From Sound to Sight: Towards AI-authored Music Videos

Leo Vitasovic[1,2] Stella Graßhof[1,2] Agnes Mercedes Kloft[3] Ville V. Lehtola[4] Martin Cunneen[5] Glenn McGarry[6] Justyna Starostka[1] Kun Li[4] Sami S. Brandt[1]

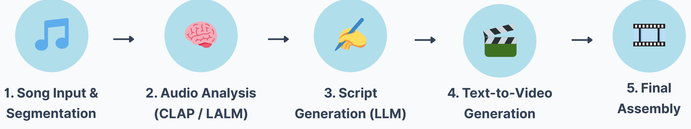
XTREME
Funded by the European Union
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Abstract

Conventional music visualization systems often depend on hand-crafted, abstract transformations with limited expressive power. This research introduces **two novel pipelines** to automatically generate complete **music videos** from any user-specified song, whether instrumental or vocal. By leveraging off-the-shelf deep learning models, the system analyzes audio for musical qualities like **emotion** and **instrumental patterns**, distils them into textual scene descriptions using a language model, and then generates corresponding video clips. A preliminary user study suggests the generated videos show potential for **storytelling**, **visual coherence**, and **emotional alignment** with the music, highlighting the power of deep generative models to expand the art of music visualization.

The Generation Pipeline

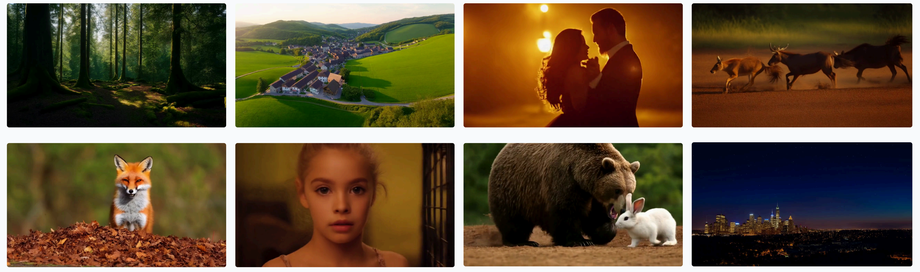
Both proposed pipelines (CLAP and LALM-based) follow a similar multi-step process to generate a music video from any audio, ensuring each stage is **human-readable** and **interpretable**.



Challenges

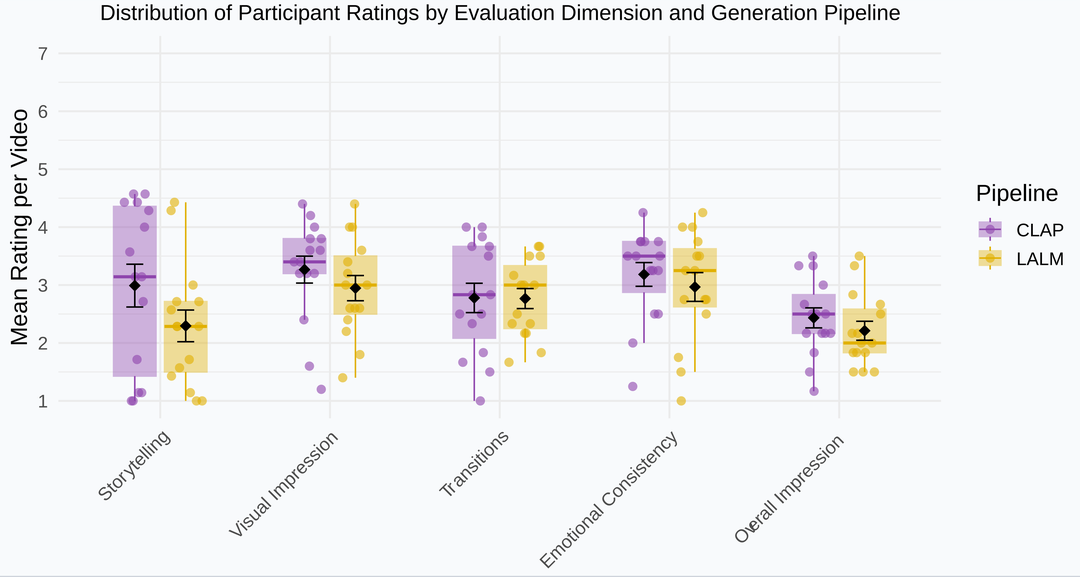
- **Expressiveness**: Older systems rely on shape/color transformations.
- **Lyric Dependency**: Existing solutions require textual input and guidance.
- **Lack of Cohesion**: Visuals often lack narrative structure and consistency.
- **Quality Assesment**: There are no metrics for music visualization.

Experiments - Selected Frames from Generated Videos



Example music video

Results - Human Evaluation



Conclusion & Future Work

The proposed pipelines demonstrate significant potential for creating conceptually coherent music videos by integrating audio analysis with LLM-driven text-to-video generation. Next steps will focus on improving **visual consistency**, incorporating **lyrics**, expanding **user studies** and exploring an end-to-end, **music-to-video** based approach without relying on text as intermediary.

Acknowledgment

This work is funded by the European Union within the Horizon Europe research and innovation program under grant agreement No. 101136006 – XTREME project, coordinated by S.S. Brandt/IT University of Copenhagen, Denmark. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, European Union can not be held responsible for them. This project was partially funded by the Pioneer Centre for AI, DNRF grant number P1.

Affiliations

[1]IT University of Copenhagen [2]Pioneer Centre for Artificial Intelligence [3]Aalto University [4]University of Twente [5]University of Limerick [6]University of Nottingham

Paper

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

IT UNIVERSITY OF COPENHAGEN