FilmComposer: LLM-Driven Music Production for Silent Film Clips

Supplementary Material

A. MusicPro-7k Details

Our dataset MusicPro-7k covers films of all common genres, and the distribution is shown in Fig. 6. Since most film genres include drama and many clips prominently feature dramatic elements, the proportion of the drama theme in the MusicPro-7k is relatively high.

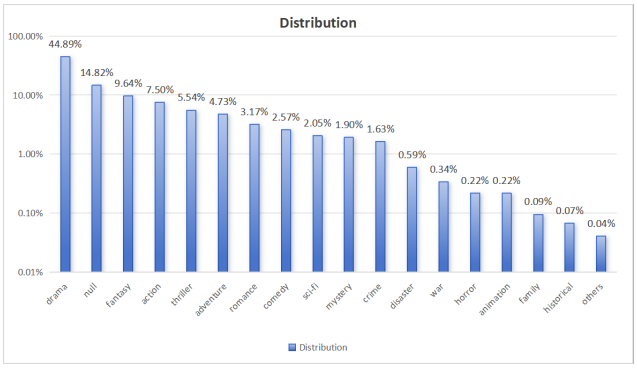


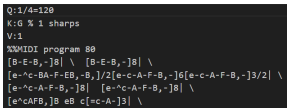
Figure 6. The distribution of MusicPro-7k. "Null" represents film clips without a clearly defined theme, while "others" includes themes such as biography, sport, and documentary. The vertical axis in the figure uses a logarithmic scale.

Musicians need to annotate visual descriptions, and to ensure the effectiveness of the annotations, predefined visual attributes need to be set to cover all film clips. The complete predefined visual attributes are present in Tab. 6.

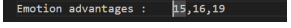
B. Experiment Details

B.1. Intermediate Results

In multi-agent assessment, arrangement and mix module, there are several intermediate outputs, such as ABC notation after the transcription of generated melody, the standard and corresponding output score of multi-agent assessment system, and the arrangement and mix scheme. Those intermediate results are shown in Fig. 7.



(a) ABC Notation



(b) Output of Evaluation Agent

(c) Arrangement and Mix Scheme

Figure 7. The output results are respectively from the transcription, multi-agent assessment system, and multi-agent arrangement and mixing system.

category	*labels*
Setting	*road, home, forest or jungle or canyon, corridor, car, sea, Kitchen and dining area, ruins, hospital, garden, space ship or outer space, hotel, workplace, bar, stairs, airport, city, ship, castle, prison, lab, beach, train station, desert, rooftop, bus, police station, elevator, stage, tunnel, alley, train, factory, square, school, playground, tent, theater, cave , church, bridge, pier, military base, night view, balcony, store, market, court, ballroom, under water, parking lot, swimming pool, rural, casino, grassland, library or bookstore, cemetery, farm, subway station, cliff, coffee shop, rain, street, null*
Brightness	*mild, bright, dull, somber, dark, glaring, contrasting*
Color Hue	*Blue, Green, Red, Pink, Yellow, Orange, Purple, Black, White, Brown, Gray*
Behavior	*shoot gun, run, call, do intimacy, kiss, fight, drive car, read, drink, smoke, climb, fall down, ride horse, eat, applaud and cheer, dance, cry, hug, write, drive plane, work, sleep, laugh, quarrel, ride motorcycle, kill or attack, pursue and arrest, play instrument, swim, fire, faint, play ball games, take shower, play games, boating, cook, sing, ride bicycle, paint, do housework, pray, sit, talk, battle, speech, goodbye, gaze, escape, open door, null*
View Scale	*long shot, full shot, medium shot, close-up shot, extreme close-up shot*
Emotion	*Dignified, Sad, Dreamy, Calm, Graceful, Joyous, Exciting, Vigorous, Nervous, Angry, Fear, Humorous, null*
Theme	*drama, fantasy, action, thriller, adventure, romance, comedy, sci-fi, mystery, crime, disaster, war, horror, animation, family, historical, biography, sport, documentary, null*

Table 6. The categories of visual descriptions and their corresponding labels.

B.2. Experiments on Agents

To validate that the agents function as intended and to elucidate the specific methodologies employed for their improvement and evaluation, we conducted comprehensive experiments. We organize the agents logically and train them by first setting a speaking order and then iteratively incorporating prompt engineering techniques - Role-Play, Few-Shot Prompting and Chain of Thought. During training, the accuracy of agents' responses in assessment, arrangement, and mix tasks improved significantly, reaching

Table 7. Agents' response accuracy during iterative optimization.

task	organized	+ speaking order	+ Role-Play	+ Few-Shot Prompting	+ Chain of Thought
assessment	77%	80%	82%	85%	92%
arrangement	47%	53%	57%	64%	89%
mix	57%	67%	70%	77%	91%

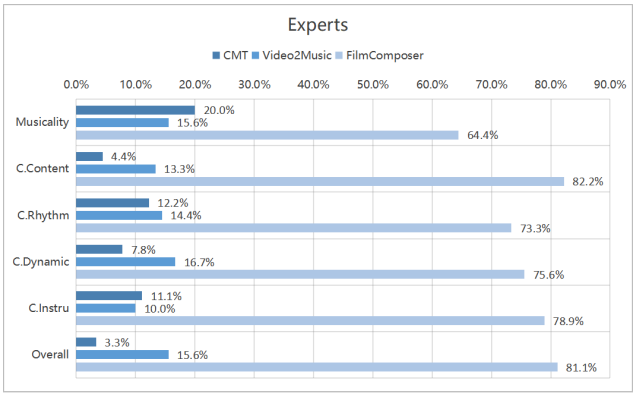


Figure 8. The bar chart of the user study of experts, reflecting the scores of the three methods across each item.

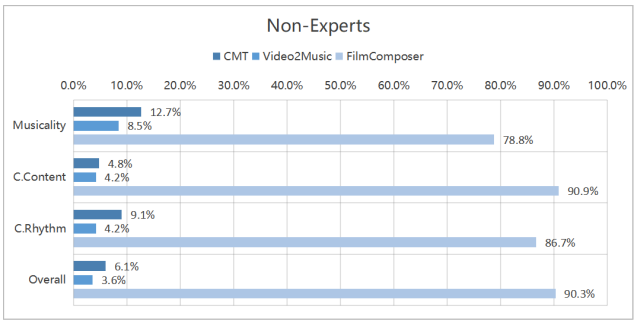


Figure 9. The bar chart of the user study of non-experts, reflecting the scores of the three methods across each item.

around 90%, shown in Tab. 7.

Upon examining the agents' behavior, we found that after setting the speaking order and incorporating Role-Play, the agents never reverse their roles. According to the data, Chain of Thought contributes the most to improving the accuracy of the agents' responses. Among the tasks, the arrangement task exhibits the lowest accuracy before training but demonstrates the most significant improvement after training.

B.3. User Study Details

In our user study, experts and non-experts were asked to select their preferred results for each item. The results were generated by FilmComposer, CMT, and Video2Music. The scores for each method in each item are presented in Fig. 8 and Fig. 9.

B.4. Final Results

Multiple experimental results and user studies yielded consistent outcomes, demonstrating the strong performance of our method. To further validate this, we provide a video demo and comparison of the results with other models on our project page, showcasing our final results.

C. Discussion

C.1. Analysis of Model Capabilities

In addition to the information provided by the video, music is also influenced by elements defined by filmmakers, such as mood, genre, and style. Therefore, during the fine-tuning process, we pay particular attention to preserving the model's ability to process professional descriptions of music as input. By simply incorporating the desired musical elements into the existing visual descriptions, the model can integrate and process these inputs to generate music that aligns with all specified descriptions.

C.2. Extensions

The richness and scalability of instruments represent a notable strength of our approach. Our method has covered 39 types of instruments and considered different playing techniques, covering instruments commonly used in film music. Moreover, adding more instruments is easy, as the selection range of Instrument-Agent can easily expand, with numerous sound sources in DAW to be integrated.

Music composition for films sometimes requires consideration of long-term coherence, as the music for a current scene may be influenced by preceding scenes or the storyline. FilmComposer can be extended to effectively capture those long-term dependencies. Module 1 (Visual Processing for Film Clips) can be augmented with a Long Video-Language Understanding model to capture long-term information, which is then fused with the information of current clip via LLMs. This hybrid approach generates effective prompts that simultaneously emphasize localized visual cues and global narrative coherence.

In addition to music, foley constitutes an indispensable auditory component in film productions. A notable strength of FilmComposer lies in its ability to extend and efficiently generate foley that is precisely synchronized with film clips. Specifically, this is achieved by incorporating an event timestamp condition, which is analyzed and ex-

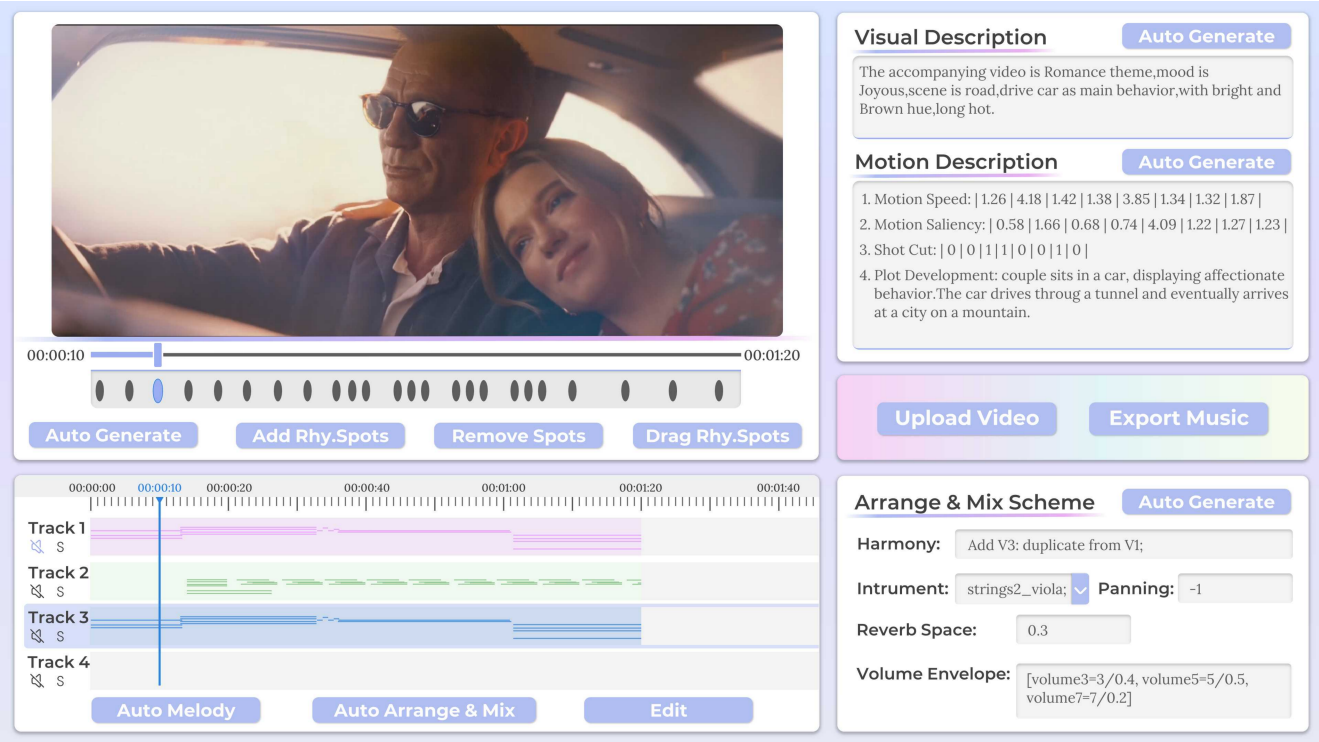


Figure 10. The FilmComposer-based interaction system interface.

tracted from the video, thereby enabling the audio generation model to produce foley with accurate temporal alignment to the event.

C.3. Limitations

Our extensive instrument repository might lead to potential incompatibility between an instrument's pitch range and melodic requirements. Accurately assessing whether a selected instrument's range fully encompasses the melodic pitch spectrum remains technically demanding. This limitation might result in the omission of notes exceeding the instrument's playable range.

Future enhancements could address this constraint. Voice-Part-Agent can be introduced to decompose music tracks into distinct voice-part roles, enabling Instrument-Agent to make more informed range-based selections. In addition, implementing algorithmic assistance in instrument selection could establish a dual verification mechanism.

D. Application Details

Fig. 10 shows the interface of the interaction system. The upper left displays the imported video, with tools below allowing for quick rhythm point generation and various adjustments. The upper right provides auto descriptions generation and easy edition. The lower right section is for creating and modifying arrangement and mix plans, and the lower left enables fast music track generation for previewing.