Bring Music The Horizon: Music-Driven 360° Video Generation

Kai Hsu, Tsai[1], Yong Wei, Fu[1], Hung I, Yang[1], and Yu-Chih Chen[1]

Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
{chris.cs12, willyfu0905.cs12, holdtensec.cs12, berriechen}@nycu.edu.tw

Abstract. Music visualization offers a powerful way to enhance listeners' understanding and experience of music by translating auditory signals into visual forms. However, most existing approaches either rely heavily on lyrics or generate flat, non-immersive videos similar to conventional music videos, which limits their ability to convey the emotional dynamics of music and provide an immersive listening experience.
We propose *Bring Music The Horizon*, an emotion-aware pipeline for music-driven 360° video generation. Given an input song, our work first estimates its emotional trajectory by predicting valence-arousal values [1] at the level of every four bars. These values are then converted into emotion-aware visual guidance using EmotiCrafter [2], and these guidance vectors can be manipulated by the SEGA framework [3], which provides fine-grained semantic control for keyframe generation. Finally, image-to-video models are applied to the generated keyframes to synthesize temporally continuous 360° videos for immersive music visualization. Our pipeline generates 360° music visualization videos that reflect the emotional progression and temporal structure of the input song. We demonstrate its capability using songs from different genres and provide qualitative comparisons with *From-Sound-To-Sight* [4], a representative audio-to-visual generation baseline, on our project page at `https://etoile-et-toi-mp3.github.io/BMTH_Project_Page/`.

Keywords: Music Visualization · 360° Video Generation · Music Emotion Recognition · Multimodal Generation

1 Introduction

1.1 Motivation

Music visualization aims to enhance how listeners perceive and experience music by translating auditory signals into visual content. However, existing approaches are often limited to fixed-view or flat videos, resembling conventional music videos rather than immersive visual experiences. Many methods also rely heavily on lyrics to infer visual semantics, which limits their applicability to instrumental music and may overlook the continuous emotional dynamics embedded in the audio itself.

To address these limitations, we propose *Bring Music The Horizon*, a music-driven pipeline for generating immersive 360° videos. Our method focuses on modeling the temporal evolution of musical emotion and synthesizing 360° visual scenes that align with both the affective and structural progression of the song.

1.2 Related Work

Recent music visualization methods combine music information retrieval (MIR) and generative models to synthesize visual content from audio. However, many approaches rely on lyrics or other surface-level semantic cues, which may weaken the connection between the generated visuals and the intrinsic musical properties such as rhythm, intensity, timbre, and emotional progression. Moreover, most prior work targets narrative-style flat music videos rather than immersive 360° visualization.

Jeong et al. [5] proposed an audio-reactive video generation model that maps audio signals to semantic and temporal visual dynamics, such as generating rain intensity according to audio magnitude. Our work shares the broader goal of audio-reactive visual generation, but focuses specifically on music-driven immersive visualization, where both emotional trajectories and musical structure are used to guide 360° video generation.

1.3 Our Pipeline and Limitations

Our pipeline consists of three main modules. The *(i) MIR module* extracts musical features such as tempo, downbeat timestamps, and valence-arousal (V-A) values [1] for each four-bar segment. The *(ii) Emotion-guided 360° Keyframe Generation module* incorporates the estimated emotional features into the visual generation process to produce affect-aligned 360° keyframes. The *(iii) 360° Video Generation module* then applies image-to-video (I2V) models to generate dynamic clips and transition clips, which are concatenated into the final immersive music visualization video.

While the proposed pipeline can generate emotion-aware and temporally structured 360° music visualization videos, several limitations remain. Similar V-A values may produce repetitive keyframes, and visible seams can appear due to boundary inconsistencies in generated 360° videos. The current output resolution is still limited, and the forward-moving visual effect cannot yet be fully controlled, which may affect motion comfort. In addition, overly long prompts may degrade 360° generation and cause the model to produce flat rather than 360°, immersive visuals.

2 Method

Given an input song S and a user-specified base prompt p_{base}, our pipeline generates an emotionally and temporally aligned 360° music visualization. As shown in Fig. 1, the song is first analyzed to extract its bar-level temporal structure and a sequence of valence-arousal (V-A) values that describe the emotional trajectory of the music. These musical cues are then used to adapt the base prompt into emotion-aware panoramic keyframes, allowing the visual content to preserve the user-provided concept while evolving with the song's emotional progression. Finally, the generated keyframes are animated and connected in musical order to produce a complete 360° video whose scene dynamics and transitions follow the structure of the input song.

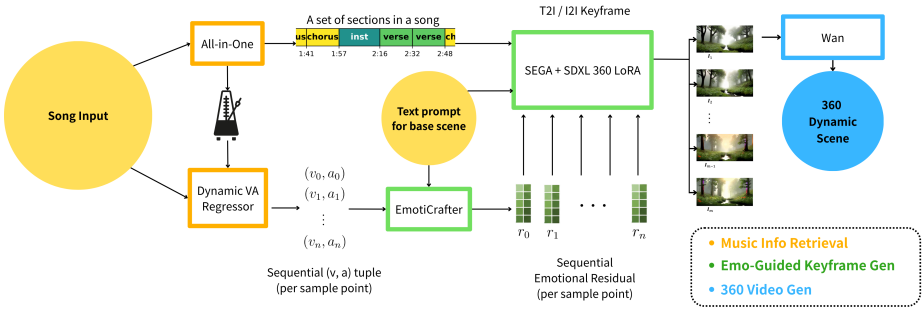

Fig. 1: Overview of the proposed music-driven 360° visualization pipeline. Given an input song and a user-specified base prompt, the MIR module extracts musical structure and downbeat-level timing cues, while the Dynamic V-A Regressor estimates a sequence of valence-arousal conditions. These emotion conditions are converted into emotional residuals by EmotiCrafter [2] and used with SEGA [3] and SDXL [6] 360° LoRA [7] to generate emotion-aware panoramic keyframes. The keyframes are then animated by Wan-based image-to-video generation to produce the final 360° dynamic visualization.

Formally, the overall process is defined as

$$(S, p_{\text{base}}) \rightarrow (B, U, E) \rightarrow K \rightarrow V_{360} \tag{1}$$

where $B = \{b_i\}_{i=1}^{N}$ denotes the extracted downbeat timestamps, $U = \{u_t\}_{t=1}^{T}$ denotes the resulting four-bar music units, and $E = \{e_t\}_{t=1}^{T}$ denotes the corresponding emotional trajectory. Each emotion condition is represented as $e_t = (v_t, a_t)$ in the V-A space. The keyframe generation stage produces a sequence of panoramic keyframes $K = \{k_t\}_{t=1}^{T}$, and the video generation stage synthesizes the final 360° music visualization V_{360}.

2.1 Music Information Retrieval (MIR)

The MIR module analyzes the input song S and extracts its temporal structure and emotional trajectory:

$$f_{\text{MIR}}(S) = (B, U, E). \tag{2}$$

We first use All-In-One [8] to estimate downbeat timestamps B and functional segment information, which provide the temporal structure of the song. Based on the extracted downbeats, we group the song into four-bar units U. For each unit u_t, our proposed Dynamic Valence-Arousal regressor predicts an emotion condition $e_t = (v_t, a_t)$, where v_t and a_t denote the valence and arousal values, respectively. This regressor is trained to estimate segment-level emotional dynamics directly from music audio, rather than relying on lyrics or manually assigned emotion labels. The resulting V-A sequence E is used as the emotional condition for subsequent visual generation, enabling the generated visuals to follow both the musical structure and the emotional progression of the song.

2.2 Emotion-guided 360° Keyframe Generation

The keyframe generation module maps each four-bar music unit to an emotion-aware panoramic keyframe:

$$f_{\text{key}}(p_{\text{base}}, e_t) = k_t. \tag{3}$$

For each unit u_t, the base prompt p_{base} and its corresponding emotion condition e_t are fed into a retrained EmotiCrafter model [2] to obtain an emotional residual r_t. This residual describes how the visual semantics of the base prompt should be shifted toward the target emotion. We then use r_t as a SEGA guidance condition g_t [3], enabling fine-grained semantic control during the diffusion process. Finally, SDXL [6] with a 360° LoRA adapter [7] is used to generate the panoramic keyframe k_t. This design allows each keyframe to maintain the user-specified scene concept while reflecting the emotion predicted from the corresponding music segment. Repeating this process for all T units yields the keyframe sequence $K = \{k_t\}_{t=1}^{T}$.

2.3 360° Video Generation

The video generation module transforms the keyframe sequence K into the final immersive visualization:

$$f_{\text{video}}(K) = V_{360}. \tag{4}$$

Each keyframe k_t anchors one four-bar video unit. The first three bars are synthesized as a dynamic scene d_t using Wan-I2V [9]. The final bar is generated as a transition clip ρ_t using Wan-flf2v [9], which connects the current dynamic scene to the next keyframe k_{t+1}. By concatenating all dynamic scenes $\{d_t\}_{t=1}^{T}$ and transition clips $\{\rho_t\}_{t=1}^{T}$ in temporal order, the module produces a complete 360° music visualization video V_{360}, in which scene changes are aligned with the temporal structure and emotional flow of the input song.

3 Results and Conclusion

The experimental results show that the proposed pipeline can generate 360° music visualization videos that are directly viewable in VR headsets. The generated scenes follow the temporal structure of the input songs, while the visual atmosphere and scene transitions reflect the estimated emotional trajectory at different musical segments. These results demonstrate the potential of extending conventional music visualization from flat videos to immersive VR experiences.

Overall, this work presents an emotion-aware and structure-aligned pipeline for music-driven 360° video generation. By integrating dynamic music emotion prediction, emotion-guided keyframe generation, and 360° image-to-video synthesis, the proposed method provides a practical framework for immersive music visualization. In addition, retraining EmotiCrafter with a custom dataset helps reduce undesired human-activity artifacts from the original model, resulting in cleaner and more scene-focused visual outputs. We believe this work establishes a promising foundation for future research on cross-modal and immersive content creation.

Acknowledgements. This work was supported by the MOE Yushan Young Scholar Program under Grant MOE-114-YSFEE-0010-008-P1 and NSTC Taiwan under Grant NSTC 115-2813-C-A49-161-E.

References

1. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology **39**, 1161–1178 (12 1980). https://doi.org/10.1037/h0077714
2. Dang, S., He, Y., Ling, L., Qian, Z., Zhao, N., Cao, N.: Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15218–15228 (2025)
3. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing text-to-image models using semantic guidance (2023), https://arxiv.org/abs/2301.12247
4. Vitasovic, L., Graßhof, S., Kloft, A.M., Lehtola, V.V., Cunneen, M., Starostka, J., McGarry, G., Li, K., Brandt, S.S.: From sound to sight: Towards ai-authored music videos (2025), https://arxiv.org/abs/2509.00029
5. Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J.: The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7822–7832 (2023)
6. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023), https://arxiv.org/abs/2307.01952
7. artificialguybr: 360Redmond: 360 degree panorama LoRA for SDXL. https://huggingface.co/artificialguybr/360Redmond (2024)
8. Kim, T., Nam, J.: All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2023)
9. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314