Augmenting Human Creativity with Generative AI

Hao-Wen (Herman) Dong

Department of Performing Arts Technology
School of Music, Theatre & Dance
University of Michigan
hermandong.com

March 5, 2025

My Background

Electrical Engineering

a female cat engineer making an electric chip in a classroom

Music

a cat playing heavy metal

Computer Science

a cat engineer debugging on laptop

Made in September 2023!

Made with Adobe Firefly

My Background

Electrical Engineering

a cat engineer making an
electric chip in a classroom

Music

a cat playing heavy metal

January 2025

Made in ~~September 2023~~!

Computer Science

a cat engineer debugging
Python programs on a laptop

Made with Adobe Firefly

Analytic AI vs Generative AI

Made with Adobe Firefly

Music & Technology Co-evolves

Hildegard Dodel, Public domain, via Wikimedia Commons.
Taken at Hamamatsu Museum of Musical Instruments, August 2019.
yan, CC BY-SA 4.0, via Wikimedia Commons.

Music & AI

(Source: Yamaha)

(Source: Sankei Shimbun)

(Shlizerman et al., 2019)

(Source: Robot Gizmos)

(Source: NBC DFW)

yamaha.com/en/news_release/2018/18013101/
sankei.com/article/20240113-CQCOSQHJWFIYPJJKZDCITRTRVI/
roboticgizmos.com/shimon-musical-robot-deep-learning/
nbcdfw.com/entertainment/the-scene/how-verdigris-ensemble-is-using-ai-to-create-a-new-concert-experience/3366031/
Eli Shlizerman, Lucio M. Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman, "Audio to Body Dynamics," *CVPR*, 2018.

Generative AI for Content Creation

(Source: The Denver Post)

(Source: UploadVR)

(Source: Descript)

(Source: Daily Bruin)

(Source: Wikimedia Commons)

Universitaetsmedizin, CC BY-SA 4.0, via Wikimedia Commons
uploadvr.com/iron-man-vr-breaks-free-from-cords-load-screens-on-quest-2/
descript.com/blog/article/what-is-the-best-audio-interface-for-recording-a-podcast
denverpost.com/2019/08/02/colorado-symphony-movie-scores-harry-potter-star-wars/
dailybruin.com/2023/08/04/theater-review-the-musical-les-misrables-offers-stellar-displays-and-impassioned-vocals

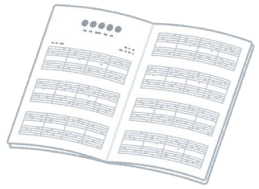
Challenges in Generative AI for Content Generation

- **New application domains may require new generative models**
- **Professionals need assistive tools that augment their creativity and productivity** in addition to fully automated tools
- **Certain media require handling multimodal data streams** at the same time

Augmenting Human Creativity with AI

- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024, AIMG 2024), **text-to-symbolic music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)
- **AI-Assisted Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **Queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WSS 2023, WASPAA 2023), **text-to-music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)

Music Generation – Four Paradigms



Symbolic music generation

Audio-domain music generation

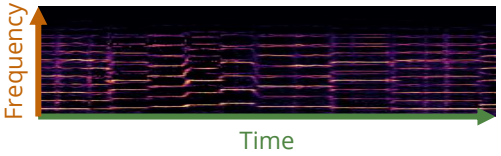
Text-based

Image-based

Time series-based

Image-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```



MIDI

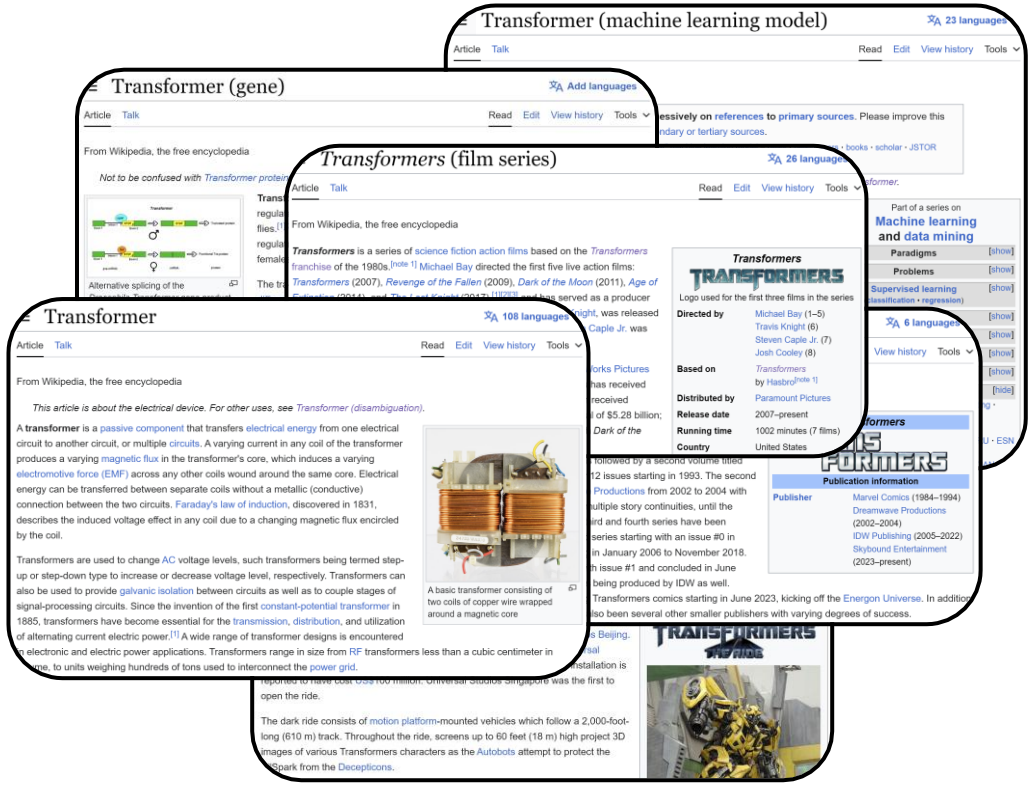
Piano roll

Waveform

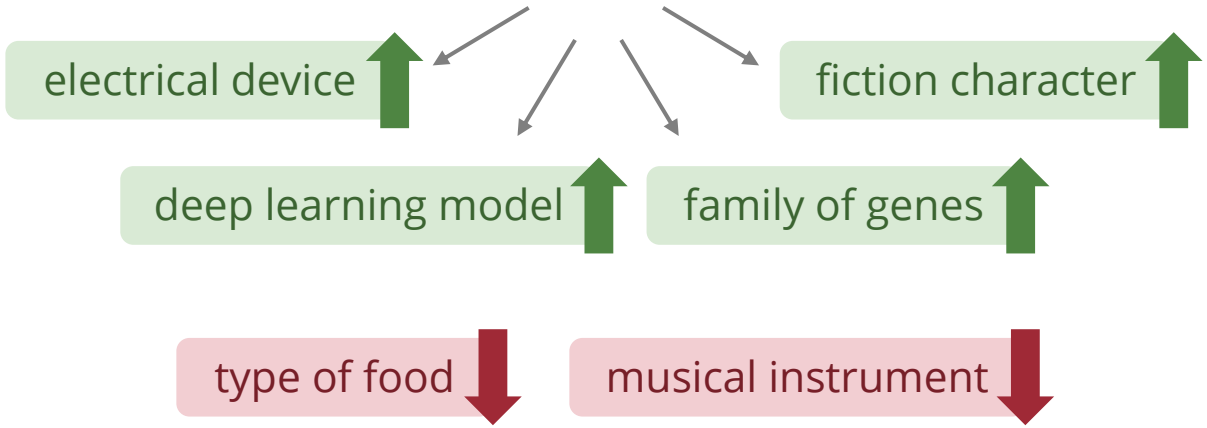
Spectrogram

Generating Text using Language Models

- Predicting the next word given the past sequence of words

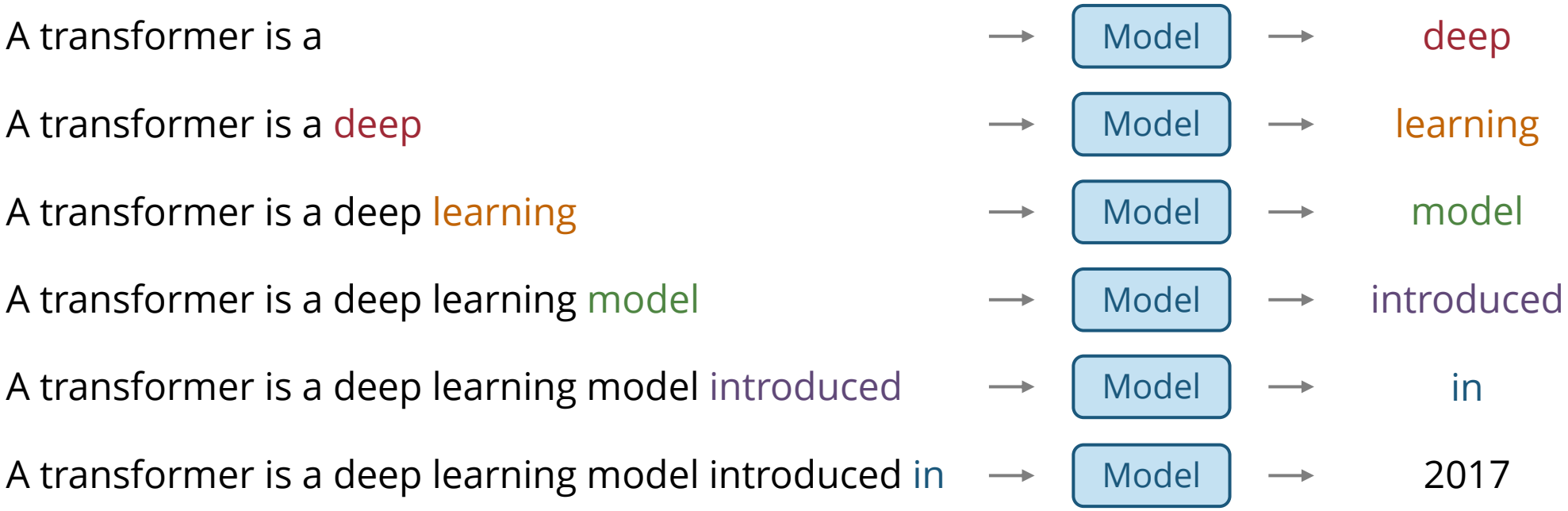


A transformer is a _____



Generating Text using Language Models

- How do we generate a new sentence with a language model?



Multitrack Music Transformer

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

University of California San Diego





UC San Diego

Designing a Machine-readable Music Language

- We represent a music piece as a sequence of "**super words**"

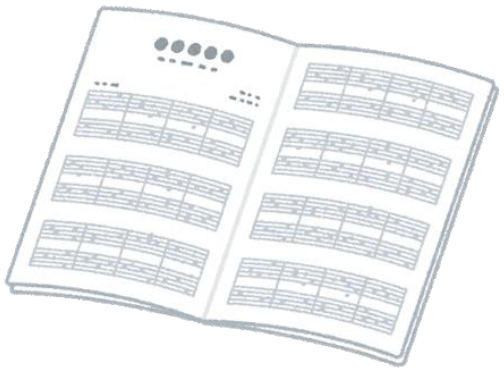
$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

- Each super word $\mathbf{x}_i$ encodes:

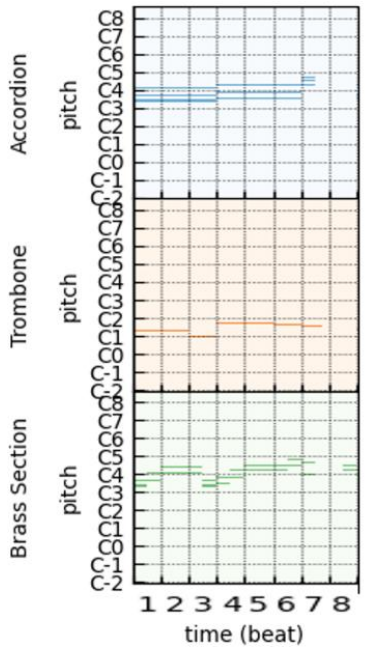
$$\mathbf{x}_i = (x_i^{type}, x_i^{beat}, x_i^{position}, x_i^{pitch}, x_i^{duration}, x_i^{instrument})$$

Specify note & instrument information

- Structural
  - Start of song
  - Start of notes
  - End of song
- Data
  - Instrument
  - Note



An Example of the Proposed Representation



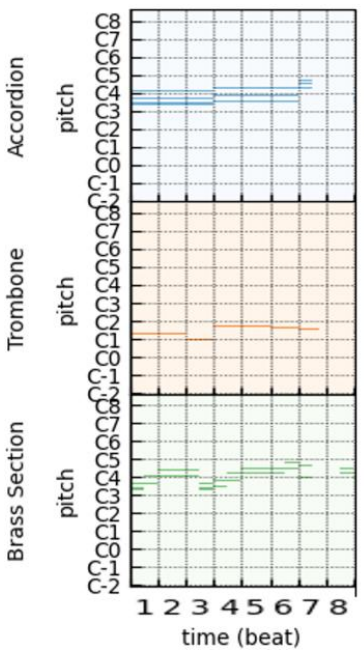
Structural events

```
(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
```

Instrument events

Note events

An Example of the Proposed Representation



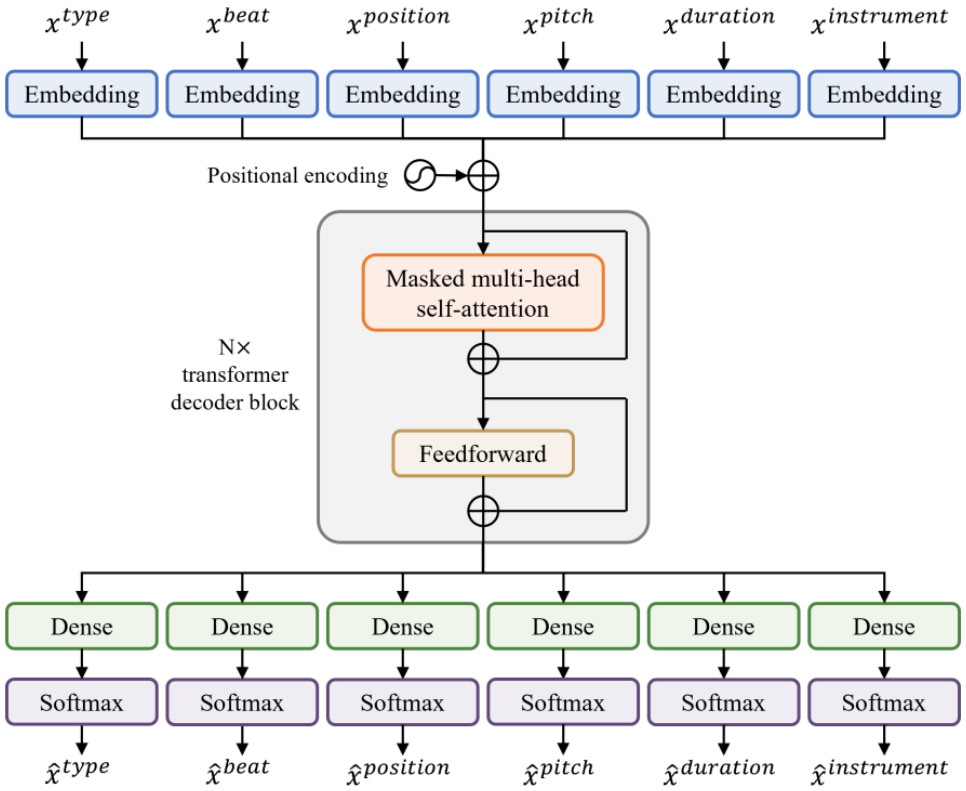

```
(0,  0,  0,  0,  0,  0)
(1,  0,  0,  0,  0, 15)
(1,  0,  0,  0,  0, 36)
(1,  0,  0,  0,  0, 39)
(2,  0,  0,  0,  0,  0)
(3,  1,  1, 41, 15, 36)
(3,  1,  1, 65,  4, 39)
(3,  1,  1, 65, 17, 15)
(3,  1,  1, 68,  4, 39)
(3,  1,  1, 68, 17, 15)
(3,  1,  1, 73, 17, 15)
(3,  1, 13, 68,  4, 39)
(3,  1, 13, 73,  4, 39)
(3,  2,  1, 73, 12, 39)
(3,  2,  1, 77, 12, 39)
            ...
(4,  0,  0,  0,  0,  0)
```

```
Start of song
Instrument: accordion
Instrument: trombone
Instrument: brasses
Start of notes
Note: beat=1, position=1,  pitch=E2, duration=48, instrument=trombone
Note: beat=1, position=1,  pitch=E4, duration=12, instrument=brasses
Note: beat=1, position=1,  pitch=E4, duration=72, instrument=accordion
Note: beat=1, position=1,  pitch=G4, duration=12, instrument=brasses
Note: beat=1, position=1,  pitch=G4, duration=72, instrument=accordion
Note: beat=1, position=1,  pitch=C5, duration=72, instrument=accordion
Note: beat=1, position=13, pitch=G4, duration=12, instrument=brasses
Note: beat=1, position=13, pitch=C5, duration=12, instrument=brasses
Note: beat=2, position=1,  pitch=C5, duration=36, instrument=brasses
Note: beat=2, position=1,  pitch=E5, duration=36, instrument=brasses
...
End of song
```

Multitrack Music Transformer (MMT)

- A decoder-only transformer model
- Predicts six fields at the same time
- Trained autoregressively

Note-by-note



Symbolic Orchestral Database (SOD)

- 5,743 orchestral pieces (**357 hours** in total)
- Contains various ensembles: choir, string quartet, symphony, etc.

Example Results

Unconditional generation

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
```

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
```

Only needs to train ONE model!

Example Results

Unconditional generation

Instrument-informed generation

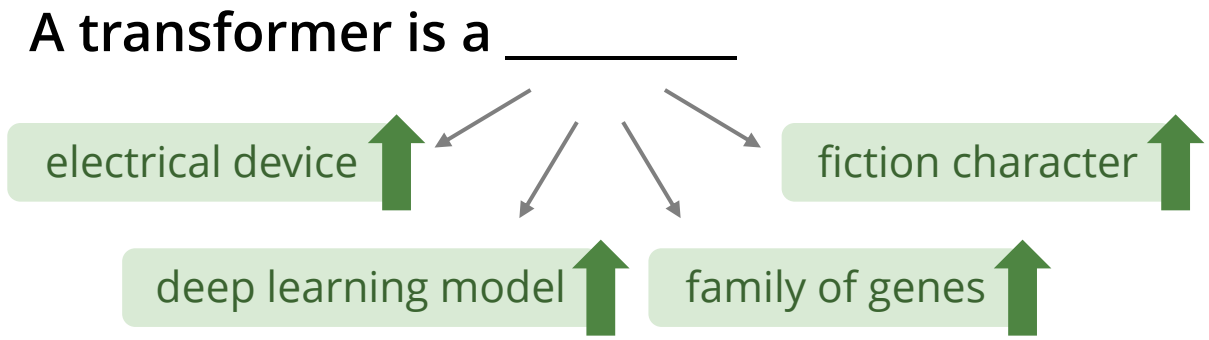
church-organ, viola, contrabass, strings, voices, horn, oboe

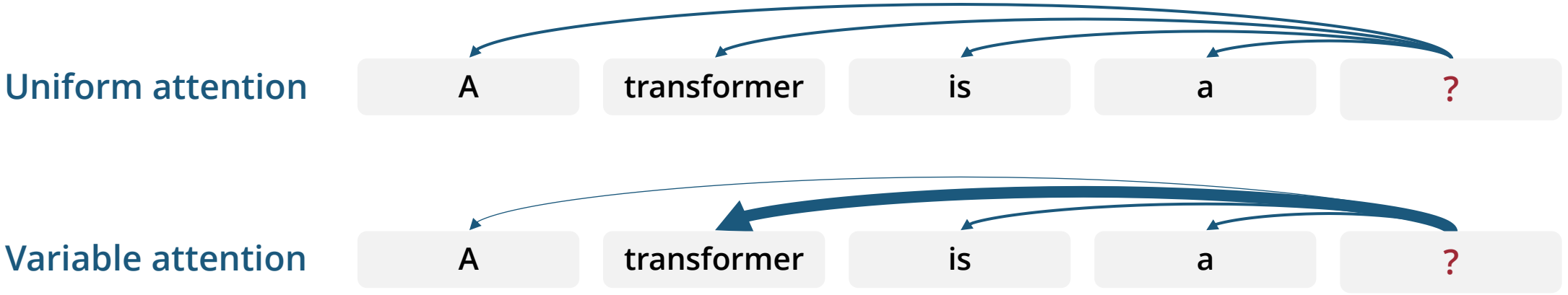
4-beat continuation

Mozart's
Eine kleine Nachtmusik

OttoPilot33, "Eine Kleine Nachtmusik - By My MIDI Virtual Orchestra." *YouTube*, 2019.

The Magic of Transformers – Self-attention Mechanism

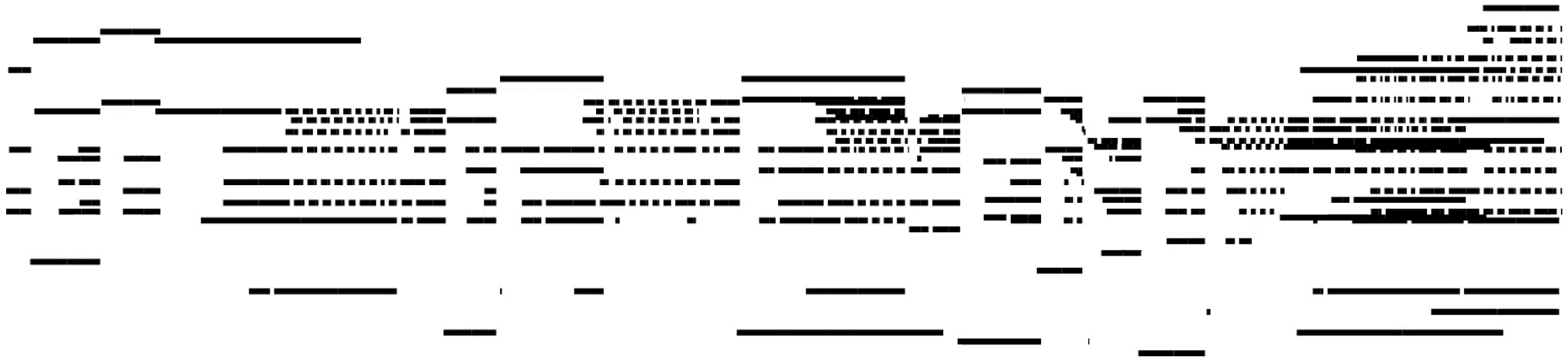




Transformers learn what to attend to from big data!

Visualizing Musical Self-attention (Huang et al., 2018)

(Each color represents an attention head)

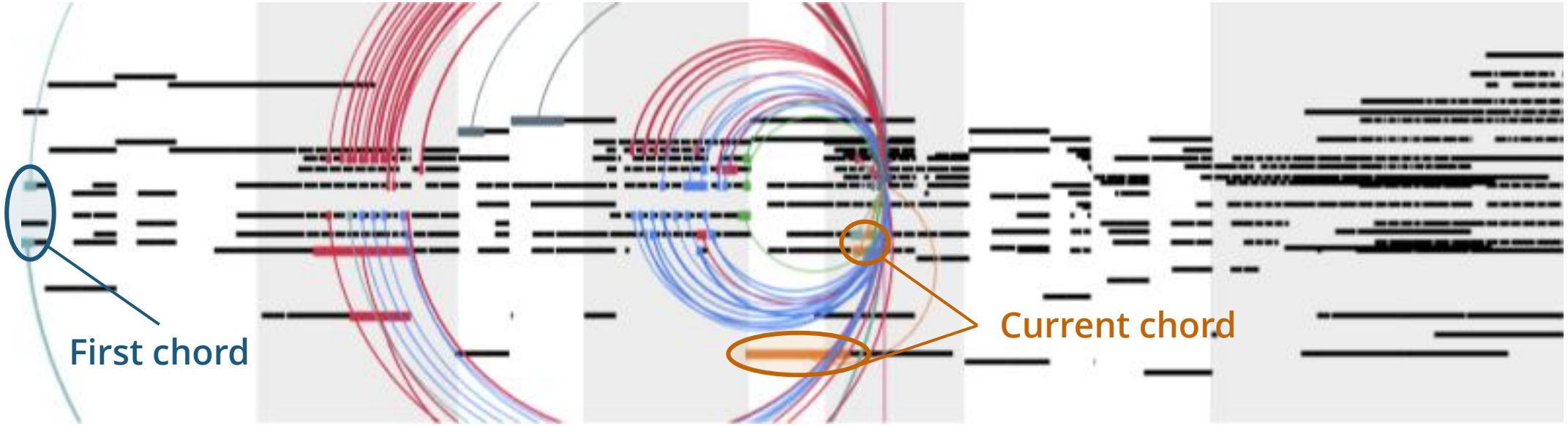


(Source: Huang et al., 2018)

Cheng-Zhi Anna Huang, Ian Simon, and Monica Dinculescu, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Visualizing Musical Self-attention (Huang et al., 2018)

(Each color represents an attention head)



(Source: Huang et al., 2018)

Can we go beyond case studies?

Cheng-Zhi Anna Huang, Ian Simon, and Monica Dinculescu, "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Systematically Analyzing Musical Self-attention

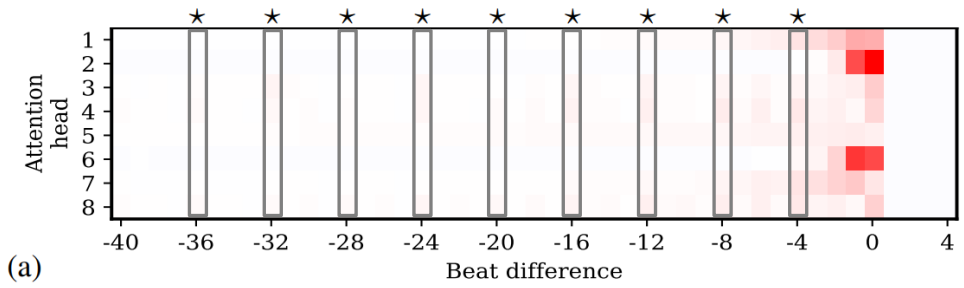
- We proposed two new quantities for measuring **mean relative attention**

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

- The MMT model attends more to notes

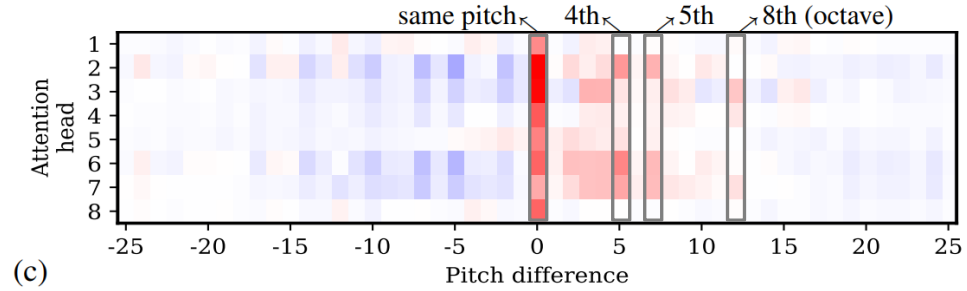
that are $4N$ beats away in the past

Positive and negative mean relative attention gain



that has a pitch in an octave above which forms a consonant interval

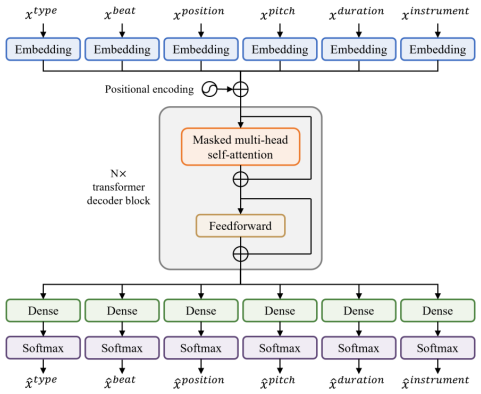
Positive and negative mean relative attention gain



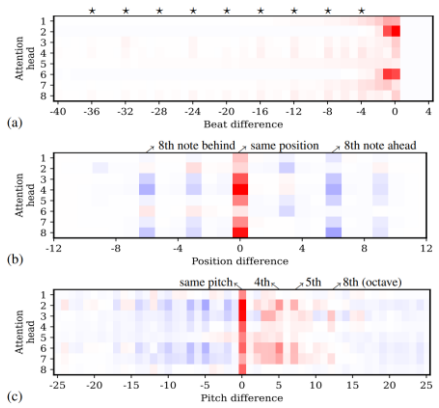
Summary

- **State-of-the-art orchestral music generation model**
- Presented the **first systematic analysis** of **musical self-attention**
- Showed that MMT learns a **relative self-attention for beat and pitch**

Multitrack Music Transformer



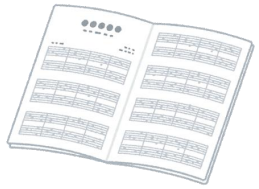
Musical Self-attention



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

UC San Diego

Music Generation – Four Paradigms



Symbolic music generation

Audio-domain music generation

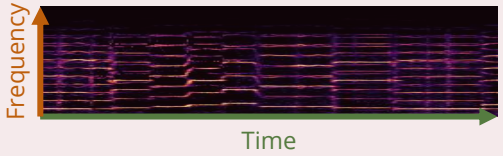
Text-based

Image-based

Time series-based

Image-based

```
Program_change_0,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_60, Time_shift_2, Note_off_60,
Note_on_76, Time_shift_2, Note_off_67,
Note_on_67, Time_shift_2, Note_off_67,
...
```



MIDI

Piano roll

Waveform

Spectrogram

Types of Audio

Speech

(Source: Wikimedia Commons)

Music

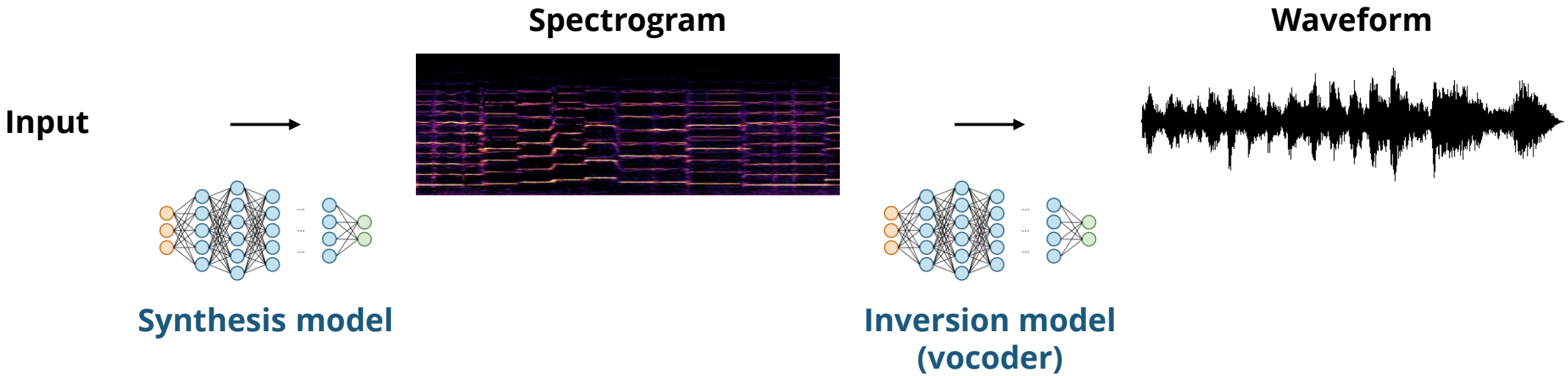
(Source: Wikimedia Commons)

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.

Frequency-domain Audio Synthesis

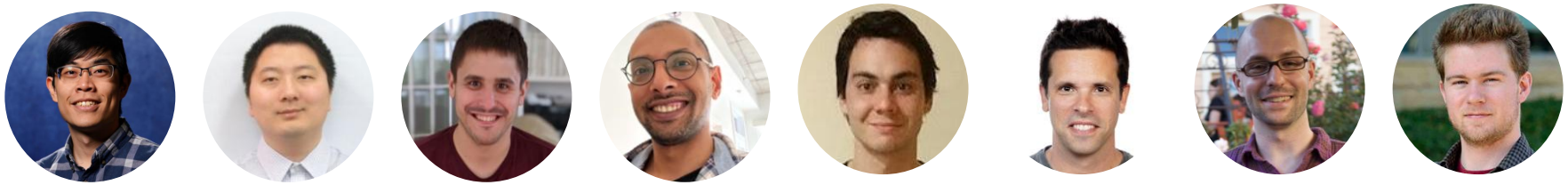


IEEE
WASPAA
2023

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



Dolby UC San Diego

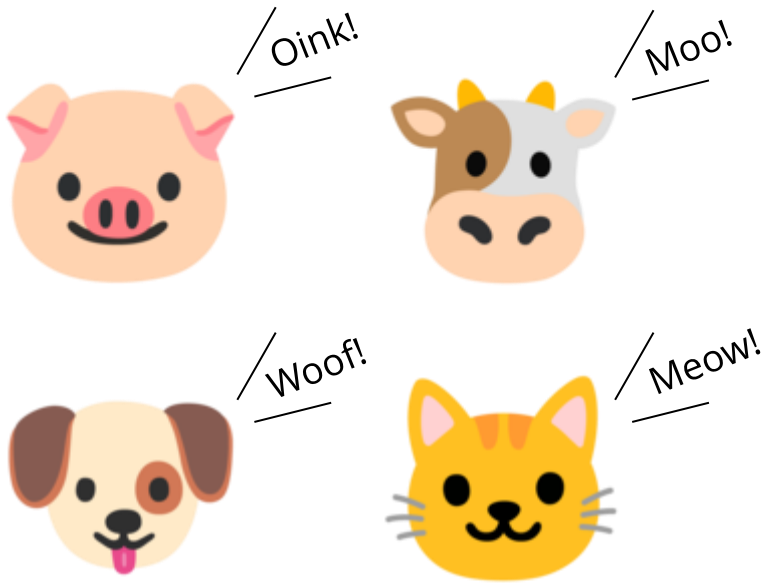
What is Text-to-Audio Synthesis?

- **Goal**: Given a text query, generate the corresponding sounds

(These samples are generated by our proposed model.)

Learning Sounds from Observations

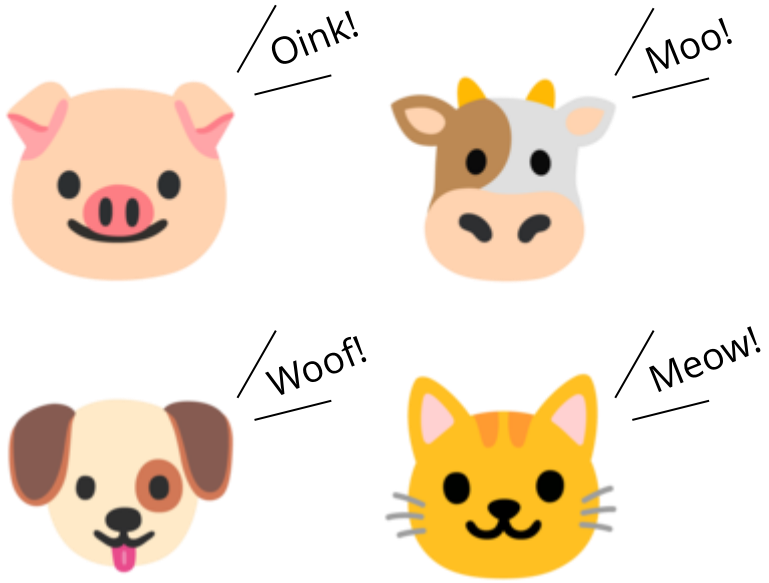
- Watching a dog barking, humans can *associate the barking sound to the dog*



What does the fox say?

Learning Sounds from Noisy Videos

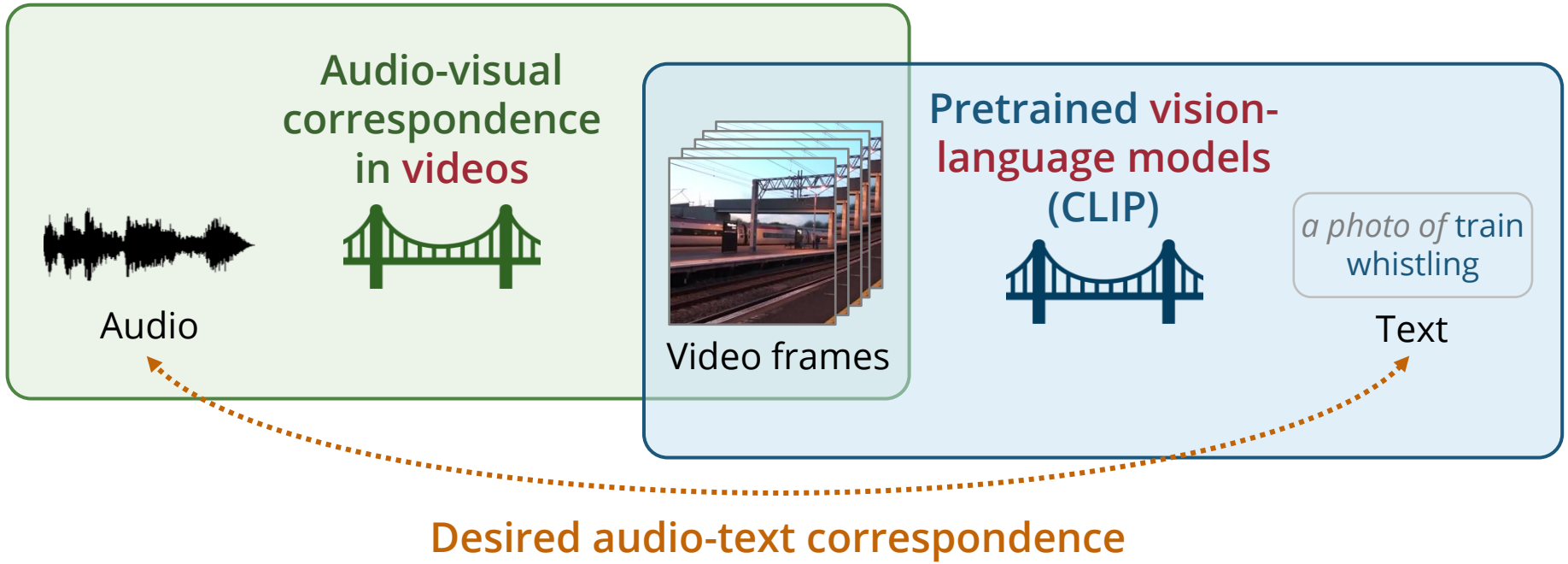
- Watching a dog barking, humans can *associate the barking sound to the dog*



Can machines learn to synthesize sounds from watching *noisy* videos?

URURU TV, "Excited Fox Noises - Fox Laughing and Squeaking," *YouTube*, 2019.

Leveraging the Visual Domain as a Bridge



No text-audio pairs required!

Scalable to large video datasets!

Data

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

(172K videos, 310 classes)

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

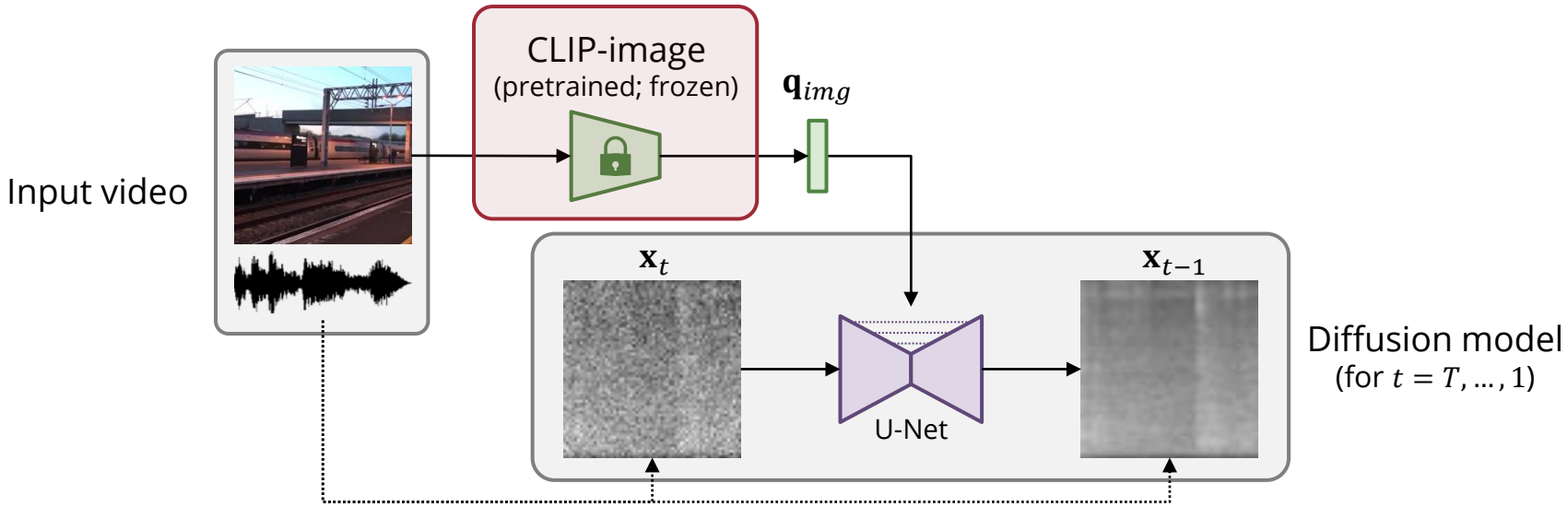
Music instrument playing videos

(1,055 videos, 21 instruments)

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The Sound of Pixels," *ECCV*, 2018.
Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A Large-Scale Audio-Visual Dataset," *ICASSP*, 2020.

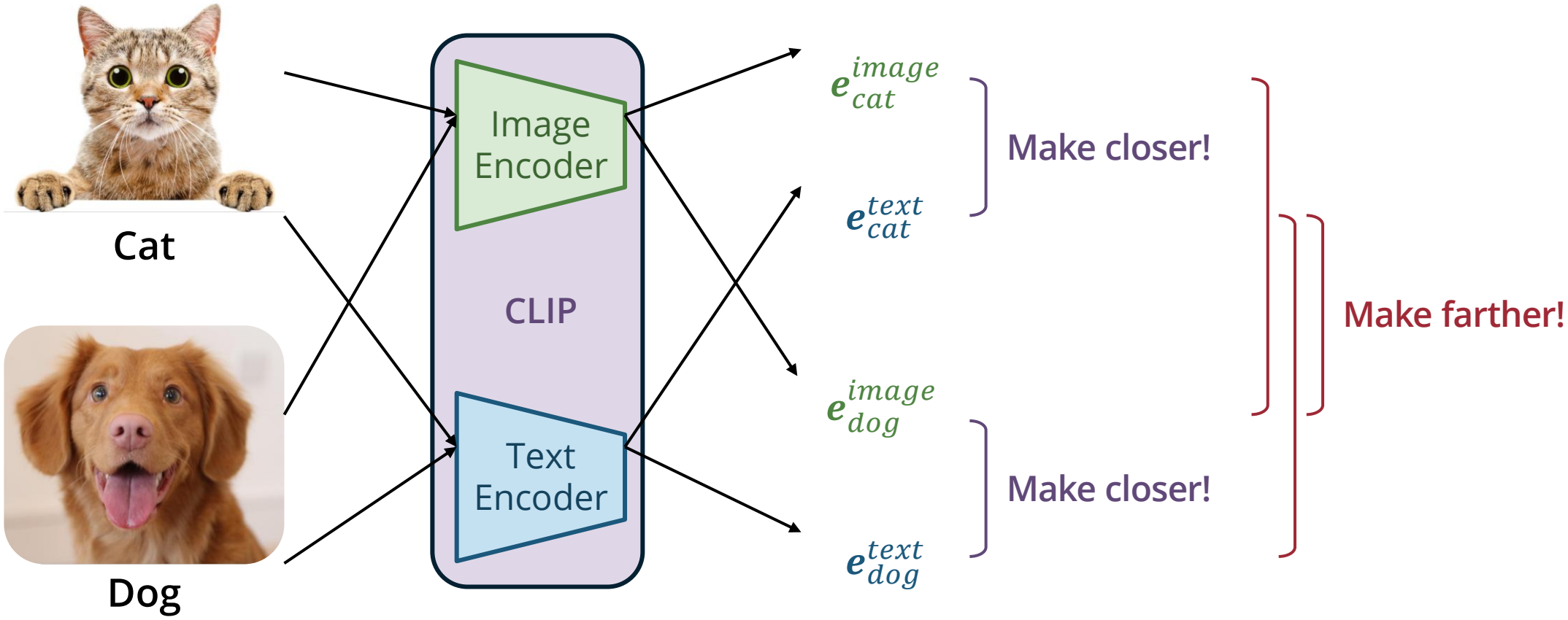
Training an Image-to-Audio Synthesis Model

- We start by training an image-to-audio synthesis model



CLIP (Contrastive Language-Image Pretraining)

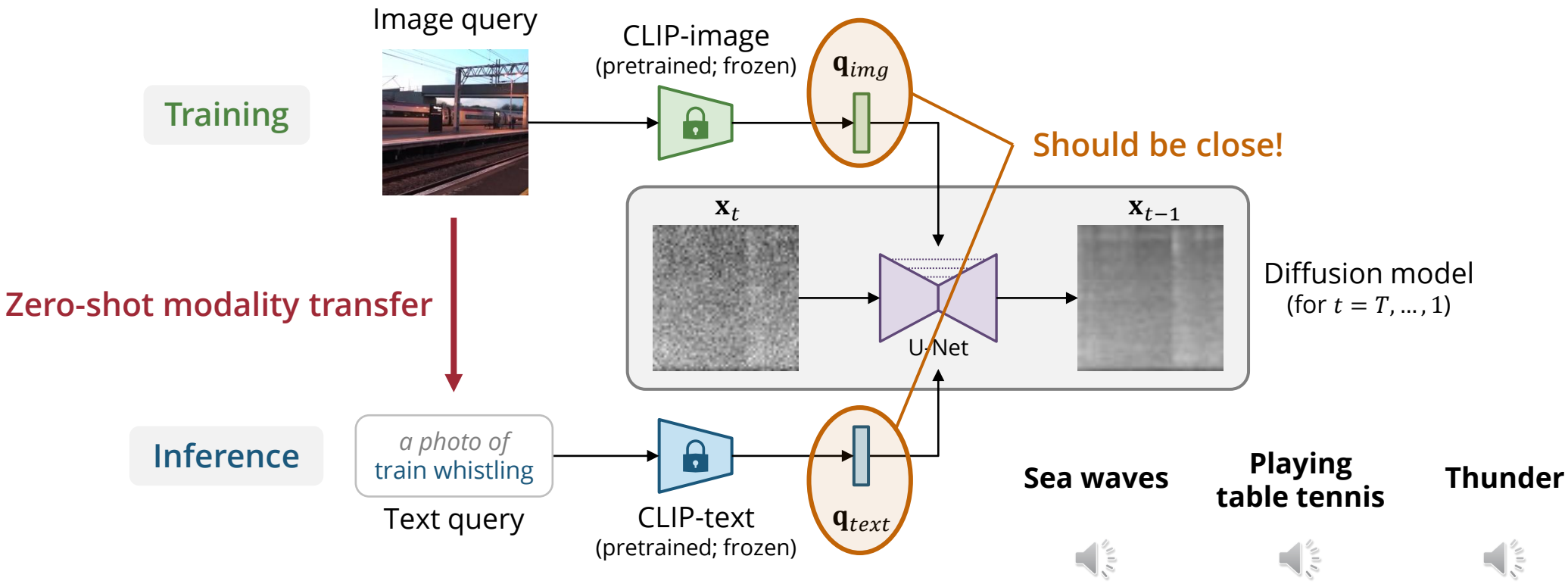
- Learn a shared embedding space for images and texts via *contrastive learning*



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

Inference – Zero-shot Modality Transfer

- We switch to a pretrained CLIP-text encoder for text-to-sound synthesis



Inference – Zero-shot Modality Transfer

- We switch to a pretrained CLIP-text encoder for text-to-sound synthesis

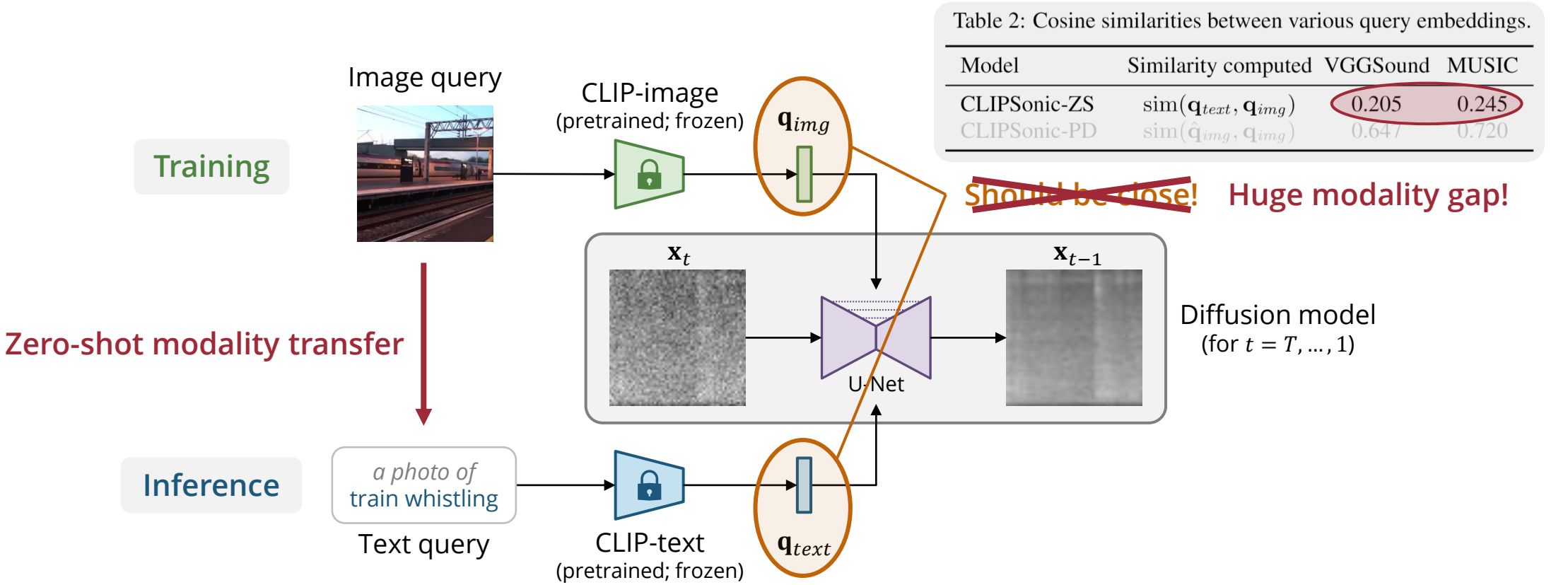
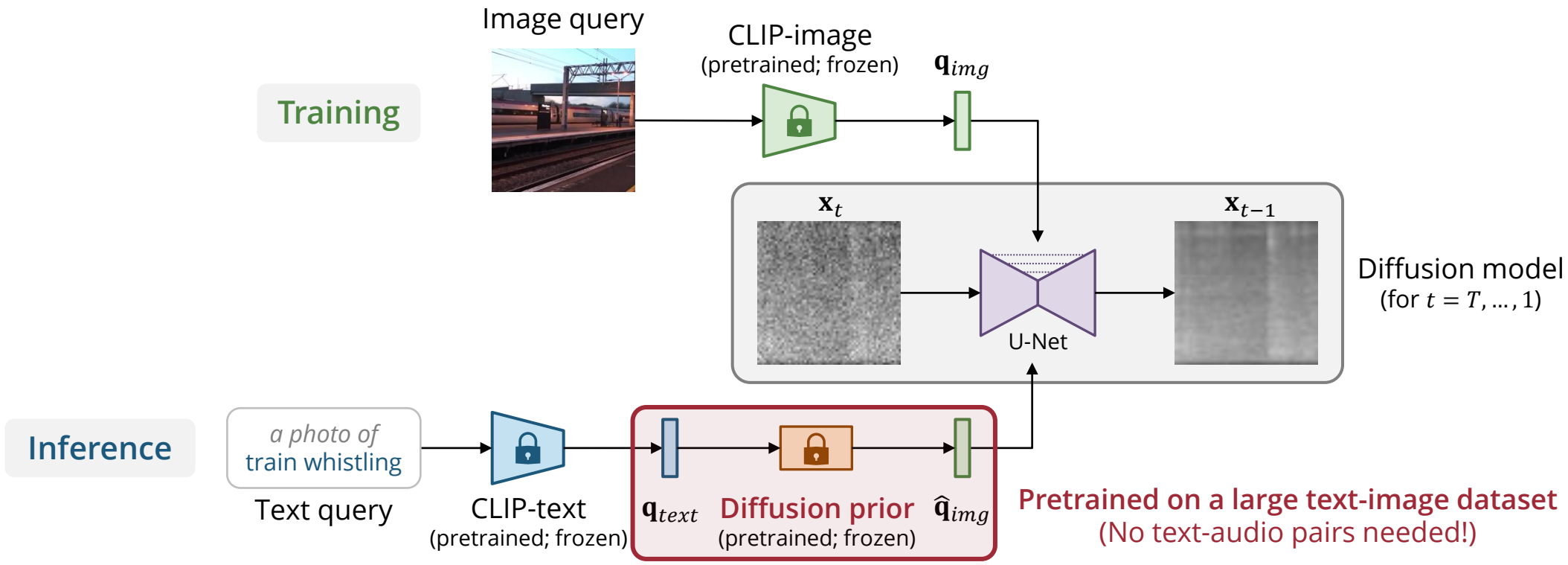


Table 2: Cosine similarities between various query embeddings.

Model	Similarity computed	VGGSound	MUSIC
CLIPSonic-ZS	$\text{sim}(\mathbf{q}_{text}, \mathbf{q}_{img})$	0.205	0.245
CLIPSonic-PD	$\text{sim}(\mathbf{q}_{img}, \mathbf{q}_{img})$	0.647	0.720

Leveraging Diffusion Prior to Close the Modality Gap

- We adopt a pretrained diffusion prior model to reduce the modality gap



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

Example Text-to-Audio Synthesis Results

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

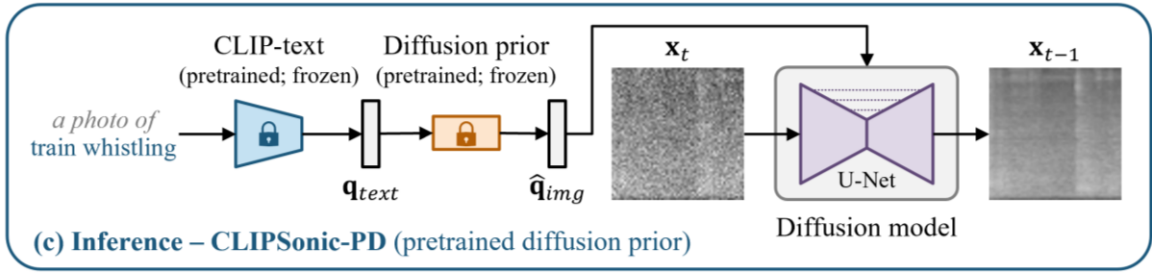
Playing violin fiddle

Example Image-to-Audio Synthesis Results (Out-of-distribution)

(Then!) State-of-the-art image-to-audio synthesis performance!

Contributions

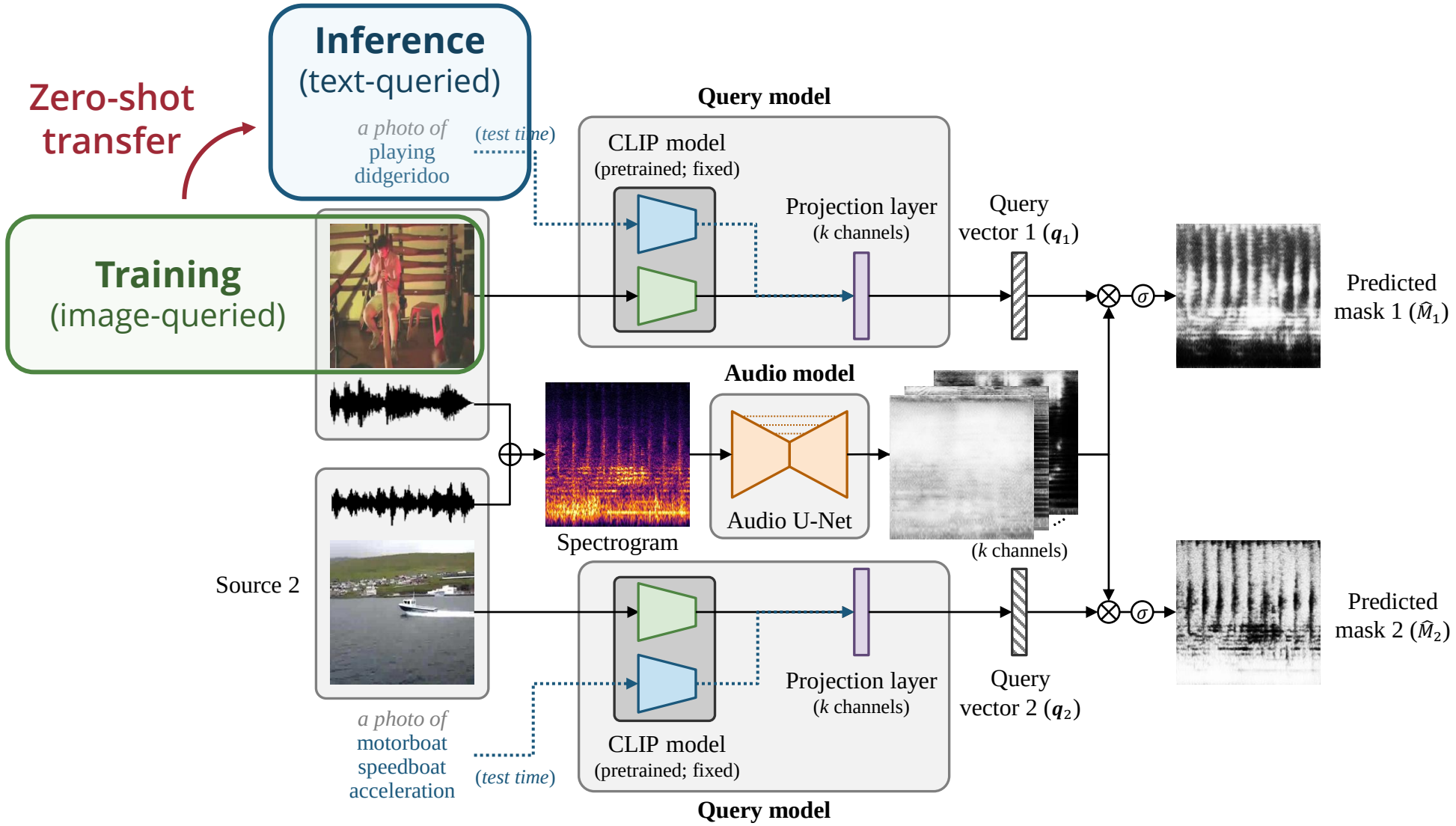
- First text-to-audio synthesis model that **requires *no* text-audio pairs**
- **Strong text-to-audio synthesis performance** without text-audio data
- (Then!) **State-of-the-art image-to-audio synthesis performance**



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

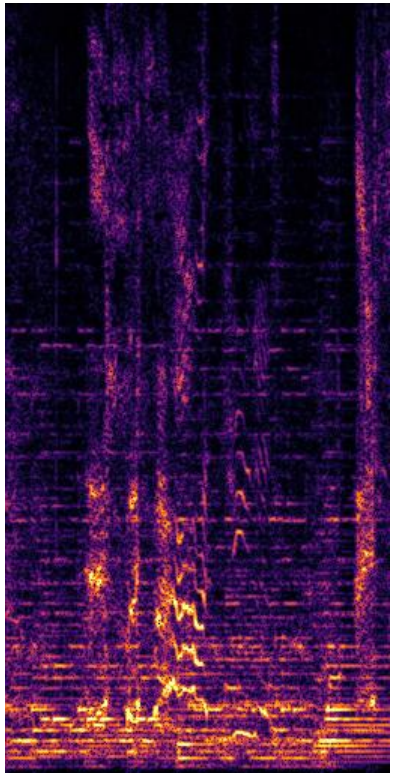
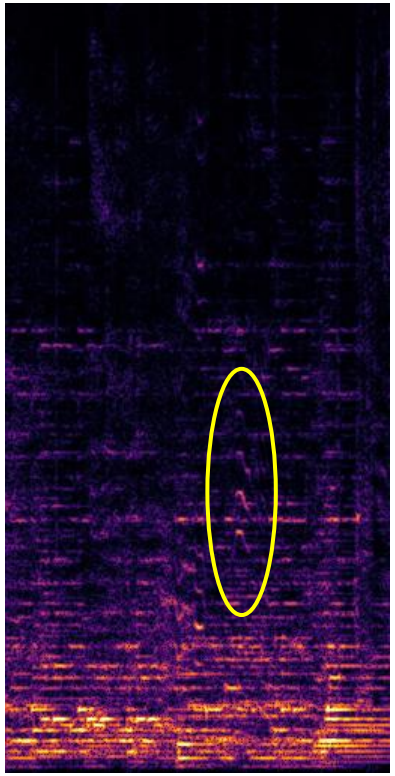
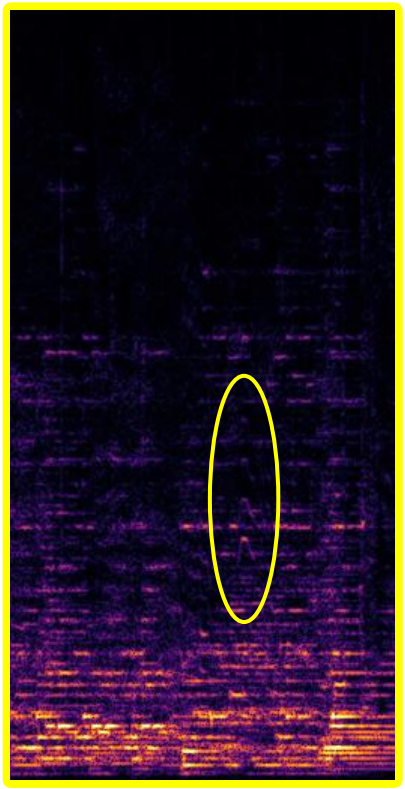
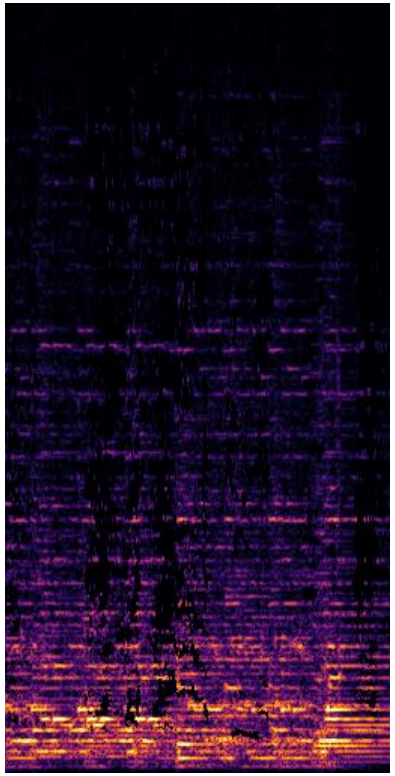
Dolby UC San Diego

CLIPSep: Text-queried Sound Separation (ICLR 2023)



CLIPSep: Text-queried Sound Separation (**ICLR 2023**)

Query: "*playing harpsichord*"

Mixture	**CLIPSep**	**CLIPSep-NIT**	**Ground truth**
			

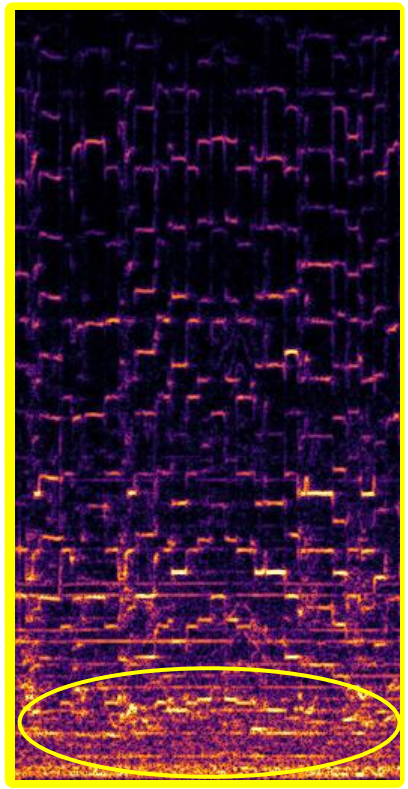
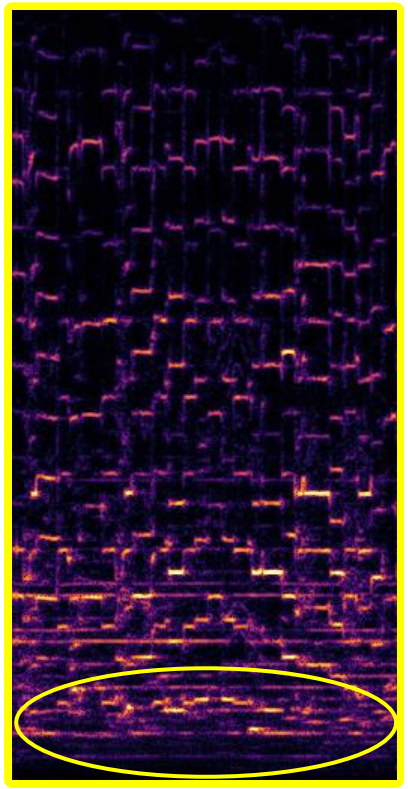
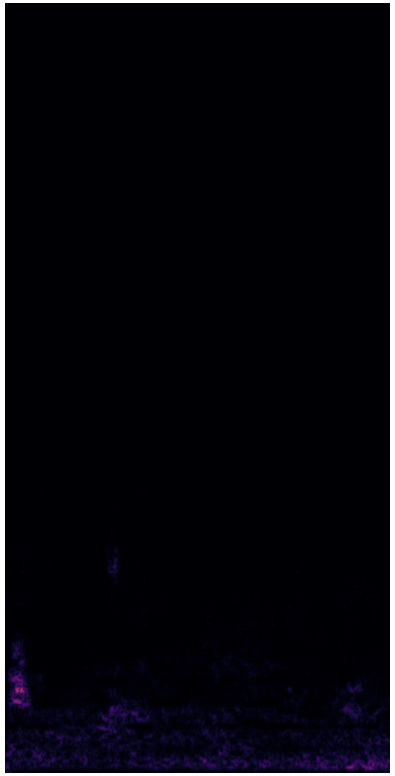
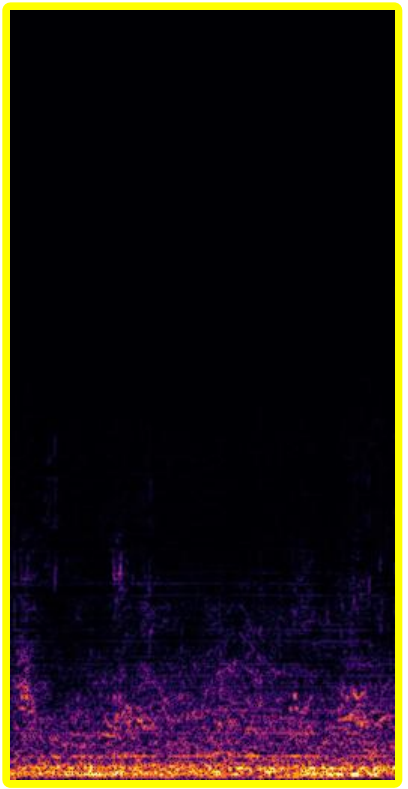
CLIPSep: Noise Removal (**ICLR 2023**)





SONY UC San Diego

Query: "*playing bagpipe*"

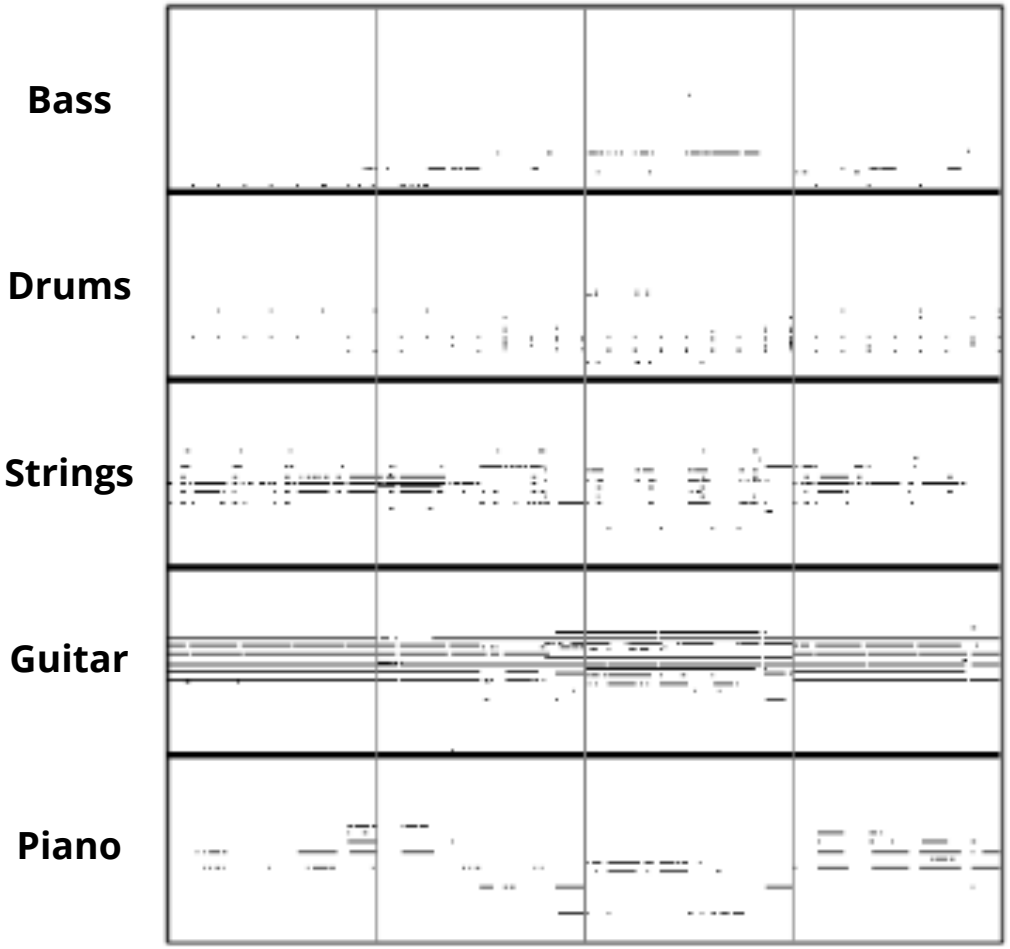
Mixture	Prediction	Noise head 1	Noise head 2
			

Augmenting Human Creativity with AI

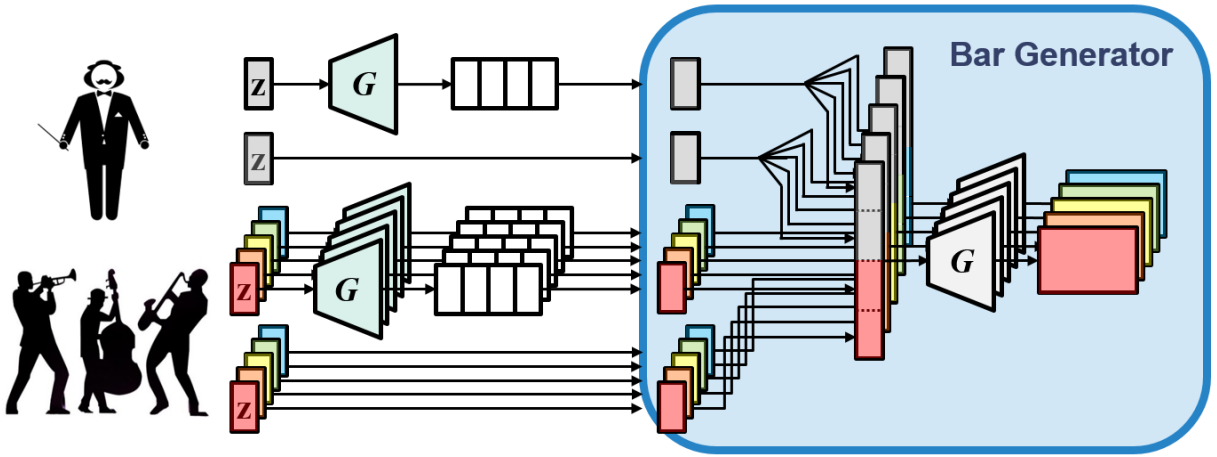
- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024, AIMG 2024), **text-to-symbolic music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)
- **AI-Assisted Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **Queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WSS 2023, WASPAA 2023), **text-to-music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)

Generating Multi-instrument Music using GANs (**AAAI 2018**)

Multitrack Piano Roll



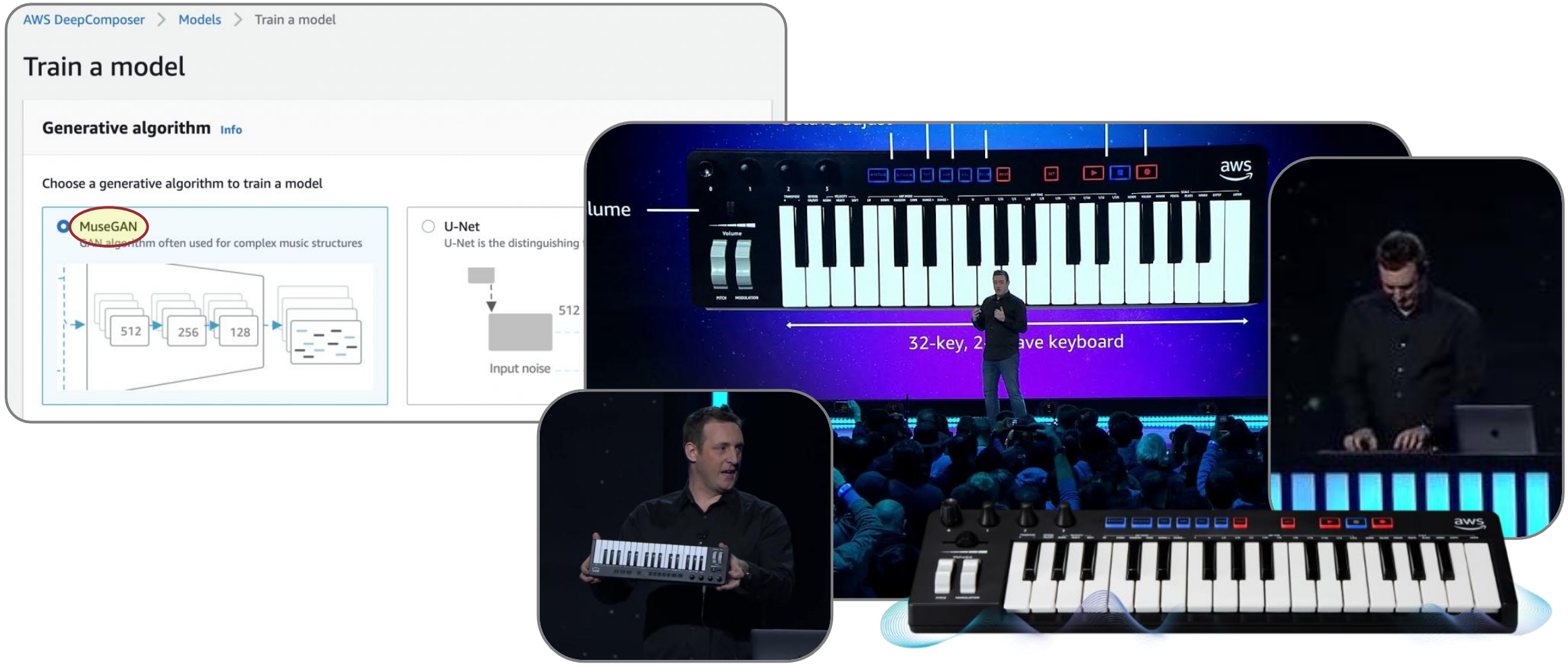
MuseGAN Generator



ACADEMIA SINICA

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI*, 2018.

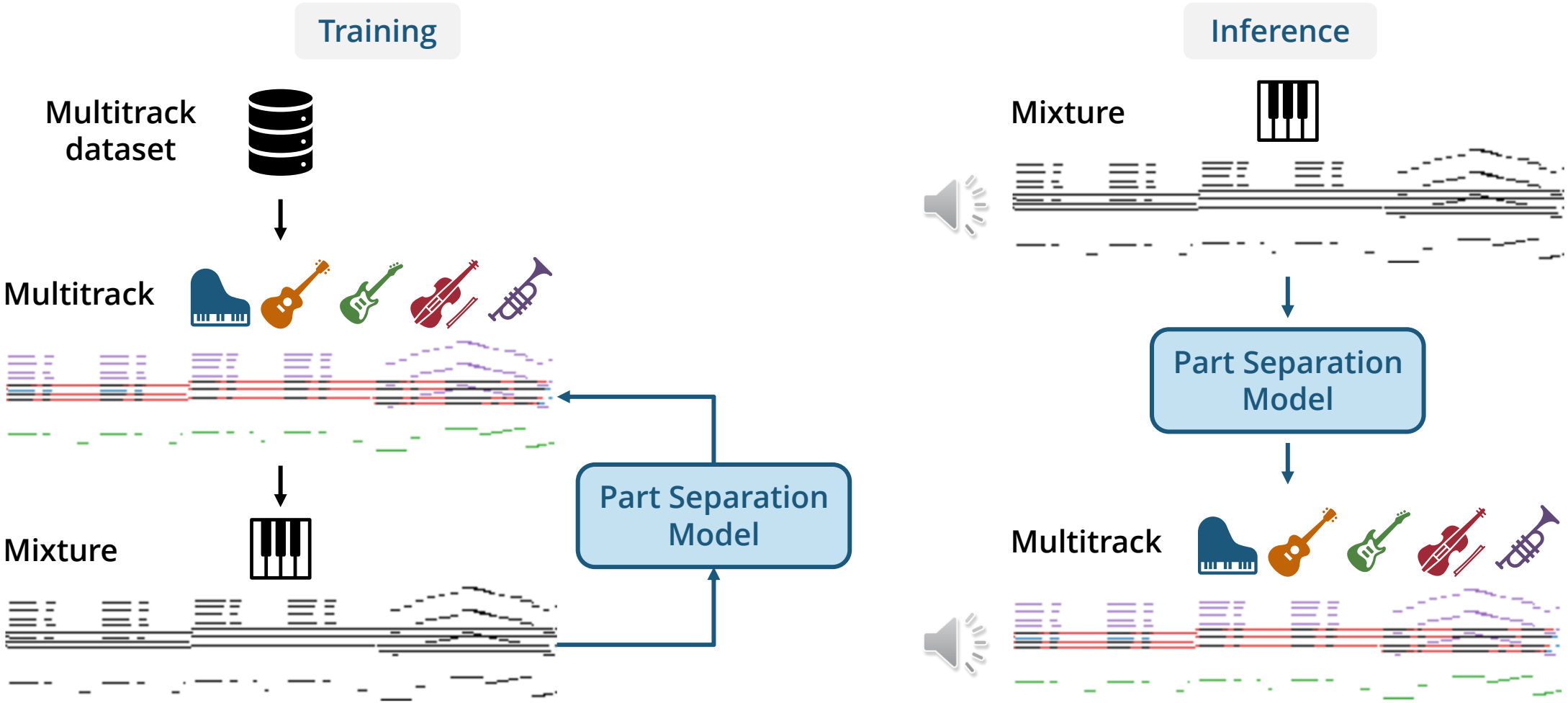
MuseGAN Features in AWS DeepComposer (2020)



amazon.com/dp/B07YGZ4V5B/
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," *AWS News Blog*, April 2, 2020.

Automatic Instrumentation (**ISMIR 2021**)

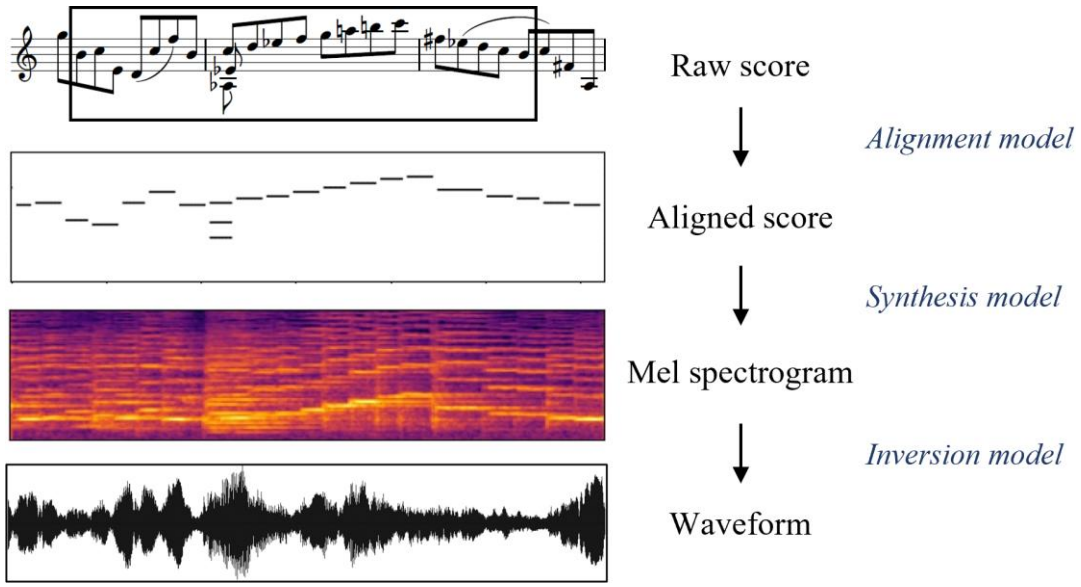
Stanford UC San Diego



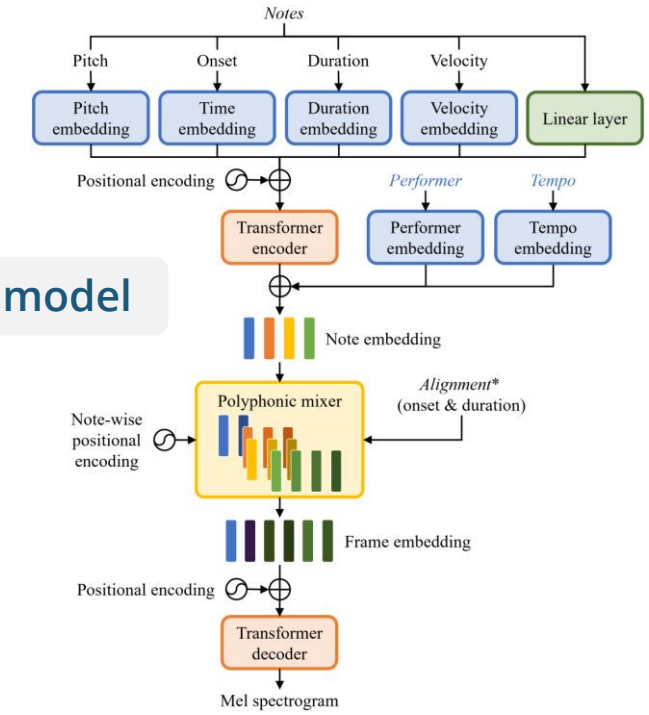
Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

Synthesizing Expressive Violin Performance (**ICASSP 2022**)

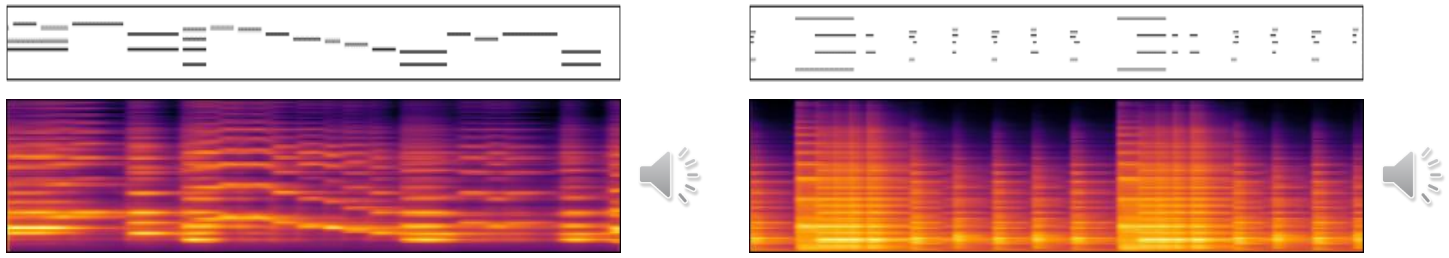
Performance synthesis



TTS-based model



Example results

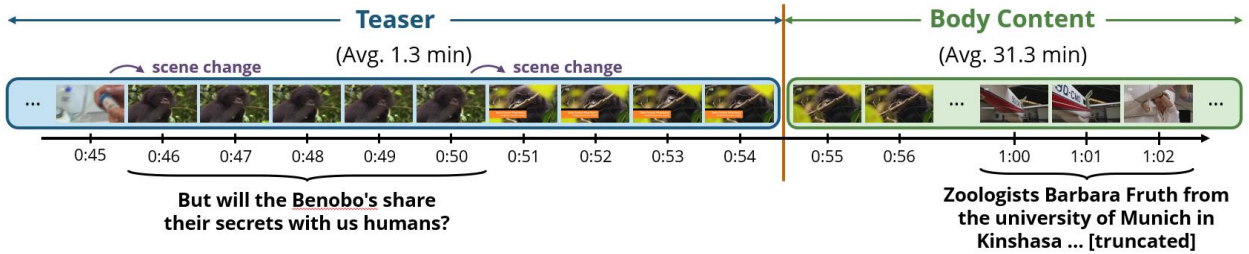


Dolby UC San Diego

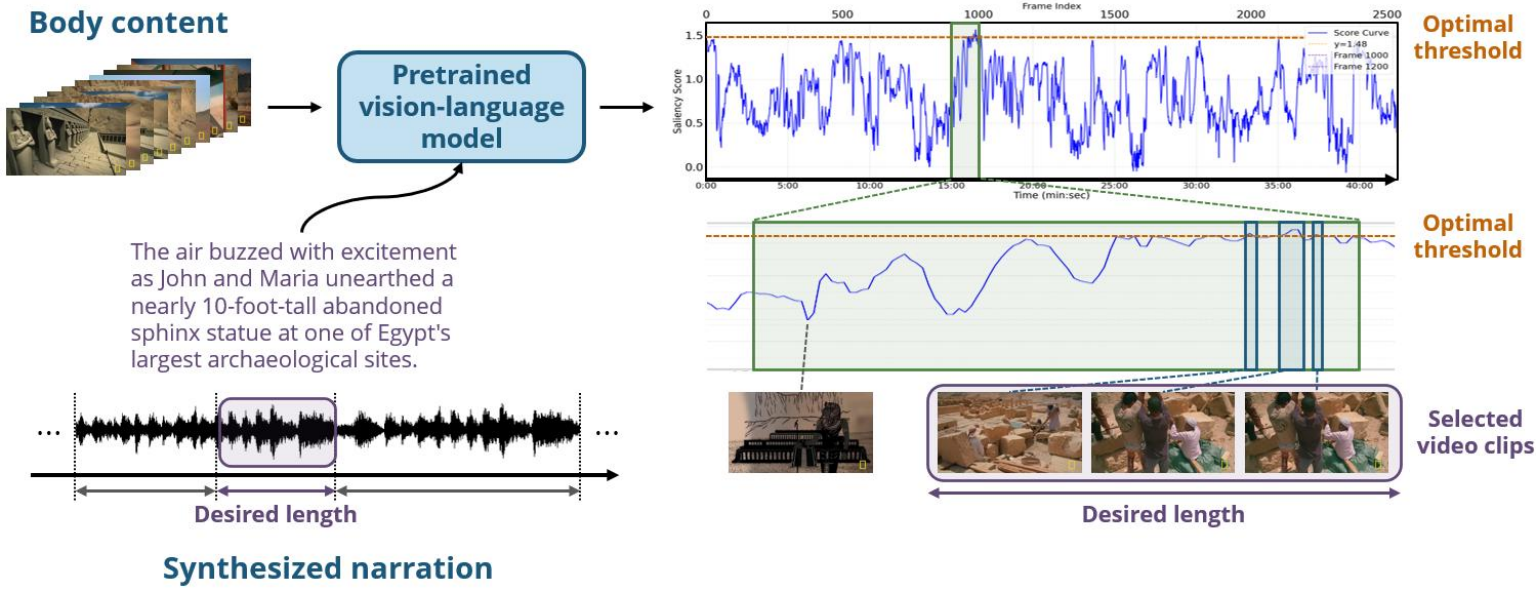
Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *ICASSP*, 2022.

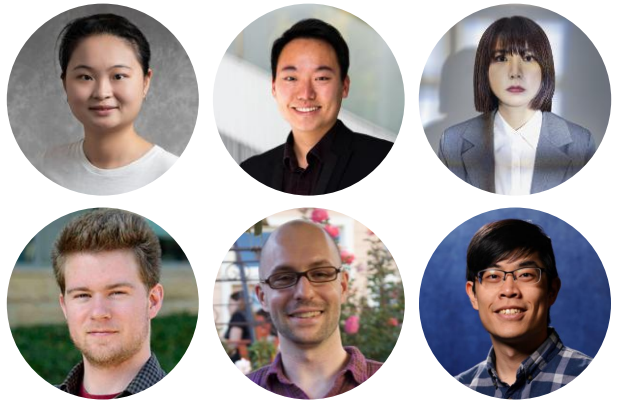
Generating Teasers for Documentaries (**ICLR 2025**)

Data



Narration-video matching model



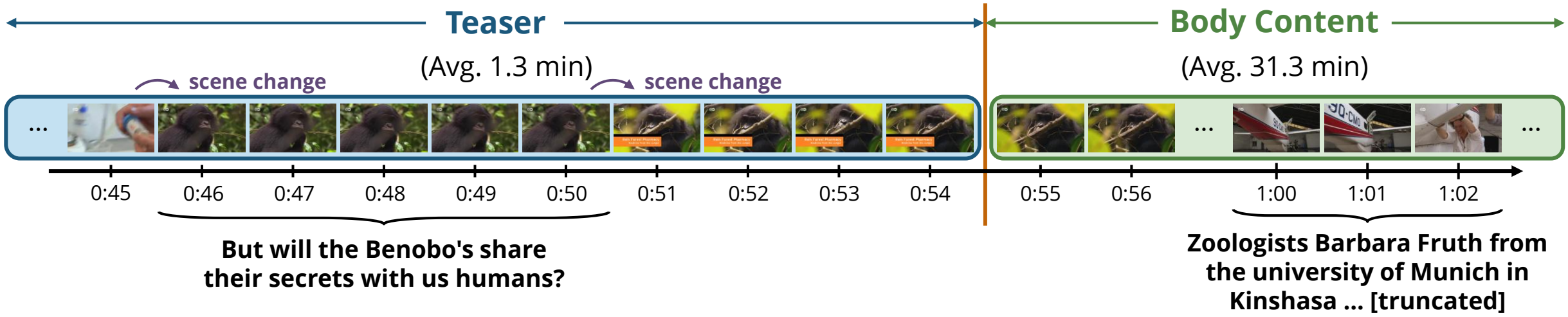


Duke UNIVERSITY MIT UC San Diego

UNIVERSITY OF MICHIGAN

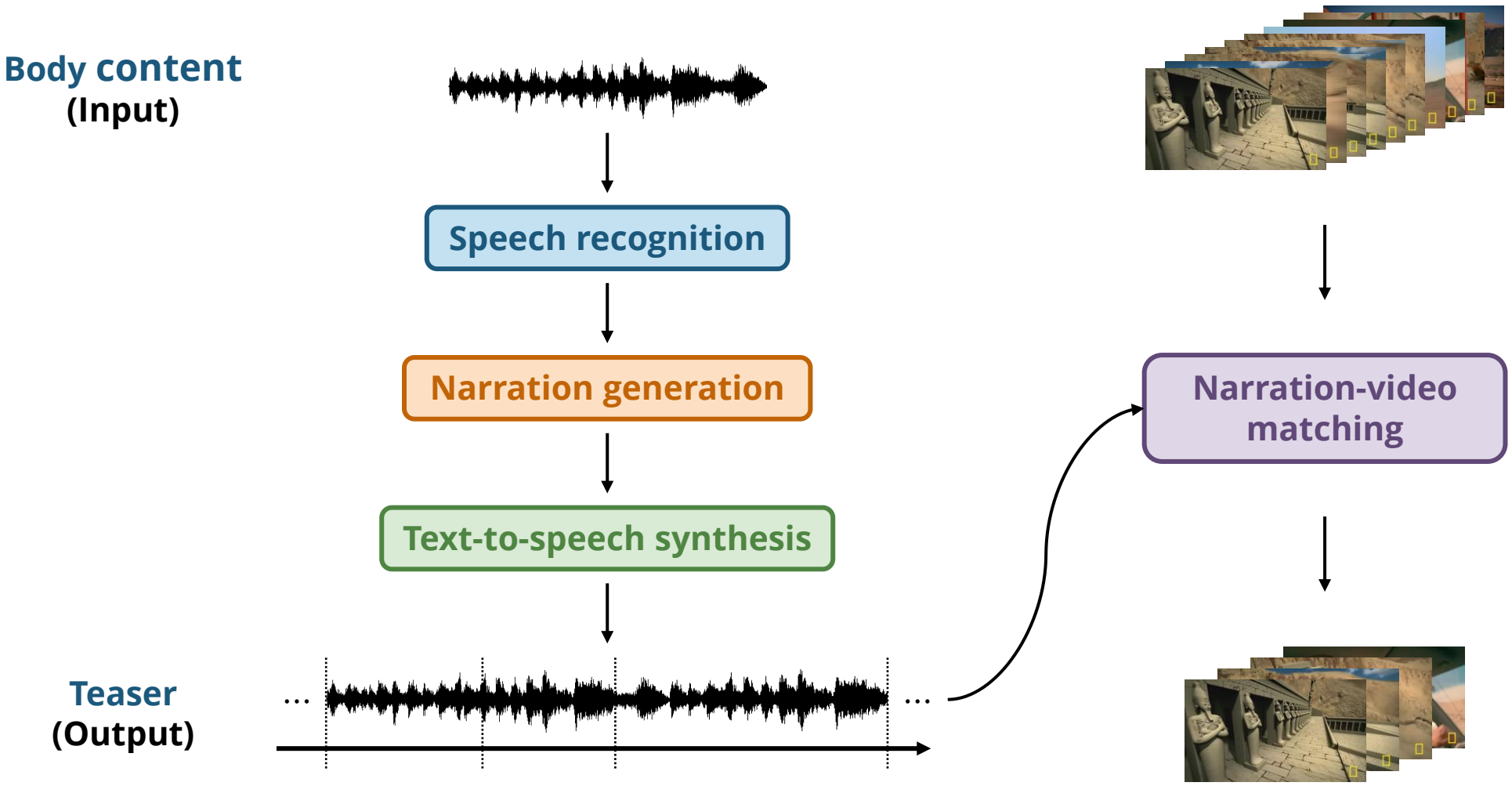
Weihan Xu, Paul Pu Liang, Haven Kim, Julian McAuley, Taylor Berg-Kirkpatrick, and Hao-Wen Dong, "TeaserGen: Generating Teasers for Long Documentaries," *ICLR*, 2025.

DocumentaryNet: A New Dataset for Documentaries



1,269 documentaries paired with teasers
(Source: DW, PBS, National Geographic; 400hr)

A Narration-Centered Approach to Teaser Generation



Augmenting Human Creativity with AI

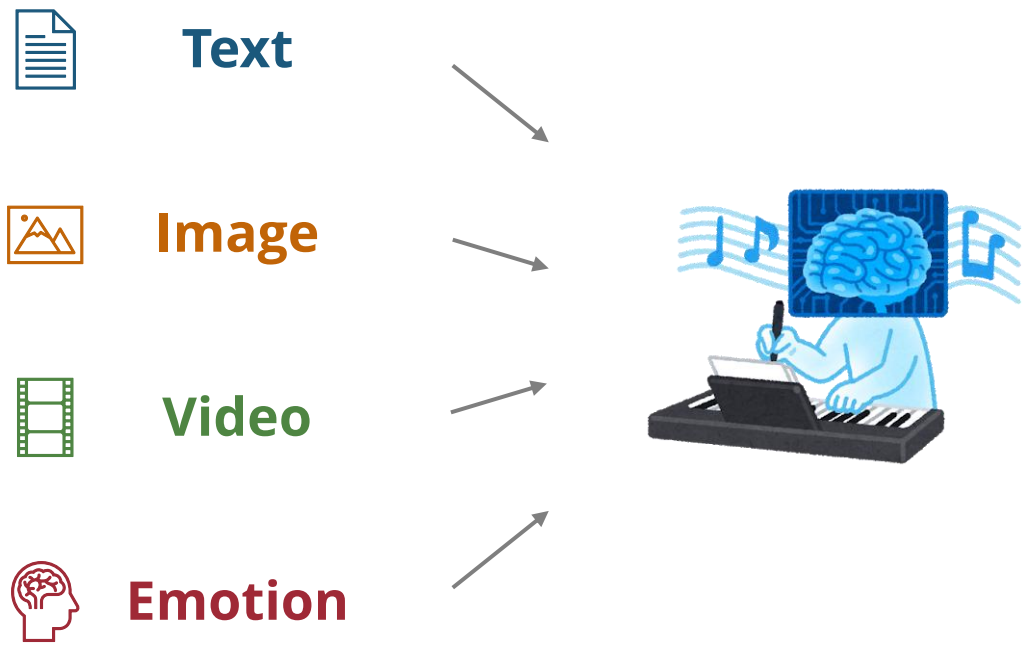
- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024, AIMG 2024), **text-to-symbolic music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)
- **AI-Assisted Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **Queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WSS 2023, WASPAA 2023), **text-to-music generation** (ISMIR LBD 2024, arXiv 2024), **documentary teaser generation** (ICLR 2025)

Future Directions

Augmenting Human Creativity with AI

- **Multimodal generative AI for content creation**
- **Human-AI co-creative tools for music, audio and video creation**
- **Human-like machine leaning algorithms for music, movies and arts**

Multimodal Inputs for Generative Music AI



WavJourney: Compositional Audio Creation (Liu et al., 2023)

Large language models
(GPT-4)

Pretrained generative audio models
(MusicGen, AudioLDM, Bark)

Instructions → **Audio Script** → **Audio**

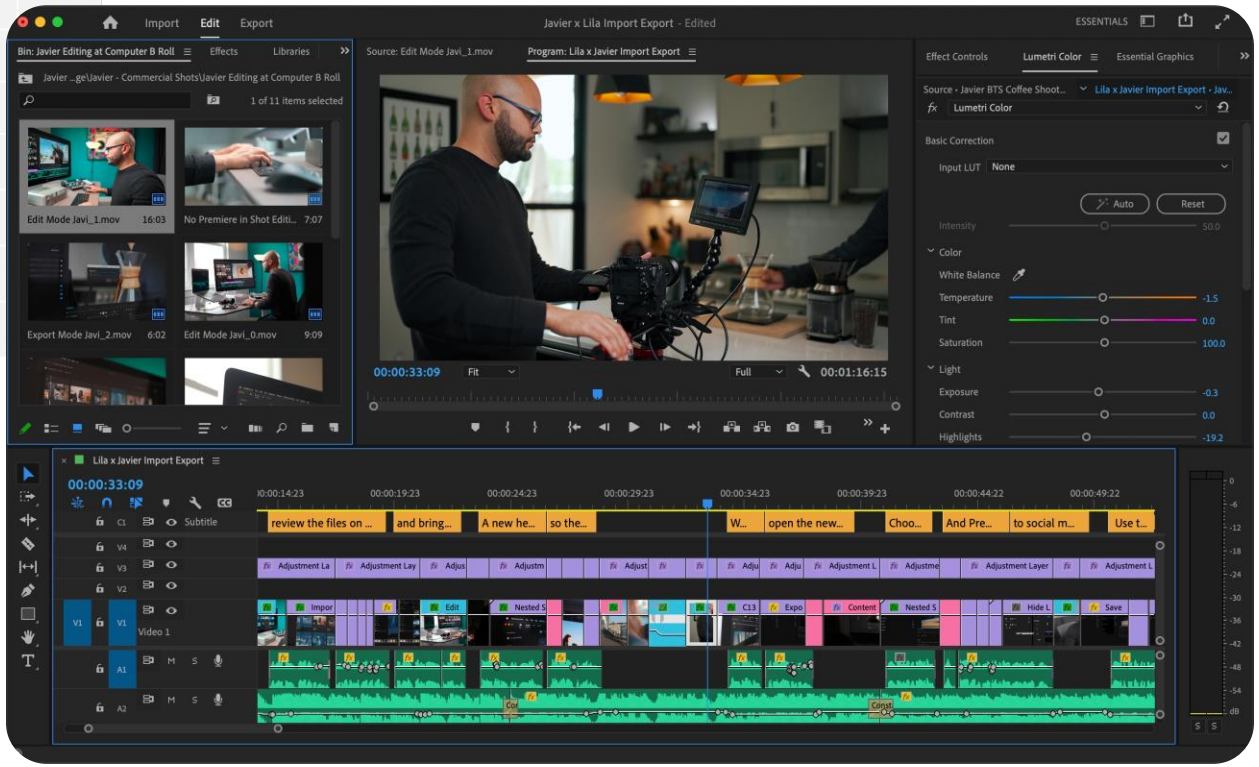
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect 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

Interactable intermediate outputs

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D. Plumbley, and Wenwu Wanga, "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

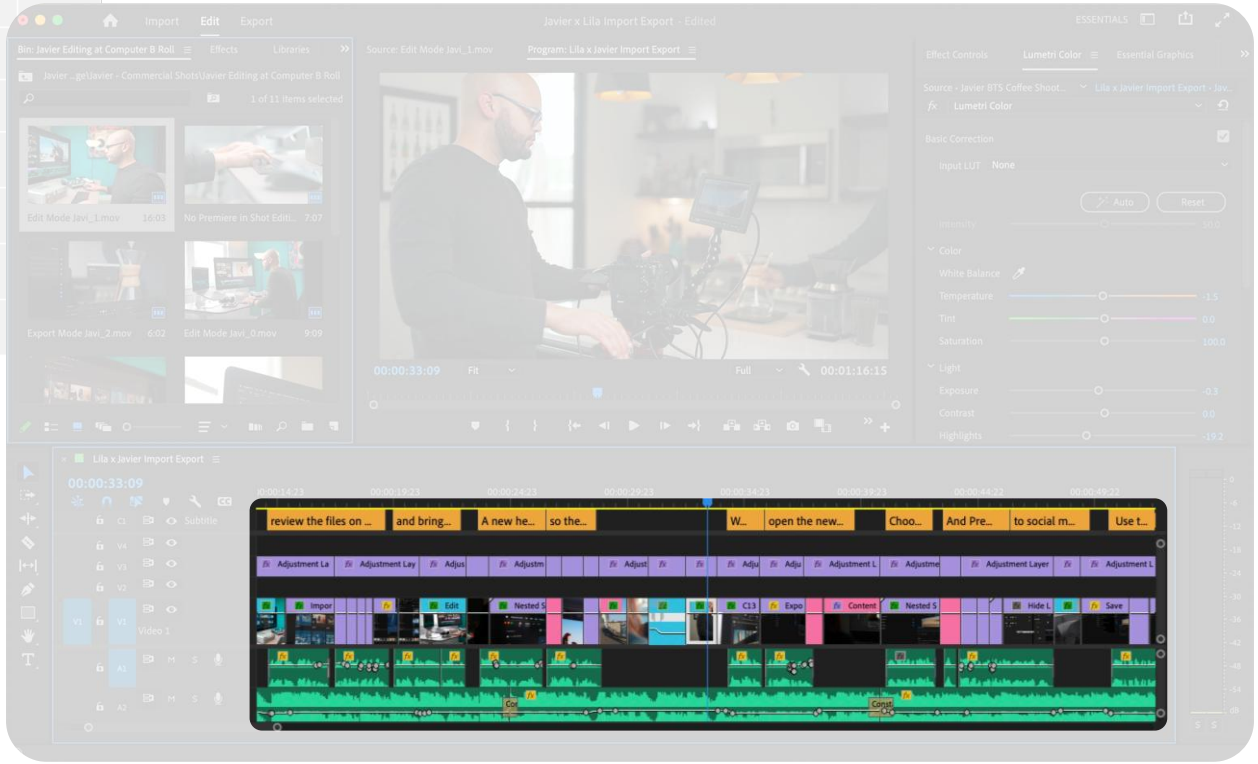
Integrating Generative AI into the Creative Workflow

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 …	



Integrating Generative AI into the Creative Workflow

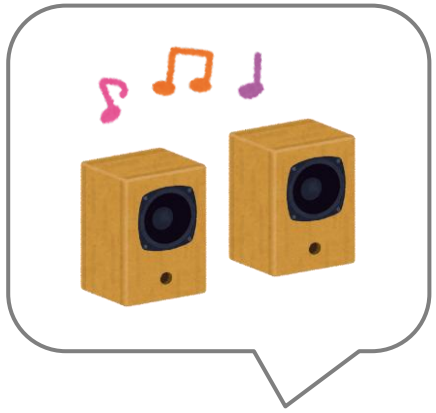
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with 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 …	



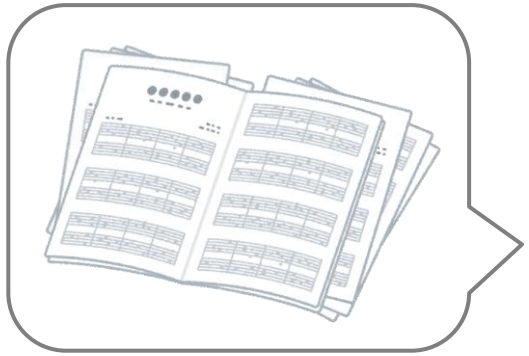
Integration into professional creative workflow

Human-inspired Machine Learning for Music & Audio

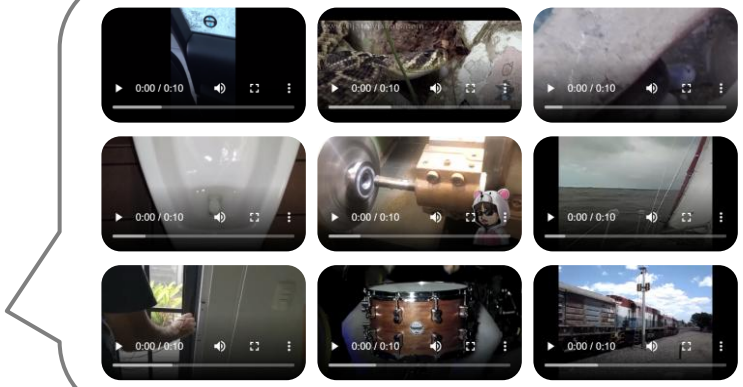
Learning from listening



Learning from reading



Learning from watching



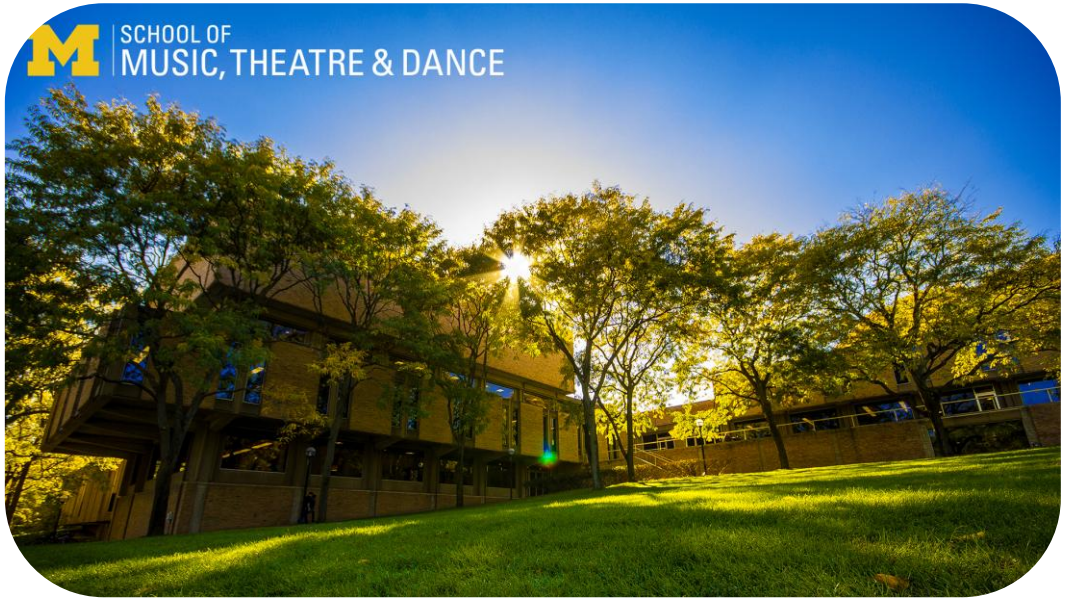
Augmenting Human Creativity with AI

- **Multimodal generative AI for content creation**
- **Human-AI co-creative tools for music, audio and video creation**
- **Human-like machine leaning algorithms for music, movies and arts**

AI Music @ Michigan

Hao-Wen Dong

Julie Zhu



SCHOOL OF MUSIC, THEATRE & DANCE
PERFORMING ARTS TECHNOLOGY
UNIVERSITY OF MICHIGAN

Augmenting Human Creativity with AI

Nothing would have been possible without all my fantastic collaborators!

UC San Diego · 中央研究院 ACADEMIA SINICA · Dolby · SONY · amazon · Ministry of Education · erc

hermandong.com / hwdong@umich.edu

UNIVERSITY OF MICHIGAN