Video-Guided Text-to-Music Generation Using Public Domain Movie Collections

Haven Kim [1], Zachary Novack [1], Weihan Xu [2], Julian McAuley [1], Hao-Wen Dong [3]
[1] University of California San Diego [2] Duke University [3] University of Michigan

Scan Me!

TL;DR

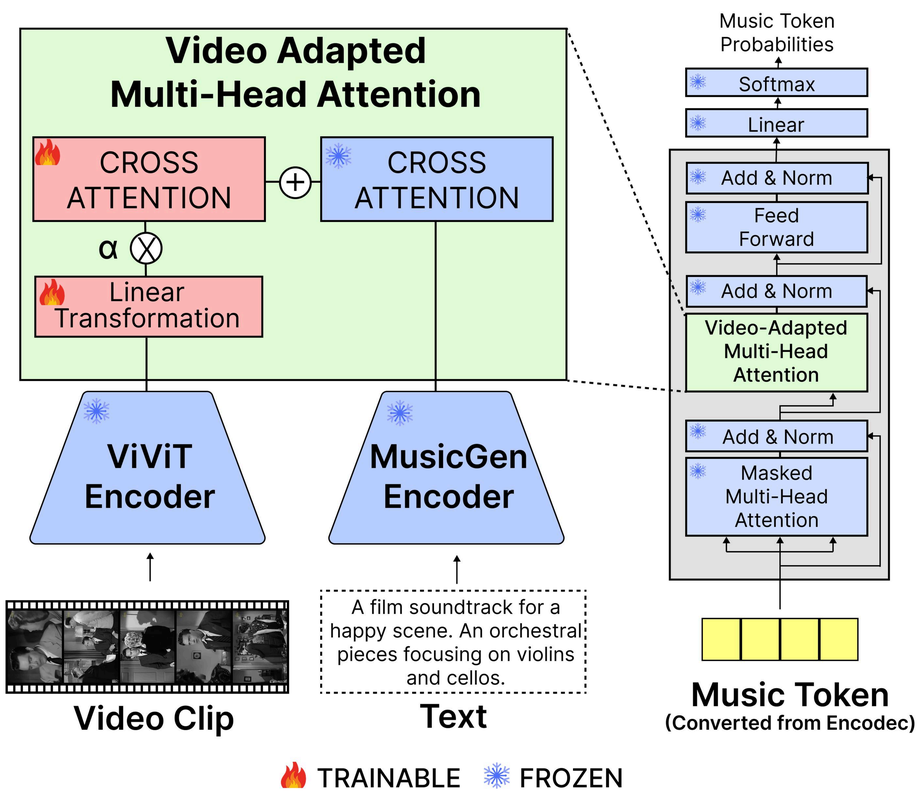
We made a music-video dataset from public domain or CC-licensed films (**OSSL**) and built a music generation model upon it. It is easy to download! Scan the QR code for details!

Open Screen Soundtrack Library (OSSL)

- A dataset consisting of movie clips from public domain films paired with high-quality separately downloaded soundtracks
- Human-annotated mood information.
- **Self-hosted** (i.e., without requiring separate download)
- Along with two separate evaluation datasets!
  - OES-Pub: from public domain films
  - OES-Com: from commercial films

Attributes	OSSL	OES-Pub	OES-Com
Number of samples	736	100	100
Number of unique films	299	76	37
Average clip duration	178.47 sec	30 sec	30 sec
Total duration	36.49 hours	0.83 hours	0.83 hours

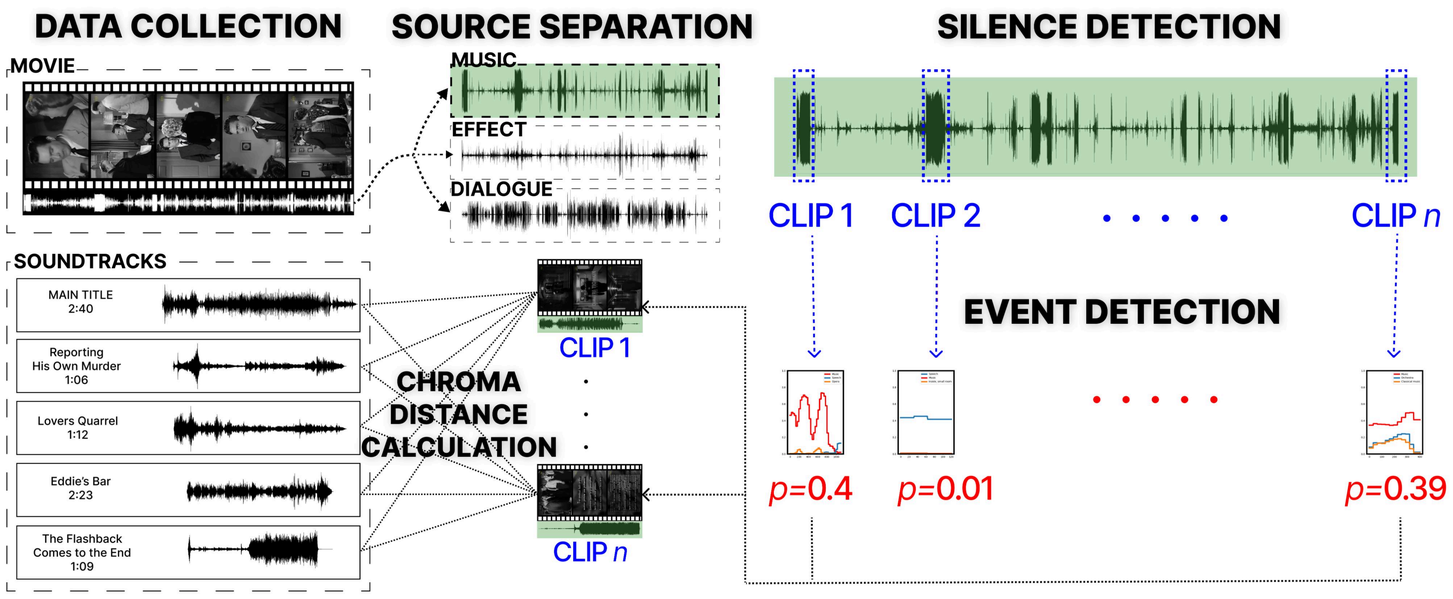
Extending MusicGen with a Video Adapter



Our video adapter applies cross-attention to linearly transformed video embeddings, scaled by a trainable parameter α, and integrates it into the MusicGen's original cross-attention mechanism.

$$\text{head}_i = \text{Attention}(x\mathbf{W_i^{(q)}}, z_t\mathbf{W_i^{(k)}}, z_t\mathbf{W_i^{(v)}})$$
$$+\alpha \times \text{Attention}(x\tilde{\mathbf{W}}_\mathbf{i}^{(\mathbf{q})}, \tilde{z_v}\tilde{\mathbf{W}}_\mathbf{i}^{(\mathbf{k})}, \tilde{z_v}\tilde{\mathbf{W}}_\mathbf{i}^{(\mathbf{v})})$$

Dataset Construction Methodology



- Download public domain movies and soundtracks separately
- Source-separate movie audio into music and other elements
- Verify the presence of musical content using an event detection algorithm
- Match movie clips with their corresponding soundtracks with chroma similarity
- Human inspection
- Mood annotation by Humans based on Russell's 4Q.

Evaluation Results

- Acronyms
  - `S`, `M`: Small, Medium
  - `Base`: MusicGen
  - `Text`: Fine-tuned with text
  - `Multi`: Fine-tuned with text and video

	OSSL Fine-tuned	Video Adapter Integrated	Objective										Subjective		
			Distributional Fidelity				Paired Fidelity				Diversity		Human Ratings		
			FAD ↓		Precision ↑		Similarity ↑		KL ↓		Recall ↑		Mood ↑	Genre ↑	Quality ↑
			pub	com	pub	com	pub	com	pub	com	pub	com	avg ± CI	avg ± CI	avg ± CI
S-Base	✗	✗	64.91	77.99	22.00	14.00	41.55	34.77	1.20	1.93	4.78	6.20	4.53 ± 0.91	5.27 ± 1.04	5.67 ± 0.99
S-Text	✓	✗	61.98	75.75	20.00	14.00	42.44	34.81	1.13	1.97	9.90	**19.00**	5.77 ± 1.73	6.00 ± 1.11	6.20 ± 0.76
S-Multi	✓	✓	64.39	75.59	16.00	14.00	43.36	36.09	1.15	1.98	7.50	3.30	4.93 ± 0.99	5.87 ± 1.00	6.10 ± 0.91
M-Base	✗	✗	60.91	76.79	21.00	12.00	43.61	34.39	1.06	1.88	8.04	8.46	5.13 ± 0.94	5.60 ± 1.12	6.20 ± 0.84
M-Text	✓	✗	61.15	77.79	24.00	17.00	**45.31**	33.72	1.04	1.90	7.28	13.78	5.20 ± 0.99	6.03 ± 0.98	6.00 ± 1.02
M-Multi	✓	✓	**59.51**	**73.26**	**25.00**	**21.00**	**45.31**	**36.25**	**1.00**	**1.81**	**9.96**	8.72	6.20 ± 1.05	6.70 ± 1.06	7.07 ± 0.93

We evaluate the models using FAD, Precision, Recall, cosine similarity between audio embeddings, KL Divergence using a audio classifier, as well as subjective evaluations based on mood, genre, and quality.