FrameBeats: A prototype that compose the sound effects for input visuals

Sen Zhang*
Xi'an Jiaotong-Liverpool University

Chengyang Song†
Xi'an Jiaotong-Liverpool University

Lingyun Yu‡
Xi'an Jiaotong-Liverpool University

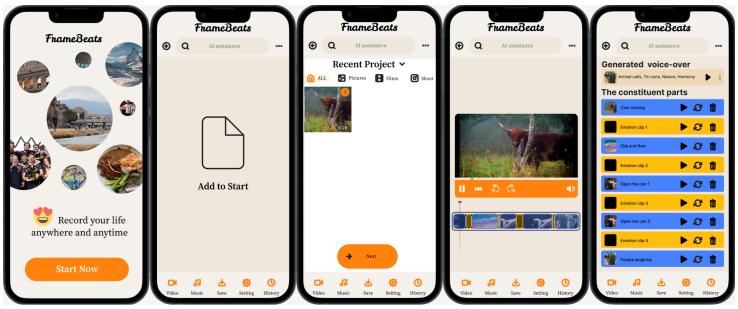

Figure 1: The prototype-FrameBeats

Abstract

This study investigates the challenges users encounter when matching sound effects to visual elements, such as videos, through a questionnaire survey. Based on the insights gathered, we developed a prototype system that automatically generates sound effects corresponding to input video and image content. We evaluate the system's effectiveness and discuss potential avenues for future research and development.

Index Terms: AI generated music, Music visualization

1 Introduction

Sound effects play an important role in videos [2]. With the emergence of short videos, the demand for sound effects has also increased. AI generation tools are becoming increasingly popular among individuals with limited professional expertise who require sound effects that align with their video content, thanks to the AI strong analytical features and user-friendly design [1]. As one of the most popular apps for non-experts to share their videos, TikTok has introduced a feature that allows users to upload videos and automatically generates music for them. The app analyzes the elements in the video and uses them as the basis for creating the accompanying music. However, adding lyrics directly is too rigid, and in some situations, the generated content is too abrupt. We conducted a user requirements test with 36 participants to explore the requirements for composing music and to gather design problems. We have discovered that many users experience issues with the **Efficiency** and **Effectiveness** of composing sound to visuals. To solve these issues, we design a prototype. Our prototype aims to use AI to generate appropriate sound effects for recognized elements and the transition sound effects from the entire video, which is a form that is more similar to traditional artificial composing.

2 Related Work

The traditional sound design work can improve the perception of the audience. Wierzbicki [5] explored combining sound with emotion makes good effects on the audience. Görne [3] also showed that the sound effects related to the context can improve the feelings of the audience. The article from Flückiger [2] pointed out that the sound effects with the film objects can create complex environments to inspire the audience's recall of objects related to the sound effects. These guidelines are intended for editors and not non-expert users. Nakashima et al. [4] preliminarily verified the feasibility of AI personalized music generation. Xie et al. [6] explored the use of LLM to implement composing music for silent films. Moreover, Xu et al. [7] found that non-experts can enhance emotional engagement by incorporating personal multimodal materials into complete music clips. AI can now be used to generate music efficiently, but most designs yield random results or separate generations without considering a cohesive song structure. Our project aims to combine AI-generated music with ease of use and the positive impact of compositing sound throughout the video. We seek to utilize AI to create sound effects that enhance the audience's experience based on the input visuals.

3 Formative Study

We conducted an online survey with 9 checkbox questions about the frequency and usage of existing editing tools, along with 2 open-ended questions addressing concerns about composing sound for videos and experiences with music generation functions. A total of 36 participants between the ages of 18 and 25 took part in this survey. All participants have experience in posting short videos, while five of them are full-time content creators. The survey reveals that 32 participants have encountered difficulties in finding suitable sound effects to match their video clips. Three participants further reported that the sound effects they found often did not align with the duration of their videos, requiring additional editing effort. Moreover, two professional video bloggers expressed a preference for avoiding sound effects that are overused by others. Thirty-one participants agreed that using AI-generated sound effects could help resolve these issues and improve their editing efficiency.

Participants also reported the issues they encountered while using existing generative AI tools. Twenty-seven participants acknowledged having used AI-supported tools for audio material generation, such as TikTok's AI music generation feature. Twenty-six of them indicated that when the input includes multiple visual elements, the AI-generated results can occasionally appear unnatural. They expressed a desire for smoother transitions between elements and more seamless integration of the generated content.

Our formative study reveals two issues with using AI to generate sound effects for videos. (1) **Efficiency**: Users spend a lot of time

*Sen.Zhang22@student.xjtlu.edu.cn
†Chengyang.Song22@student.xjtlu.edu.cn
‡Lingyun.Yu@xjtlu.edu.cn

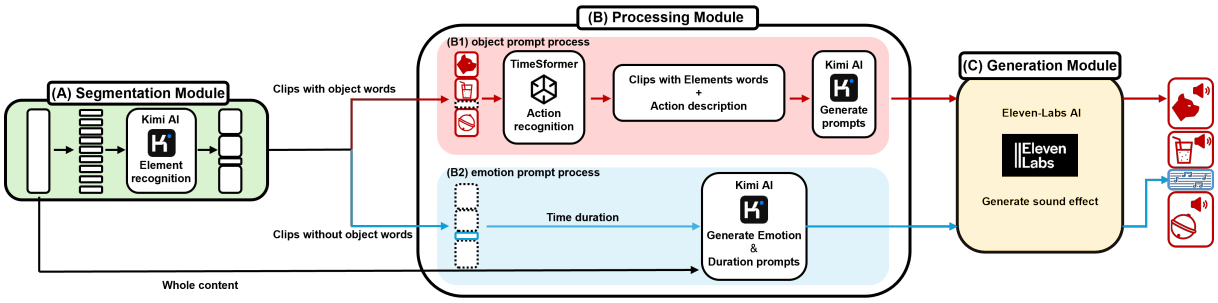

Figure 2: Prototype Framework

finding the appropriate sound and modifying the materials into a format that meets their requirements also demands additional time. (2) **Effectiveness**: The connection between elements is problematic. A direct connection may disrupt the user experience.

4 Prototype Design

To address these issues, we created a prototype to streamline the creation of sound effects based on visual elements, as illustrated in Figure 1. As Figure 2 shows, the prototype consists of three modules: **(A) Segmentation Module**, **(B) Processing Modules**, and **(C) Generation Module**. In the **(A) Segmentation Module**, the input images or videos are segmented into individual frames. The frames are then processed by the Kimi AI model, which identifies the objects in each frame and returns the description of the objects. Frames with similar descriptions are grouped into a single unit. This module will also segment the input visuals into clips containing recognized objects and clips where no objects are identified. Next, the two types of clips are handled by two separate **(B) Processing modules**. The segments containing the objects will be processed by the **(B1) Object prompt** process. The video clips are input to the TimeSformer AI model to recognize specific actions, and then the words of objects and actions are added to the Kimi AI model to formulate the prompt. For example, the object of the clips is the horse and its action is running. The duration of segment is five seconds. So the prompt to the model is "Give me a sound effect of a running horse for 5 seconds". In the **(C) Generation module**, the ElevenLabs AI model accepts the prompt, generating the sound effects of running. If the module does not contain any actions, it will just give the clips with objects words to the Kimi AI model to generate the prompt. For instance, the segment is only the horse with no actions for 5 seconds. The prompt will be "Give me a sound effect of a horse for 5 seconds". The ElevenLabs AI model in the **(C) Generation module** may generate the sound of the neigh. The second type of clips is the segment without objects. The segments will be received by the **(B2) Emotion prompt** process. Due to the lack of specific objects in these segments. The Kimi AI model will combine emotion clues of the entire visuals and the time duration to generate the prompts. For example, the whole video is consist of the clips with warm color background. The prompt generated by Kimi may be "Give me the happy sound effects for 5 seconds." The ElevenLabs AI model in the **(C) Generation module** generates the corresponding sound effects according to the prompt. The results from the **(C) Generation module** are added to the audio track. After all the clips have been handled, the whole composing sound are integrated. Moreover, the prototype provides a preview of generated composition result and the detailed list of generated sound effects. Users can listen to these results separately and decide to delete it or regenerate it.

Therefore, our system is capable of extracting key elements from input images or videos and leveraging LLM to expand these elements into descriptive sentences for prompt. These prompts are then used to generate the sound effects, which supports two types of music creation: **Object-driven composition**: It will generate the sound effects according to the descriptive sentences with objects and actions(optional); **Emotion-driven composition**: It will combine the entire visuals emotion to generate the sound effects.

5 Conclusion

We developed a prototype that processes visual inputs and generates corresponding sound effects. It facilitates seamless transitions between different visual elements. By reducing the time spent sourcing and editing assets, our prototype enhances user efficiency and supports more effective creative workflows. However, we acknowledge certain limitations to be improved in future work. For instance, the current prototype only supports video inputs with consistent emotional tone. When a video contains multiple emotional cues or has varying emotional tones throughout, the generated sound effects may conflict with the user's actual perception. Additionally, users may desire more fine-grained control over the generated content, which necessitates the integration of other input modalities, such as text, to more effectively convey the sound effects or emotional expressions they seek.

References

[1] C.-c. K. Chang. The creative commons solution: Protecting copyright in short-form videos on social media platforms. *International Journal of Law Management and Humanities*, 6(3):583–615, 2023. 1

[2] B. Flückiger. Sound effects: Strategies for sound effects in film. In G. Harper, R. Doughty, and J. Eisentraut, eds., *Sound and Music in Film and Visual Media: An Overview*, pp. 151–179. Continuum, New York, NY, 2009. 1

[3] T. Görne. The emotional impact of sound: A short theory of film sound design. In P. Kessling and T. Görne, eds., *KLG 2017. klingt gut! 2017 – International Symposium on Sound*, vol. 1 of *EPiC Series in Technology*, pp. 17–30. EasyChair, Manchester, UK, 2019. doi: 10.29007/jk8h 1

[4] S. Nakashima, Y. Imamura, S. Ogawa, and M. Fukumoto. Generation of appropriate user chord development based on interactive genetic algorithm. In *2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 450–453, 2010. doi: 10.1109/3PGCIC.2010.76 1

[5] J. Wierzbicki. *Sound Effects/Sound Affects: 'Meaningful' Noise in the Cinema*, pp. 153–168. Palgrave Macmillan UK, London, 2016. doi: 10.1057/978-1-137-51680-0_11 1

[6] Z. Xie, Q. He, Y. Zhu, Q. He, and M. Li. Filmcomposer: Llm-driven music production for silent film clips. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 13519–13528, June 2025. 1

[7] W. Xu, L. Zhao, H. Song, X. Song, Z. Lu, Y. Liu, M. Chen, E. G. Lim, and L. Yu. Mozualization: Crafting music and visual representation with multimodal ai. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2025. 1