Video-to-Music Generation

Scott Merrill[1], smerrill@unc.edu; Titus Spielvogel[1], tis@cs.unc.edu

1.Department of Computer Science, UNC-Chapel Hill

Poster Session for COMP 790 – Vision Transformer

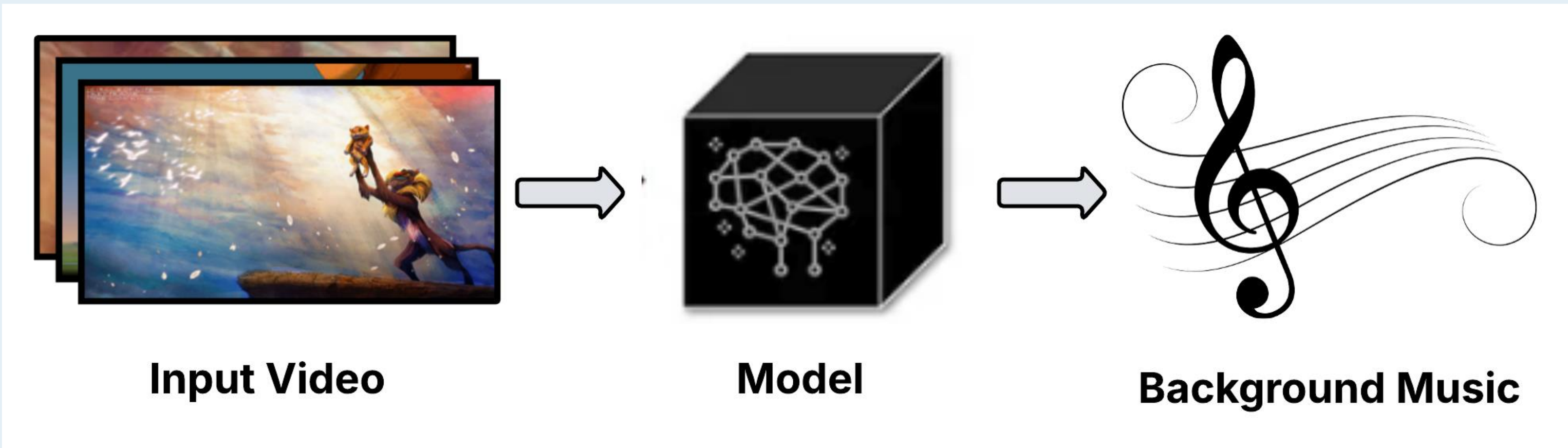
UNC DEPARTMENT OF COMPUTER SCIENCE

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

INTRODUCTION

Video-to-Music Generation

- Given an input video, automatically **generate background music** that **aligns with video content**



Why Music Matters in Video

- Music **enriches video content** by enhancing emotion, rhythm, and immersion.
- Well-synced music and visuals **drive engagement and virality.**

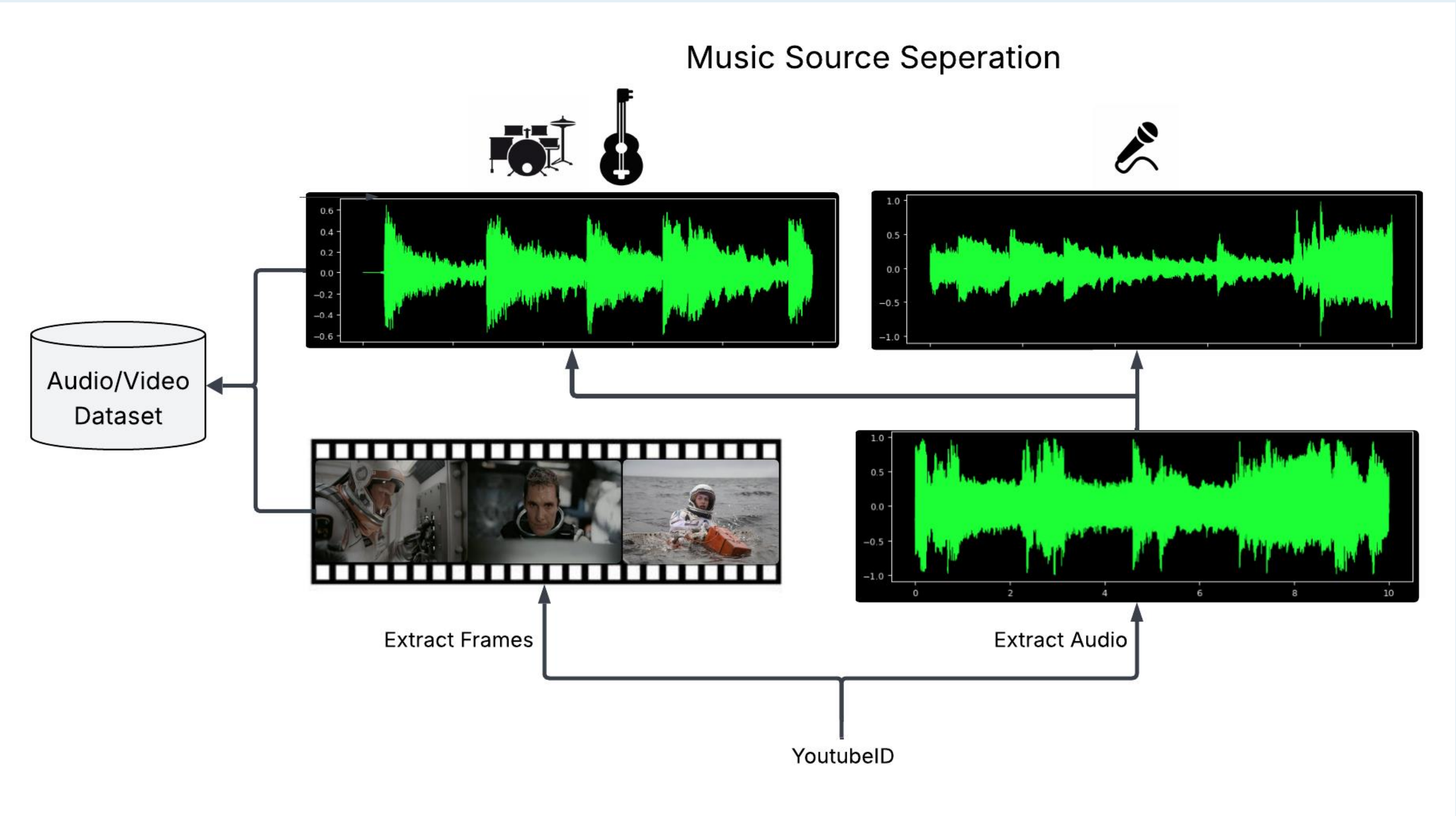
What Makes this Challenging:

- Temporal synchronization (e.g., beats matching motion)
- Musical Coherence (e.g. structure and emotional flow)
- These two goals may be **at odds with each other**

METHOD

Data Collection

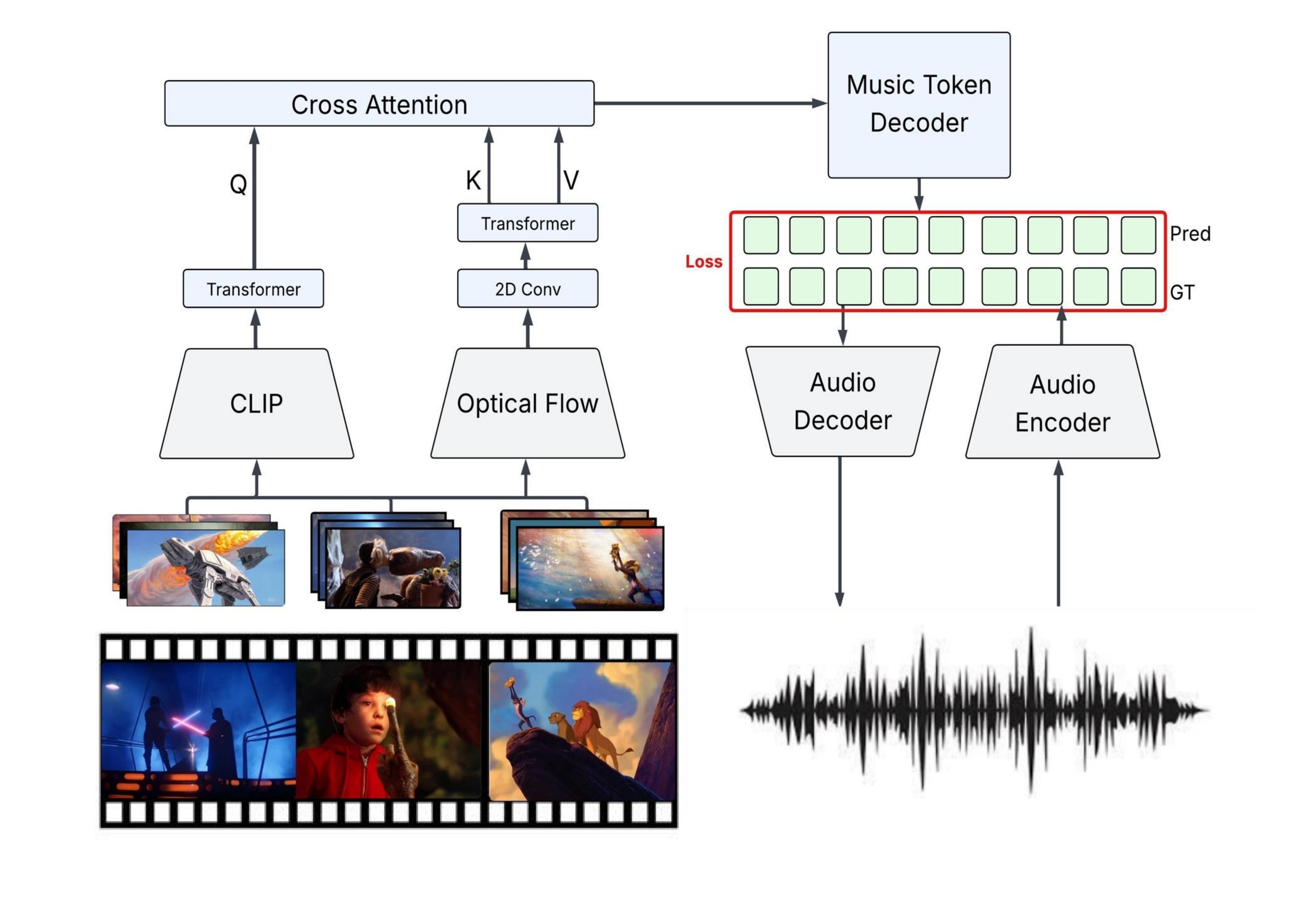
- We used a subset of the ~20,000 video-music pairs from V2M20K Dataset (13293 Training, 3798 Evaluation, 1899 Validation)
- This dataset was **carefully curated** to **have high audio-video alignment**
- Each sample includes the first **30 seconds of high-quality, stylistically diverse video–music content** (e.g., trailers, ads, documentaries).
- Sourced from YouTube using yt-dlp and took several strategies to **avoid throttle limits and evade detection**
- Video frames were extracted and **audio source separation** techniques were used to **remove vocals and retain instrumental tracks**



Model Architecture

Hypothesis: To aligned background music we need to capture both **long-range visual cues** with **fine-grained motion features .**

- **CLIP Embeddings :** Capture **longer-range visual** semantics
- **Optical Flow Embeddings:** Capture **fine-grained motion** dynamics
- **Cross-Attention Fusion: Combines** CLIP and Optical Flow embeddings and **projects** them to *token decoder's* Embedding Space
- **Msuic Token Decoder:** Predicts next audio tokens conditioned on fused video context



RESULTS

Baseline:

- We benchmark our model against *VidMuse*
- Like our approach, *VidMuse* **conditions MusicGen** using video-based embeddings
- It uses a **rolling attention window** over short- and long-term video features to model temporal context
- Effective, but **computationally expensive** and **limited** by a **fixed context window**

Metrics

- **Kullback-Leibler Divergence (KLD):** Measures divergence between the statistical distributions of generated and real audio.
- **Fréchet Audio Distance (FAD):** Measures how close generated audio is to real audio in a learned feature space (VGGish embeddings).
- **Chroma Cosine Similarity:** Compares the harmonic content of two audio clips using 12-dimensional chroma vectors (one per pitch class).

Model	Params	Training Samples	FAD ↓	KLD(P‖Q)↓	KLD(Q‖P)↓	Chroma ↑
VidMuse-M	1.9B	360k	**2.13**	**1.32**	**0.98**	0.056
Our Model-S w/o motion						
Our Model-S	487M	20k	2.87	1.48	1.08	**0.058**

Table 1. Comparison of our model against VidMuse-M across various evaluation metrics. While VidMuse-M benefits from a significantly larger parameter count and training dataset, our model demonstrates competitive performance—particularly in Chroma similarity, which reflects musical coherence and alignment..

- **Key Insight**: Our model has **5x fewer parameters** and is trained on **8x less data** yet **remains competitive** with *VidMuse*.
- It's able to outperform the baseline on **chroma similarity**, the **most musically meaningful metric**

Design Advantages:

- **Simplified Temporal Attention**: Cross-attention along only the temporal axis simplifies learning and may enhance alignment.
- **Rich Feature Input**: Use of CLIP embeddings + optical flow provides **better motion and semantic context** than *VidMuse*'s rolling window strategy.
- **High-Quality Dataset**: V2M dataset's **strong audio-visual alignment reduces noise**, enabling high performance even with fewer training samples.

CONCLUSION & FUTURE WORK

Conclusion

- Both **fine-grained** motion and **long-range context** is necessary for video-to-music generation
- **Excessively large** datasets may be **unnecessary if data quality is high**
- Given our performance against V2M, **spatial fusion** of embeddings may be **unnecessary**

1 Expand Evaluation

- Include **addition quantitative metrics** such as AV-Align
- Conduct Human Centered Evaluations
- Source additional benchmark models to compare to

2 Text-Based Conditioning

- **Fuse** video and motion embeddings with **text embeddings**
- Enable users to **condition** output on **text** as well as video