T-WGAN: A Transformer-Wasserstein GAN Approach for Melody Generation with Structural and Rhythmic Fidelity

Chunxiao Zhao
Zhengzhou Health College, Ministry of Health and Humanities Education (Art Education Center), Zhengzhou 450064, China
E-mail: zhaozhaoran1215@163.com

Keywords: music melody generation, deep neural network, transformer, generative adversarial network, rhythm consistency

Received: September 10, 2025

To address the current challenges of deep learning music generation models in capturing long-range dependencies, ensuring generation diversity, and maintaining training stability, this study proposes an optimized music melody generation model—T-WGAN. The model deeply integrates Transformer and Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). This study first preprocesses the large-scale Lakh MIDI dataset, extracts single-track main melodies, and converts them into symbolic sequences using a REMI-based event representation. On this basis, the model innovatively adopts a generator based on Transformer decoder to learn the long-range structure of melodies. It also uses a critic based on Transformer encoder for stable adversarial training under the WGAN-GP framework to enhance the diversity and authenticity of generated melodies. Experimental results show that T-WGAN performs excellently on multiple key evaluation metrics. T-WGAN achieves a Rhythmic Consistency Rate (RCR) of 85.17%, significantly higher than baseline models (e.g., Transformer's 75.68%). Its score on Fréchet Distance for Music (FDM) drops to 31.02, proving that the generated melodies are closer to real music in feature distribution. The conclusion indicates that the proposed T-WGAN model successfully addresses the three core issues in melody generation—structural integrity, diversity, and training stability—synergistically. The findings provide an effective technical approach for generating high-quality music melodies with both structural logic and innovation.

Povzetek: Študija predstavlja model T-WGAN, ki izboljša generiranje glasbenih melodij z večjo raznolikostjo, boljšo strukturo in stabilnejšim učenjem.

1 Introduction

With the rapid development of artificial intelligence (AI) technology, its applications have expanded from traditional computing and analysis tasks to the field of artistic creation, which is a domain previously considered unique to human creativity. Among these applications, automatic music generation, as a typical representative of the intersection between AI and art, aims to create musical works with logic, aesthetic value, and emotional expressiveness through algorithms and models. It provides musicians with powerful auxiliary creation tools and opens new paradigms for content production in industries such as games, film and television scoring, and digital entertainment [1, 2]. However, music is not a simple accumulation of musical notes, but a time-series art that contains complex hierarchical structures, long-range dependencies, and subtle emotional connotations. Therefore, how to use deep neural network (DNN) to generate melodies with long-duration music with structural integrity, logical consistency, and originality has become a core scientific problem urgently needing to be solved in this field.

Amid the wave of deep learning, music generation technology has undergone evolution, from Recurrent Neural Network (RNN) and their variant, Long Short-Term Memory (LSTM) network, to various advanced generation models [3]. Early RNN/LSTM models, leveraging their ability to process sequential data, made significant progress in learning the local continuity of melodies. However, its inherent vanishing gradient problem and limitations in modeling long-range dependencies hinder its performance in generating structurally coherent long-duration music. To improve generation quality, researchers introduced models such as Generative Adversarial Network (GAN) and Variational Autoencoder (VAE). These models have shown great potential in enhancing the realism and diversity of generated samples [4]. Nevertheless, the training process of standard GAN is extremely unstable and often falls into the dilemma of mode collapse, resulting in monotonous and tedious generated melodies. In addition, although self-attention models represented by

Transformer can excellently capture long-range dependencies, when used independently for generation tasks, they still face challenges in training stability and generated sample diversity, making it difficult to effectively avoid the monotony of generated content. Therefore, how to integrate the advantages of different models and avoid their respective shortcomings constitutes the starting point of this study.

To address the above challenges, this study proposes an optimized music melody generation model based on the deep fusion of Transformer and Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP), which is called T-WGAN. The core innovation of this model lies in combining the strong global context capture capability of Transformer with the stable and high-quality generation training framework of WGAN-GP. This study aims to verify the following core hypotheses through experiments:

1) To evaluate the role of integrating Transformer as a generator in enhancing the coherence and accuracy of melodies in terms of long-range structure and rhythmic patterns.

2) To verify the effectiveness of the WGAN-GP training framework in overcoming the training instability of traditional GANs, avoiding mode collapse, and enhancing melody diversity, especially through comparison with the standalone Transformer model.

This study will conduct performance evaluation based on the large-scale Lakh MIDI dataset. The purposes are to explore an effective path for generating high-quality music melodies with both structural integrity and innovation, and to provide new insights and empirical support for the application of AI in the field of artistic creation.

2 Related work

2.1 Review of the research status of music generation

As reviewed in Wei et al., AI music technology was undergoing a transformation from an auxiliary "tool" to a "creator" with independent creative potential, and its application scenarios had been widely extended to fields such as digital entertainment, art education, and content creation [5]. The current research status presents a trend of high diversification and multimodal fusion. On one hand, research efforts continue to deepen in the sub-fields of music generation. For instance, Li et al. designed a generation method based on deep reinforcement learning for Xi'an Drum Music [6]. On the other hand, researchers are actively exploring the connection between music and other modal information, aiming to achieve "controllable generation" with stronger context awareness. For example, Zheng and Li realized real-time emotion-based piano music generation through GAN [7]. Huang et al. further extended the control conditions to the visual domain, using Creative-GAN to generate artistic music [8]. Kang et al. even utilized a multimodal Transformer model to successfully generate background music that aligns with the atmosphere of video content directly from the video itself [9].

2.2 Review of the research status of music generation based on DNN

DNN is the absolute core of current music generation research. Table 1 shows the research of related scholars.

Although Transformer performs excellently in structural modeling, when used independently for generation tasks, it still faces two key challenges: 1) the problem of training instability, which is likely to cause fluctuations in generation quality; 2) insufficient generation diversity, that is, the model may converge to some common and safe melody patterns, lacking innovation.

Table 1: Summary of the characteristics and limitations of music generation technology route based on DNN.

Technical route	Core ideas	Advantages	Main limitation	Representative literature
General framework	Sort out the representation, algorithms, evaluation methods, and challenges of symbolic music generation.	Provide a comprehensive theoretical framework and methodology in the field.	-	Ji et al. [10]
Adversarial network	Learn data distribution through the adversarial game between the generator and the discriminator.	Generate samples with high authenticity and the ability to learn complex distributions.	Unstable training, prone to mode collapse and poor diversity.	Liu [11], Tanawala and Dalwadi [12]
Attention mechanism	Use the self-attention mechanism to directly calculate the long-range	Effectively capture global structures and long-range dependencies.	Unstable training during independent generation, with a	Zhang et al. [13], Li and Sung [14], Wu et al. [15]

	relationships between elements in the sequence.		tendency to produce repetitive content.	
Multi-technology fusion	Combine deep learning with other technologies (e.g., genetic algorithms).	Enrich the technical paths of music generation and potentially solve specific problems.	Increased model complexity and training difficulty.	Majidi and Toroghi [16]

2.3 Research gap and innovation of this study

Through a review of existing DNN-based music generation technical routes, it is evident that although various models have achieved significant progress in specific aspects, the current research field lacks an integrated solution that can synergistically address the three core challenges: long-range structural dependencies, generation diversity, and training stability. The proposed T-WGAN model aims to fill this critical gap. Its core innovation lies in the strong integration of Transformer into the stable and efficient adversarial training framework of WGAN-GP, with the expectation of providing a reliable reference direction for the generation of high-quality music melodies.

3 The design and evaluation method of melody generation model

3.1 The overall framework design analysis of melody generation model

To effectively address the key issues in music melody generation, including insufficient modeling of long-range dependencies, lack of diversity in generated melodies, and instability in adversarial training, this study proposes the T-WGAN model, which integrates Transformer with the WGAN-GP mechanism [17]. The overall framework of the T-WGAN model is shown in Fig. 1.

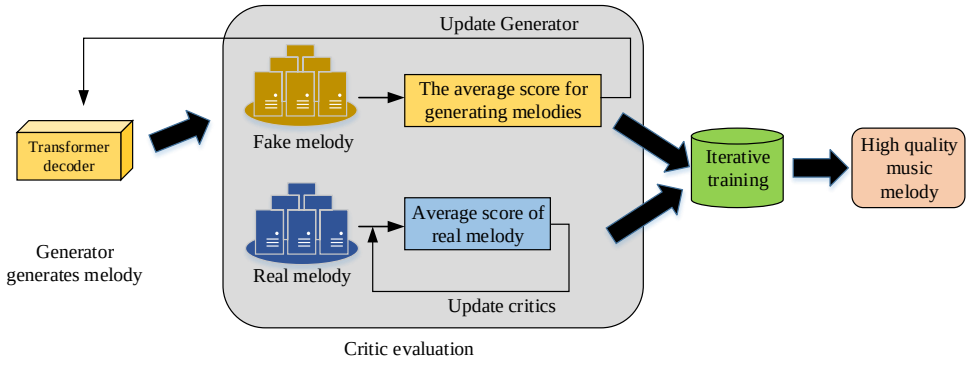

Figure 1: Schematic diagram of melody generation model architecture based on T-WGAN.

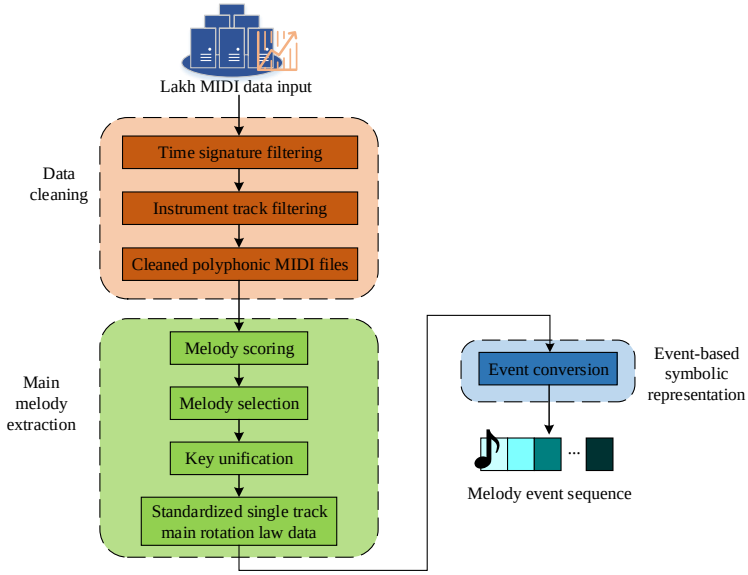

Figure 2: Schematic diagram of pretreatment process of training data.

3.2 Training data preprocessing and symbolic representation

This study adopts The Lakh MIDI Dataset v0.1 (LMD) (https://colinraffel.com/projects/lmd/) as the original data source for the melody generation task. This section focuses on a three-step process of "data cleaning-melody extraction-symbolic representation" (Fig. 2), and finally obtains high-quality single-melody sequence data suitable for T-WGAN model training.

In the data cleaning stage, first, music pieces are filtered based on time signature information, retaining only samples with the common 4/4-time signature to avoid rhythm alignment difficulties caused by complex time signatures. Then, MIDI tracks are classified by instrument, only instruments with strong melodic properties (such as piano, flute, and violin) are retained, while percussion instruments and chord filling tracks are filtered out. Next, the "main melody extraction algorithm" is used to select the track with the strongest melodic feature. Specifically, a heuristic scoring function based on pitch range and track activity is adopted to score all non-percussion tracks of each piece of music. The track with the highest score is selected as the main melody track, as shown in equation (1):

$$S_i = \alpha \cdot \text{PitchRange}(i) + \beta \cdot \text{NoteDensity}(i) \quad (1)$$

S_i denotes the melody score of the i-th track. $\text{PitchRange}(i)$ is the pitch range of this track. NoteDensity(i) is the number of notes per unit time. The empirical weights α and β are set to 0.6 and 0.4 respectively, aiming to balance the preference for melodies with wider melodic contours and richer rhythmic activities. In addition, to eliminate the burden on the model caused by different keys, all melodies are transposed to C major or a minor. The original key is identified using a key estimation tool (such as the key.estimate()) function provided by the music21 library) and converted through a transposition function to ensure the consistency and generalization ability of the input space.

To enhance the model's ability to understand music structure, this study adopts an event-based symbolic method to encode melody data, replacing the fixed-time grid representation of traditional Piano Roll. For example, a note event representing "playing a note with Pitch 60, Duration of 1/8 beat, and Velocity 80 at the 3rd Position of the 2nd Bar" can be encoded into the event sequence shown in equation (2):

$$\text{Bar}_2 \rightarrow \text{Position}_3 \rightarrow \text{Pitch}_{60} \rightarrow \text{Duration}_{1/8} \rightarrow \text{Velocity}_{80} \quad (2)$$

This encoding method not only preserves the structural information of notes but also effectively models the interactive relationships across dimensions such as rhythm, dynamics, and time, adapting to the Transformer structure's requirement for discrete sequence modeling. For example, the event dictionary maps MIDI pitches to tokens from Pitch_21 to Pitch_108, and quantized positions based on sixteenth notes to tokens from Pos_0 to Pos_15. By analogy, a vocabulary containing approximately 300 independent events is constructed. For rare events not covered by the dictionary (such as irregular time signatures or microtones), the preprocessing pipeline filters them out to maintain the consistency of the symbolic space. Overlapping notes (polyphony) appearing in the main melody track are processed by prioritizing the highest pitch at any time step, thereby ensuring the generation of monophonic melody sequences suitable for the focus of this study.

3.3 Optimization strategy analysis of melody generation model based on T-WGAN

The proposed T-WGAN model deeply integrates the Transformer architecture [18] with the optimization mechanism of WGAN-GP, constructing a music generation network capable of capturing melodic structures, ensuring training stability, and enhancing generation diversity. Its main components include a generator built based on the Transformer decoder and a critic built based on the Transformer encoder. These two components form a typical adversarial learning architecture as shown in Figure 3. The specific architectural details of the generator and the critic are presented in Table 2. Both the generator and the critic adopt standard positional encoding to inject sequence order information, and apply layer normalization within each Transformer block, which is consistent with the original Transformer architecture.

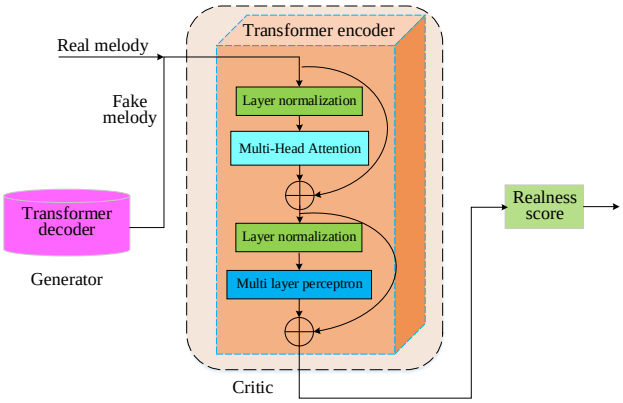

Figure 3: Schematic diagram of antagonistic learning architecture of melody generation model based on T-WGAN.

Table 2: Architecture detail table of T-WGAN model components

Parameter	Generator (Transformer decoder)	Critics (Transformer encoder)
Embedding dimension	256	256
Number of layers	6	6
Attention heads	8	8
Feed-forward dim	1024	1024
Activation function	ReLU	LeakyReLU
Dropout rate	0.1	0.1

In this architecture, the task of the generator is to learn the mapping function G(z) from the random noise z∼N(0,I) in the potential space to the melody event space, and gradually generate the melody sequence by autoregressive method, as shown in equation (3):

$$x = G(z) = (x_1, x_2, \ldots, x_T) \tag{3}$$

$x_t \in V$ represents the event generated at the t time step (such as Pitch64, Duration1/8, etc.), and V is the event dictionary. Specifically, the generator predicts that the event probability distribution of each time step is equation (4):

$$P(x_t | x_{<t}, z) = Softmax(f_\theta(x_{<t}, z)) \tag{4}$$

f_θ represents the output function modeled by the Transformer decoder, and the parameter is θ. The generation process of the whole sequence is to maximize conditional likelihood, as shown in equation (5):

$$L_{MLE} = \sum_{t=1}^{T} logP(x_t | x_{<t}, z) \tag{5}$$

This autoregressive modeling method relies on Transformer's Multi-Head Attention mechanism, which can effectively capture the long-term dependence in melody structure.

In confrontation training, the Critic's goal is to learn a real function $D(x) \in R$, which is used to measure whether the input melody is close to the real data. Different from the traditional binary classifier in GAN, WGAN introduces the Wasserstein-1 distance between the real sample distribution P_r and the generated sample distribution P_g as the optimization objective, which is defined as equation (6):

$$W(P_r, P_g) = \sup_{\|f\|_L \le 1} E_{x \sim P_r}[f(x)] - E_{\tilde{x} \sim P_g}[f(\tilde{x})] \tag{6}$$

f is a 1-Lipschitz continuous function, and Critic D(x)≈f(x). The objective is to maximize this expected difference, such that real samples receive higher scores and generated samples receive lower scores. To ensure the Critic satisfies the 1-Lipschitz constraint, WGAN-GP replaces the previous weight clipping method by introducing a GP term. The final objective function is shown in equation (7):

$$L_D = E_{\tilde{x} \sim P_g}[D(\tilde{x})] - E_{x \sim P_r}[D(x)] + \lambda E_{\hat{x} \sim P_{\hat{x}}}\left(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1\right)^2 \tag{7}$$

λ refers to the GP coefficient, which is taken as λ = 10 in this study. $\hat{x}$ is an intermediate sample obtained by linear interpolation between the real sample $x \sim P_r$ and the generated sample $\tilde{x} \sim P_g$, as shown in equation (8):

$$\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}, \quad \varepsilon \sim U[0,1] \tag{8}$$

The goal of the Critic's update is to minimize L_G, while the optimization goal of the generator is to maximize the Critic's score on the generated samples, as shown in equation (9):

$$L_G = -E_{\tilde{x} \sim P_g}[D(\tilde{x})] \tag{9}$$

In this way, and generators constitute a "zero-sum game" optimization process, which is expressed as $\min_G \max_D L_D$. During the adversarial training process, this study adopts an update ratio of 5:1 between the Critic and the Generator—i.e., the Critic is updated five times for each update of the Generator. While the optimal update ratio may vary depending on specific tasks and model architectures, a low ratio (e.g., 1:1) may lead to insufficient training of the Critic. This results in unstable gradient directions provided by the Critic, thereby causing training collapse. An excessively high ratio, however, will unnecessarily increase computational costs. Therefore, the adoption of the fully validated 5:1 ratio is a reliable compromise between ensuring gradient quality and maintaining reasonable training efficiency. It is particularly suitable for symbolic sequence generation tasks that require refined gradient guidance, effectively avoiding mode collapse and promoting diversity.

Algorithm 1: T-WGAN Training Procedure

Require: Generator G with parameters θg, Critic C with parameters θc, real data distribution Pr, noise dimension dz, batch size m, Critic update iterations n_{critic}, GP coefficient λ.

Require: Adam optimizers opt$_g$, opt$_c$.

while not converged **do**

  for t=1,…,n_{critic} **do**

    Sample a batch of real melodies $\{x^{(i)}\}_{i=1}^{m} \sim Pr$

    Sample a batch of noise vectors $\{z^{(i)}\}_{i=1}^{m} \sim N(0, I)$

    Generate a batch of fake melodies

    Sample $\varepsilon \sim U[0,1]$

    Create interpolated samples $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}$

    $C_{\text{real}} = C(x)$

    $C_{\text{fake}} = C(\tilde{x})$

    // Calculate Gradient Penalty

    $g = \nabla_{\hat{x}} C(\hat{x})$

    $L_{GP} = \lambda \cdot (\| g \|_2 - 1)^2$

    // Critic Loss

    $L_C = E[C_{fake}] - E[C_{real}] + E[L_{GP}]$

    // Update Critic

    Zero gradients of opt$_c$.

    L_C.backward()

    opt$_c$.step()

  end for

  // Update Generator

  Sample a batch of noise vectors $\{z^{(i)}\}_{i=1}^{m} \sim N(0, I)$ ·

28: Generate a batch of fake melodies $\tilde{x} = G(z)$
29:
30: // Generator Loss
31: $L_G = -E[C(\tilde{x})]$
32:
33: // Update Generator
34: Zero gradients of opt_g
35: L_G .backward()
36: opt_g. step()
37: **end while**

Among them, the pseudocode flow of the melody generation model based on T-WGAN constructed in this study is shown in algorithm 1.

3.4 Experimental evaluation

In this study, the Lakh MIDI Dataset v0.1 is selected as the experimental dataset, which contains more than 170,000 music works in MIDI format, covering various styles such as pop, jazz, and classical. To ensure training quality, the data is cleaned and filtered, mainly including: 1) Removing samples with a duration of less than 30 seconds or fewer than 200 events. 2) Standardizing the tempo of all MIDI tracks (unified to 120 BPM). 3) Retaining only the main melody track and extracting its pitch, duration, and velocity as model inputs. Finally, approximately 38,000 high-quality single-track melody segments are organized for model training and validation, with the ratio of training set, validation set, and test set being 8:1:1.

The specific software and hardware configurations as well as hyperparameter settings are presented in Table 3. These parameter settings combine the common configurations of Transformer models in the field of music generation and have been fine-tuned through preliminary experiments to balance model performance and computational efficiency on the Lakh MIDI dataset. For example, 6 network layers and 8 attention heads represent a compromise between capturing long-range dependencies and preventing overfitting. In the symbolic representation, the MIDI timeline is quantized to a resolution of sixteenth notes. Each discrete step in the generated sequence, referred to as a 'Time Step', corresponds to such a time quantum. Therefore, in a standard 4/4-time piece, one measure consists of 16-time steps.

In the model evaluation phase, the proposed T-WGAN model algorithm is compared with Transformer, MuseGAN [19], MusicVAE [20], Pop Music Accompaniment Generator (PopMAG) [21], and the model algorithm proposed by Wu et al. (2024) in related fields. The following metrics are used to measure performance: Note Repetition Rate (NRR), Mean Pitch Error (MPE), Rhythmic Consistency Rate (RCR), and Fréchet Distance for Music (FDM).

These metrics are selected in this study to comprehensively evaluate the quality of generated melodies from multiple dimensions: NRR is used to measure melody diversity and avoid monotony and repetition. MPE calculates the mean absolute error between the pitch predicted by the model and the real pitch (both represented by MIDI pitch numbers) at each time step. RCR examines the coherence of rhythmic structures. FDM evaluates similarity by first encoding generated and real melody segments into a pre-trained music feature embedding space (features are usually extracted using audio or symbolic music models such as VGGish), then calculating the Fréchet distance between these two sets of feature distributions. These metrics together form a multi-dimensional evaluation system.

Table 3: Software and hardware configuration and hyperparameter setting.

Category	Parameter term	Setting value/description
Data preprocessing	Minimum number of events	≥ 200
	Minimum duration	≥ 30 seconds
	Rhythm unification	120 BPM
	Input feature	Pitch, Duration, Velocity
Software environment	Framework	PyTorch 2.1
	Operating system	Ubuntu 22.04
Hardware environment	GPU	NVIDIA A100 80GB
	CPU	Intel Xeon Platinum
Hyperparameter	Learning rate	0.0005
	Batch Size	64
	Maximum training rounds	100
	Optimizer	Adam
	Loss function	Cross-Entropy
	Stop strategy	Early Stopping (Verification set 5 rounds without improvement termination)

4 Result and discussion

4.1 Pitch and diversity analysis under different algorithms

Comparing this model algorithm with other algorithms, the results of pitch and diversity generated by music melody are shown in Fig. 4 and Fig. 5.

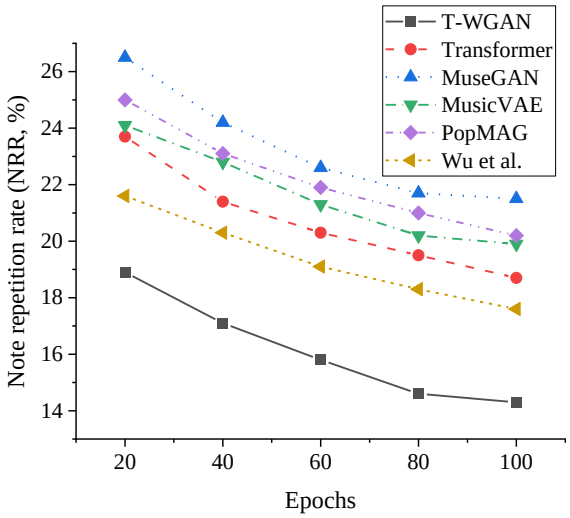

Figure 4: Results of note repetition rate generated by music melody under different algorithms.

In Fig. 4, throughout different training epochs, the proposed T-WGAN model consistently outperforms all comparative models in the NRR metric, demonstrating a stronger ability to generate melodic diversity. In the early stage of training (20 epochs), the NRR of T-WGAN is 18.9%, which is already significantly lower than that of Transformer (23.7%) and MuseGAN (26.5%). As training progresses, the NRR of T-WGAN continues to decrease steadily to 14.3%, outperforming the second-best model proposed by Wu et al. (17.6%). This advantage can be attributed to the introduction of the WGAN-GP training framework. By optimizing the Wasserstein distance, this framework effectively alleviates the mode collapse problem, prompting the generator to explore a broader melodic space. This avoids the monotonous and repetitive patterns that standalone Transformer models tend to fall into due to error accumulation in autoregressive generation.

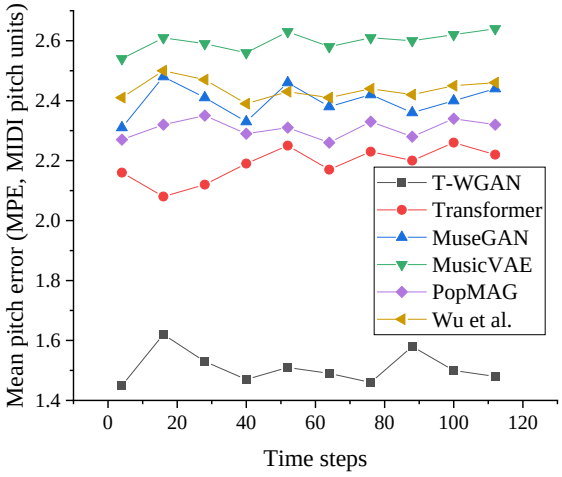

Figure 5: Results of average pitch error generated by music melody under different algorithms.

Meanwhile, data in Fig. 5 shows that the T-WGAN model maintains a significantly lower MPE than comparative models across all time steps, demonstrating excellent accuracy in melodic pitch prediction. In short-range contexts (e.g., Time step=4), the MPE of T-WGAN is 1.45, far lower than that of Transformer (2.16), MuseGAN (2.31), and MusicVAE (2.54). Even in long-range contexts (e.g., Time step=112), the error value of T-WGAN remains at a low level of 1.48, while the errors of other models are generally higher than 2.3. This is mainly due to the Transformer architecture serving as the generator backbone. Its self-attention mechanism can effectively capture long-range dependencies in music sequences, enabling the model to consider broader contextual information when predicting the current pitch, thus ensuring the coherence of the melody in global structure and the accuracy of local pitches.

4.2 Evaluation and analysis of music melody generation effect under different algorithms

The proposed model algorithm is compared with other algorithms, and the results of rhythm consistency and distribution similarity of music melody generation are further evaluated, as shown in Figs. 6 and 7.

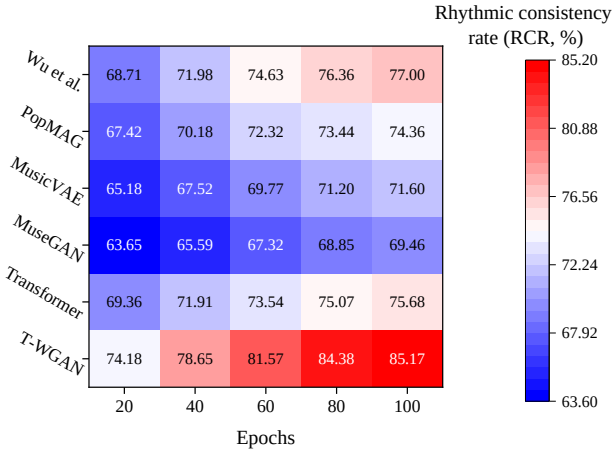

Figure 6: Results of rhythm retention rate of music melody generation under different algorithms.

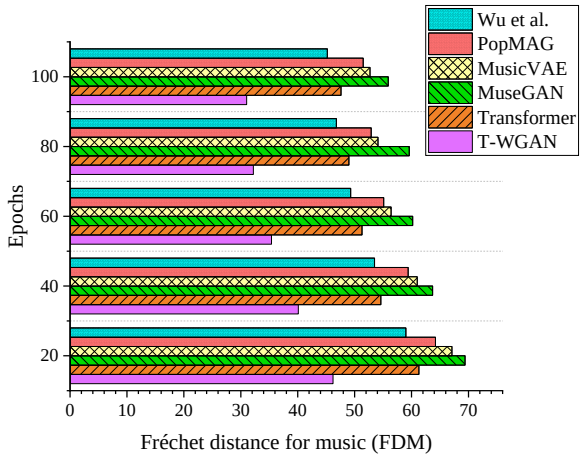

Figure 7: Results of generation fragment similarity of music melody generation under different algorithms.

In Figs. 6 and 7, the T-WGAN model demonstrates significant advantages in two core metrics: rhythmic consistency and musical fidelity. Regarding RCR, T-WGAN maintains a leading position across all training epochs. Its RCR steadily increases from 74.18% at the 20th epoch to 85.17% at the 100th epoch, significantly outperforming Transformer (75.68%), MuseGAN (69.46%), and the method proposed by Wu et al. (77.00%). As for the FDM metric, it comprehensively evaluates musical fidelity by measuring the distribution distance between generated samples and real samples in the feature space. The FDM score of T-WGAN steadily decreases from an initial 46.20 to 31.02, which is significantly better than all comparative models, such as Transformer (47.60) and MuseGAN (55.90). The substantial reduction in the FDM score holds important musical significance: it indicates that the melodies generated by T-WGAN are closer to real music in the joint distribution of multi-dimensional features, including pitch, rhythm, and dynamics.

To further verify the reliability of the performance improvement of the T-WGAN model, paired t-tests are conducted between its final results (after 100 training epochs) on the four core evaluation metrics and those of all baseline models. The test results are presented in Table 4, with all p-values evaluated at a significance level of $\alpha = 0.05$. The results indicate that the superiority of T-WGAN is not merely due to random fluctuations or accidental outcomes but is statistically significant. This finding strongly supports the conclusion of this study: the T-WGAN model significantly outperforms existing methods in overall performance.

Table 4: Paired t-test results of T-WGAN and baseline model on each index.

Contrast model	NRR (%)	MPE	RCR (%)	FDM
Transformer	0.002	< 0.001	0.001	< 0.001
MuseGAN	< 0.001	< 0.001	< 0.001	< 0.001
MusicVAE	< 0.001	< 0.001	< 0.001	< 0.001
PopMAG (MuseFlow)	0.003	< 0.001	< 0.001	< 0.001
Wu et al. [15]	0.009	< 0.001	< 0.001	< 0.001

4.3 Ablation experiment

To systematically evaluate the individual contributions of the Transformer architecture and the WGAN-GP training framework in the T-WGAN model, the following ablation study is designed. Quantitative analysis is conducted on the expected performance of each variant model, as detailed in Table 5.

Table 5: Variant model of ablation research and numerical analysis of expected performance.

Model variant	NRR (%)	MPE	RCR (%)	FDM
1. Transformer-only	23.7	2.16	75.68	47.6
2. Transformer + vanilla GAN	~20.0 (unstable)	> 3.0 (bad)	~65.0 (bad)	> 60.0 (bad)
3. WGAN-GP + RNN Generator	15.5	2.55	68.2	50.5
4. T-WGAN (the proposed model)	14.3	1.48	85.17	31.02

In Table 5, the ablation analysis clearly demonstrates the complementarity and necessity of each component in the T-WGAN model. The Transformer-only baseline performs well in the structural metric (RCR) but underperforms in diversity (NRR) and fidelity (FDM). The introduction of a standard GAN (Transformer + vanilla GAN) may lead to an overall deterioration of all performance metrics due to extremely unstable training. On the other hand, using the WGAN-GP framework alone with a weaker RNN generator (WGAN-GP + RNN Generator) can ensure generation diversity (low NRR) but sacrifices the ability to capture long-range musical structures (low RCR). Therefore, only the combination of Transformer and WGAN-GP can achieve synergistic improvements across all key dimensions, thereby generating high-quality melodies that possess both structural integrity and content novelty.

4.4 Discussion

The experimental results of this study indicate that the success of T-WGAN stems from the synergistic advantages of its architecture. As the generator backbone, the Transformer ensures the coherence of melodies in long-range structures, which is consistent with the finding of Agarwal and Sultanova [22] that emphasizes the core role of the Transformer. Meanwhile, the WGAN-GP training framework effectively overcomes mode collapse and significantly enhances generation diversity. The high RCR and low MPE are manifestations of the model's mastery of musical "grammar" (i.e., structural rules), as it can generate long sequences with coherent rhythms and accurate pitches. This suggests the model is not merely imitating notes but learning their inherent structural relationships. Compared with the method of improving stability through ensemble learning proposed by Nag et al. [23], this deep integration of "structure awareness" and "stable generation" achieves more efficient performance improvement within a single model. Nevertheless, the model still has limitations. For instance, this study relies entirely on objective metrics and lacks human subjective evaluation of the aesthetic value of generated music. As investigated by Dong [24], future work can explore the use of heuristic methods such as genetic algorithms for post-processing optimization of

generated melodies, and must introduce subjective listening experiments to construct a more comprehensive evaluation system.

5 Conclusion

This study successfully constructs a deeply integrated model based on T-WGAN, which combines the powerful long-range dependency capture capability of Transformer with the stable training framework of WGAN-GP. Experimental results show that T-WGAN significantly outperforms existing baseline models across multiple objective metrics, such as note diversity, pitch accuracy, rhythmic consistency, and generation fidelity. With a pitch error of only 1.45, it demonstrates advancement in generating high-quality, structurally complete, and innovative melodies.

Nevertheless, this study still has clear limitations. First, the model currently focuses on monophonic melody generation and has not been extended to more complex polyphonic music creation such as harmony and orchestration. Second, the evaluation system of this study relies entirely on objective metrics, lacking subjective evaluations by human listeners on the aesthetic value and emotional expression of generated music—this is a key link in measuring the success of music creation. Third, the study fails to provide detailed ablation experiments to accurately quantify the individual contributions of the Transformer and WGAN-GP to the final performance.

Looking forward, future research can be deepened in the following three aspects: 1) Extend the T-WGAN framework to multi-track and polyphonic music generation tasks; 2) Introduce conditional variables such as style and emotion to explore controllable music generation, thereby enhancing the model's practicality and interactivity; 3) Combine subjective listening experiments with objective metrics to construct a more comprehensive evaluation system for music generation quality to further promote the application of AI in the field of music creation.

References

[1] Sadiku M N, Ajayi S A, Sadiku J O. Artificial intelligence in media and entertainment. International Journal of Scientific and Applied Research (IJSAR), 2025, 5(5): 1-4. https://doi.org/10.54756/IJSAR.2025.5.1

[2] Kwiecień J, Skrzyński P, Chmiel W, Dąbrowski A, Szadkowski B, Pluta M. Technical, musical, and legal aspects of an ai-aided algorithmic music production system. Applied Sciences, 2024, 14(9): 3541. https://doi.org/10.3390/app14093541

[3] Kumar J, Goomer R, Singh AK. Long short-term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. Procedia Computer Science, 2018, 125: 676-682. https://doi.org/10.1016/j.procs.2017.12.087

[4] Dash A, Agres K. AI-based affective music generation systems: A review of methods and challenges. ACM Computing Surveys, 2024, 56(11): 287. https://doi.org/10.1145/3672554

[5] Wei L, Yu Y, Qin Y, Zhang S. From tools to creators: A review on the development and application of artificial intelligence music generation. Information, 2025, 16(8): 656. https://doi.org/10.3390/info16080656

[6] Li P, Liang T, Cao Y, Wang X, Wu X, Lei L. A novel Xi'an drum music generation method based on Bi-LSTM deep reinforcement learning. Applied Intelligence, 2024, 54(1): 80-94. https://doi.org/10.1007/s10489-023-05195-y

[7] Zheng L, Li C. Real-time emotion-based piano music generation using generative adversarial network (GAN). IEEE Access, 2024, 12: 87489-87500. https://doi.org/10.1109/ACCESS.2024.3414673

[8] Huang J, Huang X, Yang L, Tao Z. Dance-conditioned artistic music generation by creative-GAN. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2024, E107.A (5): 836-844. https://doi.org/10.1587/transfun.2023EAP1059

[9] Kang J, Poria S, Herremans D. Video 2 music: Suitable music generation from videos using an affective multimodal transformer model. Expert Systems with Applications, 2024, 249: 123640. https://doi.org/10.1016/j.eswa.2024.123640

[10] Ji S, Yang X, Luo J. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. ACM Computing Surveys, 2023, 56(1): 7. https://doi.org/10.1145/3597493

[11] Liu W. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. The Journal of Supercomputing, 2023, 79(6): 6560-6582. https://doi.org/10.1007/s11227-022-04914-5

[12] Tanawala B A, Dalwadi D C. Harmonic synergy: Leveraging deep convolutional networks, LSTMs, and RNNs for multi-genre piano roll generation with GANs. SN Computer Science, 2025, 6(4): 297. https://doi.org/10.1007/s42979-025-03855-z

[13] Zhang Y, Zhou Y, Lv X, Li J, Lu H, Su Y, Yang H. TARREAN: a novel transformer with a gate recurrent unit for stylized music generation. Sensors, 2025, 25(2): 386. https://doi.org/10.3390/s25020386

[14] Li S, Sung Y. MRBERT: pre-training of melody and rhythm for automatic music generation. Mathematics, 2023, 11(4): 798. https://doi.org/10.3390/math11040798

[15] Wu S L, Donahue C, Watanabe S, Bryan N J. Music controlnet: Multiple time-varying controls for music generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2692-2703. https://doi.org/10.1109/TASLP.2024.3399026

[16] Majidi M, Toroghi R M. A combination of multi-objective genetic algorithm and deep learning for

music harmony generation. Multimedia tools and applications, 2023, 82(2): 2419-2435. https://doi.org/10.1007/s11042-022-13329-6

[17] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of Wasserstein Gans. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach California USA, December 4-9, 2017, pp. 5769-5779. https://dl.acm.org/doi/10.5555/3295222.3295327

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach California USA, December 4-9, 2017, pp. 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349

[19] Wang F. Application of artificial intelligence-based music generation technology in popular music production. Journal of Combinatorial Mathematics and Combinatorial Computing, 2025, 127: 655-671. https://doi.org/10.61091/jcmcc127a-038

[20] Tie Y, Guo X, Zhang D, Tie J, Qi L, Lu Y. Hybrid learning module-based transformer for multitrack music generation with music theory. IEEE Transactions on Computational Social Systems, 2024, 12(2), 862-872. https://doi.org/10.1109/TCSS.2024.3486604

[21] Zhu H, Liu Q, Yuan NJ, Zhang K, Zhou G, Chen E. Pop music generation: From melody to multi-style arrangement. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(5): 54. https://doi.org/10.1145/3374915

[22] Agarwal S, Sultanova N. Music generation through transformers. International Journal of Data Science and Advanced Analytics, 2024, 6(6): 302-306. https://doi.org/10.69511/ijdsaa.v6i6.231

[23] Nag B, Middya A I, Roy S. Melody generation based on deep ensemble learning using varying temporal context length. Multimedia Tools and Applications, 2024, 83(27): 69647-69668. https://doi.org/10.1007/s11042-024-18270-4

[24] Dong L. Using deep learning and genetic algorithms for melody generation and optimization in music. Soft Computing, 2023, 27(22): 17419-17433. https://doi.org/10.1007/s00500-023-09135-3