FilmComposer: LLM-Driven Music Production for Silent Film Clips

Zhifeng Xie[1,2], Qile He[1], Youjia Zhu[1], Qiwei He[1], Mengtian Li[1,2†]
[1]Shanghai University
[2]Shanghai Engineering Research Center of Motion Picture Special Effects
{zhifeng_xie,shu_hql,zyj028,heqiwei,mtli}@shu.edu.cn

Abstract

*In this work, we implement music production for silent film clips using LLM-driven method. Given the strong professional demands of film music production, we propose the **FilmComposer**, simulating the actual workflows of professional musicians. FilmComposer is the first to combine large generative models with a multi-agent approach, leveraging the advantages of both waveform music and symbolic music generation. Additionally, FilmComposer is the first to focus on the three core elements of music production for film—audio quality, musicality, and musical development—and introduces various controls, such as rhythm, semantics, and visuals, to enhance these key aspects. Specifically, FilmComposer consists of the visual processing module, rhythm-controllable MusicGen, and multi-agent assessment, arrangement and mix. In addition, our framework can seamlessly integrate into the actual music production pipeline and allows user intervention in every step, providing strong interactivity and a high degree of creative freedom. Furthermore, we propose **MusicPro-7k** which includes 7,418 film clips, music, description, rhythm spots and main melody, considering the lack of a professional and high-quality film music dataset. Finally, both the standard metrics and the new specialized metrics we propose demonstrate that the music generated by our model achieves state-of-the-art performance in terms of quality, consistency with video, diversity, musicality, and musical development. Project page:* `https://apple-jun.github.io/FilmComposer.github.io/`

1. Introduction

The music in films can enhance visual content, guide emotions, and support narrative flow. However, manual music production is time-consuming and requires specialized expertise. Therefore, developing a fast, effective and user-friendly AI approach for film music production would assist experts in producing music more efficiently, while also enabling non-experts to create high-quality music with ease.

†Corresponding author.

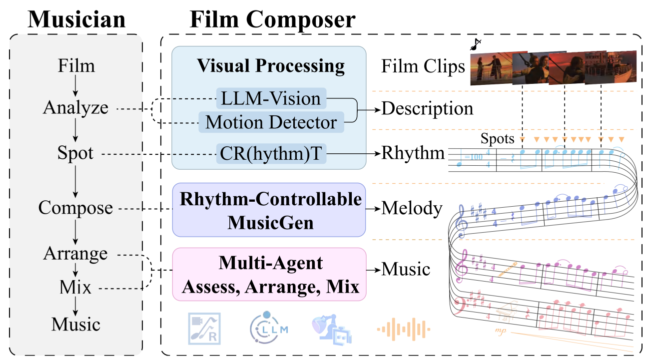

Figure 1. **A schematic of our work.** The left column illustrates the actual steps taken by human musicians in music production, while the middle column represents the corresponding simulated blocks in the FilmComposer. The inputs and outputs at each stage are depicted on the right, visually demonstrating how film clips are gradually transformed into the final music.

The challenge of AI music production lies in meeting the high standards of quality, musicality, and development. Typical music for film demands at least a 48kHz sampling rate and 24-bit depth, which current models are still far away from. It requires not only technical excellence but also heightened musicality, encompassing both expression and aesthetics. The most overlooked one, Musical development, refers to the evolution of themes and motifs over time, greatly contributing to audio-visual consistency.

Previous work focused on optimizing audio token compression to improve music quality. However, even the most advanced token interleaving patterns [7] and stable audio techniques [11] fail to meet the required standard. Since the introduction of MusicLM [2], large language models have gained significant attention, notably improving musicality. Following the open-sourcing of Meta's MusicGen [7], large language models have been widely integrated into various frameworks [21, 26]. While musicality has improved, control remains limited. Some research has explored adding conditions, such as rhythm [9, 21, 24, 45], timing [11], and melody [7, 21]. Yet, these methods tend to either compromise musicality or only achieve superficial control, lacking effective musical development. In summary, breakthroughs have been made, but it remains insufficient for the specific demands of music production for film.

To address the challenges above, we propose **FilmComposer** which imitates musicians to produce music for film, including spotting, composition, arrangement and mix. Corresponding to the general three steps, FilmComposer consists of visual processing, rhythm-controllable MusicGen, and multi-agent assessment, arrangement and mix. To train FilmComposer, We also construct a professional film music dataset named **MusicPro-7k**, comprising 7K video-music pairs, descriptions, main melodies, and rhythm spots.

Specifically, we process film clips to build description and precise rhythm control for spotting. Next, we integrate a rhythm conditioner into MusicGen and fine-tune it to compose in sync with rhythm cues, ensuring strong musicality and video alignment. This adaptation makes the model the first large language model for film-to-music generation, with generating music from visual input lowering the threshold of large language model usage. Additionally, we designed an agent-based musicality metric to filter generated melodies, ensuring a main melody with strong musicality. Next, we develop a multi-agent system based on large language models to arrange the generated main melody, which means organizing musical elements to support thematic, emotional and narrative progression. It enhances the musicality and provides excellent musical development. Then the multi-agent system operates a digital audio workstation (DAW) to arrange, mix, and finally output a high-quality and ready-to-use music for silent film clips. Our contributions are summarized as follows:

- We propose the **FilmComposer**, functioning as a musician, highly interactive and seamlessly integrated into actual production pipelines.
- We construct Rhythm-Controllable MusicGen and fine-tune it by self-created multifunctional film music dataset **MusicPro-7k**, enabling it to generate music that aligns with film clips while maintaining strong musicality.
- We integrate a multi-agent system to enhance musical development, resulting in music that excels in quality, diversity, musicality, and audio-visual consistency, as validated by newly proposed evaluations and user study.

2. Related Work

2.1. Music Generation

Waveform music generation. The question of whether to use symbolic music generation or waveform music generation remains a contentious issue. Waveform music benefits from its prevalence, supported by a wealth of generation models [2, 7, 11, 18, 21, 25, 34] and vast amounts of music data [2, 11, 18, 34], resulting in typically richer outputs. However, it suffers from generally poor sound quality, frequent noise and distortion, and less effective control.

Symbolic music generation. Symbolic music records how music is performed and thus captures the essence of music more closely, allowing for finer control and resulting in more pleasing melodies [6, 8, 22, 41]. Leveraging the unique characteristics of symbolic music, some studies have moved away from commonly used models like transformers and diffusion for music generation. Instead, they employ the multi-agent approach, achieving impressive results with greater simplicity [6, 8, 41]. However, symbolic music generation relies on MIDI datasets, which are scarce and resource-intensive to build. As a result, symbolic music generation models struggle to achieve large-scale deployment, leading to repetitive and monotonous music.

Aiming to harness the richness of waveform music while leveraging the high quality of symbolic music, we propose FilmComposer, which effectively integrates both strategies.

2.2. Video to Music Task

Video to music generation. The task of generating music for human-centric videos has been around for some time, primarily focusing on dance and concert [12, 42, 44], where clear human movements dictate the musical rhythm. Subsequently, CMT [9] pioneered the concept of video background music generation, enabling music creation for more flexible video formats, particularly suitable for short videos. CMT focuses on the rhythm alignment, connecting video and music by extracting their corresponding features. Diffbgm [24] continued to explore rhythm control, yielding new advancements. Additionally, certain studies [19, 26, 45] have focused on semantic alignment, with M2UGen [26] incorporating text controls for music generation, achieving impressive results. VidMuse [36] has begun exploring more specialized scenarios, such as music generation for advertisements and trailers, employing long-short sequence modeling to achieve multiple controls simultaneously.

Music retrieval. Music Retrieval also receives attention in recent studies [30, 35, 40, 43], with some works examining the relationship between video, audio and language [15, 37], others guiding music recommendations based on description [27] or dialogue [10].

However, the works above ignored exploring film music or considering the actual requirement of music production. Therefore, our work is the first to focus specifically on music production for film, requiring the generated music to be more professional and closely aligned with various aspects of film clips. Meanwhile, our approach integrates more user-friendly rhythm control and text control, resulting in more refined results.

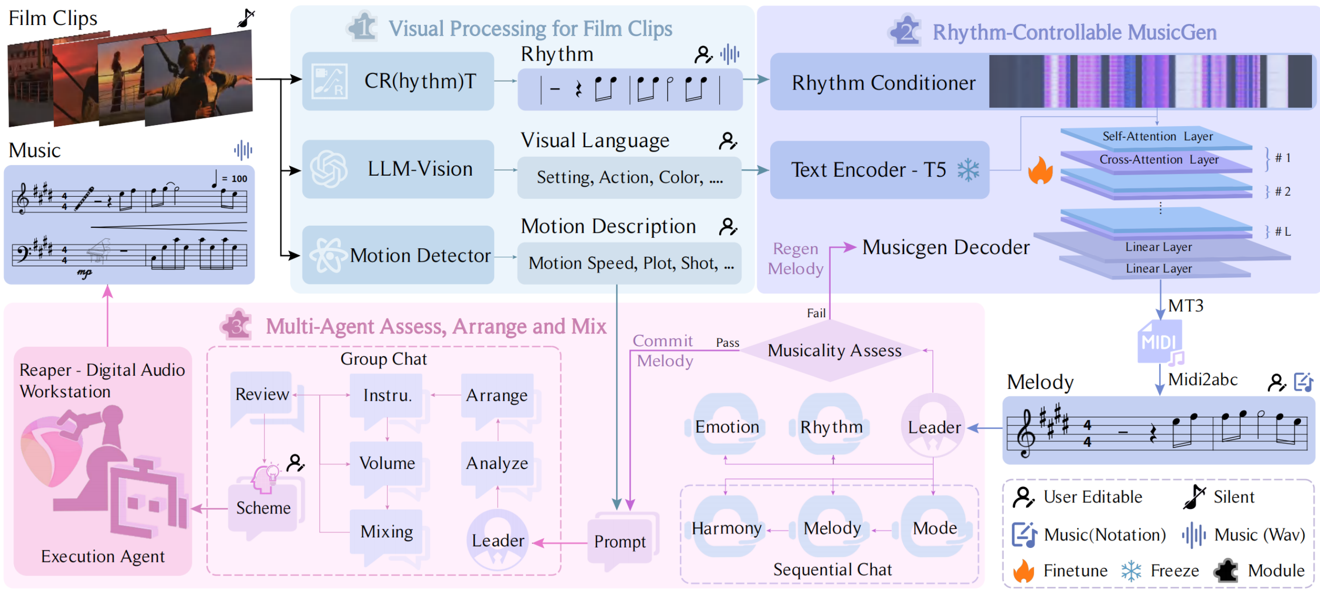

Figure 2. **The framework of FilmComposer.** Three large color blocks represent the three main modules, through which the input Film clips pass sequentially, ultimately outputting a waveform. The three blue blocks with musical notation illustrate the complete music production process, from setting the rhythm points, composing, to arranging and mixing.

2.3. Dataset

Most video-to-music datasets only focus on generating music for human-centric videos—dance [23, 40, 44] or concert [26]. MuVi-Sync [19] collects music videos characterized by a pop-oriented style, unsuitable for film music production. V2M [36] has gathered a more diverse set of video music data, but it primarily covers advertisements and trailers, with severe audio degradation caused by the mute operations. It leans toward advertising and differs from true film music. Since the professional requirements for music production are significantly higher than these types of music, there is still a lack of high-quality, large-scale, specialized film music datasets. To address this problem, we construct a dedicated dataset MusicPro-7k including film clips and music pairs, descriptions and rhythm spots.

3. Method

Overview. As shown in Fig. 2, FilmComposer consists of three main modules—visual processing, rhythm-controllable MusicGen and multi-agent assessment, arrangement and mix—to simulate the process of a human musician producing music for film. In Sec. 3.1, we perform visual processing on input film clips to construct rhythm conditions, shape visual language, and extract motion descriptions, simulating the analysis and spotting process. In Sec. 3.2, the Rhythm-Controllable MusicGen generates melodies with high musicality, aligned with the rhythm spots and visual language. The melody is transcribed into MIDI and ABC notation for further professional work. In Sec. 3.3, a multi-agent system assesses the melody's musicality. if it fails to meet the standard, the previous module will be ordered to generate a new melody. Otherwise, the system proceeds with arrangement and mix, mapping out a scheme and instructing a digital audio workstation to execute it, ultimately generating cinematic-quality music.

3.1. Visual Processing for Film Clips

When generating music for film clips, three key aspects must be considered: first, the music should align with the visual content to enhance the atmosphere; second, it must follow the pacing of the film clips; and third, it should reflect higher-level film semantics, such as emotions, themes, and plot development. To address this, we do visual processing to extract these three components from the film clips.

First, we modify CMT [9] into the CRT—Controllable Rhythm Transformer, which processes video and generates rhythm points. To be specific, we introduce an algorithm that extracts the main melody from symbolic music, and then flatten it into a rhythm. The algorithm calculates track coverage and note ratios, using instrument performance characteristics to identify the lead instrument. The coverage $Coverage_i$ for the i-th track is:

$$Coverage_i = \frac{T_i}{T_{music}}, \quad (1)$$

where T_i is the total time during which the notes of the i-th track are played, and T_{music} is the music duration.

The note ratio R_i for the i-th track is calculated as:

$$R_i = \frac{n_i}{\sum_{i=1}^{n} n_i}, \quad (2)$$

where n_i is the number of notes in the i-th track.

If the lead instrument's coverage is insufficient, we incorporate additional tracks until the total coverage $Coverage$ meets the threshold, computed as:

$$Coverage = \frac{1}{T_{music}} \sum_{i=1}^{2n-1} \mathbb{I}(N(t_i) > 0) \cdot (t_{i+1} - t_i), \quad (3)$$

where T_{music} is total music duration; $\{t_i\}_{i=1}^{2n}$ is ascending sequence of all note-onset and note-offset timestamps; $N(t)$ is active note count at t, incremented for note-onset and decremented for note-offset; $\mathbb{I}(\cdot)$ is indicator function.

Second, drawing from audio-visual language and music emotion analysis theory [16, 17, 33], we designed visual attributes: setting, brightness, color hue, action, emotional tone, view scale, and theme. ChatGPT-4V [1] is used to efficiently recognize these attributes from selected frames.

The elements extracted above are enough to guide the composition, but more is needed to guide the arrangement and mix. So third, We employ open-source algorithms [1, 4, 5] to extract motion speed, motion saliency, shot cut, and plot development. Motion speed S_{motion} for each video segment is calculated as:

$$S_{motion} = \frac{1}{N} \sum_{i=1}^{N} M_i, \quad (4)$$

where M_i is the optical flow magnitude at the i-th frame.

The motion saliency $S_{saliency}$ is the mean of the positive changes in optical flow magnitude between consecutive frames—the average increase in motion intensity:

$$S_{saliency} = \frac{1}{N^+} \sum_{i=1}^{N^+} (M_i - M_{i-1})_+. \quad (5)$$

3.2. Rhythm-Controllable MusicGen

Once rhythm spots and semantic information are ready, large language models can be applied to generate the corresponding main melody. Since there is no music generation model that can take visual language as input and be controlled by rhythm spots, we construct Rhythm-Controllable MusicGen and train it with MusicPro-7k.

Rhythm-Controllable MusicGen is composed of the rhythm conditioner, the T5 text encoder [31], and the MusicGen decoder [7]. In general, we explore joint conditioning of both the chromagram of the input rhythm spots and visual description to exert control over the melodic structure. The output Y_{output} can be calculated as:

$$Y_{output} = \text{Decoder}(C_{rhythm}, C_{description}, Y), \quad (6)$$

where the rhythm condition C_{rhythm} and visual description $C_{description}$ is passed as prefix to the transformer input Y - codebook projections along with the positional embeddings. specifically, we employ the prepend method for condition fuser, sequentially concatenating the processed rhythm condition, text condition, and transformer input.

The input passes through a transformer consisting of L layers, each with dimension D. Every layer incorporates a causal self-attention block and a cross-attention block. Lastly, the layer concludes with a fully connected block composed of a linear layer from D to $4{\cdot}D$ channels, followed by a ReLU activation, and then another linear layer back to D.

We focus on implementing the rhythm conditioner that processes audio inputs for model conditioning. This specific conditioner operates based on the chromagram, which reflects the detailed beat characteristics of the audio. The input waveform of rhythm spots is tokenized and the chromagram features are extracted from it. These chromagram embeddings are projected to match the required output dimensions. A downsampling factor is applied to adjust the chromagram length, and a mask ensures that only valid audio regions are considered. The system supports both real-time chromagram extraction and precomputed chromagrams, enhancing flexibility in handling different audio inputs.

3.3. Multi-Agent Assess, Arrange and Mix

To make the following operations efficient, we use the MT3 model [13] to convert the music to the MIDI format and then use Midi2abc [39] to convert it into ABC notation-a musical notation format based on the ASCII character set, aligned with the musician's composition process.

To ensure the musicality of the melody generated, we design a multi-agent assessment system based on AutoGen framework [38]. If the melody meets the criteria, it proceeds to the arrangement part; otherwise, the rhythm-controllable MusicGen module regenerates it. Each agent acts as a reviewer, guided by Role-play prompt engineering, evaluating aspects such as mode, melody, harmony, rhythm, and emotional expression based on music theory. Given that mode influences melody, and melody influences harmony, we design a sequential chat to link the mode agent, melody agent and harmony agent, ensuring they respond in reference to higher-level agents. The assessment criteria for each agent are detailed in Tab. 1.

Next, we apply a multi-agent system for arrangement and mix. Since this process is challenging even for human musicians, we utilize prompt engineering techniques, including Role-play, Chain of Thought (CoT), and Few-Shot Prompting, instructing a group chat for this task. The agents receive prompts consisting of motion descriptions and melody in ABC notation, and collaborate on the final arrangement, with each agent's role outlined in Tab. 2.

In the group chat, the speaking order of each agent follows the actual sequence of music production, namely, analyzing the needs first, arranging and orchestrating next, then adjusting the volume to enhance dynamics, and finally mix-

Name	*Assessment Criteria*
Mode	*1.The mode is clear; 2.The tonality is stable.*
Melody	*3.Reasonable chord progression; 4.The chords are diverse, not monotonous; 5.Appropriate variation; 6.Appropriate repetition.*
Harmony	*7.very harmonious; 8.rich, not monotonous; 9.The orchestration is reasonable; 10.the instruments collaborate well.*
Rhythm	*11.the rhythm is clear; 12.the beat is consistently the same. 13.The rhythmic pattern has appropriate variation; 14. The rhythmic pattern has appropriate repetition.*
Emotion	*15. The mode matches the emotion; 16. The chord progression matches the emotion; 17.the rhythm matches the emotion; 18. The choice of instruments matches the emotion; 19. The playing techniques fit the emotion (e.g., staccato, legato).*

Table 1. Agent names and their corresponding evaluation criteria.

Name	*Duties*
Analyze	*1.Summarize the development characteristics of the video; 2.Specify whether each measure should be forte or piano.*
Arrange	*1.Specify which measures of which tracks should be softened; 2.Duplicate tracks to increase harmony.*
Instrument	*Assign appropriate instrument.*
volume	*Design the volume envelope.*
Mixing	*1.Set the pan; 2.Set the reverb level.*
Reviewer	*1.Review whether the arrangement and mix match the video in every aspect. 2.Examine the playing techniques and effectiveness of the instruments. 3.Develop the final scheme.*

Table 2. Arrangement and mix agent names and their duties.

ing. Besides, the process of orchestration and dynamic adjustment needs to be guided by the reviewer agent.

Since some tasks are difficult for the agent, CoT is used to break down steps and improve performance. For example, it is observed that the agent is somewhat sluggish in calculating, so the Analyze-Agent is instructed to first count harmonies per measure before further analyses. To prevent excessive instrumentation, the Instrument-Agent summarizes the original instrument types and makes selections accordingly. To guide agents' responses and ensure consistency, Few-Shot Prompting provides response templates, resulting in high-quality, well-structured outputs. The output scheme of arrangement and mix is shown in Fig. 3.

Finally, an execution agent will convert this scheme into a form that Reaper—a lightweight, edit-friendly digital audio workstation—can understand, and thus guide Reaper to arrange and mix the melody of MIDI form by coding. Once completed, Reaper will output a piece of cinematic-grade music of exceptional quality, rich in musicality and perfectly aligned with the development of the film clips!

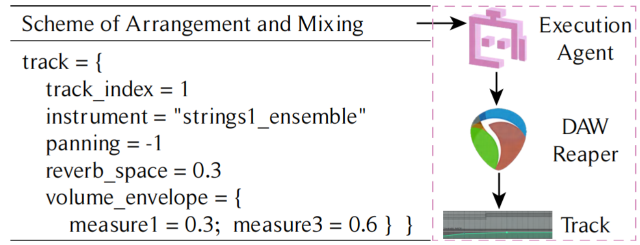

Figure 3. The output scheme from the multi-agent arrangement and mix system, together with subsequent operation.

4. MusicPro-7k Dataset

Given the lack of a suitable dataset to train FilmComposer, we construct a large-scale film music dataset MusicPro-7k featuring about 7,418 samples, each with film clip, high-quality music, visual description, music description, main melody, and rhythm spots. MusicPro-7k is distinguished by its high quality, high level of professional expertise, and multi-functionality. The structure and construction method of MusicPro-7k are illustrated in Fig. 4, and more details can be found in supplementary materials.

Instead of extracting music from videos and applying noise reduction as done in earlier approaches, we invited musicians to find music specifically for silent film clips. This approach avoids the poor audio quality and degradation caused by previous extraction methods.

Film Clips. 7K Film clips are derived from LSMDC [32], selecting video segments over 8 seconds long from numerous classic films. The source films cover all common genres, including drama, thriller, comedy, romance, and more.

Music. Musicians followed these steps: (1) For clips with pre-existing music, source their original audio; (2) If the original music could not be located or was under copyright protection, select a similar piece based on theoretical similarity and functional consistency as criteria; (3) For clips without pre-existing music, select suitable music from the music library, based on semantic, rhythmic, and progressive alignment with the clip, as well as all other relevant aspects.

Description. Musicians were also asked to annotate visual descriptions for each video based on predefined visual attributes, referring to audiovisual language theory and presented in supplementary materials. During the rhythm-controllable MusicGen fine-tuning, using the full visual description directly proved too broad in scope. To address this, we introduced music descriptions and removed components in stages, creating a smoother training process. The music descriptions, also annotated by the musicians, covered aspects such as genre, instrumentation, mood, and tempo.

Rhythm Spots. We first obtained preliminary main melodies and rhythm spots by transcribing the music and applying algorithms. Musicians then refined these using

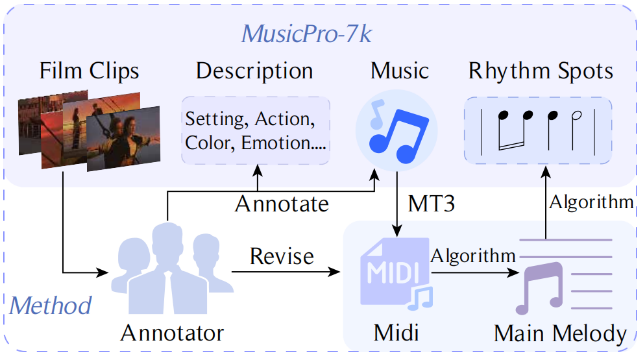

Figure 4. The structure and construction method of MusicPro-7k, which consists of film clips, description, music and rhythm spots.

professional music production software, achieving high-quality, highly aligned results. This part of the dataset serves as key data for guiding large language models in rhythm alignment, marking a pioneering effort that we believe will significantly advance this field in the future.

5. Experiment

5.1. Experiment Setup

We trained FilmComposer on MusicPro-7k, continuing from MusicGen-Melody [7]. It converges around 150 epochs on dadam optimizer [28] with 2000 updates per epoch. The batch size is 4 and learning rate starts at e-1 and smaller after. Gradient clipping is applied with a maximum norm of 1.0. The Adam optimizer is configured with betas [0.9, 0.95], weight decay of 0.1, and epsilon set to 1e-8. A cosine learning rate schedule is employed, incorporating a warm-up phase of 4,000 steps, 0.0 minimum learning rate ratio and 1.0 cycle length. We also used cache that stored audio input to improve the training process. The training takes approximately six days using one NVIDIA A6000 GPU. During the evaluation phase, we established a test set that encompasses all thematic styles, ensuring the fairness of the evaluation process.

5.2. Experiment Evaluations

For Music Quality, we use the Fréchet Audio Distance (FAD) [20] to measure the distance between feature distributions of real and generated audio. We also use the sampling rate to intuitively reflect whether the audio quality has reached a usable standard.

For Video-Music Correspondence, we use ImageBind ranking scores[14] to assess relative rankings of audiovisual consistency, adhering to the official procedure of applying the softmax operation.

For Diversity, where the attraction of music lies, we propose a new metric *Chroma-based Diversity* to reflect. Chroma-based Diversity D is calculated by averaging the pairwise differences between all distinct pairs:

$$D = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} Similarity(C_i, C_j), \quad (7)$$

where N is the total number of chroma feature matrices; C_i is the i-th chroma feature matrix in the set; $Similarity(C_i, C_j)$ is the overall similarity over all segments between C_i and C_j.

Previous work focuses on enhancing the quality and rhythm control of music generation, while they ignore the musicality and development of the music. As a more specialized form of music, the musicality and developmental aspects are particularly significant for film music.

For Musicality, We propose a new metric named *multi-agent music proficiency evaluation* to measure the musicality of music generated. In detail, the music will be transformed to ABC notation and then be delivered to the multi-agent, being evaluated in terms of mode, melody, harmony, rhythm and emotional expression. We conduct user study to certify the accuracy of this metric.

For development, we select three quantifiable aspects for evaluation—*Rhythm control*, *Dynamic Variation*, *Instrumentation*. The measurement of rhythm control involves using Madmom [3] to extract beat points—acoustic events occurring in the audio, such as drum hits or the onset of notes—and then comparing them using cross-correlation. A higher value in the cross-correlation indicates better rhythm control. To elucidate, cross-correlation assesses the similarity between two signals by shifting one signal relative to the other and calculating the dot product at each shift position, thus identifying whether the two signals exhibit similar patterns or shapes at specific points. Then it can be calculated as:

$$R_{xy}(k) = \sum_{t} x(t)y(t+k), \quad (8)$$

where $x(t)$ and $y(t)$ represent two audio sequences; k is the lag applied to $y(t)$ relative to $x(t)$.

To measure Dynamic Variation, decibel changes are used to represent dynamics, where lower cosine similarity values indicate greater similarity in dynamic variation. For Instrumentation, the Musicnn [29] model with certain non-instrument tags removed is used to extract the distribution of instruments in the music, and where lower cosine similarity values indicate greater similarity in instrumentation.

5.3. Experiments Results

Quantitative Analysis. As shown in Tab. 3, FilmComposer reaches state-of-the-art performance in every aspect. Firstly, FilmComposer gets the lowest KL and rhythm-controllable MusicGen is second only to it, meaning that our method generates the most similar results to ground

Method	KL↓	FAD↓	SR↑	ImageBind↑	Diversity↑	Musicality↑	Rhythm↑	Dynamic↓	Instru.↓
GT	0.000	0.000	48K	0.328	0.451	14.40	4042	0.000	0.000
CMT [9]	1.554	0.644	-	0.104	0.361	10.02	3153	0.801	0.519
Video2Music[19]	1.435	0.987	-	0.025	0.335	9.84	2001	1.018	0.535
M2UGen [26]	1.569	0.306	32K	0.114	0.373	9.40	3392	1.070	0.510
VidMuse [36]	2.035	0.376	32K	0.070	0.153	8.02	2467	0.892	0.527
MusicGen [7]	1.368	0.456	32K	0.108	0.133	8.80	3050	1.147	0.546
RC-MusicGen	1.320	0.219	32K	0.120	0.385	9.42	3646	0.820	0.506
FilmComposer	**1.209**	**0.207**	**48K**	**0.131**	**0.444**	**10.78**	**3834**	**0.767**	**0.434**

Table 3. **Quantitative Results.** The metrics from left to right are: KL, FAD, Sampling Rate, ImageBind ranking scores, Diversity, Musicality, Rhythm Control, Dynamic Variation, Instrumentation. RC-MusicGen is the abbreviation for Rhythm-Controllable MusicGen.

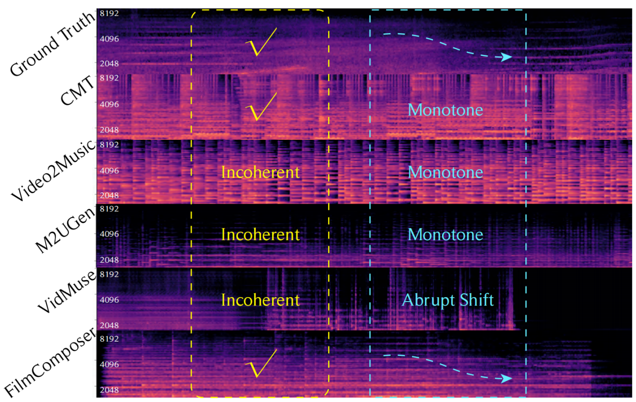

Figure 5. Qualitative Comparison results on spectrograms. The yellow box indicates where the music generated by Video2Music, M2UGen and VidMuse exhibits incoherence. The blue boxes show that CMT, Video2Music and M2UGen are generally monotone, while VidMuse presents abrupt shifts. In contrast, the ground truth and FilmComposer demonstrate clear layering.

truth. The lowest FAD and the highest sampling rate indicate the high audio quality of final results generated by FilmComposer. The sampling rate has reached the cinematic standard after multi-agent arranging and mixing, which other models cannot implement. CMT and Video2Music generate symbolic music that has no sampling rate. Next, the higher ImageBind ranking scores than in other works proves that FilmComposer reaches the highest relevance between music and film clips. Rhythm-controllable MusicGen has already achieved a high diversity score, and with the addition of multi-agent arrangement, FilmComposer further improves it significantly, reaching a level close to that of the ground truth, indicating that our results exhibit excellent richness and diversity.

Besides these basic aspects, music generated by FilmComposer also reaches the greatest level of musicality and development. Since CMT and Video2Music generate symbolic music, their results are more accurate when transcribed into ABC notation, giving them an advantage in notation-based musicality evaluation. However, despite FilmComposer generating waveform music, it ultimately surpasses all other methods in this regard, demonstrating its strong musicality in music generated. Next, the advantages of musical development are reflected in three aspects—rhythm control, dynamic variation, and instrumentation. Rhythm control is one of the major focuses of our work, and the metrics show that FilmComposer surpasses CMT and VidMuse which also focus on rhythm control, demonstrating the precision and effectiveness of our approach in rhythm control. For dynamic variation similarity, we find that CMT is a strong competitor and rhythm-controllable MusicGen, which does not specifically address dynamics, performed worse than CMT. However, FilmComposer ultimately surpassed it, achieving the best dynamic variation. In terms of instrumentation similarity, FilmComposer far outperforms all other methods, proving the notable contribution of the multi-agent arrangement in instrumentation.

Qualitative Analysis. Our qualitative analysis, presented in Fig. 5, highlights some limitations in other works. For CMT [9], the music is generally monotone. Similarly, Video2Music [19] produces monotonous outputs with excessive repetition. M2UGen's [26] music cannot develop with video. VidMuse's [19] music lacks substantive melody, with excessive noise and abnormal silences. Instead, our results are smooth and most similar to the ground truth in terms of rhythm and dynamics.

5.4. Ablation Studies

The rhythm-controllable MusicGen takes rhythm spots and description as conditions, so we conduct ablation studies to prove the effectiveness of these two modules. Additionally, we verify the crucial role of the multi-agent system in FilmComposer. The rhythm control modules, text control modules and multi-agent are removed in turn, and the evaluation results are shown in Tab. 4.

Text control. When the text control module is added, FAD metrics improve largely, which indicates increasing audio quality. It's noteworthy that the text control module also helps improve the rhythm control, as visual description can help the model decide the proper rhythm patterns. When the text control module is working, ImageBind gets a very high

Text	Rhythm	Agent	FAD↓	Rhythm↑	ImageBind↑
✓	×	×	0.246	2978	0.254
×	✓	×	0.319	2836	0.163
✓	✓	×	0.219	3646	0.265
✓	✓	✓	**0.207**	**3834**	**0.318**

Table 4. **Results of ablation study.** FAD validates improvement in audio quality, Rhythm control metrics assess the effectiveness of rhythm control, and ImageBind ranking scores verify the enhancement of visual-audio relevance.

ranking score, proving its notable contribution to enhancing the correspondence between music and film clips.

Rhythm control. After adding rhythm control module, the rhythm control metrics has been greatly improved, proving rhythm condition's significance. The rhythm control module also helps increase the audio quality and correspondence since the FAD metrics and ImageBind ranking rate improve.

Multi-agent. The inclusion of the agent led to improvements in FAD, Rhythm control, ImageBind ranking score, proving its effectiveness in enhancing audio quality, rhythm control and visual-audio relevance.

5.5. User Study

The best way to measure the professionalism of music production remains user study. We invited a total of 51 people, including 18 experts with professional backgrounds in film music production and over five years of professional experience, to evaluate music generated by FilmComposer, CMT [9] and Video2Music [19] in several aspects and give their preference separately. The aspects are: (1) Musicality: the innate ability to perceive, understand, and express musical elements such as mode, melody, harmony and rhythm; (2) Content Correspondence: The match between music and video in terms of emotion and semantics; (3) Rhythm correspondence: the alignment of rhythmic patterns with video; (4) Dynamic correspondence: the alignment of varying levels of loudness and intensity with video; (5) Instrumentation correspondence: the alignment of instrument usage with video. (6) Overall: the general assessment of music.

Results. Tab. 5 shows the results. More than 60% of both experts and non-experts preferred the music generated by FilmComposer for each item. This demonstrates FilmComposer's significant advantages in musicality, multiple aspects of correspondence, and overall quality. The user study yield results consistent with previous experiments, validating the reliability of our newly proposed metrics in Sec. 5.2.

6. Application

Since our framework imitates musicians, it can integrate into the actual music production process seamlessly and easily. So we design a system for convenient human-computer interaction, allowing users to edit and optimize almost every step to create music tailored to their needs.

Metrics	Experts	Non-Experts
Musicality	64.4%	78.8%
C.Content	82.2%	90.9%
C.Rhythm	73.3%	86.7%
C.Dynamic	75.6%	-
C.Instru	78.9%	-
Overall	81.1%	90.3%

Table 5. **Results of user study.** The data indicate the percentage of participants favoring FilmComposer. Non-Experts evaluate intuitive audiovisual alignment, without needing to consider more specialized musical knowledge. C. represents for Correspondence.

More details can be found in supplementary materials. **(1) Rhythm spotting:** Users can freely and conveniently drag, delete, or add rhythm spots to create their desired pattern, since rhythm is represented by a series of spots. **(2) Content descriptions:** The visual and motion descriptions are both in text forms, easy for users to modify. **(3) Musical notation:** After the music is transcribed into ABC notation, users can directly edit it to gain their desired melody. **(4) Arrangement and mix scheme:** Users can modify it before it is executed by the Execution Agent.

Non-experts can also use our system to easily and conveniently produce various types of music, such as background music for personal projects or social media content. The system can also serve educational purposes, helping beginners quickly understand the music production process and learn techniques through hands-on experimentation.

7. Conclusion

In this work, we propose the FilmComposer framework and the first film music dataset MusicPro-7k, implementing the film music production task and resulting in state-of-the-art performance. Our LLM-driven innovative approach has been proven highly effective across multiple metrics, which imitates musicians to conduct music production and combines symbolic and waveform music generation. Finally, the music produced by our method meets a cinematic standard suitable for direct use, exhibiting high quality, strong musicality, development, correspondence with film clips, and diversity, effectively addressing the challenges of high professionalism in film music production. FilmComposer can seamlessly integrate into the actual music production process, allowing user intervention in various aspects. This provides strong interactivity and a high degree of creative freedom. For discussions on model analysis, extensions, and limitations, please refer to the supplementary materials.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62402306) and Natural Science Foundation of Shanghai (No.24ZR1422400).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1, 2

[3] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178, 2016. 6

[4] G Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[5] Breakthrough. Pyscenedetect: Video scene cut detection tool, 2023. Version 0.6, Available at `https://github.com/Breakthrough/PySceneDetect`. 4

[6] Junhao Cheng, Xi Lu, Hanhui Li, Khun Loun Zai, Baiqiao Yin, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Autostudio: Crafting consistent subjects in multi-turn interactive image generation. *arXiv preprint arXiv:2406.01388*, 2024. 2

[7] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 4, 6, 7

[8] Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, et al. Composerx: Multi-agent symbolic music composition with llms. *arXiv preprint arXiv:2404.18081*, 2024. 2

[9] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045, 2021. 1, 2, 3, 7, 8

[10] Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12775–12785, 2024. 2

[11] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024. 1, 2

[12] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020. 2

[13] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. Mt3: Multi-task multitrack music transcription. *arXiv preprint arXiv:2111.03017*, 2021. 4

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 6

[15] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2

[16] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):166335, 2022. 4

[17] Kate Hevner. The affective character of the major and minor modes in music. *The American Journal of Psychology*, 47 (1):103–118, 1935. 4

[18] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 2

[19] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249:123640, 2024. 2, 3, 7, 8

[20] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018. 6

[21] Yun-Han Lan, Wen-Yi Hsiao, Hao-Chung Cheng, and Yi-Hsuan Yang. Musicongen: Rhythm and chord control for transformer-based text-to-music generation. *arXiv preprint arXiv:2407.15060*, 2024. 1, 2

[22] Hyun Lee, Taehyun Kim, Hyolim Kang, Minjoo Ki, Hyeonchan Hwang, Sharang Han, Seon Joo Kim, et al. Commu: Dataset for combinatorial music generation. *Advances in Neural Information Processing Systems*, 35:39103–39114, 2022. 2

[23] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 3

[24] Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. Diff-bgm: A diffusion model for video background music generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27348–27357, 2024. 1, 2

[25] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 2

[26] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. M^2 ugen: Multi-modal music understanding and

generation with the power of large language models. *arXiv preprint arXiv:2311.11255*, 2023. 1, 2, 3, 7

[27] Daniel McKee, Justin Salamon, Josef Sivic, and Bryan Russell. Language-guided music recommendation for video via prompt analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14784–14793, 2023. 2

[28] Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *IEEE Transactions on Signal Processing*, 70:6065–6079, 2022. 6

[29] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*, 2019. 6

[30] Laure Prétet, Gael Richard, and Geoffroy Peeters. Cross-modal music-video recommendation: A study of design choices. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021. 2

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4

[32] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017. 5

[33] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 4

[34] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, 2024. 2

[35] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10564–10574, 2022. 2

[36] Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Xiaoqiang Huang, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. *arXiv preprint arXiv:2406.04321*, 2024. 2, 3, 7

[37] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 2

[38] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023. 4

[39] xlvector. abcmidi. Online, 2016. `https://github.com/xlvector/abcmidi`. 4

[40] Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. Cross-modal variational auto-encoder for content-based micro-video background music recommendation. *IEEE Transactions on Multimedia*, 25:515–528, 2021. 2, 3

[41] Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. Musicagent: An ai agent for music understanding and generation with large language models. *arXiv preprint arXiv:2310.11954*, 2023. 2

[42] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao. Long-term rhythmic video soundtracker. In *International Conference on Machine Learning*, pages 40339–40353. PMLR, 2023. 2

[43] Donghuo Zeng, Yi Yu, and Keizo Oyama. Audio-visual embedding for cross-modal music video retrieval through supervised deep cca. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 143–150. IEEE, 2018. 2

[44] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision*, pages 182–199. Springer, 2022. 2, 3

[45] Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15637–15647, 2023. 1, 2