


## Article

# Double Linear Transformer for Background Music Generation from Videos

Xueting Yang , Ying Yu \* and Xiaoyu Wu

Faculty of Information and Communication Engineering, Communication University of China, Beijing 100024, China; yangxueting@cuc.edu.cn (X.Y.); wuxiaoyu@cuc.edu.cn (X.W.)

\* Correspondence: yuying@cuc.edu.cn; Tel.: +86-10-6577-9427

**Abstract:** Many music generation research works have achieved effective performance, while rarely combining music with given videos. We propose a model with two linear Transformers to generate background music according to a given video. To enhance the melodic quality of the generated music, we firstly input note-related and rhythm-related music features separately into each Transformer network. In particular, we pay attention to the connection and the independence of music features. Then, in order to generate the music that matches the given video, the current state-of-the-art cross-modal inference method is set up to establish the relationship between visual mode and sound mode. Subjective and objective experiment indicate that the generated background music matches the video well and is also melodious.

**Keywords:** video background music generation; music feature extraction; linear Transformer



**Citation:** Yang, X.; Yu, Y.; Wu, X. Double Linear Transformer for Background Music Generation from Videos. *Appl. Sci.* **2022**, *12*, 5050. <https://doi.org/10.3390/app12105050>

Academic Editors: Katia Lida Kermanidis, Phivos Mylonas and Manolis Maragoudakis

Received: 22 April 2022

Accepted: 13 May 2022

Published: 17 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Music can effectively convey information and express emotions. Compared with silent videos, an appropriate background music can make the video content easier to understand and accept. However, in daily life, generating video soundtrack is often a technical and time-consuming work. It requires the selection of suitable music from a large amount of music and needs people capable of using specific tools to edit the corresponding audio paragraphs. Furthermore, the existing methods cannot automatically customize the appropriate background music for the given video. To address these problems, this paper proposes an automatic background music generation model with two linear Transformers training jointly. This method ensures the convenience in use as well as the music uniqueness. At the same time, after a large amount of data training, it ensures both the rhythmicity of the generated music and a high degree of matching with the given video.

For the tasks related to the automatic generation of video background music, there have been many excellent achievements, such as music generation and video-audio matching tasks. However, as far as we know, the combination of generated music and video associations has not been considered for most of the existing works. Many works on music generation focus on music generation itself [1,2], and recently, more studies have paid attention to controllable music generation [3–5], while seldom [6] combining music generation with videos. As a result, the generating music cannot meet the background requirement for a given video. Furthermore, since there is no paired video-background music dataset, the existing video background music generation methods [6] skillfully established the corresponding relationship between video features and music elements, and then used the video features to change the music elements for different given videos. Although these approaches have achieved breakthrough results, they have paid less attention to the relationship and the independence of musical elements, which has led to a weak melodiousness. In this article, the proposed model improves the extraction of musical elements with two linear Transformers [7] training jointly and using the above inference method to improve the rhythm of the generated music as well as matching the given video.

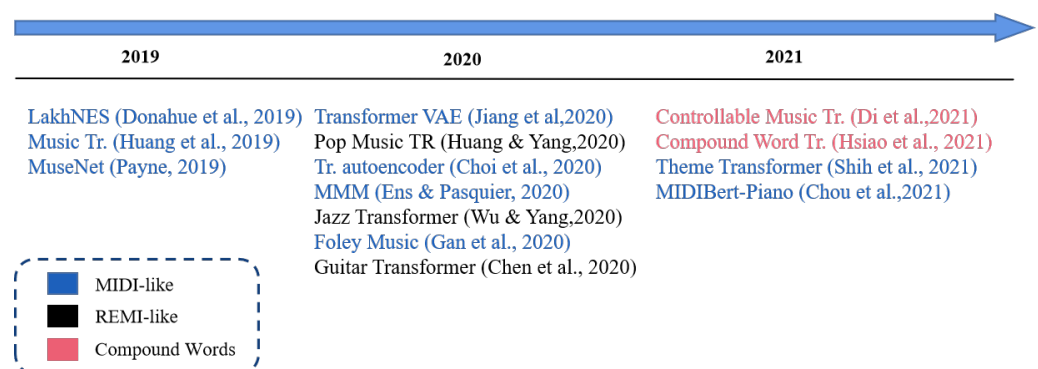
The compound word Transformer [8] was used to represent music and to consider the following factors in learning musical features part: beat, density, strength, pitch, instrument type and duration. The first three factors are related to rhythm and the last three are related to notes and using types to distinguish the grouping of factors. Compared to RNNs, the Transformer [9], a multihead model based on a self-attention network is more effective and explainable in time-based long sequence tasks such as BERT [10] and so on. The proposed model used two linear Transformers [7] with a time position to train the rhythm and note-related features separately. Taking into account the correlation between the various musical elements, the proposed model adjusted the complementary features of the two groups and trained the network in a joint way. The Lakh Pianoroll Dataset [3] was used in the training and inference stages. Furthermore, the background music was specified by using the specific features of the video. To sum up, our work has the following contributions:

1. In this paper, we propose a video background music generation model. Most background music generation works use only one Transformer to learn and extract all music elements, which leads to a weak melody. To establish the correlation and independence between rhythm-related and note-related music features, we use two linear Transformers training jointly. In particular, the two kinds of music features are put into each Transformer network separately;
2. Compared to an RNN network such as LSTM and GRU, the proposed model uses a linear Transformer, considering its lightweight and linear complexity. We use the timing information of the given video to guide the generation of music by adding a time beat encoding to the Transformer network;
3. After the model has learned the music features, we replace the density and strength features of the music with the optical flow information and rhythm of the video in the inference step, inspired by the state-of-the-art music video inference method. The proposed model combines the music feature learning step and the inference step to form a complete video background music generation model.

## 2. Related Works

### 2.1. Representation of Music

In earlier research works, music was represented primarily by MIDI-like interfaces [11], using time-shift to mark time intervals. Another representation, REMI [12], on the other hand, provides special markups for bars, chords, beat and tempo, and also uses a different way of marking time intervals than MIDI. The representation of music in some existing works can be visually seen in Figure 1 below. Compared with MIDI, REMI can better control the music structure and adjust the rhythm of the music. However, these two methods do not group the tokens according to their types.



**Figure 1.** Ways to represent music in recent studies.

In this paper, we use the compound word Transformer [8] to represent the music. Based on REMI, the compound word (CP) Transformer converts a long sequence of tokens into a composite word by grouping adjacent symbols, while filling in the missing tokens

in each time block, making each step consistent. When using the compound word (CP) Transformer to represent the music, music tokens are divided into rhythm and melody groups by referring into the Controllable Music Transformer model [6], which is helpful to consider both local classifications and overall relationship when extracting music elements.

## 2.2. Music Generation

**Music Composition.** Music composition is a challenging job for both composers and algorithms. Research works on music generation firstly focused on music generation itself, then more studies paid attention to controllable music generation. However, very few studies have been conducted to combine music models with visual models. DeepBach [1] is a deep neural network model with good composition effect which focuses on producing music in the style of Bach. MetaComposer [2] creates harmonious, pleasant and interesting compositions by using a participant-based evaluation to combine the component of its framework effectively. Different from the previous work, MuseGan [3] proposes three multitrack music generation models which provide a method for generating music with the user's control. Since then, more studies have considered controllable music generation. DeepJ [4] is an end-to-end music generative model that is capable of composing music conditioned on a specific mixture of composer styles. DeepChoir [5] is a system that can automatically generate a four-part chorus from a given piece of music. However, none of those jobs combine music generation with the given videos.

**Music Generation From Videos.** A pioneering work of video background music generation was put forward by [13]. The algorithm realized the audio generation of music performance videos collected in the laboratory. A similar tool, Audeo [14], generated background music for an input pianist video. Another work, Foley-Music [15], was proposed to generate music for a given body poses video, which achieved outstanding results in silent performance video tasks. However, the above research work is mainly about generating music for performance videos. Compared to video background music, the music of these tasks is fixed to some extent, as it can be speculated by the human pose, and its style and instrument type are also relatively fixed. In these kinds of work, it is almost impossible to complete the music for scenery videos or videos without performers. As far as we know, the only model that generates background music for general purpose videos is Controllable Music Transformer(CMT) [6]. Although the rhythm of generated music matches the video well, it is not as melodic as the training data. In this paper, we propose a model for background music generation for the given video and we improve the rhythm of the music.

## 2.3. Technical Architecture

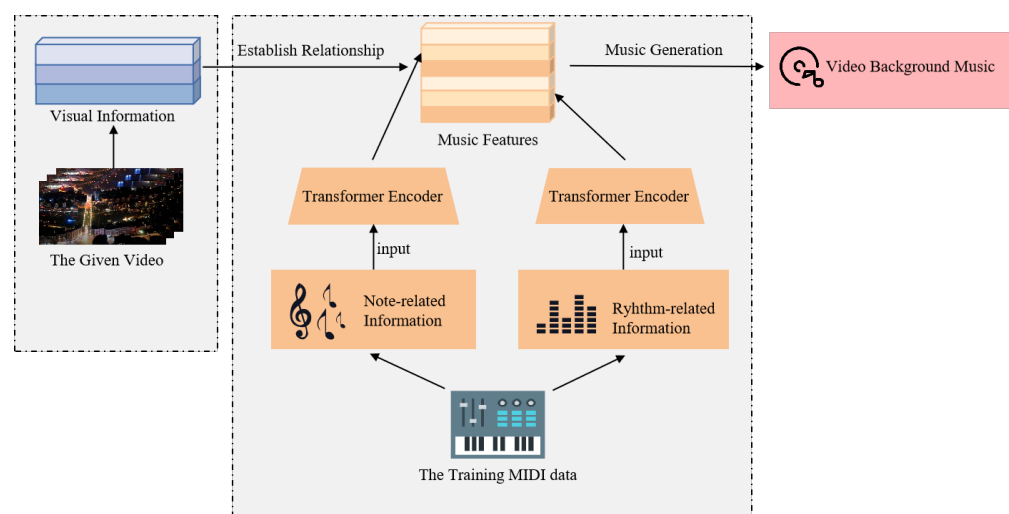
Many classical and popular music generation studies have used deep learning networks to model music notes. MuseGan [3] considered multitrack music sequences as 2D images and builds a model based on generative adversarial networks(GANs). EvoComposer [16] provided an evolutionary algorithm able to compose music. In addition, more research work is currently drawing on methods from studies with long sequence modeling, such as language identify [17], pretrained language model [10], semantic segmentation [18] and so on. Deepbach [1] is a graphical model using recurrent neural networks (RNNs) to model polyphonic music. DeepChoir [5] used two deep bidirectional RNN and a condition encoder to infer the harmonic part from the given music. Furthermore, DeepJ [4] used biaxial long short-term memory (LSTM) method and achieved valid results on the style-specific music generation task.

Apart from RNN and LSTM, the Transformer network, which has achieved compelling results in natural language processing and image processing field [19], has also widely been used in the music generation field. Music Transformer [20] employed the Transformer with a relative attention model to generate long-term piano music. Another work, Foley-Music [15], used a graph Transformer to generate music for a given body poses video. More recent works such as LakhNES [21] and Jazz Transformer [22] used Transformer-XL [23] as the backbone to complete the music generation task. Particularly, a model called Guitar

Transformer used Transformer-XL [23] to generate synthetic guitar finger-playing music was proposed by [24]. CMT used one linear Transformer [7] to extract all music features and generate background music for the given video although the melody of the resulting music was not good enough. In this article, we use two linear Transformers training jointly, which reduces the algorithm complexity compared with Transformer and use Linear attention mechanism to learn rhythm-related and note-related music features separately.

### 3. The Proposed Framework

In this paper, we propose a model to generate video background music based on training two linear Transformers jointly. The overall model structure is shown in Figure 2. The model consists of two parts. Firstly, when learning and extracting musical elements, musical features related to rhythm (beat, density and strength) are input into one Transformer, and musical elements related to notes (pitch, instrument type and duration) into the other.



**Figure 2.** The overall architecture of the proposed video background music generation method.

Through extensive training of MIDI data, the model can obtain representations of key music elements based on a given music. In the second part, the model calculates the visual characteristics and adjusts the value of music elements according to the time and motion characteristics and light flow changes of the video, and then guides the generation of background music. In the following of this section, we introduce in turn the presentation method of music data, the method of joint training and the way of using visual information to generate music.

#### 3.1. Data Representation

Inspired by PopMAG [25] and CMT [6], we chose beat, density, strength, pitch, instrument type and duration as key music elements and used the type to group the related features. In order to represent the relationship between various states in each musical feature and reduce the dimension according to the similarity between states, seven musical elements were represented by embedding vectors in Formula (1).

$$E_{i,n} = \text{Embedding}_n(\text{state}_{i,n}), n = 1, \dots, N, \quad (1)$$

In Formula (1),  $N$  represents the number of musical elements, and  $N$  was 7.  $\text{State}_{i,n}$  represents the state of the  $n$ th music element at the  $i$ th moment, and the embedding vector representation of the current music element state can be obtained through the learnable embedding layer  $\text{Embedding}(\cdot)$ . Then, the musical elements were concatenated and linearly connected with the related note and rhythm, respectively, as Formulas (2) and

(3) to obtain two preset embedded-layer-length (512) vectors, which represent rhythm and note information, respectively.

$$xt_{note} = Linear_{note}[E_{i,1} \oplus E_{i,2} \oplus E_{i,3} \oplus E_{i,type}], \quad (2)$$

$$xt_r = Linear_r[E_{i,4} \oplus E_{i,5} \oplus E_{i,6} \oplus E_{i,type}], \quad (3)$$

Taking Formula (2) as an example, the temporary embedding vector related to note was obtained using the pitch, instrument type, duration as well as the type embedding.  $Linear(*)$  represents the learnable full connection layer, the length of the input layer is the result of the concatenation for each element of the two groups, and the output vector is the preset embedding length; here it was 512.

$$xt_r = Linear_r[E_{i,4} \oplus E_{i,5} \oplus E_{i,6} \oplus E_{i,type}], \quad (4)$$

$$x_r = xt_r + \beta \cdot xt_{note}, \quad (5)$$

Next, before inputting into the linear Transformer, in Formulas (4) and (5), the hyperparameters  $\alpha$  and  $\beta$  represent the influence of rhythm features on note features and the influence of note features on rhythm features, respectively, in order to represent the relationship between the two groups and make the association between rhythm and note elements. Finally, the two embedding vectors  $x_{note}$  and  $x_r$ , which emphasize rhythm and note, were trained as an input to the Transformer network, which is described in the next section.

### 3.2. Training Two Linear Transformers Jointly with Position Encoding

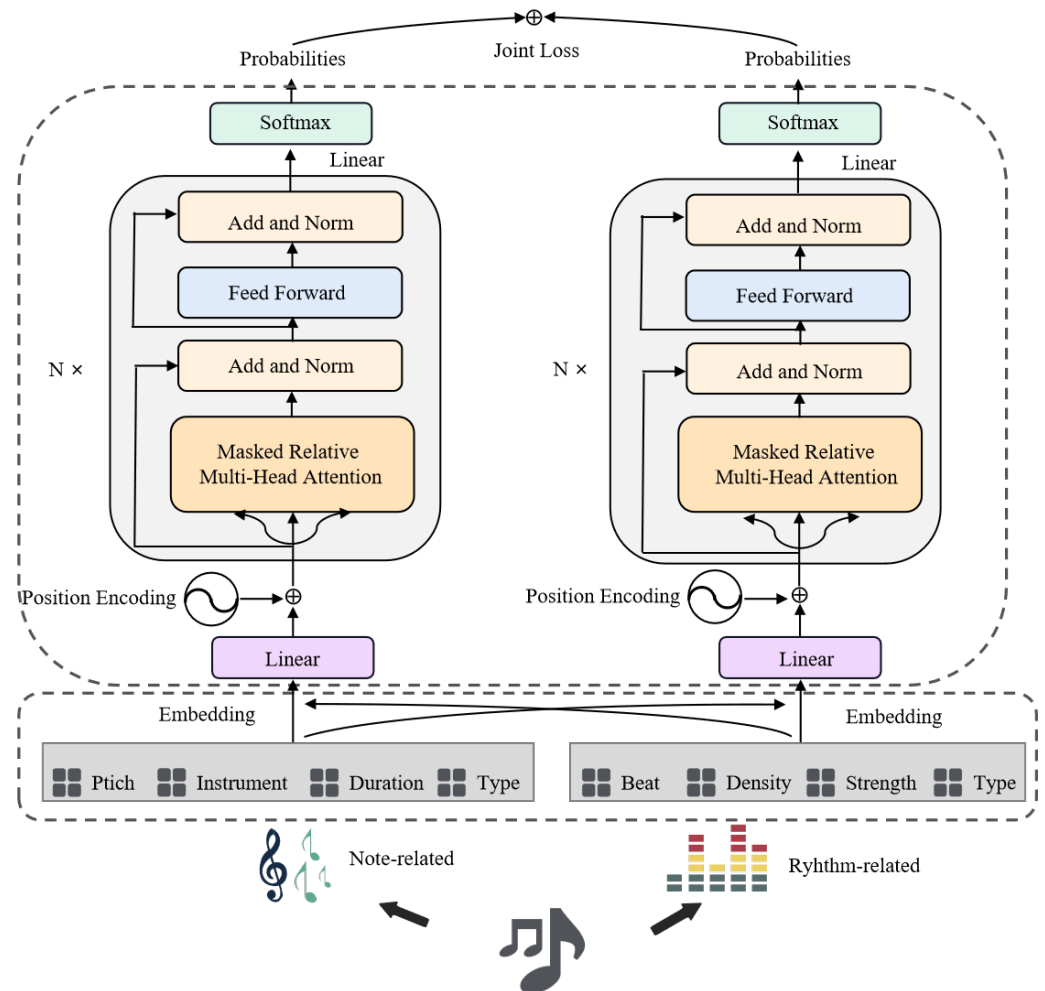
As the lightweight and linear attention mechanism of the linear Transformer, the model uses the linear Transformer as the backbone. Figure 3 depicts the structure of the model, which utilizes two linear Transformer networks corresponding to musical and rhythmic elements. In the first step, the seven musical elements were embedded and dimensioned. After the integration of the seven musical elements with note-related and rhythm-related features, respectively, the rhythm-related embedding vector was added to the note embedding vector in a certain proportion. Similarly, the note-related embedding vector was also adjusted to the rhythm embedding vector. Before blending with each other, the two individual vectors were resized to the preset embedding size by the linear Layer.

Before passing blocks in linear transformer, the time-beat position was encoded to match time, rhythm and note characteristics better. Furthermore, all musical elements in a beat should have the same position encoding, ensuring that multiple notes in the same beat are converted to the same musical fragment. After that, the time-beat position encoding was added to  $x_{note}$  and  $x_r$ .

Next, the resized embedding vector was transferred to the linear Transformer for training. Different from the traditional softmax attention, the linear Transformer uses a dot-product attention network, which enables the model to have better time and memory complexity and allows causal models for sequence generation in linear time. There are  $N$  blocks in a Transformer and  $N$  was 12 in this article. Each block was composed of a masked multiple linear attention mechanism and a position-wise forward layer. In particular, different from the application of a multihead attention in translation tasks, the encoder only focused attention on the previous notes. The mask was used to avoid the model learning the next note in advance, as we gave the target input into the Transformer at the training stage. It was added after the dot product between query and key. In this model, we use a triangular causal mask with 0 in the lower triangle and  $1 \times 10^{-9}$  in the upper triangle. After applying the mask matrix to the attention score, the upper triangle values became very low. Then, these low values became close to zero when passing through the softmax function, which meant the model could not notice the later notes.

As shown in Figure 3, after  $N$  blocks, a Linear layer and softmax function were used to calculate the result probability of the sequence and compute the possibility of each musical element. In this structural design, the model dealt with the two kinds of musical elements

separately. The proposed model computed and summed up the cross-entropy loss of each sequence to the target loss optimization. We aimed to ensure the relative independence of note-related and rhythm-related development and considered the mixed influence of them at the same time.



**Figure 3.** The structure of training two linear Transformers jointly to learn music elements.

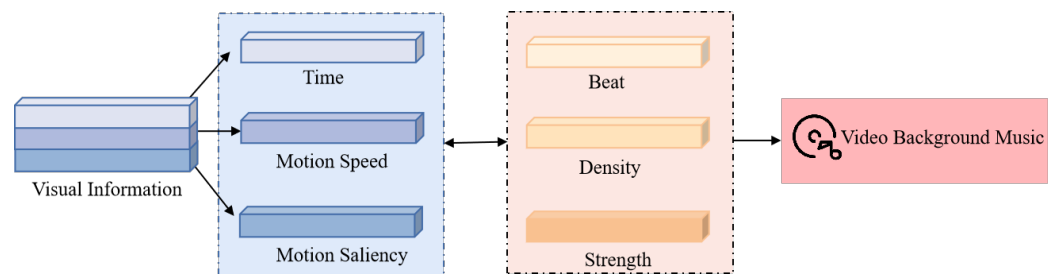
### 3.3. Video Directs the Generation of Music

As there is no paired video–background music training data to our knowledge, we used the ingenious inference method proposed in the Controllable Music Transformer [6]. In this paper, they established three rhythm relations between the given video and the generated music, as shown in Figure 4. Firstly, the model established a positive correlation between video optical flow speeds and the note density. Larger average optical flow means a faster motion in the video clip, which matches more intense background music. Secondly, the position encoding module was added to enhance the relationship between the video timing and music beats. Furthermore, the authors also built the relation between local maximum motion saliency and the note strength to speed up the tempo of the generated music when the motion changes in video clips such as shot boundaries and so on.

Aiming to introduce the video features to the music, we firstly used the Lakh Pianoroll Dataset (LPD) to train our double linear Transformer network on music feature modeling. Then, the model understood the note density and strength tokens of the generated music. The only thing we needed to do was replacing these two attributes to the calculated values that matched the given video. Through these visual-auditory connections, video variations



could effectively guide audio generation, thus ensuring that the generated music matched the rhythm of the given video.



**Figure 4.** The connection of visual and audio.

## 4. Experiment

In this paper, both subjective and objective experiments were conducted on the proposed model. In the subjective evaluation, we followed the consensual assessment technique (CAT) method to evaluate the metrics of the generated music. The main expert evaluation group consisted of five professors of music-related major at the Communication University of China and seven musicians with popular music pieces on the Internet. Firstly, participants rated the music to measure how well it matched the video compared to the music made by composers using the Likert scale five-point method. Then, they identified the generated music and music by composers from the disrupted music collection. For different scenes, the audiences voted on whether the background music was suitable for the given video, respectively. We finally calculated the Cronbach coefficient of the scores to prove the result was acceptable. As the feeling about the music varied from person to person, we also invited 23 people with no knowledge of music theory to do the same experiment. At the same time, the author paid attention to the quality of the generated music in the objective evaluation.

### 4.1. Data and Implementation Details

Referring to the Controllable Music Transformer model [6], we used the Lakh Pianoroll Dataset [3] as the source of multitrack piano roll data to train the model. This dataset is a variation of the Lakh MIDI dataset. We used a subset of Lpd-5-cleansed as our training data, which included 3038 MIDI music pieces of multitrack piano roll from all genres. Lpd-5-cleansed removed the songs whose first beat was not starting from time zero and kept only one file that had the highest confidence score in matching for each song. The validation set was a subset of the Lpd-5-cleansed training data containing 320 MIDI music pieces, and our test data were selected from Lpd-5-cleansed randomly with 304 MIDI pieces.

We used six NVIDIA Tesla P100-16GB cards on the Matpool cloud server website for the experiment with the environment of CUDA 11.1, cuDNN 8.0.5 and Python 3.7. The double linear Transformer network was used to train the data based on the framework from Figure 3. The number of layers of the Transformer was 12, that of the multiplex attention mechanism of causal-linear was 8 and the feed forward dimension was 2048. The dropout ratio was set to 0.1 with gelu [26] as the activation function in the Transformer. The initial learning rate was set to 0.0001, and we optimized the proposed model with Adam [27]. We trained 100 epochs in the LPD dataset and ran it for about 20 h in the environment of the six NVIDIA Tesla P100-16GB cards.

### 4.2. Subjective Evaluation

In the evaluation experiment, we focused more on the subjective evaluation of users. Among several subjective evaluation methodologies, the consensual assessment technique (CAT) was the most suitable method in this case. The CAT method is also called the “Gold Standard” of creativity assessment and is one of the most effective tools for measuring creative works, such as arts, engineering as well as business management. It can evaluate

our results because our creations were open, novel and appropriate. Firstly, we organized the people familiar with the field of music composition to form an evaluation team. Our team consisted of five professors of music-related major at the Communication University of China and seven musicians who had published music pieces on the Internet and had a certain popularity. Secondly, we told the audiences what aspects of the work they needed to evaluate. We chose the following metrics as evaluation standards:

1. Rhythm represents how the rhythm of the music fits into the pace of movement in the given video;
2. Emotional foil indicates whether the emotional expression of music is suitable for the content of the video;
3. Highlight says whether the music is stressed in key parts of the video;
4. Structure indicates whether the music is suitable for the background music with a crescendo at the beginning and a weakening at the end of the video.

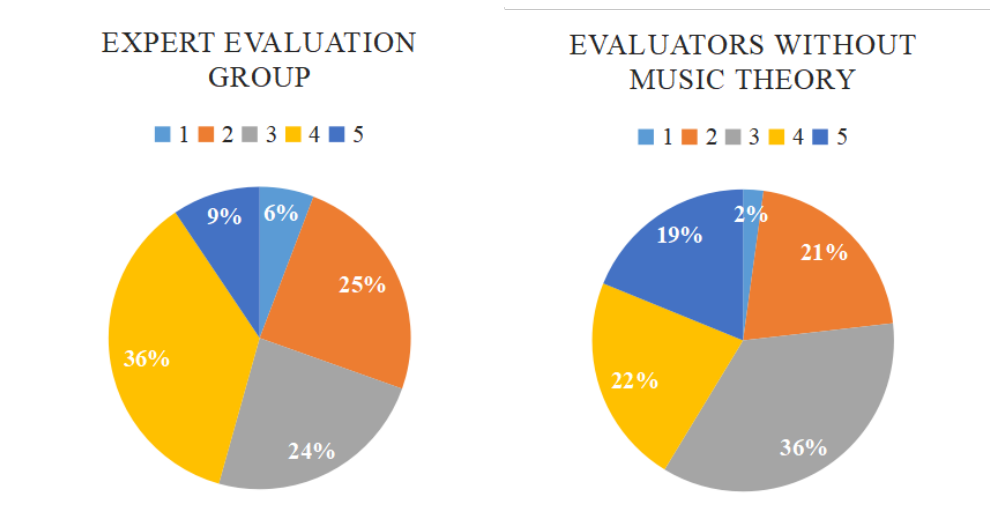
Finally, the evaluator evaluated all independent evaluation of the heard music pieces. Inspired by [1], the music pieces of five scenes were chosen, with three pieces of music for each. The selected five scenes were as follows: night scene, animals in the desert of snow mountains, battlefield, animated movie of animals and animated movie of characters. Furthermore, each piece was clipped to thirty seconds to keep the same experiment conditions. We used the Likert scale five-point method to divide the evaluation opinion into five grades, with a score ranging from one to five. In addition, considering the evaluation of music has a great relationship with subjective feelings and the feelings vary from person to person, we also chose 23 students without music knowledge background at the Communication University of China to make up our second evaluation group, and they also made independent evaluation of the same aspects.

As the CAT technique is based on the evaluators' perception of implicit theories of creativity, the subjectivity is relatively high. As a result, we calculated the raters' Cronbach score to reflect the reliability of the evaluation method after collecting the scores. From Figure 5, we find that most of the Cronbach scores of the expert group was higher than 0.7, which means the results can be accepted. In addition, we also calculated the statistics of the voting distribution for the two groups of evaluators, which is shown in Figure 6. One interesting finding is that the expert group voted less for a score of three than the other group, and they were also less likely to give a score of five.

	<b>Rhythm</b>	<b>Emotion Foil</b>	<b>Highlight</b>	<b>Structure</b>
Night scene	0.75	0.92	0.84	0.86
Desert animal	0.72	0.83	0.73	0.91
War field	0.86	0.9	0.81	0.96
Animal movie	0.81	0.75	0.64	0.84
Character movie	0.72	0.68	0.76	0.93

**Figure 5.** The Cronbach score of voting scores of the expert evaluation group.





**Figure 6.** The voting distribution of the two evaluation groups.

Then, we observed the scores on each evaluation index. Table 1 shows the average rating results of the expert group of the five scenes fit to the music generated by the algorithm, compared to the music made by the composer, calculated by Formula (6).  $N = 12$  represents the number of people of the expert group.

$$V_{ij} = \frac{\sum_{n=1}^{n=N} (S_{jm1} + S_{jm2} + S_{jm3})}{N}, \quad (6)$$

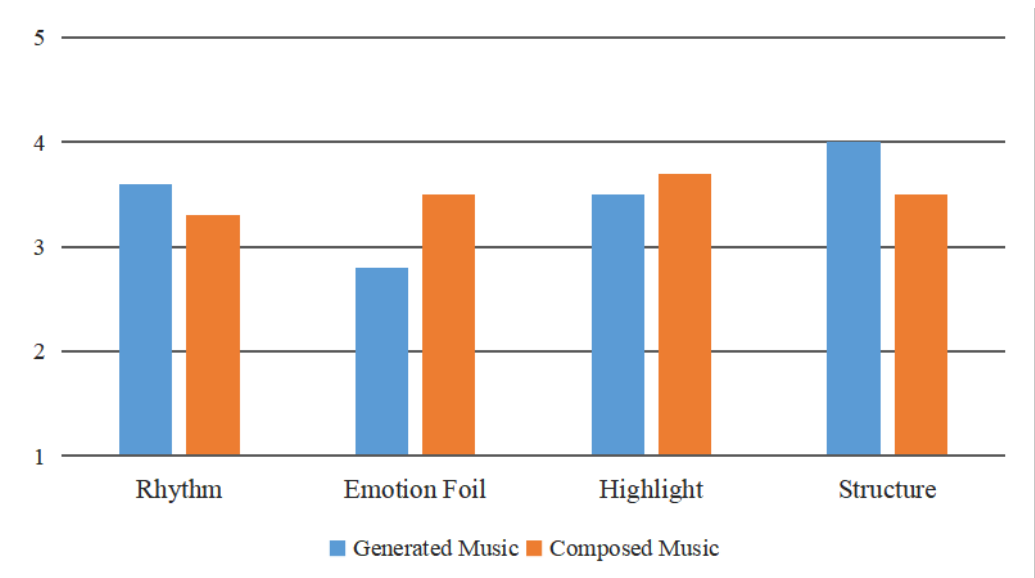
The values in Table 1 represent the expert group's score of this aspect of the generated music for each scene, and that in brackets represents the score of the music made by the composer. It can be seen from the table that although the overall score of the music generated by the algorithm did not exceed that of the existing composers, the model improved the results in highlight and structure indexes, indicating that the music generated by the algorithm was more in line with the rhythm of the video than the existing music searched through a given scene using the matching function. A major factor affecting the overall score was emotion foil; the reason was that the music generated by the proposed algorithm did not take into account the content of the videos, and as a result, it had a great disadvantage in the video soundtracks with strong emotion, such as war field.

**Table 1.** The score of the expert group on algorithmically generated music and music made by composers on 4 indicators. (The values in the table represent the score of the music generated by the algorithm, and the values in brackets represent the score of the music made by the composer.)

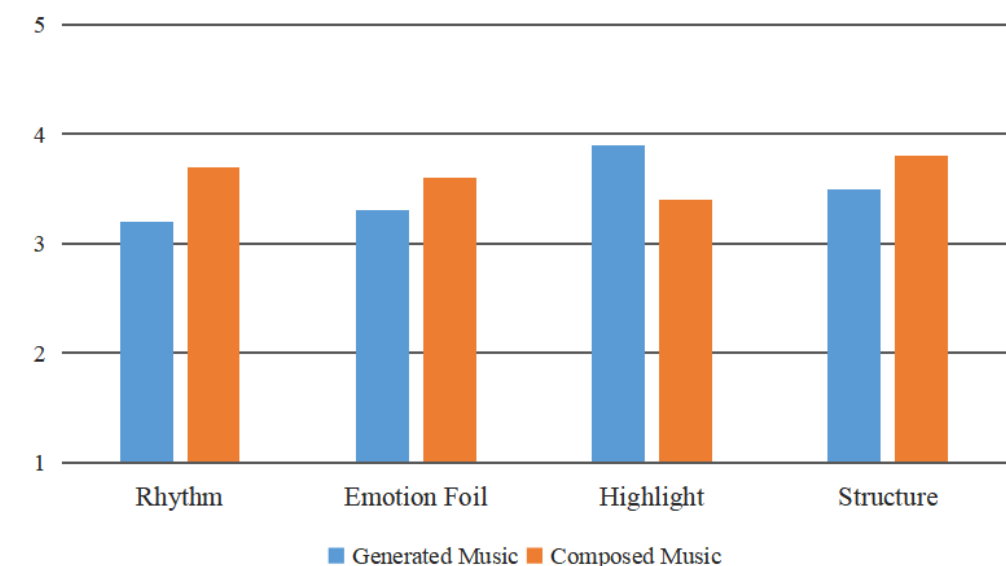
	Rhythm	Emotion Foil	Highlight	Structure	Avg.
Night scene	3.6 (3.3)	2.8 (3.5)	3.5 (3.7)	4.0 (3.5)	3.48 (3.50)
Desert animal	3.1 (3.4)	3.6 (3.1)	3.1 (3.1)	4.2 (3.6)	3.50 (3.30)
War field	2.8 (3.1)	1.4 (2.8)	3.5 (3.3)	3.1 (2.9)	2.70 (3.03)
Animal movie	3.2 (3.7)	2.4 (3.2)	3.8 (3.3)	3.8 (3.4)	3.30 (3.40)
Character movie	3.5 (3.6)	3.3 (3.5)	2.8 (3.2)	4.0 (3.6)	3.40 (3.48)
Avg.	3.24 (3.42)	2.70 (3.22)	3.34 (3.32)	3.82 (3.40)	3.28 (3.342)

At the same time, we analyzed the results from a single scenario. Figure 7 shows the average result from the expert evaluation group on the two kinds of music on different indicators, taking the night scene as an example and Figure 8 shows the average result from the group without music knowledge background. It can be concluded from the distribution in the following graphs that for the expert evaluation group, the structure score of the generated music was much larger than that of the composed music, because of the position encoder in the Transformer network. To our surprise, combined with changes of the video

motion speed, the generated music of the night scene was also slightly more rhythmic than the existing composed music selected using a matching algorithm. Different from the expert review panel, the listeners with no musical knowledge preferred the highlight of the generated music, which meant our model also improved the stress of the music by building a connection between local maximum motion saliency and note strength. However, the emotion foil score of the two groups was poorly experienced, which we need to consider in future works.



**Figure 7.** The average result from the expert evaluation group on two kind of music pieces on different indicators of night scene.



**Figure 8.** The average result from the evaluation group without music knowledge background on two kind of music pieces on different indicators of night scene.

Apart from the CAT method, we also compared the generated music with existing matching music of the composers by mixing the generated music and the music written by the composer into a set and labeled it. Then, we randomly selected the music for the two groups of participants and asked the user to distinguish whether it was generated by model or composed by composers. At the same time, participants were also asked to vote on whether they would have liked to use the heard music as background music

for the given video. Listeners judged the current music as “It is generated by AI”, “It is generated by Human” or “Unable to distinguish”. Table 2 shows the percentage of listeners correctly identifying the music generated by the model and by the composer, as well as the percentage willing to use this type of music as background music; the percentages in brackets are the votes of evaluators without knowledge of music theory.

It can be seen from Table 2 that most listeners could identify the music generated by the model, while more users were willing to use the music generated by the model as background music, as it matched the rhythm of the video better, and most of the music’s rhythm and structure also accorded with the video scene compared to the existing music.

**Table 2.** Audience’s distinction between the two kinds of music and their willingness to use the generated music as background music (the percentages in the table represent how recognizable the music generated by the model was when voted by the expert group and those in brackets were from people with no musical background.)

	It Is by AI	It Is by Human	Want to Use?
Generated music	89% (62%)	11% (35%)	56% (75%)
Composers’ music	5% (23%)	95% (64%)	78% (63%)

#### 4.3. Objective Evaluation

Although the main purpose of this paper was to generate appropriate background music for a given video, that is, to evaluate the matching degree of music and video in subjective experiments, we also paid attention to the quality of the generated music itself. In an objective evaluation, in order to measure the melody of the music, MusDr [22] was used to calculate the following indicators to measure the quality of the music itself, including pitch class histogram entropy (H), grooving pattern similarity (GS) and structureness indicator (SI):

1. The pitch class histogram entropy measures the tonality of the music by the entropy of the pitch. The calculated method is shown in Formula (7), mentioned in Jazz Transformer; here  $P = 11$ . The clearer the music piece, the lower the histogram entropy;

$$H(\vec{h}) = - \sum_{i=0}^P h_i \log_2(h_i), \quad (7)$$

2. The grooving pattern similarity computes the music rhythmicity, music with higher melodies has higher scores;
3. The structureness indicator measures the apparent repetition over a given length of time. The closer this indicator is to the music by composers, the better.

In the experiment, five songs generated by the models were fed into MusDr for index computation. For the proposed model, the music of night scene was used to detect the indicators as this kind of background music had great scores in the subjective evaluation. A higher value was better for the sum of H and GS while a smaller value was better for the gap between the SI of the model generated music and the standard value. The standard value was obtained by taking three pieces of music from each of the five scenes and averaging the output index of MusDr. Compared with the music generated by Jazz Transformer [22], CMT [6] and VMSI [28], the sum value and the gap were improved as shown in Table 3. Here,  $CMT_1$ , representing the control attribute, was only added in training while  $CMT_2$  means the attribute worked both in the training and inference step. Although the music generated by our model did not get the highest score, the sum of H and GS reached the top. That means it is effective to learn rhythm-related and note-related music features separately with two linear Transformers training jointly.

**Table 3.** Objective comparison of music generated by different models.

	H	GS	SI	SUM of H and GS
Standard Value	4.452	0.968	0.488	5.42
Jazz Transformer ([22])	2.91	0.76	<b>0.27</b>	3.67
VMSI ([28])	3.961	0.713	0.265	4.674
CMT_1 ([6])	3.617	<b>0.81</b>	0.241	4.427
CMT_2 ([6])	<b>4.113</b>	0.599	0.2	4.712
<b>The proposed model</b>	4.028	0.729	0.221	<b>4.757</b>

## 5. Discussion and Conclusions

In this paper, we showed that the proposed model improved the rhythm and melody of music in video background music generation task by using two linear Transformers training jointly. Compared with the existing research, the proposed model dealt with musical elements of note-related and rhythm-related features separately to extract and learn music features. Then, we used the existing multimodel inference method to skillfully combine the visual features of the given video with the music features while guiding the generation of the music elements. Objective and subjective experiments showed that the proposed model could be effective in matching the background music with the given video in rhythm and notes while also being melodious.

The experimental result showed that our proposed model with two linear Transformers was better in terms of the rhythm of the generated music. Furthermore, the generated music also matched most of the given video well. However, we also found that the proposed model was flawed in videos with strong emotion such as battle scenes, as we did not focus on the understanding of the video content. Furthermore, due to the double linear Transformer structure requiring more parameters to learn, the training phase of the model took a long time. How to reduce the complexity of the model is also a question we will consider in follow-up research.

In future work, we hope to use different and lightweight Transformer networks, respectively, combining the differences of rhythm-related and note-related music elements, so that the music elements can be better extracted and learned. At the same time, we hope to capture the content and emotional message of the video and make the generated music more consistent with the content of the given video.

**Author Contributions:** writing—original draft preparation, X.Y.; writing—review and editing, Y.Y.; visualization, X.Y.; project administration, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by National Key R&D Program of China (No. 2021YFF0900701).

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Hadjeres, G.; Pachet, F.; Nielsen, F. Deepbach: A steerable model for bach chorales generation. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2017; pp. 1362–1371.
2. Scirea, M.; Togelius, J.; Eklund, P.; Risi, S. Metacompose: A compositional evolutionary music composer. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 202–217.
3. Dong, H.W.; Hsiao, W.Y.; Yang, L.C.; Yang, Y.H. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Orleans, LA, USA, 2–7 February 2018; Volume 32.
4. Mao, H.H.; Shin, T.; Cottrell, G. DeepJ: Style-specific music generation. In *Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 377–382.
5. Wu, S.; Li, X.; Sun, M. Chord-Conditioned Melody Choralization with Controllable Harmonicity and Polyphonicity. *arXiv* **2022**, arXiv:2202.08423.

6. Di, S.; Jiang, Z.; Liu, S.; Wang, Z.; Zhu, L.; He, Z.; Liu, H.; Yan, S. Video Background Music Generation with Controllable Music Transformer. In Proceedings of the 29th ACM International Conference on Multimedia, Nice, France, 21–25 October 2021; pp. 2037–2045.
7. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2020; pp. 5156–5165.
8. Hsiao, W.Y.; Liu, J.Y.; Yeh, Y.C.; Yang, Y.H. Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. *arXiv* **2021**, arXiv:2101.02402.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html> (accessed on 21 April 2012).
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Oore, S.; Simon, I.; Dieleman, S.; Eck, D.; Simonyan, K. This time with feeling: Learning expressive musical performance. *Neural Comput. Appl.* **2020**, *32*, 955–967. [[CrossRef](#)]
12. Huang, Y.S.; Yang, Y.H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1180–1188.
13. Chen, L.; Srivastava, S.; Duan, Z.; Xu, C. Deep cross-modal audio-visual generation. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 349–357.
14. Su, K.; Liu, X.; Shlizerman, E. Audeo: Audio generation for a silent performance video. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3325–3337.
15. Gan, C.; Huang, D.; Chen, P.; Tenenbaum, J.B.; Torralba, A. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 758–775.
16. De Prisco, R.; Zaccagnino, G.; Zaccagnino, R. Evocomposer: An evolutionary algorithm for 4-voice music compositions. *Evol. Comput.* **2020**, *28*, 489–530. [[CrossRef](#)] [[PubMed](#)]
17. Mohamed, A.; Okhonko, D.; Zettlemoyer, L. Transformers with convolutional context for ASR. *arXiv* **2019**, arXiv:1904.11660.
18. Zhou, T.; Wang, W.; Konukoglu, E.; Van Gool, L. Rethinking Semantic Segmentation: A Prototype View. *arXiv* **2022**, arXiv:2203.15102.
19. Wang, S.; Zhou, T.; Lu, Y.; Di, H. Detail-Preserving Transformer for Light Field Image Super-Resolution. *arXiv* **2022**, arXiv:2201.00346.
20. Huang, C.Z.A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A.M.; Hoffman, M.D.; Dinculescu, M.; Eck, D. Music transformer. *arXiv* **2018**, arXiv:1809.04281.
21. Donahue, C.; Mao, H.H.; Li, Y.E.; Cottrell, G.W.; McAuley, J. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. *arXiv* **2019**, arXiv:1907.04868.
22. Wu, S.L.; Yang, Y.H. The Jazz Transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures. *arXiv* **2020**, arXiv:2008.01307.
23. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.
24. Chen, Y.H.; Huang, Y.H.; Hsiao, W.Y.; Yang, Y.H. Automatic composition of guitar tabs by transformers and groove modeling. *arXiv* **2020**, arXiv:2008.01431.
25. Ren, Y.; He, J.; Tan, X.; Qin, T.; Zhao, Z.; Liu, T.Y. Popmag: Pop music accompaniment generation. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1198–1206.
26. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
27. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
28. Chang, C.J.; Lee, C.Y.; Yang, Y.H. Variable-length music score infilling via XLNet and musically specialized positional encoding. *arXiv* **2021**, arXiv:2108.05064.