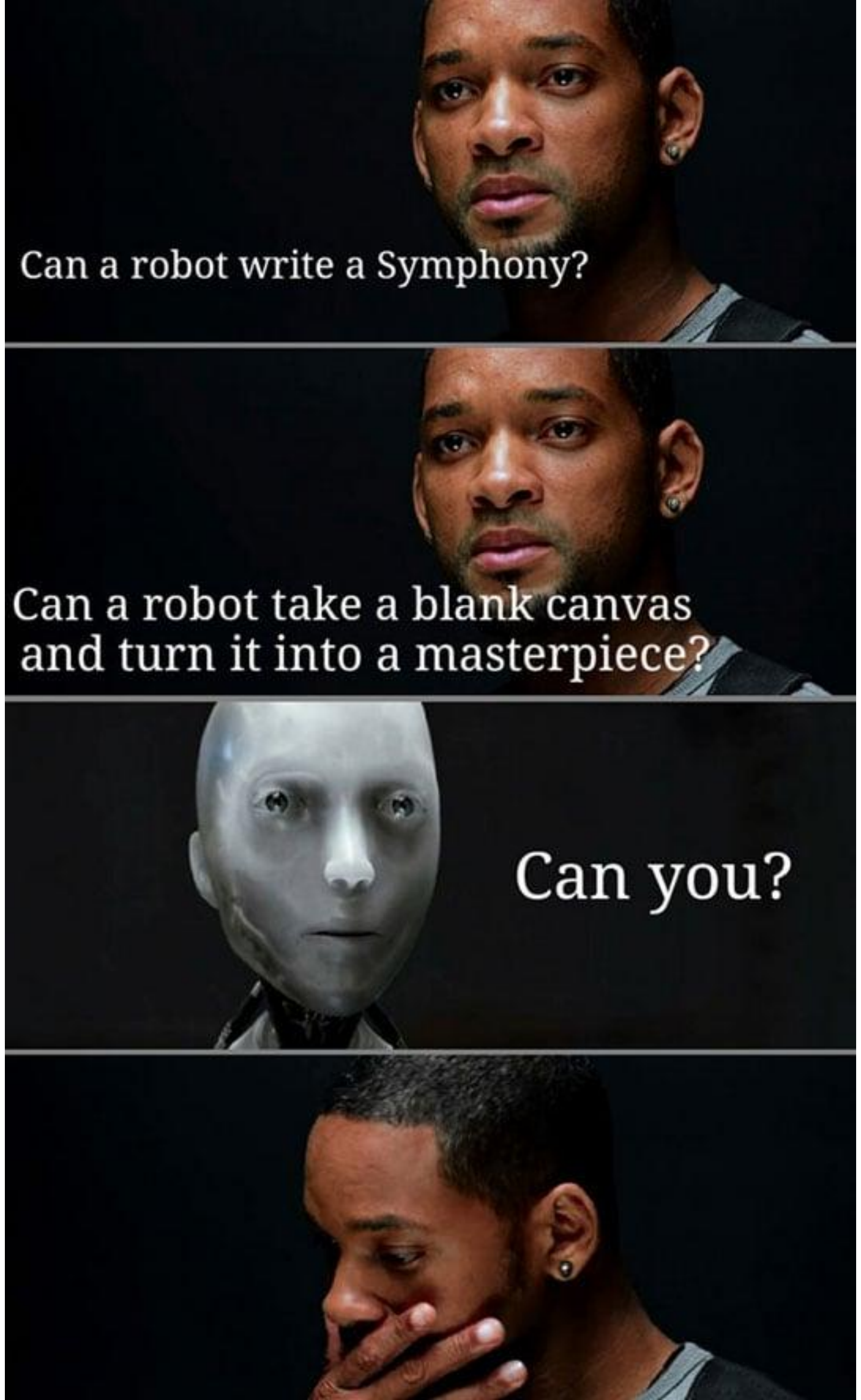
Generative AI for Music: Challenges & Opportunities

Hao-Wen (Herman) Dong

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

March 3, 2025

Can you? (I, Robot, 2004)

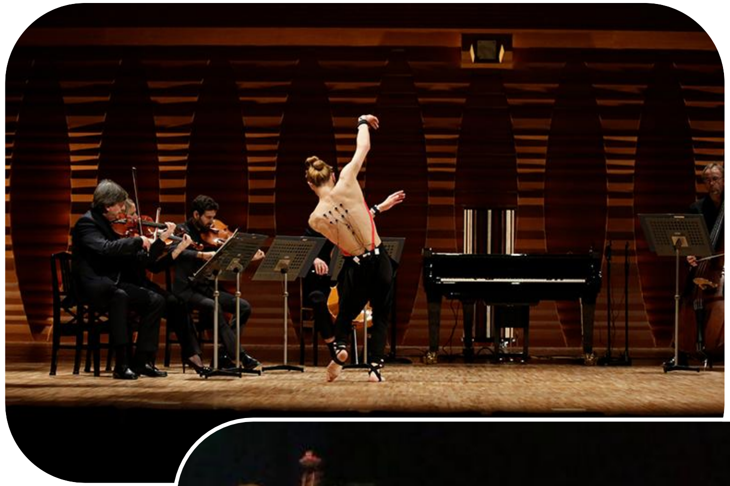


Music & Technology Co-evolves

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: Robot Gizmos)

(Source: Sankei Shimbun)

(Shlizerman et al., 2019)

(Source: NBC DFW)

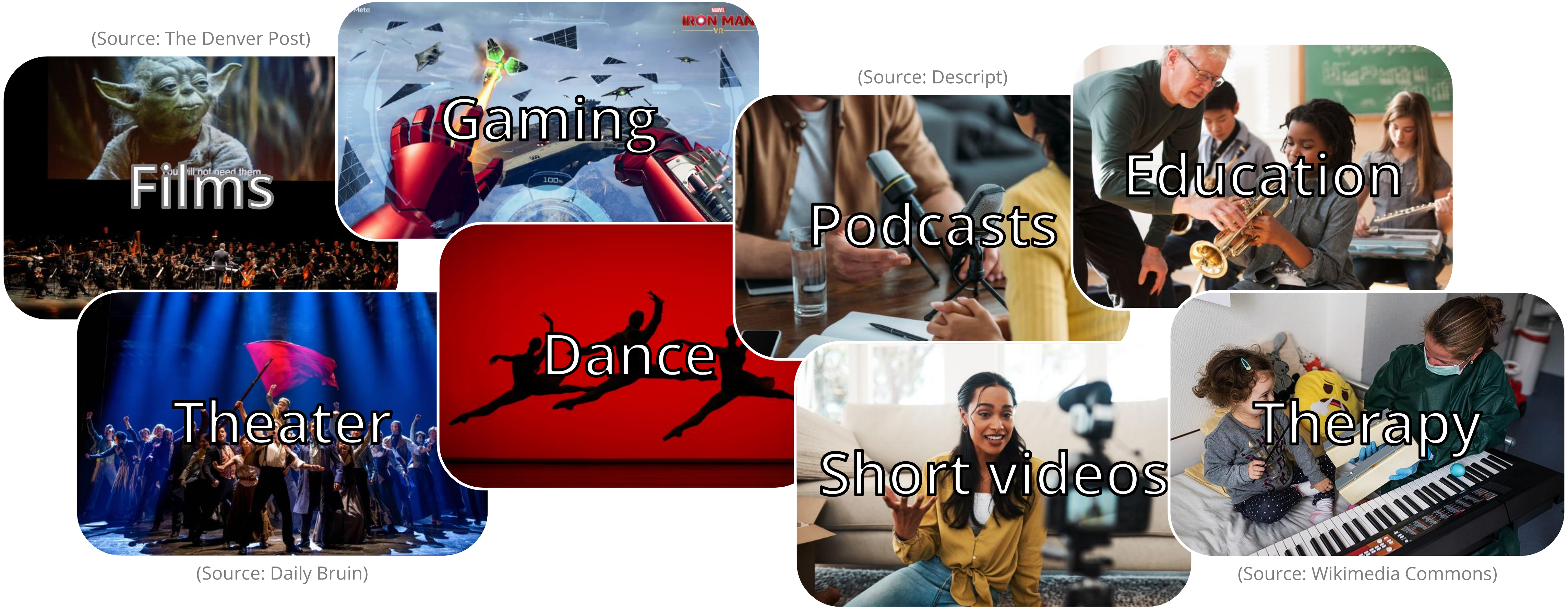
yamaha.com/en/news_release/2018/18013101/
sankei.com/article/20240113-CQCOSQHJWFIYPJJKZDCITRTRVI/
roboticgizmos.com/shimon-musical-robot-deep-learning/
nbcdfw.com/entertainment/the-scene/how-verdigris-ensemble-is-using-ai-to-create-a-new-concert-experience/3366031/
Shlizerman et al., “Audio to Body Dynamics,”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

Generative AI for Content Creation

(Source: UploadVR)

(Source: The Denver Post)

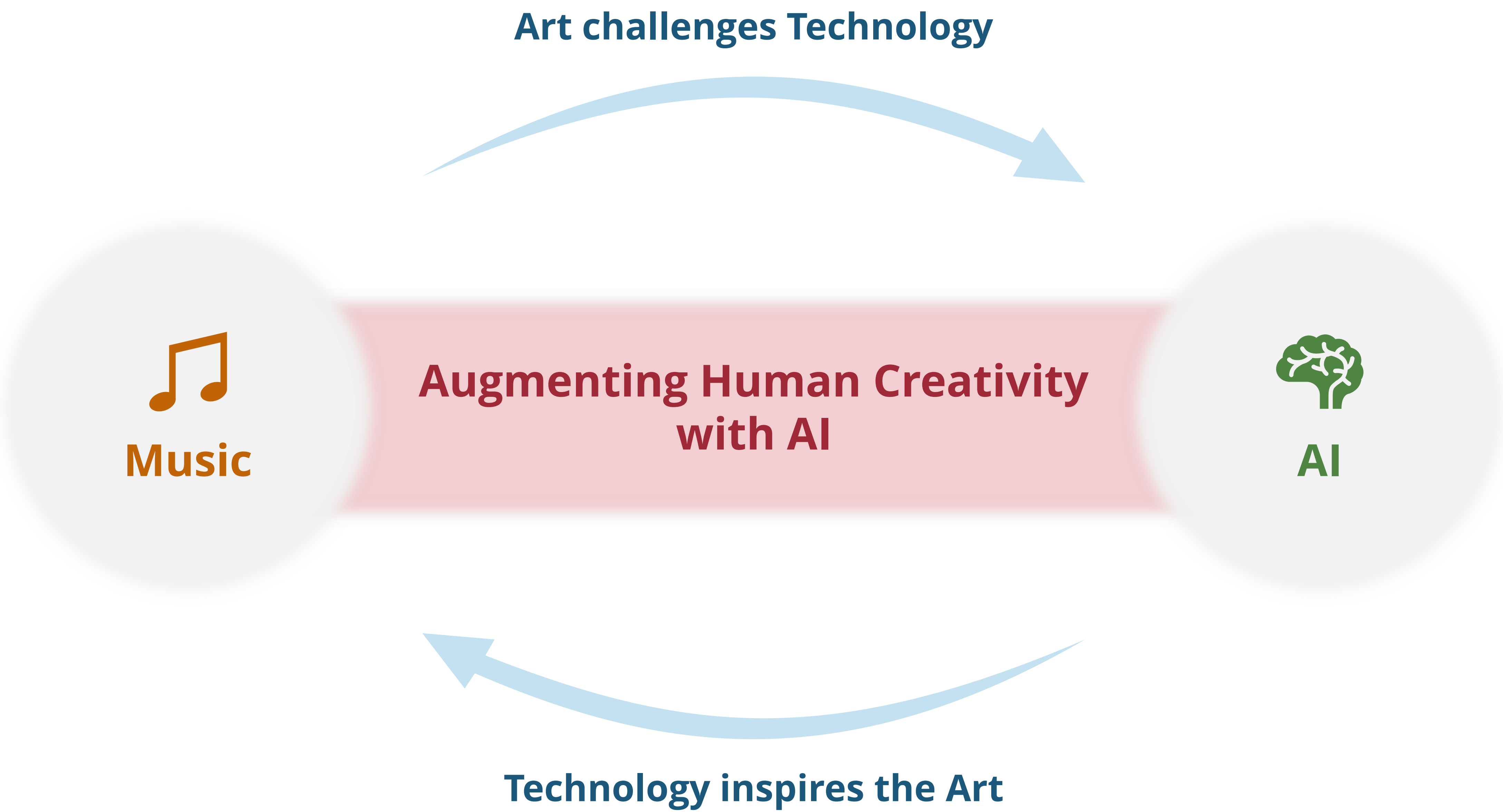
(Source: Descript)



(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


Art challenges Technology
Music
Augmenting Human Creativity with AI
AI
Technology inspires the Art

My Research on AI for Music

- **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)
- **Expressive violin performance synthesis** (ICASSP 2022, ICASSP 2025)
- **Music instrumentation** (ISMIR 2021)
- **Music harmonization** (JNMR 2020)
- **Music LLM** (NLP4MusA 2024, ICASSP 2025)
- **Choral music separation** (ISMIR 2022)
- **Optical music recognition** (ISMIR 2021)

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

Multitrack Music Transformer

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

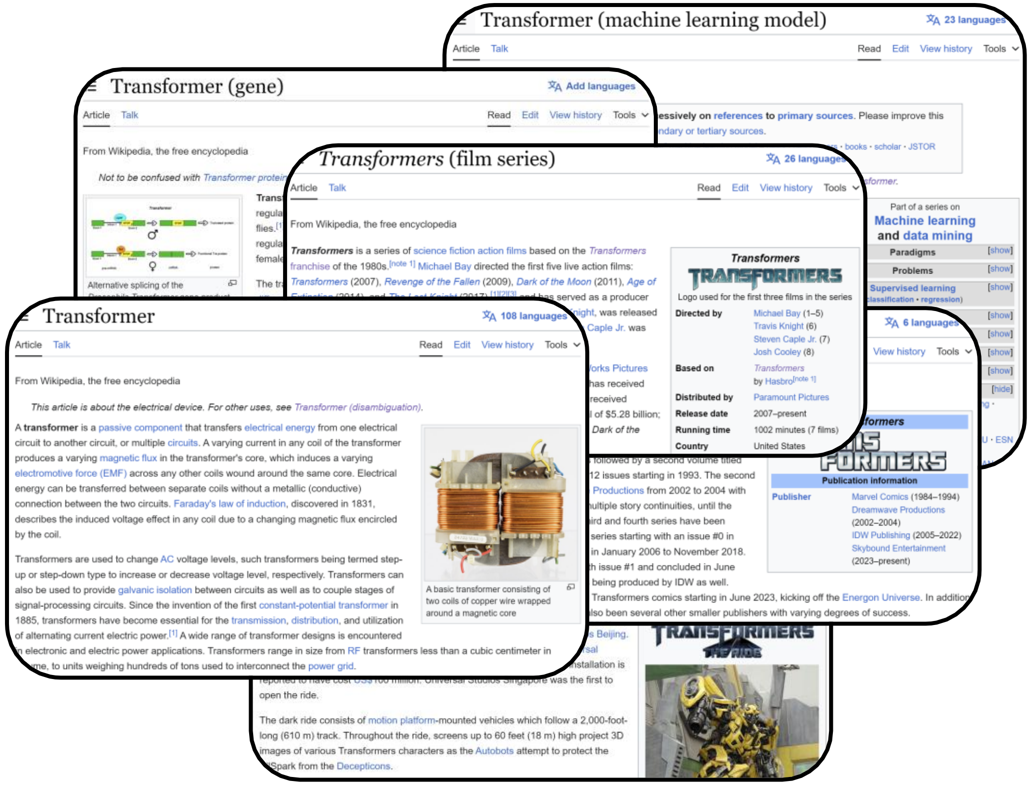
University of California San Diego

UC San Diego

Generating Text using Language Models

- Predicting the next word given the past sequence of words



A transformer is a _________

electrical device ↑
fiction character ↑
deep learning model ↑
family of genes ↑
type of food ↓
musical instrument ↓

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**"

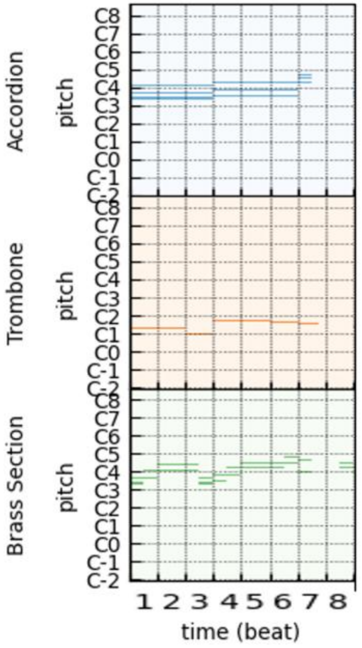
$$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$$

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

$$\mathbf{x}_i = (\underbrace{x_i^{\mathit{type}}}, \underbrace{x_i^{\mathit{beat}}, x_i^{\mathit{position}}, x_i^{\mathit{pitch}}, x_i^{\mathit{duration}}, x_i^{\mathit{instrument}}}_{\text{Specify note \& instrument information}})$$

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

An Example of the Proposed Representation



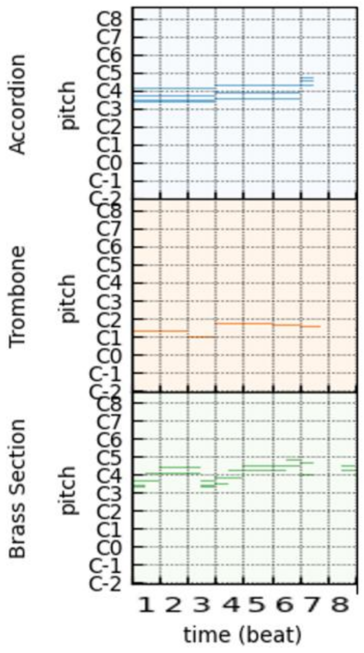
Structural events

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



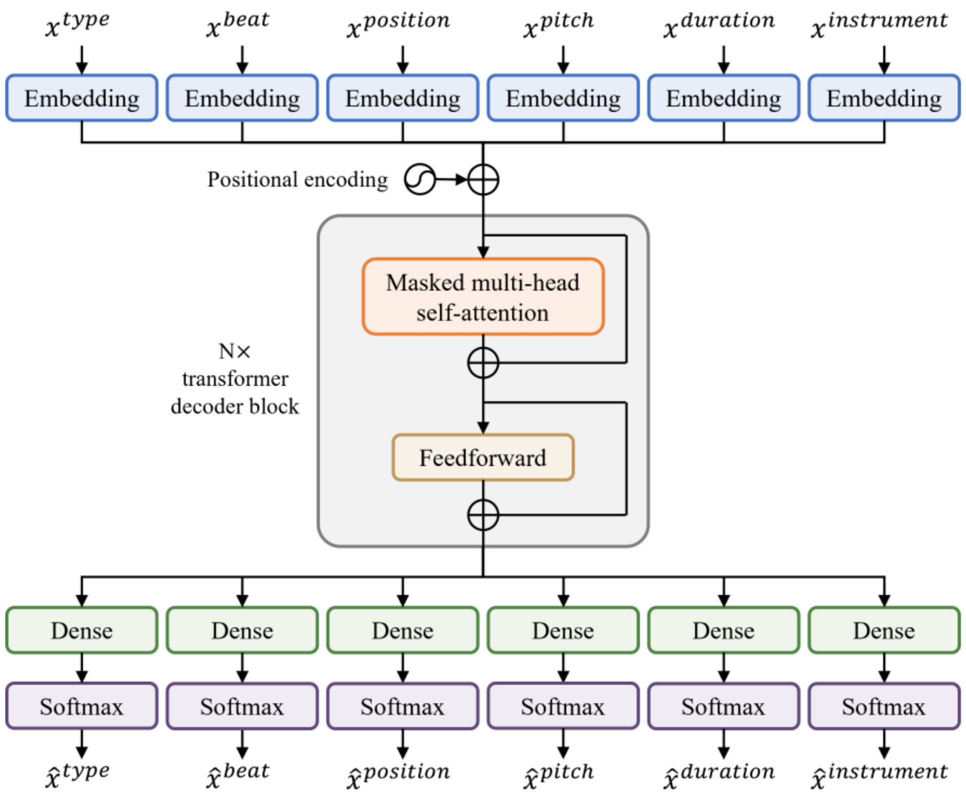

```
(0, 0,  0,  0,  0,  0)
(1, 0,  0,  0,  0, 15)
(1, 0,  0,  0,  0, 36)
(1, 0,  0,  0,  0, 39)
(2, 0,  0,  0,  0,  0)
(3, 1,  1, 41, 15, 36)
(3, 1,  1, 55,  4, 39)
(3, 1,  1, 55, 17, 15)
(3, 1,  1, 58,  4, 39)
(3, 1,  1, 58, 17, 15)
(3, 1,  1, 73, 17, 15)
(3, 1, 13, 58,  4, 39)
(3, 1, 13, 73,  4, 39)
(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

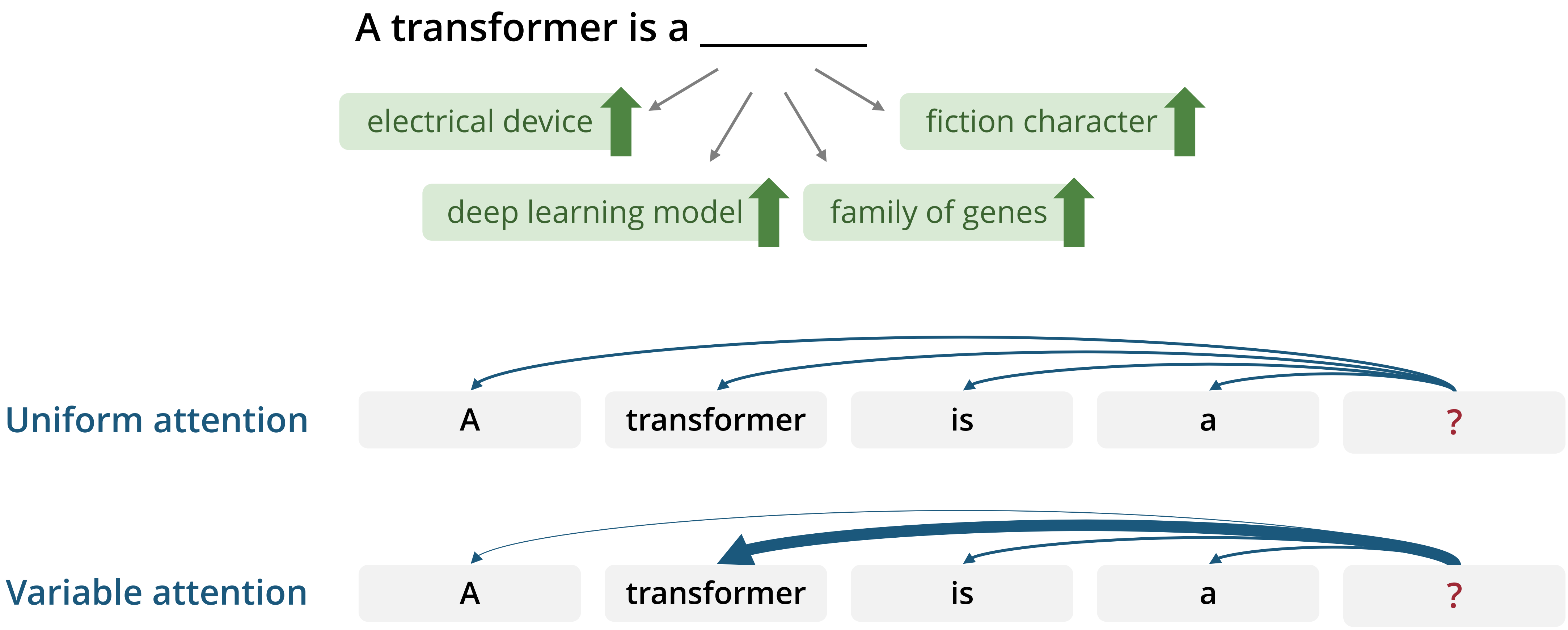
church-organ, viola, contrabass, strings, voices, horn, oboe

4-beat continuation

Mozart's
Eine kleine Nachtmusik

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

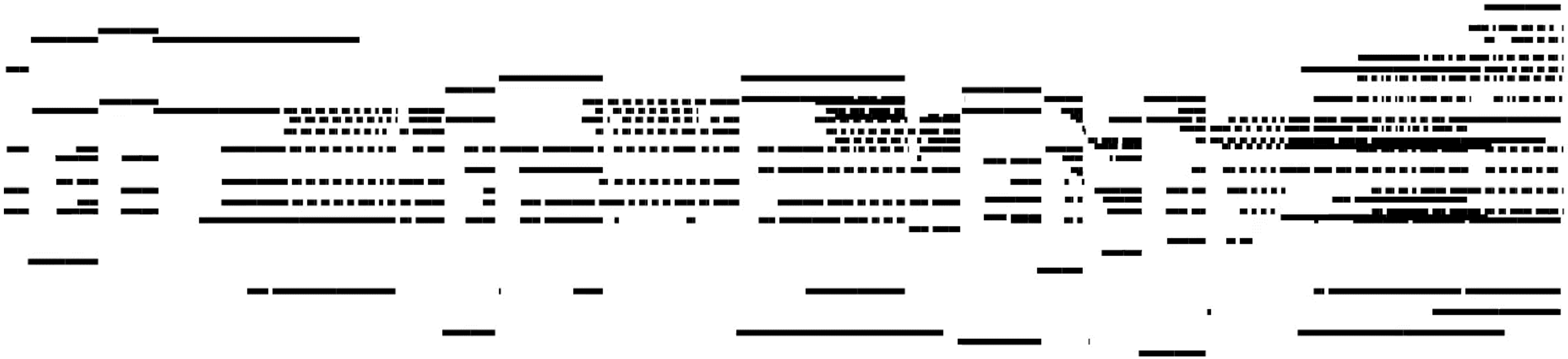
The Magic of Transformers – Self-attention Mechanism



Transformers learn what to attend to from big data!

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

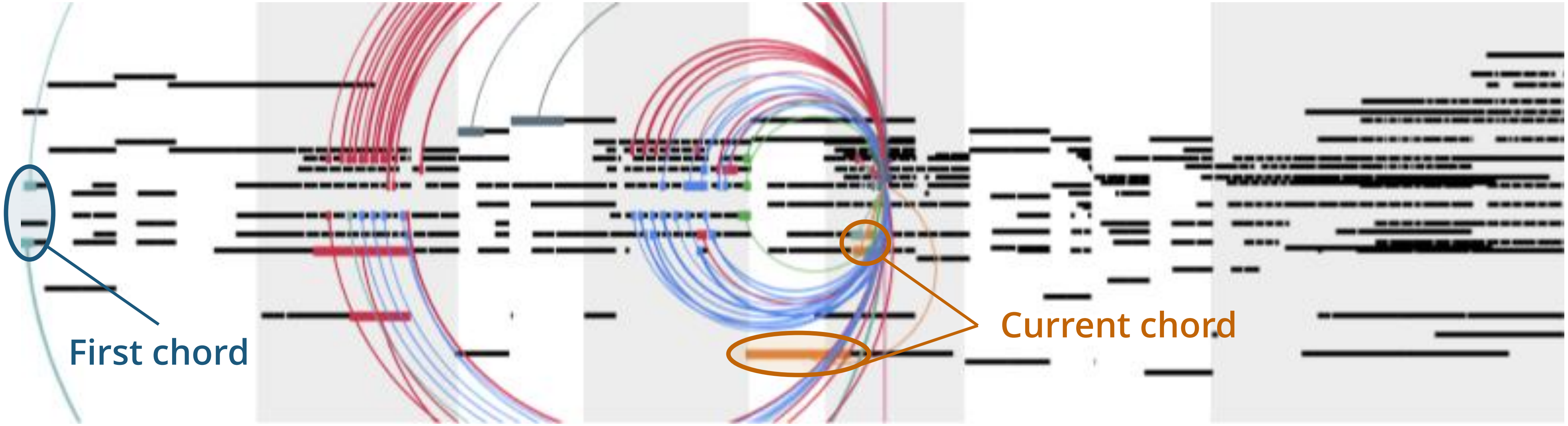
(Each color represents an attention head)



(Source: Huang et al., 2018)

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?

Systematically Analyzing Musical Self-attention

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

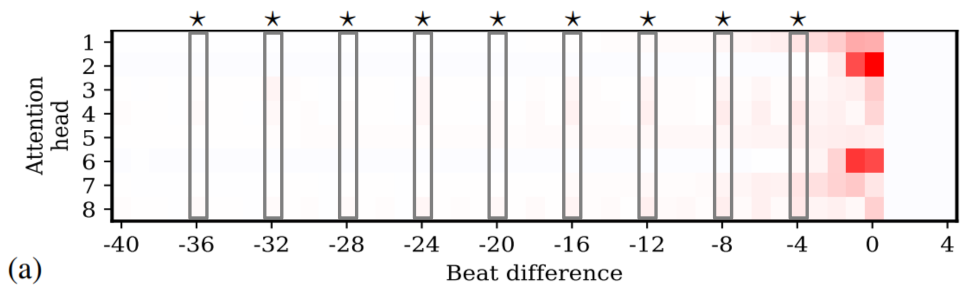
$$\gamma_k^{(d)} = \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})\, \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})} \qquad \tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{\mathbf{x}\in\mathcal{D}} \sum_{s>t} \mathbb{1}_{x_t^{(d)}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

- The MMT model attends more to notes

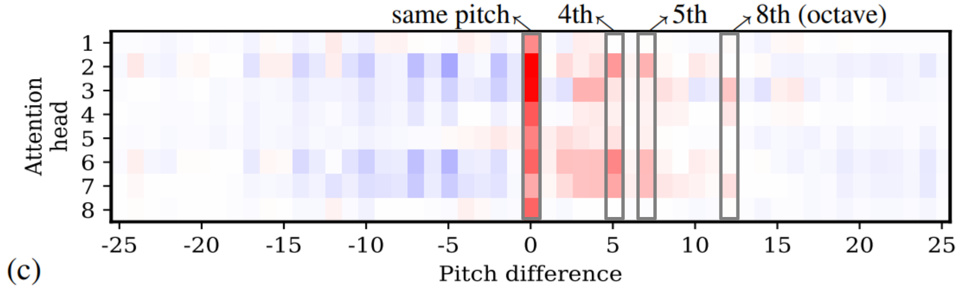
that are $4N$ beats away in the past

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

Positive and negative mean relative attention gain



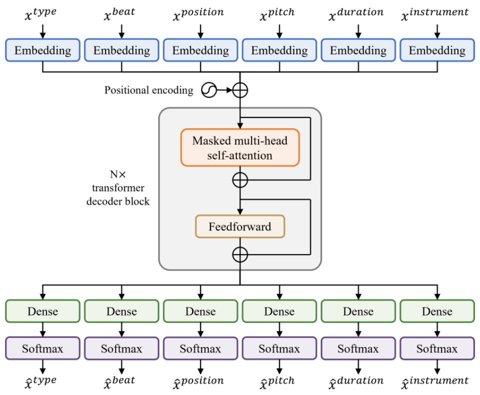
Positive and negative mean relative attention gain



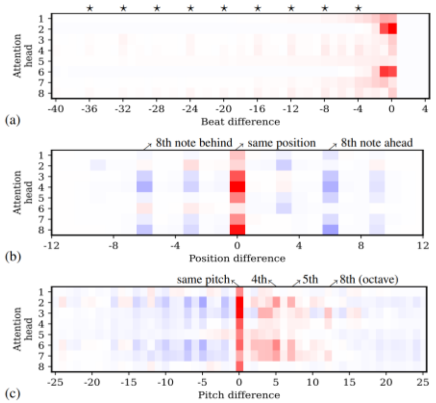
Summary

- **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

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]

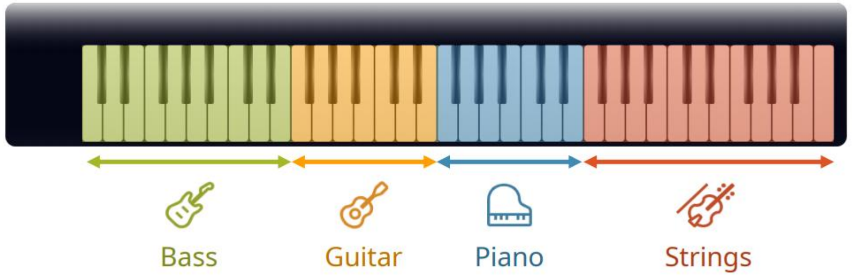
[1] University of California San Diego [2] Stanford University

Automatic Instrumentation

- **Goal**: Dynamically **assign instruments** to notes in solo music

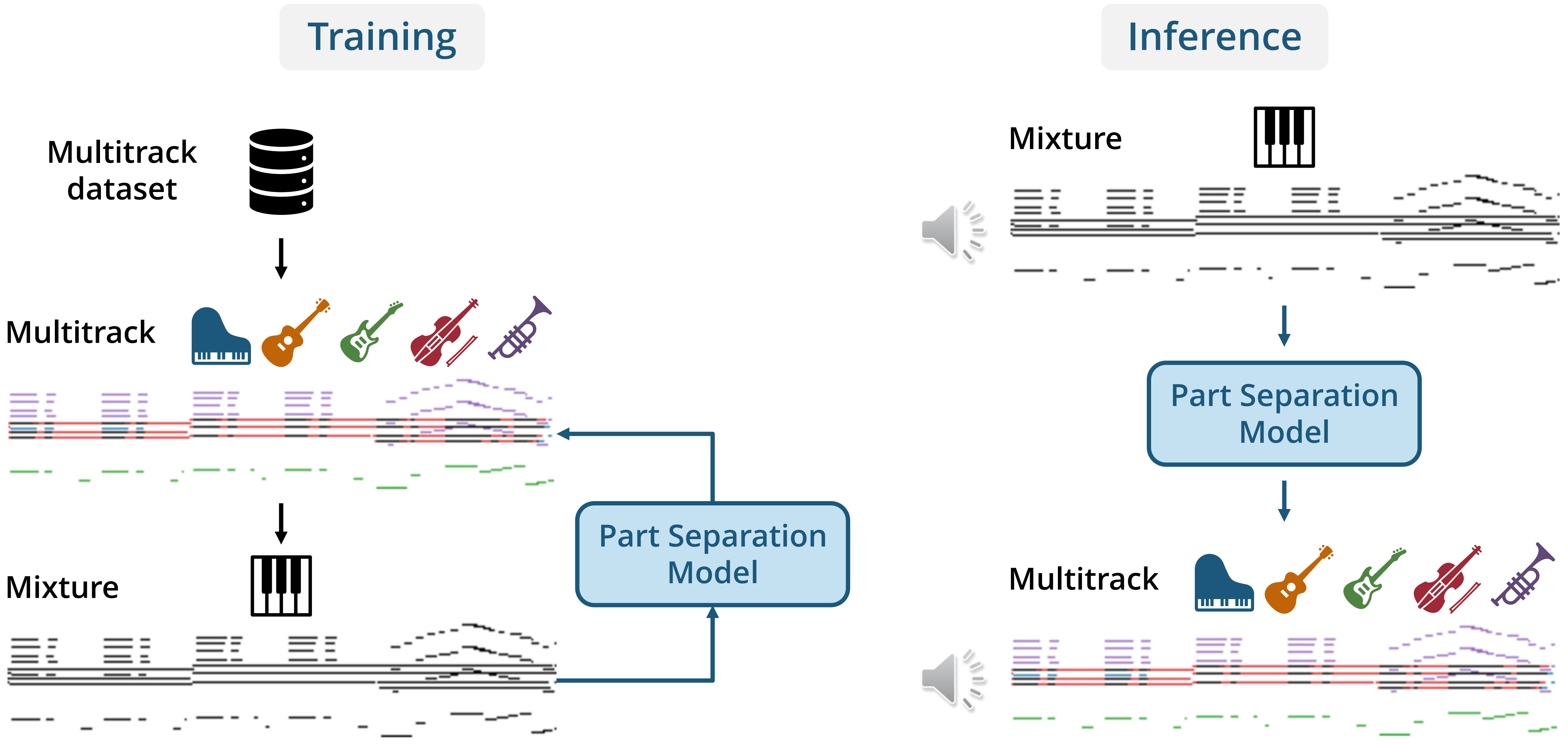
Intelligent musical instruments



Assistive composing tools

How can we acquire paired data?

Learning Automatic Instrumentation without Paired Data

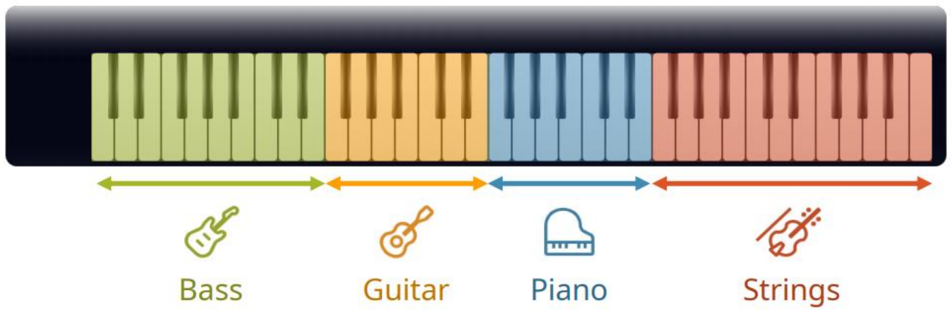


Online vs Offline Models

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

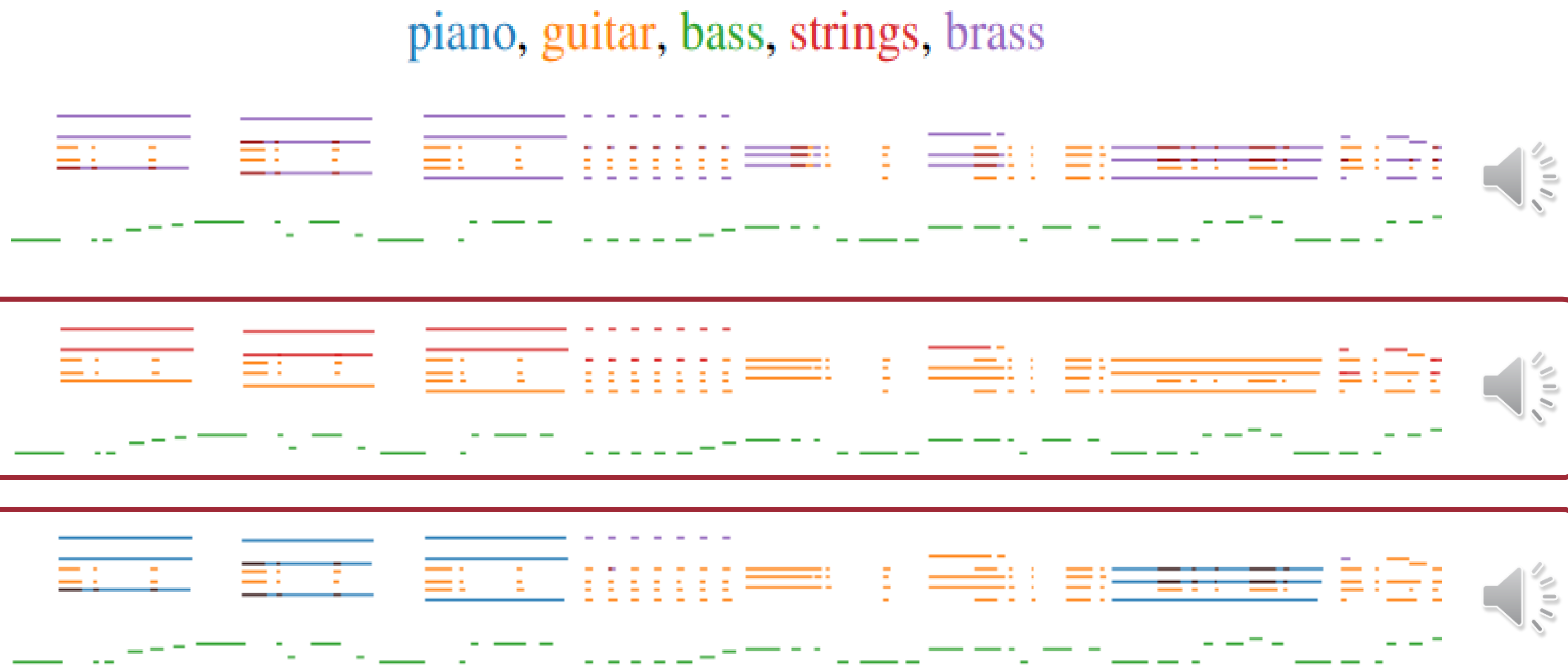
- **Time** Onset time (in time step)
- **Pitch** Pitch as a MIDI note number
- **Duration** Note length (in time step)
- **Frequency** Frequency of the pitch (in Hz)
- **Beat** Onset time (in beat)
- **Position** Position within a beat (in time step)

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



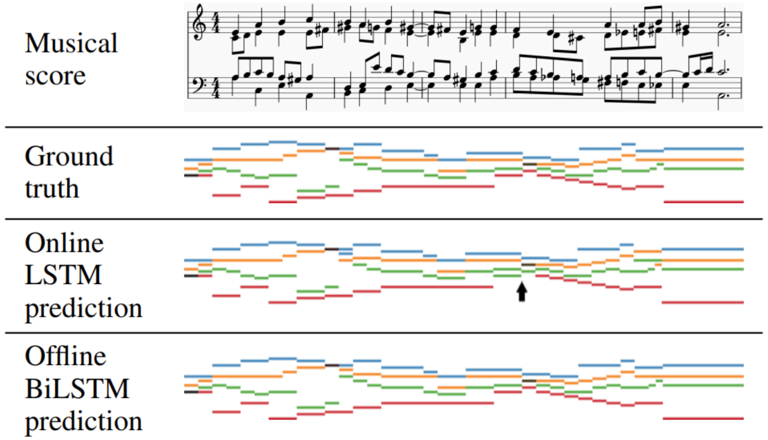
Original

LSTM
(w/o entry hints)

BiLSTM
(w/ entry hints)

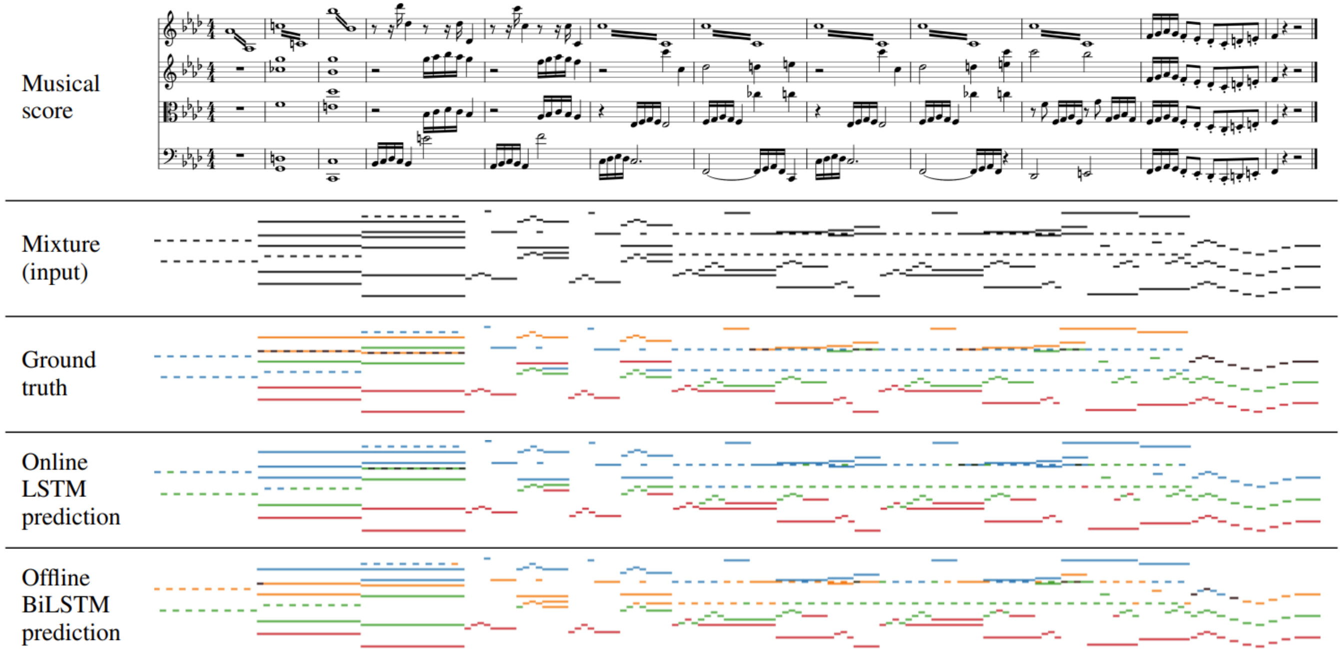
More Results

Bach chorales



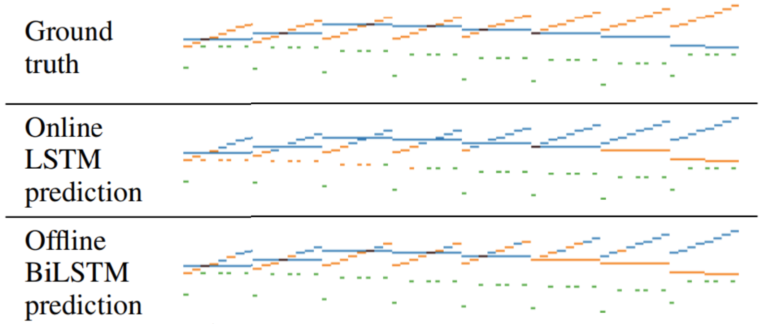
(Audio available. [1] Colors: soprano, alto, tenor, bass.)

String quartets



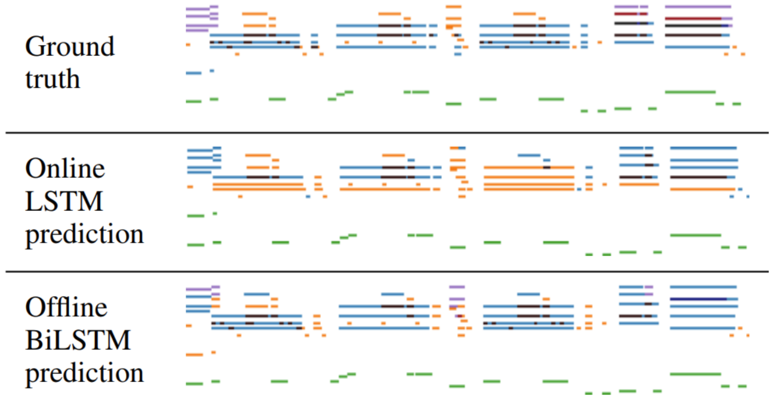
(Audio available. [1] Colors: first violin, second violin, viola, cello.)

Game music



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

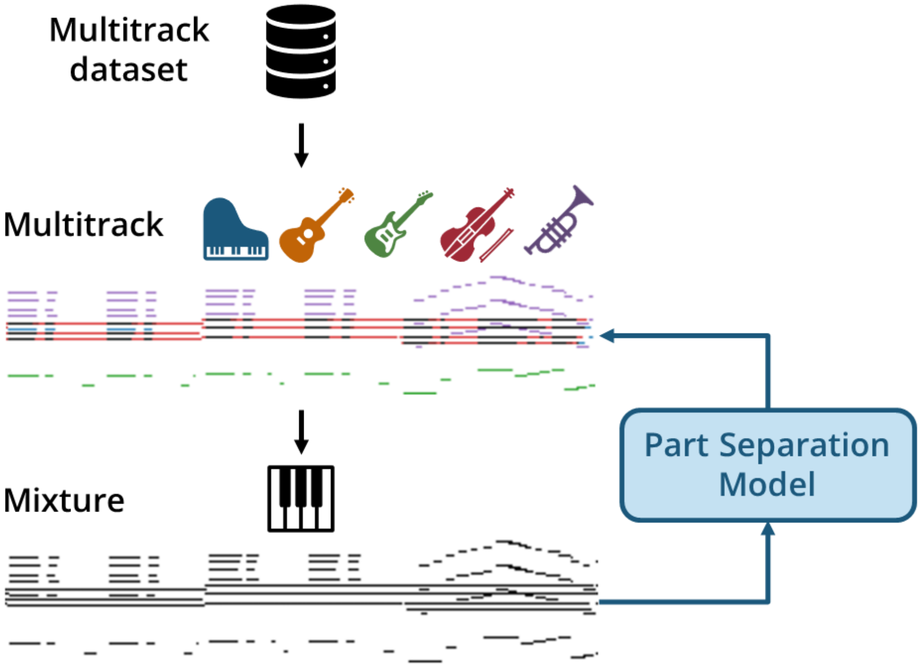
Pop music



(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**



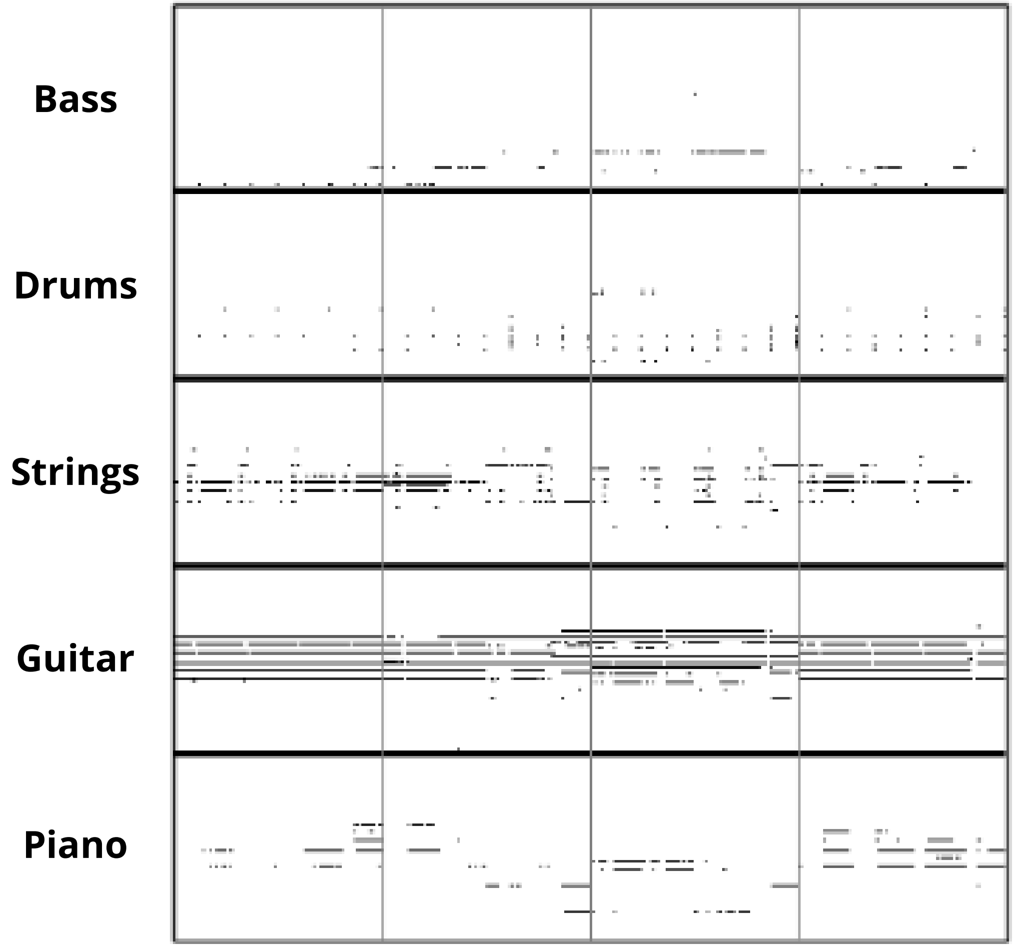
Paper: arxiv.org/abs/2107.05916
Demo: salu133445.github.io/arranger
Code: github.com/salu133445/arranger



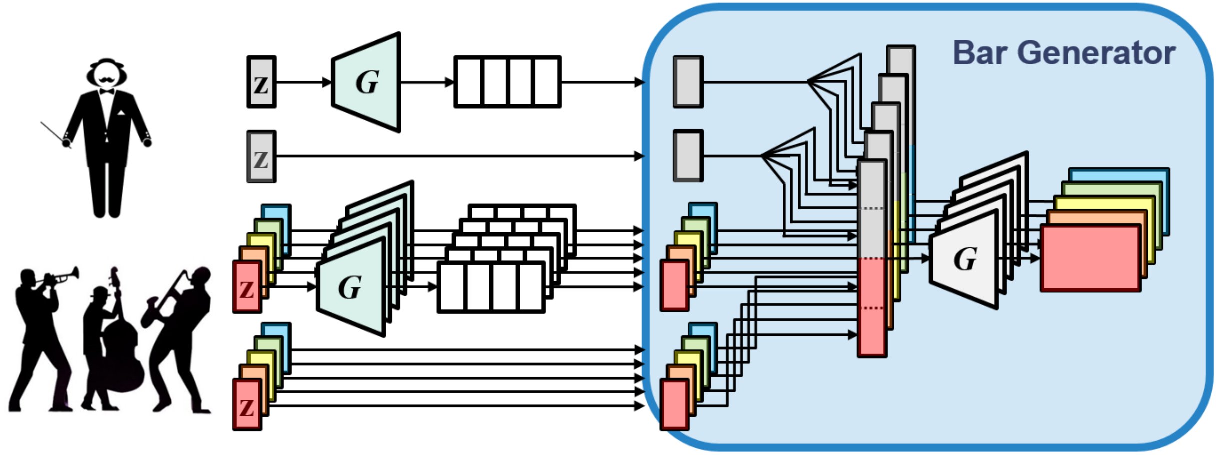

UC San Diego Stanford

Generating Multi-instrument Music using GANs (AAAI 2018)

Multitrack Piano Roll

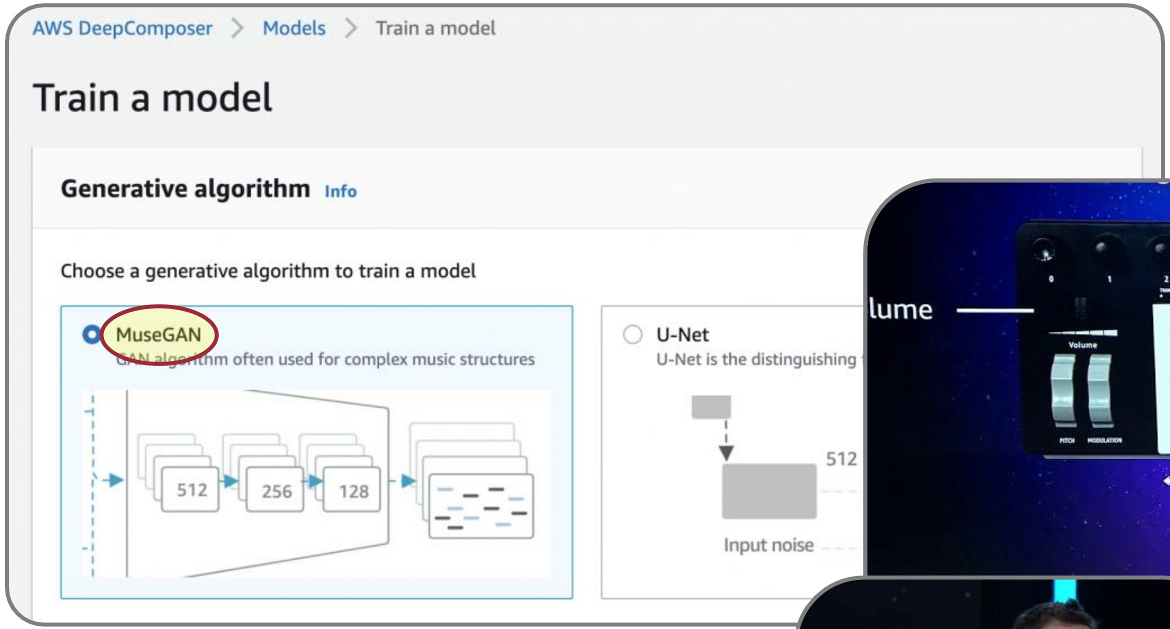


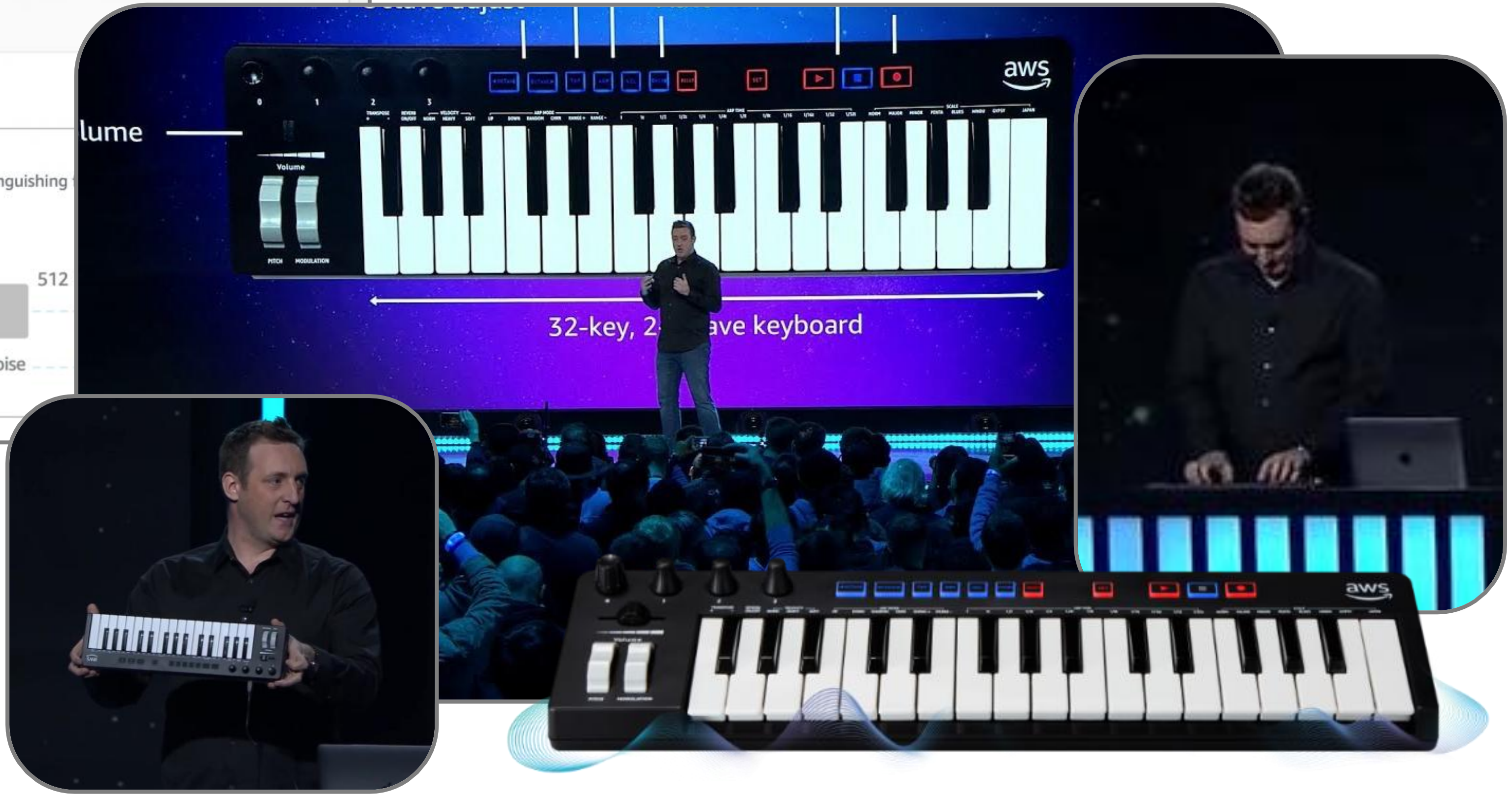
MuseGAN Generator



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)

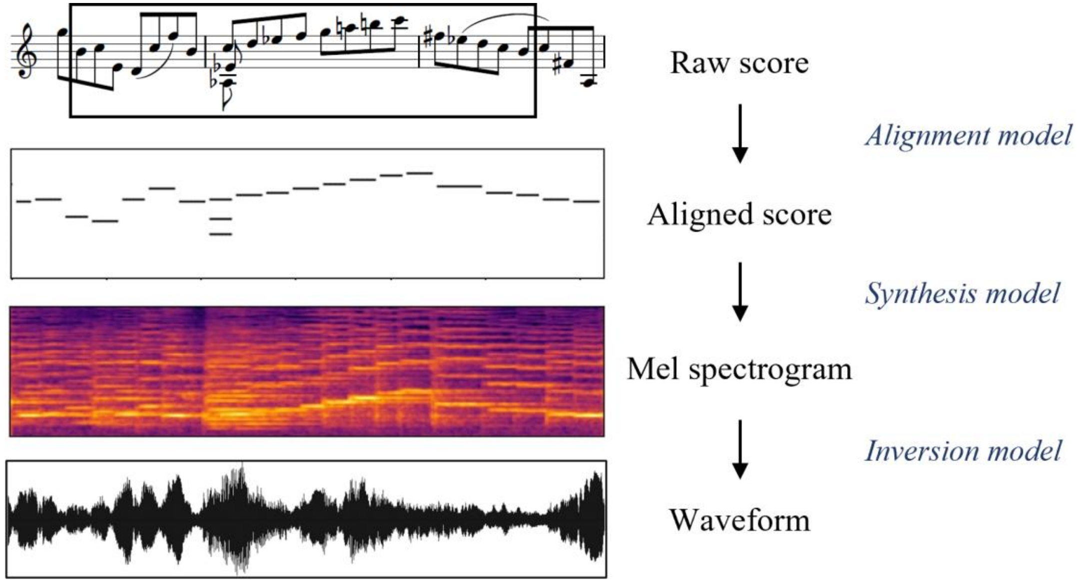




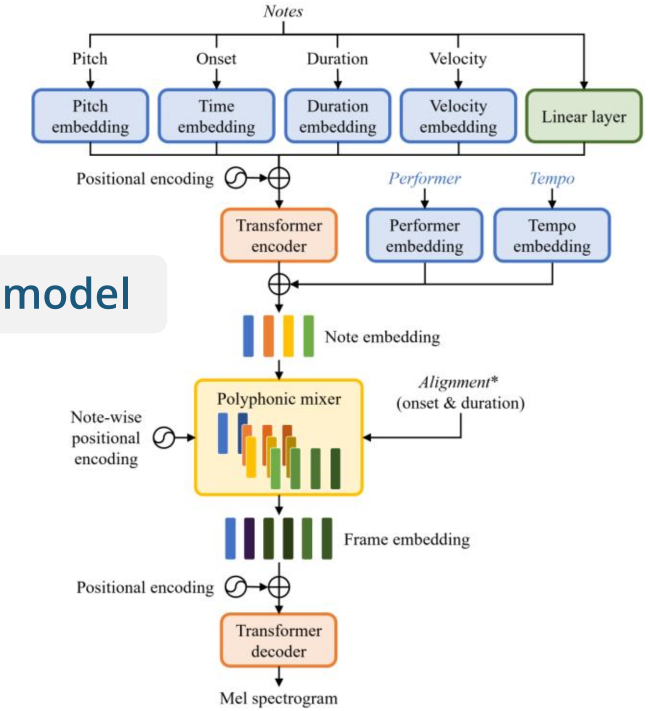
amazon.com/dp/B07YGZ4V5B/
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," *AWS News Blog*, April 2, 2020.

Synthesizing Expressive Violin Performance (ICASSP 2022)

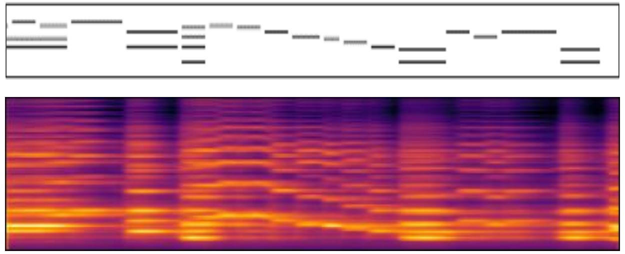
Performance synthesis

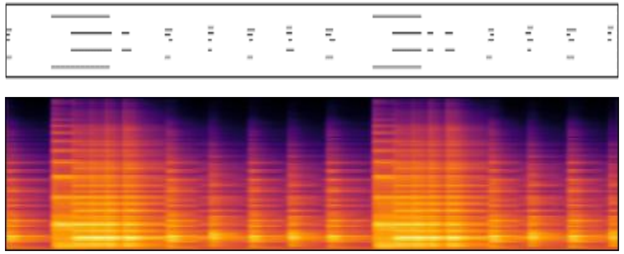


TTS-based model



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.

Challenges & Opportunities

The Five Challenges

- **Representations**
- **Usability**
- **Creativity**
- **Multimodality**
- **Personalization**

Challenge 1: Representations

How can we best represent music for machine learning?

Music Generation – Four Paradigms

Symbolic music generation

Audio-domain music generation

Text-based

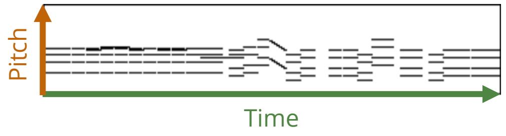
Image-based

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

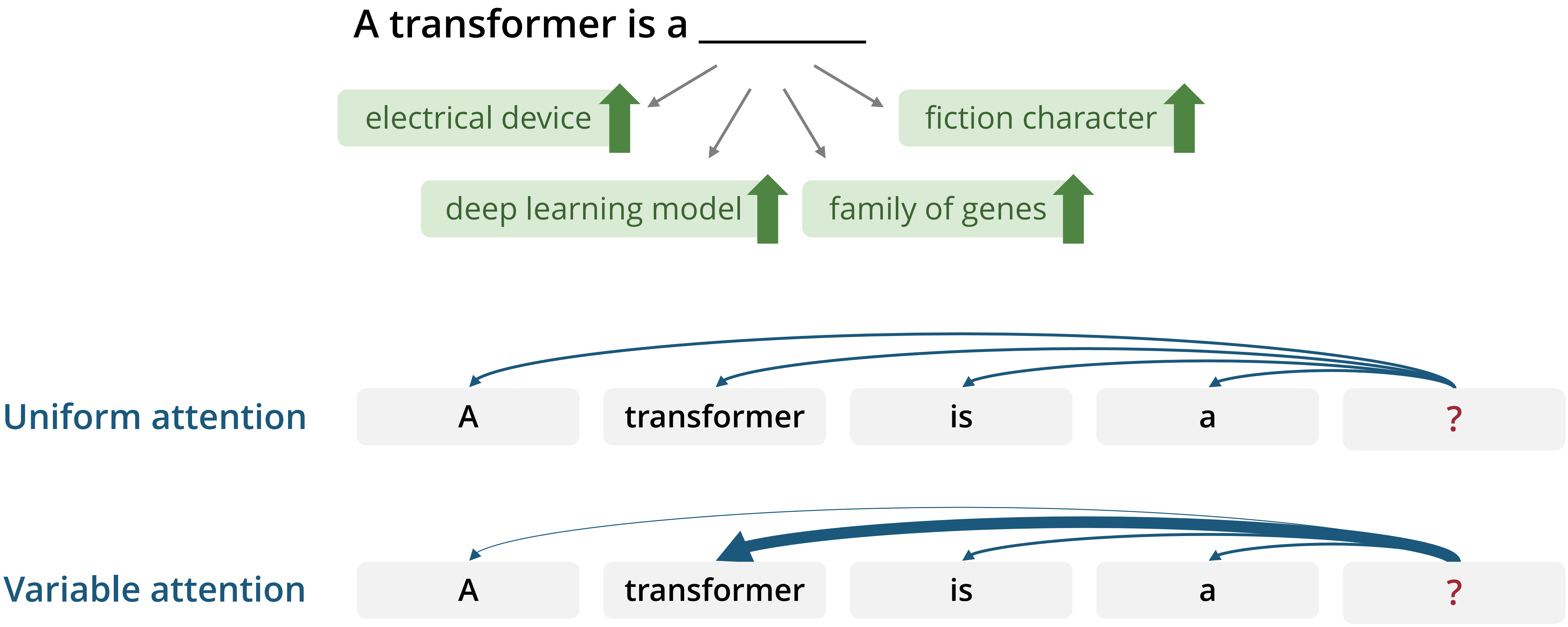


Piano roll

Waveform

Spectrogram

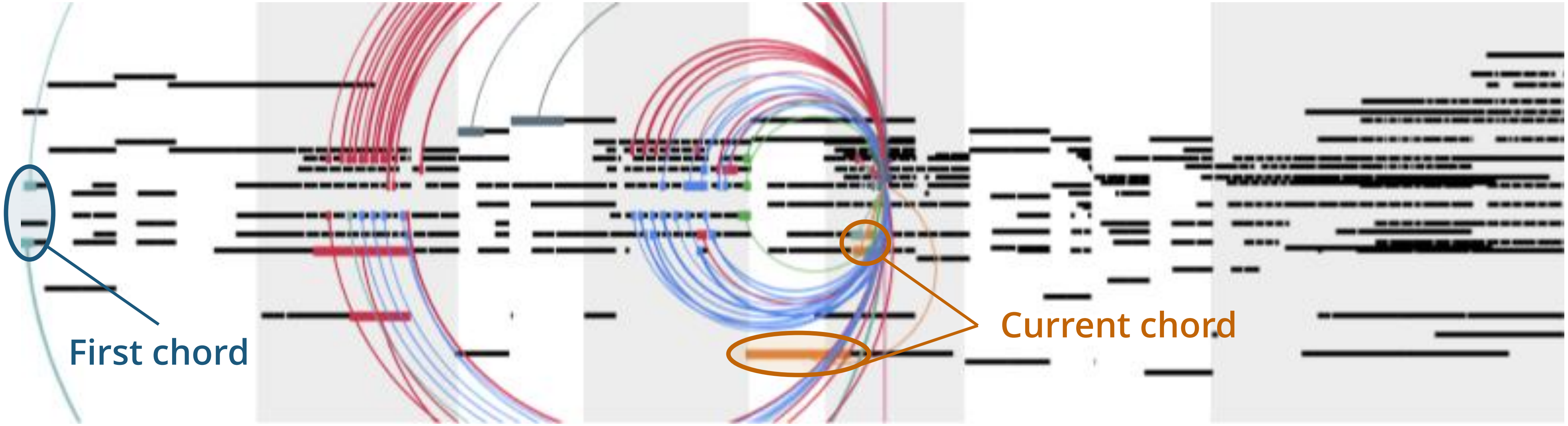
The Magic of Transformers – Self-attention Mechanism



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.

Systematically Analyzing Musical Self-attention

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

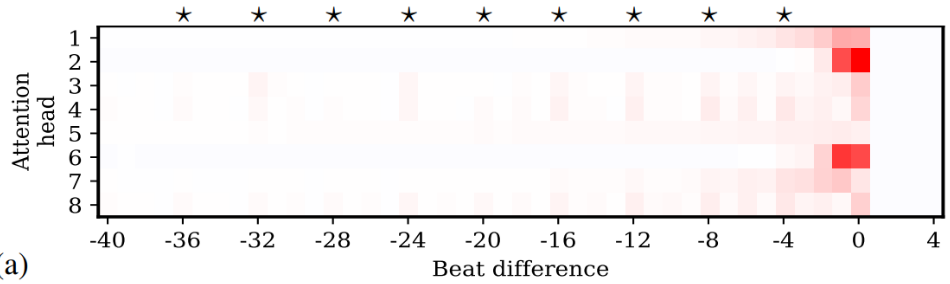
$$\gamma_k^{(d)} = \frac{\sum_{\mathbf{x} \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x}) \, \mathbb{1}_{x_t^{(d)} - x_s^{(d)} = k}}{\sum_{\mathbf{x} \in \mathcal{D}} \sum_{s>t} a_{s,t}(\mathbf{x})} \qquad \tilde{\gamma}_k^{(d)} = \gamma_k^{(d)} - \frac{\sum_{\mathbf{x} \in \mathcal{D}} \sum_{s>t} \mathbb{1}_{x_t^{(d)}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

- The MMT model attends more to notes

that are $4N$ beats away in the past

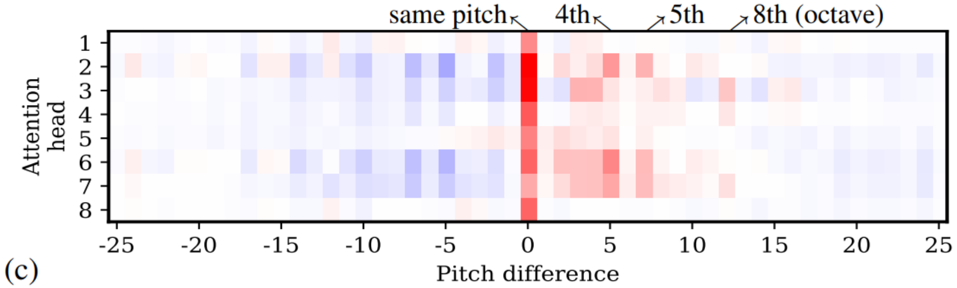
that has a pitch in an octave above which forms a consonant interval

Positive and negative mean relative attention gain



(a)

Positive and negative mean relative attention gain



(c)

Music Generation – Four Paradigms

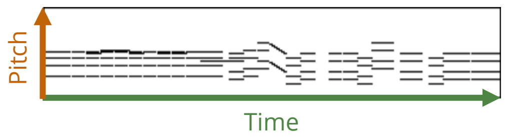
Symbolic music generation

Text-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

Audio-domain music generation

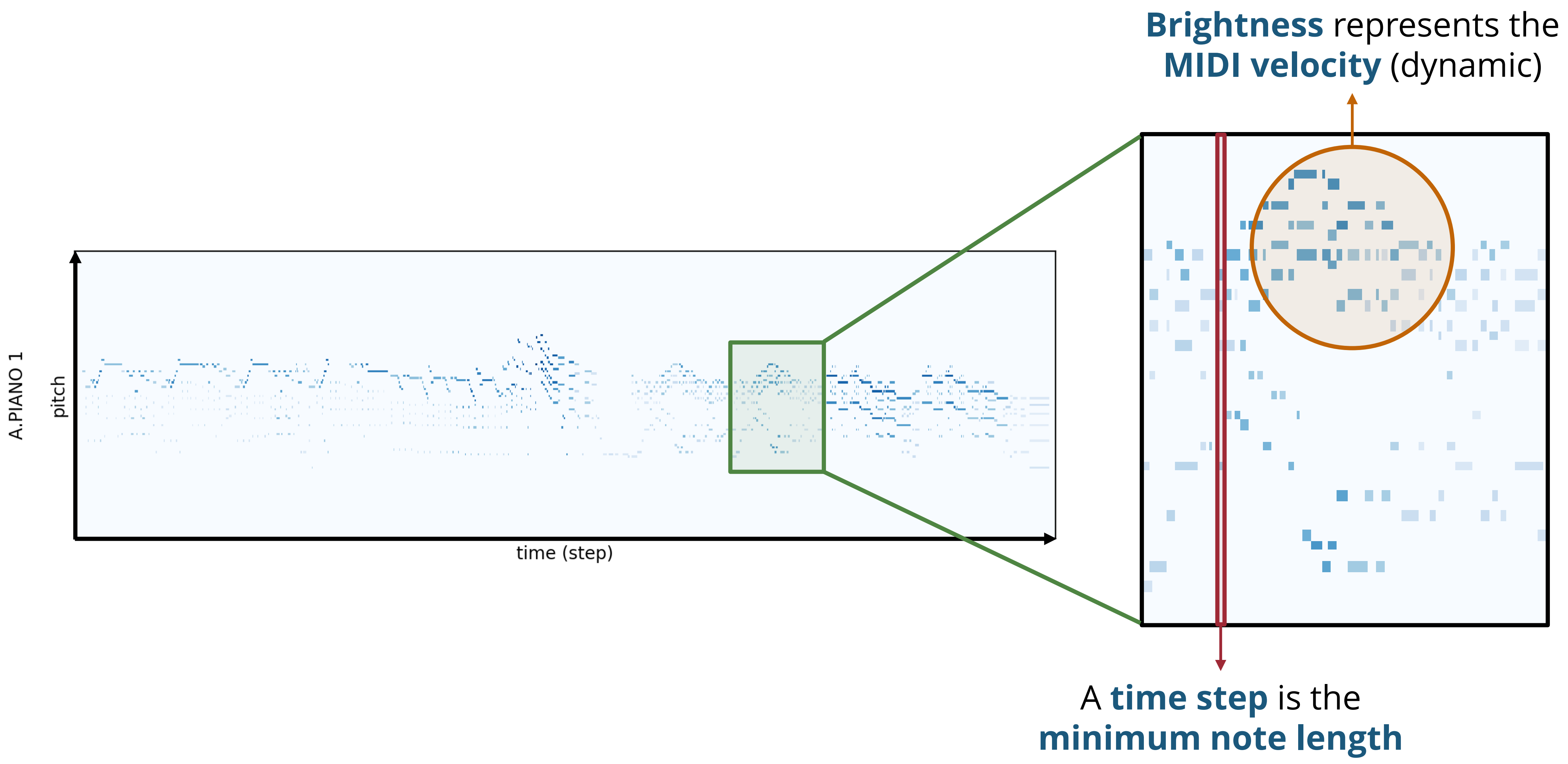
Time series-based

Image-based

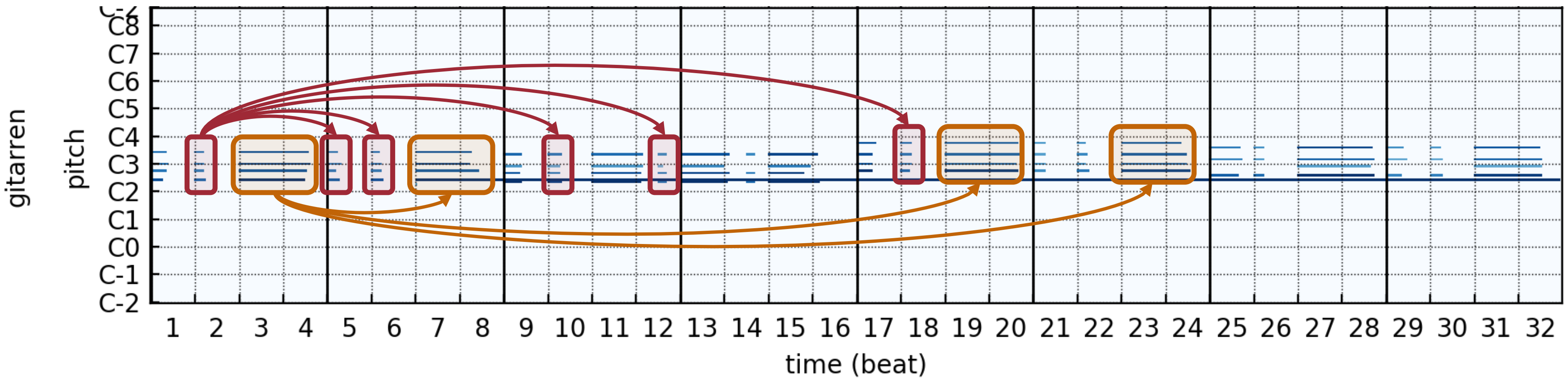
Waveform

Spectrogram

Piano Roll Representation



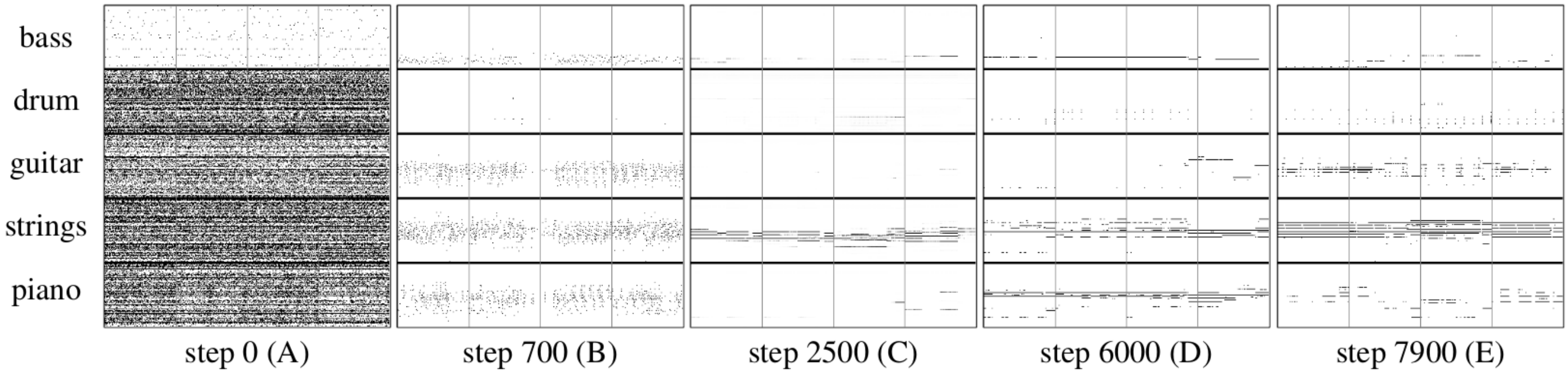
Why Piano Rolls?



Many musical patterns like melodies, chords, scales and arpeggios are translational invariant in the temporal and pitch axes

MuseGAN (Dong et al., 2018)

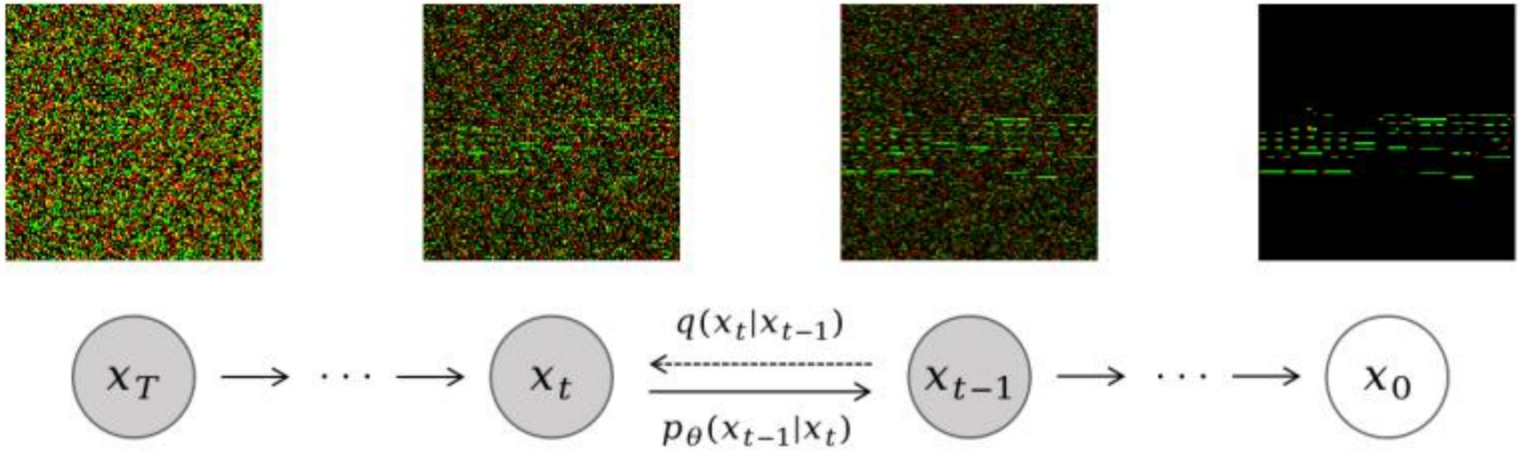
Examples of generated music



(Source: Dong et al., 2018)

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.

Polyffusion (Min et al., 2023)

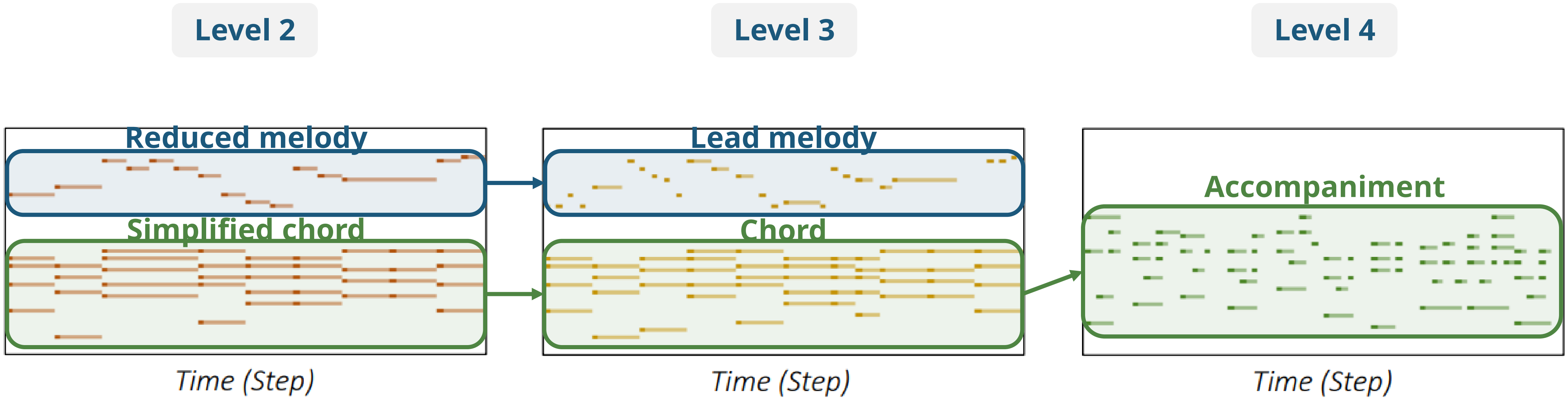


(Source: Min et al., 2023)

polyffusion.github.io

Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao, "Polyffusion: A Diffusion Model for Polyphonic Score Generation with Internal and External Controls," *ISMIR*, 2023.

Example: Cascaded Diffusion Models (Wang et al., 2024)



(Source: Wang et al., 2024)

wholesonggen.github.io

Ziyu Wang, Lejun Min, and Gus Xia, "Whole-Song Hierarchical Generation of Symbolic Music Using Cascaded Diffusion Models," *ICLR*, 2024.

Music Generation – Four Paradigms

Symbolic music generation

Audio-domain music generation

Text-based

Image-based

Time series-based

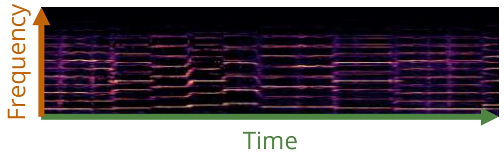
Image-based

```
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

Challenge 1: Representations

How can we best represent music for machine learning?

Challenge 2: Multimodality

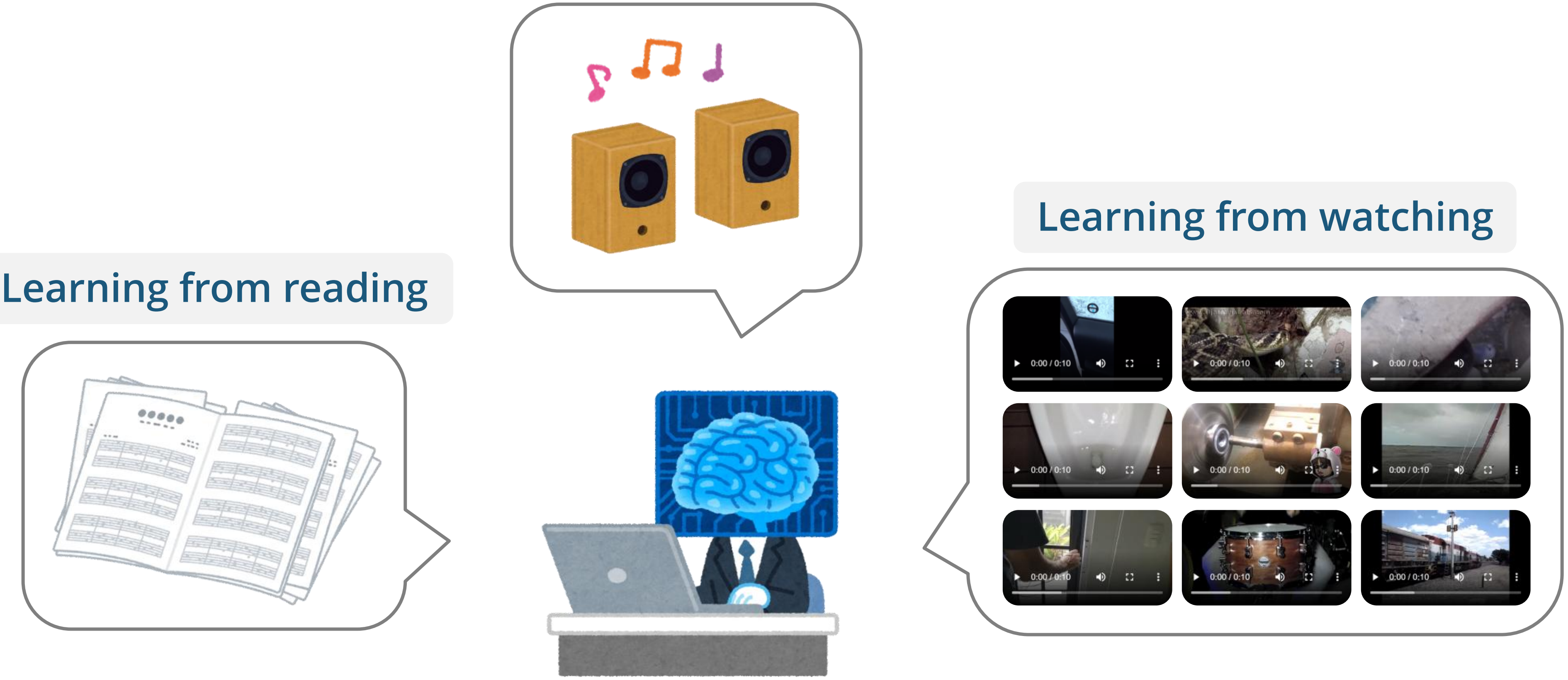
Can AI learn to create music by “listening to” music rather than “reading” music?

Human-inspired Machine Learning for Music & Audio

Learning from listening

Learning from watching

Learning from reading



A Baseline through Music Transcription

- Apply a music transcript model to acquire symbolic music data from audio
- But can we directly **learn to compose symbolic music through "listening to music" and "practicing music," just like how humans do**?

Multimodal Inputs for Generative Music AI

Text

Image

Video

Emotion

Challenge 2: Multimodality

Can AI learn to create music by “listening to” music rather than “reading” music?

Challenge 3: Usability

How can AI music tools be integrated into an artist's creative workflow?

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

Large language models
(GPT-4)

Pretrained generative audio models
(MusicGen, AudioLDM, Bark)

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

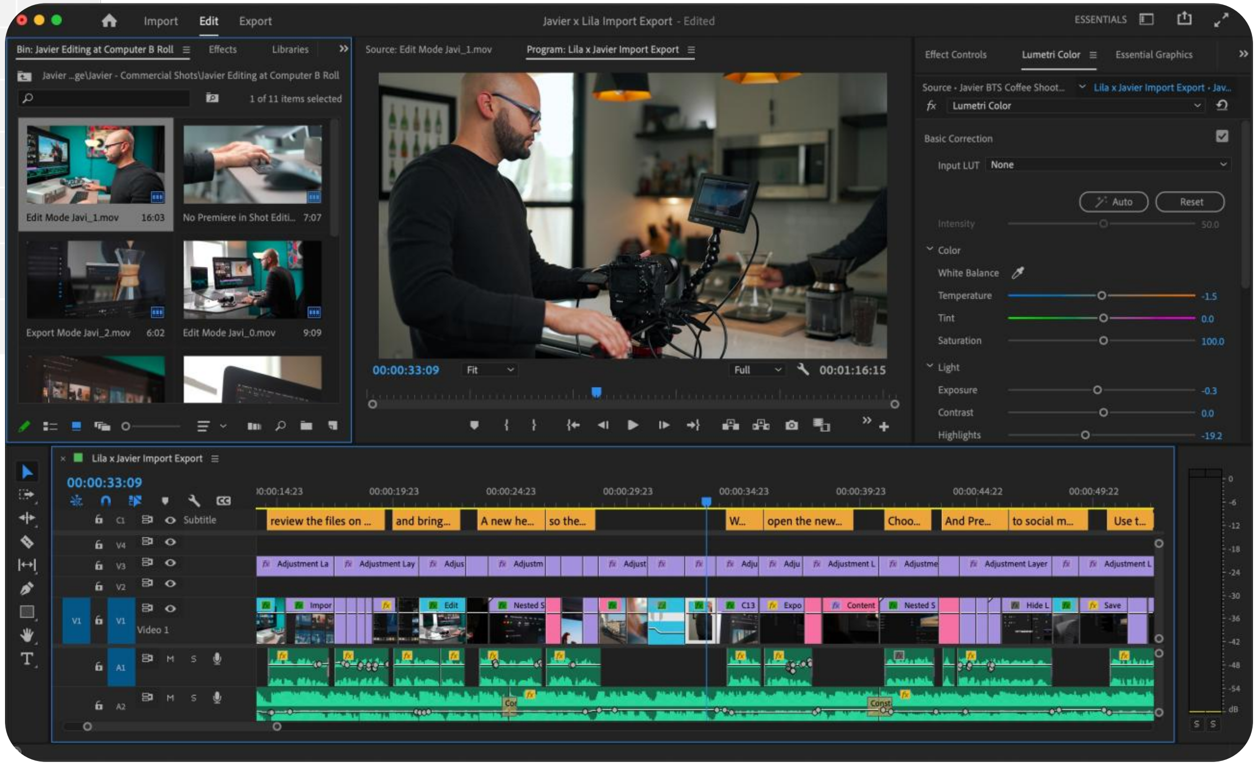
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect 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, Wenwu Wanga, "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

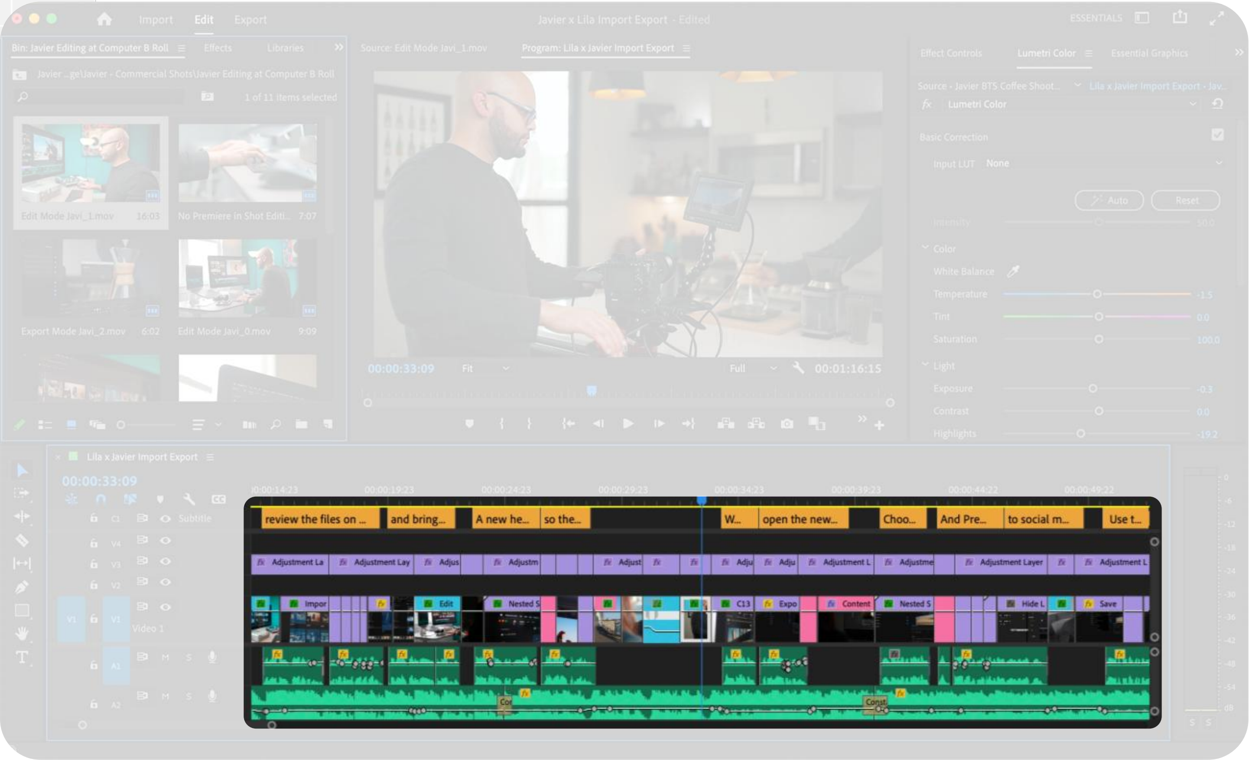
Integrating Generative AI into the Creative Workflow

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 …	



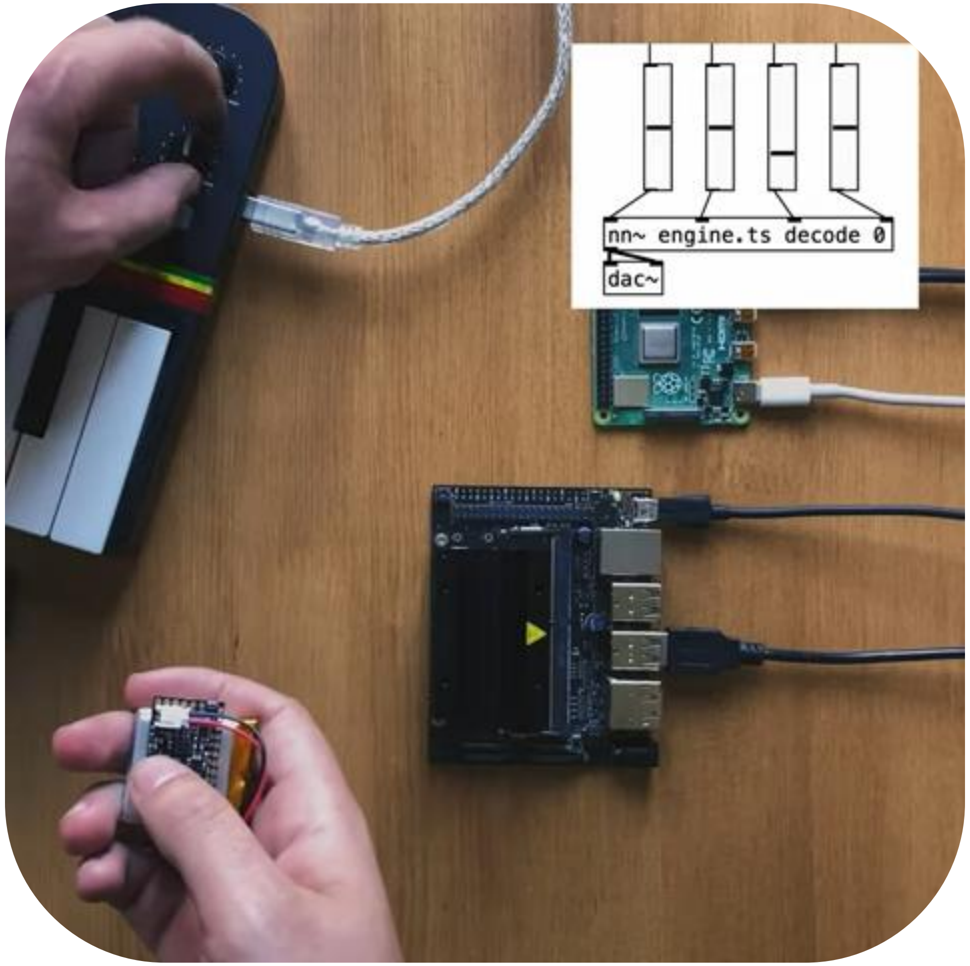
Integrating Generative AI into the Creative Workflow

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

RAVE: Real-time Audio Synthesis (Caillon & Esling, 2022)



youtu.be/jAIRf4nGgYI

Antoine Caillon and Philippe Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

Misusable Music Tools (Nao Tokui, 2023)

- "Throughout history, music and technology have often intertwined, with **new technologies being misused by artists** (turntables, etc)"
  — Nao Tokui, 2024
- "AI is more challenging to misuse because **it lacks a physical entity and operates as a black box.**"
  — Nao Tokui, 2024

(Source: Flintmi via Wikimedia Commons)

Nao Tokui, "Surfing human creativity with AI — A user's guide," *Lulu Press*, 2023.
x.com/naotokui_en/status/1756180102344499338
Flintmi, CC BY 3.0, via Wikimedia Commons

Challenge 3: Usability

How can AI music tools be integrated into an artist's creative workflow?

Challenge 4: Personalization

How can we make “my personal AI music tools”?

YACHT & Google Magenta

“**The band first took all 82 songs from their back catalog** and isolated each part, from bass lines to vocal melodies to drum rhythms; they then took those isolated parts and broke them up into four-bar loops. Then, **they put those loops into the machine learning model**, which **put out new melodies based on their old work**. **They did a similar process with lyrics**, **using their old songs plus other material they considered inspiring**. The final task was to pick lyrics and melodies that made sense, and pair them together to make a song.”

youtu.be/_yz8QYzcfxI

YACHT, “YACHT — SCATTERHEAD (4K Lyric Video)”, *YouTube*, July 26, 2019.
Megan Friedman, “Behind Magenta, the tech that rocked I/O,” *The Keyword*, May 20, 2019.
Adam Roberts, “YACHT's new album is powered by ML + Artists”, *Magenta Blog*, September 13, 2019.

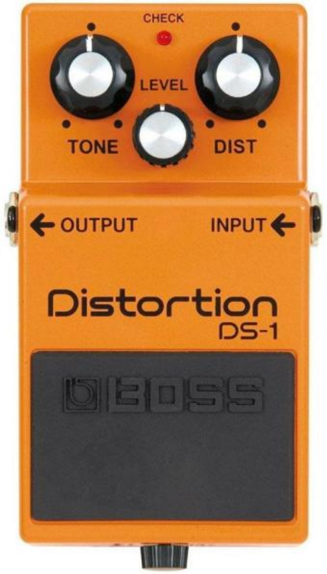
Ease of Personalization for Artists

- Through **finetuning our own models**
- Through **finetuning with live inputs**
- Python scripting vs friendly user interface

- **Can we do better?**

Overfitting vs Distortion

- Will **overfitting** be a new music expression, the “**distortion**” for AI music?



Personalized Text-to-Music Generation (Plitsis et al., 2024)

S*	A disco song with a **S***
S*	A recording of a **S*** song with a hip hop drum beat accompaniment

(Source: Plitsis et al., 2024)

Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis, "Investigating Personalization Methods in Text to Music Generation," *ICASSP*, 2024.

Challenge 4: Personalization

How can we make “my personal AI music tools”?

Challenge 5: Creativity

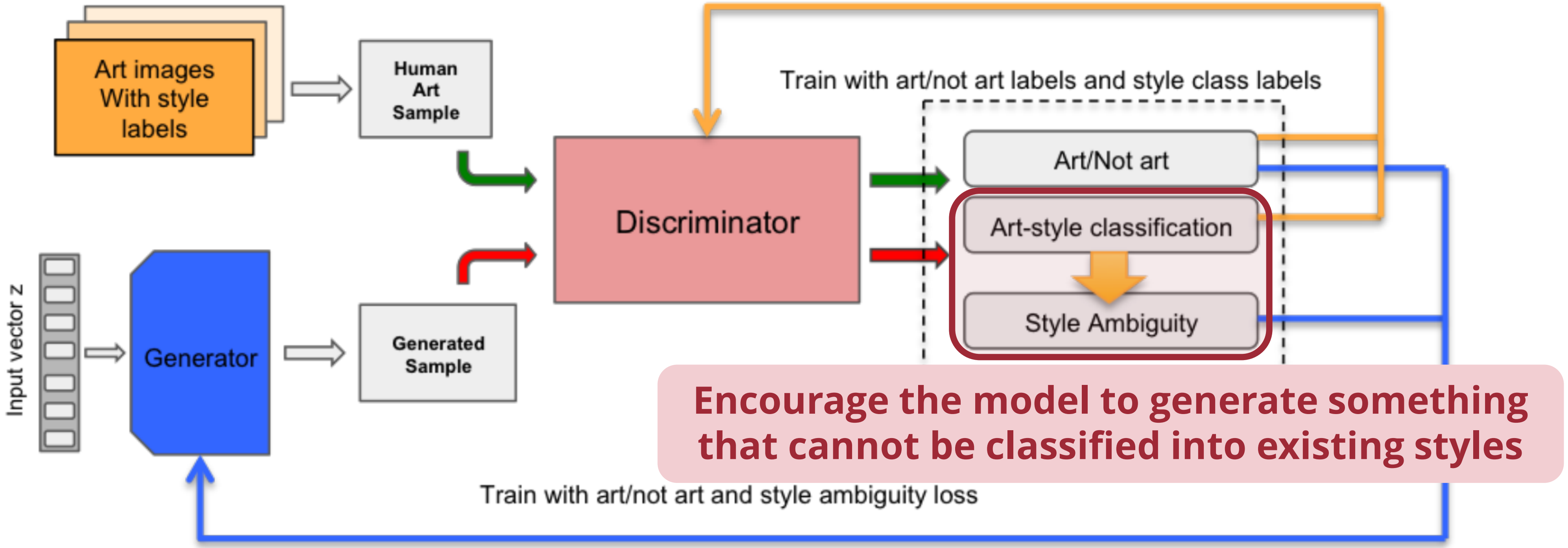
Can AI ever be creative? How can AI augment human creativity?

The Curse of Machine Learning

- As the old saying goes, "**Artificial intelligence is only as good as the data it learns from**."
- Machine learning models are trained to approximate some distribution in its formal formulation.
- This seems to contradict the idea of creativity that requires **the ability to extrapolate** and **think out of the box**.

- **Can AI ever be creative?**

Creative Adversarial Network (Elgammal et al., 2017)

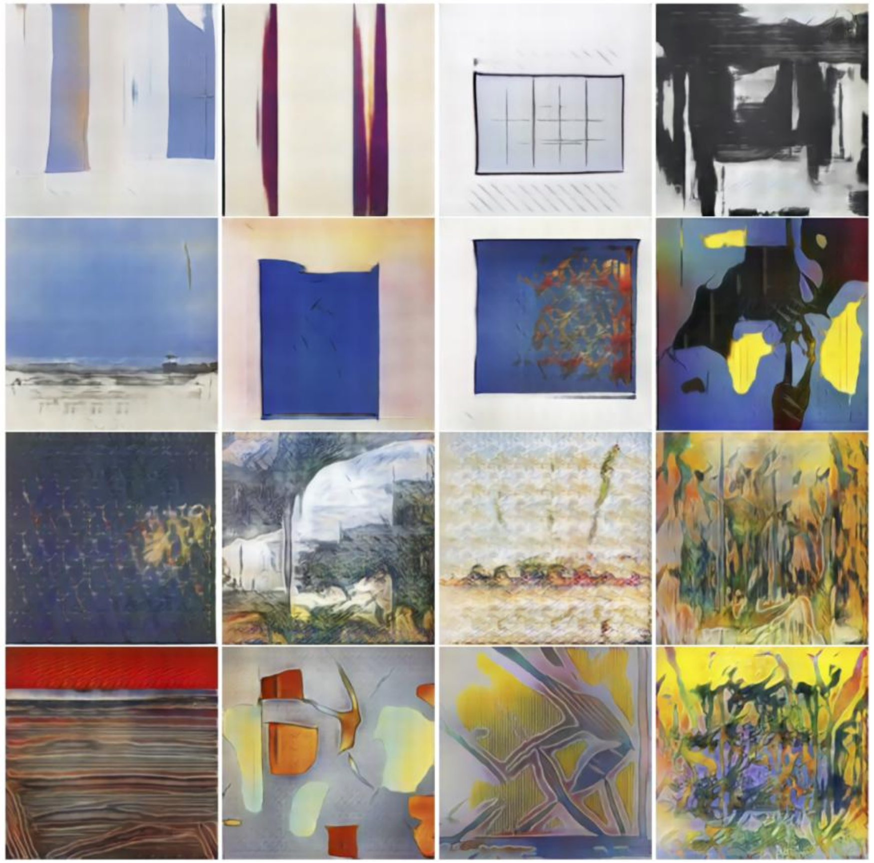


(Source: Elgammal et al., 2017)

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone, "CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms," *ICCC*, 2017.

Creative Adversarial Network (Elgammal et al., 2017)

Example generated images

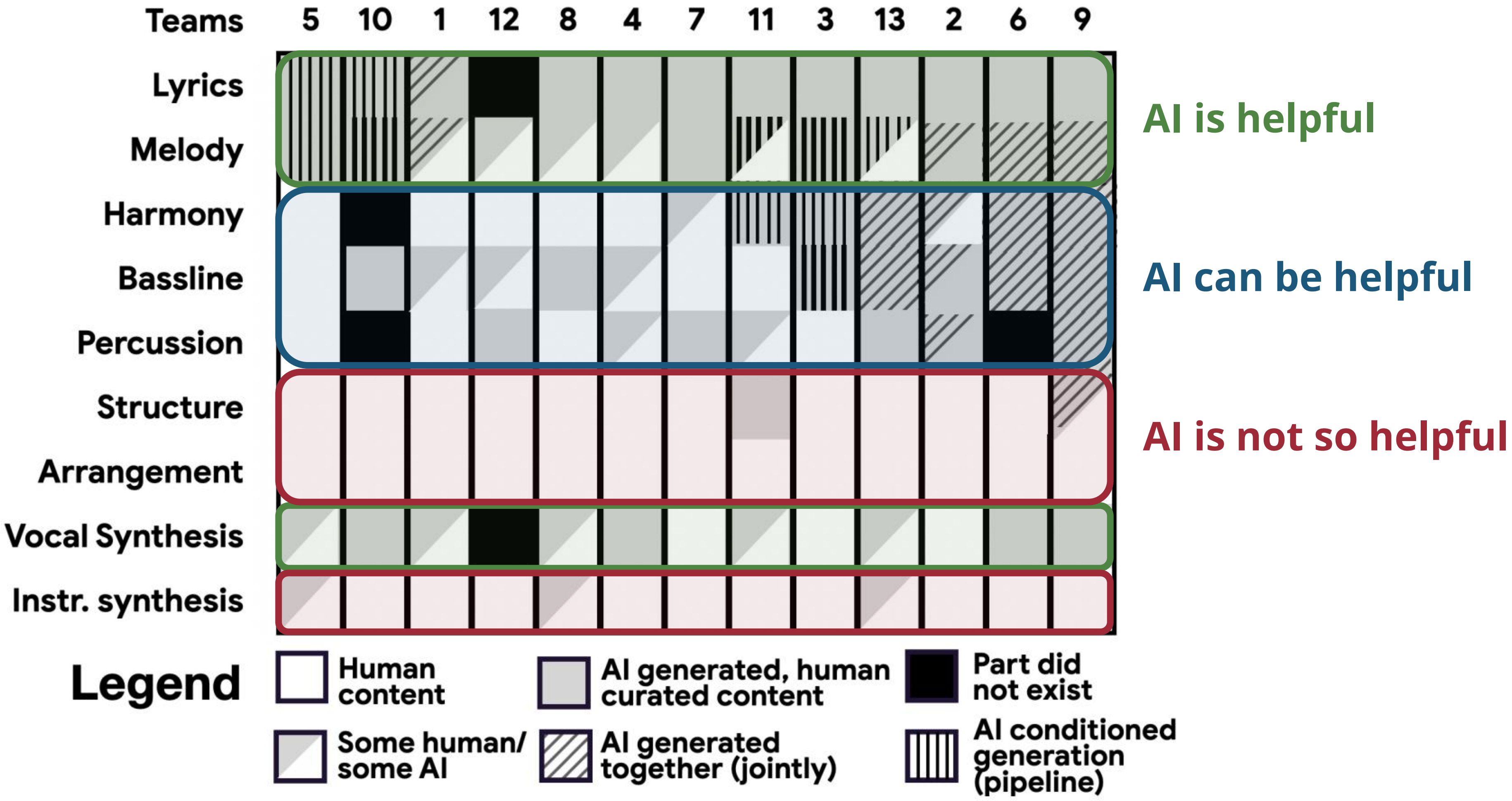


Best samples

(Source: Elgammal et al., 2017)

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone, "CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms," *ICCC*, 2017.

How can AI Augment Human Creativity?



(Source: Huang et al., 2020)

Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai, "AI Song Contest: Human-AI Co-Creation in Songwriting," *ISMIR*, 2020.

Creativity vs Art

Creativity is allowing yourself to make mistakes.
Art is knowing which ones to keep.

— Scott Adams

Challenge 5: Creativity

Can AI ever be creative? How can AI augment human creativity?

The Five Challenges

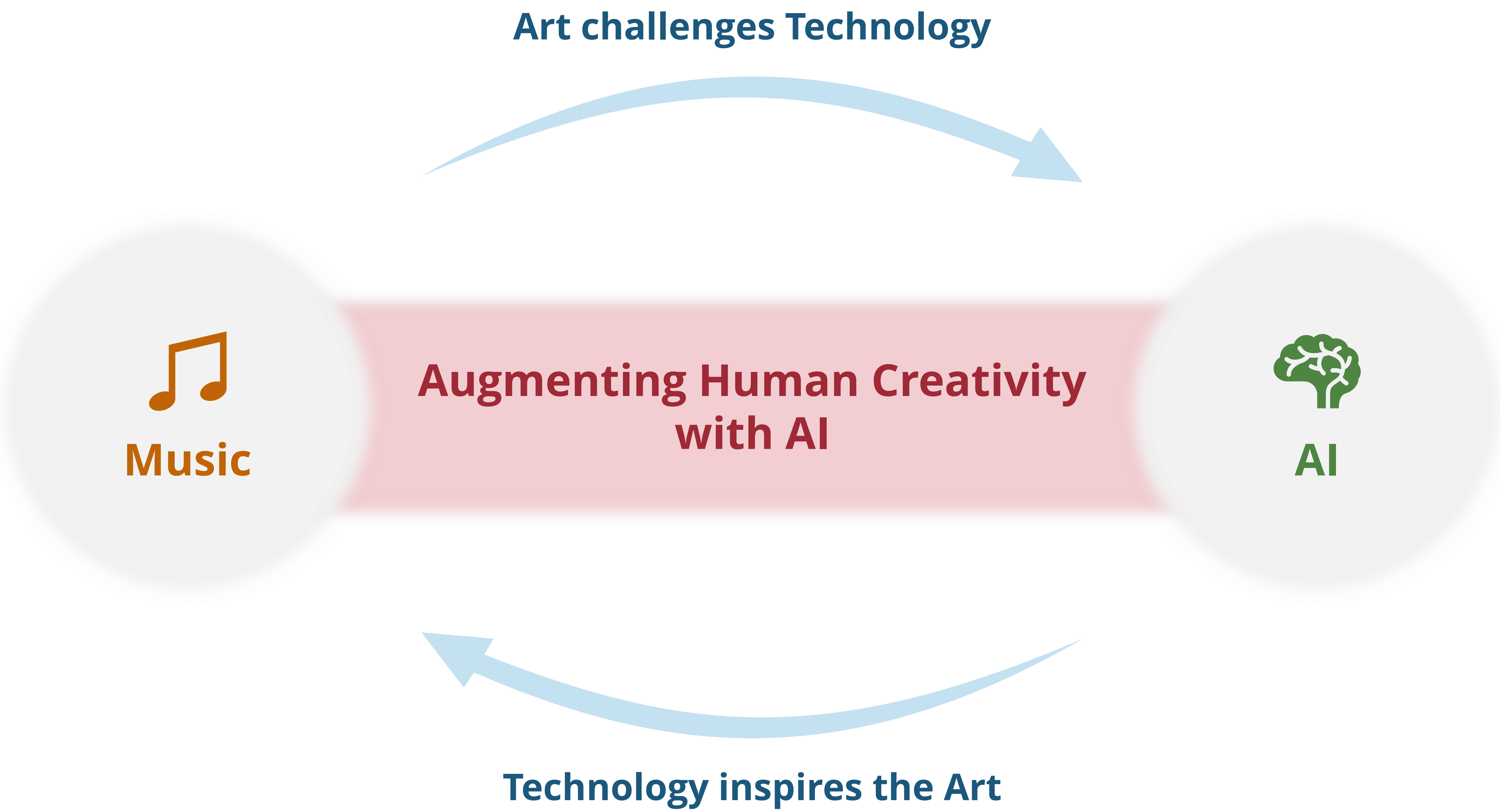
Representations Multimodality Usability Personalization Creativity

- **Representations**: How can we best represent music for machine learning?
- **Multimodality**: Can AI learn to create music by “listening to” music rather than “reading” music?
- **Usability**: How can AI music tools be integrated into an artist’s creative workflow?
- **Personalization**: How can we make “my personal AI music tools”?
- **Creativity**: Can AI ever be creative? How can AI augment human creativity?

Conclusion

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.


Art challenges Technology
Music
Augmenting Human Creativity with AI
AI
Technology inspires the Art

The Five Challenges

Representations

Usability

Creativity

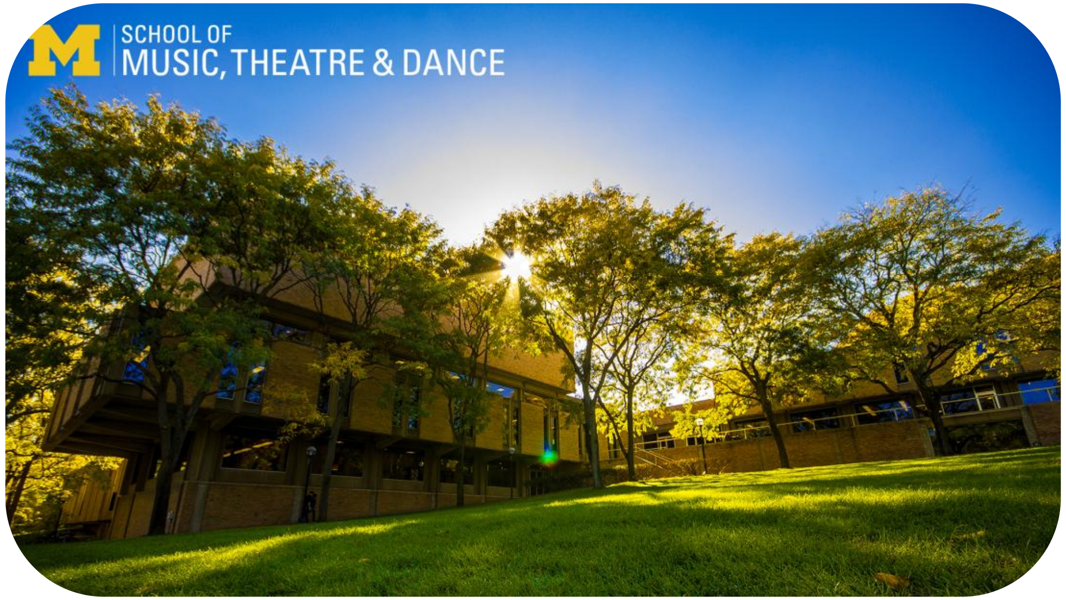
Multimodality

Personalization

AI Music @ Michigan

Hao-Wen Dong

Julie Zhu



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

Generative AI for Music: Challenges & Opportunities

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 教育部 · erc

hermandong.com / hwdong@umich.edu