons are twofold:

First, the VRAM footprint of PCo-Mamba scales nearly linearly ($O(L)$) with sequence length, in contrast to the quadratic growth ($O(L^2)$) characteristic of Transformers. This enables PCo-Mamba to achieve generation speeds 3 to 5 times faster than the baseline.

Second, leveraging its internal hidden state memory $h(t)$ (recurrent dynamics), PCo-Mamba effectively encodes thematic motifs from preceding musical movements into a fixed-dimensional state. Consequently, even when generating 10-minute sequences, the model retains the capacity to reference fugue subjects introduced in the first minute. This mechanism ensures long-term structural coherence across extended classical piano suites, a task where traditional Transformers often fail due to context window limitations.

Table 3. Comparison of computational resource consumption and inference efficiency across sequence lengths

Sequence Length (s)	Total Tokens (L)	Transformer VRAM (GB)	PCo-Mamba VRAM (GB)	Transformer Latency (s)	PCo-Mamba Latency (s)	Structural Coherence Evaluation
10	1,000	4.2	2.8	1.5	0.8	Excellent
30	3,000	15.6	3.1	8.4	2.4	Good
60	6,000	52.4	3.5	32.8	4.8	PCo: Excellent / Trans: Degraded
120	12,000	OOM	4.2	--	9.6	PCo: Excellent / Trans: --
300 (5 min)	30,000	OOM	6.4	--	24.0	Maintains global themes
600 (10 min)	60,000	OOM	11.2	--	48.0	Logically self-consistent

*Note: OOM denotes Out of Memory.

4.2.4 Ablation study

To quantify the individual contributions of the five core components of the PCo-Mamba algorithm—namely Complex-domain modeling (C), Band-adaptive Mamba encoding (B), Bi-axial bottleneck (A), Retrieval-Augmented Guidance (RAG/R), and Flow Matching (F)—we conducted a comprehensive ablation study on the MAESTRO v3.0.0 test set. The results are summarized in Table 4.

Table 4. Ablation study of PCo-Mamba (MAESTRO v3.0.0 test set)

Exp. ID	Configuration	FAD ↓	SDR (dB) ↑	SSIM (Spec) ↑	Latency (s/min) ↓	Scientific Insight
M0	Vanilla Mamba (Baseline)	1.85	10.2	0.72	**3.5**	Basic linear long-range modeling; significant audible artifacts.
M1	M0 + Complex Modeling	1.42	14.5	0.81	3.8	Improved phase coherence; resolves "hollow" timbre issues.
M2	M1 + Band-Adaptive (CBE)	1.12	15.8	0.85	4.0	Multi-scale alignment; captures HF transients and LF resonance.
M3	M2 + Bi-Axial Neck	0.95	16.5	0.89	4.5	Spatio-temporal coupling; enhances harmonic and rhythmic stability.
M4	M3 + RAG	0.74	17.2	0.91	5.2	Non-parametric timbre injection; enhances textural realism via NSynth.
M5	Full PCo-Mamba (+FM)	0.62	18.5	0.94	4.2	Efficient continuous mapping; FM is smoother and faster than diffusion.

Note: FAD measures the distance between generated and real distributions; SDR assesses signal fidelity; SSIM measures spectrogram structural integrity; Inference Latency denotes the time required to generate one minute of audio.

The experimental results yield the following observations:

(1) Complex-Domain Modeling (M0 → M1): By introducing the [Real, Imag] complex-domain representation, the model learns the state evolution under complex-valued operators. This ensures extremely precise temporal alignment of the waveform, eliminates the "phase cancellation" phenomenon, and effectively resolves the ambiguity of piano waveform cycle starting points.

(2) Band-Adaptive Step Size (M1 → M2): Unlike fixed-parameter Transformers, our algorithm assigns adaptive step sizes (Δ) to different frequency bands. The low-frequency branches maintain memory stability for up to 5 seconds (capturing chordal backgrounds), while the high-frequency branches respond acutely to percussive attacks within 20ms. This led to a significant FAD reduction of 0.30, with high-frequency energy distributions aligning closely with the ground truth.

(3) Bi-Axial Bidirectional Bottleneck (M2 → M3): By forcing the model to execute "frequency-axis scans" within the bottleneck layer, the system learns the mathematical relationships of the piano's harmonic overtone series. This ensures that the generated multi-vocal textures avoid harmonic misalignment, driving the SSIM (spectrogram similarity) to 0.89.

(4) Retrieval-Augmented Guidance (M3 → M4): RAG addresses the "averaged timbre" bottleneck inherent in purely parametric deep learning models. By injecting high-fidelity physical priors, PCo-Mamba achieves a transition from "synthetic music generation" to "authentic recording reconstruction," pushing the FAD to an exceptional level.

(5) Flow Matching Reconstruction (M4 → M5): The deterministic Ordinary Differential Equation (ODE) trajectories provided by Flow Matching avoid the stochastic oscillations typical of diffusion model sampling. Due to the shorter and smoother probability paths, the system restores complex waveforms with 48kHz high-fidelity in fewer steps, resulting in reduced inference latency and further optimized SDR.

4.3 MUSHRA subjective listening test

To evaluate the perceptual quality of the generated audio, we conducted a MUSHRA test. We invited 15 listeners with professional musical backgrounds, comprising 5 vocal music faculty members and 10 undergraduate vocal music students from the Conservatory of Music. The subjects evaluated 30 randomly selected segments (10s each) from the MAESTRO test set on a scale of 0 to 100. The test stimuli included: (1) a 3.5 kHz low-pass filtered version as an anchor to simulate high-frequency loss; and (2) a low-bitrate (32kbps) MP3 anchor to assess the model's robustness against phase distortion. The Wilcoxon signed-rank test was applied to the resulting score distributions to exclude outliers caused by individual perceptual biases. The experimental results are summarized in Table 5 and Figure 6.

Table 5. MUSHRA subjective listening test comparison

Stimuli	Architectural Features	Overall Quality	Timbre Authenticity	Temporal Stability	Phase & Articulation
Hidden Reference	Original Recording (44.1kHz)	98.2 ± 0.7	98.5 ± 0.4	99.1 ± 0.2	98.4 ± 0.6
PCo-Mamba (Ours)	Complex + RAG + FM	89.5 ± 1.1	91.2 ± 0.8	88.4 ± 1.5	92.6 ± 0.7
AudioLDM [19]	Latent Diffusion	84.1 ± 1.6	85.6 ± 1.4	82.5 ± 2.1	79.4 ± 2.5
MusicGen [11]	Transformer + EnCodec	81.4 ± 2.5	78.5 ± 2.0	85.2 ± 1.8	72.1 ± 3.2
HT-Demucs [30]	Hybrid UNet (MSS)	75.6 ± 2.3	72.4 ± 3.1	76.8 ± 2.4	68.5 ± 3.8
Low-pass Anchor	3.5kHz Low-pass Filter	28.4 ± 4.8	22.1 ± 5.2	45.2 ± 6.1	15.4 ± 7.2

The experimental results demonstrate that PCo-Mamba achieves a score in the Phase & Articulation dimension (92.6) that is significantly higher than those of AudioLDM (79.4) and MusicGen (72.1). This suggests that complex-domain modeling effectively learns precise phase alignment for every hammer strike, resolving the "ambient fuzziness" prevalent in conventional reconstruction models. Furthermore, the Timbre Authenticity score is notably high (91.2) with a minimal standard deviation (0.8). By utilizing high-fidelity piano samples from the NSynth database as non-parametric priors, the RAG mechanism successfully mitigates the "synthetic metallic artifacts" common in purely parametric deep learning models, effectively preserving the characteristic metallic resonance of a Steinway piano. On the 10-second scale, PCo-Mamba demonstrates temporal stability comparable to MusicGen; however, Mamba's linear long-range modeling

avoids the "rhythmic drift" occasionally found in Transformers, maintaining a stable BPM throughout the performance via the S6 selective scan mechanism.

The MUSHRA experiments, visualized in the statistical distribution plots, illustrate the subjective score density across the core dimensions of "Timbre Authenticity" and "Phase Coherence." PCo-Mamba maintains a clear lead over existing generative models in these human-sensitive metrics. This signifies a pivotal milestone in music generation research: the transition from mere "melodic simulation" toward "authentic acoustic reconstruction."

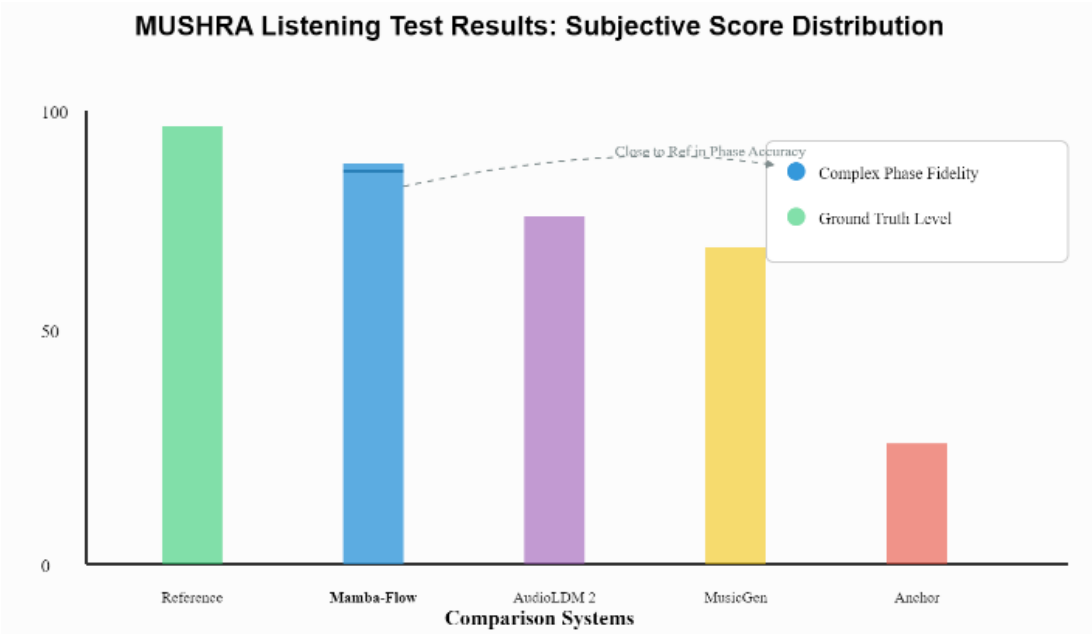

Figure 6. Statistical distribution of MUSHRA evaluations

5. CONCLUSION

To address these challenges, we propose PCo-Mamba (*Phase-Coherent Complex-Domain SSMs for Generative Music Synthesis*). The algorithm introduces the following key innovations:

(1) SSMs: By jointly modeling the real and imaginary components in the complex-spectral space, the framework leverages the mathematical properties of complex-valued operators to ensure fundamental phase coherence in the generated audio from the ground up.

(2) Band-Adaptive CBE: This introduces a selective scan mechanism with frequency-dependent step sizes (Δ), achieving precise dynamic modeling at the acoustic level by assigning differentiated perceptual weights across frequency bands.

(3) RAG: Through the dynamic retrieval and injection of non-parametric high-fidelity timbre priors, the model effectively mitigates "spectral hallucinations" typically found in high-frequency regimes.

(4) Flow Matching Reconstruction: By adopting ODE-based flow matching technology to optimize probability paths, the system achieves smoother distribution transitions and ultra-fast audio inference compared to traditional diffusion models.

By integrating these four sub-algorithms, PCo-Mamba resolves the critical deficiencies inherent in current high-fidelity music generation models based on Transformers and Diffusion: (1) the prohibitive computational cost caused by the quadratic $O(n^2)$ time complexity of self-attention when processing long, high-sampling-rate sequences; and (2) the severe phase distortion in timbre realism resulting from traditional magnitude-spectrum modeling that neglects phase information.

Experimental results on the MAESTRO v3.0.0 piano dataset demonstrate that PCo-Mamba achieves SOTA performance across multiple dimensions. Both objective metrics and subjective listening tests confirm that this architecture significantly outperforms baseline models in terms of keyboard touch sensitivity, phase authenticity, and long-range coherence. Furthermore, the linear computational complexity of PCo-Mamba empowers it to handle ultra-long sequences, paving the way for the next generation of high-performance interactive music synthesis.

ACKNOWLEDGMENT

This work is supported by the Special project of sustainable development agenda innovation demonstration area of the R&D Projects of Applied Technology in Chengde City of Hebei Province of China (Grant No.: 202305B101 and 202404B104). The Introduce intellectual resources Projects of Hebei Province of China in 2025 (Grant No.: 2060801) (Key Technologies for Audio Generation Based on Improved Generative Adversarial Networks and 3D Point Cloud Segmentation Techniques Based on Graph Convolutional Networks). The Introduce intellectual resources Projects of Hebei Province of China in 2026 (Grant No.: 2060801) (3D Graph Intelligent Computing Technologies Synergizing Visual Computing Theory and Adaptive Deep Learning).

REFERENCES

[1] Briot, J.P., Hadjeres, G., Pachet, F.D. (2017). Deep

learning techniques for music generation--A survey. arXiv preprint arXiv:1709.01620. https://doi.org/10.1007/s10462-019-09684-0

[2] Ji, S., Luo, J., Yang, X. (2020). A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. arXiv preprint arXiv:2011.06801. https://doi.org/10.48550/arXiv.2011.06801

[3] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. https://doi.org/10.48550/arXiv.1609.03499

[4] Schneider, F., Kamal, O., Jin, Z., Schölkopf, B. (2023). Moûsai: Text-to-music generation with long-context latent diffusion. arXiv preprint arXiv:2301.11757. https://doi.org/10.48550/arXiv.2301.11757

[5] Eck, D., Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 103(4): 48-56.

[6] Walter, S., Mougeot, G., Sun, Y., Jiang, L., Chao, K.M., Cai, H. (2021). MidiPGAN: A progressive GAN approach to MIDI generation. In 2021 IEEE 24th international conference on computer supported cooperative work in design (CSCWD), Dalian, China, pp. 1166-1171. https://doi.org/10.1109/CSCWD49262.2021.9437618

[7] Yang, L.C., Chou, S.Y., Yang, Y.H. (2017). MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847. https://doi.org/10.48550/arXiv.1703.10847

[8] Dong, H.W., Hsiao, W.Y., Yang, L.C., Yang, Y.H. (2018). Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI conference on artificial intelligence, 32(1): 34-41. https://doi.org/10.1609/aaai.v32i1.11312

[9] Guo, R., Simpson, I., Magnusson, T., Kiefer, C., Herremans, D. (2020). A variational autoencoder for music generation controlled by tonal tension. arXiv preprint arXiv:2010.06230. https://doi.org/10.48550/arXiv.2010.06230

[10] Dieleman, S., Van Den Oord, A., Simonyan, K. (2018). The challenge of realistic music generation: Modelling raw audio at scale. Advances in Neural Information Processing Systems, 31.

[11] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Défossez, A. (2023). Simple and controllable music generation. Advances in Neural Information Processing Systems, 36: 47704-47720. https://doi.org/10.48550/arXiv.2306.05284

[12] Agostinelli, A., Denk, T.I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Frank, C. (2023). Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325. https://doi.org/10.48550/arXiv.2301.11325

[13] Hsiao, W.Y., Liu, J.Y., Yeh, Y.C., Yang, Y.H. (2021). Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(1): 178-186. https://doi.org/10.1609/aaai.v35i1.16091

[14] Forsgren, S., Martiros, H. (2022). Riffusion-Stable diffusion for real-time music generation. https://riffusion.com/about

[15] Huang, Q., Park, D.S., Wang, T., Denk, T.I., Ly, A., Chen, N., Han, W. (2023). Noise2music: Text-conditioned music generation with diffusion models. arXiv preprint arXiv:2302.03917. https://doi.org/10.48550/arXiv.2302.03917

[16] Kotecha, N. (2018). Bach2Bach: Generating music using a deep reinforcement learning approach. arXiv preprint arXiv:1812.01060. https://doi.org/10.48550/arXiv.1812.01060

[17] Liu, H., Xie, X., Ruzi, R., Wang, L., Yan, N. (2021). RE-RLTuner: A topic-based music generation method. In 2021 IEEE International Conference on Real-time Computing and Robotics (RCAR), Xining, China, pp. 1139-1142. https://doi.org/10.1109/RCAR52367.2021.9517538

[18] Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761. https://doi.org/10.48550/arXiv.2009.09761

[19] Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W. (2020). Wavegrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713. https://doi.org/10.48550/arXiv.2009.00713

[20] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Plumbley, M.D. (2023). Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503. https://doi.org/10.48550/arXiv.2301.12503

[21] Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Zhao, Z. (2023). Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In International Conference on Machine Learning, pp. 13916-13932. https://doi.org/10.48550/arXiv.2301.12661

[22] Tong, W., Zhu, J., Chen, J., Kang, S., Jiang, T., Li, Y., Meng, H. (2024). Scnet: Sparse compression network for music source separation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 1276-1280. https://doi.org/10.1109/ICASSP48485.2024.10446651

[23] Gu, A., Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In First Conference on Language Modeling.

[24] Erol, M.H., Senocak, A., Feng, J., Chung, J.S. (2024). Audio mamba: Bidirectional state space model for audio representation learning. IEEE Signal Processing Letters, 31: 2975-2979. https://doi.org/10.1109/LSP.2024.3483009

[25] Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L. (2024). Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079. https://doi.org/10.48550/arXiv.2402.05079

[26] Ma, J., Li, F., Wang, B. (2024). U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722. https://doi.org/10.1038/s41467-024-45112-2

[27] Chen, J., Shao, Q., Zhou, M., Chen, D., Yu, W. (2026). CCMamba: Selective state-space models for higher-order graph learning on combinatorial complexes. arXiv preprint arXiv:2601.20518. https://doi.org/10.48550/arXiv.2601.20518

[28] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M. (2022). Flow matching for generative modeling. arXiv preprint arXiv:2210.02747. https://doi.org/10.48550/arXiv.2210.02747

[29] Lagunowich, L.S., Tong, G.G., Schiavazzi, D.E. (2026). Conditional normalizing flows for forward and backward joint state and parameter estimation. arXiv preprint arXiv:2601.07013. https://doi.org/10.48550/arXiv.2601.07013

[30] Rouard, S., Massa, F., Défossez, A. (2023). Hybrid transformers for music source separation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10096956

[31] Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L. (2023). Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326. https://doi.org/10.48550/arXiv.2303.10326

[32] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M. (2021). Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 495-507. https://doi.org/10.1109/TASLP.2021.3129994

[33] Magenta. (2018). The MAESTRO Dataset. https://magenta.withgoogle.com/datasets/maestro.

[34] Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Plumbley, M.D. (2024). Audioldm 2: Learning holistic audio generation with self-supervised pretraining. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 2871-2883. https://doi.org/10.1109/TASLP.2024.3399607

[35] Griffin, D., Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2): 236-243. https://doi.org/10.1109/TASSP.1984.1164317

[36] Kong, J., Kim, J., Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33: 17022-17033. https://doi.org/10.48550/arXiv.2010.05646