Final Defense for PhD in Computer Science

Generative AI for Music and Audio

Hao-Wen (Herman) Dong

UC San Diego

May 15, 2024

About Me

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

Fall 2023 — NVIDIA
Research Intern

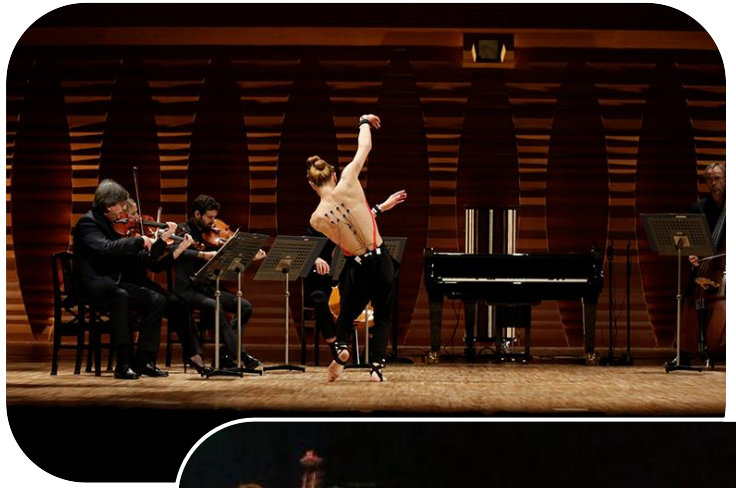
2019 – *present* — UC San Diego
Ph.D. in Computer Science (expected)

Music & Technology

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)

Shlizerman et al., "Audio to Body Dynamics," *Proc. CVPR*, 2018.
https://www.yamaha.com/en/news_release/2018/18013101/
https://www.sankei.com/article/20240113-CQCOSQHJWFIYPJJKZDCITRTRVI/
https://www.roboticgizmos.com/shimon-musical-robot-deep-learning/
https://www.nbcdfw.com/entertainment/the-scene/how-verdigris-ensemble-is-using-ai-to-create-a-new-concert-experience/3366031/

Use Cases of AI for Music & Audio

(Source: The Denver Post)

(Source: UploadVR)

(Source: Descript)

(Source: Daily Bruin)

(Source: Wikimedia Commons)

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

New technology creates new art form

AI

Empowering music and audio creation with machine learning

Music & Audio

New art form inspires new technology

My Research

Infrastructure for
Music Generation Research

How can AI help professionals and amateurs create music and audio content?

AI-powered Assistive
Music Creation Tools

AI

Music & Audio

Deep Learning for
Multitrack Music Generation

How can AI learn to create music and audio like how humans do?

Multimodal Machine Learning
for Audio & Music

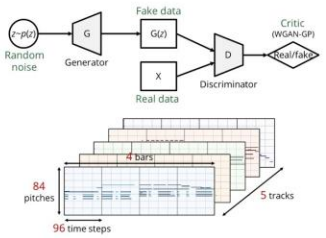
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

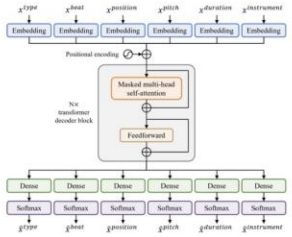
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)



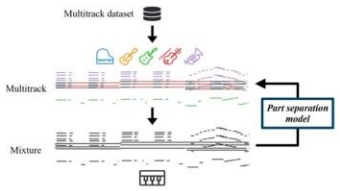
MMT
(ICASSP 2023)



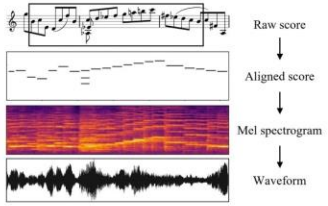
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger
(ISMIR 2021)



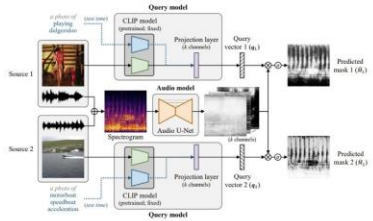
Deep Performer
(ICASSP 2022)



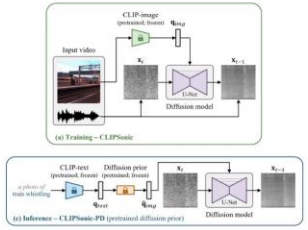
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



CLIPSonic
(WASPAA 2023)



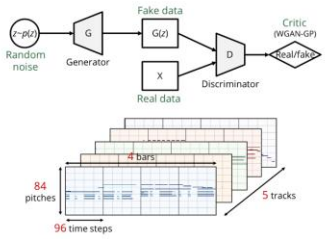
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

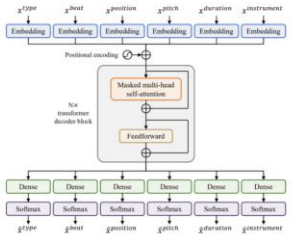
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)



MMT
(ICASSP 2023)



Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

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)

Generative AI for Music & Audio

Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)

MMT
(ICASSP 2023)

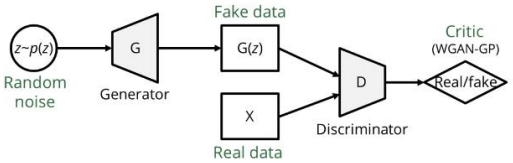
How can we build better machine learning models for music generation?

Generative AI for Music & Audio

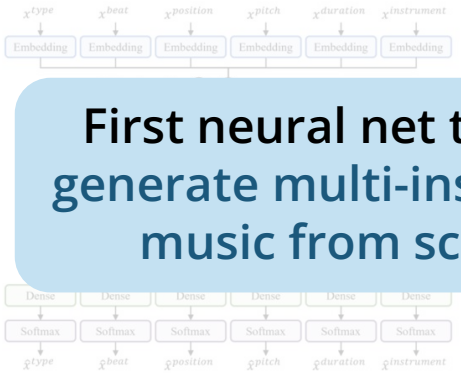
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN (AAAI 2018)



MMT (ICASSP 2023)

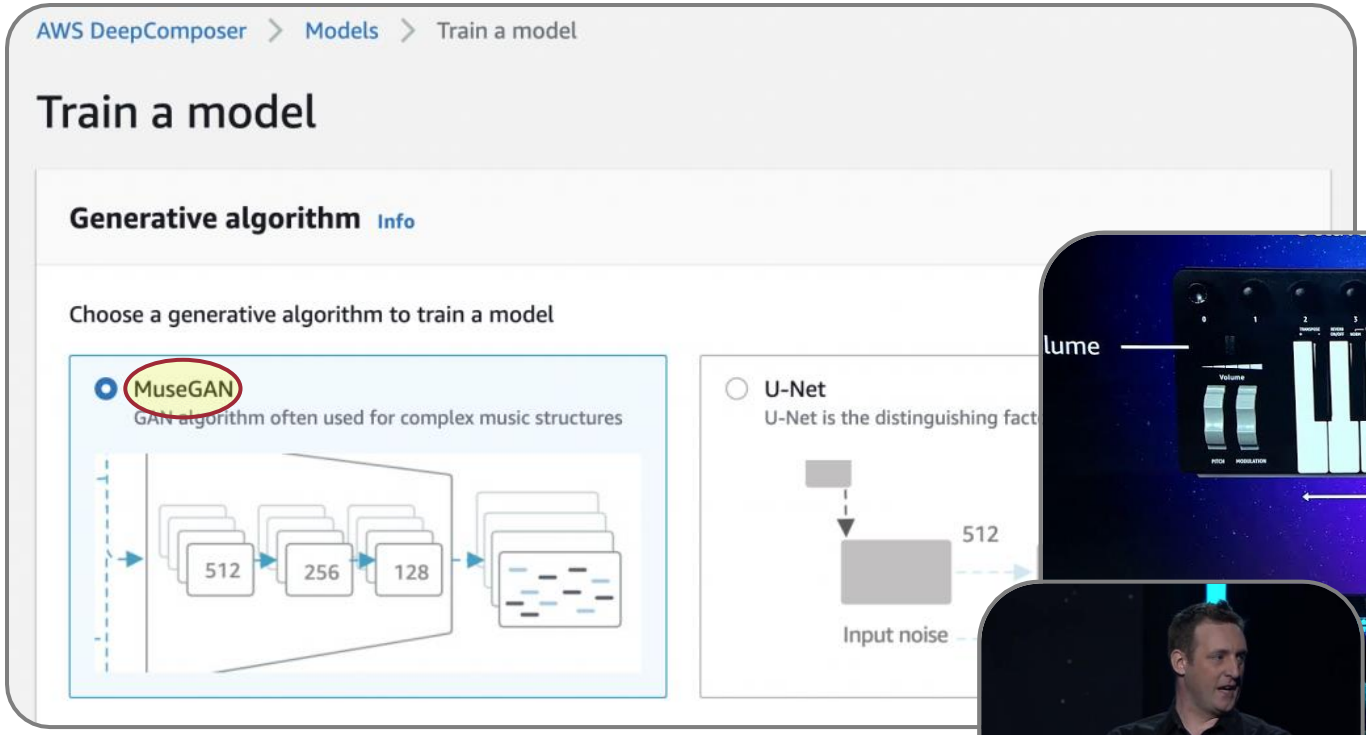


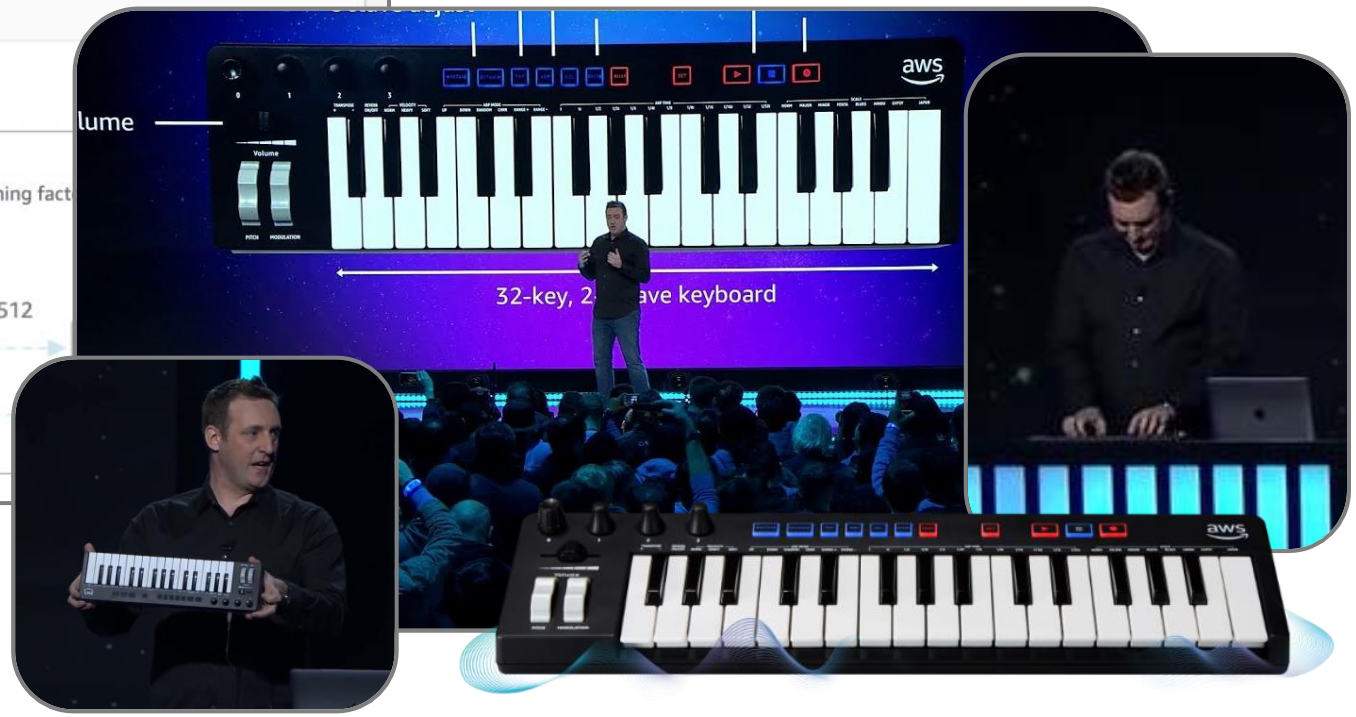
First neural net that can generate multi-instrument music from scratch

Work done before starting my PhD

Pop Music Generation

MuseGAN in AWS DeepComposer





MuseGAN featured in AWS DeepComposer!

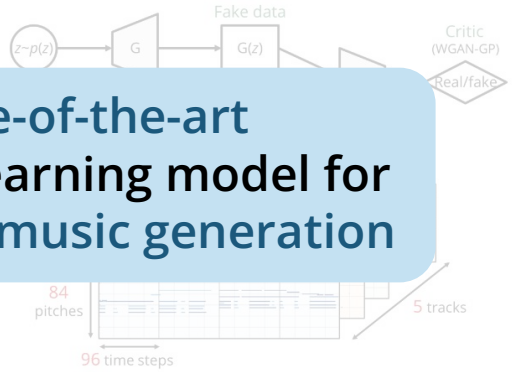
https://www.amazon.com/dp/B07YGZ4V5B/
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," *AWS News Blog*, April 2, 2020.

Generative AI for Music & Audio

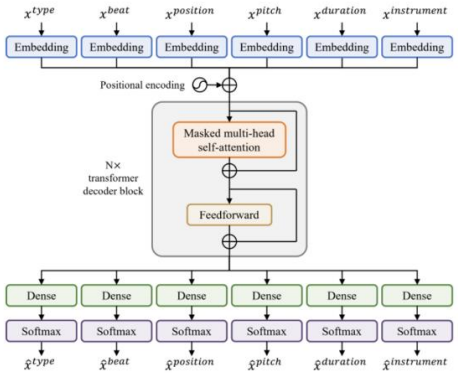
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)



MMT
(ICASSP 2023)



State-of-the-art machine learning model for orchestral music generation

Orchestral Music Generation

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

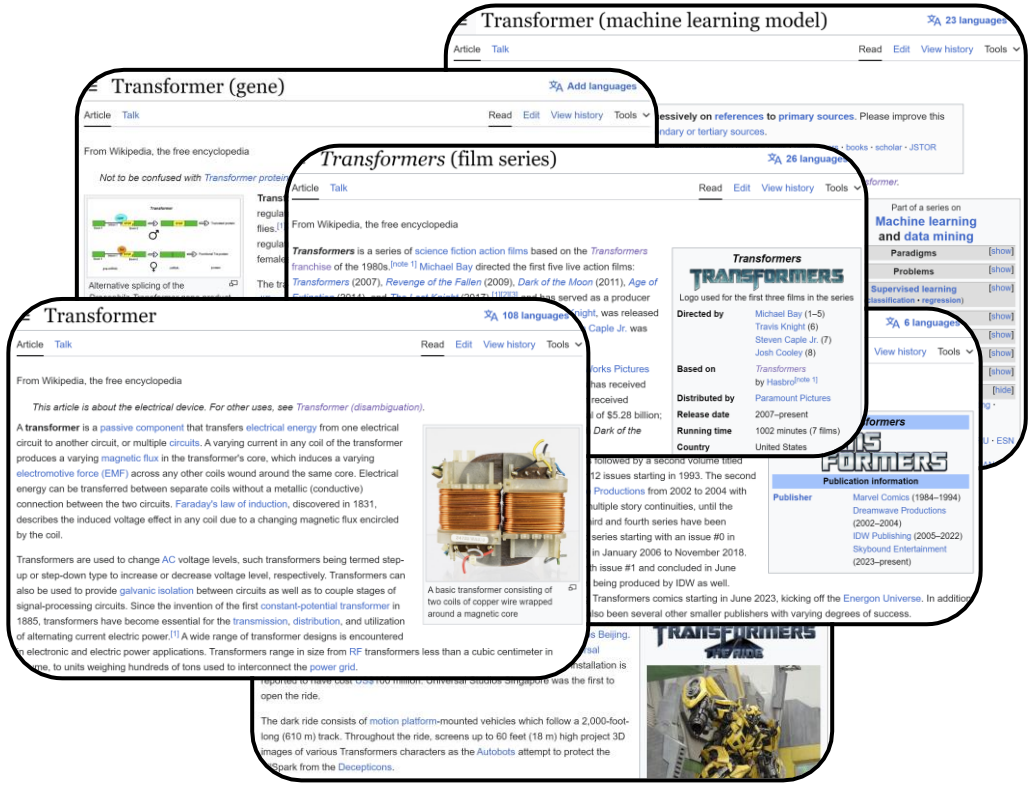
(Source: Vienna Mozart Orchestra)

3.5x longer generated samples
3.3x faster generation speed
} Critical for orchestral music!

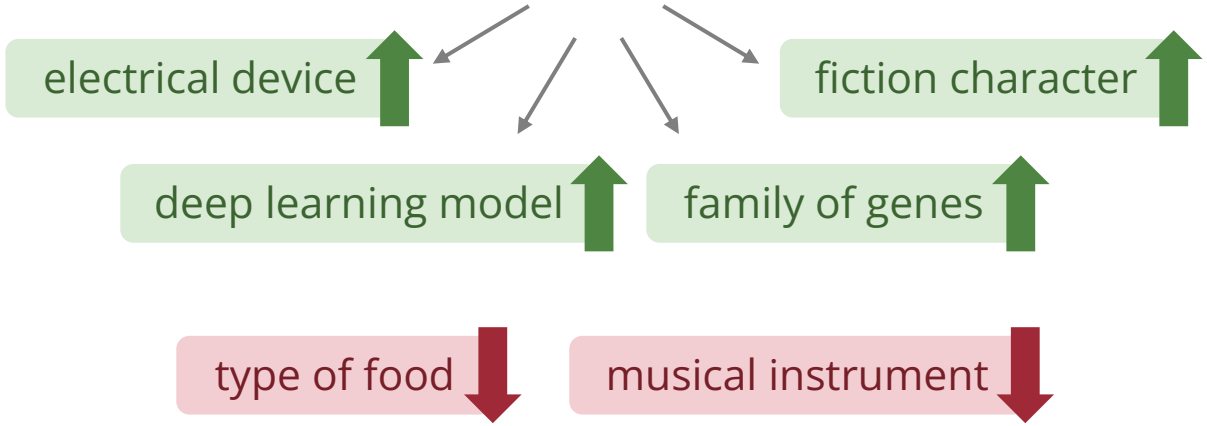
Competitive quality of generated music

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?

A transformer is a	→	Model	→	deep
A transformer is a deep	→	Model	→	learning
A transformer is a deep learning	→	Model	→	model
A transformer is a deep learning model	→	Model	→	introduced
A transformer is a deep learning model introduced	→	Model	→	in
A transformer is a deep learning model introduced in	→	Model	→	2017

Designing a Machine-readable Music Language

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

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

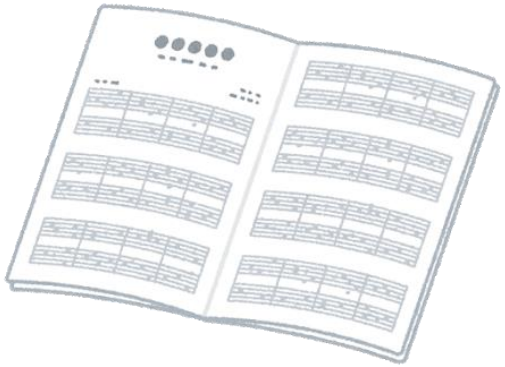
- Each super word $\mathbf{x}_i$ encodes:

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

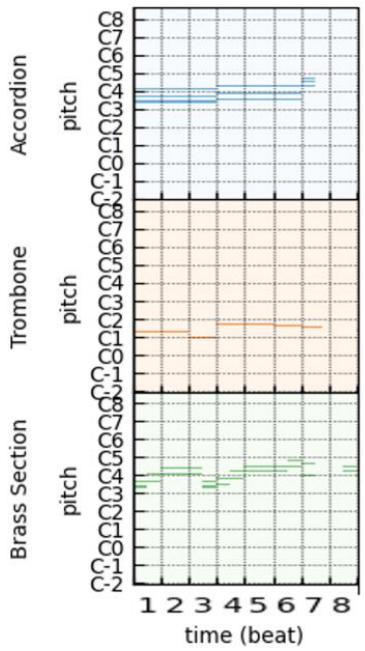
Specify note & instrument information

Structural
- Start of song
- Start of notes
- End of song

Data
- Instrument
- Note



An Example of the Proposed Representation



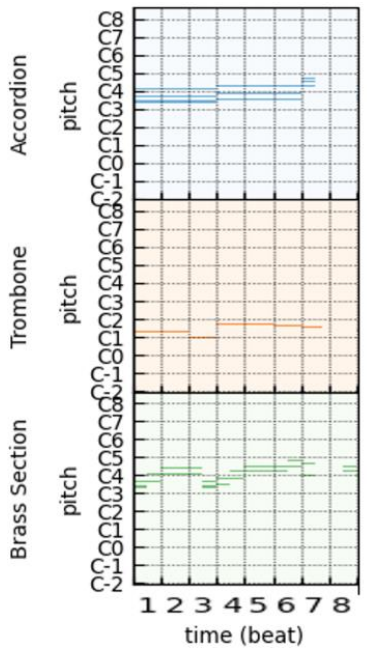
Structural events

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
```

An Example of the Proposed Representation



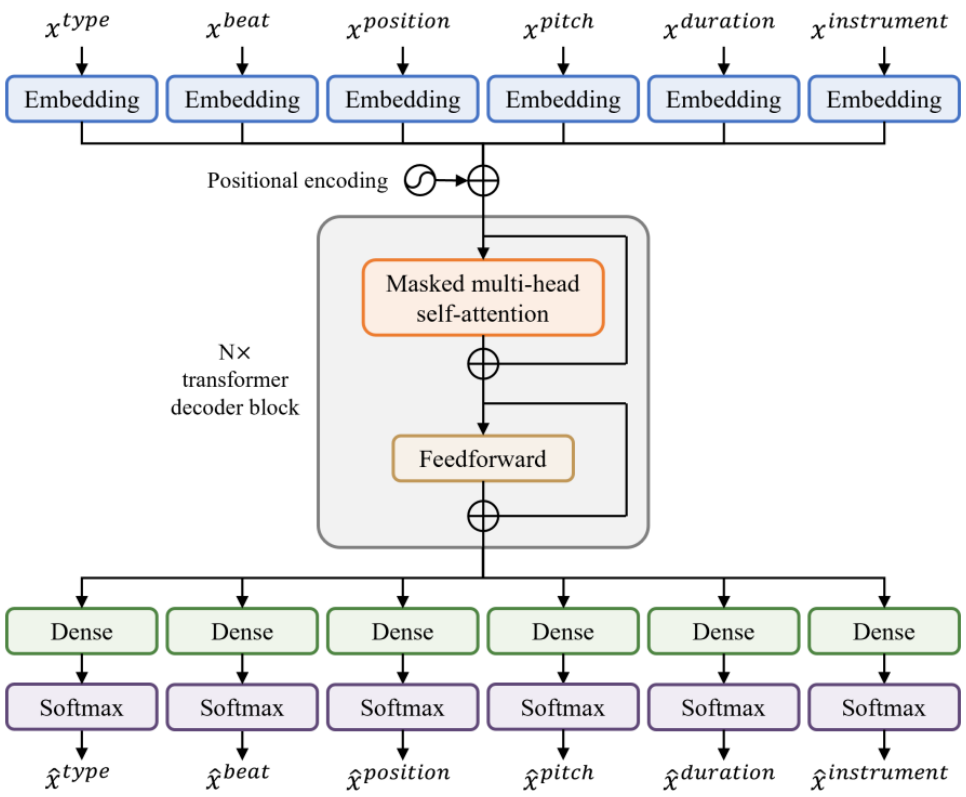

```
(0, 0,  0,  0,  0,  0)
(1, 0,  0,  0,  0, 15)
(1, 0,  0,  0,  0, 36)
(1, 0,  0,  0,  0, 39)
(2, 0,  0,  0,  0,  0)
(3, 1,  1, 41, 15, 36)
(3, 1,  1, 65,  4, 39)
(3, 1,  1, 65, 17, 15)
(3, 1,  1, 68,  4, 39)
(3, 1,  1, 68, 17, 15)
(3, 1,  1, 73, 17, 15)
(3, 1, 13, 68,  4, 39)
(3, 1, 13, 73,  4, 39)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

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

Word-by-word



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

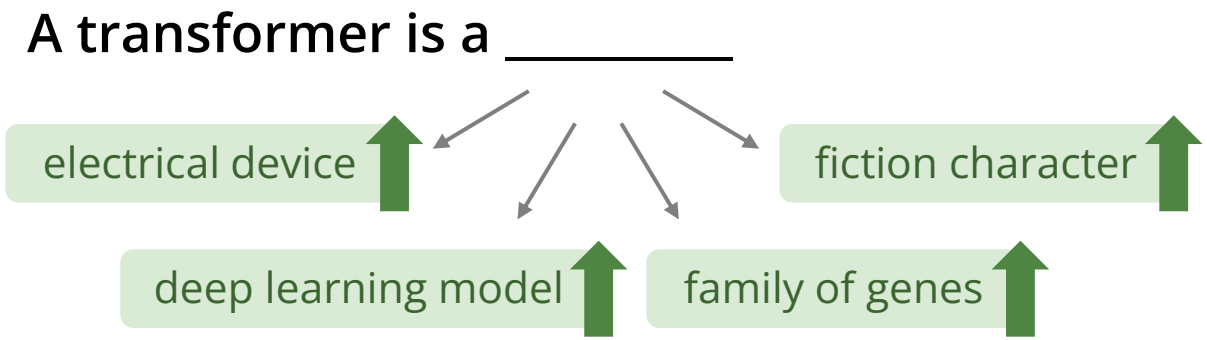
church-organ, viola, contrabass, strings, voices, horn, oboe

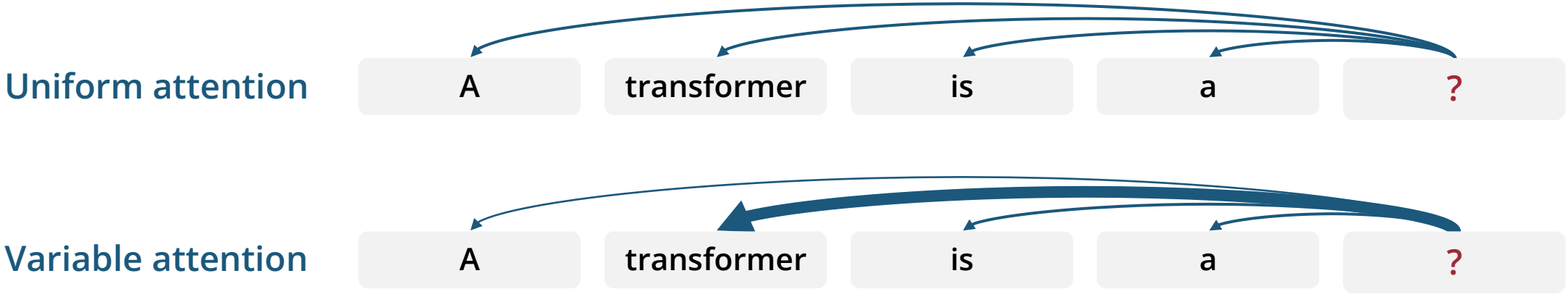
4-beat continuation

Mozart's
Eine kleine Nachtmusik

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

The Magic of Transformers – Self-attention Mechanism

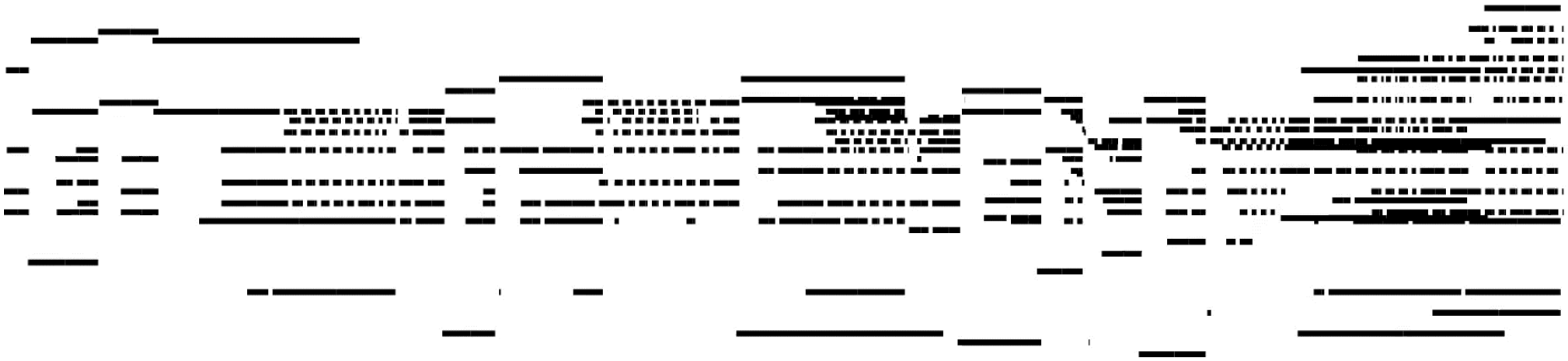




Transformers learn what to attend to from big data!

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

(Each color represents an attention head)

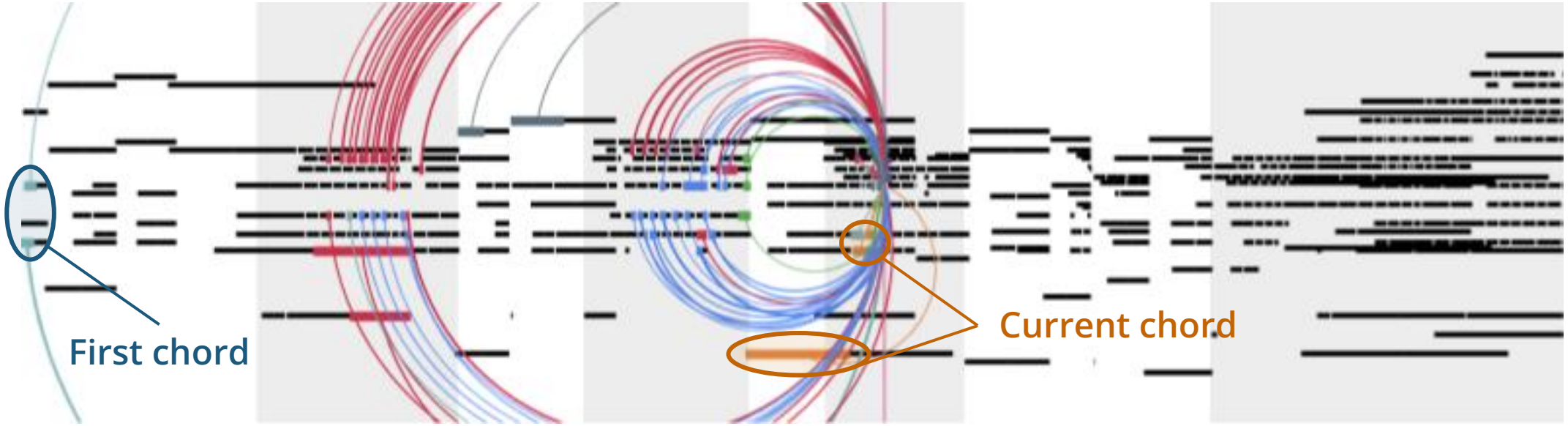


(Source: Huang et al., 2018)

Huang et al., "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?

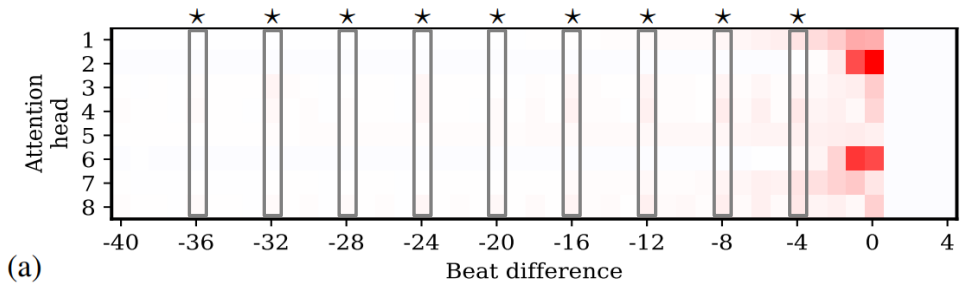
Huang et al., "Music Transformer: Generating Music with Long-Term Structure," *Magenta Blog*, December 13, 2018.

Systematically Analyzing Musical Self-attention

The MMT model attends more to notes

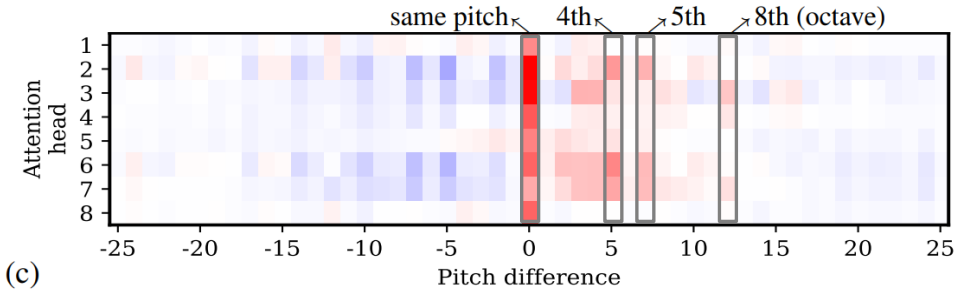
that are $4N$ beats away in the past

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

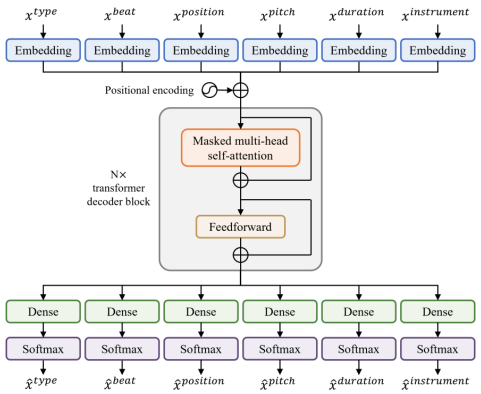


MMT learns a relative self-attention for beat and pitch!

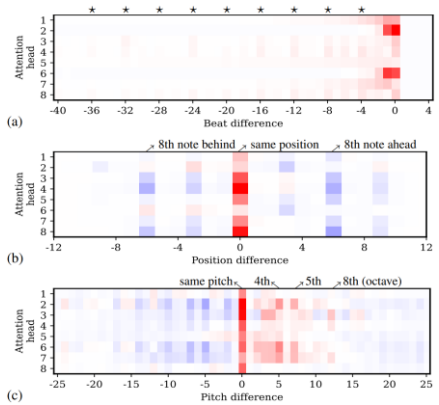
Summary

- **State-of-the-art** machine learning model for **orchestral 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

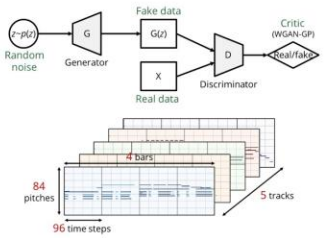
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

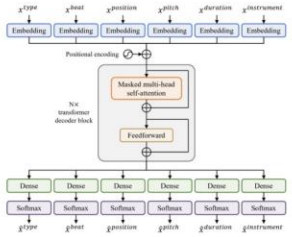
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN (AAAI 2018)



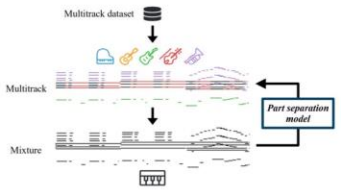
MMT (ICASSP 2023)



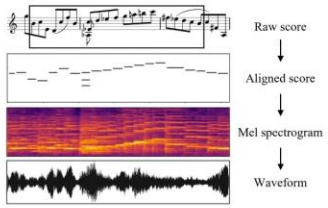
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger (ISMIR 2021)



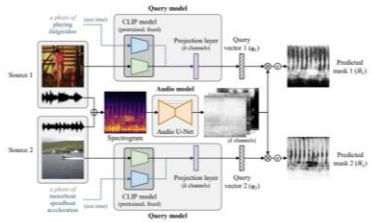
Deep Performer (ICASSP 2022)



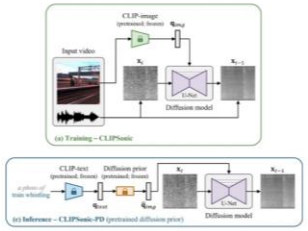
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep (ICLR 2023)



CLIPSonic (WASPAA 2023)



Generative AI for Music & Audio

Empowering music and audio creation with machine learning

Multitrack Music Generation

Advancing deep generative models for multitrack music

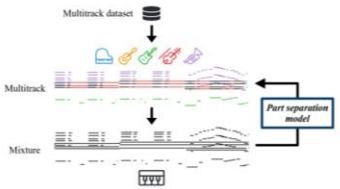
MuseGAN
(AAAI 2018)

MMT
(ICASSP 2023)

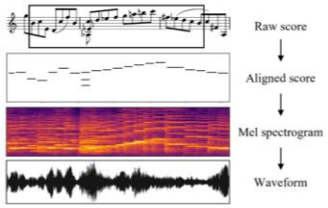
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

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)

Generative AI for Music & Audio

Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger
(ISMIR 2021)

Deep Performer
(ICASSP 2022)

How can AI help professionals and amateurs create music?

Multitrack Music Genera

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)

lal Learning for Audio & Music

und separation
sis from videos

Sep
023)

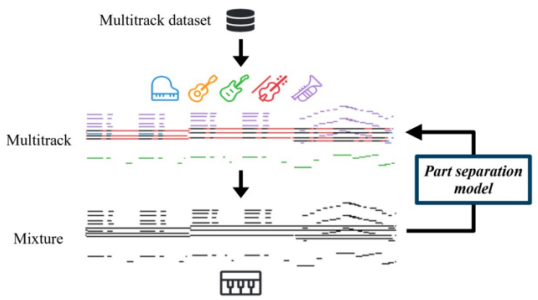
CLIPSonic
(WASPAA 2023)

Generative AI for Music & Audio

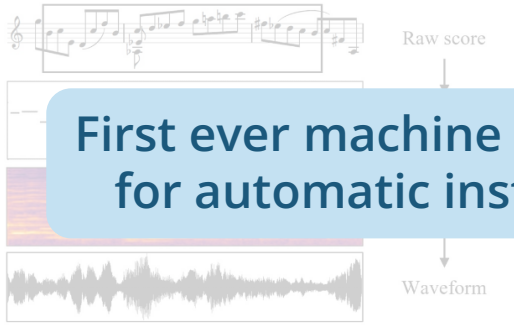
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger
(ISMIR 2021)



Deep Performer
(ICASSP 2022)



First ever machine learning model for automatic instrumentation

Automatic Instrumentation

Towards Automatic Instrumentation by Learning to Separate Parts in Multitrack Music

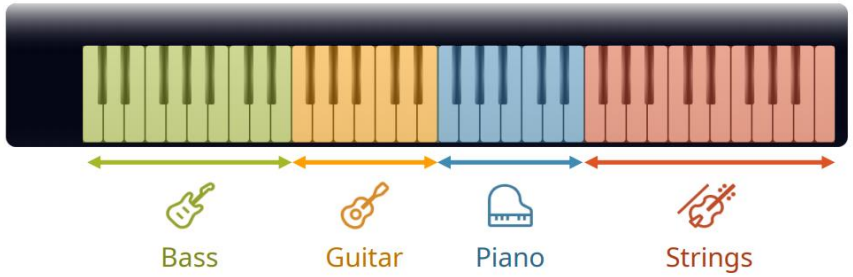
Hao-Wen Dong[1] Chris Donahue[2] Taylor Berg-Kirkpatrick[1] Julian McAuley[1]

[1] University of California San Diego [2] Stanford University

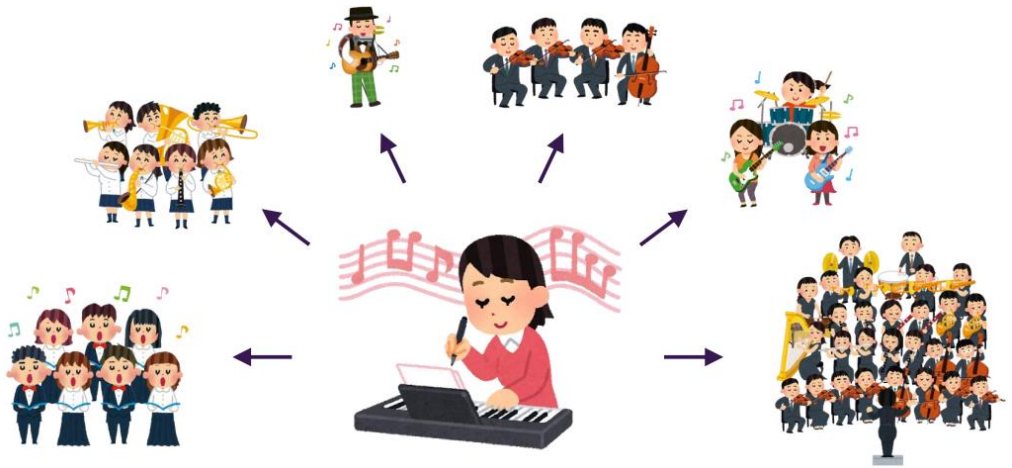
Automatic Instrumentation

- Goal: Dynamically assign instruments to notes in solo music

Intelligent musical instruments

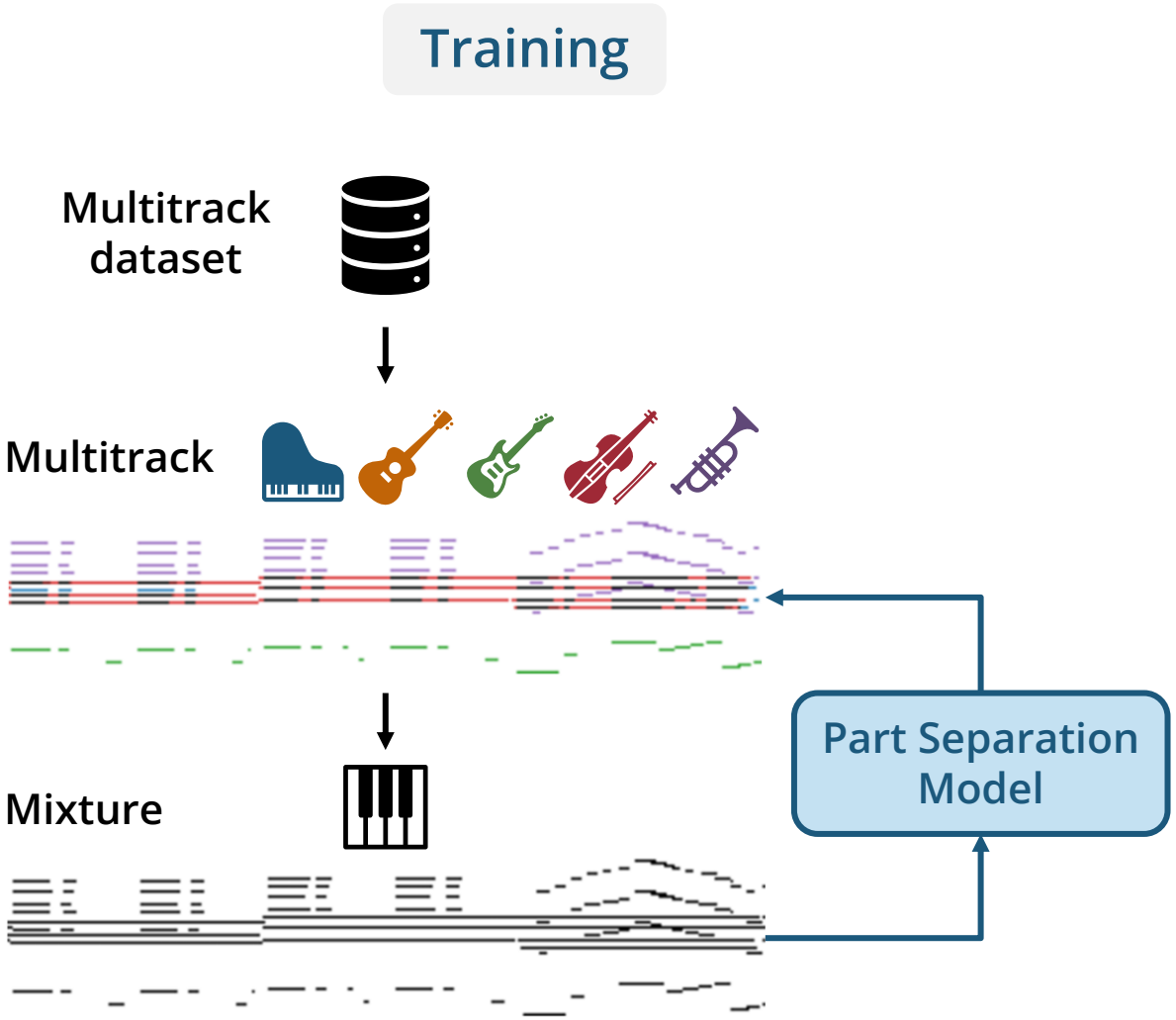


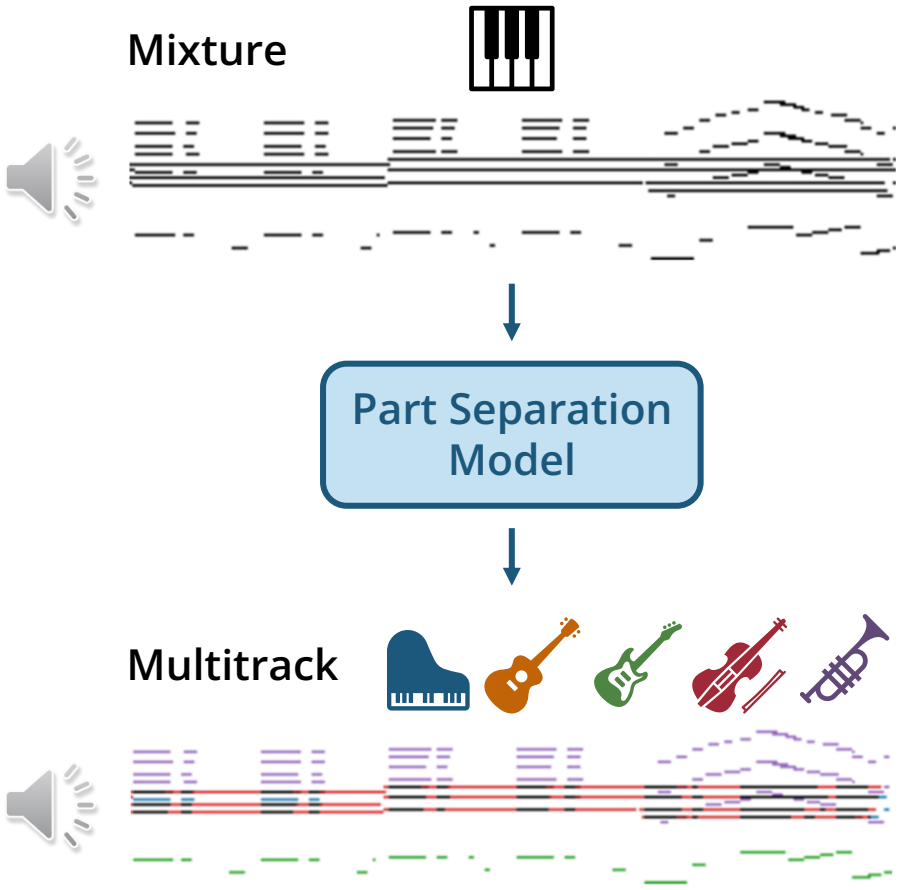
Assistive composing tools



How can we acquire paired data?

Learning Automatic Instrumentation without Paired Data



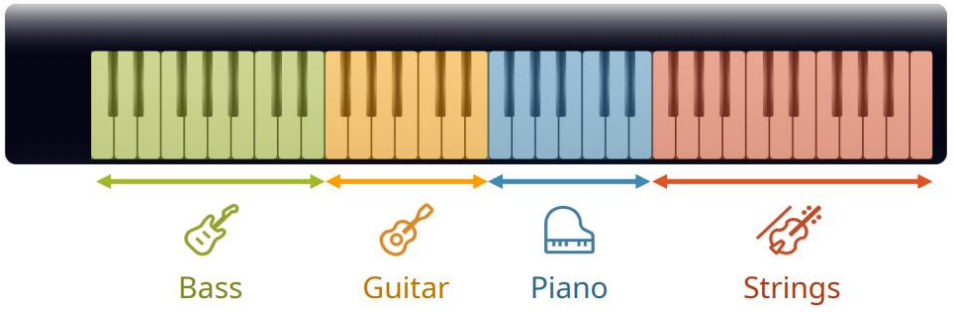


Two Types of Model

Online models

Can only look at the **past**

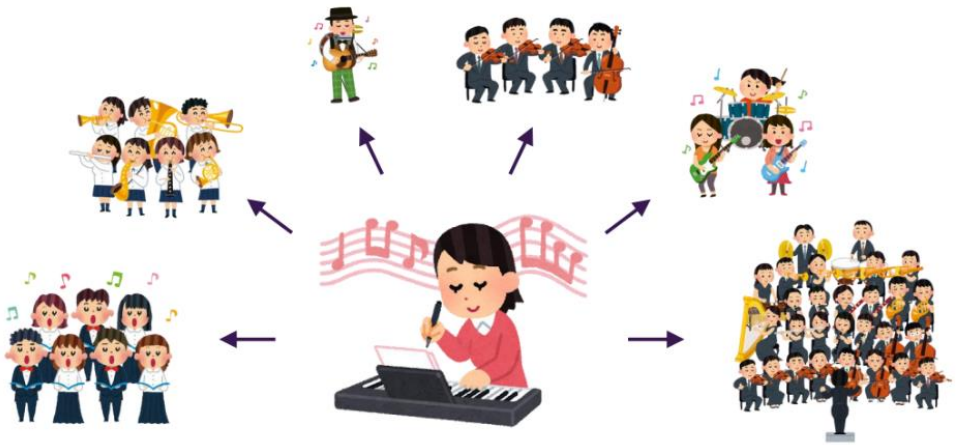
- LSTMs
- Transformer decoders



Offline models

Can look at both the **future** and the **past**

- BiLSTMs
- Transformer encoders



Representation & Datasets

A **sequence of notes** specified by

- **Time** Onset time (in time step)
- **Pitch** Pitch as a MIDI note number
- **Duration** Note length (in time step)
- **Frequency** Frequency of the pitch (in Hz)
- **Beat** Onset time (in beat)
- **Position** Position within a beat (in time step)

Representing music in a machine-readable format

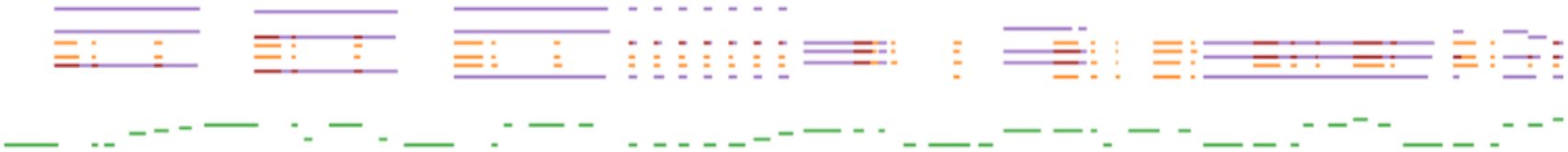
Dataset	Hours	Files	Notes	Parts	Ensemble	Most common label
Bach chorales [31]	3.23	409	96.6K	4	soprano, alto, tenor, bass	bass (27.05%)
String quartets [32]	6.31	57	226K	4	first violin, second violin, viola, cello	first violin (38.72%)
Game music [33]	45.05	4.61K	2.46M	3	pulse wave I, pulse wave II, triangle wave	pulse wave II (39.35%)
Pop music [34]	1.02K	16.2K	63.6M	5	piano, guitar, bass, strings, brass	guitar (42.50%)

Example Results

- Produce alternative convincing instrumentations for an existing arrangement

piano, guitar, bass, strings, brass

Original

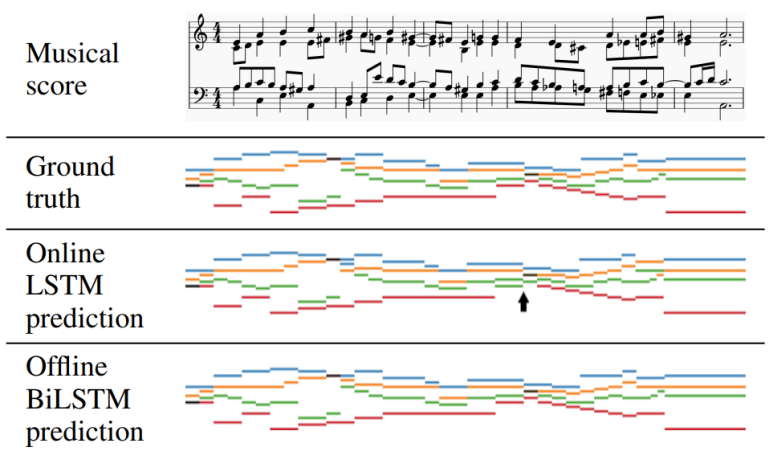


LSTM
(w/o entry hints)

BiLSTM
(w/ entry hints)

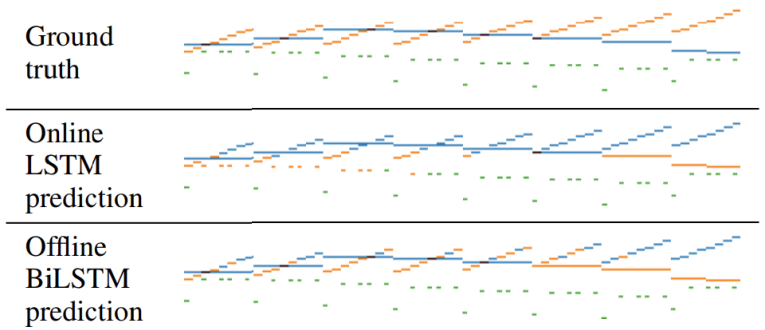
More Results

Bach chorales



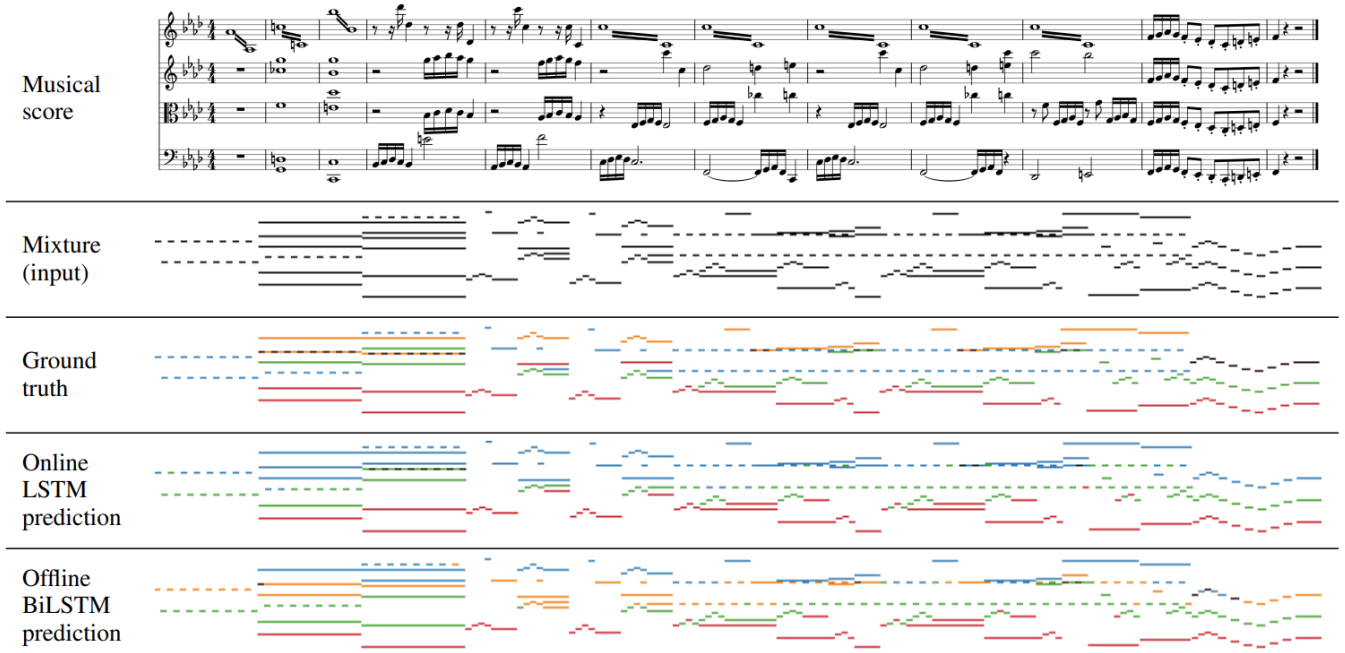
(Audio available. [1] Colors: soprano, alto, tenor, bass.)

Game music



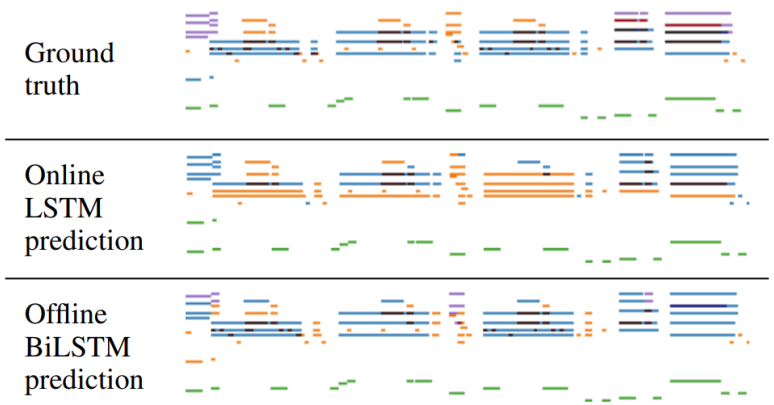
(Audio available. [1] Colors: pulse wave I, pulse wave II, triangle wave.)

String quartets



(Audio available. [1] Colors: first violin, second violin, viola, cello.)

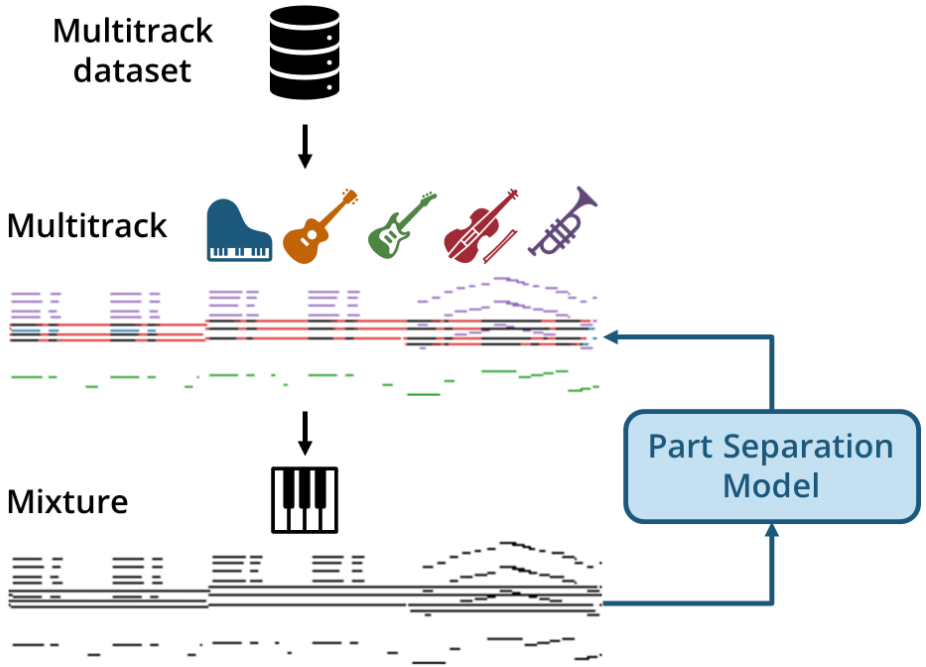
Pop music



(Audio available. [1] Colors: piano, guitar, bass, strings, brass.)

Summary

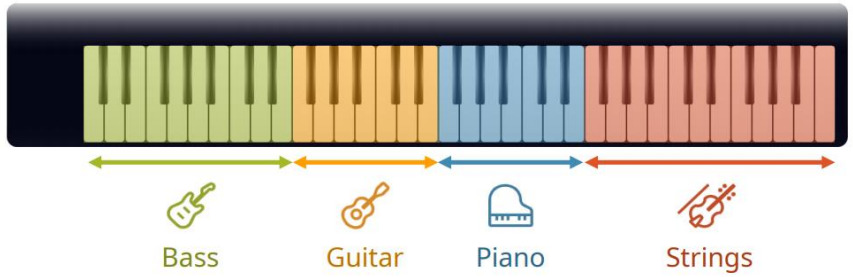
- First ever machine learning model for **automatic instrumentation**
- Potential applications in **assistive creation tools** and **intelligent keyboards**



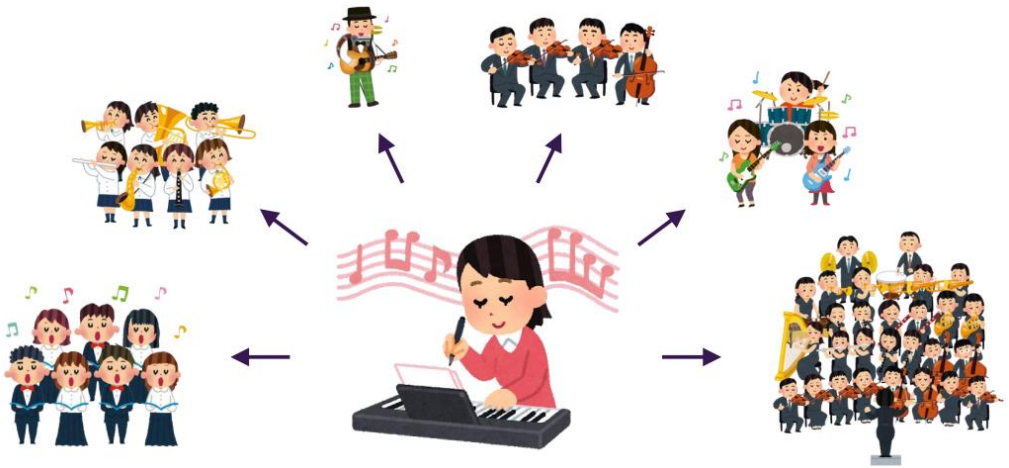
Paper: arxiv.org/abs/2107.05916
Demo: salu133445.github.io/arranger
Code: github.com/salu133445/arranger

Potential Applications of Automatic Instrumentation

Intelligent musical instruments



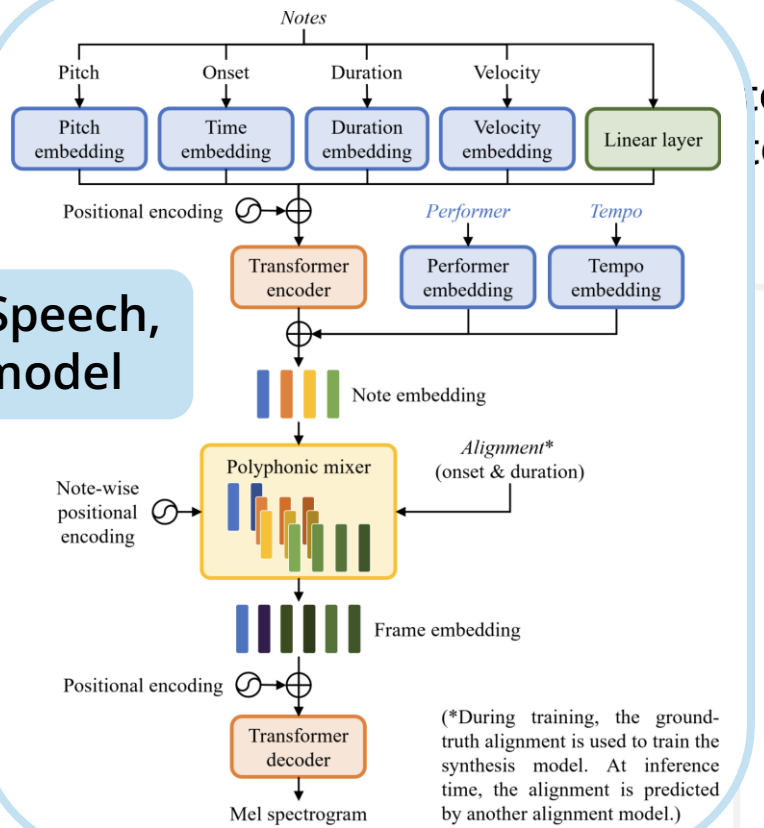
Assistive composing tools



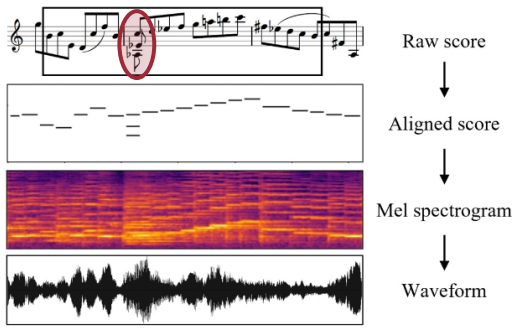
Generative AI for Music & Audio

Assistive Music Creation Tools

Adapted from FastSpeech, a text-to-speech model



Deep Performer (ICASSP 2022)



Score-to-audio synthesis

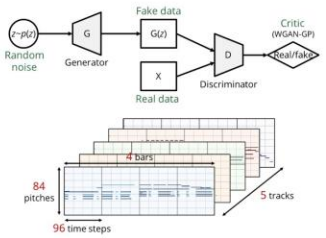
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

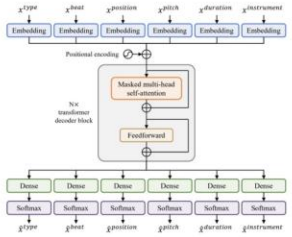
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN (AAAI 2018)



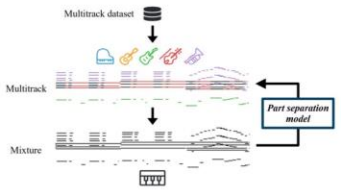
MMT (ICASSP 2023)



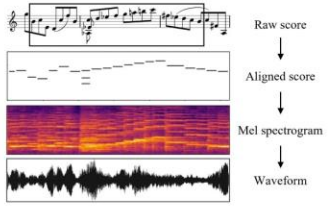
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger (ISMIR 2021)



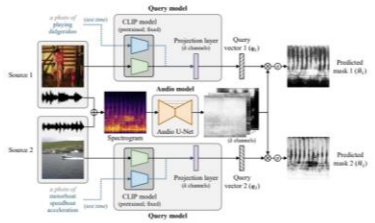
Deep Performer (ICASSP 2022)



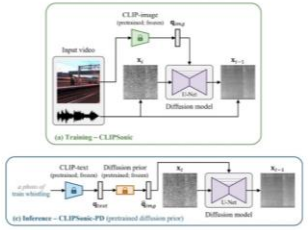
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep (ICLR 2023)



CLIPSonic (WASPAA 2023)



Generative AI for Music & Audio

Empowering music and audio creation with machine learning

Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)

MMT
(ICASSP 2023)

Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

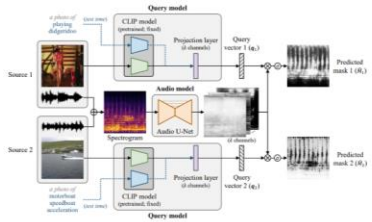
Arranger
(ISMIR 2021)

Deep Performer
(ICASSP 2022)

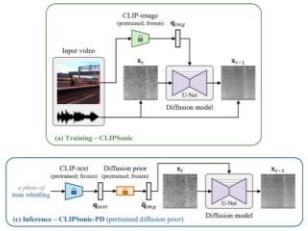
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



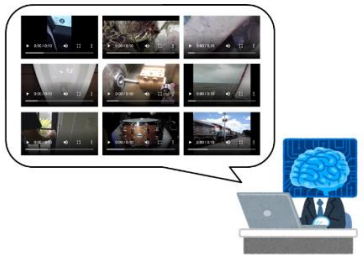
CLIPSonic
(WASPAA 2023)



Generative AI for Music & Audio

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos



CLIPSep
(ICLR 2023)

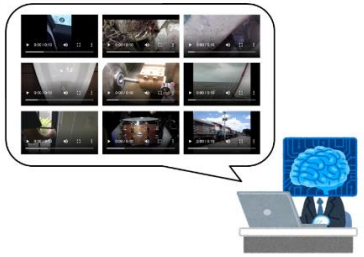
CLIPSonic
(WASPAA 2023)

How can we build AI systems that learn audio concepts like how humans do?

Generative AI for Music & Audio

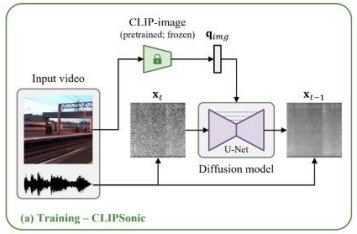
Multimodal Learning for Audio & Music

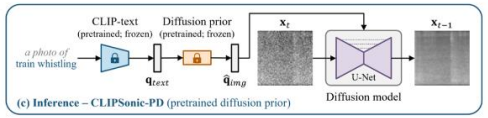
Learning sound separation and synthesis from videos



CLIPSep
(ICLR 2023)

CLIPSonic
(WASPAA 2023)





First text-to-sound synthesis model that can be trained **using only unlabeled videos**

Text-to-audio synthesis

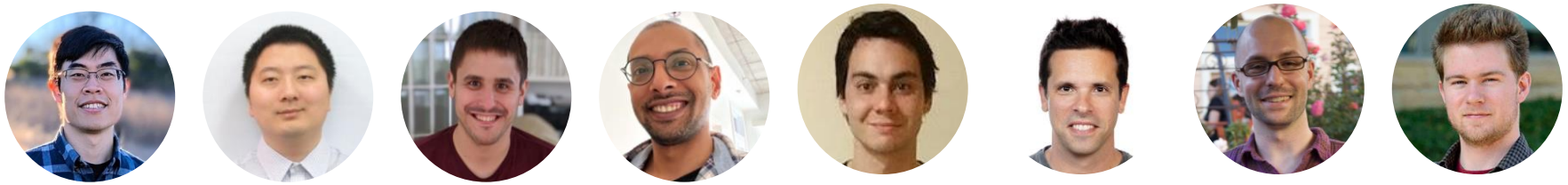
IEEE
WASPAA
2023

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

* Work done during an internship at Dolby



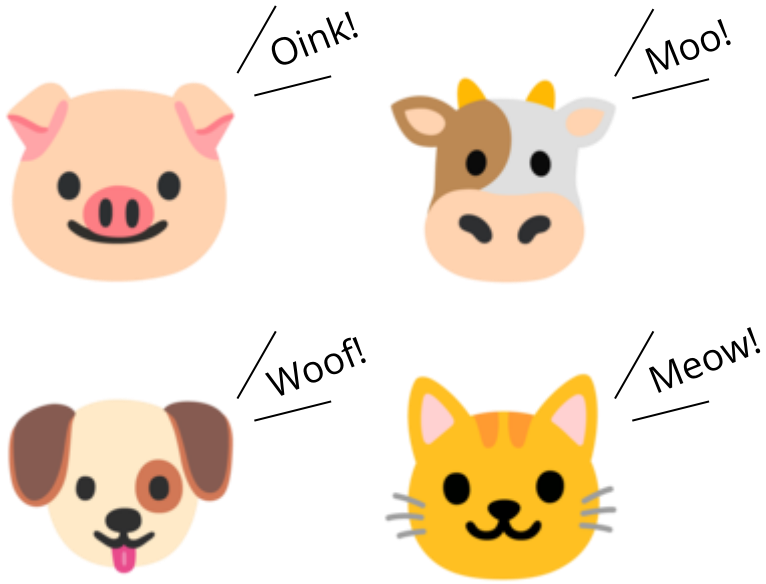
Dolby UC San Diego

What is Text-to-Audio Synthesis?

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

Learning Sounds from Observations

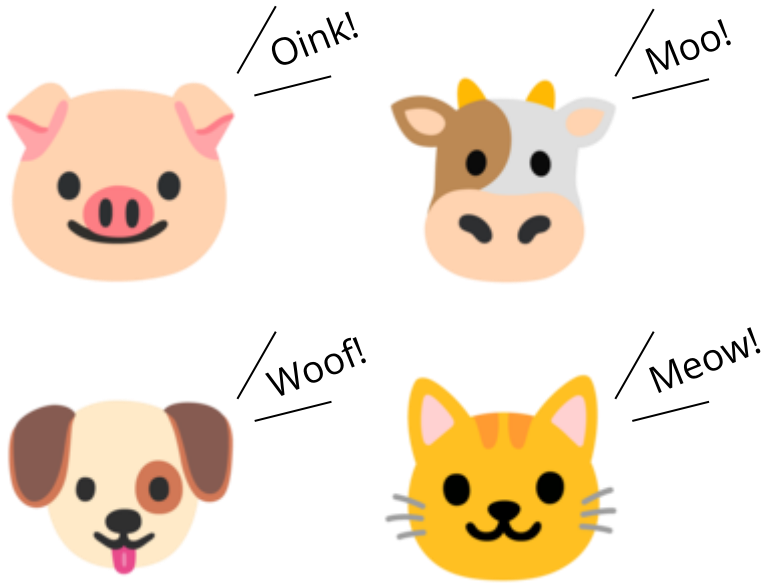
- Watching a dog barking, humans can *associate the barking sound to the dog*



What does the fox say?

Learning Sounds from Noisy Videos

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

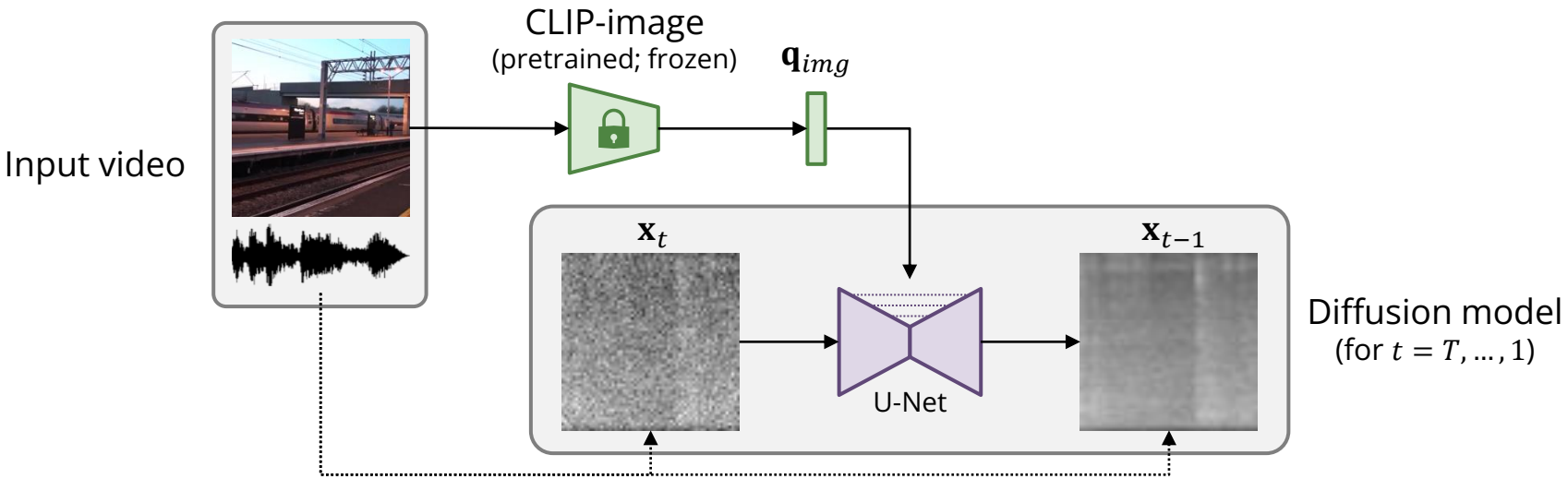


Can machines learn to synthesize sounds from watching *noisy* videos?

URURU TV, "Excited Fox Noises - Fox Laughing and Squeaking," *YouTube*, https://youtu.be/XiCdBFDfUTg, 2019.

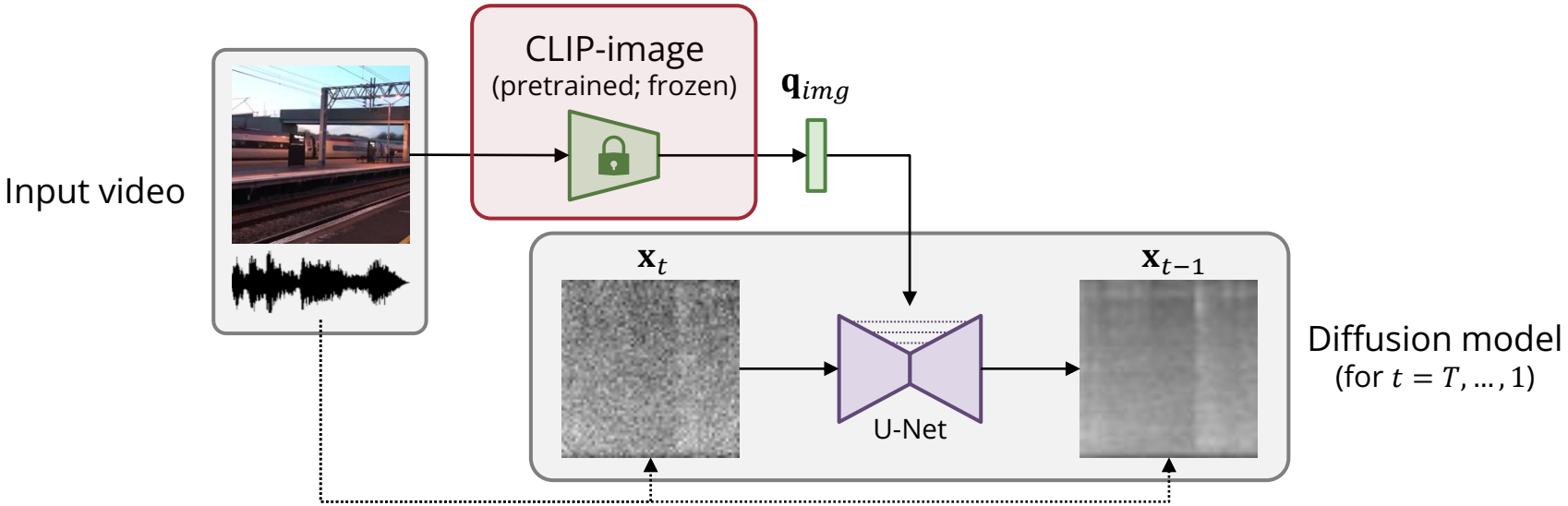
Training an Image-to-Audio Synthesis Model

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



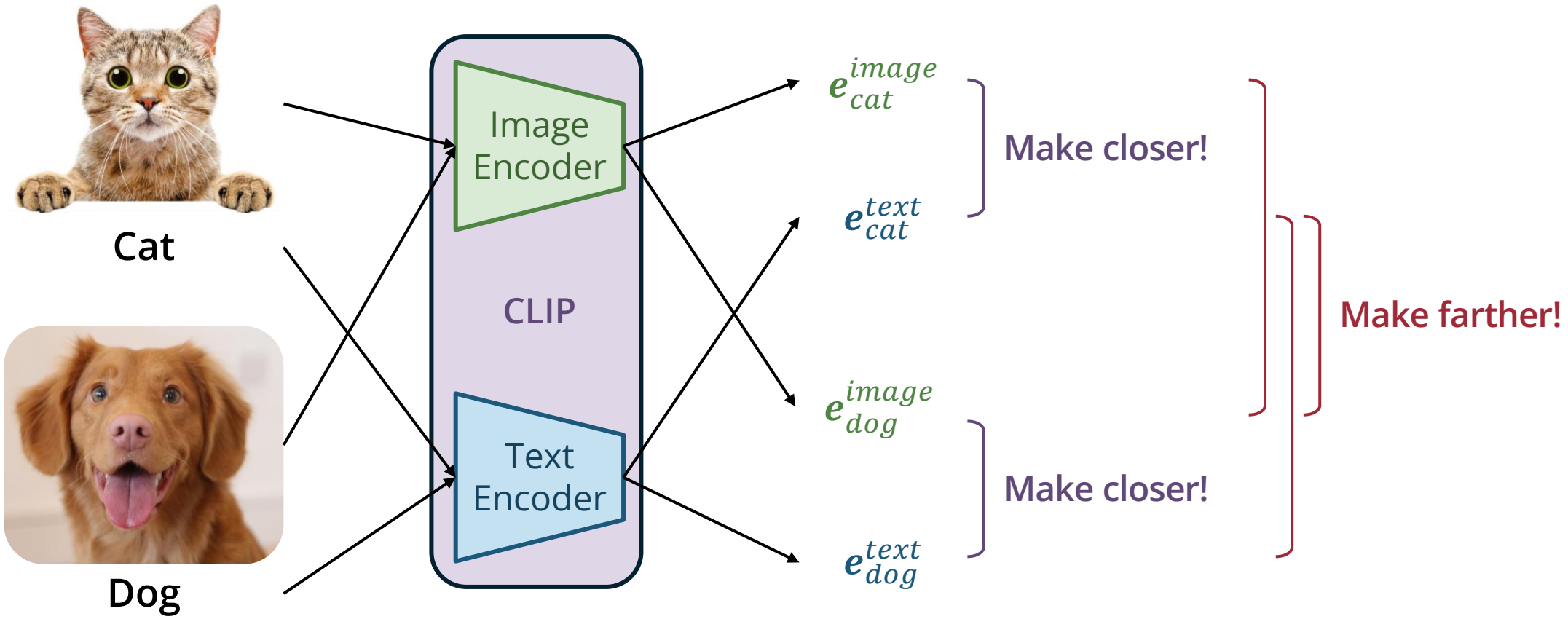
Training an Image-to-Audio Synthesis Model

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



CLIP (Contrastive Language-Image Pretraining)

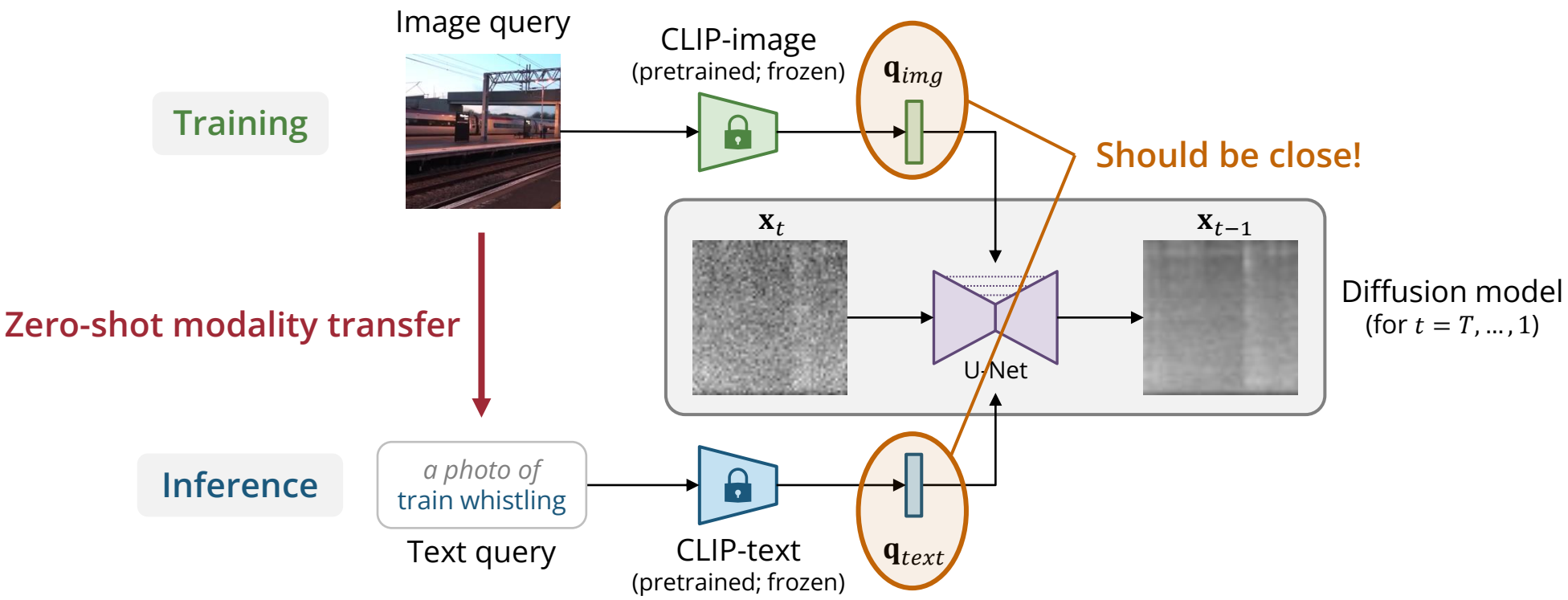
- Learn a shared embedding space for images and texts via *contrastive learning*



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

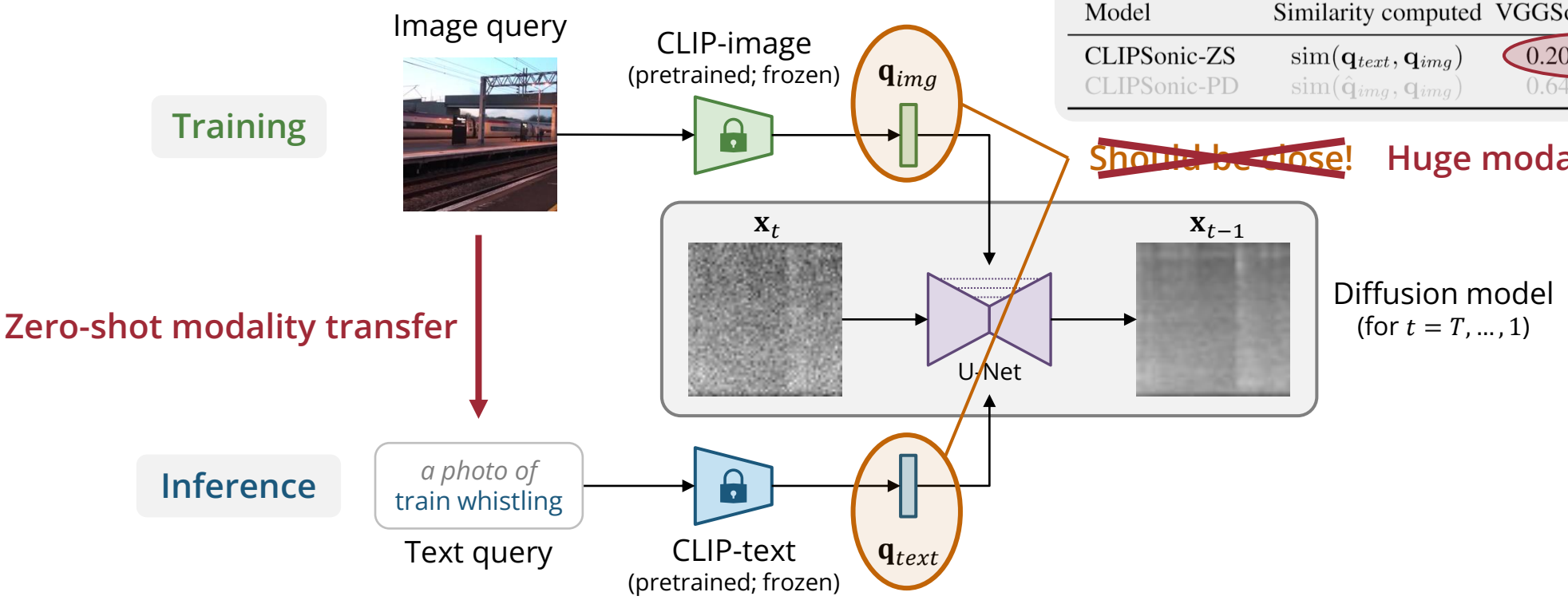


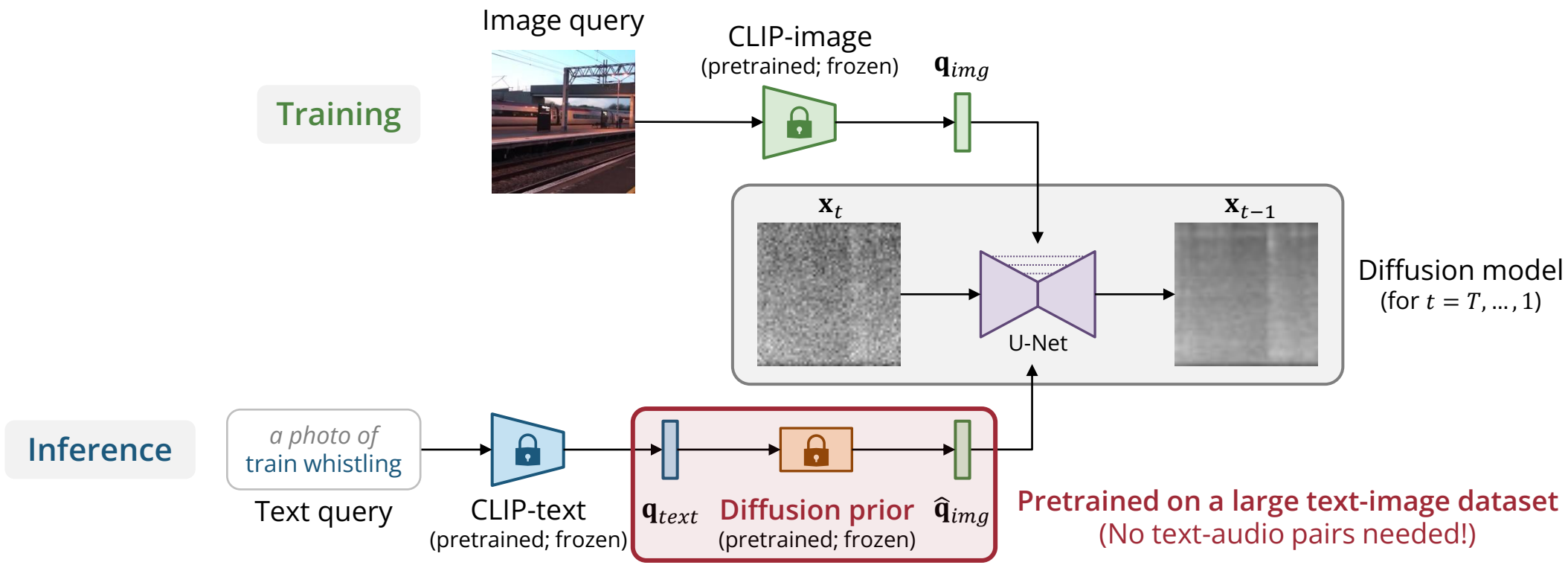
Table 2: Cosine similarities between various query embeddings.

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

~~Should be close!~~ Huge modality gap!

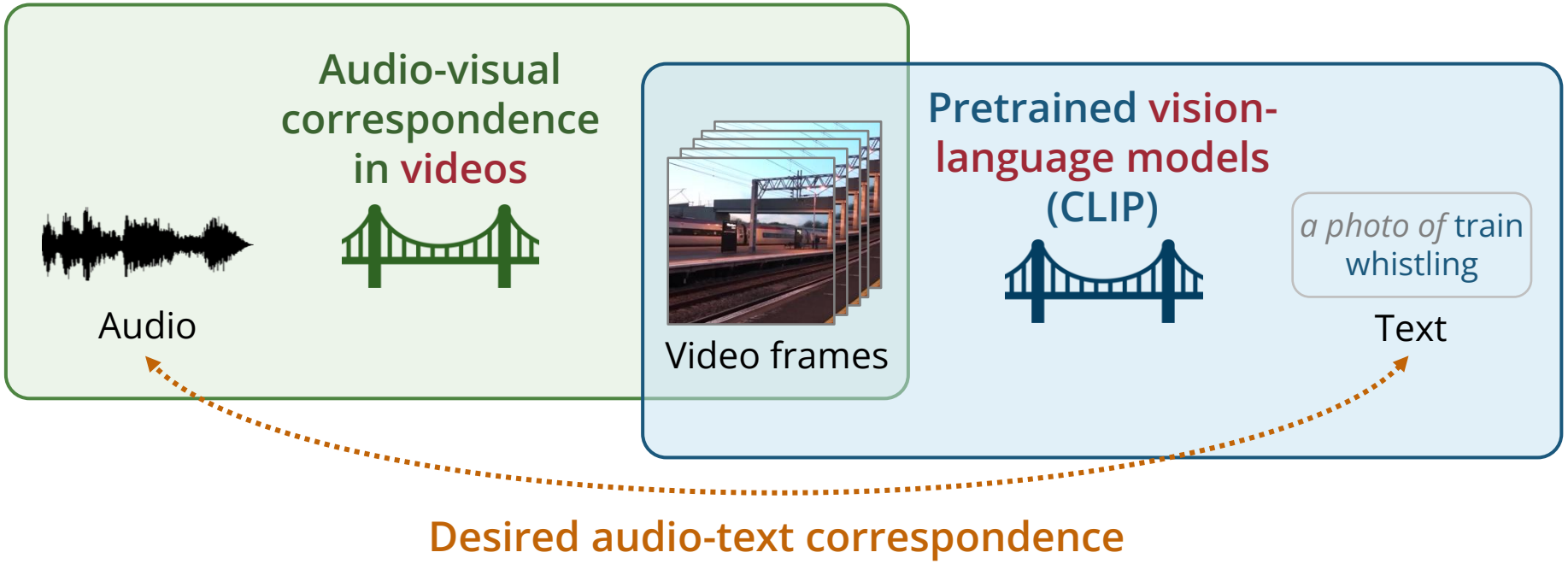
Leveraging Diffusion Prior to Close the Modality Gap

- 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.

Leveraging the Visual Domain as a Bridge



No text-audio pairs required!

Scalable to large video datasets!

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

(1,055 videos, 21 instruments)

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

(172K videos, 310 classes)

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

Example Text-to-Audio Synthesis Results

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

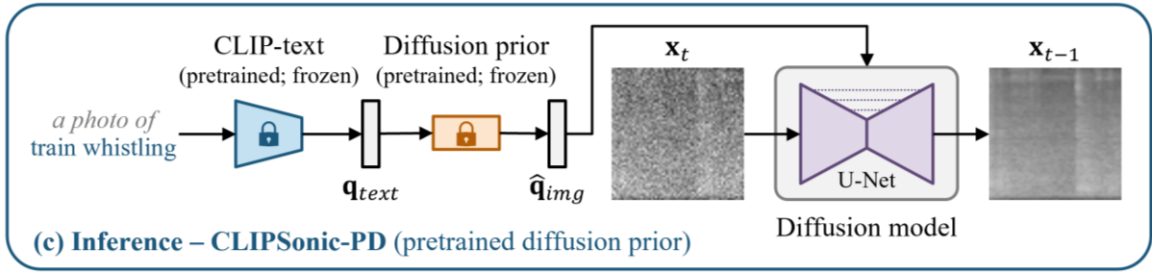
Playing violin fiddle

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

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

Summary

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

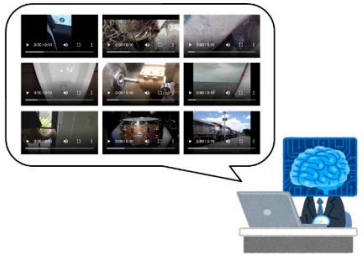


Paper: arxiv.org/abs/2306.09635
Demo: salu133445.github.io/clipsonic

Generative AI for Music & Audio

Multimodal Learning for Audio & Music

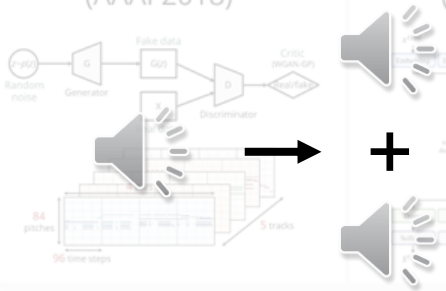
Learning sound separation and synthesis from videos



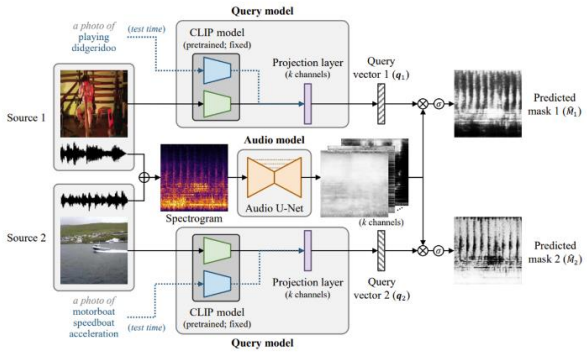
Multitrack Music Genera

Advancing deep generative models for multitrack music

Query:
"*playing harpsichord*"



CLIPSep
(ICLR 2023)



CLIPSonic
(WASPAA 2023)

First text-queried sound separation model that can be trained **using only unlabeled videos**

Text-queried sound separation

Conclusion

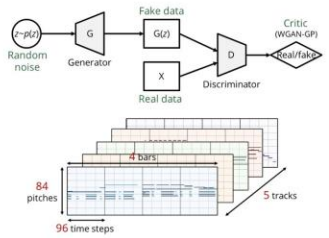
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

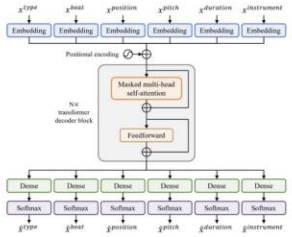
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)



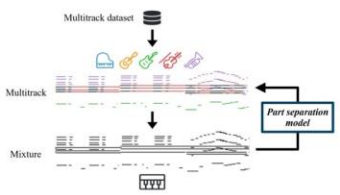
MMT
(ICASSP 2023)



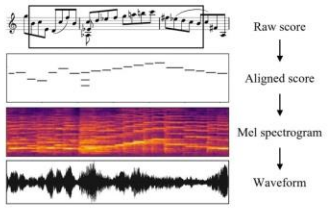
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger
(ISMIR 2021)



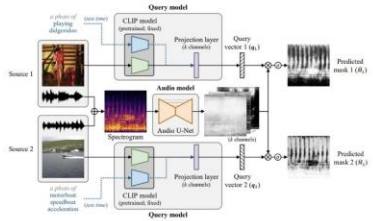
Deep Performer
(ICASSP 2022)



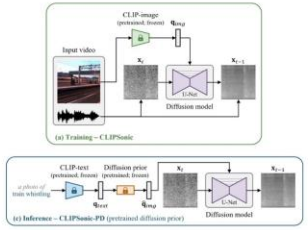
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



CLIPSonic
(WASPAA 2023)



Generative AI for Music & Audio

Empowering music and audio creation with machine learning

Multitrack Music Generation

Dong et al., *ISMIR LBD* 2017
Dong et al., *AAAI* 2018
Dong & Yang, *ISMIR* 2018
Dong et al., *ICASSP* 2023
Xu et al., *AIMG* 2023
Xu et al., *under review*
Ryu et al., *under review*

Assistive Music Creation Tools

Yeh et al., *JNMR* 2021
Dong et al., *ISMIR* 2021
Dong et al., *ICASSP* 2022
Chen et al., *ISMIR* 2022

Multimodal Learning for Audio & Music

Dong et al., *ICLR* 2023
Dong et al., *WSS* 2023
Dong et al., *WASPAA* 2023

Infrastructure for Music Generation Research

Dong et al., *ISMIR* 2021
Dong et al., *ISMIR LBD* 2018

Open Science, Open Source!

OPEN SCIENCE

open source

salu133445/musegan

An AI for Music Generation

Python 1.7k 360

salu133445/muspy

A toolkit for symbolic music generation

Python 413 49

sony/CLIPSep

No description provided

Python 31 1

salu133445/bmusegan

Code for "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation"

Python 57 13

salu133445/arranger

Official Implementation of "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music" (ISMIR 2021)

Python 54 7

salu133445/mmt

Official Implementation of "Multitrack Music Transformer" (ICASSP 2023)

Python 130 22

salu133445/pypianoroll

A toolkit for working with piano rolls

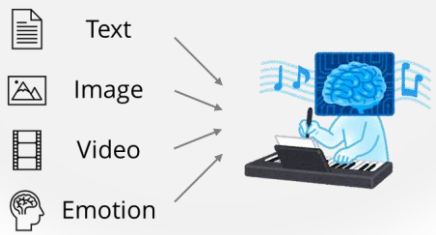
Python 132 16

All source code was made open source with MIT license whenever possible!

github.com/salu133445

Future Directions

Multimodal generative AI with music and audio



How can AI help professionals and amateurs create music and audio content?

Interactive AI tools for music & audio production

AI

Music & Audio

Human-like machine learning algorithms for music

How can AI learn to create music and audio like how humans do?

Interactive human-AI music co-creation

Video Generation without Sounds

Video → Music & sound effects
Text → Video with music & sound effects

OpenAI, Sora: Creating video from text, https://openai.com/sora, Feb 15, 2024.

Multimodal 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**

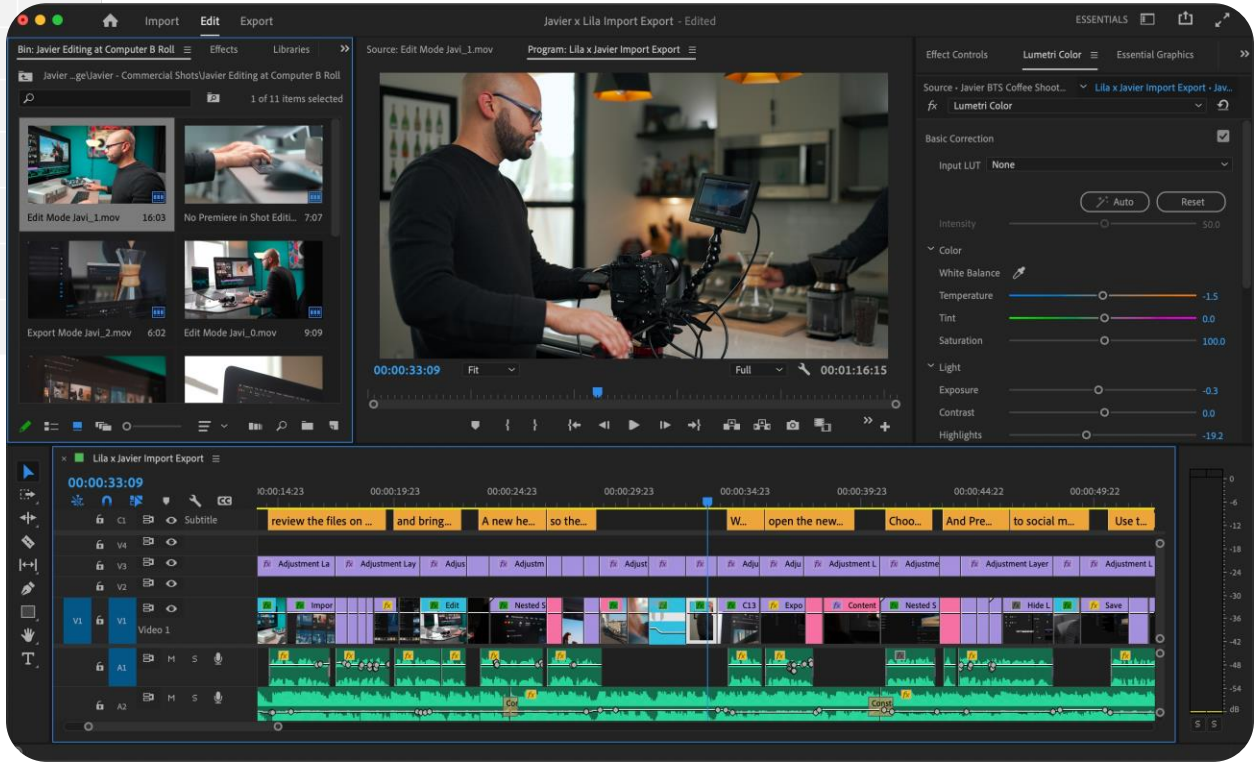
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect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

Interactable intermediate outputs

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

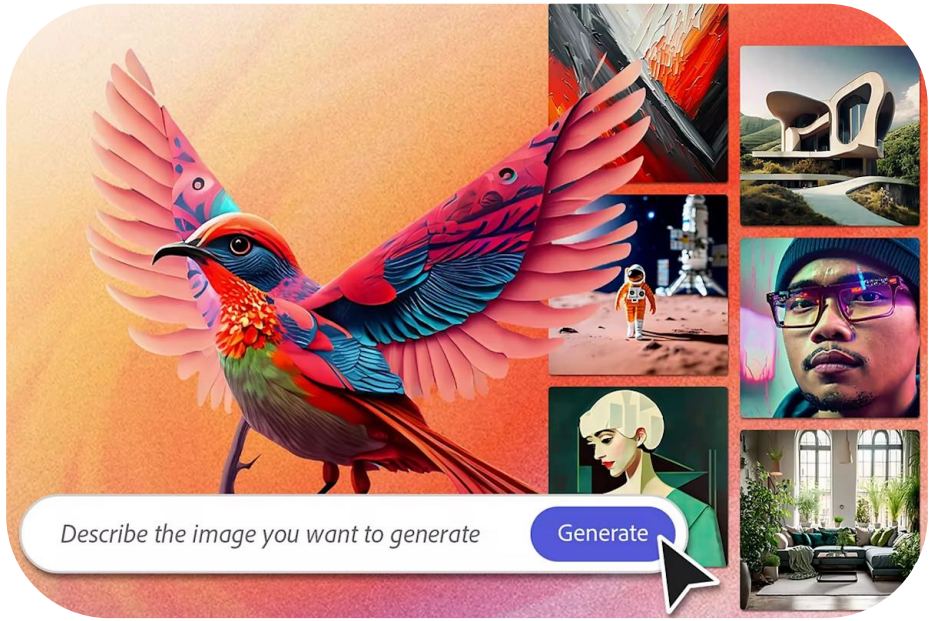
Controllable Generative AI

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 …	

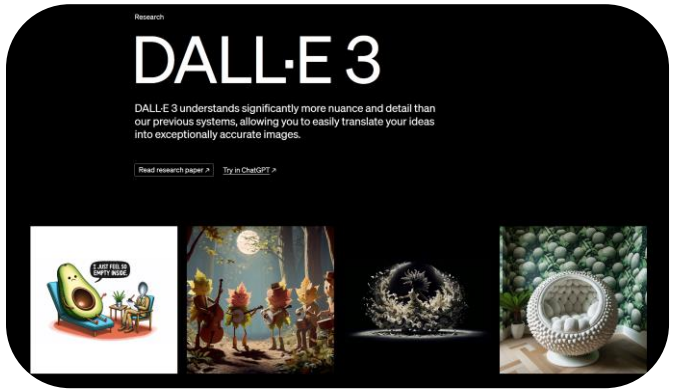


Integration into professional creative workflow

Licensing Training Data for Generative AI



Adobe + Adobe Stock



OpenAI + shutterstock

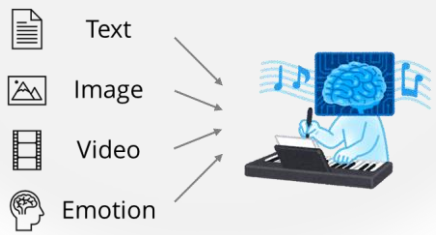
NVIDIA + gettyimages

Attributing AI-Generated Content

(Source: Wang et al., 2023)

Wang et al., "Evaluating Data Attribution for Text-to-Image Models," *ICCV*, 2023
Barnett et al., "Exploring Musical Roots: Applying Audio Embeddings to Empower Influence Attribution for a Generative Music Model," *arXiv preprint arXiv:2401.14542*, 2024.

Multimodal generative AI with music and audio



How can AI help professionals and amateurs create music and audio content?

Interactive AI tools for music & audio production

AI

Music & Audio

Human-like machine learning algorithms for music

How can AI learn to create music and audio like how humans do?

Interactive human-AI music co-creation

Acknowledgements

Thank You to My Advisors!

Thank You to All My Collaborators!

Thank You to My Internship Mentors

Dolby

SONY

amazon

Adobe

NVIDIA

Thank You to My Labmates!

BERG LAB

McAuley Lab

Thank You to My Family!

Thank You for Supporting My PhD Study

Fellowships

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Department of Electrical and Computer Engineering

J. Yang and Family Foundation

MINISTRY OF EDUCATION 教育部

Internships

Dolby

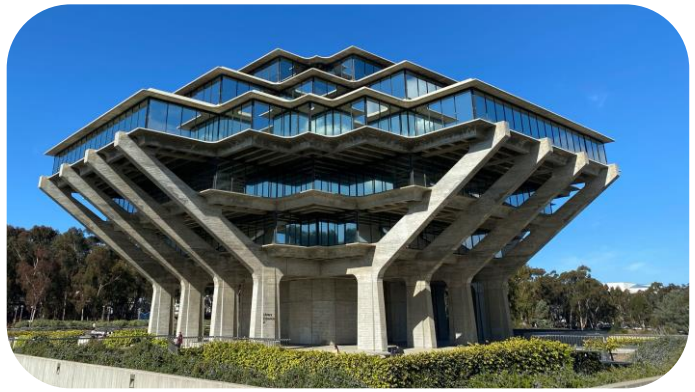
SONY

amazon

Adobe

NVIDIA

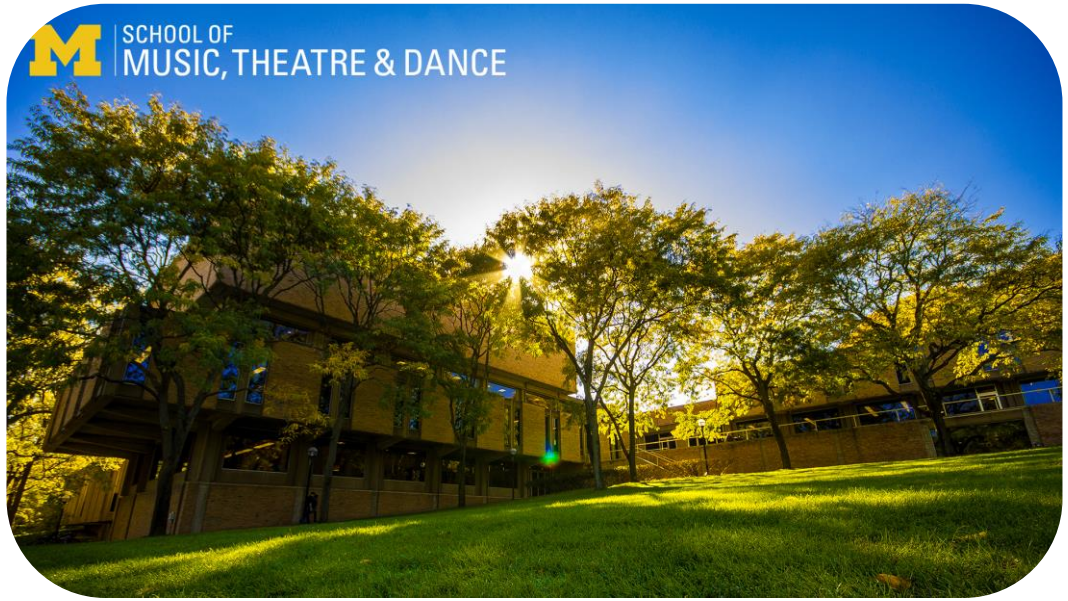
Thank you, UCSD!



UC San Diego

Next Chapter!

Incoming Assistant Professor
Department of Performing Arts Technology
School of Music, Theatre & Dance
University of Michigan



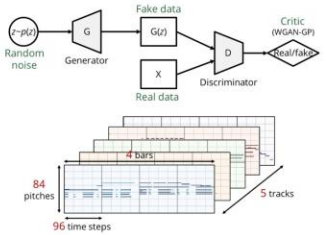
Generative AI for Music & Audio

Empowering music and audio creation with machine learning

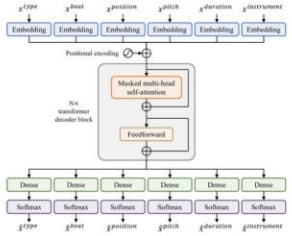
Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)



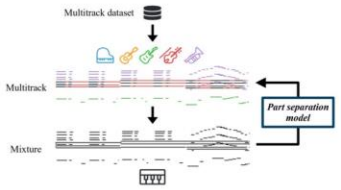
MMT
(ICASSP 2023)



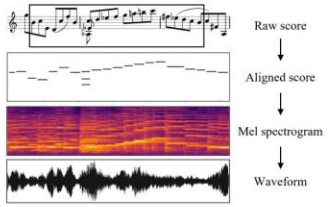
Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

Arranger
(ISMIR 2021)



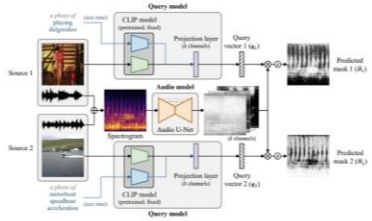
Deep Performer
(ICASSP 2022)



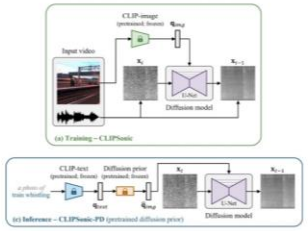
Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



CLIPSonic
(WASPAA 2023)



hwdong@ucsd.edu hermandong.com

UC San Diego