Generative AI for Music and Audio

Hao-Wen (Herman) Dong
董皓文

UC San Diego

Generative AI for Music and Audio

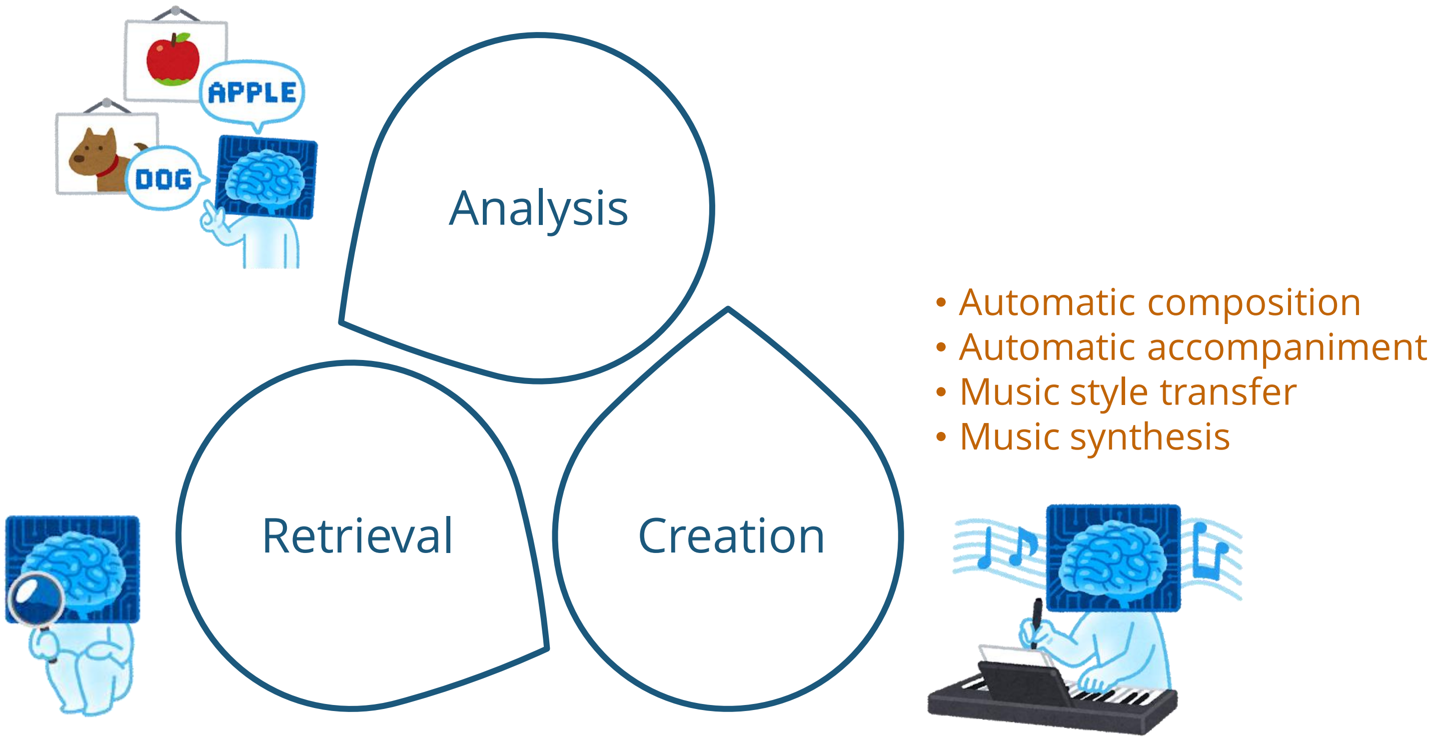
About me

Hi, I'm Herman.
I do AI x Music research.
I love music and movies!

- 2013 – 2017: 國立臺灣大學 National Taiwan University — *B.S. in Electrical Engineering*
- 2017 – 2019: 中央研究院 ACADEMIA SINICA — *Research Assistant*
- 2019 – 2021: UC San Diego — *M.S. in Computer Science*
- Summer 2019: YAMAHA — *Research Intern*
- Summer 2021: Dolby — *Deep Learning Audio Intern*
- Summer 2022: SONY — *Student Intern*
- Fall 2022: amazon — *Applied Scientist Intern*
- Winter 2023: Dolby — *Speech/Audio Deep Learning Intern*
- Summer 2023: Adobe — *Research Scientist/Engineer Intern*
- 2019 – *present*: UC San Diego — *Ph.D. in Computer Science (expected)*
- Fall 2023: NVIDIA — *Research Intern*

Music Information Research (MIR)

- *"Intelligent ways to analyze, retrieve and create music"* (Yang 2018)



Yang, "Music Information Research," *SNHCC*, TIGP, lecture notes, April 2018.

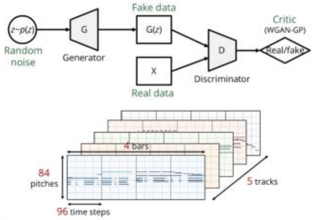
My Research

AI × Music

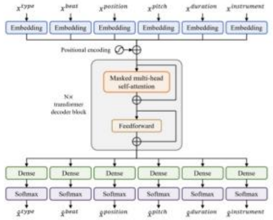
Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)



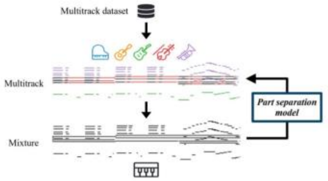
Multitrack Music Transformer
(ICASSP 2023)



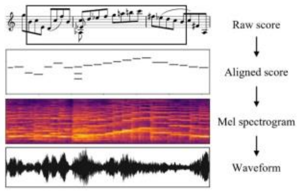
Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)



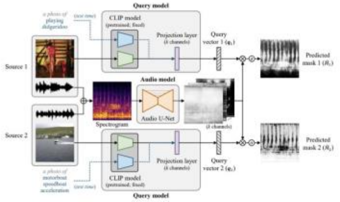
Deep Performer
(ICASSP 2022)



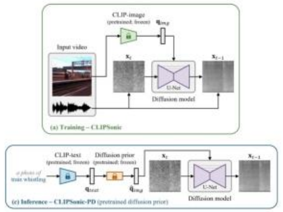
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



CLIPSonic
(WASPAA 2023)



About me

EE

a female cat engineer making
an electric chip in a classroom

Music

a cat playing heavy metal

CS

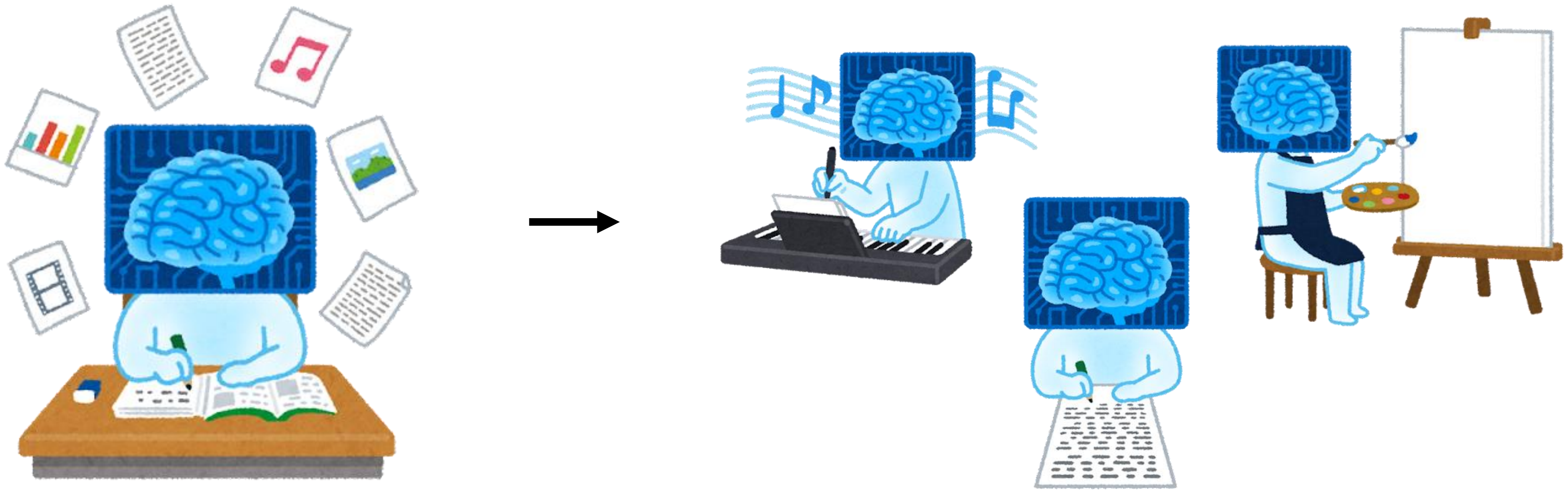
a cat engineer debugging on laptop

Made with Adobe Firefly (firefly.adobe.com) and Facebook MusicGen (huggingface.co/spaces/facebook/MusicGen)

Introduction

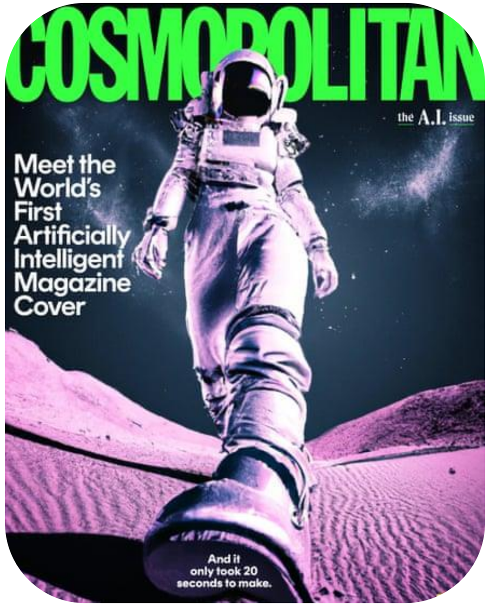
What is Generative AI?

- Generative AI is **AI capable of generating text, images, or other media**.
  - Learns the patterns and structure of their input training data
  - Generates new data that has similar characteristics



Generative AI for Visual Arts

AI made a magazine cover



(Source: Cosmopolitan)

AI won an art contest

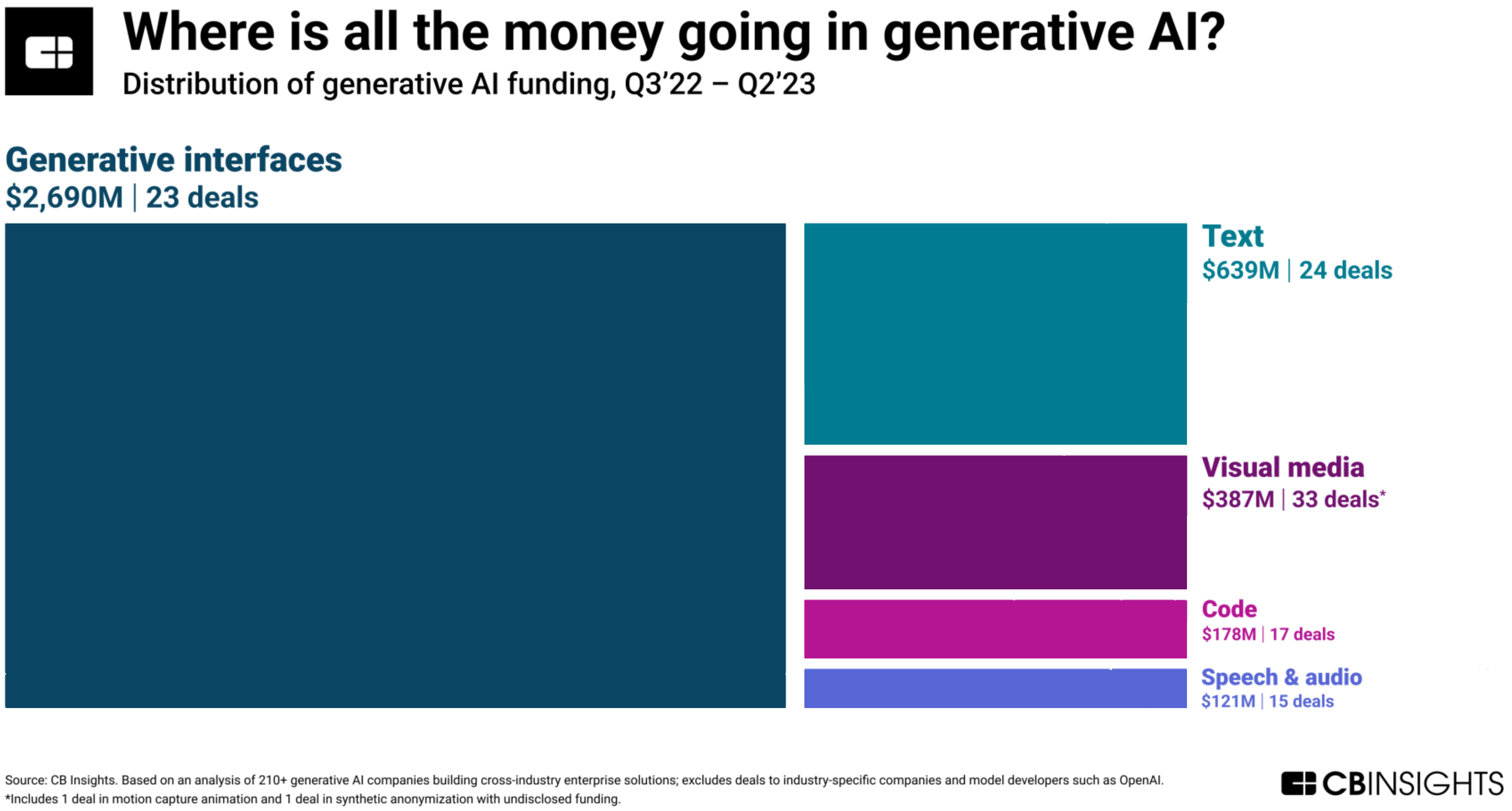
(Source: CNN Business)

AI won a photography contest

(Source: CNN)

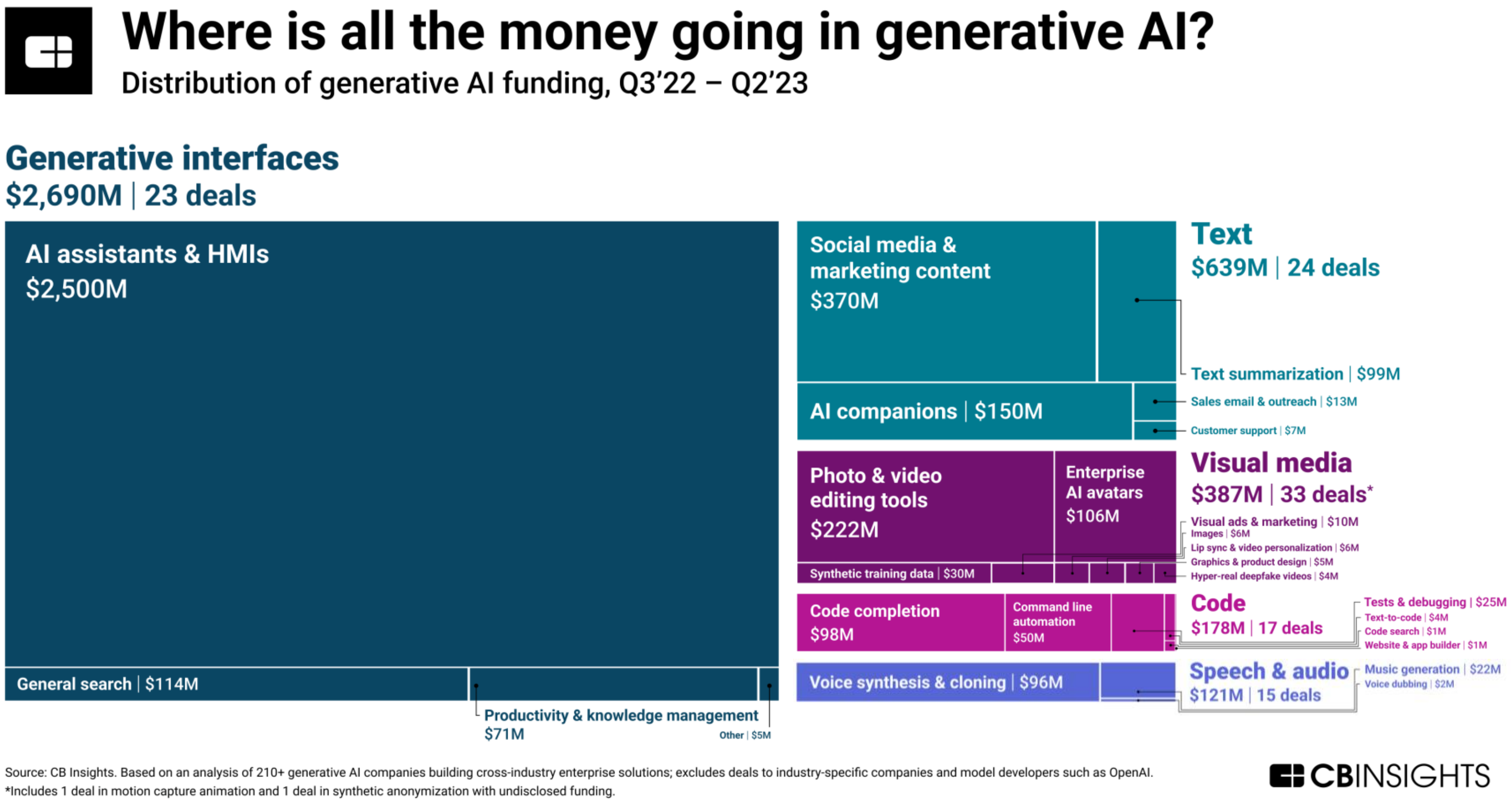
Gloria Liu, "The World's Smartest Artificial Intelligence Just Made Its First Magazine Cover," *Cosmopolitan*, June 21, 2022.
Rachel Metz, "AI won an art contest, and artists are furious," *CNN Business*, September 3, 2022.
Lianne Kolirin, "Artist rejects photo prize after AI-generated image wins award," *CNN*, April 18, 2023.

Landscape of Generative AI



CB Insights, "The generative AI landscape: Top startups, venture capital firms, and more," 2023.

Landscape of Generative AI



CB Insights, “The generative AI landscape: Top startups, venture capital firms, and more,” 2023.

Types of Audio

Speech

(Source: Wikimedia Commons)

Music

(Source: Wikimedia Commons)

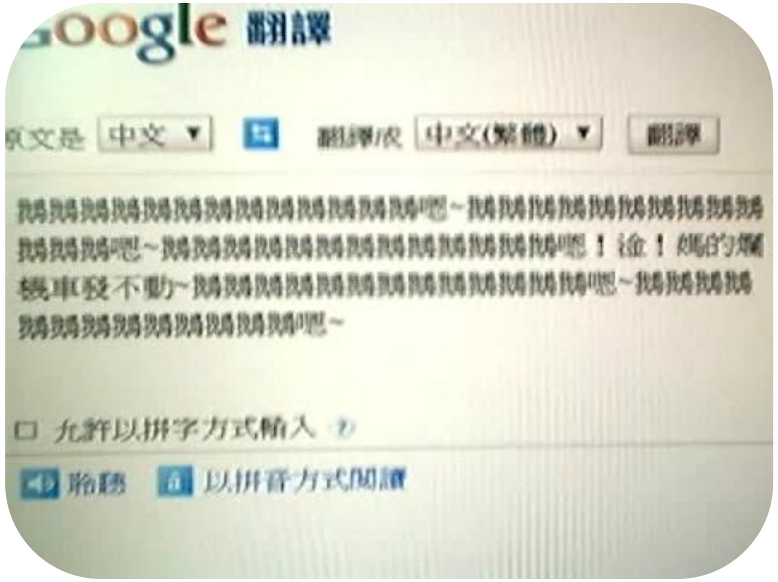
Sound effects

(Source: Wikimedia Commons)

BPJ Media Inc, CC BY-SA 3.0, via Wikimedia Commons.
Vancouver Film SchoolRetouched version by User:Quenhitran., CC BY 2.0, via Wikimedia Commons.
The Blackbird Academy, CC BY-SA 2.0, via Wikimedia Commons.
One Man Films, "One Shot - WAR ACTION SHORT FILM," *YouTube*, September 11, 2022.

Generative AI for Speech

Text-to-Speech



Voice Cloning

凌小海, "超爆笑的google翻譯~我一直狂笑!," *YouTube*, April 27, 2011.
Eric, *Twitter*, https://twitter.com/VyacheslavAI/status/1692144315055677818, August 17, 2023.

Generative AI for Music

Prompt: relaxing and smooth jazz played in a stylish cafe

Prompt: delightful country music with acoustic guitars

Prompt: cinematic and suspenseful orchestral music

Made with Facebook MusicGen (huggingface.co/spaces/facebook/MusicGen)

Generative AI for Sound Effects

Text-to-audio Synthesis

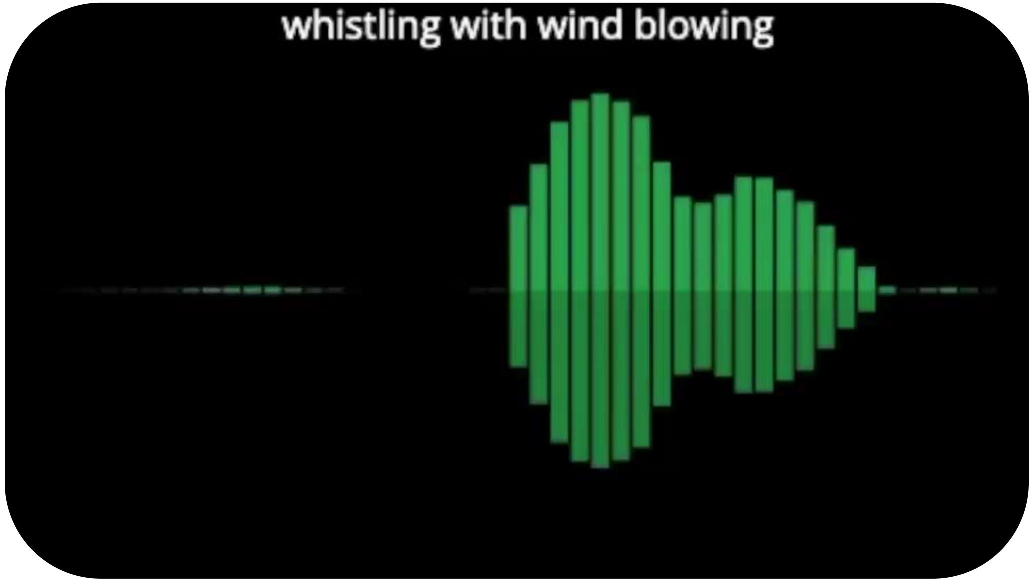


Image-to-audio Synthesis

Liu et al., "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *ICML*, 2023.
Dong et al., "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.

My Research

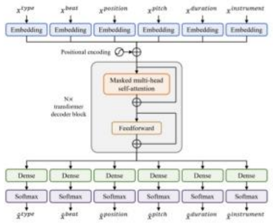
AI × Music

Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)

Multitrack Music Transformer
(ICASSP 2023)



Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)

Deep Performer
(ICASSP 2022)

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)

CLIPSonic
(WASPAA 2023)

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Multitrack Music Transformer

Hao-Wen Dong Ke Chen Shlomo Dubnov Julian McAuley Taylor Berg-Kirkpatrick

University of California San Diego

UC San Diego

Overview

Generate orchestral music

- of diverse instruments
- using a new compact representation
- with a multi-dimensional transformer

Demo

(Source: Vienna Mozart Orchestra)

Related Work (Transformers for Music Generation)

Model	Multitrack	Instrument control	Compound tokens	Generative modeling
REMI [5]				✓
MMM [10]	✓			✓
CP [6]			✓	✓
MusicBERT [15]	✓		✓	
FIGARO [11]	✓			✓
MMT (ours)	✓	✓	✓	✓

	Average sample length (sec)	Inference speed (notes per second)
MMM [10]	38.69	5.66
REMI+ [11]	28.69	3.58
MMT (ours)	**100.42**	**11.79**

→ Longer samples!
Faster inference speed!

Huang and Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," *Proc. MM*, 2020.
Ens and Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.
Hsiao et al., "Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs," *Proc. AAAI*, 2023.
Zeng et al., "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training," *Proc. Findings of ACL*, 2021.
von Rütte et al., "FIGARO: Controllable Music Generation using Learned and Expert Features," *Proc. ICLR*, 2023.

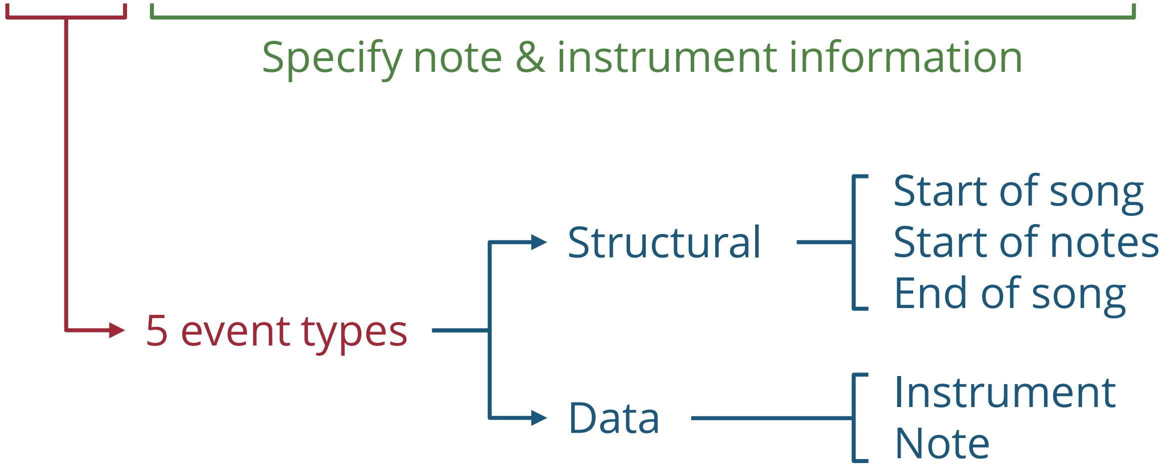
Representation

- We represent a music piece as a sequence of events

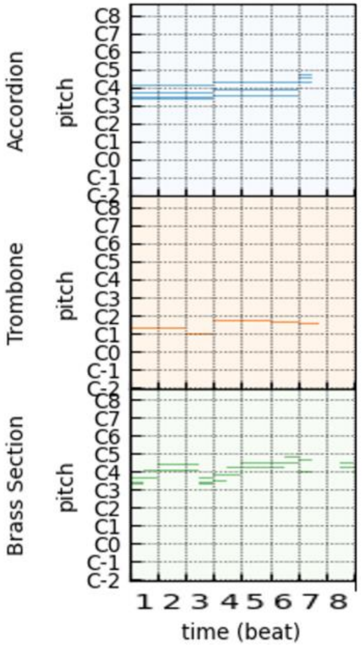
$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

- Each event $\mathbf{x}_i$ is encoded as

$$\mathbf{x}_i = (x_i^{type}, x_i^{beat}, x_i^{position}, x_i^{pitch}, x_i^{duration}, x_i^{instrument})$$



Representation (An Example)



Structural events

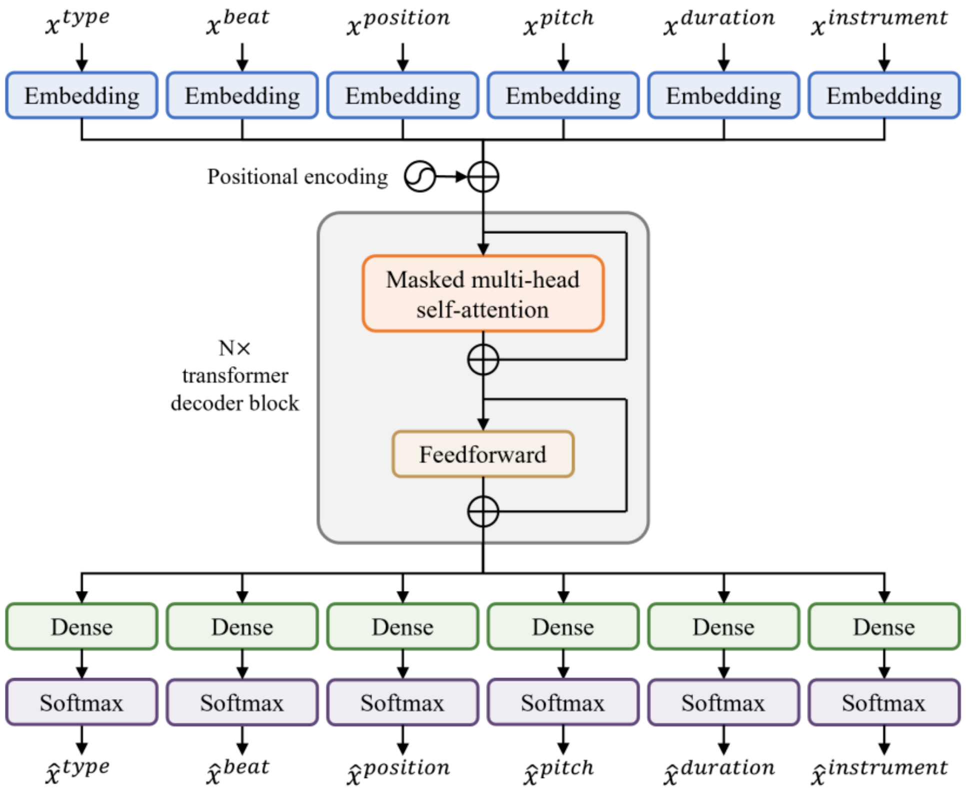
Instrument events

Note events

(0, 0, 0, 0, 0, 0)	Start of song
(1, 0, 0, 0, 0, 15)	Instrument: accordion
(1, 0, 0, 0, 0, 36)	Instrument: trombone
(1, 0, 0, 0, 0, 39)	Instrument: brasses
(2, 0, 0, 0, 0, 0)	Start of notes
(3, 1, 1, 41, 15, 36)	Note: beat=1, position=1, pitch=E2, duration=48, instrument=trombone
(3, 1, 1, 65, 4, 39)	Note: beat=1, position=1, pitch=E4, duration=12, instrument=brasses
(3, 1, 1, 65, 17, 15)	Note: beat=1, position=1, pitch=E4, duration=72, instrument=accordion
(3, 1, 1, 68, 4, 39)	Note: beat=1, position=1, pitch=G4, duration=12, instrument=brasses
(3, 1, 1, 68, 17, 15)	Note: beat=1, position=1, pitch=G4, duration=72, instrument=accordion
(3, 1, 1, 73, 17, 15)	Note: beat=1, position=1, pitch=C5, duration=72, instrument=accordion
(3, 1, 13, 68, 4, 39)	Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3, 1, 13, 73, 4, 39)	Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3, 2, 1, 73, 12, 39)	Note: beat=2, position=1, pitch=C5, duration=36, instrument=brasses
(3, 2, 1, 77, 12, 39)	Note: beat=2, position=1, pitch=E5, duration=36, instrument=brasses
...	...
(4, 0, 0, 0, 0, 0)	End of song

Multitrack Music Transformer

- A multi-dimensional decoder-only transformer model
  - Predict six fields *at the same time*
- Trained autoregressively
  - Predict the next event given past events



Three Sampling Modes

Unconditional generation

Input

```
(0, 0,  0,  0,  0,  0)  Start of song
(1, 0,  0,  0,  0, 15)  Instrument: accordion
(1, 0,  0,  0,  0, 36)  Instrument: trombone
(1, 0,  0,  0,  0, 39)  Instrument: brasses
(2, 0,  0,  0,  0,  0)  Start of notes
(3, 1,  1, 41, 15, 36)  Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39)  Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15)  Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39)  Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15)  Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15)  Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3, 1, 13, 68,  4, 39)  Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3, 1, 13, 73,  4, 39)  Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3, 2,  1, 73, 12, 39)  Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3, 2,  1, 77, 12, 39)  Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
         ...            ...
(4, 0,  0,  0,  0,  0)  End of song
```

Only need to train ONE model!

Instrument-informed generation

Input

```
(0, 0,  0,  0,  0,  0)  Start of song
(1, 0,  0,  0,  0, 15)  Instrument: accordion
(1, 0,  0,  0,  0, 36)  Instrument: trombone
(1, 0,  0,  0,  0, 39)  Instrument: brasses
(2, 0,  0,  0,  0,  0)  Start of notes
(3, 1,  1, 41, 15, 36)  Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39)  Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15)  Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39)  Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15)  Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15)  Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3, 1, 13, 68,  4, 39)  Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3, 1, 13, 73,  4, 39)  Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3, 2,  1, 73, 12, 39)  Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3, 2,  1, 77, 12, 39)  Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
         ...            ...
(4, 0,  0,  0,  0,  0)  End of song
```

N-beat continuation

Input

```
(0, 0,  0,  0,  0,  0)  Start of song
(1, 0,  0,  0,  0, 15)  Instrument: accordion
(1, 0,  0,  0,  0, 36)  Instrument: trombone
(1, 0,  0,  0,  0, 39)  Instrument: brasses
(2, 0,  0,  0,  0,  0)  Start of notes
(3, 1,  1, 41, 15, 36)  Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39)  Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15)  Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39)  Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15)  Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15)  Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
(3, 1, 13, 68,  4, 39)  Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
(3, 1, 13, 73,  4, 39)  Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
(3, 2,  1, 73, 12, 39)  Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
(3, 2,  1, 77, 12, 39)  Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
         ...            ...
(4, 0,  0,  0,  0,  0)  End of song
```

Experimental Setup

Data

- Symbolic Orchestral Database (SOD) (Crestel et al., 2017)
  - 5,743 songs, 357 hours
- Temporal resolution: 12 time steps per quarter note
- 80% training, 10% validation, 10% test
- Data augmentation
  - Randomly shift for -5~6 semitones
  - Randomly select a starting beat

Model & Training

- 6 transformer decoder blocks
- 8 attention heads
- Model dimension: 512
- Sequence length: 1024
- Maximum number of beats: 256
- Maximum training steps: 200,000

Crestel et al., "A database linking piano and orchestral MIDI scores with application to automatic projective orchestration," *Proc. ISMIR*, 2017.

Example Results

Unconditional generation

Instrument-informed generation

church-organ, viola, contrabass, strings, voices, horn, oboe

4-beat continuation

Wolfgang Amadeus Mozart's Eine kleine Nachtmusik

More audio samples

salu133445.github.io/mmt/

OttoPilot33, "Eine Kleine Nachtmusik - By My MIDI Virtual Orchestra." https://youtu.be/_g8iCdY34jc

Subjective Listening Test Results

	Number of parameters	Average sample length (sec)	Inference speed (notes per second)	Subjective listening test results			
				Coherence	Richness	Arrangement	Overall
MMM [10]	19.81 M	38.69	5.66	3.48 ± 0.35	3.05 ± 0.38	3.28 ± 0.37	3.17 ± 0.43
REMI+ [11]	20.72 M	28.69	3.58	**3.90 ± 0.52**	**3.74 ± 0.21**	**3.74 ± 0.44**	**3.77 ± 0.41**
MMT (ours)	19.94 M	**100.42**	**11.79**	3.55 ± 0.46	3.53 ± 0.35	3.40 ± 0.44	3.33 ± 0.47

2.6x/3.5x longer generated samples
(within the same sequence length)

2.1x/3.3x faster inference speed

Higher quality than MMM
Lower quality than REMI+

Ens and Pasquier, "MMM : Exploring Conditional Multi-Track Music Generation with the Transformer," *arXiv preprint arXiv:2008.06048*, 2020.
von Rütte et al., "FIGARO: Controllable Music Generation using Learned and Expert Features," *Proc. ICLR*, 2023.

Analyzing Self-attention

- *Mean relative attention* for a field d:

$$\gamma_k^{(d)} = \frac{\sum_{x\in\mathcal{D}}\sum_{s>t} a_{s,t}(\mathbf{x})\, 1_{x_t^{(d)}-x_s^{(d)}=k}}{\sum_{x\in\mathcal{D}}\sum_{s>t} a_{s,t}(\mathbf{x})}$$

Attention weight: $a_{s,t}(\mathbf{x})$

Whether the field value is of difference k: $1_{x_t^{(d)}-x_s^{(d)}=k}$

```
(0, 0,  0,  0,  0,  0)  Start of song
(1, 0,  0,  0,  0, 15)  Instrument: accordion
(1, 0,  0,  0,  0, 36)  Instrument: trombone
(1, 0,  0,  0,  0, 39)  Instrument: brasses
(2, 0,  0,  0,  0,  0)  Start of notes
(3, 1,  1, 41, 15, 36)  Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
(3, 1,  1, 65,  4, 39)  Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
(3, 1,  1, 65, 17, 15)  Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
(3, 1,  1, 68,  4, 39)  Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
(3, 1,  1, 68, 17, 15)  Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
(3, 1,  1, 73, 17, 15)  Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
```

$\gamma_{-8}^{(pitch)}$

$\gamma_{-5}^{(pitch)}$

Analyzing Self-attention

- *Mean relative attention* for a field d:

$$\gamma_k^{(d)} = \frac{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})\, 1_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{x \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})}$$

Biased towards difference that occurred more frequently!

- *Mean relative attention gain* for a field d:

$$\tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{x \in \mathcal{D}} \sum_{s>t} 1_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{x \in \mathcal{D}} \sum_{s>t} \boxed{1}}$$

Assuming a uniform attention matrix

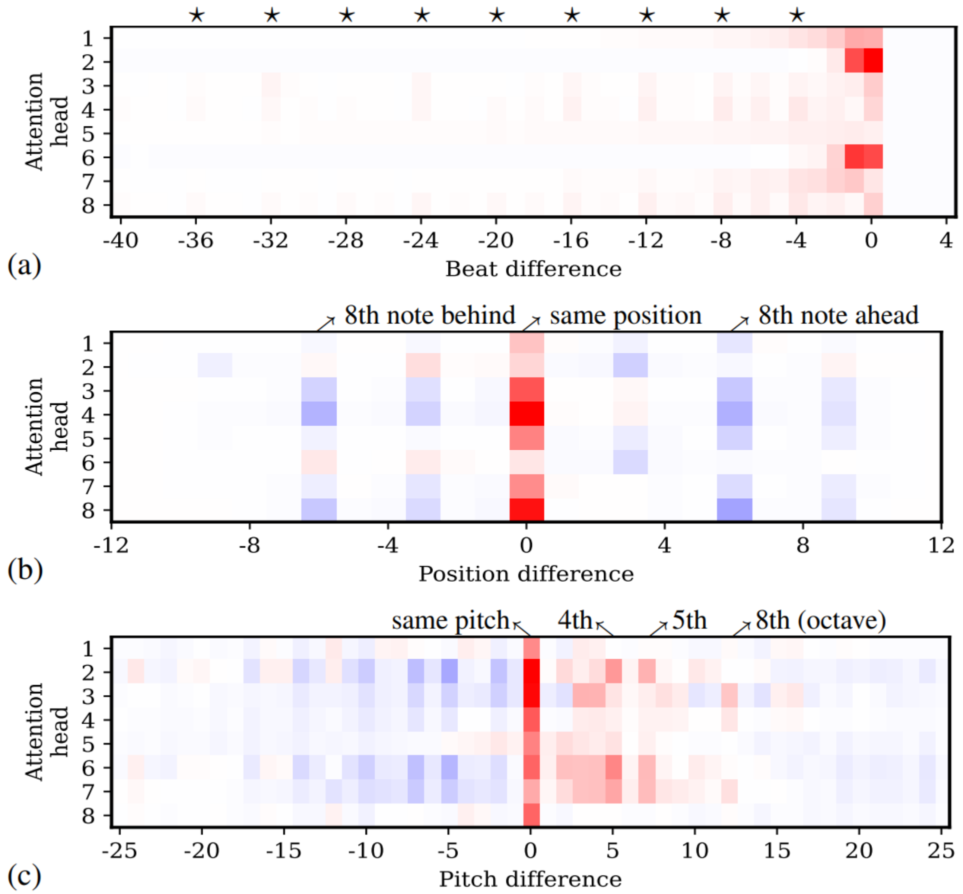
Musical Self-attention

The MMT model attends more to notes

- that are $4N$ beats away in the past
- that have the same position as the current note (A note on beat attends more to a note on beat; a note off beat attends more to a note off beat.)
- that has a pitch in an octave above which forms a consonant interval

→ MMT learns a relative self-attention for certain aspects of music, specifically, beat, position and pitch.

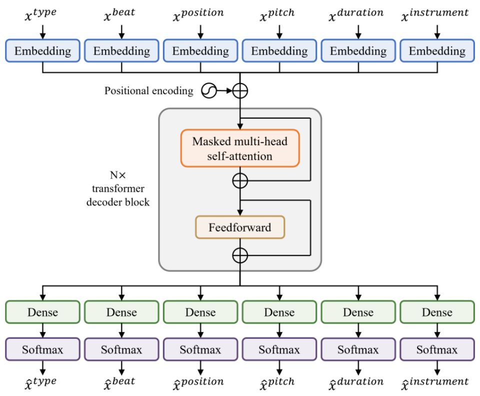
Positive and negative mean relative attention gain



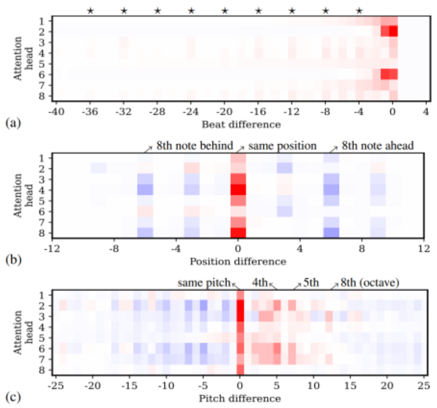
Summary

- Proposed an efficient representation and model for multitrack music generation
- Presented the first systematic analysis of musical self-attention

Multitrack Music Transformer



Musical Self-attention



Paper: arxiv.org/abs/2207.06983
Demo: salu133445.github.io/mmt/
Code: github.com/salu133445/mmt





My Research

AI × Music

Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)

Multitrack Music Transformer
(ICASSP 2023)

Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)

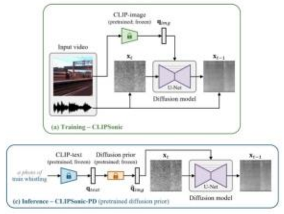
Deep Performer
(ICASSP 2022)

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)

CLIPSonic
(WASPAA 2023)

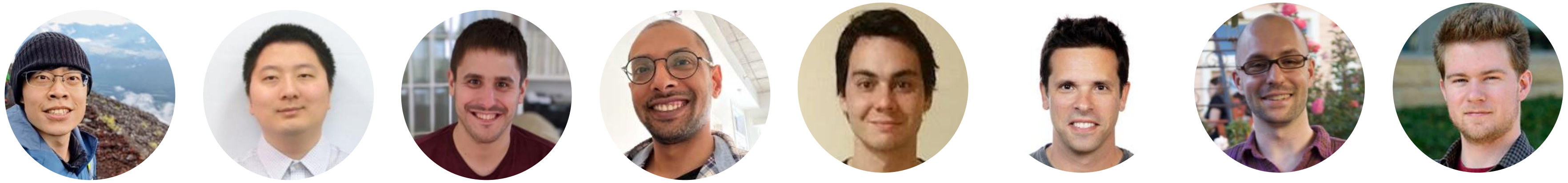


IEEE WASPAA 2023

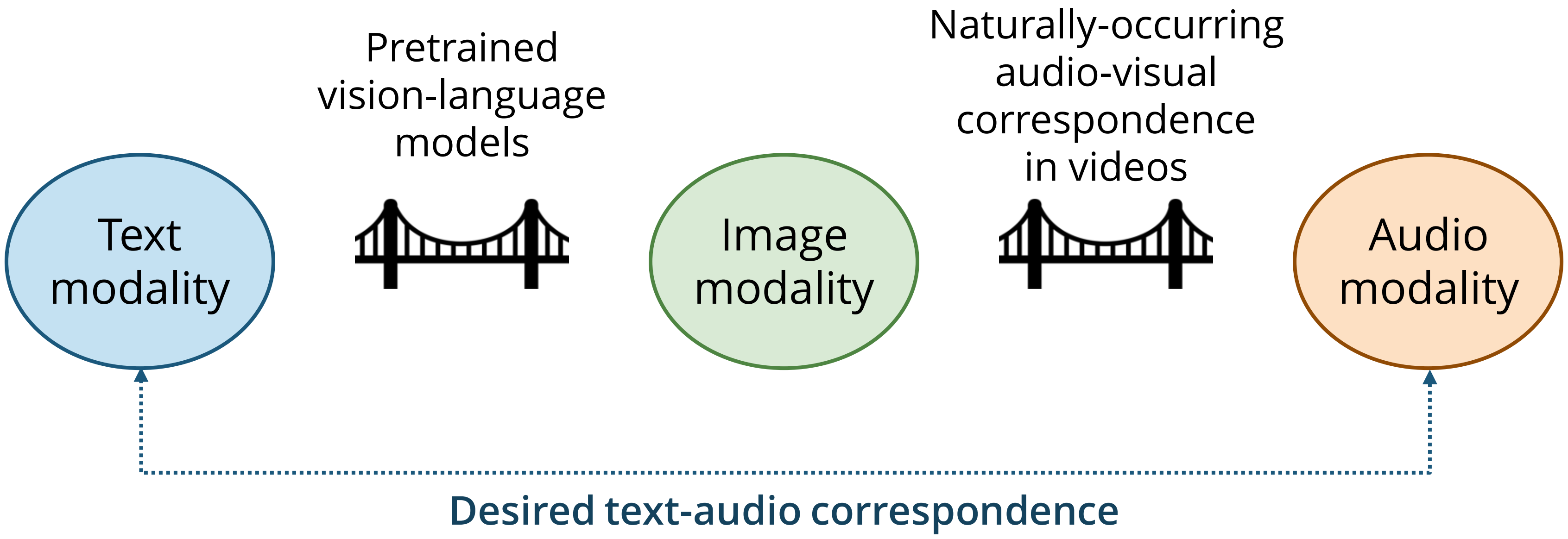
CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models

Hao-Wen Dong[1,2]* Xiaoyu Liu[1] Jordi Pons[1] Gautam Bhattacharya[1]
Santiago Pascual[1] Joan Serrà[1] Taylor Berg-Kirkpatrick[2] Julian McAuley[2]

[1] Dolby Laboratories [2] University of California San Diego
* Work done during an internship at Dolby



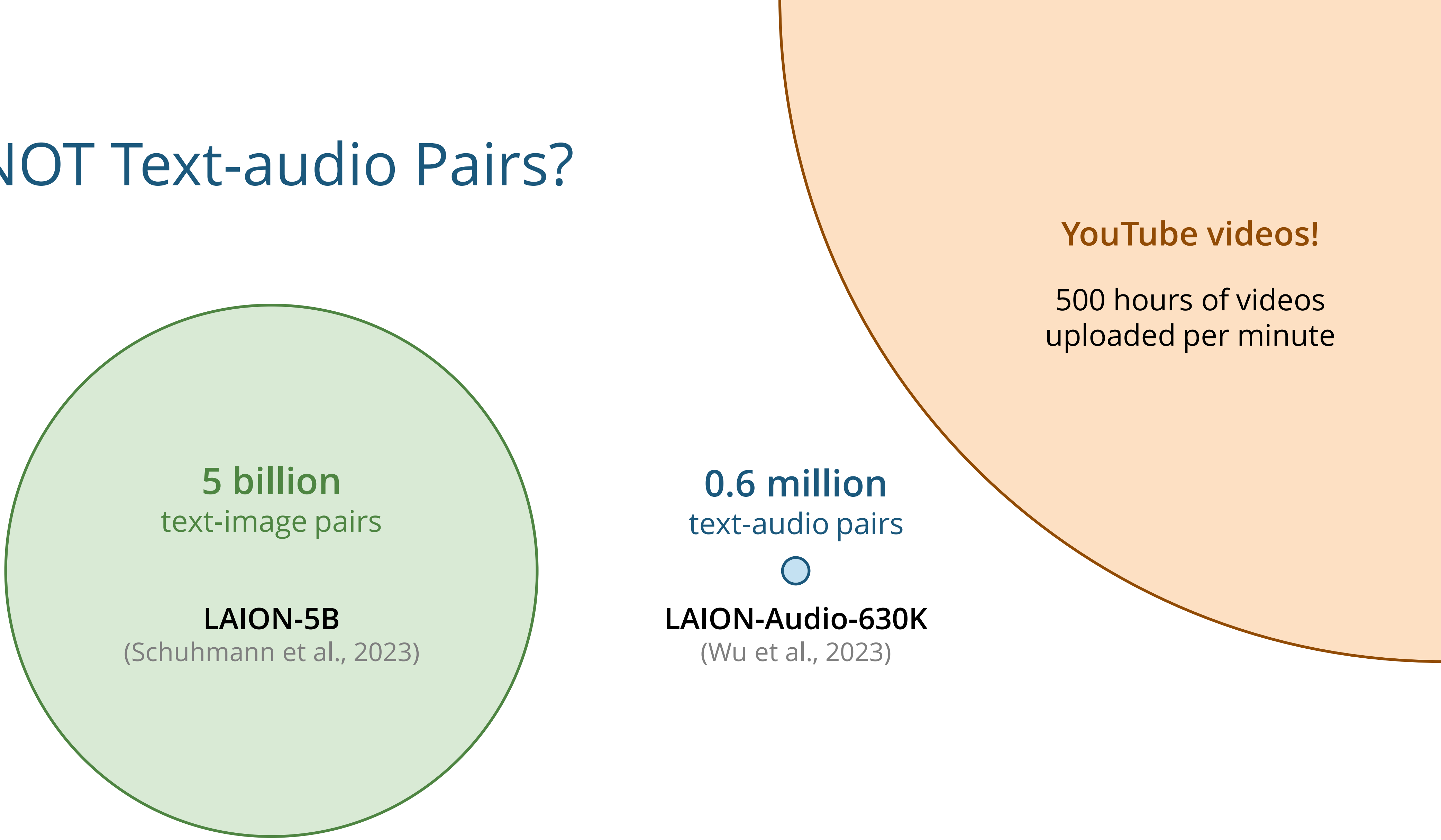
Leveraging the Visual Domain as a Bridge



No text-audio pairs required!

Scalable to large video datasets!

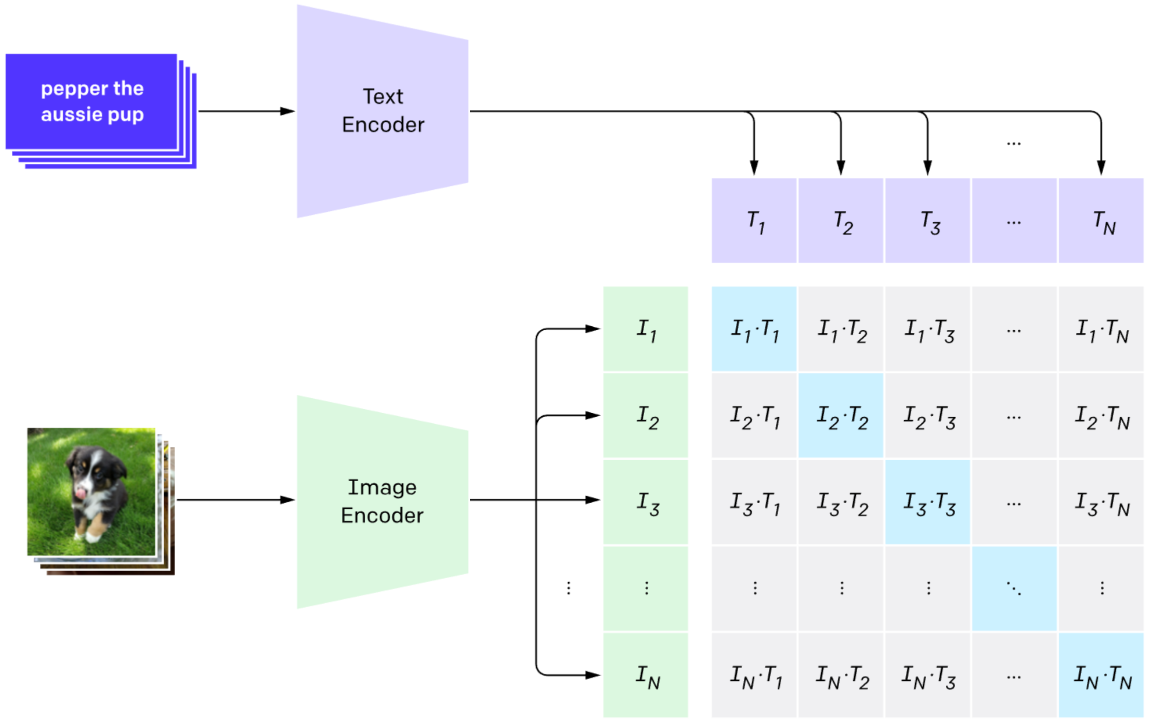
Why NOT Text-audio Pairs?



Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *Proc. NeurIPS, Datasets and Benchmarks Track*, 2023.
Wu et al., "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," *Proc. ICASSP*, 2023.

CLIP (Contrastive Language-Image Pretraining)

- Learned a shared embedding space for images and texts via *contrastive learning*

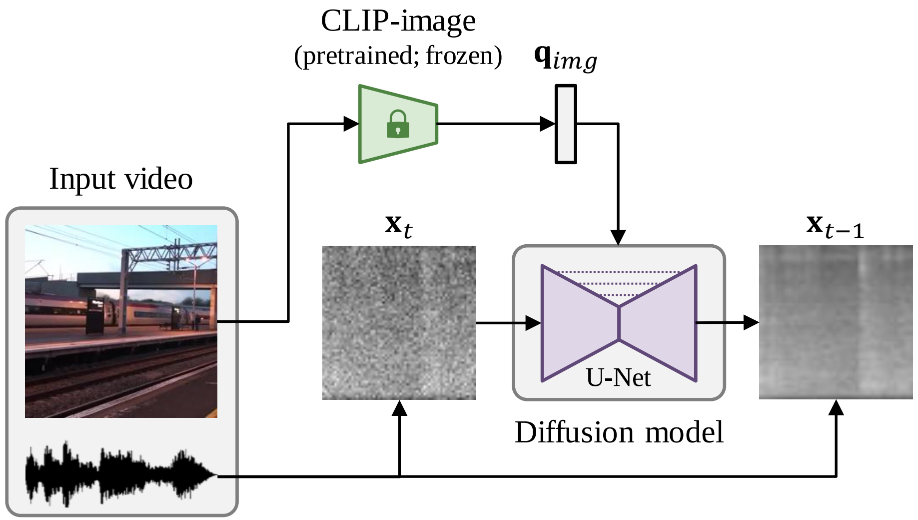


Contrastive loss

Radford et al., "CLIP: Connecting Text and Images," OpenAI Blog, January 5, 2021.
Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proc. ICML*, 2021.

CLIPSonic – Training

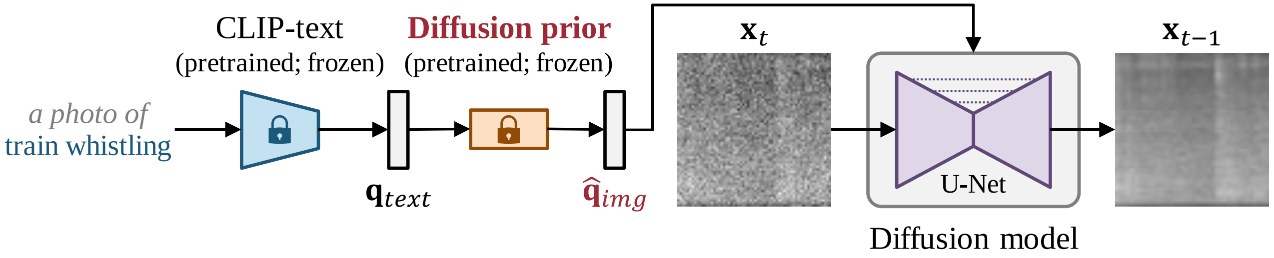
- We train the model to perform image-to-audio synthesis
  - Encode a video frame using a pretrained CLIP-image encoder (Radford et al., 2021)



Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proc. ICML*, 2021.

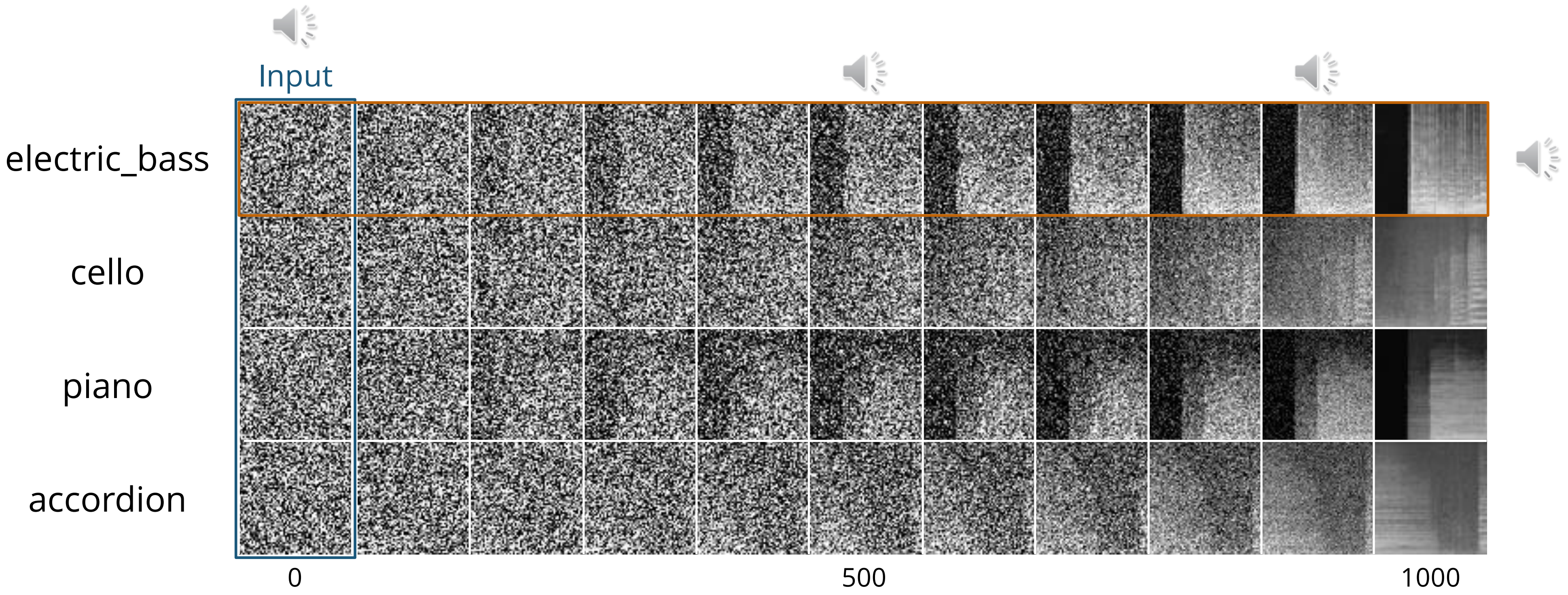
CLIPSonic – Inference

- We use a pretrained diffusion prior model (Ramesh et al., 2022)
  - To generate a CLIP-image embedding given a CLIP-text embedding



Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

CLIPSonic – Inference Examples



Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

Zhao et al., "The Sound of Pixels," *Proc. ECCV*, 2018. (dataset)
Chen et al., "VGGSound: A Large-Scale Audio-Visual Dataset," *Proc. ICASSP*, 2020. (dataset)

Text-to-Audio Synthesis Demo

Demo

Rapping	Sea waves	Smoke detector beeping	Playing table tennis	Thunder	Playing violin fiddle

Subjective Listening Test (text-to-audio synthesis)

- CLIPSonic-PD with a pretrained diffusion prior model performs significantly better
  - than its counterpart *without* a diffusion prior model (CLIPSonic-ZS)
  - in terms of fidelity on both datasets
  - in terms of relevance on MUSIC

Table 3: Listening test results for text-to-audio synthesis (MOS).

Model	VGGSound		MUSIC	
	Fidelity	Relevance	Fidelity	Relevance
CLIPSonic-ZS	2.55 ± 0.22	2.01 ± 0.27	2.98 ± 0.23	3.87 ± 0.24
CLIPSonic-PD	$\mathbf{3.04 \pm 0.20}$	2.86 ± 0.25	$\mathbf{3.67 \pm 0.18}$	3.91 ± 0.24
Ground truth	3.78 ± 0.19	3.54 ± 0.29	3.90 ± 0.17	4.34 ± 0.18

Significant performance improvement

Image-to-Audio Synthesis Demo (out-of-distribution)

Demo



CLIPSonic-IQ (ours)

Im2wav (Sheffer & Adi, 2023)

SpecVQGAN (Iashin & Rahtu, 2021)



CLIPSonic-IQ (ours)

Im2wav (Sheffer & Adi, 2023)

SpecVQGAN (Iashin & Rahtu, 2021)

Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *Proc. ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *Proc. BMVC*, 2021.

Subjective Listening Test (image-to-audio synthesis)

- CLIPSonic-IQ significantly outperforms im2wav and SpecVQGAN in audio fidelity
- CLIPSonic-IQ significantly outperforms SpecVQGAN in text-audio relevance
- CLIPSonic-IQ is competitive against im2wav in text-audio relevance

Table 4: Listening test results for image-to-audio synthesis (MOS).

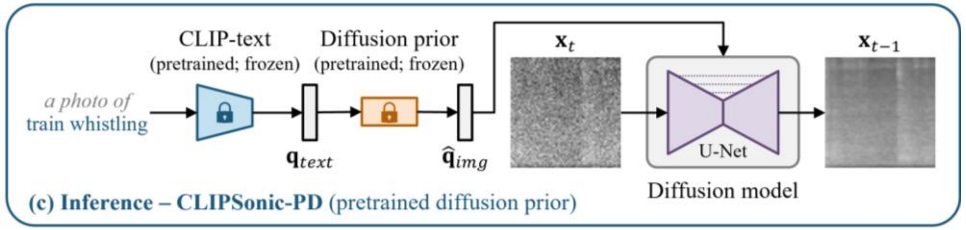
Model	Fidelity	Relevance
CLIPSonic-IQ (image-queried)	$\mathbf{3.29 \pm 0.16}$	3.80 ± 0.19
SpecVQGAN [20]	2.15 ± 0.17	2.54 ± 0.23
im2wav [21]	2.19 ± 0.15	$\mathbf{3.90 \pm 0.22}$

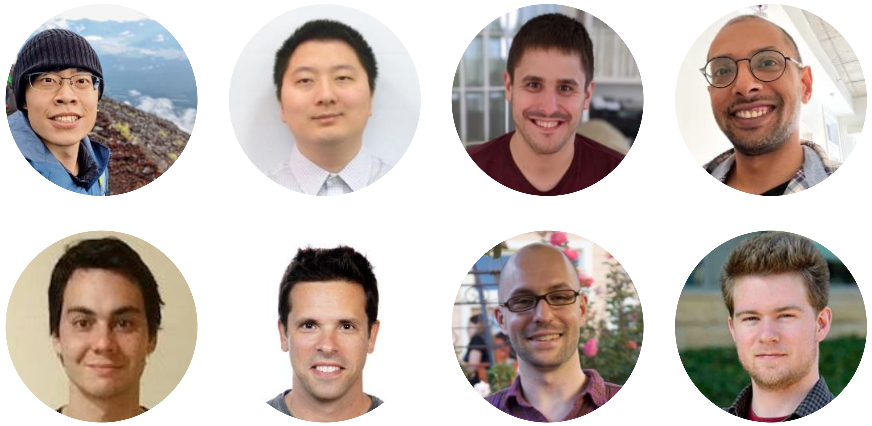
State-of-the-art image-to-audio performance!

Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *Proc. ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *Proc. BMVC*, 2021.

Summary

- Proposed a new text-to-audio synthesis model that requires *no* text-audio pairs
- CLIPSonic-PD achieves good performance in objective and subjective evaluations
- CLIPSonic-IQ achieves state-of-the-art performance in image-to-audio synthesis



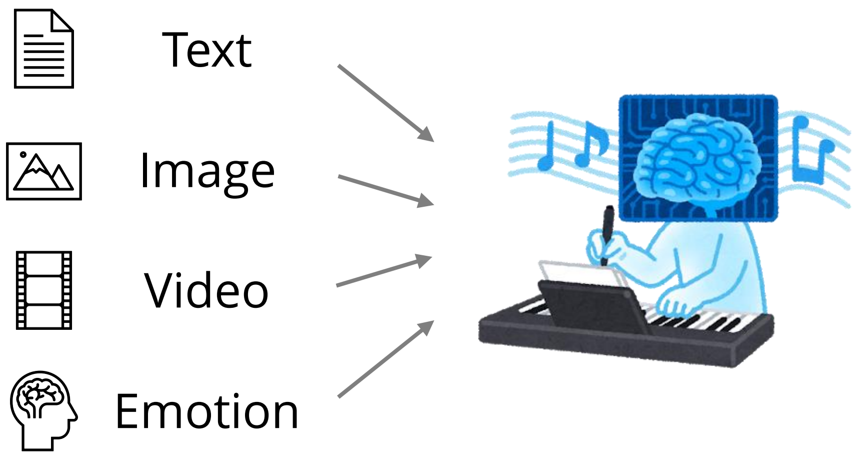


Paper: arxiv.org/abs/2306.09635
Demo: salu133445.github.io/clipsonic

Future of Generative AI

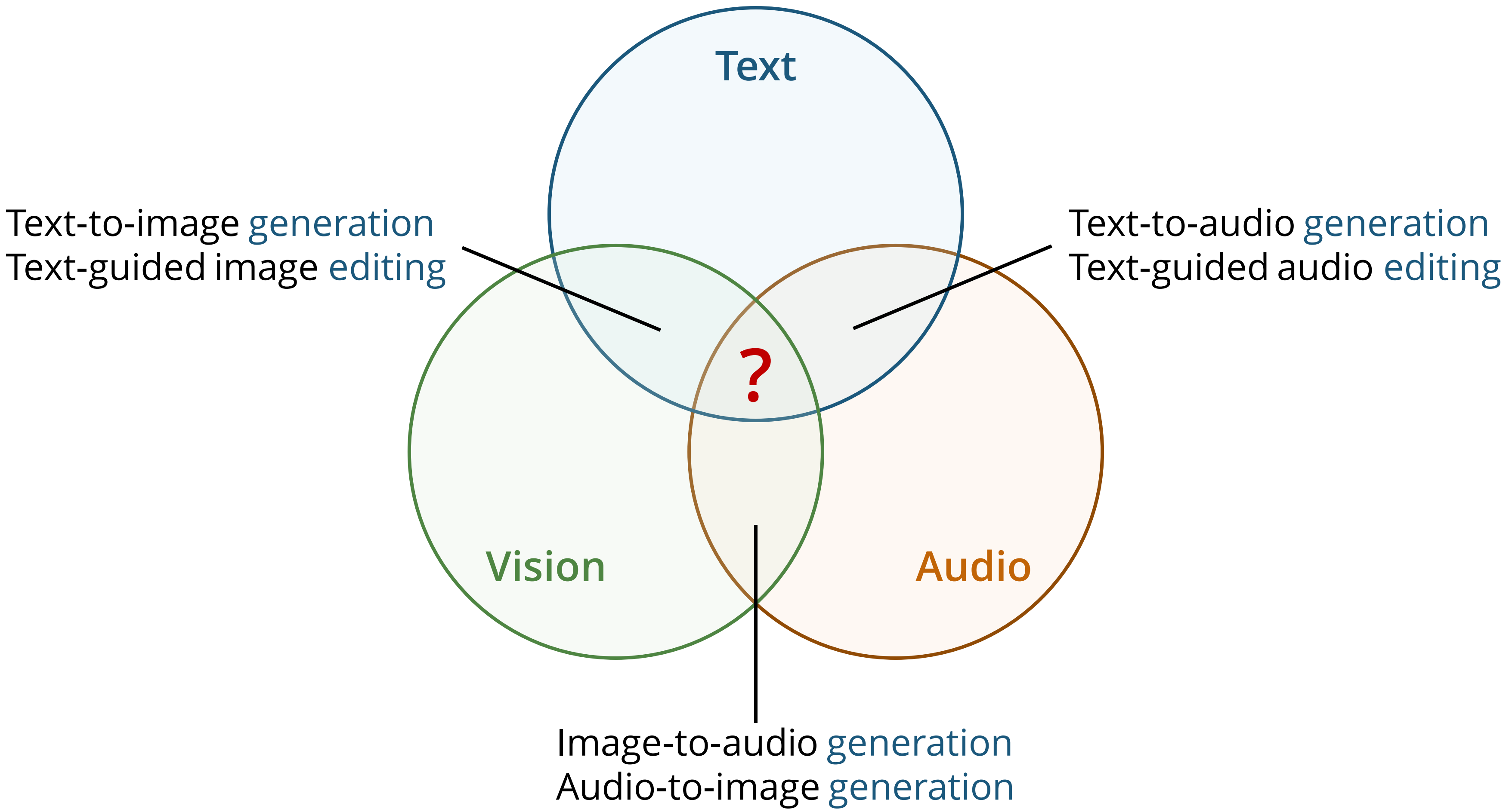
Challenges

Multimodality



Usability

Multimodal Generative AI



Multimodal generative AI for Ads

Video **Runway Gen-2**

Music **MusicGen**

進捗共有チャンネル, "Runway Gen-2 Sample / Tiger Whiskey," *YouTube*, July 9, 2023.
Kaoru Naito, *Twitter*, https://twitter.com/ka0ru_1620/status/1678313226453520385, July 10, 2023.

Multimodal generative AI for Films

Visuals	**Midjourney**
Video	**Runway**
Narration (script)	**ChatGPT**
Narration (voice)	**ElevenLabs**
Sound effects	**Audiocraft**

Sagar Amruth, *Twitter*, https://twitter.com/sagartheamruth/status/1688530899271364608, August 7, 2023.


Mumbai, the city of dreams.

Generative AI for News

Generate an audio in Science Fiction theme: Mars News reporting that Humans send light-speed probe to Alpha Centauri.
Start with news anchor, followed by a reporter interviewing a chief engineer from an organization that built this probe, founded by United Earth and Mars Government, and end with the news anchor again.

Script	**GPT-4**
Music	**MusicGen**
Narration	**Bark**
Sound effects	**AudioLDM**

Liu et al., "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

Controllable Generative AI

Large language models (GPT-4)

Pretrained generative audio models (MusicGen, AudioLDM, Bark)

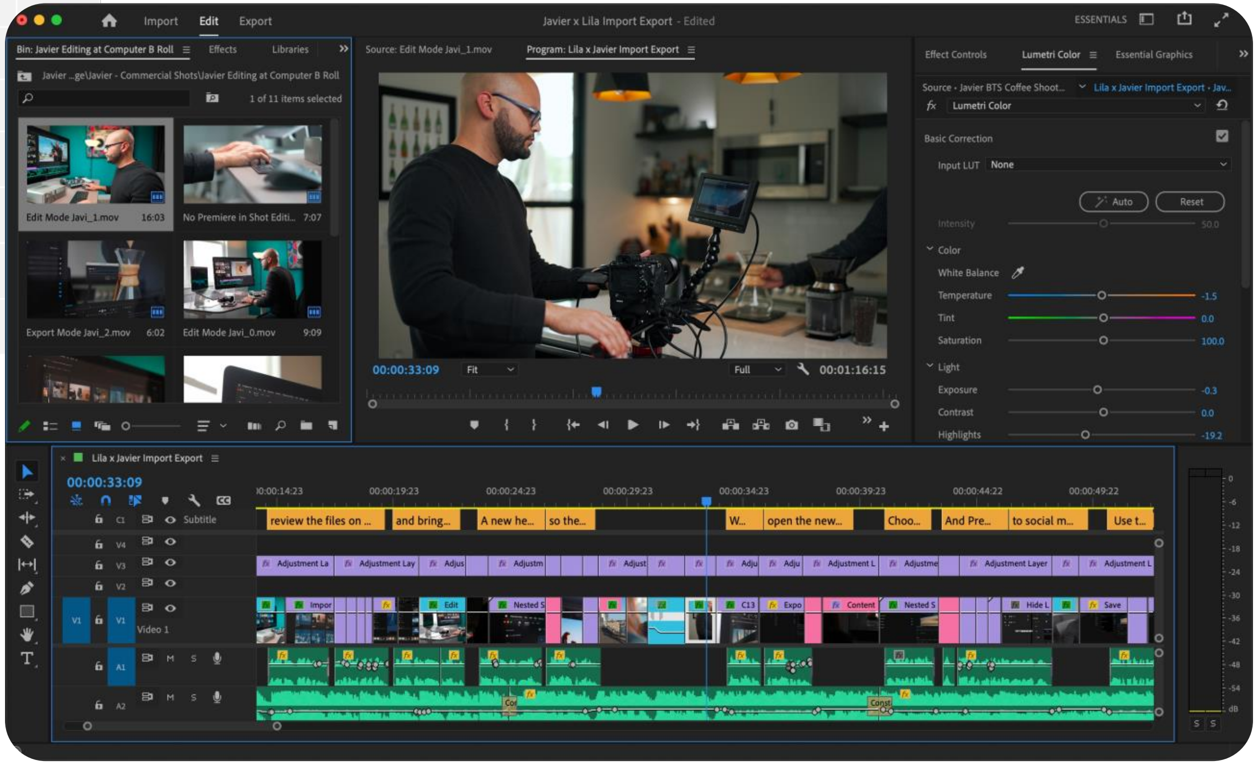
Instructions → Audio Script → Audio

Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with our on-site reporter ...	Auto
Sound effect	Foreground	N/A	N/A	-35	N/A	Transition swoosh.	1
Sound effect	Background	2	N/A	-30	Begin	Background noise of busy engineering office.	Auto
Speech	Foreground	N/A	Reporter	-15	N/A	We're here at the headquarters of …	Auto
Speech	Foreground	N/A	Director	-15	N/A	Thank you, so it's a fantastic …	Auto
Speech	Foreground	N/A	Reporter	-15	N/A	This is truly an impressive feat …	Auto

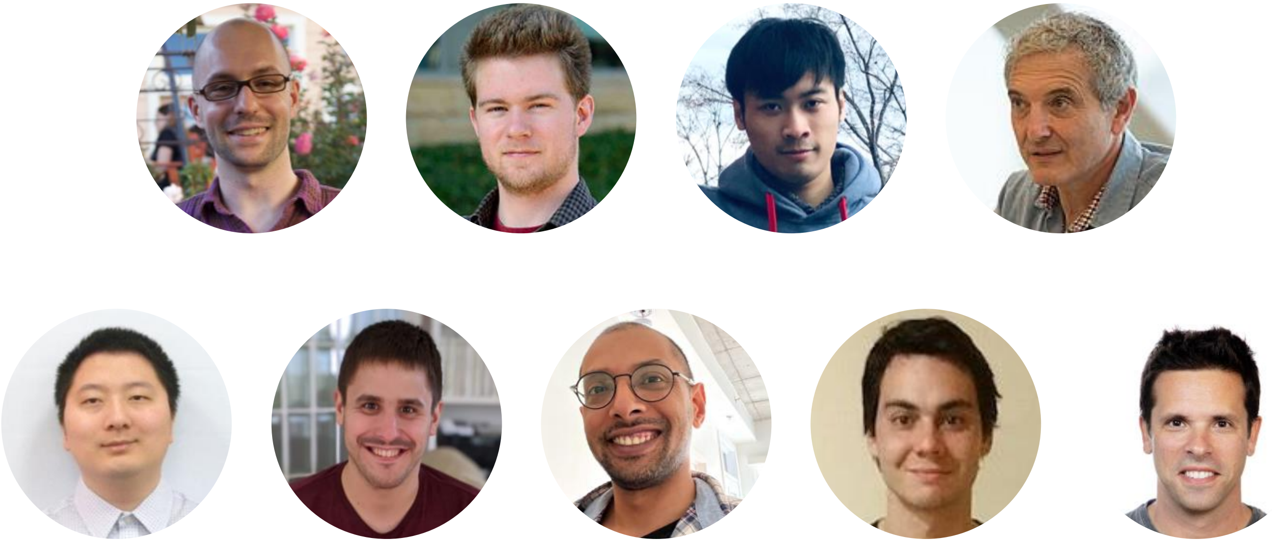
Liu et al., "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

Controllable Generative AI

Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with our on-site reporter ...	
Sound effect	Foreground	N/A	N/A	-35	N/A	Transition swoosh.	
Sound effect	Background	2	N/A	-30	Begin	Background noise of busy engineering office.	
Speech	Foreground	N/A	Reporter	-15	N/A	We're here at the headquarters of …	
Speech	Foreground	N/A	Director	-15	N/A	Thank you, so it's a fantastic …	
Speech	Foreground	N/A	Reporter	-15	N/A	This is truly an impressive feat …	



Acknowledgements



Thank you!