OPEN SCREEN SOUNDTRACK LIBRARY VERSION 2

Haven Kim[1] **Leduo Chen**[1] **Bill Wang**[1] **Hao-Wen Dong**[2] **Julian McAuley**[1]

[1] University of California San Diego
[2] University of Michigan

ABSTRACT

Despite growing interest in video-to-music generation systems, their application in film production remains limited, primarily due to the lack of large-scale datasets containing aligned pairs of movie clips and soundtracks. Although prior work has attempted to construct such a dataset [1], this comprises only 36.5 hours of data, which is insufficient for training robust models. In this paper, we present **Open Screen Soundtrack Library Version 2**, a novel dataset comprising pairs of video clips from films and their corresponding soundtracks, curated with a novel methodology that automatically identifies and extracts soundtrack segments from video clips, consisting of 552.70 hours and 76,408 video clips sourced from both public domain movies as well as commercial ones from a publicly available dataset [2].

1. INTRODUCTION

Video-to-music generation systems have gained increasing attention in both the audio and symbolic domains. However, their application in film production remains limited, as most prior work has focused on use cases such as music videos [3–5], advertisements [6], or user-generated content [7]. While some studies have utilized trailer data for video-to-music generation [6], trailer music possesses unique characteristics that are significantly different from main movie soundtracks.

To date, we have identified only two studies that specifically address video-to-music generation for film production—one in the symbolic domain [8] and the other in the audio domain [1]. Despite the success of the former, the symbolic-domain approach requires expert knowledge to convert symbolic outputs into usable soundtracks, making it impractical for film producers. The latter leveraged video information to generate soundtracks directly; however, it relied heavily on textual information due to the limited size of its dataset, which is only about 36.5 hours.

Table 1. Comparison of video-music datasets available as of August 2025. Approximately half of our dataset is self-hosted.

Dataset	Self-Hosted	Video Content	Length (Hours)
HIMV-200K [9]	✗	Music Video, User-Generated Video	-
URMP [11]	✗	Music Performance	33.5
TikTok [12]	✗	Dance Video	1.5
SymMV [3]	✗	Music Video	76.5
MuVi-Sync [13]	✗	Music Video	-
BGM909 [10]	✗	Music Video	-
VidMuse [6]	✗	Music Video, Advertisements, Trailer	18k
OSSL [1]	✓	Films	36.5
OSSL-V2 (ours)	✓ and ✗ (partial)	Films	522.7

To overcome these limitations, we constructed a large-scale dataset consisting of aligned movie clips and corresponding soundtracks, totaling 552.70 hours and 76,408 video clips. Using a novel methodology, we accurately identified and extracted soundtrack segments from video content. We believe that this dataset will enable training video-to-music generation models tailored specifically for film production applications.

2. MUSIC-VIDEO DATASETS

In this section, we review datasets containing aligned pairs of music and video clips in the audio domain. These datasets span various types of video content, including music videos [3, 5, 6, 9, 10], musical performance recordings [11], and user-generated content [9]. The only dataset that contains film clips and their corresponding soundtracks is the Open Screen Soundtrack Library, which comprises 36.5 hours of music–movie clip pairs sourced from public domain films [1], making it self-hosted (i.e., users do not need to undergo a separate procedure to download the dataset, such as web scraping YouTube). In contrast, our dataset, Open Screen Soundtrack Library Version 2, is partially self-hosted, as it draws from both public domain and commercial films. However, our dataset is significantly larger in scale. A comprehensive comparison of video-music datasets is provided in Table 1.

3. DATASET CONSTRUCTION

Our music-movie clip dataset is constructed from two types of movie data. The former comprises 1,886 public

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

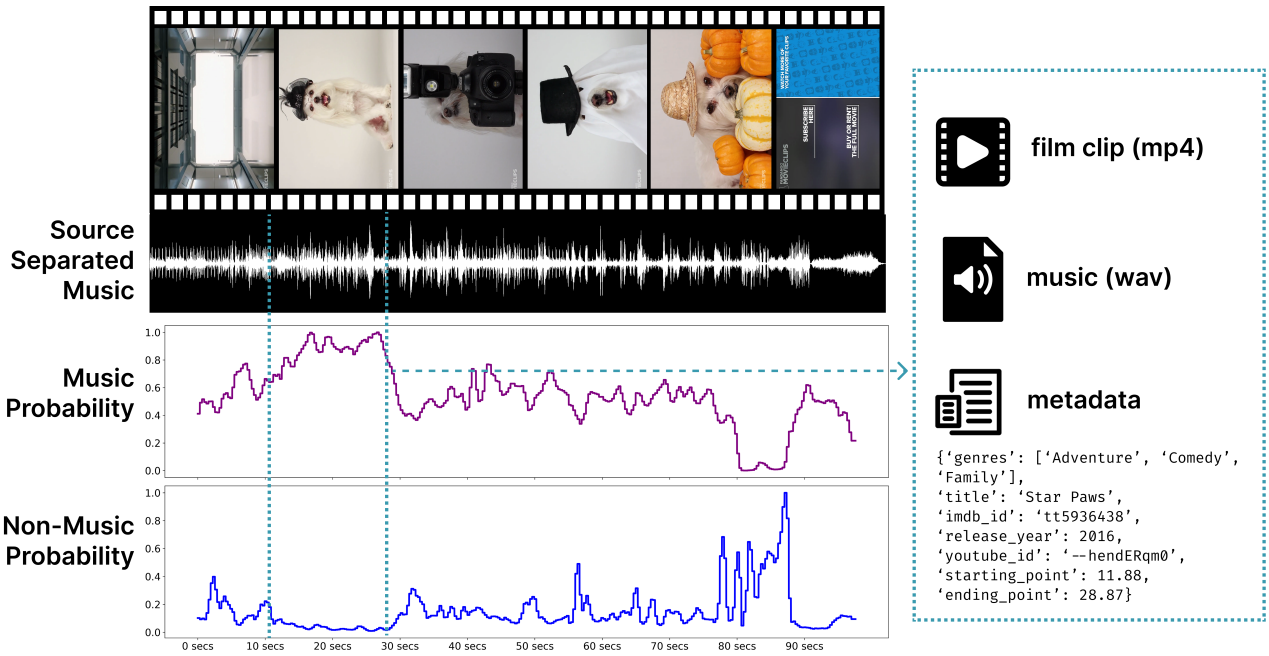

Figure 1. Illustration of our methodology for constructing the Open Screen Soundtrack Library Version 2. We apply an open-source event detection model to source-separated music to ensure that the music in our dataset does not contain any non-musical components, and extract segments where the music probability exceeds the non-music probability and the non-music one is lower than 0.05 for at least 10 seconds.

domain films downloaded from YouTube [1], and the latter is derived from a publicly available movie dataset, the Condensed Movies Dataset [2]. Our dataset construction process consists of two main components: source separation and event detection, as illustrated in Figure 1 .

In the first step, we applied an open-source separation model [14] in order to extract music from each movie clip's audio track. This model offers a high-quality processing option that requires three times longer than the default option. We selected the high-quality option for audio source separation because our objective is to create a music-movie clip dataset with the highest possible quality.

In the second step, we employed an event detection model to estimate the probability distribution of event types in source-separated musical tracks. This step was essential because the source separation model, even when using a high-quality option, was not perfect; source-separated music often contained non-musical events. To address this, we used an open-source automatic event detection model [15], from which we identified 157 out of 527 categories as musical events (e.g., "trance music"). We defined the music probability as the sum of probabilities for the 157 musical events, and the non-music probability as the sum of probabilities for the 370 non-musical events. We extracted segments where the music probability exceeded the non-music probability for at least 10 consecutive seconds. However, this failed to filter out cases where both musical and non-musical events were prominent (e.g., music probability of 0.8 and non-music probability of 0.7). Therefore, we applied an additional filter to exclude cases where the non-music probability exceeded 0.05.

[1] This part is self-hosted, meaning that readers do not need to undergo a separate download process such as web scraping.

Table 2. Statistics of the Open Screen Soundtrack Library Version 2. To obtain commercial movie clips, we used a list of YouTube IDs from the Condensed Movies Dataset [2] and scraped the corresponding clips from the web.

	Public Domain	**Commercial [2]**	**Total**
Number of Clips	35,705	40,703	76,408
Number of Unique Films	1,886	2,633	4,519
Average Length (seconds)	28.77	23.65	26.04
Total Length(hours)	285.31	267.39	552.70

This approach yielded a total of 76,408 video clips with source-separated soundtracks (processed using the high-quality option) averaging 26.04 seconds in length, along with rich metadata such as genres, release year, and title. Detailed dataset statistics are presented in Table 2.

4. CONCLUSION

In this paper, we introduced the Open Screen Soundtrack Library Version 2, a large-scale dataset of paired movie clips and their corresponding soundtracks, constructed using a novel methodology that automatically identifies and extracts soundtrack segments from video clips. We believe this dataset will facilitate the training of video-to-music generation systems with applications in film production. Although our focus was specifically on film clips, we also want to emphasize the broad generalizability of our methodology, which is also applicable to other types of video content, such as vlogs, highlighting its potential for constructing diverse music–video datasets.

5. REFERENCES

[1] H. Kim, Z. Novack, W. Xu, J. McAuley, and H.-W. Dong, "Video-guided text-to-music generation using public domain movie collections," *arXiv preprint arXiv:2506.12573*, 2025. [Online]. Available: https://arxiv.org/abs/2506.12573

[2] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," *arXiv preprint arXiv:2005.04208*, 2020. [Online]. Available: https://arxiv.org/abs/2005.04208

[3] L. Zhuo, Z. Wang, B. Wang, Y. Liao, C. Bao, S. Peng, S. Han, A. Zhang, F. Fang, and S. Liu, "Video background music generation: Dataset, method and evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 637–15 647.

[4] K. Su, J. Y. Li, Q. Huang, D. Kuzmin, J. Lee, C. Donahue, F. Sha, A. Jansen, Y. Wang, M. Verzetti *et al.*, "V2meow: Meowing to the visual beat via video-to-music generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4952–4960.

[5] R. Li, S. Zheng, X. Cheng, Z. Zhang, S. Ji, and Z. Zhao, "Muvi: Video-to-music generation with semantic alignment and rhythmic synchronization," *arXiv preprint arXiv:2410.12957*, 2024.

[6] Z. Tian, Z. Liu, R. Yuan, J. Pan, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, "Vidmuse: A simple video-to-music generation framework with long-short-term modeling," *arXiv preprint arXiv:2406.04321*, 2024.

[7] H. Zuo, W. You, J. Wu, S. Ren, P. Chen, M. Zhou, Y. Lu, and L. Sun, "Gvmgen: A general video-to-music generation model with hierarchical attentions," *arXiv preprint arXiv:2501.09972*, 2025.

[8] Z. Xie, Q. He, Y. Zhu, Q. He, and M. Li, "Filmcomposer: Llm-driven music production for silent film clips," 2025. [Online]. Available: https://arxiv.org/abs/2503.08147

[9] S. Hong, W. Im, and H. S. Yang, "Content-based video-music retrieval using soft intra-modal structure constraint," 2017. [Online]. Available: https://arxiv.org/abs/1704.06761

[10] S. Li, Y. Qin, M. Zheng, X. Jin, and Y. Liu, "Diff-bgm: A diffusion model for video background music generation," 2024.

[11] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, p. 522–535, Feb. 2019. [Online]. Available: http://dx.doi.org/10.1109/TMM.2018.2856090

[12] Y. Zhu, K. Olszewski, Y. Wu, P. Achlioptas, M. Chai, Y. Yan, and S. Tulyakov, "Quantized gan for complex music generation from dance videos," 2022. [Online]. Available: https://arxiv.org/abs/2204.00604

[13] J. Kang, S. Poria, and D. Herremans, "Video2music: Suitable music generation from videos using an affective multimodal transformer model," *Expert Systems with Applications*, vol. 249, p. 123640, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2024.123640

[14] R. Solovyev, A. Stempkovskiy, and T. Habruseva, "Benchmarks and leaderboards for sound demixing tasks," 2023.

[15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.