

Article

MusicDiffusionNet: Enhancing Text-to-Music Generation with Adaptive Style and Multi-Scale Temporal Mixup Strategies

Leiheng Xu ¹, Jiancong Chen ¹, Chengcheng Li ^{1,*}  and Jinsong Liang ² ¹ College of Modern Science and Technology, China Jiliang University, Hangzhou 310018, China² School of Energy & Environmental Engineering, Hebei University of Technology, Tianjin 300130, China

* Correspondence: lcc_1027@cjlu.edu.cn

Abstract

Text-to-music generation aims to automatically produce audio content with semantic consistency and coherent musical structure based on natural language descriptions. However, existing methods still face challenges in terms of style diversity, rhythmic consistency, and long-term structural modeling. To address these issues, we propose a novel text-to-music generation model, termed MusicDiffusionNet (MDN), which integrates diffusion models with the WaveNet architecture to jointly model musical semantics and temporal structure in a continuous latent space. By decoupling high-level semantic conditioning from low-level audio generation, MDN enhances its ability to model long-range musical structure while improving semantic alignment between text and generated music with stable generation behavior. Building upon this framework, we further design two complementary mixing strategies to improve generation quality and structural coherence. Adaptive Style Mixing (ASM) performs weighted interpolation among stylistically similar music samples in the style embedding space, incorporating key and harmonic compatibility constraints to expand the style distribution while avoiding dissonance. Multi-scale Temporal Mixing (MTM) adopts beat-aware temporal decomposition, mixing, and reorganization across multiple time scales, thereby enhancing the modeling of both local and global temporal variations while preserving rhythmic periodicity and musical groove. Both strategies are integrated into the diffusion process as conditional augmentation mechanisms, contributing to improved learning stability and representational capacity under limited data conditions. Experimental results on the Audiosstock dataset demonstrate that MDN and its mixing strategies achieve consistent improvements across multiple objective metrics, including generation quality, style diversity, and rhythmic coherence, validating the effectiveness of the proposed approach for text-to-music generation.

Keywords: music generation; text-to-music; diffusion model; deep learningAcademic Editor: Douglas
O'Shaughnessy

Received: 18 January 2026

Revised: 13 February 2026

Accepted: 17 February 2026

Published: 20 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article
distributed under the terms andconditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the continuous exploration and advancement of artificial intelligence technologies, text-guided generation tasks have achieved remarkable progress across multiple domains, including text-to-image, text-to-video, and text-to-audio generation [1]. Among them, text-to-music generation, as a specialized form of audio generation, has attracted significant attention from the research community due to its broad application potential in personalized content creation, virtual characters, film scoring, and game scenarios [2].

Compared to traditional symbolic or probabilistic modeling approaches, diffusion models [3] gradually approximate the data distribution within continuous latent spaces,

enabling more effective modeling of complex structures and long-range dependencies. Their success in image generation, particularly in terms of generation quality and semantic consistency, has been widely validated, providing valuable insights for addressing challenges such as complex structural modeling and semantic alignment in music generation. Furthermore, in the audio and music generation domain, diffusion models have gradually expanded from natural sound modeling to more structurally complex music generation tasks, demonstrating strong potential in continuous-time modeling and fine-grained detail preservation. In text-to-music generation scenarios, diffusion models must additionally address multiple challenges, including text–music semantic alignment, style control, and long-term temporal structure modeling [4].

Although diffusion models provide a promising foundation for text-to-music generation, existing diffusion-based approaches still exhibit several methodological limitations, particularly under limited paired data and high requirements for musical diversity and structural coherence. These limitations stem not only from data scarcity, but also from the design choices of current diffusion-based generation frameworks.

- (1) Scarcity of training data and limited style-level generalization. High-quality text–music paired data remain scarce due to the structural complexity and subjectivity of music, as well as the difficulty of fine-grained annotation. Under such conditions, most diffusion-based models rely on single-sample conditional training, lacking explicit mechanisms to expand the style distribution in latent space, which often leads to overfitting and reduced stylistic diversity [5].
- (2) Insufficient modeling of multi-scale temporal structure. While diffusion models are effective at capturing local acoustic details, most existing approaches perform denoising over uniformly sampled temporal representations, without explicitly modeling hierarchical musical structures such as beats, bars, and long-range rhythmic patterns. This limitation becomes more pronounced in small-sample settings, frequently resulting in unstable rhythm and fragmented musical form [6].
- (3) Limited controllability of originality and diversity. Due to the highly structured nature of music, diffusion models trained on limited data are prone to memorization and segment-level repetition. However, most existing frameworks [7,8] treat originality as an implicit consequence of stochastic sampling, rather than addressing it through explicit structural or distribution-level mechanisms, which raises concerns regarding diversity and potential copyright risks.

The above challenges reveal two key research gaps in diffusion-based text-to-music generation: (i) the lack of explicit mechanisms for expanding the conditional style distribution under limited paired data, and (ii) the absence of temporally structured modeling and augmentation strategies that account for music’s multi-scale rhythmic organization.

To address the above challenges, this paper proposes a text-conditioned music generation framework termed MDN. MDN adopts diffusion models as the core generative mechanism and performs conditional modeling in a continuous audio latent space. A modular design is employed to decouple text semantic modeling, audio latent representation generation, and audio decoding. In our implementation, MDN is adapted from the conditional diffusion framework of Stable Diffusion [9] and incorporates WaveNet [10] to model temporal continuity and rhythmic details of audio signals, enabling the reconstruction of high-fidelity waveform audio from diffusion-generated latent representations. The main contributions of this work are summarized as follows.

- We propose a general text-conditioned music generation framework, MDN. Centered on diffusion models operating in a continuous audio latent space, MDN decouples text semantic modeling, audio latent representation generation, and audio decoding through a modular design, providing a general and extensible modeling framework for text-to-music generation.
- Within the MDN framework, we systematically design the music generation process from the perspectives of style modeling and temporal structure modeling. We introduce two mechanisms, ASM and MTM, to alleviate issues such as high repetition and unstable temporal structure under small-sample conditions, thereby improving the diversity and structural consistency of generated music.
- We conduct comprehensive experimental evaluations on the Audiostock dataset, comparing the proposed method with multiple existing baseline models across several dimensions, including semantic alignment, originality, style diversity, and plagiarism risk. The results demonstrate the effectiveness of the proposed approach.

2. Related Works

Early approaches [11] to text-to-music generation were primarily based on rule-based systems and symbolic-level modeling, where predefined music theory rules or probabilistic models were applied to symbolic representations such as MIDI to control melody, harmony, and rhythmic structure. These methods typically relied on handcrafted rules derived from tonal harmony, chord progression patterns, and rhythmic templates, enabling explicit control over musical form and structure. While such approaches provided a degree of interpretability and controllability, they were limited in their ability to model complex musical textures, expressive timing, and long-range dependencies, and they struggled to generalize to diverse musical styles or to incorporate semantic conditioning from textual descriptions.

Subsequently, the application of machine learning has significantly advanced music generation technology. Early methods relied on statistical approaches such as Markov models [12] and genetic algorithms [13] to model musical structures, but their limited contextual modeling ability made it difficult to maintain long-range coherence and align with natural language semantics. Recurrent neural networks (RNNs) [14] and long short-term memory (LSTM) networks [15] later improved sequence modeling and were gradually extended to text-to-music generation, enabling the inclusion of richer musical elements. These models were gradually extended to text-to-music generation by incorporating conditional inputs and embedding representations, enabling the generation of richer musical elements, smoother temporal transitions, and more expressive dynamics compared with earlier statistical approaches.

With the rise of deep learning, generative adversarial networks (GANs) [16] and Transformers [17] further enhanced the realism and diversity of generated music. More recently, diffusion models have emerged as a powerful paradigm, showing strong performance in cross-modal generation, particularly for producing high-quality and semantically consistent audio from textual or multimodal inputs. For example, Huang et al. [6] propose Noise2Music, which uses cascaded diffusion models to generate music from text prompts with good semantic alignment. Schneider et al. [18] propose Moûsai, a two-stage latent diffusion model capable of generating long, high-quality music efficiently from text descriptions. Liu et al. [19] introduce AudioLCM, which accelerates text-to-audio generation by incorporating latent consistency models while maintaining high audio quality. Lanzendörfer et al. [20] propose DiscoDiff, a hierarchical latent diffusion model that improves audio quality and text–music alignment through a coarse-to-fine generation strategy. Kim et al. introduce AudioLDM [21], a low-rank adaptation to fine-tune selected attention and projection layers, enabling efficient genre-aware text-to-music generation with

improved controllability and semantic alignment under limited data and computational cost. These methods model music signals in continuous latent spaces, further improving audio quality and temporal coherence in the generated results.

3. Methodology

As illustrated by the overall workflow in Figure 1, the proposed framework can be understood as a three-stage process. First, MDN performs conditional modeling in a continuous audio latent space, where text semantic modeling, audio latent representation generation, and audio decoding are decoupled to improve stability and long-context modeling capability. Second, two complementary mixing mechanisms are introduced to enhance generation quality. ASM interpolates and combines music segments of different styles in the embedding space based on a Beta distribution [22], thereby expanding the generative distribution at the stylistic semantic level and improving diversity. In parallel, MTM operates across multiple temporal scales by segmenting music clips, injecting noise, and reorganizing temporal structures, enabling coherent melody and rhythm at both micro- and macro-structural levels. Finally, the model is trained in two phases on the Audiostock dataset: the base diffusion model is first trained, followed by further training with ASM and MTM enabled, after which the trained model is used for conditional music generation.

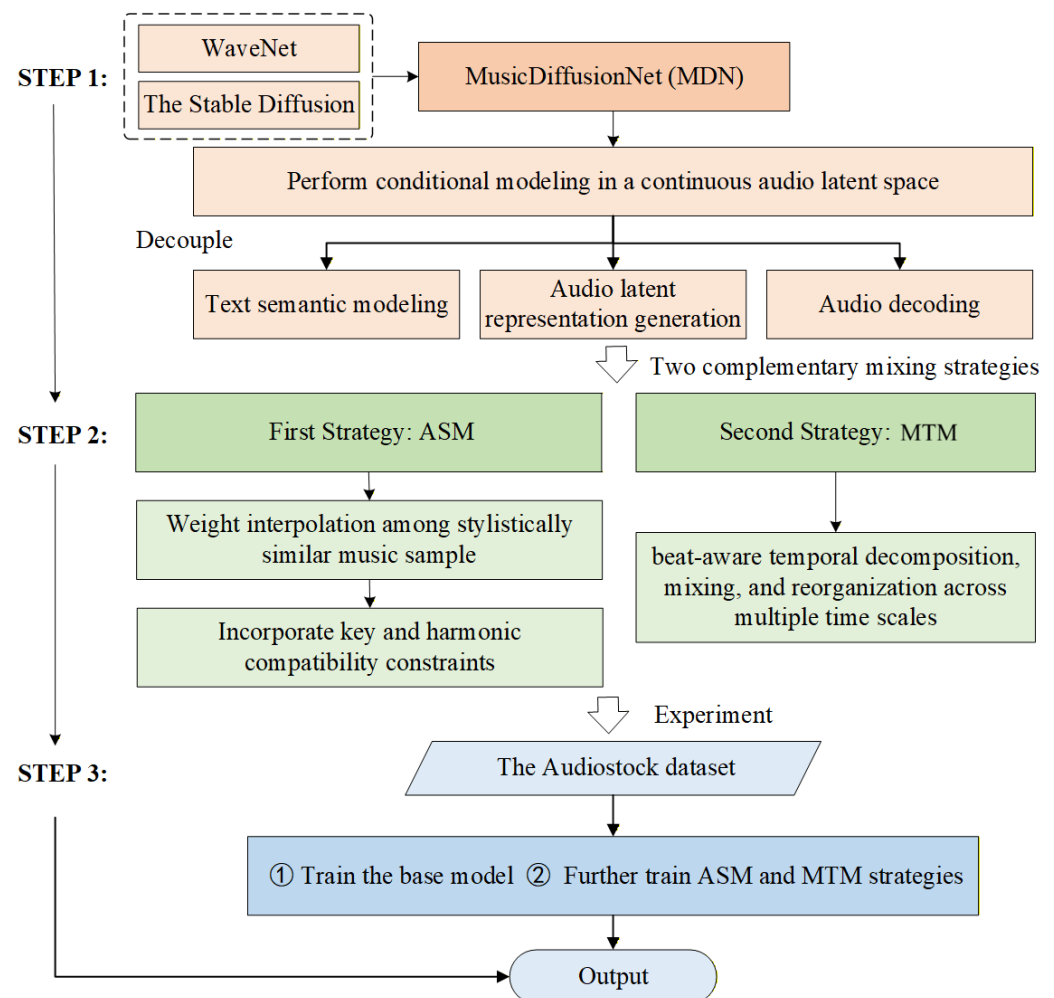


Figure 1. Workflow of the proposed MusicDiffusionNet framework.

As illustrated by the overall architecture in Figure 2, in terms of architectural design, MDN adopts a continuous latent space representation that supports long-context modeling, thereby enhancing its ability to capture long-range musical structures. Meanwhile, by decoupling audio latent representation modeling from high-level semantic modeling and incorporating Transformer-based conditional control, the model improves semantic alignment while maintaining generation stability. During inference, a parallelization and caching strategy based on Fast WaveNet [23] is employed to alleviate the computational overhead introduced by sample-by-sample decoding and to improve overall decoding efficiency.

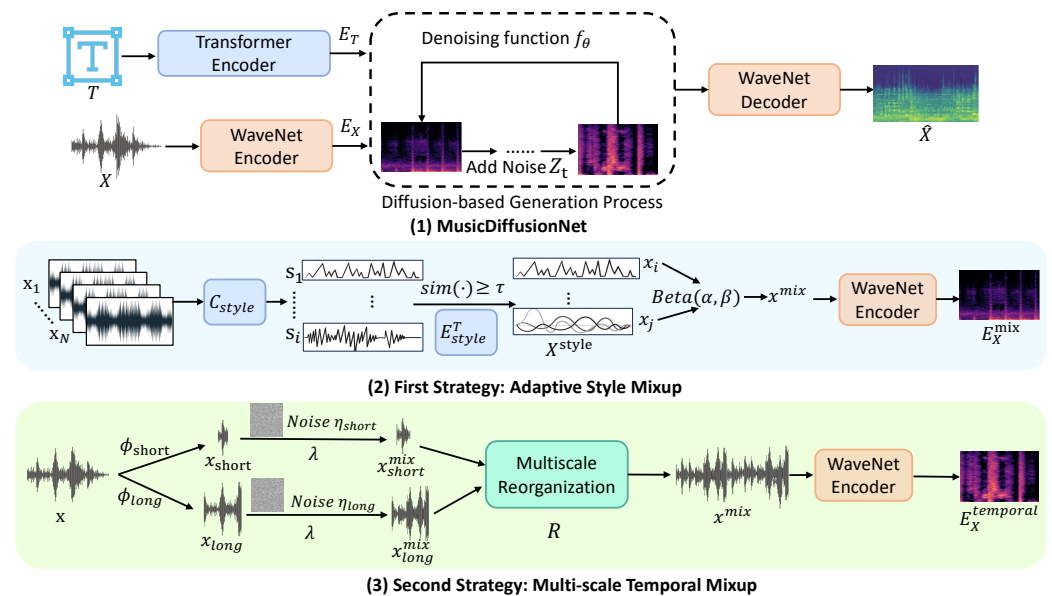


Figure 2. The diagram illustrates the architecture of MDN, which combines ASM and MTM for text-to-music generation.

3.1. MusicDiffusionNet

3.1.1. Text and Audio Feature Extraction Module

The input text sequence $T = \{t_1, t_2, \dots, t_n\}$ is first encoded through a pre-trained Transformer encoder, mapped to an embedding vector $E_T \in \mathbb{R}^{n \times d}$ in a high-dimensional semantic space:

$$E_T = \text{LayerNorm}(\text{MultiHeadAttention}(T, T, T) + T). \quad (1)$$

This encoder adopts a fine-tuned T5-small architecture specifically tailored for music semantic tasks. Meanwhile, the input audio signal X is processed through a WaveNet encoder to extract a time–frequency feature matrix $E_X \in \mathbb{R}^{m \times d}$:

$$E_X = \text{Conv1D}(\sigma(\text{ResidualBlock}(X))). \quad (2)$$

The WaveNet encoder, with its efficient expressive capability in audio modelling, effectively captures rhythm, melody, and timbre information in music signals.

It is worth noting that, compared to GAN-based vocoders [24] or neural audio codecs [25], WaveNet is not the most efficient choice in terms of generation speed or perceptual audio quality. However, its autoregressive, sample-by-sample modeling paradigm provides explicit temporal continuity constraints during audio generation, resulting in higher stability when modeling fine-grained rhythmic variations and complex temporal dependencies. This property is particularly beneficial in generation scenarios involving multi-time-scale recomposition or non-stationary structures, as it helps mitigate rhythmic

jitter and transient discontinuities. Furthermore, the proposed framework adopts a modular architecture, allowing this component to be flexibly replaced with other neural audio decoders or vocoders under different task requirements and computational constraints, without affecting the overall diffusion modeling pipeline.

3.1.2. Diffusion-Based Generation Process

During the diffusion phase, the model starts from y_0 (target audio embedding) and generates intermediate variables z_t through forward diffusion by adding noise:

$$q(z_t|y_0) = \mathcal{N}(z_t; \sqrt{\alpha_t}y_0, (1 - \alpha_t)I), \quad (3)$$

where α_t is the noise intensity adjustment factor.

In the denoising step, the signal is gradually restored by learning the denoising function:

$$p_\theta(z_{t-1}|z_t, T, E_T, E_X) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, T, E_T, E_X), \sigma_t^2 I), \quad (4)$$

where f_0 combines text and audio embeddings to enhance cross-modal semantic consistency.

3.1.3. Audio Decoding Module

Given an input text sequence

$$T = \{t_1, t_2, \dots, t_n\}, \quad (5)$$

the text is first encoded by a pre-trained Transformer encoder and mapped into a high-dimensional semantic embedding space:

$$E_T \in \mathbb{R}^{n \times d}. \quad (6)$$

Specifically, the text embedding is obtained as:

$$E_T = \text{LayerNorm}(\text{MultiHeadAttention}(T, T, T) + T). \quad (7)$$

The encoder adopts a T5-small architecture fine-tuned for music-related semantic tasks.

Meanwhile, the input audio signal X is processed by a WaveNet encoder to extract a time–frequency feature representation:

$$E_X \in \mathbb{R}^{m \times d}, \quad (8)$$

which is computed as:

$$E_X = \text{Conv1D}(\sigma(\text{ResidualBlock}(X))). \quad (9)$$

Benefiting from its strong representational capacity for audio signals, the WaveNet encoder effectively captures rhythmic, melodic, and timbral characteristics of music.

3.2. The First Strategy: ASM

To alleviate data scarcity and overfitting in diffusion-based text-to-music generation, MDN introduces a style-aware mixing mechanism termed Adaptive Style Mixing (ASM), as illustrated in Algorithm 1. ASM is designed to enhance stylistic diversity while maintaining semantic consistency with the text condition, by expanding the effective style distribution within the conditional generation process.

Algorithm 1. ASM

Require: Dataset $X = \{x_i\}_{i=1}^N$, text prompt $\mathcal{T}$, style classifier C_{style}
Ensure: Mixed audio embedding E_X

- 1: $E \leftarrow \text{Text2StyleEmb}(\mathcal{T})$
- 2: **for all** $x_i \in X$ **do**
- 3: $s_i \leftarrow C_{\text{style}}(x_i)$
- 4: **end for**
- 5: $X_{\text{style}} \leftarrow \{x_i \mid \text{Sim}(E, s_i) \geq \tau\}$
- 6: **Select** $(x_i, x_j) \in X_{\text{style}}$ with key compatibility
- 7: $\lambda \sim \text{Beta}(\alpha, \beta)$
- 8: $x_{\text{mix}} \leftarrow \lambda x_i + (1 - \lambda)x_j$
- 9: $E_X \leftarrow W_{\text{audio}}(x_{\text{mix}})$
- 10: **return** E_X

3.2.1. Style Classification and Embedding Extraction

Given a collection of music samples

$$X = \{x_1, x_2, \dots, x_N\}, \quad (10)$$

a pre-trained style classifier C_{style} is used to extract style embeddings:

$$s_i = C_{\text{style}}(x_i), \quad \forall i \in \{1, 2, \dots, N\}. \quad (11)$$

The classifier is based on a convolutional neural network and pre-trained on publicly available labeled music datasets (e.g., GTZAN) to extract fixed-length semantic representations encoding genre and emotion. To reduce style ambiguity, we perform segment normalization and single-style pre-filtering to improve label consistency.

3.2.2. Adaptive Style Sample Selection

Given a target style embedding $E \in \mathbb{R}^d$ derived from the text prompt, we select a subset of stylistically similar samples:

$$X_{\text{style}} = \{x_i \in X \mid \text{Sim}(E, s_i) \geq \tau\}, \quad (12)$$

where $\text{Sim}(\cdot, \cdot)$ denotes cosine similarity. The threshold τ is determined empirically or by selecting the top 10% most similar samples.

To further ensure harmonic compatibility, we introduce a key-consistency constraint. Using chroma-based key estimation methods (e.g., the Krumhansl–Schmuckler algorithm), each sample is assigned a musical key k_i . A sample pair (x_i, x_j) is allowed for mixing only if

$$\Delta k = \text{dist}_{\text{circle-of-fifths}}(k_i, k_j) \leq \delta, \quad (13)$$

where $\text{dist}_{\text{circle-of-fifths}}(\cdot, \cdot)$ measures key distance on the circle of fifths and δ is the maximum allowable deviation (set to two fifths in this work). This constraint prevents mixing between highly incompatible keys (e.g., C major and F# minor), thereby reducing dissonance. To avoid style bias and mode collapse, we limit the maximum number of samples per sub-style.

3.2.3. Style Mixing and Interpolation

For each qualified sample pair (x_i, x_j) , we apply Beta-distribution-based interpolation:

$$x_{\text{mix}} = \lambda x_i + (1 - \lambda)x_j, \quad \lambda \sim \text{Beta}(\alpha, \beta). \quad (14)$$

If minor key mismatches remain, we perform pitch alignment using a phase-vocoder-based pitch-shifting method, restricting shifts to within ± 2 semitones.

To mitigate formant distortion, we apply the following constraints:

- Pitch adjustment is performed in the Mel-spectrogram domain rather than direct waveform resampling;
- For vocal or polyphonic segments, only integer-semitone shifts are applied;
- Post-shift spectral normalization and smoothing are used to suppress abnormal resonance amplification.

This design preserves timbral structure while ensuring harmonic alignment. Additionally, we constrain the style similarity between mixed samples as

$$0.3 \leq \text{Sim}(s_i, s_j) \leq 0.7, \quad (15)$$

to ensure sufficient stylistic tension.

3.2.4. Embedding Transformation

The mixed audio sample x_{mix} is encoded by the WaveNet encoder

$$E_X = W_{\text{audio}}(x_{\text{mix}}), \quad (16)$$

producing an audio embedding that preserves fine-grained temporal and pitch structures. This embedding, together with the text embedding E_T , serves as conditional input to the diffusion model.

In the ASM module, style representations form the basis for similarity-based selection and mixing. Specifically, audio samples are converted into Mel spectrograms and processed by a CNN to extract fixed-length style embeddings $s_i \in \mathbb{R}^{d_s}$ (see Figure 3). Mel spectrograms provide a robust mid-level representation aligned with human auditory perception while preserving pitch and rhythmic information [26]. The CNN used for style embedding extraction is a lightweight convolutional architecture consisting of multiple convolutional blocks with batch normalization and pooling, designed to capture timbral, rhythmic, and spectral patterns at different time–frequency scales. Its design follows widely adopted architectures in music tagging and emotion recognition tasks [27], which have demonstrated strong generalization and discriminative capability. In our implementation, this CNN operates within the VGGish-based feature extraction framework illustrated in Figure 3, where the convolutional layers follow the general structure of VGG-style networks adapted for audio analysis. VGGish serves as the backbone for extracting perceptually meaningful embeddings, while the additional processing layers refine these representations for style similarity estimation. During preprocessing, samples with multi-label ambiguity, overlapping styles, or excessive noise are removed, and preliminary rhythm and pitch normalization is applied using the librosa Python library (version 0.11.0).

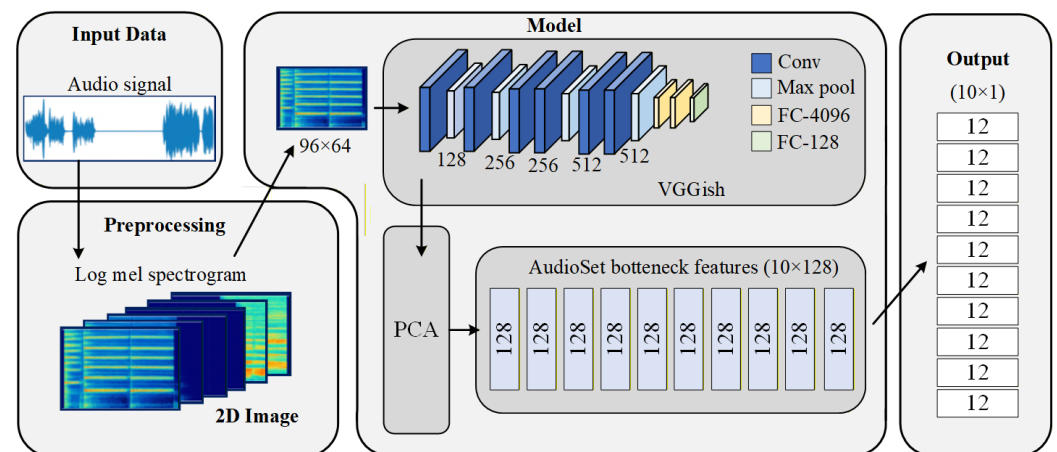


Figure 3. Audio style embedding extraction flowchart. The original audio signal is first framed, windowed, and processed through a Mel filter bank to obtain the Mel spectrogram. This spectrogram is input as a two-dimensional image into the convolutional neural network, which, after multiple layers of convolution and pooling, outputs a fixed-length style embedding vector used to describe the style attributes of the music (such as genre, emotion, etc.).

3.3. The Second Strategy: MTM

To enhance the model's ability to capture musical structure and rhythmic consistency across multiple temporal scales, we propose a beat-aware Multi-Time-Scale Mixing (MTM) mechanism, as illustrated in Algorithm 2. MTM performs structured temporal decomposition and recombination while explicitly introducing beat and downbeat alignment constraints, thereby preventing arbitrary temporal concatenation from disrupting musical groove and rhythmic continuity. The overall procedure consists of four stages.

Algorithm 2. MTM

Require: Music samples x_1, x_2 , mixing parameters (α, β) , noise scale σ , WaveNet encoder W_{audio}

Ensure: Temporal audio embedding E_X^{temporal}

- 1: $B \leftarrow \text{BeatTrack}(x_1); D \leftarrow \text{DownbeatEst}(x_1)$
 - 2: $(x_{1,\text{short}}, x_{1,\text{long}}) \leftarrow \text{Decompose}(x_1, B)$
 - 3: $(x_{2,\text{short}}, x_{2,\text{long}}) \leftarrow \text{Decompose}(\text{TempoAlign}(x_2, x_1), B)$
 - 4: **for** $k \in \{\text{short}, \text{long}\}$ **do**
 - 5: $\lambda_k \sim \text{Beta}(\alpha, \beta); \eta_k \sim \mathcal{N}(0, \sigma^2)$
 - 6: $x_k^{\text{mix}} \leftarrow \lambda_k x_{1,k} + (1 - \lambda_k) x_{2,k} + \eta_k$
 - 7: **end for**
 - 8: $x^{\text{mix}} \leftarrow R(x_{\text{short}}^{\text{mix}}, x_{\text{long}}^{\text{mix}}, B, D)$ {downbeat-constrained recombination}
 - 9: $E_X^{\text{temporal}} \leftarrow W_{\text{audio}}(x^{\text{mix}})$
 - 10: **return** E_X^{temporal}
-

3.3.1. Multi-Time-Scale Decomposition

Given a music sample

$$x \in \mathbb{R}^{T \times d_x},$$

we first estimate its beat sequence using a beat-tracking algorithm based on spectral energy and rhythmic periodicity (e.g., `beat_track` in `librosa`)

$$B = \{b_1, b_2, \dots, b_M\},$$

where b_i denotes the position of the i -th beat. Downbeat positions are further estimated based on beat strength and periodic stability to capture bar-level structure.

As illustrated in Figure 4, we perform beat-aligned multi-scale temporal decomposition using beats as alignment anchors

$$\phi_k(x, t_i) = \sum_t \omega_k(t - t_i) \cdot x(t), \quad k \in \{\text{short, long}\},$$

where the support of the temporal weighting function $\omega_k(\cdot)$ is constrained to integer numbers of beats or bars. The short-term scale corresponds to intra-beat or beat-level windows, capturing local rhythmic patterns and transient events (e.g., drum hits and ornaments), while the long-term scale spans multiple beats or bars to model global musical structures such as sectional repetition and rhythmic evolution [28]. This beat-aligned decomposition expands the effective receptive field while preserving the original rhythmic skeleton.

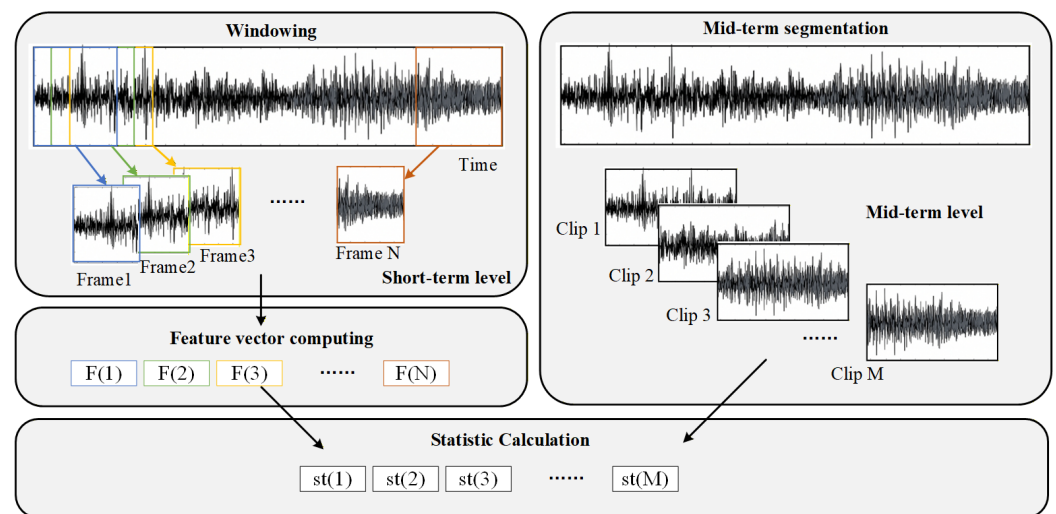


Figure 4. Schematic diagram of dual-scale temporal decomposition. The upper part shows the original audio waveform, which is processed through short-term frame-level windowing to generate multiple short-term feature vectors. The lower part shows mid-term segmentation, which is used to capture structural information over a larger range. This figure effectively demonstrates how the model extracts features at different temporal granularities, providing a foundation for the subsequent mixing and recombination of short-term and long-term scales.

3.3.2. Beat-Aligned Short-Term and Long-Term Mixing

Given two music samples x_1 and x_2 , mixing is performed only on beat-synchronized multi-scale representations:

$$x_k^{\text{mix}} = \lambda_k x_{1,k} + (1 - \lambda_k) x_{2,k} + \eta_k, \quad k \in \{\text{short, long}\},$$

where $\lambda_k \sim \text{Beta}(\alpha, \beta)$ and $\eta_k \sim \mathcal{N}(0, \sigma^2)$.

Before mixing, linear time-stretching is applied to align x_1 and x_2 to a common tempo (BPM), ensuring one-to-one correspondence between beat positions at each temporal scale. Unlike ASM, MTM does not focus on stylistic variation; instead, it perturbs temporal organization while preserving rhythmic periodicity, thereby improving robustness to rhythmic variation and structural diversity.

3.3.3. Downbeat-Constrained Multi-Scale Recomposition

After multi-scale interpolation, temporal segments are aggregated using a recomposition function $R(\cdot)$

$$R : \quad x^{\text{mix}} = \sum_{j=1}^N \gamma_j x_j^{\text{mix}},$$

where the weights γ_j depend not only on temporal position but also explicitly on beat and downbeat information. Higher weights are assigned near downbeat locations to preserve bar boundaries and rhythmic accents, while more flexible mixing is allowed at non-accented positions to enhance local variability. This strategy is inspired by time-weighted reconstruction techniques in speech enhancement and music generation, enabling multi-scale temporal perturbation while maintaining beat- and bar-level structural consistency.

3.3.4. Rhythm-Preserving Audio Embedding

Finally, the beat-aligned and recomposed mixed audio sample x^{mix} is processed by the WaveNet encoder

$$E_X^{\text{temporal}} = W_{\text{audio}}(x^{\text{mix}}),$$

producing an audio embedding that simultaneously encodes fine-grained rhythmic details and global temporal structure. This embedding, together with the textual semantic embedding E_T , serves as conditional input to the diffusion model for final music generation.

By introducing explicit beat and downbeat alignment constraints in MTM, the model learns more diverse temporal structures without compromising musical groove. Although WaveNet incurs relatively high computational cost, its strong capability in rhythmic modeling and waveform detail preservation makes it well suited as the temporal encoder within MTM.

MDN enhances long-term music structure modeling by adopting diffusion-based modeling in a continuous latent space and decoupling high-level semantic conditioning from low-level audio generation. Building upon this foundation, we introduce two complementary strategies: ASM and MTM. ASM expands the stylistic distribution in the style embedding space by incorporating tonal and harmonic compatibility constraints, while MTM improves rhythmic coherence and structural stability through beat-aware multi-scale temporal modeling. Together, these designs enable MDN to generate music that is both diverse and structurally consistent while preserving musicality and groove.

4. Experimental Setup and Evaluation

We trained and evaluated the MDN model on the Audiostock dataset [5], which contains 9000 training samples and 1000 test samples, totaling about 46.3 h. While Audiostock is primarily an audio dataset, the samples used in this work are exclusively musical audio tracks with accompanying metadata and textual descriptions. These descriptions provide information about the music's mood, instrumentation, and intended use case, forming the basis for the text-to-music generation task. All audio was uniformly sampled at 16 kHz mono, pre-processed with dynamic range compression and mute removal. Mel spectra were used as feature representations and embeddings were extracted by a pre-trained WaveNet encoder [29]. The Transformer part of MusicDiffusionNet uses a 12-layer encoder with 8 heads per layer and an embedding dimension of 512. A 20-layer residual block was used for the WaveNet encoder and decoder. The diffusion model uses an inverse process with 1000 time steps. The model is trained in two phases: first the base model is trained and then ASM and MTM strategies are introduced for further training. Training is performed on eight NVIDIA A100 GPUs. The batch size is 256, the initial learning rate is 2×10^{-4} and gradually decays to 1×10^{-6} ; using an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, the weights are decayed to 10^{-4} , the optimizer is Adam, 800 Epochs are trained, and the loss function is the mean square error (MSE).

Evaluation metrics include Fréchet Audio Distance (FAD), Precision–Recall Distributions (PRD), FDvgg, FDpann, Inception Score (IS), and Kullback–Leibler Divergence (KLD). These metrics evaluate generation quality from a distributional perspective in learned audio embedding spaces. FAD, FDvgg, and FDpann measure the distance between the

distributions of generated and real music using pre-trained audio feature extractors that encode high-level semantic attributes such as timbre, rhythm, instrumentation, and genre. The Fréchet distance between two multivariate Gaussian distributions with means μ_r, μ_g and covariance matrices Σ_r, Σ_g is computed as

$$FD = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (17)$$

Lower values indicate higher audio fidelity and better semantic consistency with real music. FAD, FDvgg, and FDpann differ primarily in the feature extractor used to compute embeddings.

PRD further analyzes the trade-off between sample quality (precision) and diversity (recall). Precision and recall are estimated by comparing generated and real sample distributions in the embedding space

$$\text{Precision} = \frac{|G \cap R|}{|G|}, \quad \text{Recall} = \frac{|G \cap R|}{|R|} \quad (18)$$

where G and R denote the generated and real sample manifolds, respectively.

IS assesses both confidence and diversity of predictions made by a pre-trained audio classifier and is defined as

$$IS = \exp(\mathbb{E}_x [D_{KL}(p(y|x) \| p(y))]) \quad (19)$$

where $p(y|x)$ is the conditional label distribution predicted by the classifier and $p(y)$ is the marginal distribution over generated samples.

KLD measures the divergence between two probability distributions:

$$D_{KL}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (20)$$

Lower KLD indicates improved distributional alignment and reduced overfitting in conditional generation.

4.1. Generation Quality

We evaluated the generation quality of the MDN model through Table 1, focusing on the impact of two strategies, ASM and MTM, on the quality of the generated music, and compared it with the baseline model (AudioLDM, MusicGen, Riffusion, DiscoDiff). For reference-based metrics such as FDvgg and FDpann, we use a held-out validation subset of 500 real music samples from Audiostock as the reference set to compute Fréchet distances. The experimental results show that the MDN model combining ASM and MTM strategies performs superiorly on several evaluation metrics. In particular, on FDvgg and FDpann, the music generated by MDN is closer to the real samples; on IS metrics, MDN exhibits higher diversity and quality; on KL Divergence, the music distribution generated by the MDN model is closer to the real data. While traditional plagiarism detection in music relies on fingerprint-based systems or human judgment, our evaluation approximates originality using distributional similarity metrics. Although these do not directly measure plagiarism, they offer insights into distributional shifts that may correlate with originality.

Overall, the results indicate that mixing strategies in representation space can effectively improve diffusion-based text-to-music generation without modifying the core architecture. ASM mainly enhances stylistic diversity, while MTM improves temporal coherence, and their combination provides complementary gains in both quality and structure. These findings highlight the importance of jointly modeling stylistic variation and

temporal structure, and suggest that modular mixing strategies are a practical way to improve conditional audio generation.

Table 1. Evaluation of generation quality for MDN and baseline models.

Model	FDvgg ↓	FDpann ↓	IS ↑	KL Div. ↓
Riffusion (2022)	68.95	10.77	1.34	5.00
Moûsai (2023)	26.67	2.40	1.81	3.60
MusicGen (2024)	25.19	2.17	1.72	3.10
DiscoDiff (2025)	24.23	2.20	1.81	3.12
MDN (No Mixup)	26.67	2.40	1.81	3.80
MDN w/ASM	24.95	2.31	1.79	3.40
MDN w/MTM	23.54	2.28	1.83	3.32
MDN w/ASM + MTM	22.81	2.15	1.84	3.20

Figure 5 shows a comprehensive comparison of the performance of our proposed MDN model and its different variants across four key metrics: FDvgg, FDpann, IS, and KLD. As clearly shown in the Figure 5, the “MDN + ASM + MTM” combination achieves the best performance across all metrics, with FDvgg and FDpann values of 22.81 and 2.15, respectively, significantly outperforming other baselines (e.g., MusicGen’s FDvgg is 25.19 and FDpann is 2.17). Additionally, its IS reaches 1.84, and KLD drops to 3.20, balancing both generation quality and distribution consistency. This result clearly demonstrates that the ASM strategy (Adaptive Style Mixup) significantly enhances the semantic consistency and stylistic diversity of generated samples; the MTM strategy (Multiscale Temporal Mixup) effectively improves rhythmic structural coherence, preventing fragmented generation; and combining both (ASM + MTM) achieves comprehensive optimization of style and structure, effectively mitigating overfitting issues and enhancing the originality and expressiveness of generated music.

Overall, these results suggest that improving stylistic diversity and temporal structure is critical for diffusion-based text-to-music generation. ASM contributes mainly to expanding stylistic variation, while MTM enhances temporal coherence and structural consistency, and their combination provides complementary gains. From a methodological perspective, this demonstrates that mixing strategies in representation space can effectively improve generation quality without modifying the core diffusion architecture, offering a practical approach for enhancing conditional audio generation.

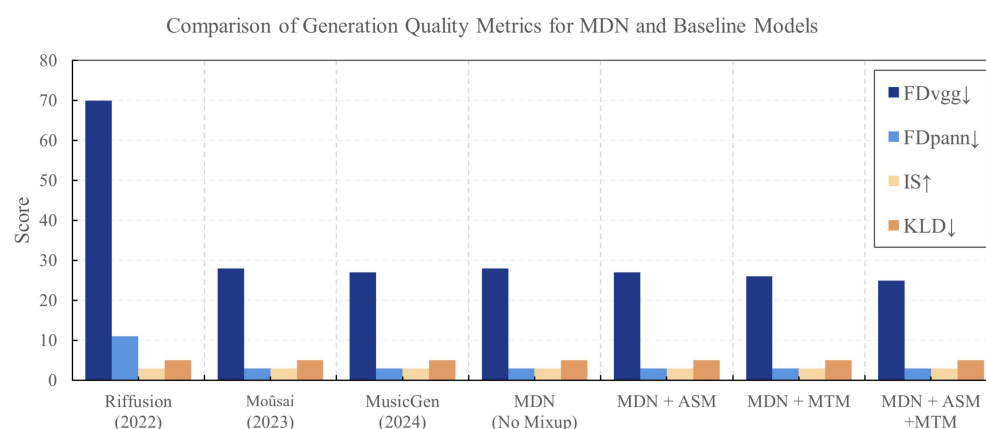


Figure 5. Comparison of generation quality metrics for MDN and baseline models.

4.2. Ablation Study

4.2.1. Ablation Study on Mixed Strategy

We evaluate the generative effectiveness of the MDN model using well-established metrics commonly used in text-to-audio and text-to-music generation tasks. Table 2 highlights the contributions of ASM and MTM strategies to relevance, novelty, and plagiarism risk. While ASM improves semantic alignment, MTM enhances novelty and reduces plagiarism risk. Combining the two strategies yields the best overall performance.

Table 2. Text–audio relevance, novelty, and plagiarism risk improvements by ASM and MTM strategies.

Strategy	Relevance (IS $\uparrow$)	Novelty (KLD $\downarrow$)	Plagiarism Risk (FDpann $\downarrow$)
Baseline (No Mixup)	1.81	3.80	2.40
ASM	1.79	3.40	2.31
MTM	1.83	3.32	2.28
ASM + MTM	1.84	3.20	2.15

To evaluate the impact of ASM and MTM strategies on stylistic and temporal features of the generated music, we use two additional metrics: style diversity and temporal consistency. Style diversity is computed based on feature embeddings extracted from generated audio, reflecting the variety in genres, instrumentation, and emotional tones. Higher diversity indicates that the model can generate more varied and creative music. Temporal consistency, on the other hand, measures the coherence of rhythmic and structural patterns across short and long timescales, which ensures that the generated music feels cohesive and well-organized. Together, these metrics provide a comprehensive view of how well the MDN model generates diverse yet rhythmically consistent music. Table 3 demonstrates the improvements in style diversity and temporal consistency with ASM and MTM strategies. ASM focuses on stylistic variation, while MTM ensures rhythmic coherence.

Table 3. Style diversity and temporal consistency analysis for MDN variants.

Model	Style Diversity ($\uparrow$)	Temporal Consistency ($\uparrow$)
MDN (No Mixup)	0.72	0.64
ASM	0.85	0.67
MTM	0.78	0.85
ASM + MTM	0.92	0.89

In Figure 6, a comparison of spectrograms generated with different strategies shows that music generated without using the mixing strategies exhibits higher similarity to the training data. In contrast, the ASM and MTM strategies produce music with more stylistic and rhythmic variations. Combining ASM and MTM further enhances originality and complexity, effectively avoiding excessive similarity to the training data and mitigating the risk of plagiarism.

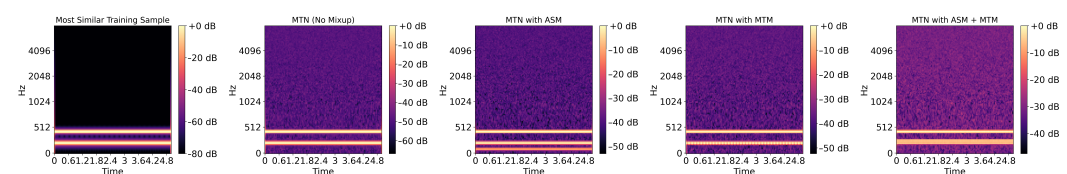


Figure 6. Comparison of spectrograms for MDN with different mixup strategies and the most similar training sample.

4.2.2. Ablation on Baseline Models

To evaluate whether the performance of the proposed framework is inherently dependent on the WaveNet decoder, we conduct a decoder replacement ablation study by strictly fixing all components of the generation pipeline except the waveform decoder. Specifically, the diffusion model produces the same continuous audio representation y_0 at the final denoising step, while the module mapping y_0 to the time-domain waveform is replaced by different decoders, including a WaveNet decoder, a GAN-based vocoder, and neural audio codec decoders (EnCodec and DAC). For codec-based decoders, a lightweight adapter is introduced to map y_0 to the corresponding codec latent space, with the decoder parameters frozen to isolate the effect of the decoder itself. Quantitative results in Table 4 show that the WaveNet decoder achieves the best Fréchet distance and KLD scores, indicating more faithful reconstruction of temporal structure, but incurs the highest inference cost. In contrast, GAN-based and codec-based decoders substantially improve inference efficiency while maintaining competitive generation quality, demonstrating that the proposed diffusion-based framework is not intrinsically tied to a WaveNet decoder and can be effectively paired with modern, efficient audio decoders used in recent large-scale music generation systems.

Table 4. Decoder replacement ablation study. All models share the same diffusion backbone and training configuration; only the waveform decoder is replaced. Bold values indicate the best results.

Decoder	FD _{VGG} ↓	FD _{PANN} ↓	IS ↑	KLD ↓	RTF ↓	Params (M)
WaveNet (D-WN)	18.7	15.9	7.42	0.082	1.34	32.1
GAN Vocoder (D-GAN)	19.8	16.7	7.61	0.094	0.21	13.5
EnCodec (D-ENC)	21.2	18.4	7.08	0.118	0.09	9.7
DAC (D-DAC)	20.6	17.9	7.15	0.110	0.06	11.2

4.2.3. Ablation on Beta Distribution Parameters

Both ASM and MTM sample the mixing coefficient $\lambda \sim \text{Beta}(\alpha, \beta)$, where (α, β) controls the interpolation strength. To verify the impact of these parameters, we conduct an ablation study by varying (α, β) while keeping all other training settings unchanged. As shown in Table 5, the symmetric configuration (1, 1) achieves the best overall performance, obtaining the lowest FD_{vgg} (22.81) and FD_{pann} (2.15), as well as the best IS (1.84) and KLD (3.20). A slightly more concentrated symmetric distribution (2, 2) yields comparable but slightly inferior results, suggesting that moderate interpolation is beneficial while overly constrained mixing may reduce diversity. When α and β are smaller than 1, e.g., (0.5, 0.5), performance degrades across all metrics, indicating that overly extreme interpolation ratios are less effective. Asymmetric settings such as (2, 0.5) and (0.5, 2) also lead to consistently worse results, implying that balanced mixing between paired samples is important for stable training and improved generalization. Based on these observations, we adopt $(\alpha, \beta) = (1, 1)$ in all experiments unless otherwise specified.

Table 5. Ablation results on Beta distribution parameters. Moderate symmetric settings achieve the best trade-off between generation quality and diversity. Bold values indicate the best results.

(α, β)	FD _{vgg} ↓	FD _{pann} ↓	IS ↑	KLD ↓
(0.5, 0.5)	23.47	2.32	1.79	3.45
(1, 1)	22.81	2.15	1.84	3.20
(2, 2)	22.95	2.18	1.83	3.24
(2, 0.5)	23.18	2.27	1.81	3.33
(0.5, 2)	23.26	2.29	1.80	3.38

5. Conclusions

This paper proposes a novel text-to-music generation framework, MDN, which integrates Stable Diffusion and WaveNet architectures and introduces two diffusion-aware mixing mechanisms, ASM and MTM. These strategies effectively address key challenges in text-to-music generation under small-scale datasets, including homogenization, limited diversity, and insufficient originality. Across multiple objective and subjective evaluation metrics, MDN demonstrates superior performance compared with existing methods, particularly in reducing similarity between generated music and training samples, thereby mitigating overfitting and the risk of generative plagiarism. Experimental results show that ASM enhances stylistic diversity and semantic coverage, while MTM improves the modeling of musical structure, leading to stronger rhythmic coherence and temporal organization. Quantitatively, ASM reduces FDvgg from 26.67 to 24.95, MTM further improves performance (FDvgg 23.54, IS 1.83), and the combined ASM + MTM achieves the best overall results (FDvgg 22.81, FDpann 2.15, IS 1.84, KLD 3.20). The integration of Beta-distribution-based interpolation, multi-scale temporal modeling, and WaveNet-based temporal encoding further enables a balanced trade-off between rhythmic continuity, structural integrity, and stylistic control.

Beyond its empirical performance, this work highlights the importance of explicitly modeling style diversity and multi-scale temporal structure within conditional diffusion frameworks for music generation. Future work will explore extending MDN to larger and more diverse text–music datasets with richer annotations, developing more comprehensive evaluation protocols that combine objective metrics with human listening studies, and investigating lightweight decoding strategies to support real-time or low-latency music generation. Incorporating higher-level musical controls, such as harmony progression and instrumentation constraints, also represents a promising direction for improving controllability and practical applicability.

Despite its effectiveness, the proposed framework has several limitations. First, the performance of MDN remains dependent on the quality and coverage of available text–music paired data, and rare or fine-grained musical concepts may not be fully captured under limited data conditions. Second, while distribution-based evaluation metrics provide scalable and reproducible assessment, they cannot fully reflect human perceptual judgments, particularly for long-term musical structure and emotional expressiveness. Finally, the computational overhead introduced by diffusion-based generation and WaveNet decoding currently limits real-time deployment, motivating further research into more efficient model architectures and inference strategies.

Author Contributions: L.X. and J.C. jointly conceived and designed the study. L.X. and J.L. drafted the original manuscript and coordinated the revision process. J.C. and C.L. performed the experiments and conducted data analysis. C.L. assisted with data processing and figure preparation. J.L. contributed to the interpretation of the results and provided critical revisions of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by General research projects of Zhejiang Provincial Department of Education Y20257906 and the APC was funded by C.L.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to their large volume and storage limitations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Jiang, R.; Mou, X. The Analysis of Multi-Track Music Generation With Deep Learning Models in Music Production Process. *IEEE Access* **2024**, *12*, 110322–110330. [\[CrossRef\]](#)
- Briot, J.P.; Hadjeres, G.; Pachet, F.D. *Deep Learning Techniques for Music Generation*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 1.
- Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.A.; Li, S.Z. A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 2814–2830. [\[CrossRef\]](#)
- Xing, Z.; Feng, Q.; Chen, H.; Dai, Q.; Hu, H.; Xu, H.; Wu, Z.; Jiang, Y.G. A survey on video diffusion models. *ACM Comput. Surv.* **2024**, *57*, 1–42. [\[CrossRef\]](#)
- Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
- Huang, Q.; Park, D.S.; Wang, T.; Denk, T.I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv* **2023**, arXiv:2302.03917.
- Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; Goldstein, T. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 6048–6058.
- Will, J. Rage against the machine: Copyright infringement in AI-generated music. *J. Intell. Prop. L.* **2024**, *31*, 378.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 10684–10695.
- Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499. [\[CrossRef\]](#)
- Scirea, M.; Barros, G.A.; Shaker, N.; Togelius, J. SMUG: Scientific Music Generator. In *Proceedings of the ICCG, Park City, UT, USA, 29 June–2 July 2015*; pp. 204–211.
- Van Der Merwe, A.; Schulze, W. Music generation with markov models. *IEEE Multimed.* **2010**, *18*, 78–85. [\[CrossRef\]](#)
- Cope, D. Experiments in musical intelligence (EMI): Non-linear linguistic-based composition. *J. New Music. Res.* **1989**, *18*, 117–139. [\[CrossRef\]](#)
- Wu, J.; Hu, C.; Wang, Y.; Hu, X.; Zhu, J. A hierarchical recurrent neural network for symbolic melody generation. *IEEE Trans. Cybern.* **2019**, *50*, 2749–2757. [\[CrossRef\]](#) [\[PubMed\]](#)
- Choi, K.; Fazekas, G.; Sandler, M. Text-based LSTM networks for automatic music composition. *arXiv* **2016**, arXiv:1604.05358. [\[CrossRef\]](#)
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [\[CrossRef\]](#)
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Volume 30.
- Schneider, F.; Kamal, O.; Jin, Z.; Schölkopf, B. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024*; pp. 8050–8068.
- Liu, H.; Huang, R.; Liu, Y.; Cao, H.; Wang, J.; Cheng, X.; Zheng, S.; Zhao, Z. AudioLCM: Text-to-Audio Generation with Latent Consistency Models. *arXiv* **2024**, arXiv:2406.00356.
- Lanzendörfer, L.A.; Lu, T.; Perraudin, N.; Herremans, D.; Wattenhofer, R. Coarse-to-Fine Text-to-Music Latent Diffusion. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
- Kim, S.; Kim, G.; Yagishita, S.; Han, D.; Im, J.; Sung, Y. Enhancing Diffusion-Based Music Generation Performance with LoRA. *Appl. Sci.* **2025**, *15*, 8646. [\[CrossRef\]](#)
- McDonald, J.B.; Xu, Y.J. A generalization of the beta distribution with applications. *J. Econom.* **1995**, *66*, 133–152. [\[CrossRef\]](#)
- Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Driessche, G.; Lockhart, E.; Cobo, L.; Stimmer, F.; et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; PMLR: Cambridge, MA, USA, 2018; pp. 3918–3926.
- Song, K.; Zhang, Y.; Lei, Y.; Cong, J.; Li, H.; Xie, L.; He, G.; Bai, J. DSPGAN: A gan-based universal vocoder for high-fidelity tts by time-frequency domain supervision from dsp. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.

25. Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *30*, 495–507. [[CrossRef](#)]
26. Tzanetakis, G.; Cook, P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 293–302. [[CrossRef](#)]
27. Choi, K.; Fazekas, G.; Sandler, M.; Cho, K. Convolutional recurrent neural networks for music classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2392–2396.
28. Chen, K.; Wu, Y.; Liu, H.; Nezhurina, M.; Berg-Kirkpatrick, T.; Dubnov, S. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1206–1210.
29. Kleijn, W.B.; Lim, F.S.; Luebs, A.; Skoglund, J.; Stimberg, F.; Wang, Q.; Walters, T.C. Wavenet based low rate speech coding. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 676–680.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.