ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

The Inheritance Path of Ethnic Music Culture in Pre-school Music Curriculum of Colleges and Universities

Na Li[1,*]

[1] School of Normal, Jingchu University of Technology, Jingmen, Hubei, 448002, China

SUMMARY: *Ethnic music, as a valuable treasure created in the long-term historical development of the Chinese nation, has the unique cultural elements of the Chinese nation in its musical form, artistic value and ideological concepts, and its integration into the music curriculum of preschool education in colleges and universities can help students to deepen their understanding of the ethnic culture and promote the inheritance of ethnic music culture. For this reason, the study builds an immersive situation in the ethnic music classroom of preschool education in colleges and universities, and explores the effective path of the music generation model in promoting the reform of experiential ethnic music teaching. Based on the elaboration of related music knowledge and MIDI file format, we innovatively constructed a multi-track ethnomusicological music generation model Tr-MTMG based on Transformer, which consists of three parts: data preprocessing network, learning network and generating network, and is capable of generating multi-track music. Through a series of experiments, in the Lakh MIDI dataset, the entropy value of the level histogram of this model is 2.8504, which is closest to the real samples, and it proves that the model based on this paper can generate more harmonized and more realistic music. The music created by this model achieved a high score of 105.65 in the Turing test. After multiple evaluations, it is verified that the model proposed in this paper can generate good quality music, and can be effectively applied to music tasks in pre-school education in universities.*

KEYWORDS: *transformer; Tr-MTMG model; music generation; ethnic music*

1 Introduction

With the progress of modern society and the increase in attention to quality education, it makes people pay more and more attention to the comprehensive development of students' moral, intellectual, physical, social, aesthetic and labor, and is no longer limited to evaluating students by their grades. The improvement of students' aesthetic quality can help students obtain a broader space for development, so that they can learn happily and actively [1]. Integrating ethnic music into the teaching content of the preschool education professional curriculum can effectively enrich the teaching content of the preschool education professional curriculum, and also allow students to keep up with the modern concept of educational development and the new direction of talent cultivation, so that students can form an advanced educational philosophy [2-4]. Effectively expanding the scope of dissemination of folk music, enriching the form of dissemination of folk music, so that students can not only listen to the beautiful melodies of folk music, but also understand the cultural connotations of folk music, combined with national characteristics and customs to feel the charm of traditional Chinese culture, and

*margielee118@163.com
https://doi.org/10.65102/is2026044
Published: 30 April, 2026

consciously carry out the exploration of the cultural background and melodic characteristics behind folk music [5].

However, judging from the current situation, preschool education majors in colleges and universities do not pay enough attention to the effects and roles of ethnic music education, and therefore fail to fully explore and utilize relevant ethnic music materials [6]. Some teachers fail to fully understand the artistic characteristics of ethnic music, resulting in ethnic music education has been in a passive position, the purpose of ethnic music teaching is difficult to put into practice, and the teaching effect is unsatisfactory [7]. In addition, some college teachers do not have a deep enough understanding of folk music, and their teaching methods are relatively single, and they do not have the skills to use folk music to carry out activities, and in teaching practice, they cannot lead by example, and they are not clear about the important value of folk music in traditional culture, so they cannot flexibly utilize different folk music materials in actual teaching [8-10]. Ethnic music is rich in content, including folk songs and dances, local operas, folk ballads and so on. Different forms of ethnic music, its musical structure, artistic style, rhythm and melody are different, which makes it difficult to carry out preschool education teaching activities [11]. At this stage, the material content of ethnic music education for preschool education majors in colleges and universities is not rich enough, and there is a thin teaching content and too few musical works. In the actual teaching process, preschool education teachers in the process of ethnic music education, often due to the lack of material content, can not effectively integrate ethnic music resources and teaching activities [12-15]. Due to the uniqueness of ethnic music education, classical teaching activities usually take a lot of time, and the heavy workload of some teachers themselves, coupled with the lack of material content, leads to the poor quality of ethnic music education [16-19]. Therefore, in the above context, it is of far-reaching significance to study in depth the inheritance path of ethnic music culture in the music curriculum of preschool education in colleges and universities.

The article first explores methods of creating immersive contexts in college preschool ethnic music classrooms. Music fundamentals and the MIDI file format are described. The optimization model Linear Transformer, which can speed up training and reduce memory usage, is briefly introduced. A music generation model, Tr-MTMG, is proposed, which consists of a data preprocessing network, a learning network, and a generative network, and the learning network consists of a Cross-trackattention mechanism, which is an improvement of Transformer, to learn the information between the tracks of different instruments. And the ability of text continuation of GPT model is utilized in the generative part for music continuation. In order to demonstrate the effectiveness and superiority of the multi-track music generation model Tr-MTMG proposed in this paper, model performance evaluation is carried out. Finally, a combination of objective and subjective assessment is used to comprehensively evaluate the quality of the music generated based on the model of this paper.

2 Immersive Context Creation in the Ethnic Music Classroom of Pre-school Education in Colleges and Universities

2.1 Construction of art resources with modern technical support

In the teaching practice of music appreciation in pre-school education in colleges and universities, multimedia technology and virtual reality and other modern means are fully utilized to build a digital art resource library. Taking the teaching of “Fengyang Flower Drum” as an example, VR technology is utilized to reproduce the ancient art of dance, and through

panoramic audio-visual experience, students are helped to deeply understand the artistic charm of Han folk customs. In preschool music classroom teaching, digital audio processing technology is used to demonstrate and analyze different vocal methods and breathing techniques. With the help of audio capture equipment, record students' singing data to assist them in mastering the correct vocalization. Establish a digital resource library for vocal training, and systematically collect and organize demonstration videos of different singing styles, such as American, ethnic, and popular. At the same time, a cloud-based teaching platform is built to integrate the resources of basic vocal training, breath practice, resonance training and other high-quality courses to realize the sharing of teaching resources. Develop intelligent music teaching software, analyze students' singing data, and provide a scientific basis for teaching students according to their aptitude. Through the virtual concert hall, students can enjoy the performances of famous vocal artists at home and abroad to further improve their vocal singing ability.

2.2 Integration of local music culture into artistic context creation

Based on the cultural characteristics of mountainous music, local music elements are integrated into the creation of situations in teaching. In the music courses of pre-school education in colleges and universities, local ethnic musicians are invited to enter the classroom to guide students in terms of pharyngeal and true/false voice transitions and other skills, helping them to experience the vocal characteristics of ethnic music, so as to feel the unique charms of the art of ethnic music. Organize students to participate in local folk music picking activities, collect and organize different performance clips, and make a display wall for vocal training, so that classroom teaching can be closely integrated with local folk music culture. In the process of teaching folk music, we integrate the singing techniques of local folk music, and help students to grasp the characteristics of different vocal cadences through comparative exercises. Establishing ethnic music art pavilions and regularly inviting ethnic musicians from all over the world to carry out lectures and demonstrations on vocal techniques to create a strong artistic atmosphere and inspire students to pass on their passion for traditional ethnic music and art.

2.3 Classroom design for multidimensional perceptual experiences

According to the physical and mental characteristics of students, design multi-level and multi-angle perception and experience activities. In the teaching of folk music, students are guided to perceive the unique tonal characteristics of folk music by means of audio-visual combination. The use of picaresque techniques helps students understand the meaning of the lyrics and organize imitation singing. When teaching folk music, design a group cooperative inquiry session and use artificial intelligence to generate music with folk style, which will be introduced in detail later. Students can analyze the characteristics of the vocal weave of folk music, feel its harmonic effect, and carry out improvisation activities to experience the process of creating polyphonic music. For the teaching of folk music, we design a number of independent tunes to deepen students' understanding of the characteristics of folk music through the reproduction of soundscapes and contextual performances. Innovative teaching methods, such as role-playing and drama, are used to enhance students' emotional experience of the music.

3 Research on Intelligent Ethnic Music Generation Model

In order to generate more high-quality ethnomusic works, this section innovatively constructs an intelligent ethnomusic generation model Tr-MTMG based on Transformer, which is capable of generating music including three instruments, namely, piano, guitar and bass, and it consists

of three parts, namely, the data preprocessing network, the learning network and the generation network.

3.1 Relevant musical knowledge

3.1.1 Basic concepts of music

Music consists of many elements, including pitch, timbre, tempo and other note descriptive information. These elements can be further combined to form the harmony, rhythm, and melody of music, which ultimately constitute a complete musical work. Pitch reflects how high or low the music is in frequency. Tone refers to the quality or sound of the voice or instrument, and tempo refers to the amount of force used to play a note. Different timbres can make a particular piece of music sound distinctive, even if it has the same pitch and intensity. In addition, different players playing the same notes on the same instrument may sound different because of differences in playing technique or in the craftsmanship of the instrument. Two players playing the same note on the same instrument may also sound different because of the way the instruments are played. It is because of these properties that the study of audio signals can be modeled less well because of these nuances, whereas symbolic music can be modeled consistently because it contains only musical information.

3.1.2 MIDI format

The ethnomusicological raw data studied in this paper is in MIDI format, a protocol that allows computers, musical instruments, and other hardware to communicate, and is primarily used to solve the problem of communication standards between the growing number of digital ethnomusicological hardware.Rather than recording sound signals from microphone recordings, MIDI records ethnomusicological performance information. This information includes the pitch of each note, start time, stop time, and other relevant performance information. Recording folk music using MIDI has a number of features. First of all, a single MIDI file is usually only a few kilobytes in size, whereas audio signals require dozens of times the disk footprint to record the same content, and the more compact format allows MIDI files to have faster transmission efficiency and smaller storage space on the Internet, which is one of the important significance of the study of symbolic folk music in this paper. Secondly, MIDI records information about notes rather than specific sounds, which makes it easy to manipulate and modify notes without re-recording folk music. At the same time, one can send the notes recorded by MIDI to different instruments to change the overall sound of folk music, which brings more abundant and convenient ways of using folk music for presentation.

3.2 Transformer model

Recurrent neural networks need to read the inputs sequentially along the order of the input and output sequences, and at each moment the model generates the current moment's hidden layer state h_t by combining the previous hidden layer state h_{t-1} and the input at moment t according to a transition function. This sequential computation of results prevents the model from being computed in parallel, making training long sequences time-consuming and limiting the possibility of large-scale pre-training of the model. Although there are techniques and methods to improve the computational efficiency of recurrent neural networks, the constraint of not being able to compute in parallel remains unsolved.

In order to solve these problems, the Transformer model was created. The structure of the Transformer model is shown in Figure 1.

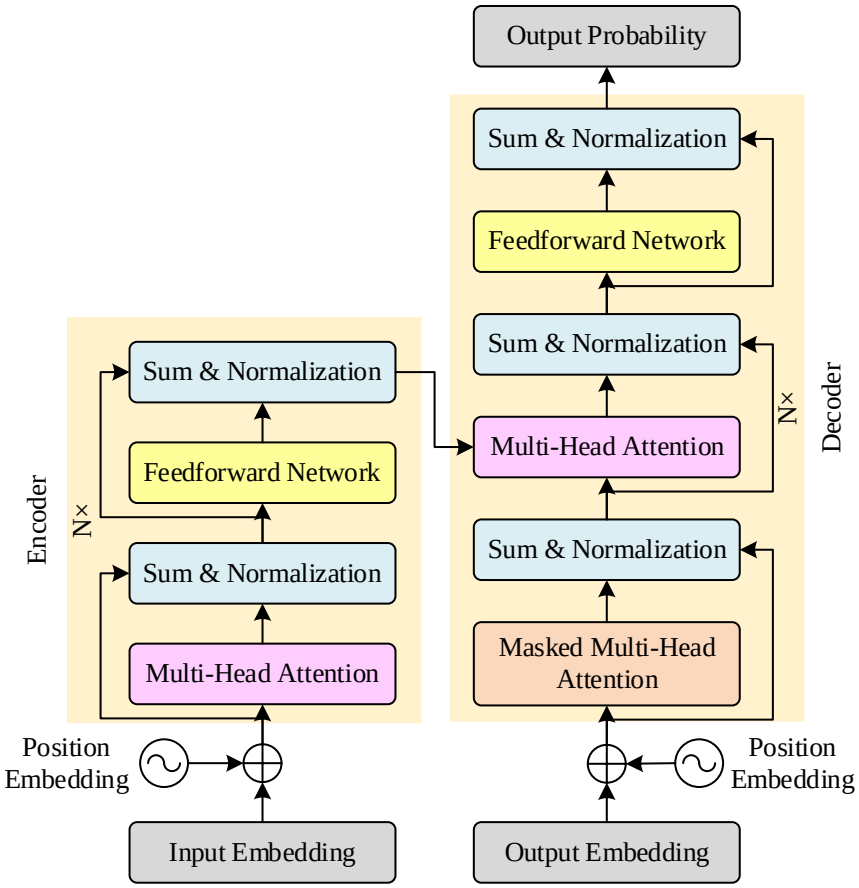

Figure 1: Transformer structure diagram

As can be seen from the figure, the Transformer model is also a classical encoder-decoder structure. The decoder part consists of multiple layers of the same structure stacked on top of each other, and each layer contains multiple submodules. The first part of the sub-module is the multi-head attention, which is then passed into a fully connected feed-forward network layer. Both sub-modules are followed by residual concatenation and layer normalization to ensure that the model still has good performance in the case of multi-layer stacking. The structure of the decoder is similar to that of the encoder, but a mask operation is added to the front-most multi-head self-attention layer, and the subsequent multi-head self-attention no longer only reads the inputs of the model in front of it, but also processes the inputs of the encoder, realizing the interactions between the encoder and the decoder.The actual inputs of the Transformer model are obtained by summing up the word embeddings and the positional embeddings of the inputs. The word embeddings can be directly used in the existing popular Word2Vec and other models of open-source data, or can be directly randomly initialized before the start of training.The Transformer model adds positional encoding at the bottom of the encoder and decoder. The positional encoding has the same word embedding dimension as the word embedding, so the two can be added directly. In the Transformer model, the positional embeddings are computed using sine and cosine functions with different frequencies, and the odd dimensions are computed as shown in Equation (1):

$$PE_{(pos,2i)} = \sin\left(pos / 10000^{2i/d_{model}}\right) \tag{1}$$

The even dimensions are calculated as shown in equation (2):

$$PE_{(pos,2i+1)} = \cos\left(pos / 10000^{2i/d_{model}} \right) \tag{2}$$

where pos is the position of the input word in the original sequence and i is the dimension. That is, each dimension of the position encoding corresponds to a sinusoidal curve. In this computational approach, for any offset k, PE_{pos+k} can be expressed as a linear function of PE_{pos} using the triangular sum-difference product formula, so that the positional embeddings of the added length can also be computed using the above formula when sequences longer than the training data appear.

After processing the inputs, they need to be fed into the self-attention layer. The self-attention layer gets the output based on the query matrix and a set of key-value pair matrix mappings. All three matrix vectors are obtained from the input word embeddings and update their parameters separately during training.

Specifically, the Transformer model uses a scaled dot product attention computation method, and the scaled dot product attention is shown in Fig. 2.

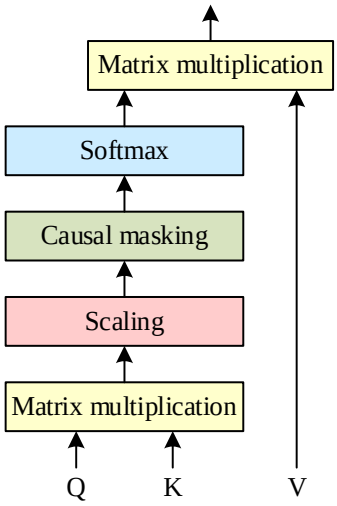

Figure 2: Zoom the dot product attention

The input consists of a query matrix of dimension d_k, a key matrix and a value matrix of dimension d_v, which is computed as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}} \right)V \tag{3}$$

By calculating the dot product of the query matrix and the key matrix, and then dividing the obtained value by $\sqrt{d_k}$, the softmax function is finally used to obtain the weight of the value and multiply it with the value matrix to obtain the actual attention score. Because when the dimension d_k is large, the result of the inner product of the query matrix and the key matrix is large will cause the result of the softmax function to appear in the gradient is small, which affects the updating of the parameters, so in the calculation of the dot product will be divided by $\sqrt{d_k}$ to carry out the scaling of the attention weights.

In the actual training, the Transformer model will splice multiple self-attention layers

horizontally, calculate the attention scores of more than one at the same time each time, and then splice them together to synthesize a single attention score to be passed to the later model:

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\cdots,head_h\right)W^O \tag{4}$$

where $head_i$ is the computation result of each attention head, and W^O is the parameter of the connected linear layer, which is used to adjust the data dimension to facilitate the subsequent computation. Multi-head attention expands the sensory field by allowing each attention layer to use different parameters during training so as to focus on different information, thus enhancing the model's ability to acquire information.

In order to solve the training problem that occurs after stacking multi-layer modules, the Transformer model has a residual connection after calculating the attention scores to ensure that the network only focuses on the part of the current difference, and then uses layer normalization to accelerate the convergence speed of each layer.

In addition to the attention sublayer, each layer in the encoder and decoder contains a fully connected feedforward network. This feedforward network consists of two linear transformations with a ReLU activation function in the middle, computed as shown below:

$$FFN\left(x\right)=\max\left(0,xW_1+b_1\right)W_2+b_2 \tag{5}$$

3.3 Ethnic music generation model based on Tr-MTMG

3.3.1 Data processing and presentation

This paper focuses on the representation and study of three kinds of tracks: piano, guitar and bass, and the MIDI format of folk music can save the basic information of folk music and can be processed. In terms of data set selection, this chapter chooses the Lakh MIDI data set which has a large amount of data and meets the experimental requirements, it contains 176533 multi-track ethnomusicos in MIDI format with labeled instrumental tracks, but the types of ethnomusicos in the data set are messy, so it is necessary to preprocess the data set. In this paper, we use Pretty_midi to filter the dataset according to the labeled instrumental tracks, retaining the folk music in 4/4 time, filtering the folk music with large differences in duration and pattern, and selecting the folk music that contains three instruments at the same time, namely, piano, guitar and bass, according to the instruments. Then, the required three instruments were retained and all other instruments were deleted from the selected folk music, and 34622 multi-track and single-track folk music datasets satisfying the modeling requirements were obtained by dividing the three instruments in the folk music into tracks. Finally, 24636 multi-track ethnomusic songs were used as the training set and 10000 multi-track ethnomusic songs were used as the test set. In terms of data representation, firstly, three kinds of note features of mono-track folk music, namely pitch, onset time and duration, are extracted in this chapter, and the extracted features are expressed in the form of text and transformed into text sequences by using the method of Notewise57. These notes can be represented by events, which may occur randomly at a certain moment, and multiple notes can be played at the same moment. The information about these events can be recorded using “Notewise”.

3.3.2 Network modeling

(1) Learning network

The framework of the Tr-MTMG network model is shown in Fig. 3. The learning network consists of 6 Encoding layer sub-networks, each of which can learn the audio tracks from real

samples by two-by-two interaction. Each layer contains six Encoding modules, and the Cross-track attention mechanism is an important part of the Encoding module. The difference between Self-attention mechanism and Cross-track attention mechanism is that Self-attention mechanism mainly learns the information of the sequence itself, while Cross-track attention mechanism mainly learns the information of different sequences. Therefore, in the case of learning between different tracks, this paper selects the Cross-track attention mechanism for learning between tracks.

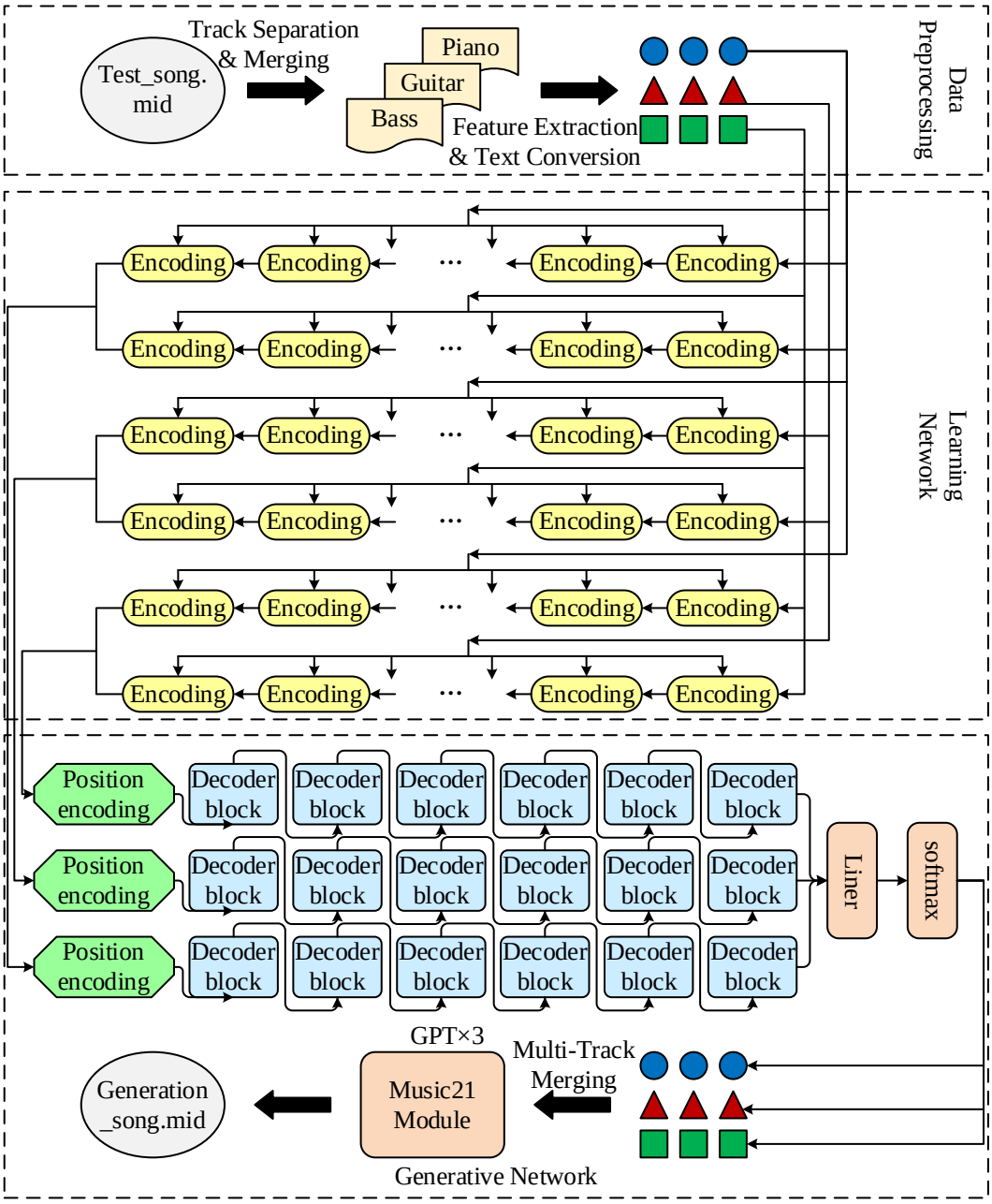

Figure 3: The Tr-MTMG network model framework

In this section, we take the piano track and the guitar track as an example, the piano track is treated as a learning sequence and the guitar track is treated as a learned sequence, and the Encoding structure is shown in Figure 4. The piano sequence and guitar sequence are denoted

by $X_p \in R^{T_p \times d_p}$, $X_g \in R^{T_g \times d_g}$, respectively. $T_{(\cdot)}$ is denoted as the length of the sequence, and $d_{(\cdot)}$ denotes the dimension of the feature.

In the sequence learning process, the Cross-track attention mechanism treats the query as the dot product of the inputs of the learning target and the input transformation matrix, i.e., $Q_p = X_p W_{Q_p}$, and the key-value pairs as the dot product of the inputs of the learned target and the key-value pair transformation matrix, i.e., $K_g = X_g W_{K_g}$ and $V_g = X_g W_{V_g}$, where $W_{Q_p} \in R^{d_p \times d_k}$ are the transformation matrix weights of the piano input sequence query, and $W_{K_g} \in R^{d_g \times d_k}$ and $W_{V_g} \in R^{d_g \times d_v}$ are the transformation matrix weights of the key-value pairs of the guitar input sequence, respectively. That is, the value of Cross-track attention for learning guitar sequence track information from piano sequence can be expressed as Equation (6):

$$Z_{p \to g} = CT_{p \to g} attention\left(X_p, X_g\right) = soft\max\left(\frac{Q_p K_g^T}{\sqrt{d_k}}\right) V_g \tag{6}$$

where $Z_{p \to g}$ is one of the Multihead Cross-track attention mechanisms, as shown in Eq. (8), which can be regarded as $head_h$. Then, h Cross-track attention values are spliced, and then linear activation is applied to them, as shown in Eq. (7), to obtain the Multihead Cross-track attention value $Multihead(h)$:

$$Multihead\left(h\right) = W\left[head_1; head_2; \cdots; head_h\right] \tag{7}$$

$$head_h = CTattention\left(Q_p, K_g, V_g\right) \tag{8}$$

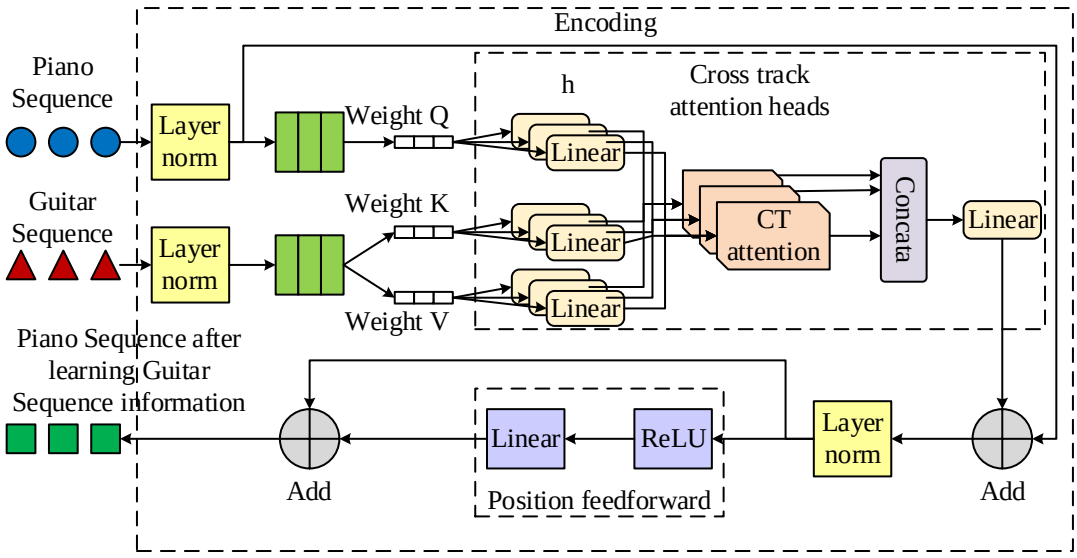

Figure 4: Encoding structure

From the Tr-MTMG network framework shown in Fig. 4, it can be learned that the three tracks need six Encoding layers to learn from each other, and each layer contains six Encoding modules, and when the piano sequence is the source sequence, and the guitar sequence and the

bass sequence are the target sequences, the model needs to have to have two Encoding layer subnetworks, and one layer is used for the piano sequence to learn the information of the guitar sequence, as shown in Eq. (9), and the model needs to have two encoding layer subnetworks. as in Eq. (9). One layer is used for the piano sequence to learn the information of the bass sequence, as in Equation (10). Where $\hat{Z}_{p\to g}^{[i]}$ is the piano sequence after the piano sequence learns the guitar sequence, and the piano sequence is output after passing through the Cross-track attention mechanism of the i-layer polytope, $i=\{1,2,3,4,5,6\}$.

$$\hat{Z}_{p\to g}^{[i]} = CT_{p\to g}^{[i]}\left(LN\left(Z_{p\to g}^{[i-1]}\right), LN\left(Z_{p}^{[0]}\right)\right) + LN\left(Z_{p\to g}^{[i-1]}\right) \tag{9}$$

$$\hat{Z}_{p\to b}^{[i]} = CT_{p\to b}^{[i]}\left(LN\left(Z_{p\to b}^{[i-1]}\right), LN\left(Z_{p}^{[0]}\right)\right) + LN\left(Z_{p\to b}^{[i-1]}\right) \tag{10}$$

After obtaining the value of the multi-head Cross-track attention sequence through the formula, then let the output sequence go through the layer normalization to get the output sequence with the same dimension as the input sequence, and use the obtained output sequence as the input of the feed-forward sub-layer, after the feed-forward sub-layer and then the output sequence with the same dimension as the input sequence through the residual connection, so as to obtain the sequences $Z_{p\to g}^{[i]}$ and $Z_{p\to b}^{[i]}$, which are the output sequences after encoding in i layers. They are Eq. (11) and Eq. (12), respectively.

$$Z_{p\to g}^{[i]} = f_{p\to g}^{[i]}\left(LN\left(\hat{Z}_{p\to g}^{[i]}\right)\right) + LN\left(\hat{Z}_{p\to g}^{[i]}\right) \tag{11}$$

$$Z_{p\to b}^{[i]} = f_{p\to b}^{[i]}\left(LN\left(\hat{Z}_{p\to b}^{[i]}\right)\right) + LN\left(\hat{Z}_{p\to b}^{[i]}\right) \tag{12}$$

After obtaining two output sequences $Z_{p\to g}^{[i]}$ and $Z_{p\to b}^{[i]}$ for learning other tracks, they are spliced as in Eq. (13), and finally output a one-dimensional piano sequence Z_p that contains information about both the guitar sequence and the bass sequence. The guitar sequence Z_g, the bass sequence Z_b are similar to the piano sequence Z_p.

$$Z_p = Concat\left(Z_{p\to g}^{[i]}, Z_{p\to b}^{[i]}\right) \tag{13}$$

(2) Generating the network

The GPT architecture is shown in Fig. 5. It consists of an embedding layer, six decoder modules and a linear Softmax layer, each decoder module consists of eight 256-dimensional Self-attention layers and 1024-dimensional feed-forward sub-layers, which is able to predict the next moment state based on the previous state.GPT not only increases the dependency between sequence contexts, but also solves the complexity problems arising from recurrent neural networks, long and short-term memory GPT not only increases the dependency between sequence contexts, but also solves the complexity problems of recurrent neural networks, long and short-term memory networks, etc. Another important part of the decoding module of GPT is the masked multi-head attention mechanism, its function is that when learning the above information, each character only pays attention to its own current character and all the characters predicted before it, so that the model can predict the next moment state based on the previous state. The role of positional encoding is to superimpose a fixed vector to each input

word vector to represent its positional tag, and output the corresponding attention value of the sequence after 6 layers of decoder module, and through the linear layer and Softmax layer, to get the next moment of the character state.

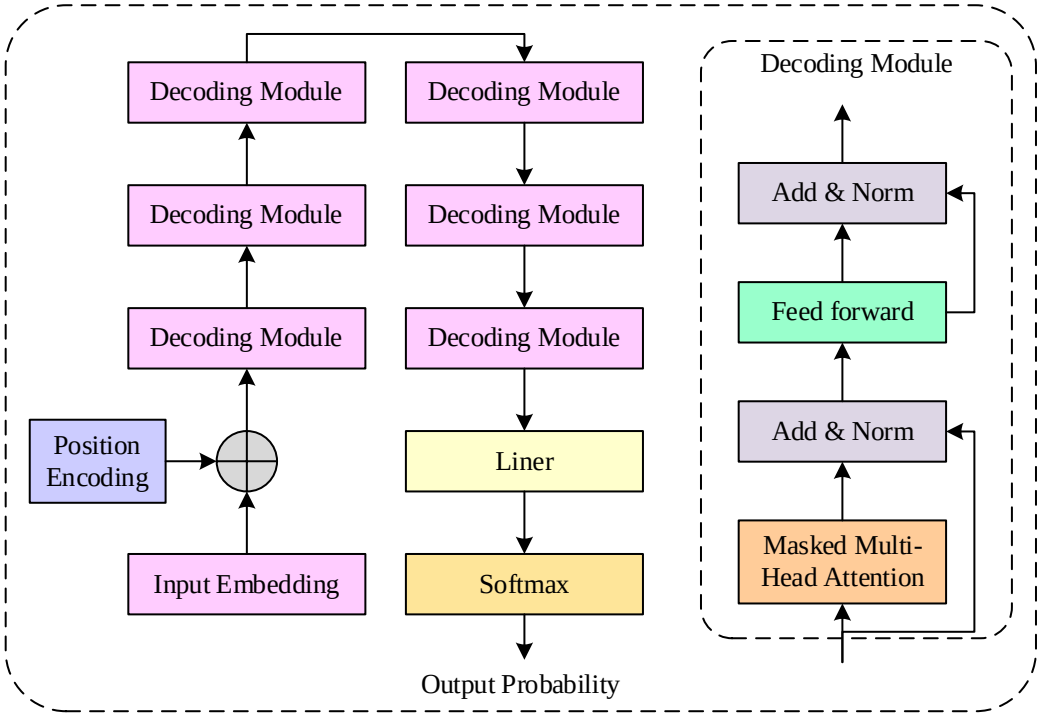

Figure 5: GPT Architecture

4 Experimentation and Evaluation of Intelligent Ethnic Music Generation

4.1 Model performance evaluation experiments

4.1.1 Experimental setup

The codes of the experiments were all realized through python language, and all models were built by Pytorch for deep learning models, and finally trained and tested on NVIDIA GeForce RTX 2080 Ti graphics card. For the preprocessing of the ethnomusic data and the code of some objective evaluation indexes, the open source toolkit MusPyss for symbolic ethnomusic generation was used, which provides the necessary tools for the development of the ethnomusic generation system, including dataset management, data 1/O, data preprocessing, and model evaluation, as well as interfaces to common symbolic ethnomusic formats and other symbolic ethnomusic libraries. In addition, MusPy can be used for the implementation of common ethnomusic data representations for ethnomusic generation, including pitch-based, event-based, piano-roll-based and annotation-based data representations.

(1) Dataset

Three publicly available datasets, Lakh MIDI, LPD-5-cleansed, and Free MIDI, were chosen for the experiments. All three datasets are popular training datasets in the field of multi-track ethnic music generation today. After the style matching operation of data preprocessing, the unsuccessful tracks are filtered out, and then the tracks containing empty tracks or tracks with less than 20 notes are filtered out. The final statistics of the three datasets are shown in

Table 1.

Table 1: Experimental data statistics

Dataset	Scale	Duration	Average duration
Lakh MIDI	1465 songs	100 hours	4.12 Minute/songs
LPD-5-cleansed	698 songs	46hours	3.95 Minute/songs
Free MIDI	720 songs	51 hours	4.16 Minute/songs

(2) Objective evaluation indexes

The four objective evaluation metrics selected in this paper include tone level histogram entropy (PCHE), rhythmic pattern similarity (GPS), self-similarity matrix (SSM), and structural index (SI). All four evaluation indexes are closer to the true value representing the higher quality of the generated folk music.

(3) Comparison Model

In this paper, three excellent multi-track ethnomusic generation models are chosen for comparison, including MuseGAN, MIDI-Sandwich2 and Museformer.

(4) Model implementation details

The embedding layer dimensions and sampling strategy settings for each token are shown in Table 2. For the three comparison models, MuseGAN, MIDI-Sandwich2, and Museformer, the data representation, network architecture, and various parameter settings of the original authors were followed.

Table 2: The embedding layer dimension and sampling strategy Settings for each token

Token type	Number of tokens	Dimension of the embedding layer	Sampling strategy	
			τ	p
Type	5(+1)	36	1.0	0.8
Bar/Beat	16(+1)	68	1.4	1.0
Chord	26(+1)	68	1.0	0.8
Track	5(+1)	36	1.0	0.8
Pitch	130(+1)	512	1.0	0.8
Duration	16(+1)	130	2.0	0.8
Velocity	130(+1)	130	6.0	1.0
Genre	6(+1)	32	1.0	1.0

4.1.2 Model performance evaluation

Three multi-track ethnic music generation models, MuseGAN, MIDI-Sandwich2, and Museformer, were selected as comparison models in the model evaluation experiments. And three public datasets, Lakh MIDI, LPD-5-cleansed, and Free MIDI, were used for training and testing.

The experiment randomly selected 60 real ethnomusic samples in the selected dataset as a comparison, and compared the score results with the real ethnomusic samples. The experiment uses the evaluation indexes introduced above for all the generated samples and real folk music samples, in which the structural indexes take three kinds of intervals SJ_3^8, SJ_8^{15}, SJ_{15}, and SJ_{15}, which are used for evaluating the short, medium, and long structures of the folk music samples respectively, and the results for the Lakh MIDI, LPD- 5-cleansed, and Free MIDI datasets, respectively, are shown in Tables 3 to 5, where H stands for tone level histogram entropy, gs stands for rhythmic pattern similarity, and SJ stands for structural metrics.

From the experimental results in the tables, it can be seen that the scores of the Tr-MTMG network model proposed in this paper are optimal in several metrics in the three datasets, which indicates the feasibility and superiority of the model. Specifically, for the level histogram entropy, compared with MuseGAN, MIDI-Sandwich2 and Museformer models, the level histogram entropy of this paper's model in the Lakh MIDI dataset is 2.8504, with a score closest to the real samples, indicating that the model based on this paper is able to generate more harmonic and realistic music. Overall, the Tr-MTMG network model proposed in this paper is significantly ahead of the other three music generation models in all three time scales and still performs well in the generation of long repetitive segments,which indicates that the Tr-MTMG network model has a better ability to learn the structure of the music,and can generate higher-quality multi-track ethnomusicology music.

Table 3: Experimental results of performance evaluation of Lakh MIDI dataset model

Indicator	G	gs	SJ_3^8	SJ_8^{15}	SJ_{15}
MuseGAN	3.6852	0.6195	0.2612	0.2677	0.2146
MIDI-Sandwich2	3.394	0.665	0.32	0.2463	0.1035
Museformer	2.5242	0.7981	0.3494	0.3053	0.2247
Ours	2.8504	0.8186	0.4096	0.4199	0.3371
Real Data	2.9875	0.8306	0.434	0.4428	0.4366

Table 4: Experimental results of performance evaluation

Indicator	H	gs	SJ_3^8	SJ_8^{15}	SJ_{15}
MuseGAN	3.4218	0.6642	0.275	0.2077	0.1888
MIDI-Sandwich2	3.6468	0.6215	0.2818	0.2277	0.1439
Museformer	2.5355	0.8478	0.3136	0.2881	0.2078
Ours	2.7066	0.8117	0.4057	0.3631	0.3096
Real Data	2.8922	0.8273	0.4514	0.4356	0.428

Table 5: Experimental results of performance evaluation of the Free MIDI dataset

Indicator	H	gs	SJ_3^8	SJ_8^{15}	SJ_{15}
MuseGAN	3.6875	0.6304	0.2746	0.2599	0.1494
MIDI-Sandwich2	3.3226	0.6495	0.3011	0.1958	0.1027
Museformer	2.7974	0.789	0.3672	0.2501	0.1828
Ours	2.9367	0.7902	0.4073	0.3511	0.3083
Real Data	2.7013	0.8537	0.4653	0.4434	0.4587

4.2 Subjective and objective assessment experiments

After the automatic generation of music, the evaluation of the generated music is especially important. In this paper, we use the combination of objective and subjective evaluation to comprehensively evaluate the generated music.

4.2.1 Objective assessment indicators

In this paper, three checkpoints with loss equal to 0.1, 0.25, and 0.8 were selected to evaluate the music generated by the Tr-MTMG network model (Model1) as well as the original model (Model2) in terms of objective metrics to reflect the quality of music generated by these two models from the evaluated metrics. The average experimental results of the two groups of

models are shown in Table 6. Comparison between the two models, it can be found in the table that the music generated using the model proposed in this paper has higher scale consistency and rhythmic interval similarity as well as lower pitch class entropy than using the original model, which indicates that the music generated using the model in this paper has better rhythm and melody.

Table 6: The average experimental results of the two groups of models

Loss		0.8	0.25	0.1	Database
Number of notes	Model 1	35	30	30	30
	Model 2	36	32	29	
Number of note types	Model 1	12	12	8	8
	Model 2	12	12	8	
Scale consistency	Model 1	1.003	1.222	0.84	0.966
	Model 2	0.827	1.076	0.785	
Pitch class entropy	Model 1	3	2.71	2.765	2.685
	Model 2	2.923	2.656	2.762	
Air beat rate(%)	Model 1	0.45	0.31	0.16	0.75
	Model 2	0.24	0.29	0.08	
Rhythm interval similarity	Model 1	0.9533	0.9813	0.9931	0.985
	Model 2	0.9492	0.9878	0.9804	

The number of notes and note classes of the two models are shown in Table 7. In the table, it can be found that the music generated by using the model of this paper is more stable in terms of the number of notes and the number of note classes than the original model, and it will not generate songs with great differences, which reflects that the model of this paper has stronger stability and less error rate.

Table 7: The number of notes and the number of note classes of the two models

Indicator		Mean	Standard deviation	Maximum value	Minimum value
Number of notes	Model 1	35.1	6.2	51	26
	Model 2	35.8	7.8	56	20
Number of note types	Model 1	9.2	0.7	15	15
	Model 2	9.8	1.3	6	6

In order to verify the quality and performance of this paper's model more comprehensively, this paper compares the MiDiNet model based on generative adversarial network and the MusicVAE model based on dual-layer encoder using six objective evaluation metrics.The MiDiNet model and the MusicVAE model each generates 60 pieces of music of 32 bars, and the evaluation results of the three models with the database are shown in Table 8. It can be found through the table: using this paper's model is much better than MiDiNet and MusicVAE in the evaluation results of the generated music, this point verifies that the use of this paper's model in the music generation task, compared with the use of generative adversarial network and dual-layer encoder based on a certain degree of disparity, the use of this paper's model for the generation of music has a clear advantage.

Table 8: The evaluation results of three models and databases

Indicator name	Ours	MiDiNet	MusicVAE	Database
Number of notes	36	30	29	36
The number of note types	9	7	7	9
Scale consistency	0.927	0.947	0.877	0.99
Pitch class entropy	2.707	3.072	3.142	2.707
Aerial photography rate (%)	0.14	0.85	0.89	0.77
Rhythm interval similarity	0.9235	0.9168	0.8661	0.988

4.2.2 Music continuation

This task is mainly used to verify the integrity of the music produced by the model, the specific method is to give the model a MIDI initial excitation as the tone of the beginning of the piano song, which is about 15 seconds, and then use this paper to improve the model as well as the original model based on the given law of the tone, to continue to write the music of the 32 bars, and then this paper uses the evaluation index to compare the two music, using the two kinds of models were respectively renewed 60 songs, and the comparison results of the renewed music are shown in Table 9. From the table, it can be concluded that for the incentives given by the model beforehand, the music renewed by the model in this paper is closer to the initial incentives in terms of the number of notes and the number of note classes than the original model, because the initial incentives given by the initial incentives use fewer number of notes and the number of note classes and the overall tone is more calm, so the model proposed in this paper also uses fewer notes and note classes to complete the renewal of the music.

Table 9: Continue the comparison results of the music

Indicator name	Ours	Transformer-XL	Initial incentive
Number of notes	20	30	9
The number of note types	8	12	3
Scale consistency	1.066	0.895	0.981
Pitch class entropy	2.531	2.703	2.354
Aerial photography rate (%)	0.097	0.072	0.166
Rhythm interval similarity	1.001	1.008	0.995

The number of notes and the number of note classes of the renewed music are shown in Fig. 6. In the figure, this paper gives the distribution of the number of notes and the number of note classes of the renewed music of the two models, the horizontal coordinate is the 60 songs involved in the validation, and the vertical coordinate is the number of notes and the number of note classes. Using this paper's model compared to the original model, the renewed music has less volatility in the number of notes and note classes, and this paper's model plays more stable when music renewal is performed.

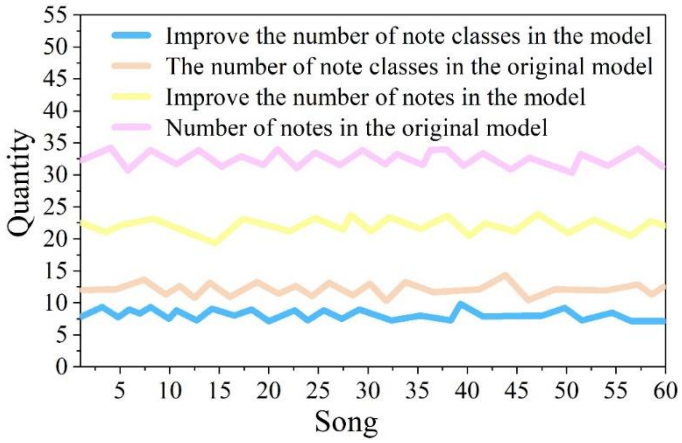

Figure 6: The number of notes and the number of notes

4.2.3 Subjective assessment of generated music

(1) On-site scoring

Following the steps described above, the scoring was weighted according to the formula, and the scoring results are shown in Fig. 7. When counting the scores, this experiment arranges the disrupted music in sequential intervals. The first of them is the music in the database, the second is the music generated with the original model,the third is the music generated with the model of this paper,and so on after that. Each type of music is 10 songs.

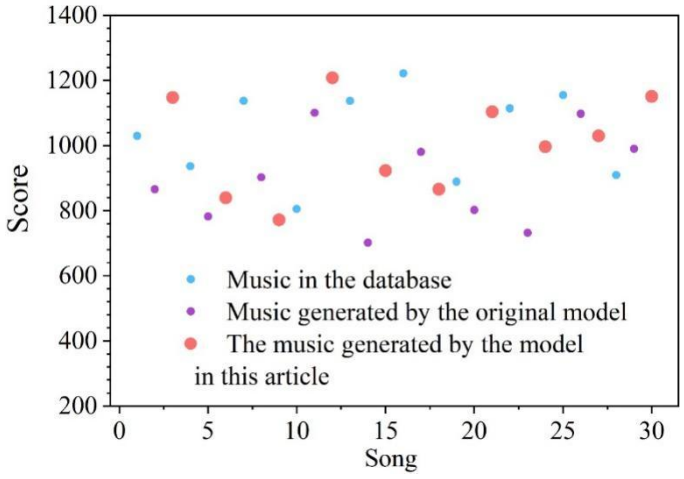

Figure 7: Scoring results

From the figure, we can get the ranking of the 5 music songs with the highest evaluation scores, and the ranking is shown in Table 10. From the table, it can be seen that 2 of the 5 music songs with the highest scores were generated using the model proposed in this paper, and the scores they achieved are close to those of the songs in the database, which indicates that the two songs have reached the level of human work songs, which are recognized by music majors and non-music majors, and it verifies that the model proposed in this paper has the ability of generating high-quality music.

Table 10: Score ranking

Ranking	Source of the song	Song Serial Number	Song score
1	Within the database	16	1221.64686
2	This model	12	1208.03622
3	Within the database	25	1154.69723
4	This model	30	1151.38653
5	Within the database	11	1100.99038

In addition to listing the five songs with the highest scores, this paper developed criteria for excellence and passing rates to continue analyzing the music. The processing results are shown in Table 11. The excellence rate of ethnic music generation based on the model of this paper reaches 40%, which is higher than the original model by 15%.

Table 11: Processing of scoring results

Music type	Average value	Standard deviation	Excellence rate(%)	Pass rate(%)
Database	974.58	142.57	55	90
This model	934.9	144.5	40	85
Original model	834.07	151.01	25	60

In order to follow up the in-depth study, this paper lists the specific scoring contents of the testers on the 30 songs, and the comparison of the scoring contents is shown in Table 12. Among the songs, the melody and rhythm scoring range is 1-3 points, and the overall structure scoring is 0-3 points. Through the on-site scoring experiment, we can see that the music majors and non-music majors affirm the model proposed in this paper, and the excellent score of the overall structure reaches 2.74, which verifies the superiority of the quality of the music generated by the model in this paper.

Table 12: Comparison of the content of the score

	Melody	Rhythm	Overall structure
Worse	1.94	1.37	2.28
Qualified	2.34	1.934	2.554
Excellence	2.74	2.664	2.74

(2) Music “Turing Test” Experiments

The summary table of the “Turing test” is shown in Table 13, and the 10 pieces of music are ranked according to “composition method” and “name”. The scores shown in the table make it easy to compare the quality of each piece of music, and the manual judgment of each piece of music is also shown, so that the results of the Turing Test can be directly expressed by comparing them with the original compositional style. MIDI3 composed by the model also achieved a high score of 105.65, while MIDI2 and MIDI10 were both considered to be artificially composed, and MIDI9, although artificially composed, was still considered by most teachers to be composed by the model in this paper, which shows that the quality of model-composed music is not necessarily poor, and that good-quality music can be recognized by teachers, and that there are also differences in the quality of artificially composed music. The quality of artificially created music varies and is not uniformly high.

Table 13: Information Summary

Number	Name	Score/artificial judgment (majority)	Composition method
1	MIDI 1	103.32/Artificial	Man-made creation
2	MIDI 4	96.66/Artificial	Man-made creation
3	MIDI 5	86.35/Artificial	Man-made creation
4	MIDI 8	109.54/Artificial	Man-made creation
5	MIDI 9	76.54/Model	Man-made creation
6	MIDI 2	99.01/Artificial	Model creation
7	MIDI 3	105.65/Artificial	Model creation
8	MIDI 6	64.25/Model	Model creation
9	MIDI 7	81.26/Model	Model creation
10	MIDI 10	93.22/Artificial	Model creation

In this paper, we also categorize the 10 pieces of music into two types according to the compositional method and in the order of scoring scores, and the music scores are compared as shown in Table 14. We can more clearly observe the difference in scores between model compositions and human compositions.

Table 14: Music score comparison

	1	2	3	4	5
Man-made creation	74.47	90.36	98.30	101.39	109.34
Model generation	59.59	80.00	89.30	95.18	102.44

Through the objective and subjective evaluation experiments, both of them verified that the quality of the ethnic music generated by using this paper's model is much higher than that generated by the original model, and the music generated by using this paper's model is more likely to be loved by the testers, and some of the music can even reach the level of fake to real. The objective and subjective evaluation results consistently show the effectiveness of this model, which confirms that this model can be well used in music generation tasks. At the same time, the feasibility of this model in the music course of semester education in colleges and universities is verified, so as to realize the inheritance of national music culture.

5 Conclusion

Ethnic music, as an important part of the music field, not only reflects the unique cultural connotations and aesthetic concepts of various ethnic groups, but also is a vivid embodiment of the national spirit. The inheritance and development of ethnic music should be highly emphasized in the music curriculum of preschool education in colleges and universities in order to enhance students' sense of national identity and pride. The article proposes an intelligent ethnic music generation model (Tr-MTMG model) and verifies the superiority of the model proposed in this paper through a series of experiments. The experimental conclusions drawn in this paper are as follows:

(1) The results on three datasets, Lakh MIDI, LPD-5-cleansed and FreeMIDI, show that the MIDI-XL model proposed in this paper achieves better results in a number of metrics compared to the three models, such as MuseGAN, etc., and the entropy value of the histogram of the level of this paper's model in the Lakh MIDI dataset is 2.8504, which is the score The score is closest to the real sample, which proves that the model based on this paper can generate higher quality music works.

(2) By organizing students' scoring, it is found that the overall structure of the music generated by the model proposed in this paper reaches 2.74 points by both music majors and non-music majors, which verifies the superiority of the quality of the music generated by the model in this paper.

In summary, the method proposed in this paper can generate ethnic music well, which not only helps students deeply appreciate the unique charm of ethnic culture, but also enhances their sense of identification with ethnic music. During the teaching period, teachers should deeply recognize the main problems in the current teaching of folk music and take effective improvement measures, not only to strengthen their own professionalism and correct the wrong concepts of education, but also to introduce diversified teaching methods and enrich the teaching content, so as to stimulate the students' interest in learning and to guide them to understand the valuable spiritual treasures contained in folk music. At the same time, teachers also need to actively integrate the ethnic music resources of various regions to broaden students' knowledge and deepen their understanding of ethnic cultural diversity. This will help realize the inheritance and development of ethnic music culture.

About the Author

Na Li graduated from Wuhan University in 2009 with a master's degree. Currently, she works at Jingchu University of Technology. Her research direction is music education.

Funding

Project Title: 2023 Philosophy and Social Sciences Research Project of the Department of Education of Hubei Province Research on the Reconstruction of High-Quality Preschool Education Teacher Training System Against the Background of Low Fertility Rate (Project No. 23D087)

References

[1] Muzyka, O., Lopatiuk, Y., Belinska, T., Belozerskaya, A., & Shvets, I. (2021). Modern aesthetic education and its further directions. Linguistics and Culture Review, 5(S4), 12-21.

[2] Fox, D. B. (2021). The musical education of early childhood majors: All God's critters got a place in the choir. Visions of Research in Music Education, 16(4), 7.

[3] Bautista, A., Yeung, J., Mclaren, M. L., & Ilari, B. (2024). Music in early childhood teacher education: Raising awareness of a worrisome reality and proposing strategies to move forward. Arts Education Policy Review, 125(3), 139-149.

[4] Bakken, L., Brown, N., & Downing, B. (2017). Early childhood education: The long-term benefits. Journal of research in Childhood Education, 31(2), 255-269.

[5] Law, W. W., & Ho, W. C. (2011). Music education in China: In search of social harmony and Chinese nationalism. British Journal of Music Education, 28(3), 371-388.

[6] Oliver-Barcelo, M., Ferrer-Ribot, M., & Jové, G. (2024). Arts education in early

childhood teacher training: An international analysis. Teaching and Teacher Education, 148, 104703.

[7] Brown, C. S., Cheddie, T. N., Horry, L. F., & Monk, J. E. (2017). Training to Be an Early Childhood Professional: Teacher Candidates' Perceptions about Their Education and Training. Journal of Education and Training Studies, 5(6), 177-186.

[8] Garvis, S., & Pendergast, D. (2011). An investigation of early childhood teacher self-efficacy beliefs in the teaching of arts education. International journal of education & the arts, 12(9), 1-15.

[9] Twigg, D., & Garvis, S. (2010). Exploring art in early childhood education. International Journal of Arts in Society.

[10] Hayes, N., Maguire, J., & O'Sullivan, C. (2021). Professional development in arts education for early childhood education: A creative exchange model. International Journal of Early Childhood, 53(2), 159-174.

[11] Leung, S. K., Wu, J., & Ho, T. H. (2025). Early childhood visual arts education: Teachers' content knowledge, pedagogical content knowledge, and challenges. The Asia-Pacific Education Researcher, 34(1), 351-363.

[12] Garvis, S. (2012). A Self-Study in Teacher Education: Learning to Teach in Higher Education after Teaching the Arts to Young Children. Online Submission.

[13] Zhang, Q., Wu, W., Jiang, K., & Shan, C. (2024). The arts in early childhood teacher education in China: a question of curriculum balance. Asia-Pacific Journal of Teacher Education, 52(1), 47-63.

[14] Hyoung-Jai, K., Min-Seo, S., & Soon-Ohk, H. (2016). The development of a program model of art convergence personality education for early childhood pre-service teacher. Advanced Science Letters, 22(11), 3426-3431.

[15] Lee, L., & Lin, C. H. (2023). Digital and Traditional Learning: Learning Styles with Music and Technology for Early Childhood Education. Engineering Proceedings, 38(1), 19.

[16] Barrett, M. S., Flynn, L. M., Brown, J. E., & Welch, G. F. (2019). Beliefs and values about music in early childhood education and care: Perspectives from practitioners. Frontiers in psychology, 10, 724.

[17] Jeremić, B., Gordić, S., Trbojević, A., Vujaković, F. J., Tubić, M., & Savić, M. V. (2025). Educators' Views on the Impact of Traditional Music on the Socio-Emotional Competencies of Preschool-Aged Children. International Journal of Cognitive Research in Science, Engineering and Education, 13(1), 83-95.

[18] Jeremić, B., Markov, Z., & Nikolić, L. (2023). Traditional music in preschool and the development of social-emotional competencies of preschool children. Društvene i humanističke studije, 8(2 (23)), 541-556.

[19] Young, S. (2016). Early childhood music education research: An overview. Research

Studies in Music Education, 38(1), 9-21.