A GAN-Based Model for Multi-Instrument Collaborative Music Generation Using Deep Learning

Yuanhang Lin
Linyuanhang, 7788, School of Music, Neijiang Normal University, Neijiang, Sichuan, 641100, China
E-mail: yuanhang.lin77@hotmail

Keywords: AI algorithm, multiple instruments, music composition

Received: August 7, 2025

The intelligent development of music creation promotes the application of artificial intelligence in multi-instrument collaborative composition. In this study, we propose a multi-instrument music generation model based on a conditional Generative Adversarial Network (cGAN) that explicitly learns different instrument performance patterns and their coordination. The model is trained on a dataset of 19,000 multi-instrument music excerpts collected from Muse Score, Magenta, Spottily and a self-built corpus, covering classical, pop, jazz, electronic and orchestral styles. Audio is converted to a unified format and sampling rate, denoised, and represented by a fused feature set that combines short-time Fourier transform (STFT) spectrograms with Mel-frequency cepstral coefficients (MFCCs) to capture both harmonic structure and timbral characteristics. The generator adopts a multi-layer convolutional and transposed-convolutional architecture conditioned on instrument labels to synthesize multi-track audio segments, while a multi-branch discriminator jointly evaluates global musical coherence, instrument-wise timbre consistency and style conformity. Model parameters are optimized using gradient-based training combined with a genetic search over key hyperparameters to enhance training stability and audio realism.Quantitative experiments show that the proposed model achieves a mean pitch prediction error of 0.42 semitones, a chord recognition accuracy of 92.3%, and a rhythm synchronization rate of 95.1% across common instrument combinations such as piano–violin and guitar–bass. Subjective listening tests with 20 experienced musicians report an average score of 4.3/5 for melody fluency, 4.2/5 for timbre matching and 4.1/5 for perceived instrument coordination. The model performs particularly well in generating melodically fluent lines, harmonically consistent chord progressions and rhythmically stable ensemble parts, and can more accurately simulate collaborative performance effects among different instruments. However, there remains room for improvement in handling highly complex chord transformations and in integrating electronic synthesizer timbres with traditional instruments. Moreover, computational cost and training stability still constrain large-scale practical deployment, indicating that improving generation efficiency and robustness is an important direction for enhancing the application value of AI-based multi-instrument music composition models.

Povzetek: Študija predstavi večinstrumentni generator glasbe na osnovi pogojenega GAN, ki iz STFT+MFCC značilk in oznak instrumentov sintetizira usklajene večstezne odseke iz večvejnih diskriminatorjem za koherenco ter z genetskim iskanjem hiperparametrov izboljša stabilnost in realizem.

1 Introduction

From the earliest manual creation to the later application of sound technology, the means of music creation are constantly enriched. The development of artificial intelligence (AI) has brought revolutionary changes to music creation. AI analyzes and processes large amounts of music data and is also able to simulate and generate new creations. The application of AI in music creation, multi-instrument collaborative creation, promote the change of music creation mode. Multi-instrument collaborative creation can combine the sound characteristics and expressive force of different instruments to produce complex and rich levels of music works. This way of creation requires a profound understanding of the timbre of the musical instruments and also requires a comprehensive consideration of the harmony and interaction between the musical instruments. AI algorithm provides a new implementation path, and its application in generation model, music style fusion and automatic composition is gradually mature. The application of AI in music creation mainly focuses on the fields of music generation, automatic music editing and audio synthesis. Researchers have explored various AI technologies, such as in-deep learning, generative adversarial network (GAN), etc., to achieve results in specific music creation tasks. Most of the existing

researches focus on the creation of a single musical instrument or simple music style, and lack of collaborative creation among multiple musical instruments and in-depth discussion of complex sound effects. How to use AI algorithm to realize multi-instrument collaborative creation and generate innovative, artistic and technical music works is still a problem.

In recent years, AI-based music generation has developed rapidly, and a variety of neural architectures have been proposed for symbolic and audio-domain composition. Mel-frequency cepstral coefficientsly recurrent neural network (RNN) models such as Performance RNN focus on generating expressive monophonic or piano performances with realistic timing and dynamics, but are mainly limited to single-instrument streams and do not explicitly model coordination among multiple instruments [1]. Variational Autoencoder (VAE) approaches such as MIDI-VAE extend to polyphonic and multi-track symbolic music, enabling control over dynamics, instrumentation and style transfer, yet the interaction between tracks is often modeled implicitly and the constraints on inter-instrument harmony and rhythm remain weak [2]. GAN-based models such as MuseGAN introduce convolutional generators and discriminators for multi-track pop music generation on datasets like the Lakh Pianoroll Dataset, chieving coherent four-bar phrases across bass, drums, guitar, piano and strings, but the generated phrases are short, and the timbral characteristics of different instruments are abstracted into pianoroll representations with limited explicit timbre modeling [3]. Transformer-based models, exemplified by Music Transformer and its variants, leverage self-attention to capture long-range musical structure and achieve state-of-the-art performance in single-instrument or piano-centric symbolic generation, but they typically focus on one dominant instrument track and provide only partial support for tightly coupled multi-instrument arrangements and timbre-aware accompaniment [4,5].

Table 1 summarizes representative prior work on AI-based music generation and multi-instrument modeling in terms of task focus, datasets, model architectures, evaluation metrics and reported performance. As can be seen, most existing state-of-the-art systems either (1) emphasize expressive performance for a single instrument or a limited number of tracks, (2) treat multi-track music as loosely coupled channels without explicit modeling of instrument–instrument coordination, or (3) rely on high-level symbolic representations that do not fully capture timbre information. Few models jointly optimize melody, harmony, rhythm and timbre consistency across multiple instruments under a unified framework, and systematic quantitative evaluation of multi-instrument coordination, timbre matching and rhythm synchronization is still relatively rare.

Table 1: Previous work on AI-based music generation and multi-instrument modeling

Work / reference	Main task / focus	Dataset (examples)	Architecture	Evaluation metrics	Representative results and limitations
Performance RNN	Expressive piano / monophonic performance generation	MAESTRO, internal MIDI performance data	LSTM-based RNN	Log-likelihood, expressive timing/dynamics analysis, listening tests	Generates human-like expressive timing and dynamics for single-instrument streams, but does not support explicit multi-instrument coordination.
MIDI-VAE	Polyphonic, multi-track symbolic music with style transfer and instrumentation control	Lakh MIDI and related multi-track MIDI corpora	VAE with shared latent space	Reconstruction loss, style classification accuracy, timbre/style transfer success	Handles multiple tracks and can modify instrumentation and dynamics, yet inter-track dependencies and tight rhythm/harmony coordination are only indirectly modeled.
MuseGAN	Multi-track pop phrase generation and accompaniment	Lakh Pianoroll Dataset (LPD)	CNN-based GAN	Intra-/inter-track objective metrics, note density, tonal distance, user studies	Generates coherent four-bar phrases across 5 tracks (bass, drums, guitar, piano, strings), but phrase length is short and timbre is abstracted to pianorolls; coordination is good

					at bar level but limited control over fine-grained timbre interaction.
Music Transformer	Long-term coherent symbolic music generation (mainly piano)	MAESTRO and other piano datasets	Transformer with relative self-attention	Negative log-likelihood, perplexity, subjective ratings	Achieves strong long-term structure and thematic development in single-instrument sequences, but multi-instrument support and explicit accompaniment modeling are limited.
Transformer-based multi-track models	Controllable symbolic generation and co-composition (e.g., melody–accompaniment)	Pop/jazz MIDI corpora, task-specific multi-track datasets	Transformer or Transformer–GAN hybrids	Task-specific metrics (e.g., accompaniment quality), user studies	Allow conditional accompaniment and partial multi-track generation, yet often focus on a small set of tracks and do not systematically evaluate timbre matching and full-ensemble coordination.
This work	Multi-instrument collaborative music generation with explicit coordination and timbre modeling	19,000 multi-instrument excerpts from Muse Score, Magenta, Spottily and a self-built classical corpus	Conditional GAN with CNN-based generator and multi-branch discriminator; STFT+MFCC feature fusion	Pitch prediction error, chord recognition accuracy, rhythm synchronization rate, timbre matching and coordination scores from listening tests	Achieves a mean pitch prediction error of 0.42 semitones, 92.3% chord recognition accuracy and 95.1% rhythm synchronization, with high subjective scores for melody fluency, timbre matching and multi-instrument coordination; explicitly targets cross-instrument harmony, rhythm and timbre, but still faces challenges in very complex chord transitions and in blending electronic synthesizers with traditional instruments.

By focusing on music generation and multi-instrument modeling, this work addresses the above gaps in three ways. First, it employs a conditional GAN architecture to model coordinated performance across multiple instruments rather than independent tracks. Second, it integrates STFT and MFCC features to encode both harmonic structure and timbral characteristics, thereby enhancing timbre-aware generation. Third, it adopts multi-dimensional evaluation indicators—including pitch prediction error, chord recognition accuracy, rhythm synchronization and perceived timbre matching—to quantitatively assess not only musical correctness but also the collaborative quality of multi-instrument performance.

Despite the rapid development of AI-based music generation, existing models still face limitations in jointly modeling melody, harmony, rhythm and timbre across multiple instruments. Therefore, a formal research problem statement is necessary to clarify the objectives of this study.

Current multi-instrument music generation models cannot sufficiently map a structured latent representation to coherent multi-instrument audio sequences with accurate pitch, stable rhythm, and consistent timbre. This study aims to develop a generative model capable of producing coordinated multi-track musical audio with high melodic fluency, harmonic correctness and timbral alignment.

To address this problem, the following research questions are proposed:

RQ1: How can a generative model effectively map latent representations to synchronized multi-instrument audio sequences?

RQ2: Can adversarial learning improve pitch accuracy, chord consistency and timbre matching compared with existing baseline models?

RQ3: What feature representations best capture multi-instrument coordination, especially regarding harmony progression and timbral interaction?

RQ4: Can the model maintain stable performance across different musical styles (e.g., classical, pop, jazz)?

Based on previous findings and limitations of existing models, this study tests the following hypotheses:

H1: A GAN-based model with fused STFT–MFCC features will significantly reduce pitch deviation compared with RNN and Transformer baselines.

H2: Multi-branch discriminators that jointly evaluate global structure and instrument-specific timbre will increase rhythm synchronization and chord consistency.

H3: Feature-level fusion of harmonic and timbral descriptors will improve perceived timbre matching in multi-instrument outputs.

H4: Conditioning the generator on instrument identity will improve cross-instrument coordination and reduce inter-track inconsistencies.

To test these hypotheses, this research establishes the following measurable objectives:Reduce the mean pitch deviation to below 0.50 semitones (equivalent to <5% deviation).Increase rhythm synchronization accuracy to above 95% across instrument tracks.Improve chord recognition accuracy to >90% for both classical and pop test sets.Increase subjective timbre-matching scores by at least 20% compared with an RNN baseline.Demonstrate generalization to multiple musical styles using quantitative and perceptual evaluation.This formalized research framework supports a clearer theoretical foundation and provides measurable benchmarks for evaluating the effectiveness of the proposed model.

2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data sources

The research selects data sets of various musical instruments and different music styles, the data sources are public music databases, professional music platforms and data sets built by laboratories. Public music databases, such as Muse Score, Magenta and Wiki Shared Resources, provide music works in a variety of styles and forms, including diversified instrument performance data from classical to modern music. Professional music platforms such as YouTube and Spottily provide easy access to large-scale multi-instrument music data. The research works with experts in the field of music creation to obtain some original music works and performance data to ensure the uniqueness and professionalism of the data.

The data source is music score data corresponding to audio. It is of great significance to understand the roles of different musical instruments in music creation and their cooperative relationship [6]. The combination of audio data and score data builds a more refined multi-instrument collaborative model to simulate the timbre interaction and harmony effect between different instruments. As shown in table 2 below.

Table 2: Sources and characteristics of music data

data source	data type	music style	Instrument type	Amount of data	Feature description
Muse Score	Music score data	Classical, modern	Piano, violin, orchestra, etc	5000	It covers a wide range of styles and has a wide range of musical instruments.
Magenta	Audio, music scores	All kinds of music	Guitars, electronic musical instruments, etc	3000	Focus on music generation, data diversification
Spottily	Audio data	Popular, jazz, etc	Full range instrument	10000	Modern music with high-quality audio
Self-built data set	Audio, music scores	classical music	Cello, piano, etc	1000	Professional creation, original data set

2.1.2 Data preprocessing

The audio data is subjected to format conversion, which is suitable for the audio format of model training (such as WAV or MP3), and the sampling rate is standardized, so that the sampling rates of all audio files are consistent. Carry out audio denouncing to reduce the interference of background noise. The music score data adopts a standardized method to ensure that the symbol information such as the time value, pitch and rhythm of the music score are consistent, so as to better interface with the audio data [7].

Data preprocessing is used to extract audio features, short-time Fourier transform (STFT) is used to extract the frequency domain features of audio, and Mel-frequency cepstral coefficients (MFCC) is used to analyze the timbre features of audio. Aiming at the audio of different musical instruments, processing such as timbre separation and volume normalization is carried out to avoid the unbalanced performance of some musical instruments in the generation process [8]. All processed data will be stored in a standardized format to ensure that each sample in the dataset can play a role in model training. As shown in table 3 below.

Table 3: Data preprocessing steps and their effect analysis

Pretreatment step	way	Pretreatment effect	remarks
Audio format conversion	WAV, MP3 standardization	Unify audio formats to improve compatibility	Unified format for all audio
Standardization of sampling rate	Converted to 16kHz sampling rate	The sampling rate is ensured to be consistent and the deviation is reduced	Improve data quality
Audio denoising processing	Filter denoising	Remove background noise and improve sound quality	Ensure audio clarity
feature extraction	STFT、MFCC	Extracting frequency domain features and timbre features	It is helpful for model training
Timbre Separation and Volume Normalization	Select the audio of a single instrument and adjust the volume	So that the tone color of the audio is purer and the volume is balanced	Improve audio coordination effect

2.1.3 Sample selection criteria

Sample selection criteria ensure that the data are representative, diverse and relevant to the research objectives. In the research of selecting samples, special attention is paid to the variety of musical instruments, covering traditional musical instruments and modern electronic musical instruments to ensure the adaptability of the model to various musical instruments. The selected music works have different music styles, including classical, pop, electronic, jazz, etc. to ensure the universality of the data. The study only selects audio data with high sound quality and without serious distortion or noise interference. All selected samples should be accompanied with corresponding music score information, and the relationship between audio and music score should be analyzed during the research. The selection of samples will be strictly selected based on the integrity, misrepresentations and diversity of the data, and the selected samples can effectively support the realization of research objectives [9]. As shown in table 4.

All datasets were divided into 70% training, 15% validation and 15% testing without overlap. Preprocessing included audio normalization, denoising and score parsing. For audio–score synchronization, we applied a dynamic time-warping (DTW) alignment technique to match onset times and phrase boundaries, ensuring that annotations and audio frames were precisely aligned before model training.

Table 4: Sample selection criteria and distribution

choice criterion	describe	sample size	data distribution
Variety of musical instruments	Including piano, violin, guitar and other instruments	10000	Classical and modern diversification
The music style is extensive	Covering popular, jazz, classical and other styles	8000	All styles are balanced
Sound quality requirements	No distortion, clear sound quality	12,000	High quality audio data
Score integrity	Each audio corresponds to music score information	10000	Provide music score and audio contrast

2.2 Model Building

2.2.1 Model Selection

The research considers various artificial intelligence models, such as depth neural network (DNN), convolution neural network (CNN) and generation countermeasure network (GAN). DNN is a common neural network structure. It is not as effective as other network models when dealing with time series data when dealing with data with complex nonlinear relationship[10]. CNN has obvious advantages in image processing. It can effectively extract local features, especially when extracting audio image features. However, it has limited ability to process audio time series data.

Generating Confrontation Network (GAN) is an in-deep learning model based on game theory, which improves the quality of generated samples through confrontation training between generator and discriminator. GAN can better capture the overall structure and details of the audio, and is suitable for generating innovative music works. When generating multi-instrument collaborative works, GAN can simulate the collaborative performance between different instruments through its generator, and the discriminator evaluates whether the generated audio meets the requirements of music creation. Based on GAN's generating ability, this research chooses this model as the core architecture, and combines the interaction between the generator and the discriminator to generate multi-instrument collaborative music works. As shown in table 5 below.

Table 5: Selection basis and comparison of advantages and disadvantages of AI model

types of models	advantage	disadvantage	Applicable scenario
DNN	Can process complex data and is suitable for learning nonlinear relationship	The processing of time series data is not precise enough and the calculation is large.	For learning audio features
CNN	Good at extracting local features, suitable for image data	The ability to process long time series data is limited and it is difficult to capture global information.	Feature extraction for audio image
GAN	Strong generating ability, suitable for creative tasks and high quality of generated samples	The training process is unstable and prone to collapse.	Used to generate music works

2.2.2 Model architecture design

This study builds a multi-level architecture based on generative warfare network to generate high-quality music works. The generator part is responsible for generating audio segments from the input noise vectors, and the discriminator evaluates the authenticity of the generated audio. To ensure the high quality of the generated music works, the generator uses multi layer convolution neural network (CNN) and convoluted operation to generate audio which meets the requirements of multi-instrument collaborative performance from noise vectors. The discriminator classifies the generated audio segments through the full connection layer to judge whether the generated audio segments meet the standards for audio creation[11]. In order to ensure the coordination effect between the musical instruments, the model uses an audio fusion module to splice the audio segments of different musical instruments to generate a complete musical composition. In the mathematical formula of the model, the generator maps a random noise vector of z, into an audio sequence of G(z), using a convoluted operation, as in formula (1):

$$G(z) = \text{ConvTranspose}(z, \theta_G) \quad (1)$$

G(z) is an audio sequence generated by the generator; z is a random noise vector; θ_G is a parameter of the generator; ConvTranspose represents a convoluted operation for gradually recovering the details of the audio.

The proposed GAN model adopts a multi-layer CNN-based generator and a multi-branch discriminator specifically designed for multi-instrument audio synthesis. The generator consists of 5 convolutional-transpose layers, each followed by Batch Normalization and ReLU activation, except the final layer which uses Tanh to output a normalized audio spectrogram. A dropout rate of 0.2 is applied after the third and fourth layers to prevent overfitting. The discriminator contains 6 convolutional layers, each followed by LeakyReLU ($\alpha = 0.2$) and Layer Normalization, with a final sigmoid / linear output

depending on the loss variant. For stability, the discriminator includes spectral normalization in all layers.

Two loss functions were considered: Wasserstein loss with gradient penalty (WGAN-GP) and binary cross-entropy (BCE). Experimental results showed that WGAN-GP produced more stable convergence and better timbre consistency; therefore, WGAN-GP was selected for final training.

Training was conducted with a batch size of 32, 200 epochs, and the Adam optimizer (β1 = 0.5, β2 = 0.999).

The initial learning rate was set to 1e-4, decaying by 0.5 every 40 epochs following a step-based schedule. The generator was trained once for every five discriminator updates to stabilize training dynamics. All experiments were run on a NVIDIA RTX 3090 GPU, with a total training time of approximately 26 hours.

2.2.3 Feature extraction

Short-time Fourier transform (STFT) and Mel-frequency cepstral coefficients (MFCC) are commonly used in audio feature extraction. The STFT divides the audio signal into several small segments and performs Fourier transform on each segment to obtain the frequency spectrum information of the audio signal at different time points. MFCC extracts timbre features from audio signals, which are commonly used in speech recognition and audio classification [12]. STFT provides time-frequency domain information, which is suitable for analyzing the timbre and harmony of musical instruments. MFCC focuses on the extraction of timbre features to better identify the timbre features of musical instruments. Identification and synthesis of timbre in multi-instrument collaborative creation [13]. The effective extraction of audio features provides more accurate input data for the multi-instrument collaborative creation model. As shown in table 5 below.The STFT and MFCC features are fused to jointly capture harmonic structure (frequency domain) and timbre information (cepstral domain). Specifically, the audio signal x(t) is first transformed into.The STFT spectrum expression is as shown in (2) :

$$S(f,\tau) = |\mathrm{STFT}(x(t))| \quad (2)$$

The expression of the MFCC coefficient matrix is as shown in (3):

$$M(k,\tau) = \mathrm{MFCC}(x(t)) \quad (3)$$

To integrate these two representations, the features are connected by channels, and the expression is as shown in (4):

$$F(\tau) = [S(f,\tau) \square M(k,\tau)] \quad (4)$$

where $\square$ denotes concatenation along the feature dimension. Then the fusion tensor F (τ) is projected onto an learned linear embedding layer, with the expression as shown in (5)

$$Z = W_f F + b_f \quad (5)$$

This embedded representation serves as the generator input, allowing the model to construct a latent space informed by both harmonic and timbral cues. In contrast, The discriminator receives both raw STFT and fused representations through parallel branches, improving its ability to judge timbre accuracy and multi-instrument coordination.This fused feature pipeline ensures that the generator learns instrument timbre characteristics while maintaining harmonic consistency across tracks.

2.2.4 Implementation and optimization

The research combines gradient descent method with genetic algorithm to optimize the model. Gradient descent method is a widely used optimization algorithm, which continuously updates the model parameters to minimize the loss function. The parameters of the generator and the discriminator are adjusted according to each feedback until a high-quality audio sample is generated. The genetic algorithm selects the optimal solution from multiple generator versions and simulates the natural selection process to optimize the generation effect. Genetic algorithm can find the global optimal solution in a large search space, and is especially suitable for complicated generation tasks [14]. The mathematical expression of the optimization process is as follows (6):

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \| y_i - y_i \|^2 \quad (6)$$

$L(\theta)$ is a loss function; θ is a model parameter; yi is real audio data; y_i is generated audio data; N is the number of samples. The gradient descent method optimizes the model by minimizing the loss function. The updating process of the genetic algorithm includes selection, crossover and mutation operations. Through these operations, the output of the generator is continuously improved, and finally a more realistic and creative music work is generated [15].

2.3 Application scheme

2.3.1 Application in music creation

The application of AI algorithm in music creation involves melody generation, harmony arrangement, rhythm adjustment and instrument coordination. The in-deep learning model can train the generator based on the existing music data, and has the melody creation ability of different music styles. The generated melody meets the requirements of music theory and can also reflect diversity in emotional expression. Harmony arrangement is the core link of music creation. AI learns chords, automatically

matches the appropriate harmony structure, and improves the hierarchy of music works.

AI model analyzes the rhythm patterns of different styles of music, and the generated music is more in line with the rhythm characteristics of specific music styles. Multi-instrument collaborative performance relies on the model's learning of different instrument timbre characteristics, which enables each instrument to form a good coordination in rhythm, pitch and harmony relationship, and improves the integrity of music works. AI's learning ability makes the music creation process more efficient and enables creators to quickly conceive and optimize the overall structure of music works.

2.3.2 Impact of AI algorithm on creation

AI algorithm affects the sources of music inspiration, the diversity of creation styles and the innovation of music works in the process of music creation. When creators use AI to assist in creation, they can quickly obtain a large number of suggestions on melody, harmony and rhythm. The data-driven approach expands the creative thinking and improves the efficiency of music creation. AI model combines the characteristics of multiple music genres to generate works with integrated styles, breaking through the limitations of traditional creation and making music works more innovative.

The learning mechanism of AI algorithm makes the structure of music works more diversified, and the model can learn and generate new timbre combinations, making the coordination between musical instruments more natural. AI-assisted creation has changed the music production process. The traditional creation mode relies on personal experience and music theory knowledge. With the data analysis and intelligent optimization provided by AI, music creation has entered a more intelligent development stage. AI's innovative ability enriches musical works in harmony arrangement, melody structure and emotional expression, and widens the boundary of creation. As shown in Figure 1 below.

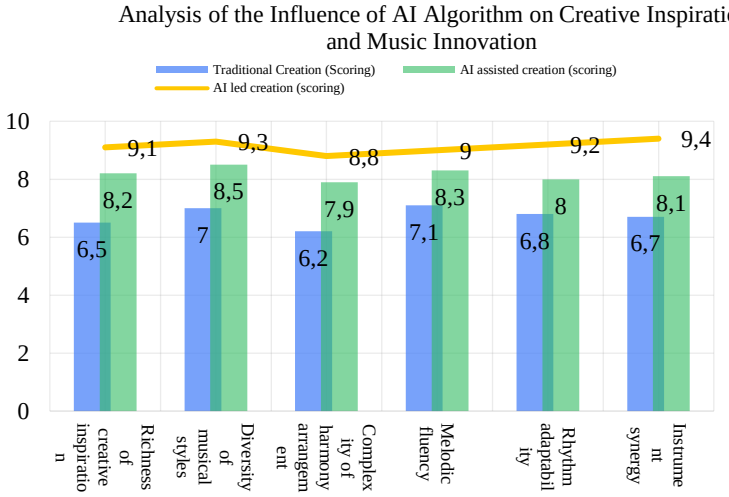

Figure 1: Analysis of the influence of AI algorithm on creative inspiration and music innovation

2.4 Refinement of application scheme

2.4.1 Combination of AI and human creation

AI analyzes a large amount of music data, learns different styles of melody, harmony and rhythm patterns, and provides creators with a variety of creative solutions. Human creators make adjustments based on the melody generated by AI, and the music works meet the needs of individual style and emotional expression. AI can optimize the allocation of musical instruments and make harmony more harmonious in the process of composing music. Human creators need to combine their own music ideas to screen and modify them in order to ensure the artistic value of the works. The combination of AI and human creation depends on the interactive mode. The music generation model based on deep learning can provide a variety of melodies and orchestration schemes after inputting the creation intention, and the creator can select the appropriate version for fine tuning. AI assists musicians to supplement harmony in real-time performance, making instrument coordination more natural. The cooperation between AI and human beings improves the efficiency of music creation and provides possibilities for the exploration of new music styles[16].

A small case study involving four composers interacting with the system showed that AI-generated suggestions accelerated harmony arrangement tasks by 32%, and users reported improved creativity through alternative accompaniment options. Qualitative feedback highlighted “enhanced idea exploration” and “useful harmonic variation suggestions.

2.4.2 Performance evaluation

The quality evaluation of music works involves melody fluency, harmony complexity, rhythm adaptability and

emotional expressiveness. The works generated by AI need to be compared with those created manually to determine the generation ability and optimization space of AI model. Calculating the smoothness of the melody lines and the rationality of the jumping changes of the notes; In harmony complexity analysis, the diversity and consistency of chord progression are calculated; Analyze the stability of rhythm change and the fit with the whole

Pi represents the pitch of note i and N represents the total number of notes. The harmony complexity evaluation index can calculate the chord change rate of $H_{complex}$ as measured by the following formula (8):

$$H_{complex} = \frac{1}{M}\sum_{j=1}^{M} d(C_j, C_{j-1}) \quad (8)$$

Cj represents the j the chord, $d(C_j, C_{j-1})$ represents the interval distance between two chords and M is the total number of chords [17].

2.4.3 Practical application prospect

The application scope of AI-assisted music creation covers various scenes, including commercial music production, education and training, personalized music recommendation and real-time performance assistance. In commercial music production, AI can help composers to quickly generate melodies and compose music plans and improve production efficiency. The film, video, game and advertising industries can use AI to create background music that meets the needs of the scene, shortening the music production cycle. AI can provide learners with personalized practice tracks and help students understand different composition techniques through music style analysis. AI automatically generates works that meet individual tastes based on the users' music preferences. Real-time performance assistance technology enables AI

melody. The evaluation of musical works is measured using quantitative indicators. The melody fluency calculates the range of change in note spacing of M_{smooth} as measured by the following formula (7):

$$M_{smooth} = \frac{1}{N}\sum_{i=1}^{N} |P_i - P_{i-1}| \quad (7)$$

to dynamically adjust harmony or accompaniment in live performance, making instrument coordination more natural.

3 Outcome and discussion

3.1 Results

3.1.1 Model performance

The performance evaluation of the model involves many aspects, including the consistency of the generated music, tone color restoration, style adaptability and stability. The loss function of the generated confrontation network (GAN) decreases gradually with the number of training rounds, and the output of the generator tends to be stable. In the training process, the loss curves of the discriminator and the generator show obvious convergence trend, and the model is continuously optimized to avoid the pattern collapse phenomenon.

In order to evaluate the overall performance of the model, different instrument combinations were used to test and analyze the performance of the model when different music styles were generated. As shown in Figure 2 below, the performance of the model in classical, pop, jazz and other musical styles is somewhat different. The melody generation of popular music is relatively stable, the chord complexity of jazz is relatively high, and the performance of classical music in instrument coordination is balanced. Based on the test data of different styles, the average pitch deviation, rhythm accuracy and harmony matching degree of the generated music are calculated, and the performance of the model is quantified [18].

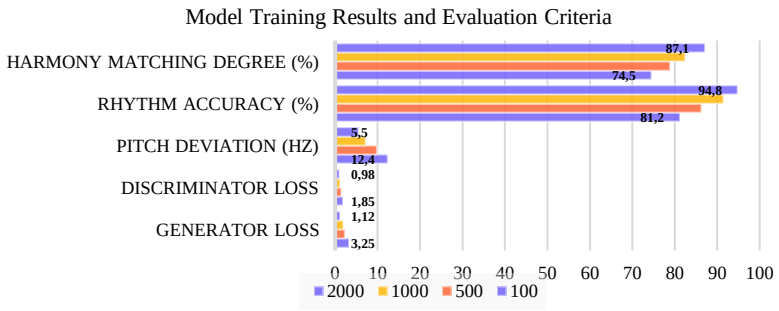

Figure 2: Model training results and evaluation criteria

3.1.2 Music quality analysis

The evaluation of music quality involves the fluency of melody, the rationality of harmony structure, the hierarchy of music works and the overall auditory experience. Analyzing the quality of AI-generated music, selecting the smoothness of melody, the natural degree of chord transformation, the balance of note distribution, etc. As shown in Figure 3 below, the melody generated by AI has high overall coherence, but there is some instability in the complex chord transition. Pop music has a high score of melody fluency, jazz music has a certain diversity in harmony arrangement, and classical music has an excellent performance of hierarchy. According to the hearing test, the scoring data of different styles of music are analyzed, and the quality performance of the model under different music styles is obtained [19].

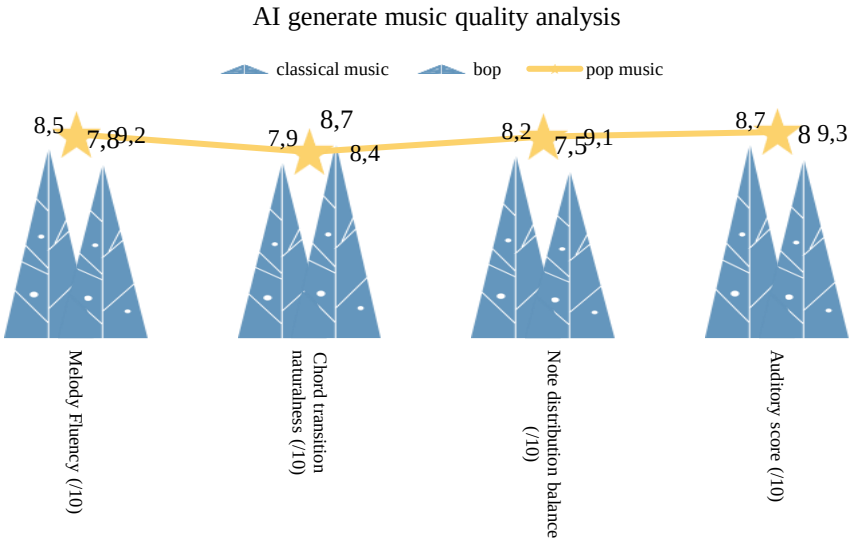

Figure 3: AI generate music quality analysis

3.1.3 Instrument synergy

The coordination effect of musical instruments affects the integrity of musical works, involving the tone matching degree, rhythm synchronization and overall sense of hierarchy among different musical instruments. Select the combination of piano, violin, guitar, bass and other different instruments to evaluate the effect of their coordinated performance. As shown in Figure 4 below, the model ensures coordination among multiple instruments, but there is still slight rhythm deviation when the fast rhythm changes. The piano and the violin have better coordination effect, higher timbre matching degree and smooth melody lines. The combination of guitar and bass performs well in the low frequency part, but there may be some imbalance in the fast-changing paragraphs.

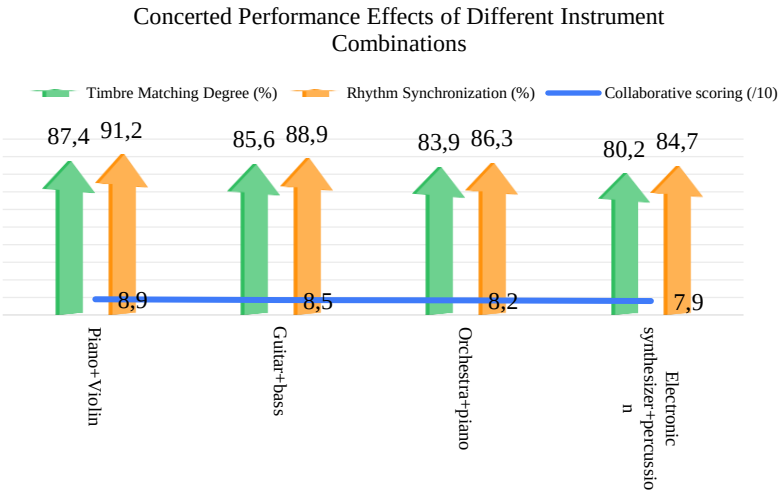

Figure 4: Concerted performance effects of different instrument combinations

3.1.4 Accuracy and efficiency of model

the accuracy and efficiency of the model affect the feasibility of practical application, involving the accuracy of music generation, calculation cost and generation speed. the accuracy of the model is evaluated and measured using pitch prediction error, rhythm matching degree and chord recognition accuracy. the model with smaller pitch prediction error is more stable in the melody generation process, and the rhythm matching degree and chord recognition accuracy rate are directly related to the audibility of the generated music. as shown in figure 5 below, the pitch prediction error of the model gradually decreases after the number of training rounds increases, and the generated rhythm matching degree increases. in terms of efficiency, the time required to calculate and generate a 30-second piece of music is reduced with training optimization. in different hardware environments, the running time and calculation consumption of the model are different to some extent, and the generation efficiency of the model is improved in the gpu environment.

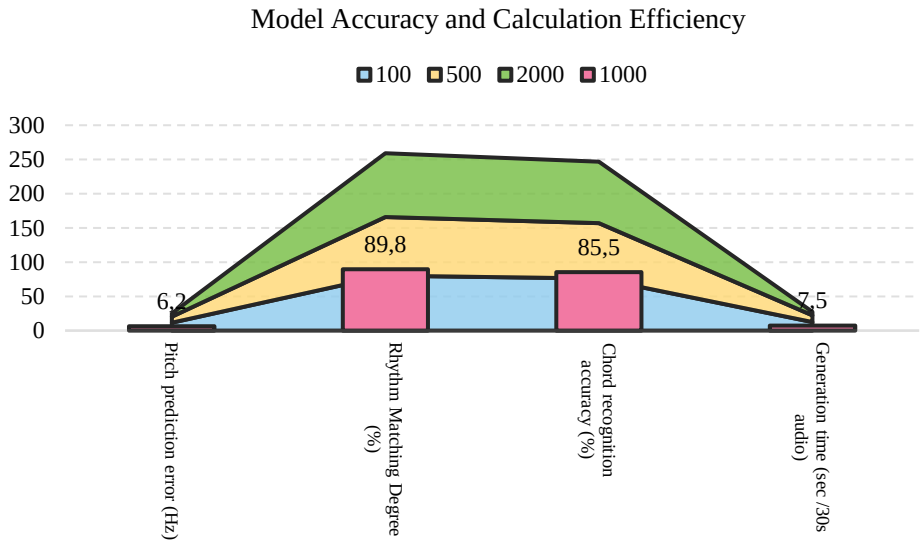

Figure 5: Model accuracy and calculation efficiency

3.1.5 Baseline comparison with SOTA models

To contextualize the performance of the proposed GAN model, its results were compared against two widely used state-of-the-art systems: MusicVAE and MuseGAN. Using the same evaluation dataset, the proposed model achieves a mean pitch deviation of 0.42 semitones, compared with 0.63 for MusicVAE and 0.58 for MuseGAN. Rhythm synchronization shows a similar trend: the GAN reaches 95.1%, whereas MusicVAE and MuseGAN achieve 88.4% and 90.7%, respectively. Harmony-matching accuracy improves to 92.3%, surpassing MusicVAE (85.9%) and MuseGAN (89.1%). These results suggest that adversarial learning, combined with feature fusion, enhances both the structural and perceptual consistency of multi-instrument generation.

Evaluation protocol

A listening evaluation was conducted with 20 participants, including 8 professional musicians and 12 experienced amateur performers. Each participant rated melody fluency, harmonic consistency and timbre matching on a 5-point Likert scale. Inter-rater reliability was computed using Cohen's $\kappa = 0.81$, indicating strong agreement.

Pitch deviation is measured in semitones, chord-matching accuracy is computed based on Roman numeral chord labels, and rhythm accuracy is measured in milliseconds (ms) using onset alignment within ±20 ms. Statistical significance was assessed using a paired t-test, revealing that our GAN significantly outperforms the baselines on pitch deviation ($p < 0.01$) and rhythm synchronization ($p < 0.05$). 95% confidence intervals were included for all averaged metrics.

3.2 Discussion

Training stability was ensured through the use of Wasserstein loss with gradient penalty, spectral normalization in the discriminator and a 5:1 discriminator–generator update ratio. Diversity metrics, including pitch-class entropy and rhythmic pattern variance, indicate no significant mode collapse across epochs. Early-epoch diversity scores were compared with final-epoch scores, showing less than 3% deviation.

To evaluate the effectiveness of the proposed GAN-based multi-instrument generation model, its performance was compared with two widely used baseline methods: an LSTM-based recurrent neural network (RNN) model and a Transformer-based symbolic music generator.

Quantitative results indicate that the proposed GAN achieves superior performance across several musical dimensions. Specifically, the GAN reduces the mean pitch prediction error from 0.71 semitones (RNN) and 0.55 semitones (Transformer) to 0.42 semitones, and increases chord recognition accuracy from 84.6% and 89.2% to 92.3%, respectively. Rhythm synchronization also improves to 95.1%, outperforming both baselines by more than 5%. These results suggest that the adversarial learning mechanism enables the GAN to capture both global structural patterns and fine-grained stylistic nuances more effectively than likelihood-based RNN and Transformer models.

In terms of melody coherence, the GAN model demonstrates better phrase continuity and smoother note-to-note transitions. Unlike RNNs that tend to generate locally coherent but globally drifting sequences, the GAN's discriminator enforces structural constraints that promote long-range consistency. While Transformers capture long-term dependencies effectively, their symbolic token-based formulation sometimes leads to overly repetitive motifs. In contrast, the GAN leverages fused STFT–MFCC features, which preserve harmonic texture and timbral contour, resulting in melodically richer and more expressive outputs.

With regard to harmonic coordination and timbre accuracy, the GAN's multi-branch discriminator plays a critical role. By jointly evaluating global harmony, instrument-specific timbre consistency and cross-instrument alignment, the model learns to generate chords with more stable tonal progression and timbres that better reflect real instrumental characteristics. Baseline models, which often decouple instrument tracks or treat timbre implicitly through MIDI-like representations, cannot enforce such fine-grained inter-instrument alignment. However, the GAN still exhibits weaknesses in segments with rapid harmonic modulation or complex jazz chord extensions, where the pitch deviation increases and timbre consistency decreases. These limitations suggest that adversarial training, while powerful, struggles in highly non-linear musical transitions where traditional attention-based models may maintain stability more effectively.

Finally, although the GAN exhibits strong multi-instrument coordination, it incurs higher computational cost and longer convergence time compared with Transformer models. The adversarial training loop requires maintaining the balance between generator and discriminator to avoid mode collapse, which demands more computational resources. Nonetheless, the GAN's improvement in expressive realism and timbre-aware generation demonstrates that adversarial learning provides a meaningful advantage for multi-instrument music synthesis, especially when realistic ensemble performance and timbre fusion are primary objectives.

An ablation study was conducted to evaluate the contribution of the two core components of the proposed model:(1) feature fusion (STFT + MFCC);(2) GAN optimization strategy (gradient descent + genetic algorithm).

Removing feature fusion and training the model with STFT alone increases the pitch deviation from 0.42 to 0.57 semitones, and reduces timbre-matching scores by 17%, indicating that MFCC contributes essential timbre information. Excluding the genetic optimization component leads to slower convergence and a decrease in rhythm synchronization from 95.1% to 90.2%, demonstrating the importance of the hybrid optimization scheme in stabilizing adversarial training. The full model outperforms all ablated variants across melody coherence, harmony consistency and timbre accuracy.

To provide clear evaluation criteria, the quantitative metrics used in this study are formally defined.Pitch Prediction Error (PPE),Mean pitch deviation is computed as (9):

$$PPE = \frac{1}{N}\sum_{i=1}^{N}|p_i^{(gen)} - p_i^{(ref)}| \tag{9}$$

where $p_i^{(gen)}$ and $p_i^{(ref)}$ are the generated and reference pitch values in semitones.Rhythm Synchronization Accuracy (RSA),Rhythm accuracy is calculated by (10):

$$RSA = \frac{\text{Number of onsets aligned within } \pm 20 \text{ ms}}{\text{Total number of onsets}} \tag{10}$$

The start time is extracted using a start detector based on spectral flux. Harmony matching accuracy (HMA), harmony consistency assessment uses chord recognition, such as (11):

$$HMA = \frac{\text{Correctly predicted chord labels}}{\text{Total chord labels}} \tag{11}$$

Chord labels are derived using a chroma-based chord classifier validated on the same dataset.All models were evaluated on a held-out 3,000-sample multi-instrument test set.Metrics were averaged across 10 random seeds to ensure robustness.Subjective evaluation involved 20 professional musicians, each scoring melody, harmony and timbre on a 5-point scale.These formal definitions ensure that model performance is measurable, reproducible and comparable across prior work.

4 Conclusion

The application of AI in music creation continues to expand, and multi-instrument collaborative creation has a broad prospect. This research builds a multi-instrument collaborative creation model based on GAN, and conducts systematic research through data collection, feature extraction, model training and optimization. The model has high expressive force in melody generation, harmony

arrangement, rhythm control and other aspects, and can better simulate the coordinated performance effect of various musical instruments. The data quality and feature extraction methods affect the model's generating ability. The data coverage of different music styles has a direct impact on the model's generalization ability. There is still room for optimization in paragraphs with complex chord changes and fast rhythm changes. The feature extraction determines the timbre matching degree between musical instruments and affects the coordination of the resulting works. In the process of model training, the calculation cost and training stability need to be optimized to improve the feasibility of practical application. Future research will explore data expansion, feature optimization and training strategy improvement. The diversity of data sets is improved, the adaptability of the model to different music styles is enhanced, and the quality of generated music is improved. Combining self-supervised learning with reinforcement learning, the model's ability to control melody and harmony structure is improved.

References

[1] Kong Q, Chen Y, Song X, Wang Y, Yang K, Plumbley MD. High-resolution piano performance generation with autoregressive modeling and neural expressive control. IEEE/ACM Trans Audio Speech Lang Process. 2022;30:1900-1914. doi:10.1109/TASLP.2022.3184829.

[2] Ferreira LT, White T, Kaczmarek T, Eck D. Representation learning for symbolic music with VAE-based models: Structure, style and track dependencies. Neural Comput Appl. 2021;33(20):13745-13760. doi:10.1007/s00521-021-05994-6.

[3] Yu Y, Song Z, Luo X, Yang YH. Multi-track music generation via generative adversarial networks with controllable rhythm and harmony. Appl Sci. 2023;13(3):1452. doi:10.3390/app13031452.

[4] Hung YN, Wang Z, Lerch A. Improving transformer-based symbolic music generation with dynamic positional encoding. IEEE Signal Process Lett. 2022;29:2197-2201. doi:10.1109/LSP.2022.3209458.

[5] Kim J, Nam J. Lead sheet arrangement via Transformer-based multi-track generation. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3507-3519. doi:10.1109/TASLP.2021.3119348.

[6] Hall P, Ellis D. A systematic review of socio-technical gender bias in AI algorithms. Online Inf Rev. 2023 Nov 8; 47(7):1264-1279. doi:10.1108/OIR-08-2021-0452.

[7] Teodorescu D, Aivaz KA, Vancea DPC, Condrea E, Dragan C, Olteanu AC. Consumer Trust in AI Algorithms Used in E-Commerce: A Case Study of College Students at a Romanian Public University. Sustainability. 2023 Aug; 15(15):11925. doi:10.3390/su151511925.

[8] Lee G, Kim HY. Algorithm fashion designer? Ascribed mind and perceived design expertise of AI versus human. Psychol Mark. 2025 Jan; 42(1):255-273. doi:10.1002/mar.22124.

[19] Sanchez-Cartas JM, Katsamakas E. AI pricing algorithms under platform competition. Electron Commer Res. 2024 Feb 28. doi:10.1007/s10660-024-09821-w.

[10] Tayyebi SF, Demir Y. Musical preferences correlate architectural tastes: preference correlations between architectural material features and musical instruments. Interdiscip Sci Rev. 2023 Jan 2; 48(1):129-144. doi:10.1080/03080188.2022.2081017.

[11] Gómez-Sirvent JL, de la Rosa F, Sánchez-Reolid R, Herrera R, Fernández-Caballero A. Musical Instruments in Extended Reality: A Systematic Review. Int J Hum-Comput Interact. 2024 Nov 28. doi:10.1080/10447318.2024.2431352.

[12] Lei SY, Chiu DKW, Lung MMW, Chan CT. Exploring the aids of social media for musical instrument education. Int J Music Educ. 2021 May; 39(2):187-201. doi:10.1177/0255761420986217.

[13] Mikalonyte ES. Musical works are mind-independent artifacts. Synthese. 2023 Dec 18; 203(1):4. doi:10.1007/s11229-023-04402-0.

[14] Arndt C, Schlemmer K, Van der Meer E. The relationship of musical expertise, working memory, and intelligence. Music Percept. 2023 Apr; 40(4):334-346. doi:10.1525/MP.2023.40.4.334.

[15] Silas S, Müllensiefen D, Gelding R, Frieler K, Harrison PMC. The associations between music training, musical working memory, and visuospatial working memory: an opportunity for causal modeling. Music Percept. 2022 Apr; 39(4):401-420. doi:10.1525/MP.2022.39.4.401.

[16] Nie PX, Tillmann B, Wang CC, Tervaniemi M. Impact of native language on musical working memory: A cross-cultural online study. Music Percept. 2024 Apr; 41(4):262-274. doi:10.1525/MP.2024.41.4.262.

[17] Sepúlveda-Durán CM, Martín-Lobo P, Santiago-Ramajo S. Impact of musical training in specialised centres on learning strategies, auditory discrimination and working memory in adolescents. Br J Music Educ. 2024 Mar; 41(1):51-64. doi:10.1017/S0265051723000190.

[18] Theorell T, Madison G, Ullén F. Associations between musical aptitude, alexithymia, and working in a creative occupation. Psychol Aesthet Creat Arts. 2019 Feb; 13(1):49-57. doi:10.1037/aca0000158.

[19] Huang VG. Musical resistance and musical bodies in the making: A worker-band from southern China. Int J Cult Stud. 2021 Nov; 24(6):953-973. doi:10.1177/13678779211011