Bring Music The Horizon: Music-Driven VR Video Generation

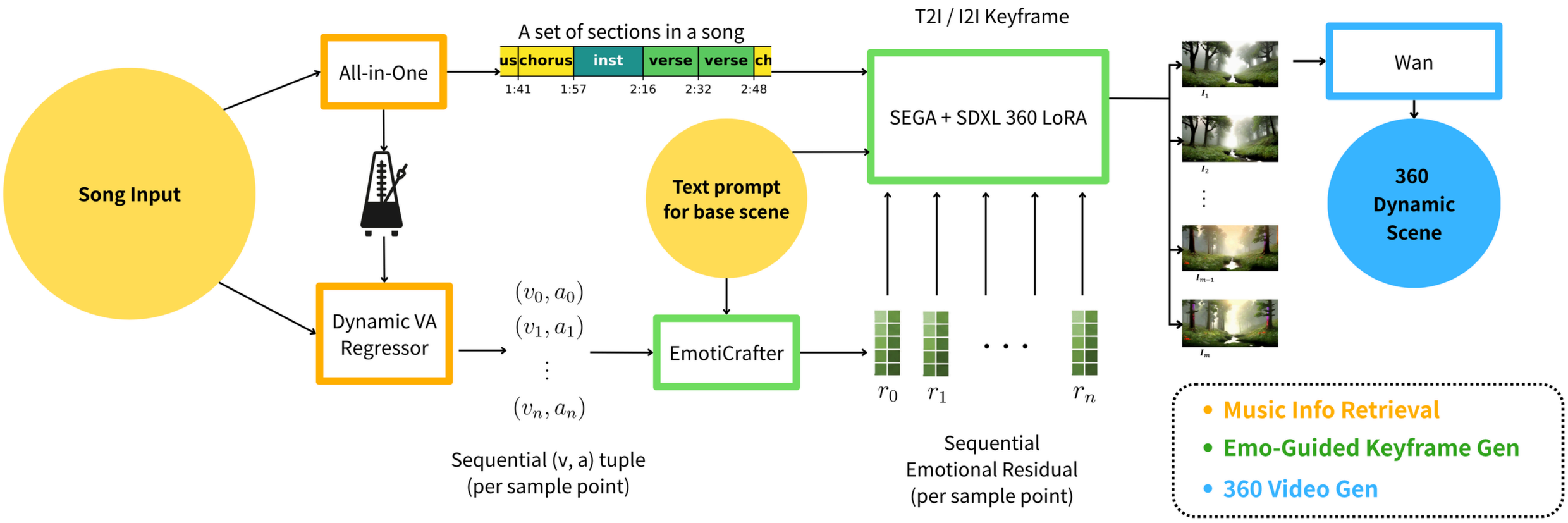
指導教授：陳昱芝　　學生：蔡凱旭、傅永威、楊弘奕

Abstract

This research advances AI music visualization from single-perspective media to musically-aware VR experience by overcoming bottlenecks in **continuous emotion modeling** and **immersive scene presentation**. By integrating Music Information Retreival (MIR), semantic embedding guidance techniques, and Diffusion models, we manage to synthesize a 360° immersive video that can respond to the emotional dynamics and musical structure of the given song. This project establishes a solid foundation for immersive content generation and future transmedia creation.

Method

Our pipeline consists of a Music Information Retrieval module, an Emotion-guided Keyframe Generation module, and a 360° video generation module. First, **All-in-one[1]** will extract the timestamp of each downbeat. Then, our **VA[2] Regressor** will extract the implied emotion, also known as the **VA value[2]**, at each downbeat. Next, **EmotiCrafter[3]** will generate the "emotional residual" based on the VA value, which is then used with **SEGA[4] guidance term** to guide the diffusion process of the keyframes. Lastly, we will use **Wan-I2V[5]** to generate 3 bars of dynamic scene for each keyframe, and then **Wan-flf2v[5]** to generate the transition video between each dynamic scene. By concatenating the dynamic scenes and their transitions, we can obtain the immersive visualization of the song in the form of 360° video.



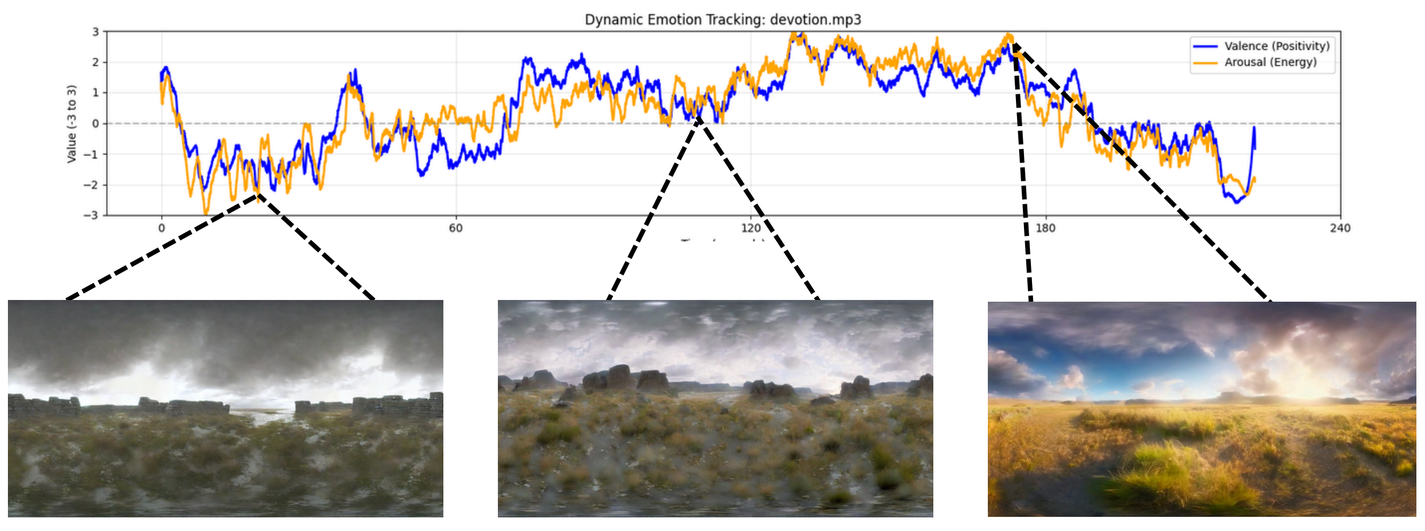
Result

This pipeline generates 360° video that can be played in a VR headset to get immersive music visualization experience. Each keyframe aligns with the emotion of their corresponding timepoints in the song.

Input

Prompt: “a beautiful prairie” / Song: Devotion [6]

VA Value



Keyframes

Key Contributions

- Immersive experience in Music Visualization.
- Novel method for dynamic VA prediction for music.
- Eliminated unwanted human-activities in **EmotiCrafter[3]** via retraining with self-generated dataset.

Demo

Reference

[1] T. Kim and J. Nam, “All-In-One Metrical And Functional Structure Analysis With Neighborhood Attentions on Demixed Audio,” IEEE WASPAA, 2023.
[2] James A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, 39(6):1161–1178, 1980.
[3] S. Dang et al., “Emoticrafter: Text-to-emotional-image generation based on valence-arousal model,” ICCV, 2025. [4] M. Brack et al., “SEGA: Instructing diffusion using semantic dimensions,” NeurIPS, 2023.
[5] Team Wan et al., “Wan: Open and Advanced Large-Scale Video Generative Models,” arXiv:2503.20314, 2025. [6] 草東沒有派對. 「還願」. 《還願（遊戲原聲帶）》, 赤燭股份有限公司, 2019 年。YouTube.