GENERATING VIDEO SOUNDTRACKS WITH CONTEXTUAL VISUAL THUMBNAILS

Mina Huh[1,2] **C. Ailie Fraser**[2]
Dingzeyu Li[2] **Mira Dontcheva**[2] **Bryan Wang**[2]
[1] The University of Texas at Austin
[2] Adobe Research
minahuh@cs.utexas.edu

ABSTRACT

Selecting a soundtrack is a critical step in video editing. However, evaluating music is a slow, sequential process. Creators must listen to tracks one by one, making direct comparison difficult and forcing them to rely on auditory memory to predict a track's impact on their video. We present *VidTune*, a system that facilitates exploration and comparison of generative soundtracks with *Visual Thumbnails*. We introduce a technique for generating contextual visual thumbnails that translate a music track's character into a stylized preview of the user's own video. Our method maps analyzed musical attributes like valence and energy to visual parameters such as color and brightness applied to a keyframe. This approach transforms soundtrack selection from a slow process of sequential listening into a rapid act of parallel visual comparison, allowing creators to more intuitively imagine each track's final impact on their video.

1. INTRODUCTION

Music is a powerful device in storytelling that can emphasize emotions and engage the audience [1]. To add effective soundtracks, video creators make strategic musical choices based on their videos – upbeat, popular music to create a pleasant atmosphere for their vlogs [2], ambient music for tutorial videos [3], and memorable hooks to build brand identity in advertisements [4]. While professional video production teams often work with dedicated composers to create music, most individual content creators search for suitable tracks in stock music libraries [5]. However, searching for effective soundtracks can be a challenging process as they have to find copyright-free tracks that match a specific style and emotion of the video.

As high-quality text-to-music models are becoming easily accessible on commercial platforms like Suno [6] and Udio [7], more content creators are using them to generate music for their videos [8]. Text-to-music models can be a powerful alternative that offers flexible customization that traditional stock music lacks. However, there are core challenges that non-expert creators face. First, creators struggle to find the right words to describe the music they want [5,9]. They often discover their preferences only after listening to several diverse alternatives, highlighting a need for broad exploration. Second, even when presented with multiple options, creators face an evaluation hurdle [10–12]. It is difficult and time-consuming to discern the nuanced differences between abstract audio tracks and predict their emotional impact on the video. Finally, refining a promising but imperfect track presents another challenge, as translating a high-level creative goal (*e.g., "make it more emotional"*) into a precise prompt modification is not intuitive.

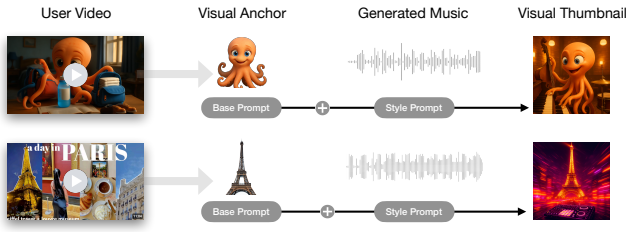

Figure 1. The core concept of our generative thumbnails. The system synthesizes a base prompt, derived from the video's visual anchor (the 'what'), with a style prompt, derived from the music's character (the 'how'). The resulting thumbnail fuses the video's content with the music's style, as seen in the transformation of an octopus character into a jazz musician or the Eiffel Tower into a neon DJ scene.

We present *VidTune* (Figure 2), a system that facilitates exploration and comparison of generative soundtracks. To support creative exploration, VidTune first expands a user's prompt to generate musically diverse alternatives. At the core of our system are *visual thumbnails*, which allow creators to quickly skim and compare different music options. These thumbnails translate a rich set of musical attributes – from high-level genre and mood to specific qualities like valence and energy – into intuitive visual parameters like artistic style, color, and brightness. By grounding these

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

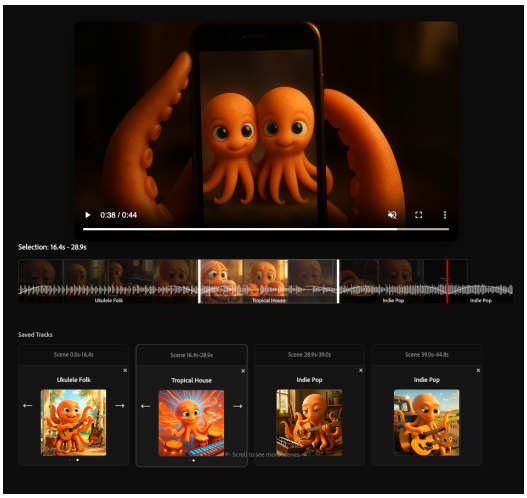

Figure 2. VidTune's interface integrates video player for users to preview music in context to the video. VidTune's visual thumbnails enable users to quickly sensemake different music options.

visuals in an anchor from the user's own video, VidTune helps creators more intuitively imagine each music's impact. Users can also preview music synchronized with their video and review VidTune's fit indicators that explain how a track matches the video's context and the user's goals.

2. PIPELINE FOR GENERATING THUMBNAILS

VidTune's visual thumbnails allow creators to instantly see and compare how different musical options would feel in the context of their actual footage (Figure 1). To generate the thumbnails, our pipeline 1) extracts the core semantic visual anchor from the video (*e.g.,* key subject or scene) and 2) analyzes each track for musical attributes like valence, energy, and tempo. Then, we generate each thumbnail by blending the musical impact directly into the style of the visual anchor.

2.1 Identifying the Visual Anchor

Our pipeline uses a large multimodal model (LMM) [13] to identify the visual anchor from the users' video. We prompt LMM to first identify one or more protagonists, such as stylized protagonists, such as animated figures, animal characters, or generated human avatars. By focusing on non-real characters, we set a clear boundary against modifying real humans' appearance and behavior from the video recording, which can lead to ethical issues [14] and uncanny valley effect [15]. In videos where no such protagonist is present, LMM first identifies the video's central theme or setting. For instance, for a travel vlog, the model would first identify the theme "Paris" then select a keyframe from the segment that best illustrates the theme (*e.g.,* a shot of the Eiffel Tower) as the visual anchor.

Musical Attribute	Visual Mapping Rule
Genre & Style	Background scene and artistic style *e.g.,* electronic → neon cityscape
Instruments	Protagonist performing instrument (size proportional to prominence)
Tempo	Implied motion *e.g.,* fast → speed lines & blur
Emotion & Mood	Facial expressions and body language matching emotion
Valence	Visual filter: hue/tint adjustment *e.g.,* positive → warm, negative → cool
Energy	Visual filter: brightness and saturation *e.g.,* high → bright, low → dim

Table 1. Mapping of musical attributes to AI-generated thumbnails in VidTune.

2.2 Analyzing Music and Generating Thumbnails

Whenever a new music track is generated, we prompt an LMM [13] to produce a detailed description of its musical character, covering the attributes in our mapping framework (Table 1). From this description, we then generate a stylistic modifier – a short phrase that translates the music's qualities into visual effects. For consistent translations, we prompt LMM with both the mapping rules (Table 1) and several in-context examples [16] of rules being applied. Finally, the stylistic modifier is appended to the visual anchor's description to construct a complete prompt for the text-to-image model [17]. For example, the anchor *'a shot of the Eiffel Tower'* is combined with a modifier derived from an upbeat electronic music to become: *'A cinematic shot of the Eiffel Tower at night. The artistic style is futuristic electronic, with the surrounding scene transformed into a vibrant neon cityscape'* (Genre & Style). *The image is bright and highly saturated with a warm color palette of gold and magenta* (Energy & Valence). *There is a dynamic sense of implied motion, with subtle light streaks and motion blur to reflect the fast tempo* (Tempo)'.

3. CONCLUSION

We introduce a technique for generating contextual visual thumbnails that translate a track's analyzed musical character into a stylized preview of the user's own footage. Our method provides a visual proxy for audio, transforming a cognitively demanding task of listening and imagining into a rapid act of parallel visual comparison. The potential of this cross-modal approach extends far beyond generative music. Our analysis pipeline can be applied to any audio track, envisioning a future where existing music libraries are browsed not just by tags, but by dynamic visual previews contextualized to the user's project.

4. ACKNOWLEDGMENTS

We thank Nick Bryan for his valuable feedback, and Gabi Duncombe for generously sharing her video assets.

5. REFERENCES

[1] S. Rubin and M. Agrawala, "Generating emotionally relevant musical scores for audio stories," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 439–448.

[2] G. Husain, W. F. Thompson, and E. G. Schellenberg, "Effects of musical tempo and mode on arousal, mood, and spatial abilities," *Music perception*, vol. 20, no. 2, pp. 151–171, 2002.

[3] A. S. Souza and L. C. L. Barbosa, "Should we turn off the music? music with lyrics interferes with cognitive tasks," *Journal of cognition*, vol. 6, no. 1, p. 24, 2023.

[4] A. SHAKIL and D. A. Siddiqui, "How jingles in advertising affect retention and recall of the product," *Shakil, A. and Siddiqui, DA (2019). How Jingles in Advertising Affect Retention and Recall of the Product. International Journal of Thesis Projects and Dissertations*, vol. 7, no. 2, pp. 20–29, 2019.

[5] E. Frid, C. Gomes, and Z. Jin, "Music creation by example," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[6] Suno, "Suno | ai music." [Online]. Available: https://suno.com/

[7] Udio, "Udio ai music generator." [Online]. Available: https://www.udio.com/

[8] Y. Lyu, H. Zhang, S. Niu, and J. Cai, "A preliminary exploration of youtubers' use of generative-ai in content creation," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–7.

[9] N. Hammad, C. A. Fraser, E. Harpstead, J. Hammer, and M. Dontcheva, ""it's more of a vibe i'm going for": Designing text-to-music generation interfaces for video creators," in *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, 2025, pp. 2738–2754.

[10] Y. Choi, J. Moon, J. Yoo, and J.-H. Hong, "Exploring the potential of music generative ai for music-making by deaf and hard of hearing people," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–20.

[11] M. Huh, Y.-H. Peng, and A. Pavel, "Genassist: Making image generation accessible," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–17.

[12] M. Huh, D. Li, K. Pimmel, H. V. Shin, A. Pavel, and M. Dontcheva, "Videodiff: Human-ai video cocreation with alternatives," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–19.

[13] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[15] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & automation magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[17] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.