Ourmuse: Plot-Specific AI Music Generation for Video Advertising

Hyeseong Park, Myung Won Raymond Jung, Sanjarbek Rakhmonov, and Sngon Kim
AKA AI
Seoul, South Korea
julie@akaintelligence.com, rjung@akaintelligence.com, sanjarbek@akaai.kr, sgkim@akaon.com

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10826083

***Abstract*—The integration of music and video is a pivotal aspect of creating impactful advertisements. This study explores the application of Artificial Intelligence (AI) in generating music tailored for video advertisements through a system we developed, named Ourmuse. By leveraging advanced Generative AI models, Ourmuse aims to enhance the synchronization between music and video content. The methodology involves extracting and analyzing still images and texture context from the video. A critical aspect of our approach is the plot-based training of the AI model by identifying transition points between video scenes, where the video is segmented into distinct parts rather than being treated as a single unit. This segmentation ensures that Ourmuse produces music that is coherent and contextually appropriate for each segment, enhancing the overall impact and cohesion of the final advertisement. By focusing on multi-modal inputs and plot-specific training, this research aims to develop a robust AI model capable of revolutionizing the music composition process for video advertisements.**

***Index Terms*—artificial intelligence (AI), video advertisements, music generation, multi-modal training, plot-based training**

I. Introduction

In the rapidly evolving landscape of digital advertising, the synchronization of music and video has become increasingly vital for creating compelling and memorable content. Music not only enhances the emotional resonance of an advertisement but also plays a crucial role in reinforcing the narrative structure and pacing of the visual content. Traditionally, the process of integrating music with video has involved either adapting existing music to fit the visual elements or composing new music specifically tailored to the pre-edited video [1]. However, these conventional methods are often resource-intensive and may not fully capture the dynamic interaction between music and visual storytelling.

Recent advancements in Artificial Intelligence (AI) offer new opportunities to optimize this process. By leveraging Generative AI models, there is potential to revolutionize the way music is composed and synchronized with video content, making the process more efficient while maintaining creative integrity [2], [3]. This paper proposes an innovative approach to AI-driven music generation for video advertisements through a system we developed, named Ourmuse, which emphasizes the importance of *plot-based training* to ensure that the music produced is contextually aligned with the narrative flow of the video.

In the context of video advertisements, two primary approaches are commonly employed: matching video to pre-existing music or composing music to align with the edited video. The former is typically used when music plays a central role in the advertisement, driving the emotional and thematic elements. In contrast, the latter approach is preferred when the video's narrative structure is paramount, requiring the music to adapt to the visual pacing and transitions. This approach often involves segmenting the music to match specific parts of the video, such as the introduction, climax, or conclusion, while using sound effects or narration to fill the remaining gaps.

Our proposed AI model, Ourmuse, is designed with these nuances in mind, recognizing the importance of plot-specific training. The model processes multi-modal inputs, including still images, textual context, and video content [4]. The purpose is to generate music that is not only coherent across the entire advertisement but also responsive to the distinct narrative segments, or plots, within the video. By training the AI on these segmented plots rather than on the entire video as a single unit, we ensure that the music generated is both contextually appropriate and enhances the overall impact of the advertisement.

The Ourmuse system leverages big data methodologies by utilizing large-scale, multi-modal datasets that include diverse video content, still images, and text-based inputs. These datasets enable Ourmuse to train on a wide variety of video types and contexts, enhancing its ability to generate contextually relevant music for a broader range of advertisements. The integration of big data is essential in refining Ourmuse's plot-specific music generation, allowing the model to adapt to the nuanced variations in advertisement formats and audiences.

Furthermore, a key challenge in developing this AI model lies in identifying and processing the transition points between different plots within a video. These transition points are critical, as they often signal shifts in tone, tempo, or narrative focus, which the music must reflect. The AI must be trained to detect these transitions and adjust the musical output accordingly, ensuring a seamless and engaging viewer experience.

This paper aims to provide a foundational framework for the development of an AI-driven music composition model tailored for video advertisements. While the model is still in its conceptual stage, the ideas presented here lay the groundwork for a more detailed exploration and eventual implementation. By focusing on plot-based training and multi-modal input processing, we propose a system that could significantly enhance

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

the way music is integrated with video content, offering new possibilities for the advertising industry.

In summary, this study introduces a novel approach to AI-driven music generation through Ourmuse, emphasizing the segmentation of video content into distinct plots for more targeted and contextually relevant music composition. This approach not only addresses the limitations of traditional methods but also paves the way for future developments in AI-based creative tools, with the potential to transform the landscape of video advertising.

II. Related Technologies

A. Music Generation

Music generation using artificial intelligence has significantly advanced in recent years, leveraging various techniques ranging from rule-based systems to more sophisticated machine learning models. Early approaches often relied on symbolic AI, where predefined rules and heuristics were used to generate music compositions. However, the advent of deep learning has transformed the field, with models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and, more recently, Transformer [5], enabling the creation of music that is more complex and emotionally resonant. The Music Transformer model, for instance, has been particularly influential in generating long-term structured music compositions by capturing dependencies over extended sequences Similarly, advancements in Generative Adversarial Networks (GANs) have facilitated the generation of high-quality, diverse musical pieces by learning from large datasets of existing music.

While significant advancements have been made in AI-driven music generation, most current approaches generate continuous music without adapting to specific narrative segments. These methods lack the flexibility required for applications such as video advertising, where synchronization with visual transitions and shifts in emotional tone is critical. Research on plot-based training, where music generation is tailored to distinct "plots" within a video, remains limited. Ourmuse addresses this gap by segmenting videos and adjusting musical elements like tempo and rhythm to fit each plot, enhancing alignment between music and video content. This targeted approach fulfills a critical need in video advertising, creating a cohesive and engaging viewer experience.

B. GPT

Large Language Models (LLMs) have been increasingly adopted across a broad range of applications [6]. In particular, Generative Pre-trained Transformers (GPT) have shown remarkable capabilities in processing and generating human-like text, which can be extended to various creative tasks, including music generation. In our model, GPT can be utilized to process multi-modal inputs, such as still images, textual context, and other settings extracted from video content, to generate music that is contextually aligned with the visual and narrative elements of the video. This approach leverages the strengths of GPT in understanding and generating coherent sequences based on diverse inputs. Additionally, tools like MuseNet, an extension of GPT, demonstrate the potential to directly generate complex, multi-instrumental music by learning from vast amounts of musical data. MuseNet can create compositions that are not only harmonically rich but also adaptable to specific contextual cues provided by video content. These GPT-based technologies offer powerful new avenues for integrating AI-driven music generation into video advertisements, enhancing the synergy between audio and visual elements.

III. Proposed Design

The proposed design for our AI-driven music generation system, named Ourmuse, focuses on the dynamic interaction between video content and music, with particular emphasis on creating music that enhances the narrative and emotional impact of video advertisements. This approach recognizes that in advertising, music often plays a critical role in conveying the message, either by driving the visual narrative or by complementing it. Depending on the specific requirements of the advertisement, music may need to be synchronized with pre-existing video content or composed in a way that allows the video to adapt to the musical structure. The overall structure of Ourmuse is depicted in Fig. 1.

A. Overview of the System

Our system is designed primarily for business-to-business applications, where the AI model acts as a collaborative tool between content creators and advertisers. The core idea is to develop a model that can process multi-modal inputs—such as still images, textual context, and other metadata extracted from the video—and generate music that is tailored to these inputs. The generated music should be contextually relevant, enhancing the visual content and ensuring a cohesive and engaging advertisement.

B. Plot-Based Training Approach

A key innovation in our design is the use of plot-based training. Rather than treating the entire video as a single unit, we segment the video into distinct plots, each representing a different part of the narrative structure. These plots might correspond to the introduction, development, climax, and conclusion of the advertisement. By training the AI on these individual segments, we aim to produce music that is not only consistent with the overall theme but also responsive to the specific needs of each plot. This segmentation is crucial because it allows the AI to adjust the musical elements—such as tempo, rhythm, and melody—according to the narrative flow, ensuring that the music aligns seamlessly with the visual transitions.

C. Multi-Modal Input Processing

Our AI model leverages Generative AI technologies, such as GPT and MuseNet, to process the multi-modal inputs. For instance, still images extracted from the video serve as a visual reference, while the textual context provides insights

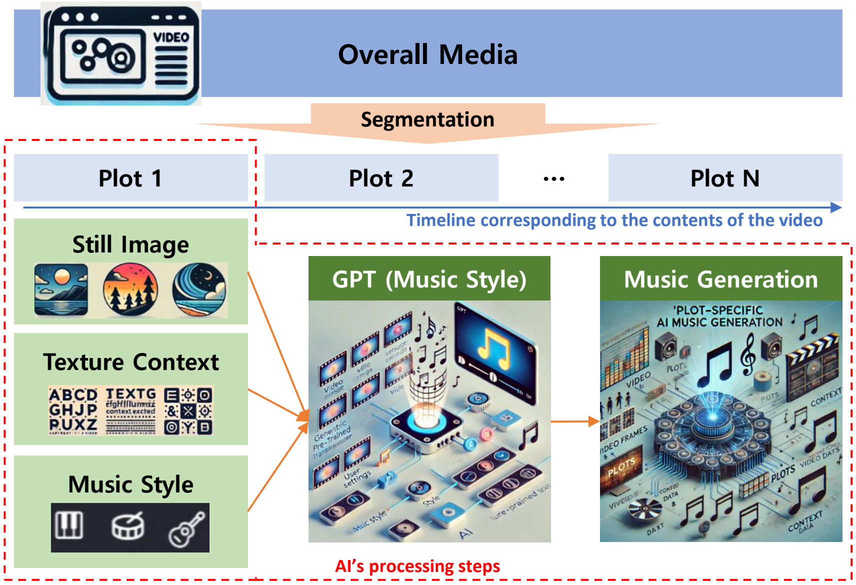

Fig. 1. The structure of the Proposed Ourmuse Design

into the thematic and emotional tone of the advertisement [7]. Additionally, metadata such as target audience demographics and brand messaging are incorporated into the input layer. The AI uses these inputs to generate music that not only fits the narrative but also resonates with the intended audience. By combining different types of data, our model can create music that is finely tuned to the specific requirements of each advertisement.

D. Transition Point Detection and Adaptation

A critical aspect of the proposed design is the detection and adaptation to transition points within the video. Transition points are moments in the video where significant shifts occur—such as changes in scene, tone, or pacing—that require corresponding changes in the music. The AI model is trained to identify these points and adjust the music accordingly. This could involve altering the tempo, introducing new musical motifs, or even changing the genre to better match the evolving narrative. By focusing on these transitions, the system ensures that the music remains engaging and contextually appropriate throughout the entire advertisement.

E. Iterative Development and Customization

Given the collaborative nature of the advertising process, our design incorporates iterative development and customization. After the initial music generation, the AI allows for feedback and adjustments from the content creators. This could involve tweaking specific segments of the music, adjusting the balance between music and sound effects, or reworking the transitions between plots. The system is designed to be flexible, enabling multiple iterations until the final product meets the creative and strategic goals of the advertisement.

F. Integration with Existing Tools

To ensure practicality and ease of use, our AI model is designed to integrate with existing music production and video editing tools. This integration allows content creators to seamlessly incorporate AI-generated music into their workflow, making it easier to produce high-quality advertisements efficiently. The system's outputs are compatible with standard audio and video formats, ensuring that the generated music can be easily incorporated into the final advertisement without additional processing.

G. Application

The proposed design offers a robust framework for AI-driven music generation tailored specifically for video advertisements. In our current trials, Ourmuse successfully generated complementary musical elements, such as piano and guitar sounds, when initial inputs included drum and bass patterns for each plot. While traditional generative AI has focused on creating full compositions, Ourmuse demonstrated its capability to produce complementary music elements based on plot-specific input conditions. Future work will involve a more detailed implementation and testing of the model, with a focus on refining the plot-based training approach and enhancing the model's ability to handle complex multi-modal inputs. Additionally, exploring new generative techniques, such as advanced Transformer models and cross-modal learning, could further improve the system's capabilities and expand its applications to other media formats.

IV. Future Research Directions

The development of AI-driven music generation models, particularly those tailored for video advertisements, such as our proposed system Ourmuse, presents several promising avenues for future research [8]. As Ourmuse continues to evolve, it is essential to explore additional aspects that can further enhance the effectiveness, flexibility, and applicability of the system. This section outlines key areas for future research that could significantly impact the development and refinement of AI-driven music composition.

A. Enhancement of Plot-Based Training Models

While plot-based training is a cornerstone of our proposed design, further research is needed to optimize this approach. One area of focus could be the development of more sophisticated algorithms for segmenting videos into plots. This involves not only recognizing the narrative structure of a video but also understanding the emotional and thematic shifts that occur within each segment. Advanced machine learning techniques, such as reinforcement learning or unsupervised learning, could be explored to improve the accuracy and granularity of plot segmentation, leading to even more contextually relevant music generation.

B. Exploration of Cross-Modal Learning

Another promising area for future research is cross-modal learning, where the AI model learns to correlate different types of inputs—such as audio, visual, and textual data—more effectively [9]. Cross-modal learning could enable the AI to better understand how music interacts with other elements of the advertisement, resulting in more cohesive and impactful outputs. By exploring techniques like multi-modal transformers or cross-attention mechanisms, researchers could develop models that are better equipped to handle the complex interplay between various forms of media.

C. Incorporation of Real-Time Feedback Loops

Incorporating real-time feedback loops into the AI model could greatly enhance its utility in a collaborative creative environment. Future research could focus on developing systems that allow content creators to provide real-time feedback during the music generation process. This could involve integrating user interfaces that enable adjustments to be made on-the-fly, with the AI model responding dynamically to these inputs. Such a system would not only streamline the creative process but also ensure that the final product aligns closely with the creative vision of the stakeholders involved.

D. Expansion to Other Media Formats

While the current focus of our design is on video advertisements, the underlying principles and technologies could be adapted for use in other media formats, such as film, television, or interactive media like video games. Future research could explore how plot-based music generation models can be tailored to suit these different contexts, each of which has unique requirements in terms of narrative structure and audience engagement. Expanding the applicability of the model could open up new opportunities for AI-driven music generation across a wide range of industries.

E. Ethical Considerations and Bias Mitigation

As AI continues to play a larger role in creative processes, ethical considerations become increasingly important. Future research should address the potential biases in AI-generated music, particularly in relation to cultural representation and diversity. Ensuring that AI models do not perpetuate harmful stereotypes or exclude certain musical traditions is crucial. Researchers should explore methods for incorporating diverse datasets and implementing bias mitigation strategies within the training process, promoting inclusivity and fairness in AI-driven music generation.

V. Conclusion

The integration of music and video is crucial in creating impactful advertisements. This paper has explored how Artificial Intelligence (AI) can enhance the synchronization between music and visual content, introducing Ourmuse, an innovative AI-driven music generation system. Ourmuse emphasizes plot-based training and multi-modal inputs to produce music that is contextually relevant and emotionally resonant.

By segmenting video content into distinct plots and utilizing advanced AI models like GPT and MuseNet, Ourmuse generates music that dynamically aligns with the narrative flow of the video. This approach offers a more flexible and efficient solution for content creators and advertisers, addressing limitations of traditional methods.

Several avenues for future research, including enhancing plot-based training and exploring cross-modal learning, have been identified to further refine Ourmuse. Although still in the conceptual stage, Ourmuse lays the groundwork for future innovations in AI-driven music composition, with the potential to transform how music is integrated into visual content.

References

[1] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[2] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[3] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[4] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath *et al.*, "Discovering language model behaviors with model-written evaluations," *arXiv preprint arXiv:2212.09251*, 2022.

[7] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[8] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, *Deep learning techniques for music generation*. Springer, 2020, vol. 1.

[9] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.