

# Augmenting Human Creativity with AI

**Hao-Wen (Herman) Dong**

Department of Performing Arts Technology  
School of Music, Theatre & Dance  
University of Michigan  
[hermandong.com](http://hermandong.com)

October 7, 2025

# About Me



 **國立臺灣大學**  
National Taiwan University  
*B.S. in Electrical Engineering*

**UC San Diego**  
*M.S. in Computer Science*

**UC San Diego**  
*Ph.D. in Computer Science*

2013 – 2017

2017 – 2019

 **中央研究院**  
ACADEMIA SINICA  
*Research Assistant*

Summer 2019

 **YAMAHA**  
*Research Intern*

2019 – 2021

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 – 2024



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

# Music & Technology Co-evolves



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



# Music & AI

(Source: Yamaha)



(Source: Sankei Shimbun)



(Source: Robot Gizmos)



(Source: NBC DFW)

[yamaha.com/en/news\\_release/2018/18013101/](https://yamaha.com/en/news_release/2018/18013101/)  
[sankei.com/article/20240113-CQCOSQHJWFIYPJJKZDCITRTRVI/](https://sankei.com/article/20240113-CQCOSQHJWFIYPJJKZDCITRTRVI/)  
[roboticgizmos.com/shimon-musical-robot-deep-learning/](https://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/](https://nbcdfw.com/entertainment/the-scene/how-verdigris-ensemble-is-using-ai-to-create-a-new-concert-experience/3366031/)



# Generative AI for Music, Audio & Video Creation



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**

# Augmenting Human Creativity with AI

- **Generative Models for Music Creation**

- **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024), **text-to-music generation** (ISMIR 2025), **video-to-music generation** (ISMIR 2025), **symbolic music processing tools** (ISMIR LBD 2019, ISMIR 2020)

- **AI-assisted Music Creation Tools**

- **Expressive violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020), **a cappella source separation** (ISMIR LBD 2025)

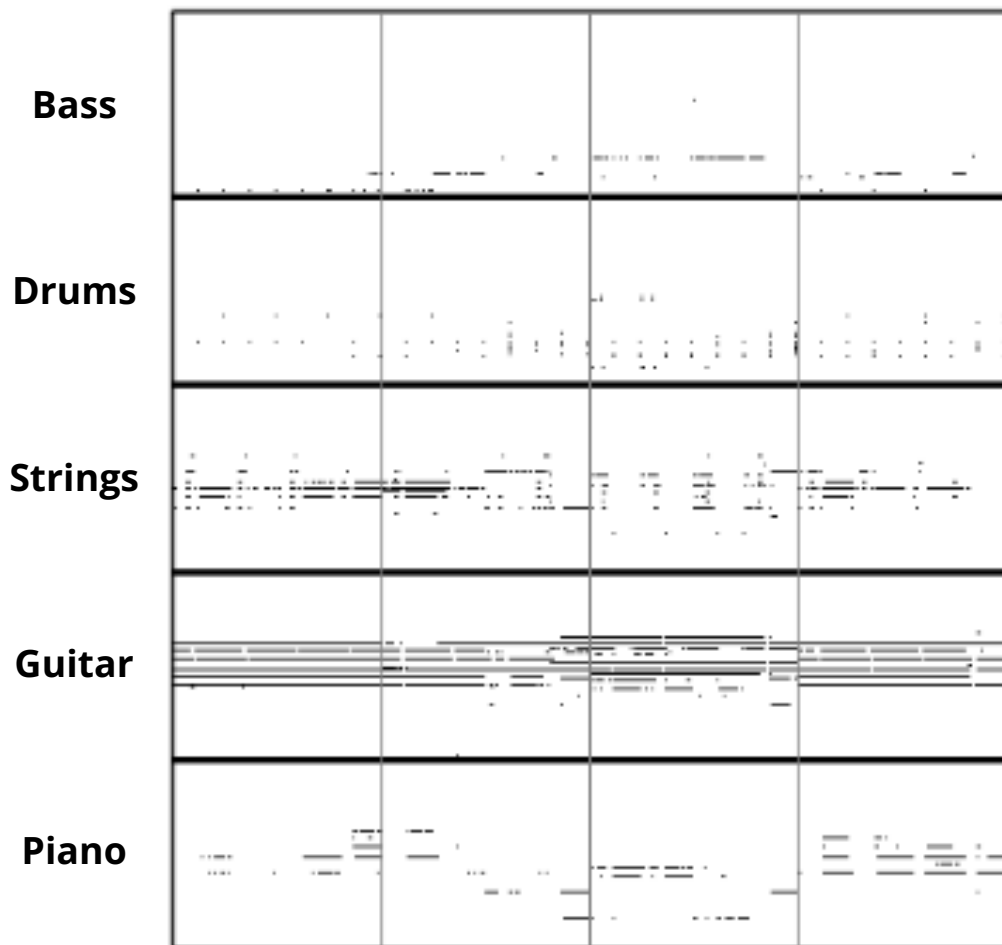
- **Multimodal Generative Models for Content Creation**

- **Long-to-short video editing** (ICLR 2025, NeurIPS 2025), **text-queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WASPAA 2023)

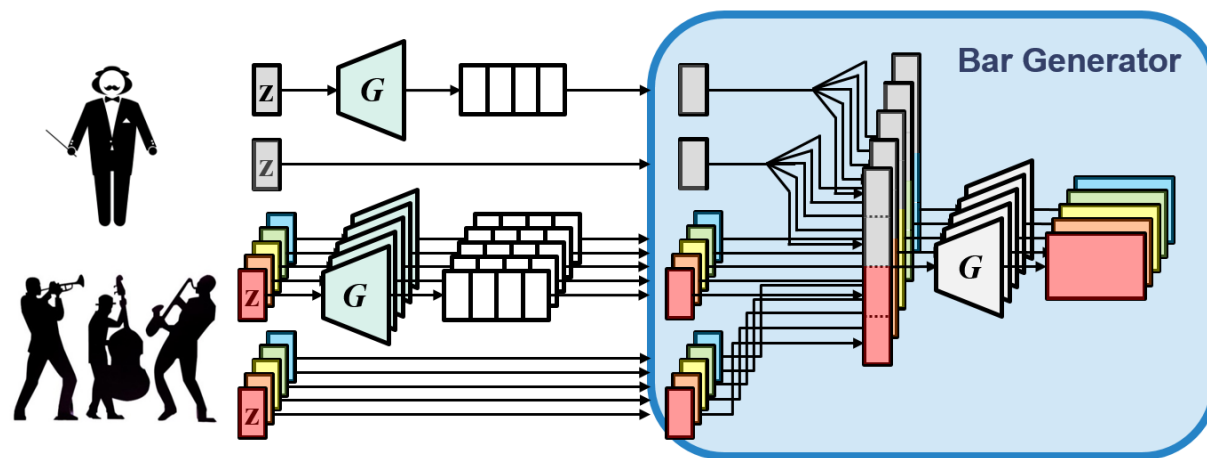


# Generating Multi-instrument Music using GANs (AAAI 2018)

Multitrack Piano Roll



MuseGAN Generator



# MuseGAN Features in AWS DeepComposer (2020)

AWS DeepComposer > Models > Train a model

## Train a model

**Generative algorithm** [Info](#)

Choose a generative algorithm to train a model

☒ **MuseGAN**  
GAN algorithm often used for complex music structures

☐ **U-Net**  
U-Net is the distinguishing

512

Input noise

Volume

32-key, 2-octave keyboard

The image is a composite of several elements. On the left, a screenshot of the AWS DeepComposer web interface shows the 'Train a model' page. The 'Generative algorithm' section has 'MuseGAN' selected, with a note that it's often used for complex music structures. A diagram shows a sequence of layers with 512, 256, and 128 units. Another diagram shows 'Input noise' being processed by a 'U-Net' block. In the center, a large screen displays a 32-key, 2-octave keyboard. A man is standing in front of the screen, gesturing towards it. To the right, a smaller inset shows a man playing a similar keyboard. At the bottom, another man is shown holding a 32-key, 2-octave keyboard. The overall theme is the integration of AWS DeepComposer's MuseGAN feature into a physical keyboard interface.

[amazon.com/dp/B07YGGZ4V5B/](https://amazon.com/dp/B07YGGZ4V5B/)

Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," AWS News Blog, April 2, 2020.

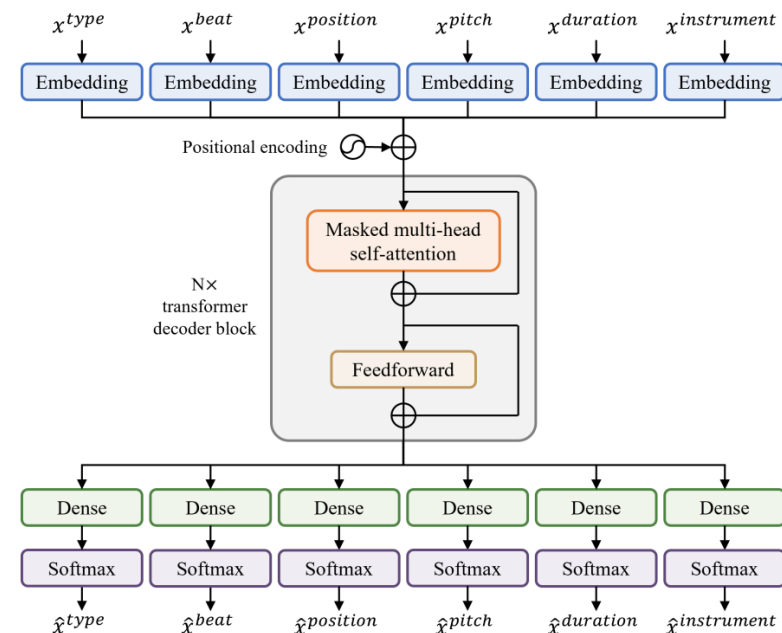
# Generating Multitrack Music with Transformers (ICASSP 2023)

## Multitrack Music Representation

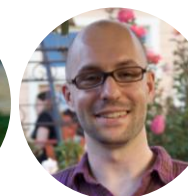
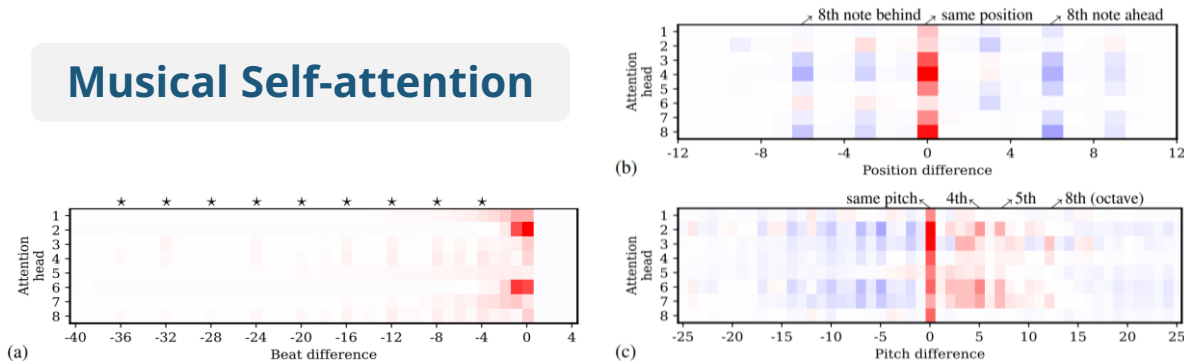
(0, 0, 0, 0, 0, 0) Start of song  
(1, 0, 0, 0, 0, 15) Instrument: accordion  
(1, 0, 0, 0, 0, 36) Instrument: trombone  
(1, 0, 0, 0, 0, 39) Instrument: brasses  
(2, 0, 0, 0, 0, 0) Start of notes  
(3, 1, 1, 41, 15, 36) Note: beat=1, position=1, pitch=E2, duration=48, instrument=trombone  
(3, 1, 1,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



## Multitrack Music Transformer



## Musical Self-attention



UC San Diego



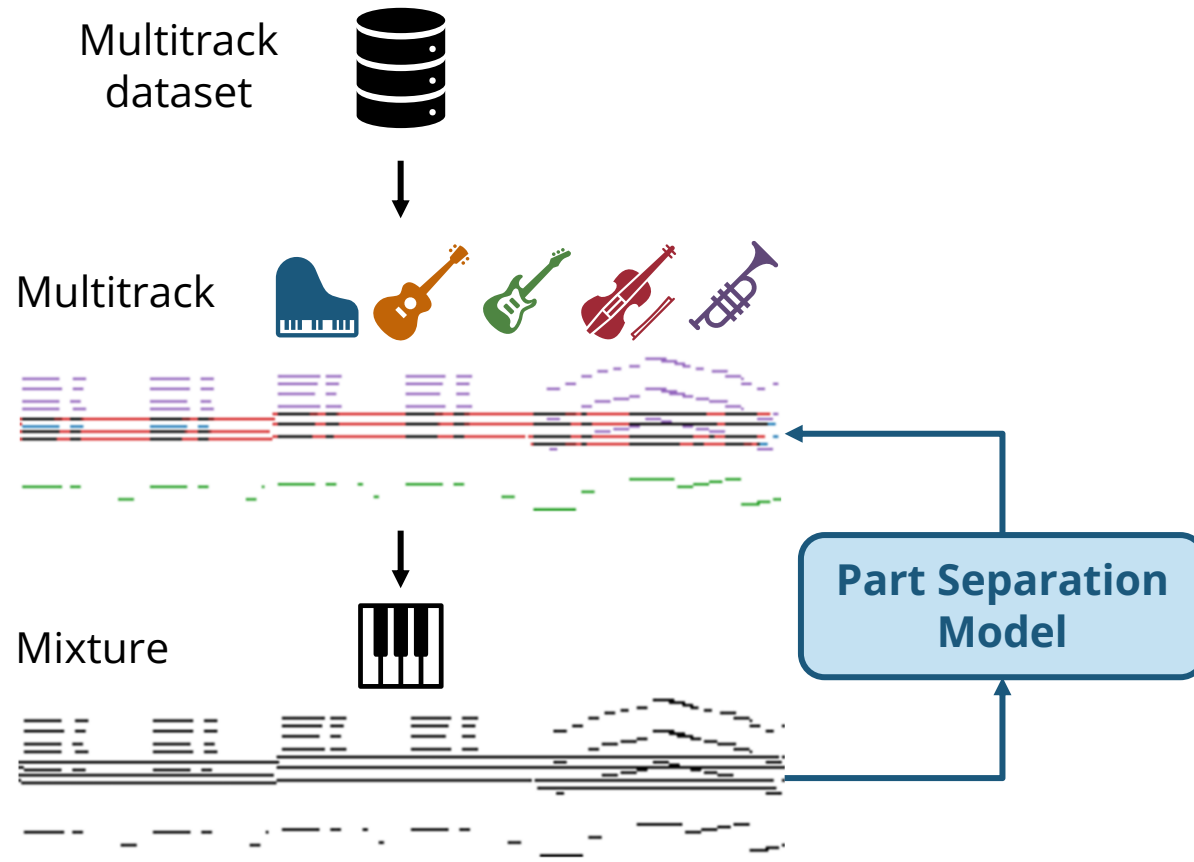
# Automatic Instrumentation (ISMIR 2021)



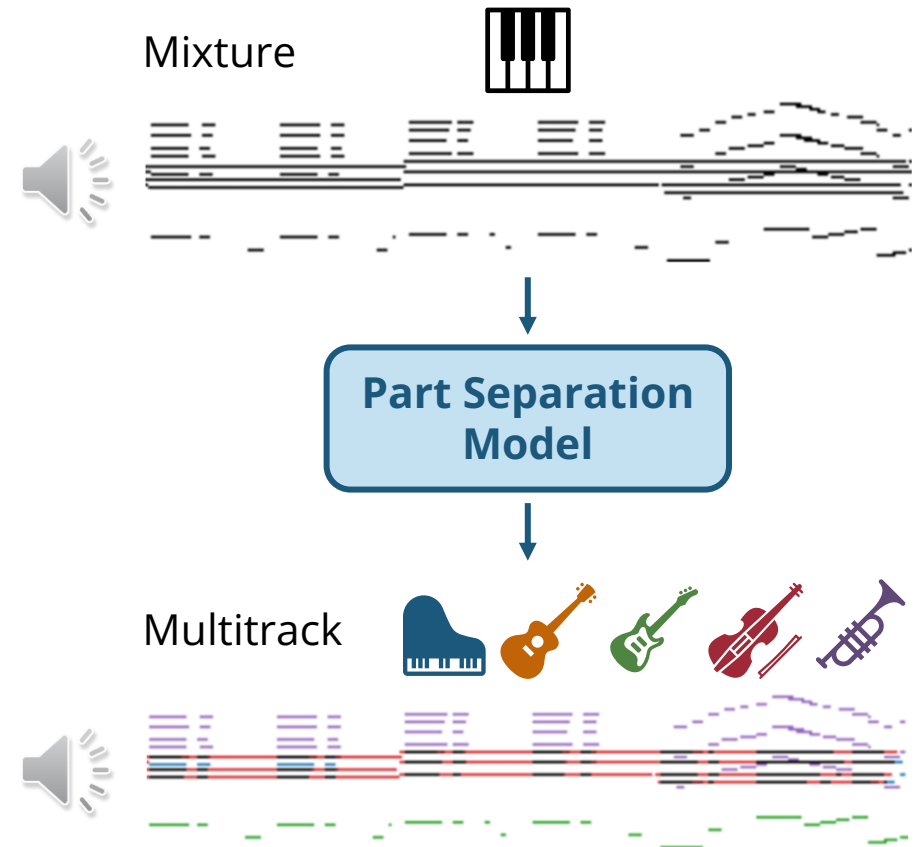
Stanford

UC San Diego

## Training

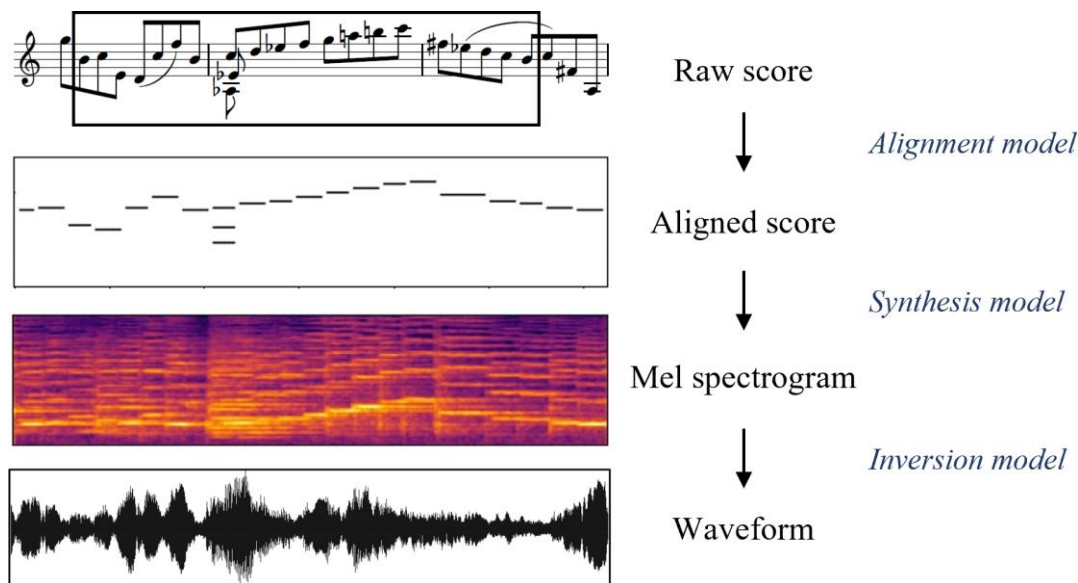


## Inference

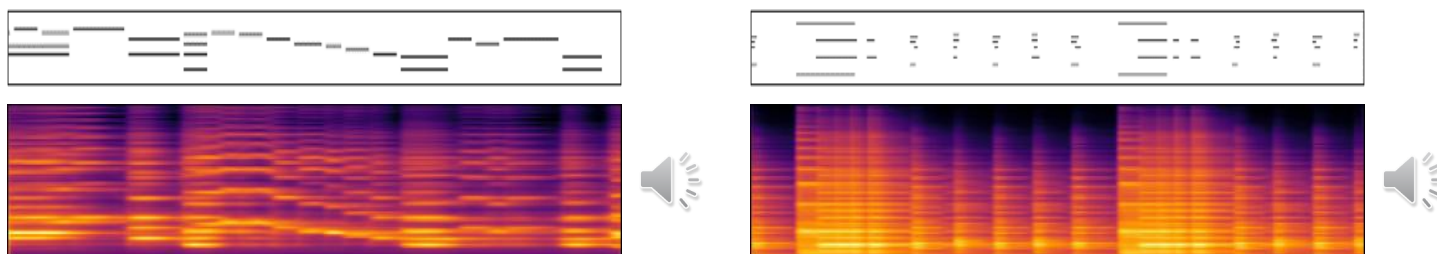


# Synthesizing Expressive Violin Performance (ICASSP 2022)

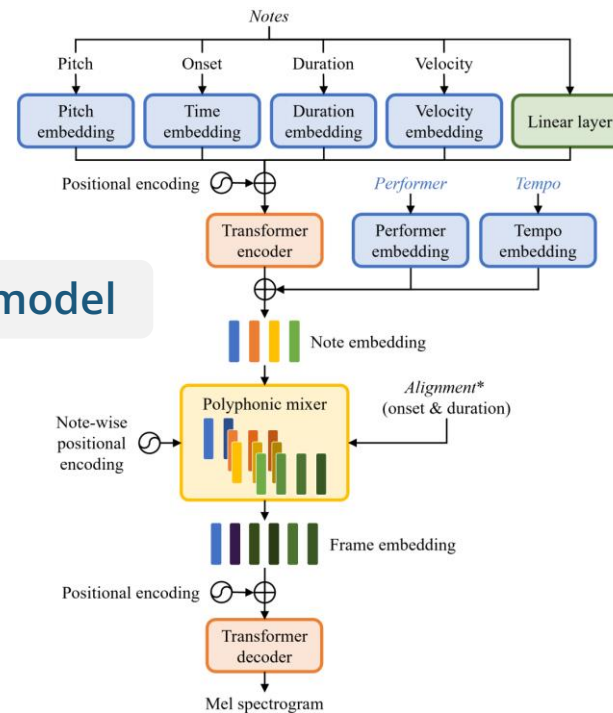
## Performance synthesis



## Example results



## TTS-based model

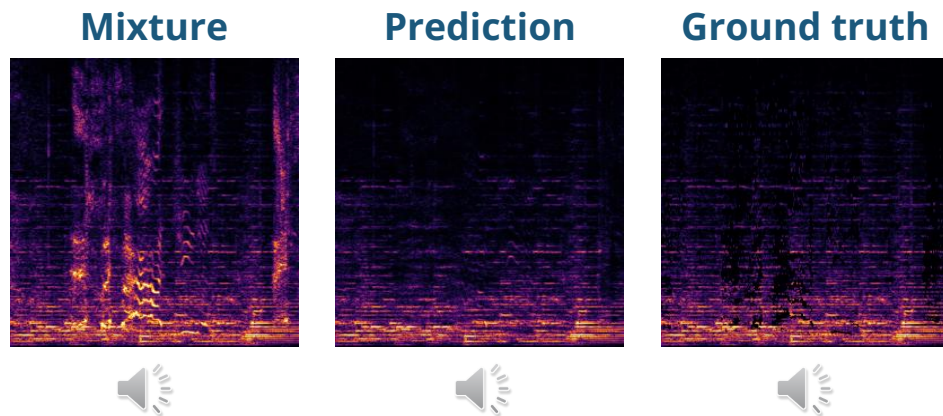


Dolby

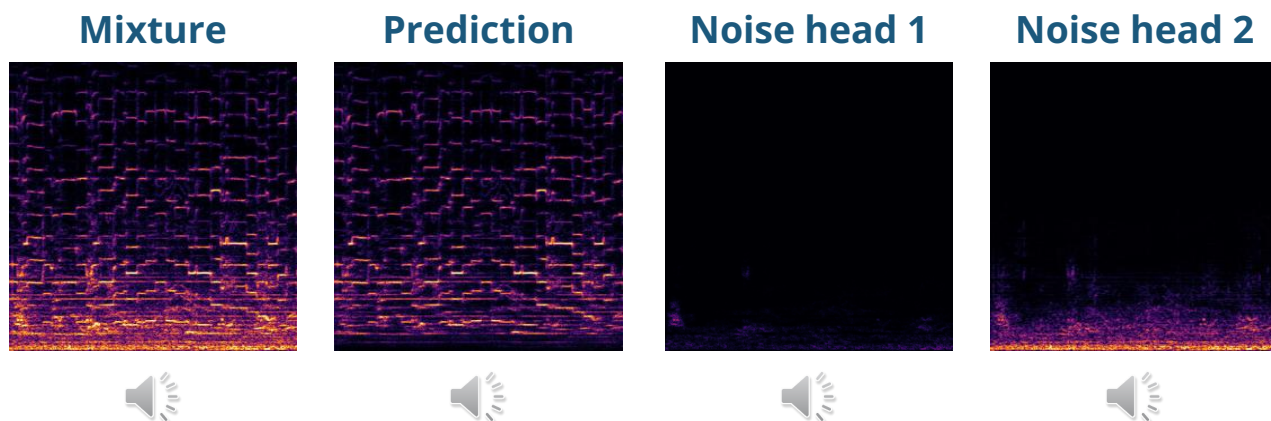
UC San Diego

# Text-queried Sound Separation (ICLR 2023)

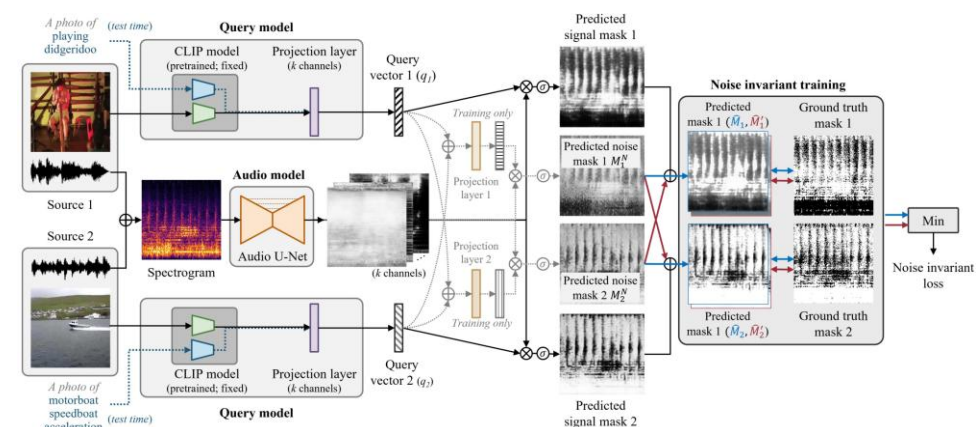
Query: *"playing harpsichord"*



Query: *"playing bagpipe"*



## Text-queried sound separation model



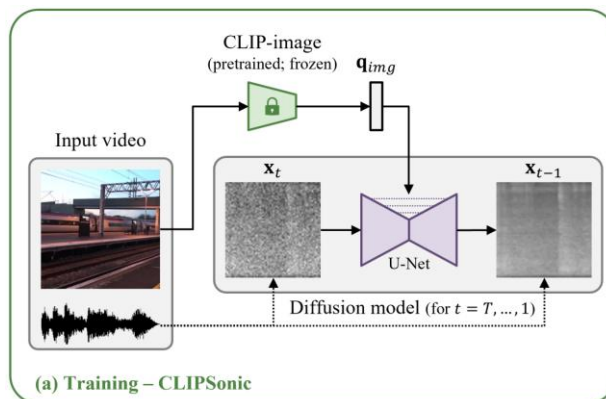


# Text-to-Audio Synthesis (WASPAA 2023)

## Learning Sounds from Noisy Videos



## Training

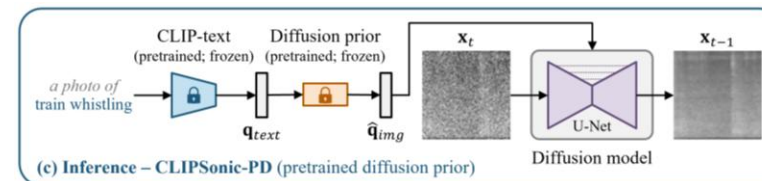


## Image-to-sound results

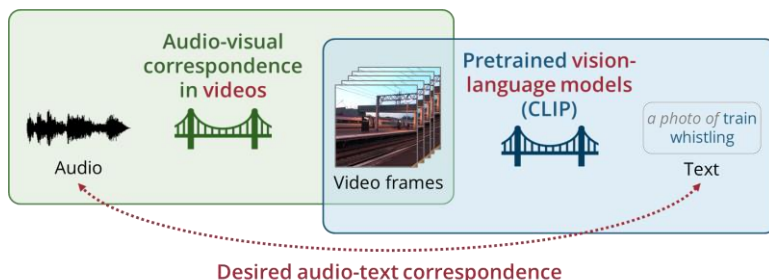


UC San Diego

## Inference



## Text-to-sound results

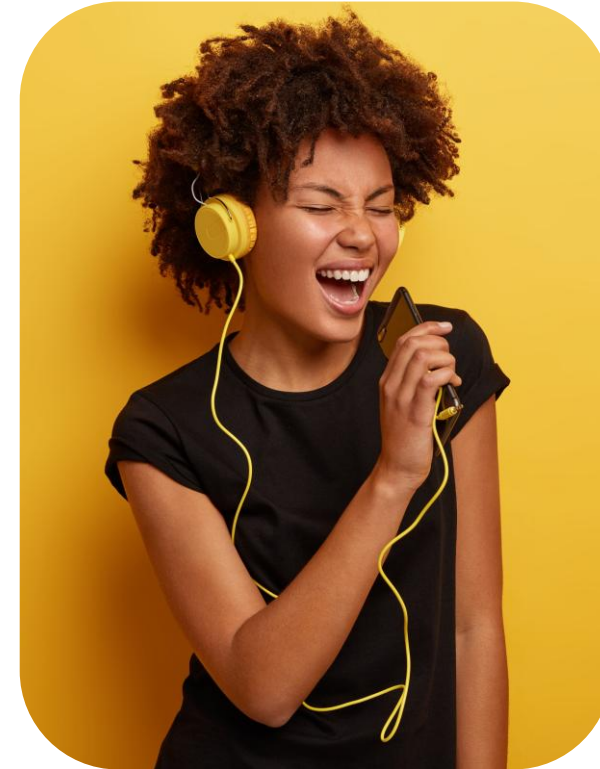


🔥 Ongoing Work: A Cappella Vocal Coach 🔥

# AcaMate: AI-assisted A Cappella Practice App



**Seagull-K** from Hsinchu, Taiwan

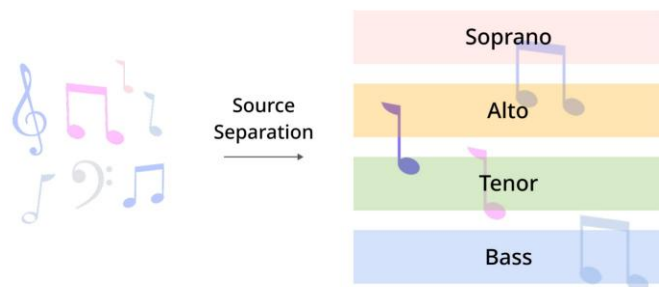


**How can we best support  
a novice a cappella singer in  
practicing their singing skills?**

# AcaMate: AI-assisted A Cappella Practice App



## Preparing voice parts



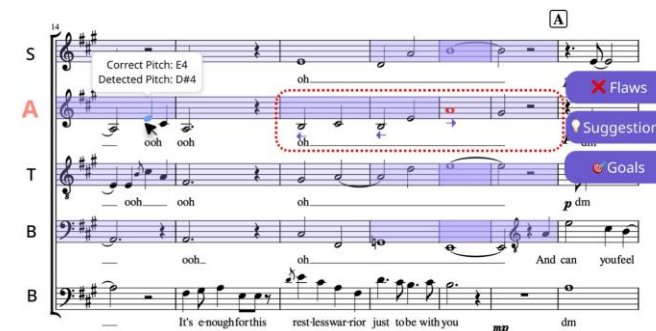
**A** Before Singing

## Highlighting musical patterns



**B** During Singing

## Providing visual feedback (on pitch, rhythm, dynamics)



**C** After Singing

**D** Iterative Singing

System-Assisted Deliberate Practice



# AcaMate: AI-assisted A Cappella Practice App



ACAMATE

**Pitch errors** **Rhythm errors** **A**

S  
Correct Pitch: E4  
Detected Pitch: D#4  
oh p dm

A  
ooh ooh p dm

T  
ooh ooh oh p dm

B  
ooh oh And can you feel

It's e-nough for this rest-less war-rior just to be with you mp dm

## High-level suggestions

### What needs to be improved

Adjust timing: avoid rushing into the first note of each measure and keep a steady tempo throughout the segment.

Correct pitch accuracy on the highest note (slightly sharp A4).

Align dynamics with other parts by building a stronger crescendo instead of decrescendo, especially emphasizing strength in the last two measures.

### Suggestions

Mute the Bass Part

Accept

Lower Soprano Volume

Accept

### Goals:

- ★ Rhythm
- ★ Pitch
- ★ Dynamics

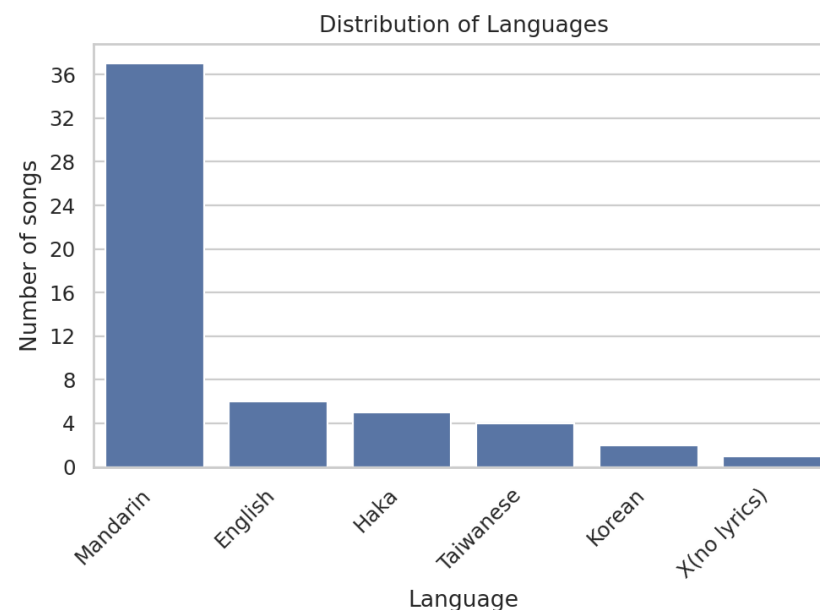
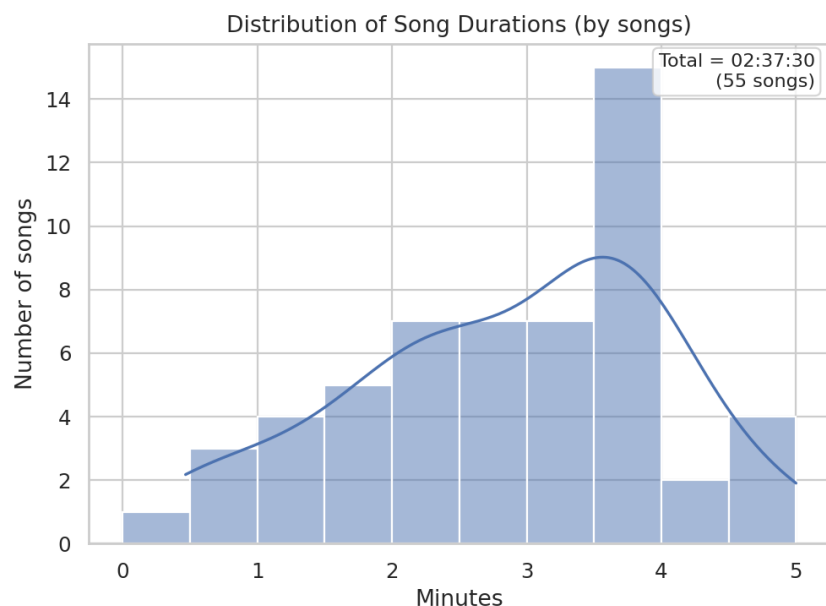
Practice again! →

Practice this part

# ACappellaSet: Studio Recordings with Stems



- **55** studio-quality a cappella songs **with stems** performed by 3 groups
- **2.6 hours** in total
- **Five languages:** Mandarin, English, Hakka, Taiwanese, and Korean






# Finetuning A Cappella Source Separation Models







| Model                 | VP   | Other | All  |
|-----------------------|------|-------|------|
| Pretrained (official) | 5.22 | 10.66 | 7.94 |
| Pretrained (drum)     | 3.66 | 9.24  | 6.45 |
| Fine-tuned (ours)     | 7.62 | 11.63 | 9.62 |


+2.4 dB +1 dB



**Vocal percussion****SATB**



**Pretrained**



**Finetuned**

**Ground truth**

**Vocal percussion****SATB**

**Pretrained**

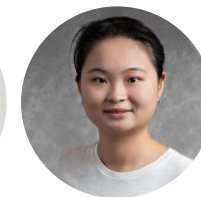
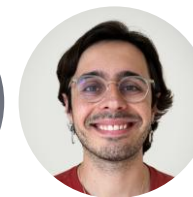
**Finetuned**

**Ground truth**

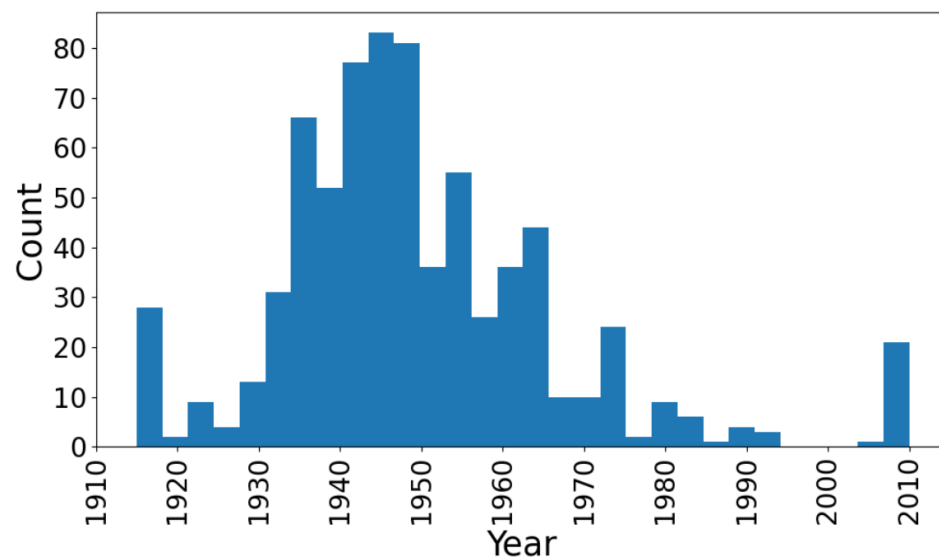
🔥 Ongoing Work: AI-assisted Film Scoring 🔥



# Open Screen Soundtrack Library (OSSL)



- **736 video clips** from **299 films** in **public domain** or **CC-licensed**
- **36.5 hours** in total
- **Mood annotations** as Russell's 4Q (arousal-valence model)

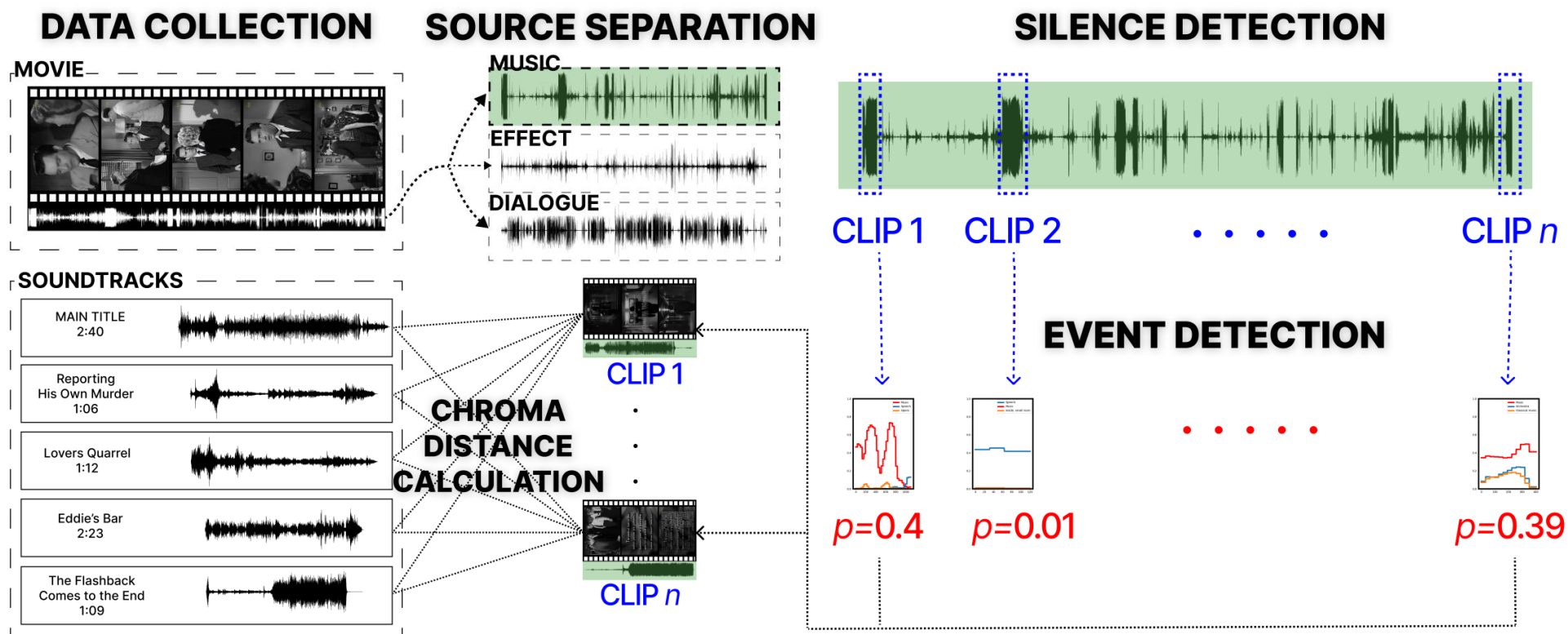


# | Open Screen Soundtrack Library (OSSL)

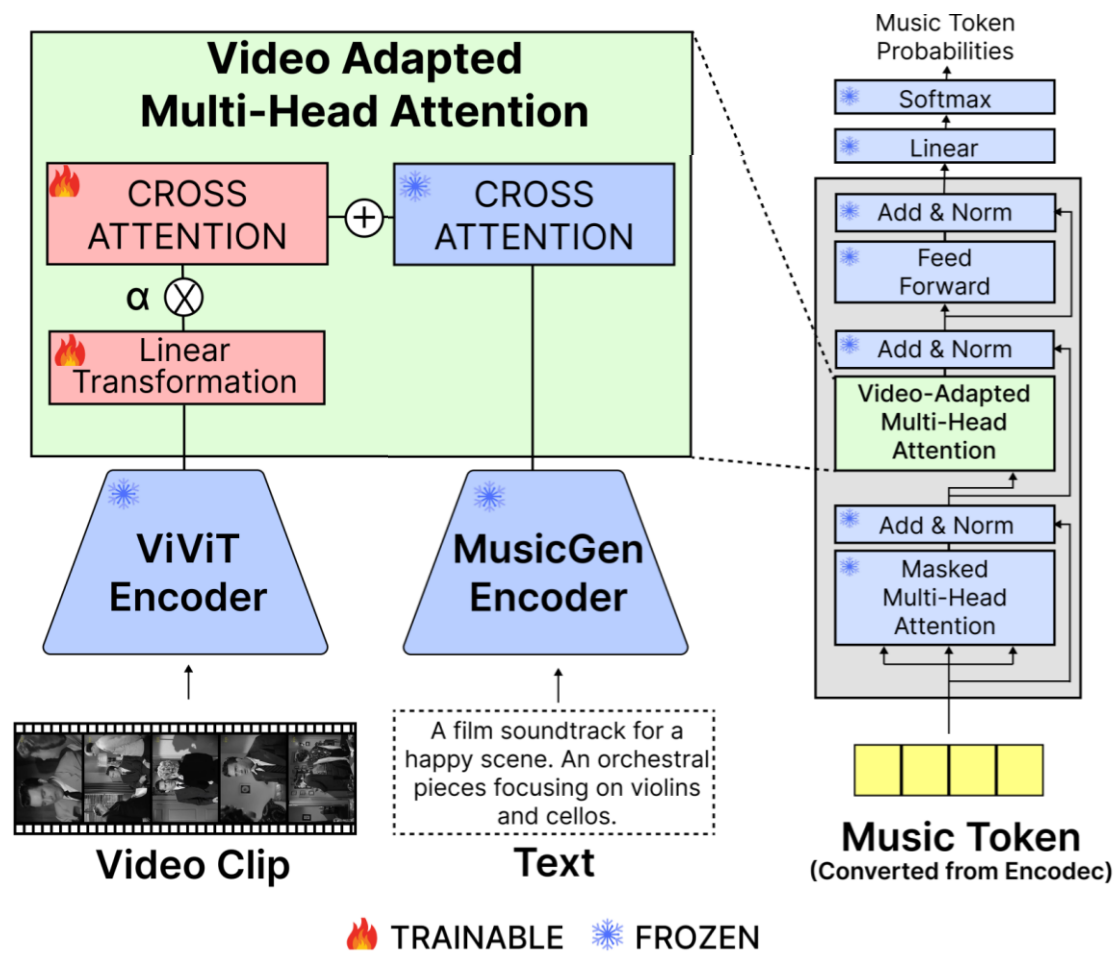
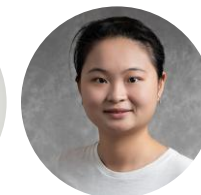
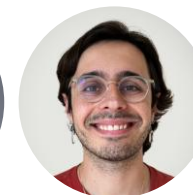


[youtu.be/DjBVqhErShM](https://youtu.be/DjBVqhErShM)

# Matching Soundtracks to Video Clips



# Video-Guided Text-to-Music Generation





# | Video-Guided Text-to-Music Generation



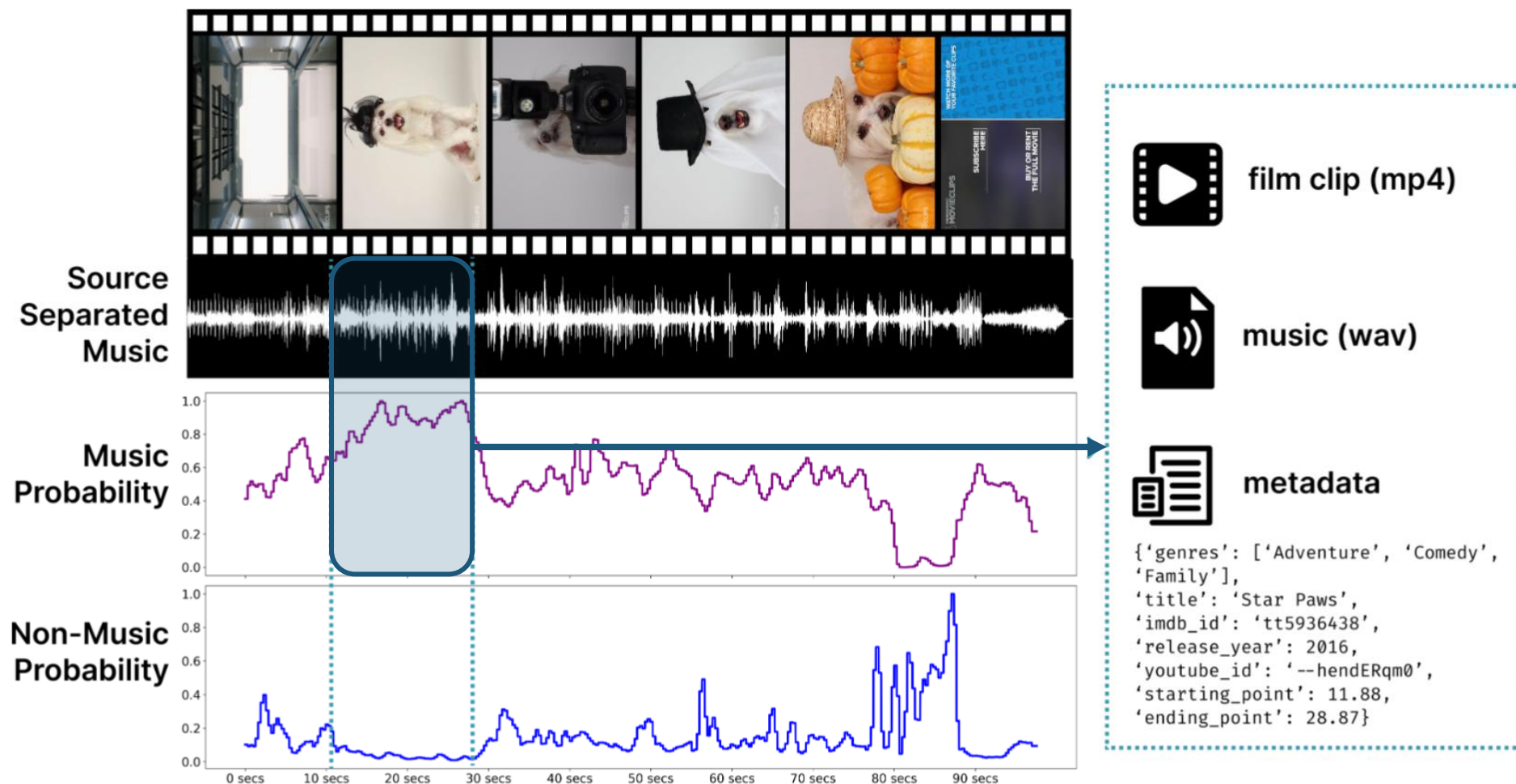
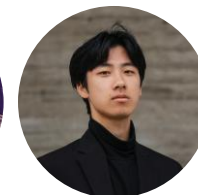
## Video-Guided Text-to-Music Generation Using Public Domain Movie Collections

(ISMIR 2025)

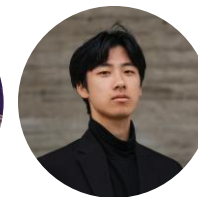
Haven Kim, Zachary Novack, Weihan Xu,  
Julian McAuley, Hao-Wen Dong

[youtu.be/S0BMicbdzmg](https://youtu.be/S0BMicbdzmg)

# Extending OSSL to OSSL v2



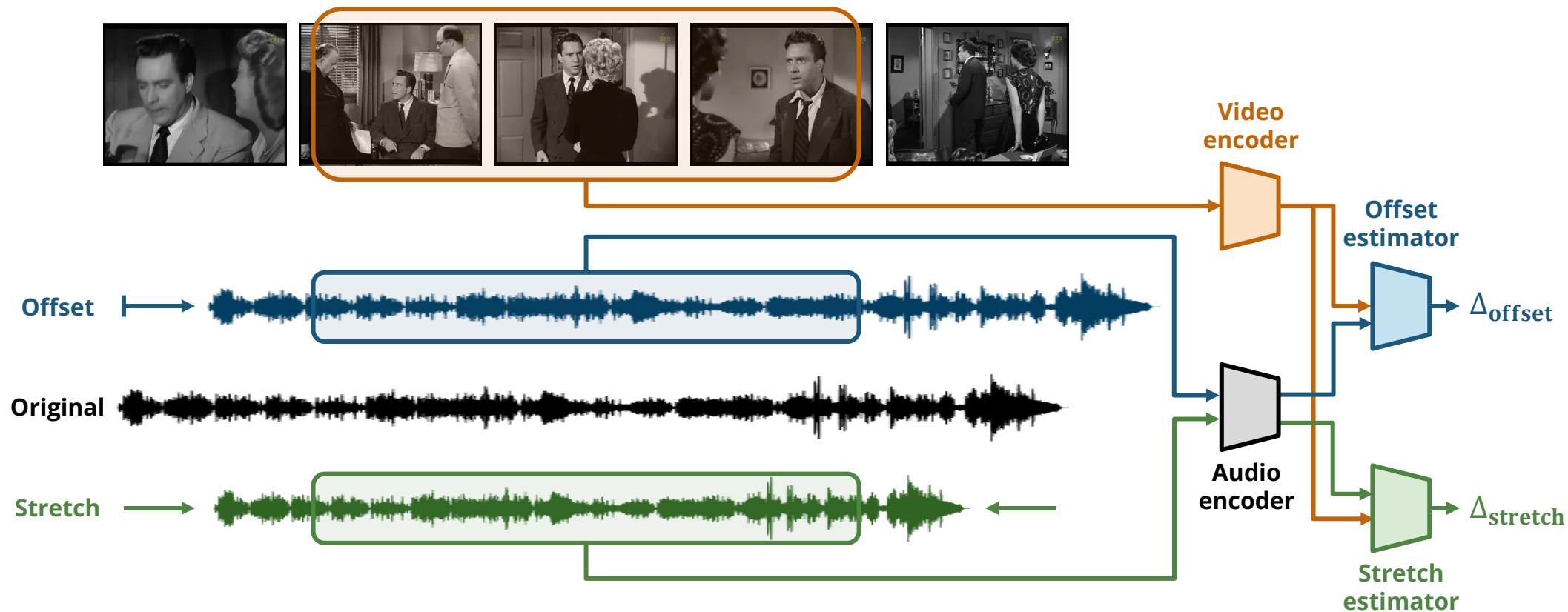
# Extending OSSL to OSSL v2



|                          | Public Domain | Commercial [2] | Total  |
|--------------------------|---------------|----------------|--------|
| Number of Clips          | 35,705        | 40,703         | 76,408 |
| Number of Unique Films   | 1,886         | 2,633          | 4,519  |
| Average Length (seconds) | 28.77         | 23.65          | 26.04  |
| Total Length(hours)      | 285.31        | 267.39         | 552.70 |

Stay tuned! 🤖

# Future Work: Measuring Video-Music Alignment



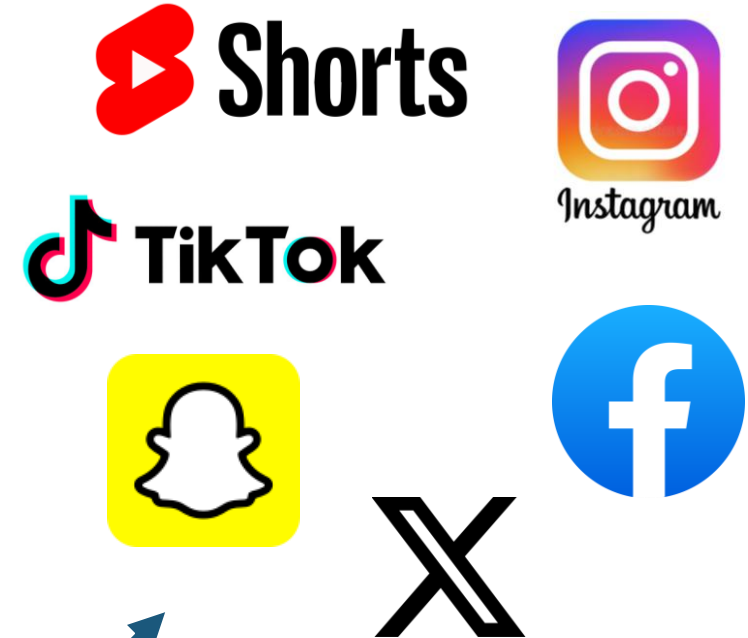


🔥 Ongoing Work: AI-assisted Video Editing 🔥

# Fast-growing Short Video Platforms



For content creators,  
help **promote** their  
long video contents



For content consumers,  
help **digest information**  
in a more engaging way

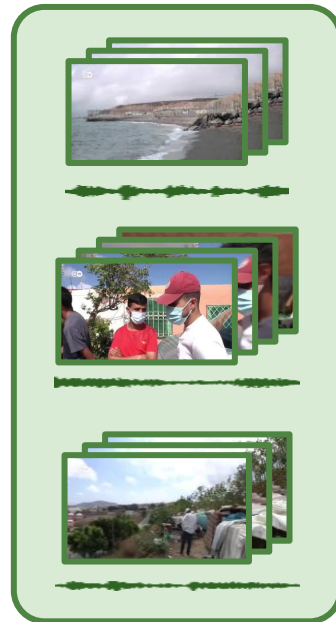
# Video Editing



**Interview footage**  
(main character)



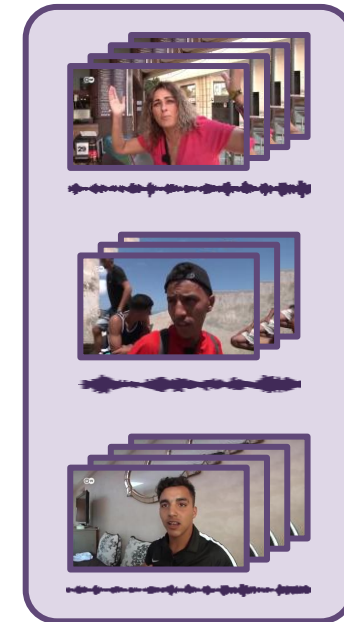
**Background footage**



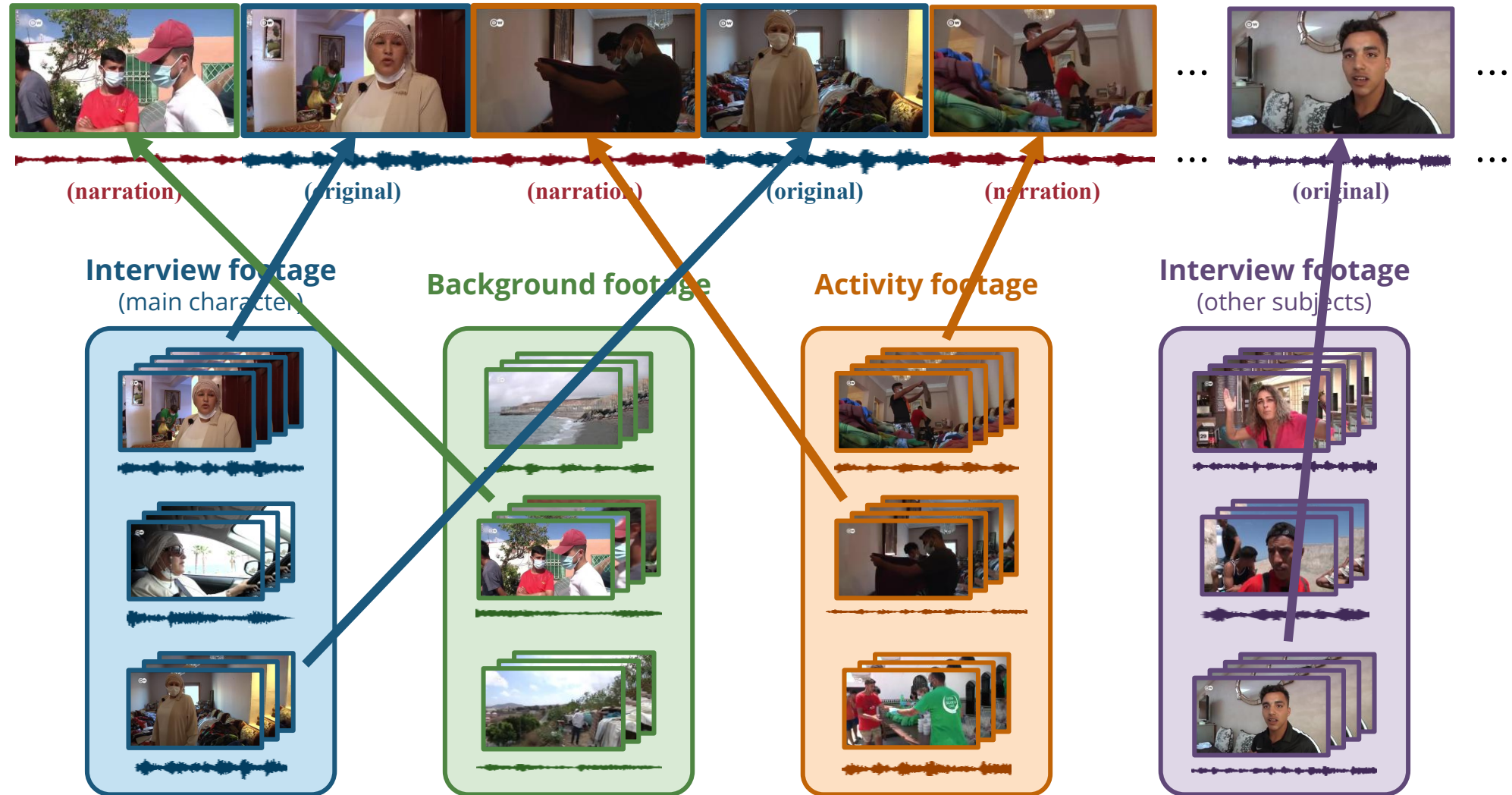
**Activity footage**



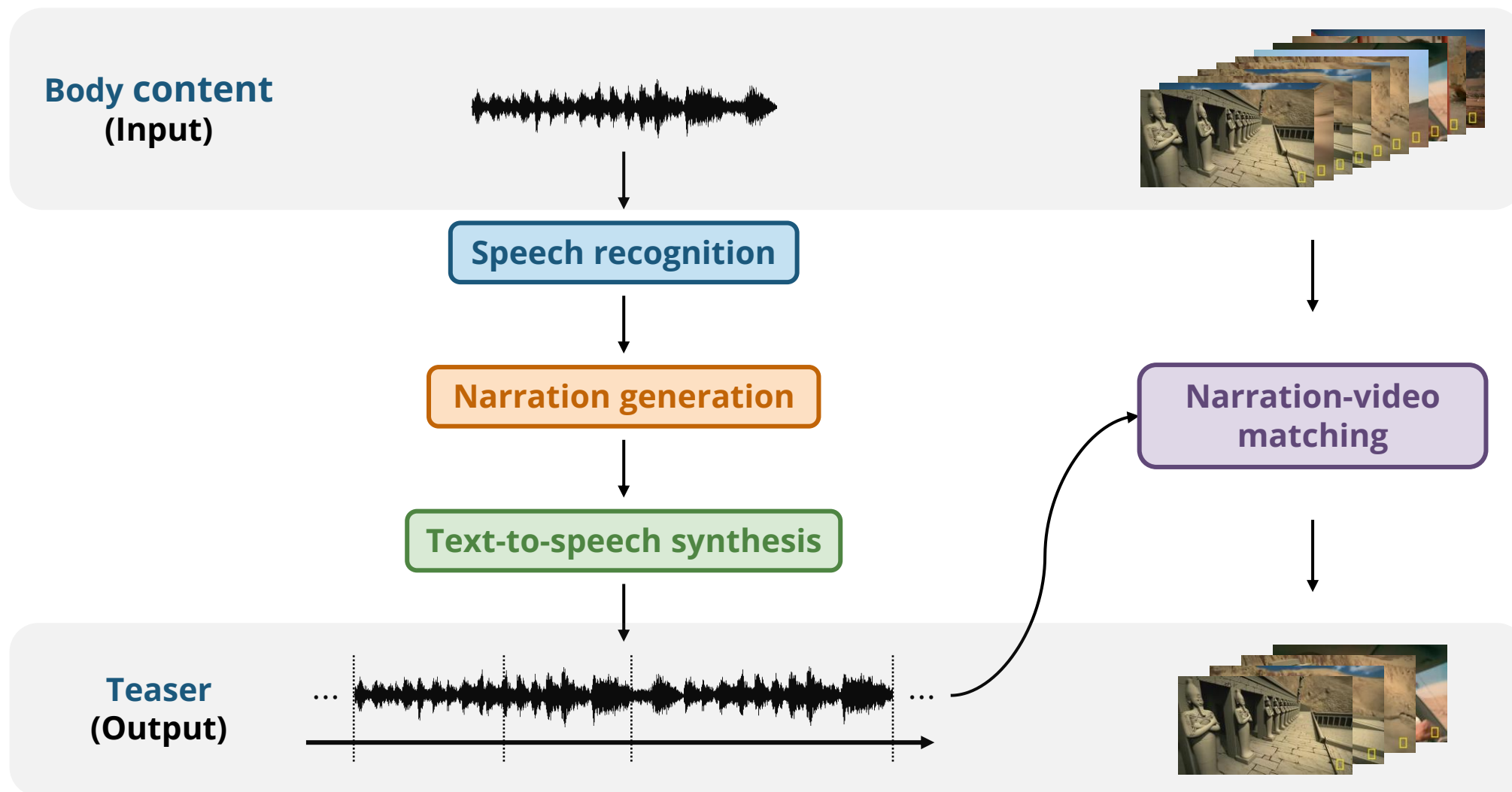
**Interview footage**  
(other subjects)



# Video Editing

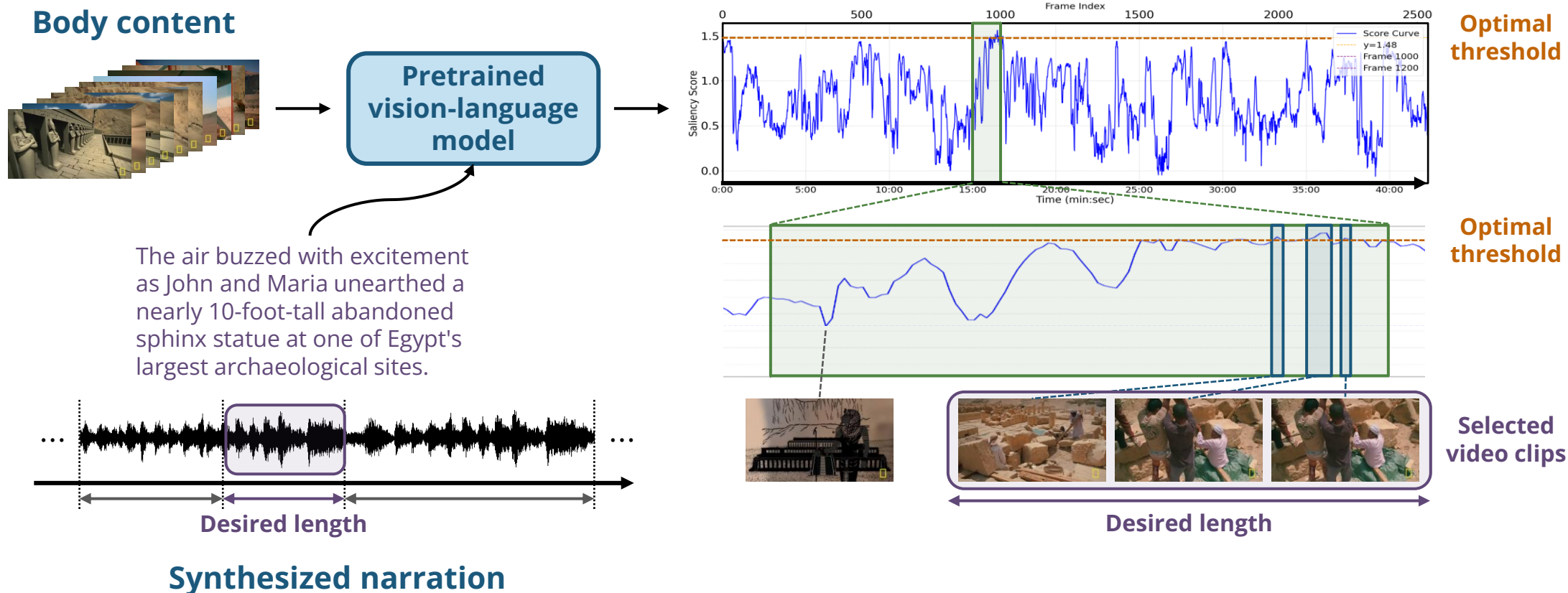


# Narration-Centered Long-to-Short Video Editing

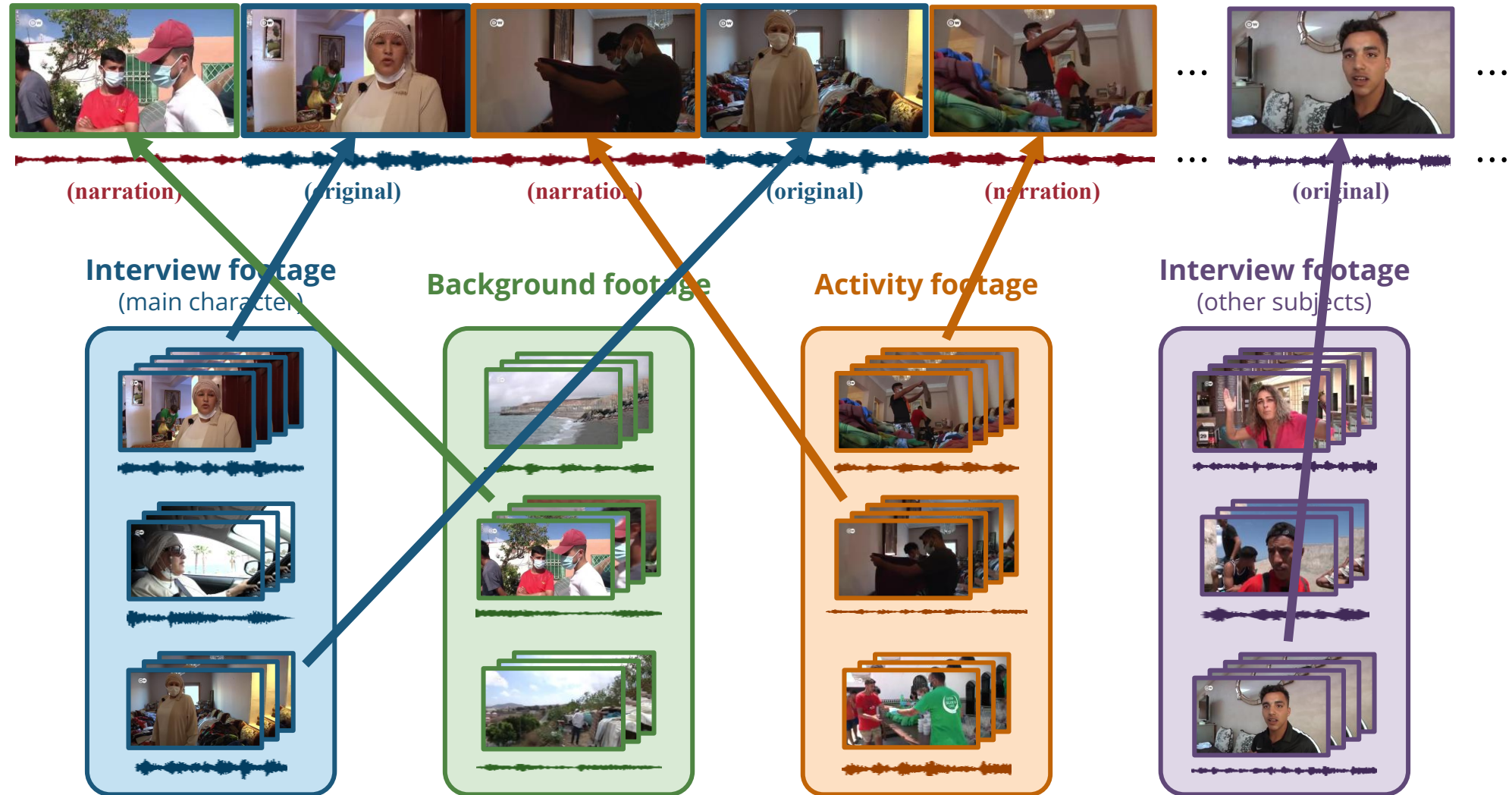




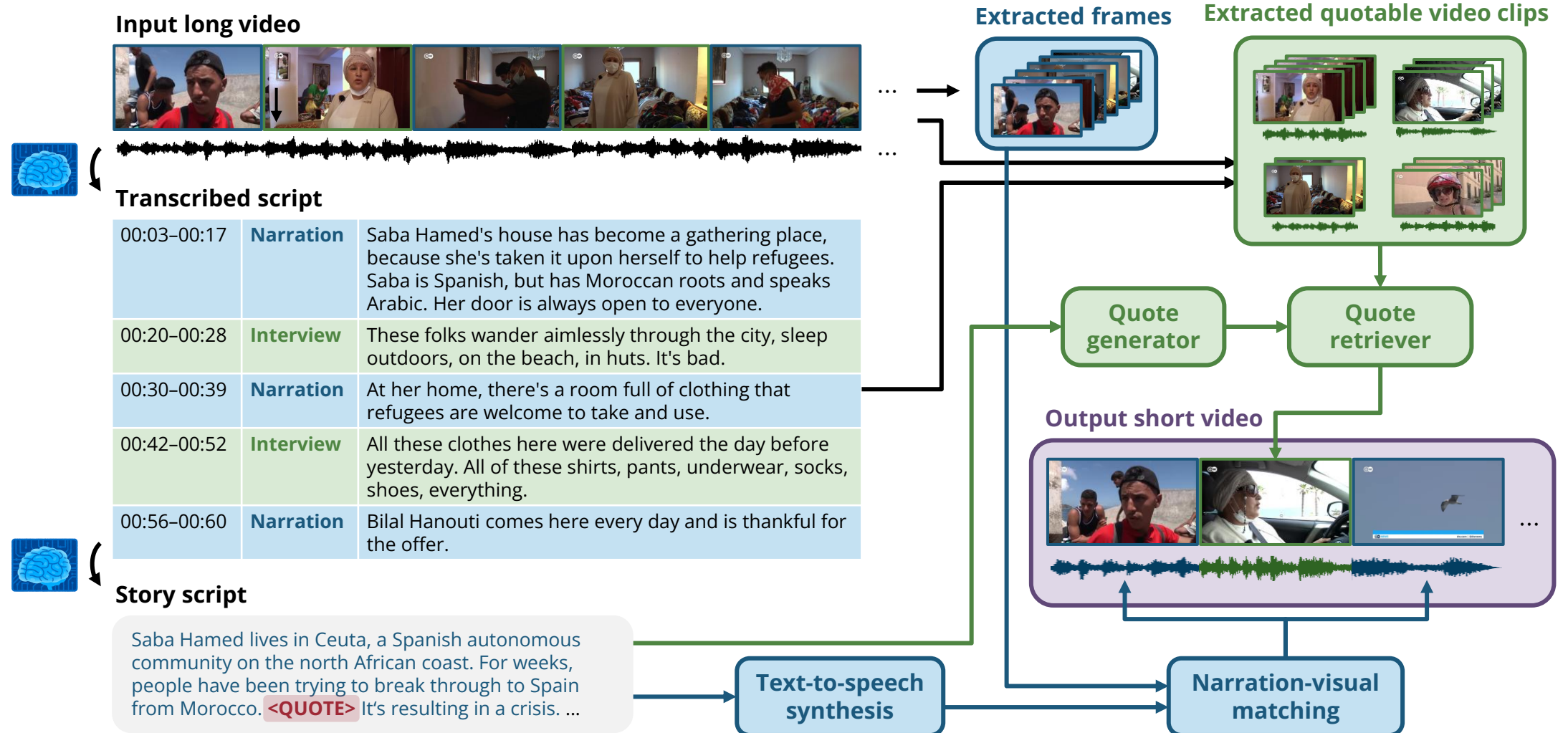
# Finding Accompanying Visuals for Narrations



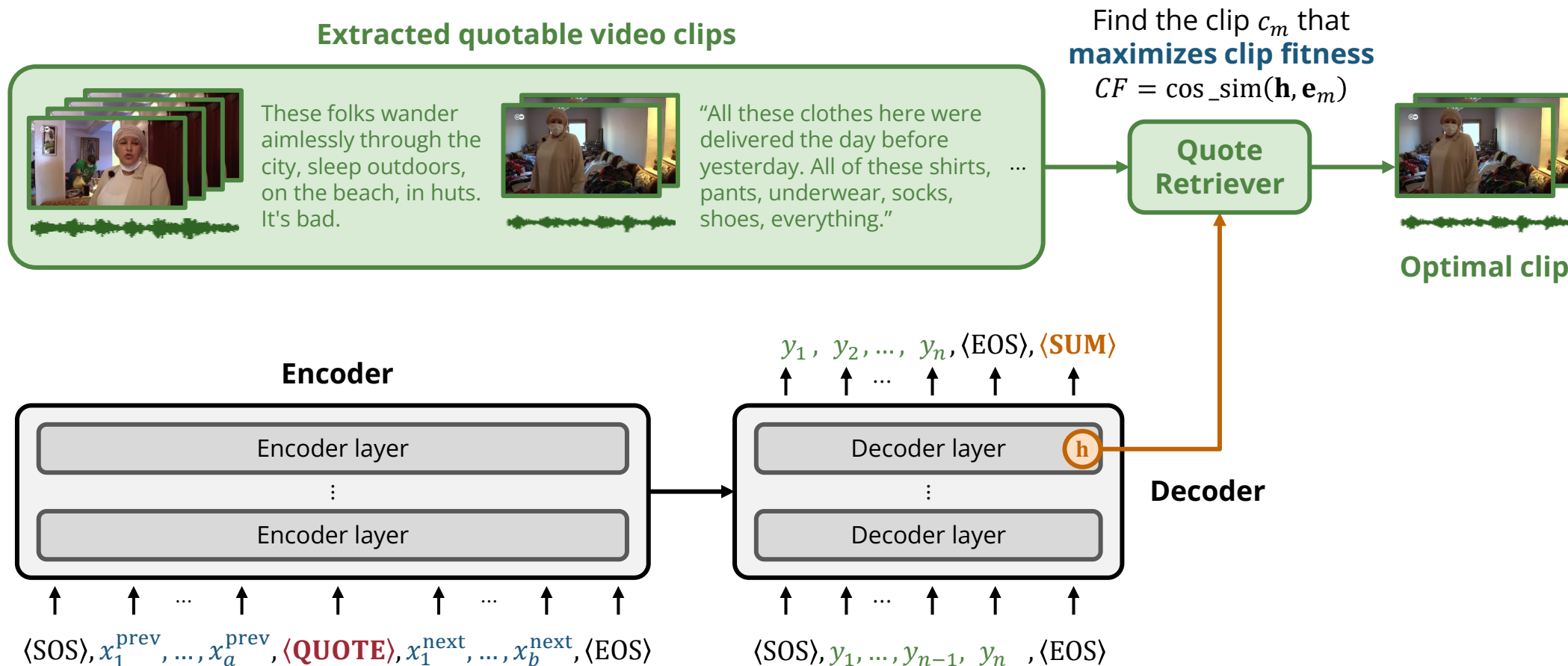
# Video Editing



# Learning to *Quote* a Video



# Retrieving a Video Quote

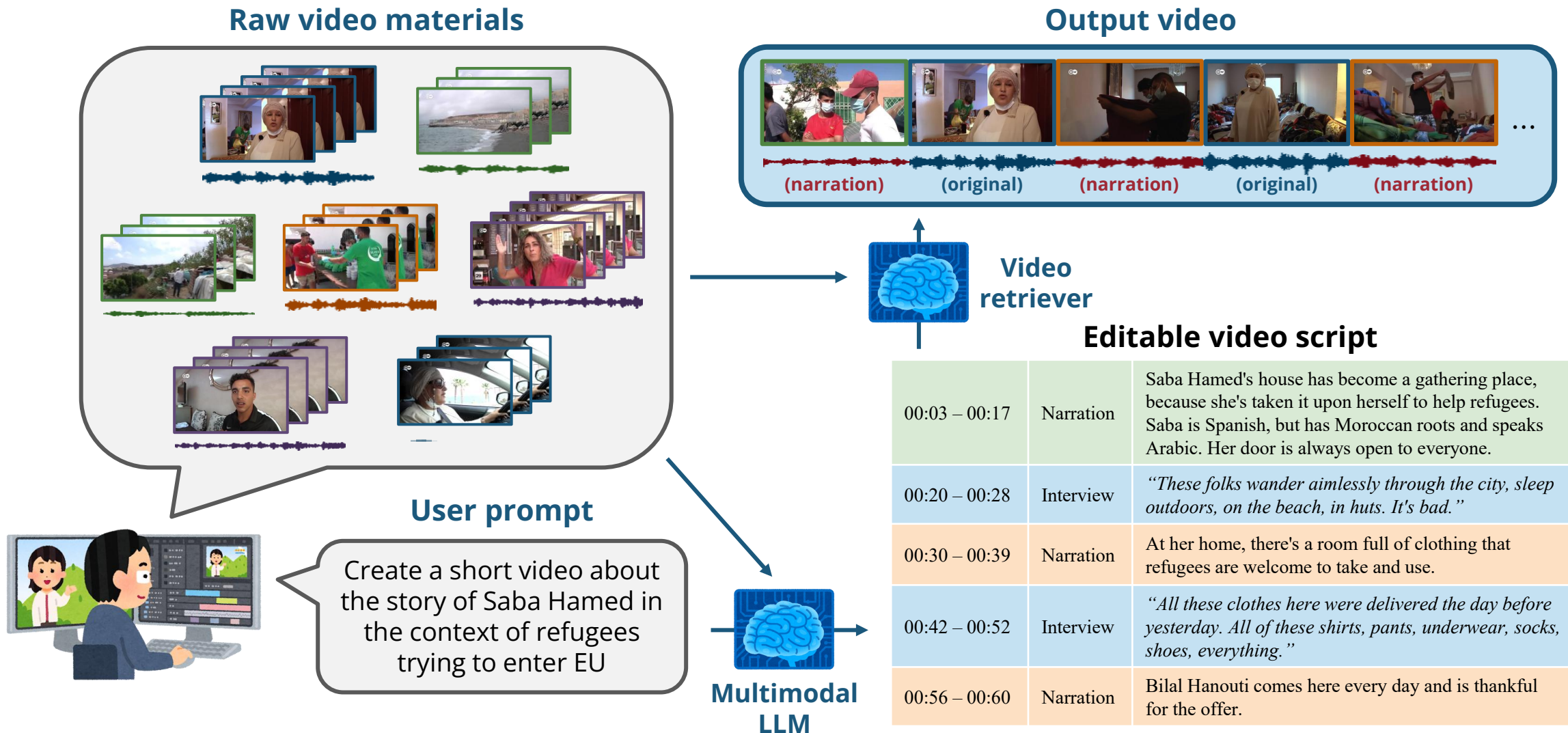


# Retrieval-Augmented → Retrieval-Embedded Generation

- Can an LLM **learn to quote** and **embed the quote properly**?
- How to **quote materials in other modalities**?
  - Audio, image, videos, sensor data, etc.
  - We need **a retriever to identify candidate quotable materials**
  - We need **a multimodal LLM that understands multimodal data** so that it can incorporate the retrieved materials and embed them properly

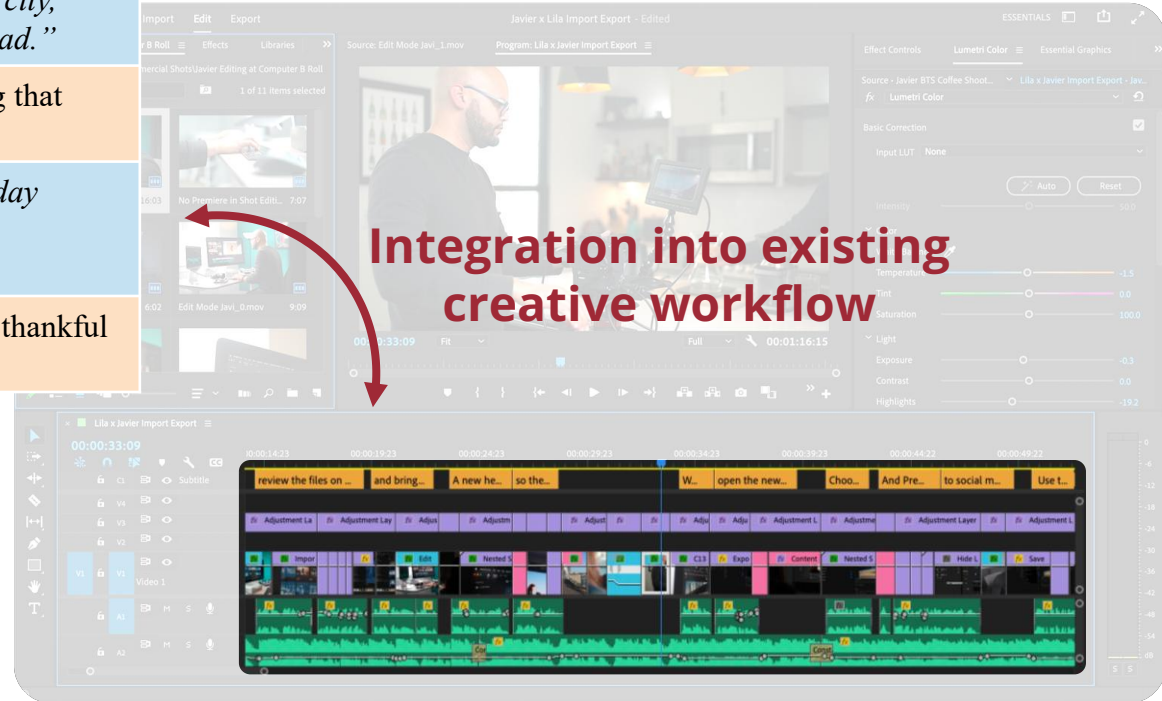
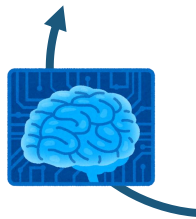


# Future Work: Multimodal RAG-based Video Editing



# Future Work: Integration into Video Editing Software

|               |           |  |
|---------------|-----------|--|
| 00:03 – 00:17 | Narration | Saba Hamed's house has become a gathering place, because she's taken it upon herself to help refugees. Saba is Spanish, but has Moroccan roots and speaks Arabic. Her door is always open to everyone. |
| 00:20 – 00:28 | Interview | <i>"These folks wander aimlessly through the city, sleep outdoors, on the beach, in huts. It's bad."</i>   |
| 00:30 – 00:39 | Narration | At her home, there's a room full of clothing that refugees are welcome to take and use.  |
| 00:42 – 00:52 | Interview | <i>"All these clothes here were delivered the day before yesterday. All of these shirts, pants, underwear, socks, shoes, everything."</i>  |
| 00:56 – 00:60 | Narration | Bilal Hanouti comes here every day and is thankful for the offer.  |



# Future Work: Integrating GenAI into Music Production



(Source: Avid)



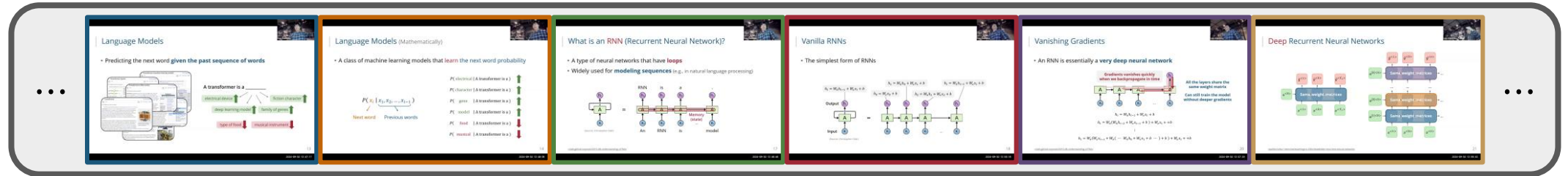
# Future Work: Integrating GenAI into Music Production



(Source: Avid)

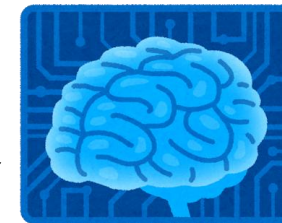
# Future Work: LectureRecap

## Lecture recording

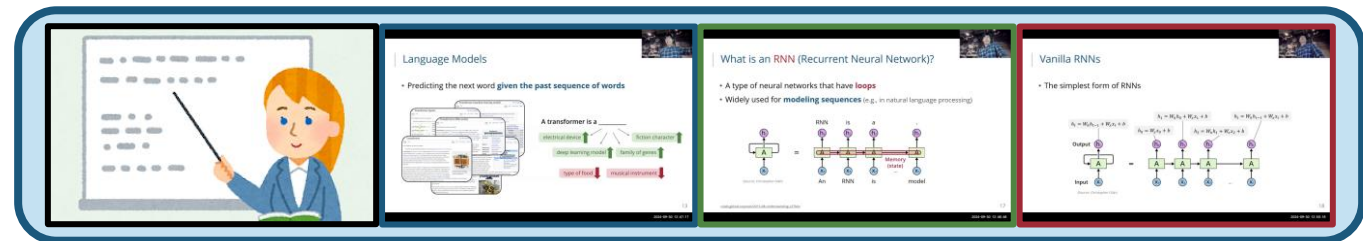


I would like to **review** the concept of **recurrent neural networks**. How does an RNN work?

Can you explain the **math** behind it?



LectureRecap

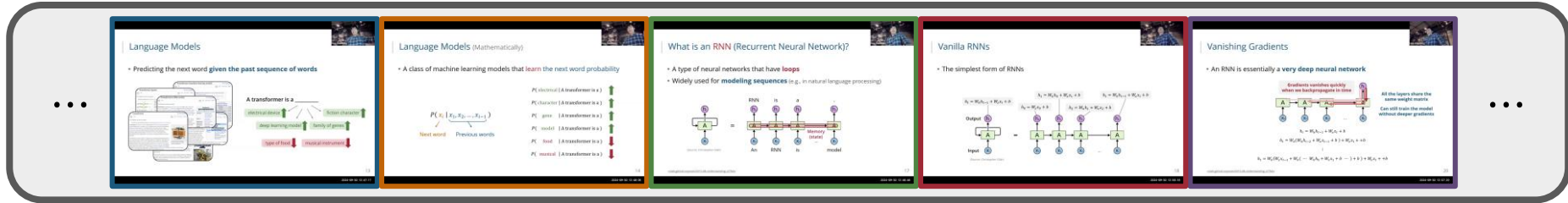


Lecture recap



# Future Work: LectureRecap

Lecture  
recording



User query

I would like to **review** the concept of **recurrent neural networks**.

Script generation

Speech recognition

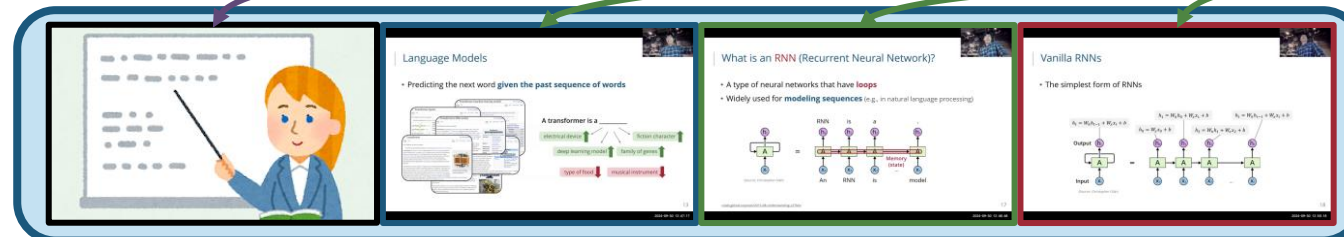
Video script

[**Narration**] Recurrent neural networks are a class of deep neural networks that ...  
[**Video clip insertion (10:24–12:48)**] Now let's first look at language models ...  
[**Video clip insertion (15:10–16:30)**] So what is a recurrent neural network? Intuitively, ...  
[**Video clip insertion (20:48–23:45)**] Mathematically, we can define an RNN as ...

Text-to-speech synthesis & talking head generation

Video clip extraction

Lecture recap



🔥 Ongoing Work: Playful Music GenAI 🔥



# Maestro VR (2024)



[youtu.be/OffnSNxidiY](https://youtu.be/OffnSNxidiY)



# SuperConductor

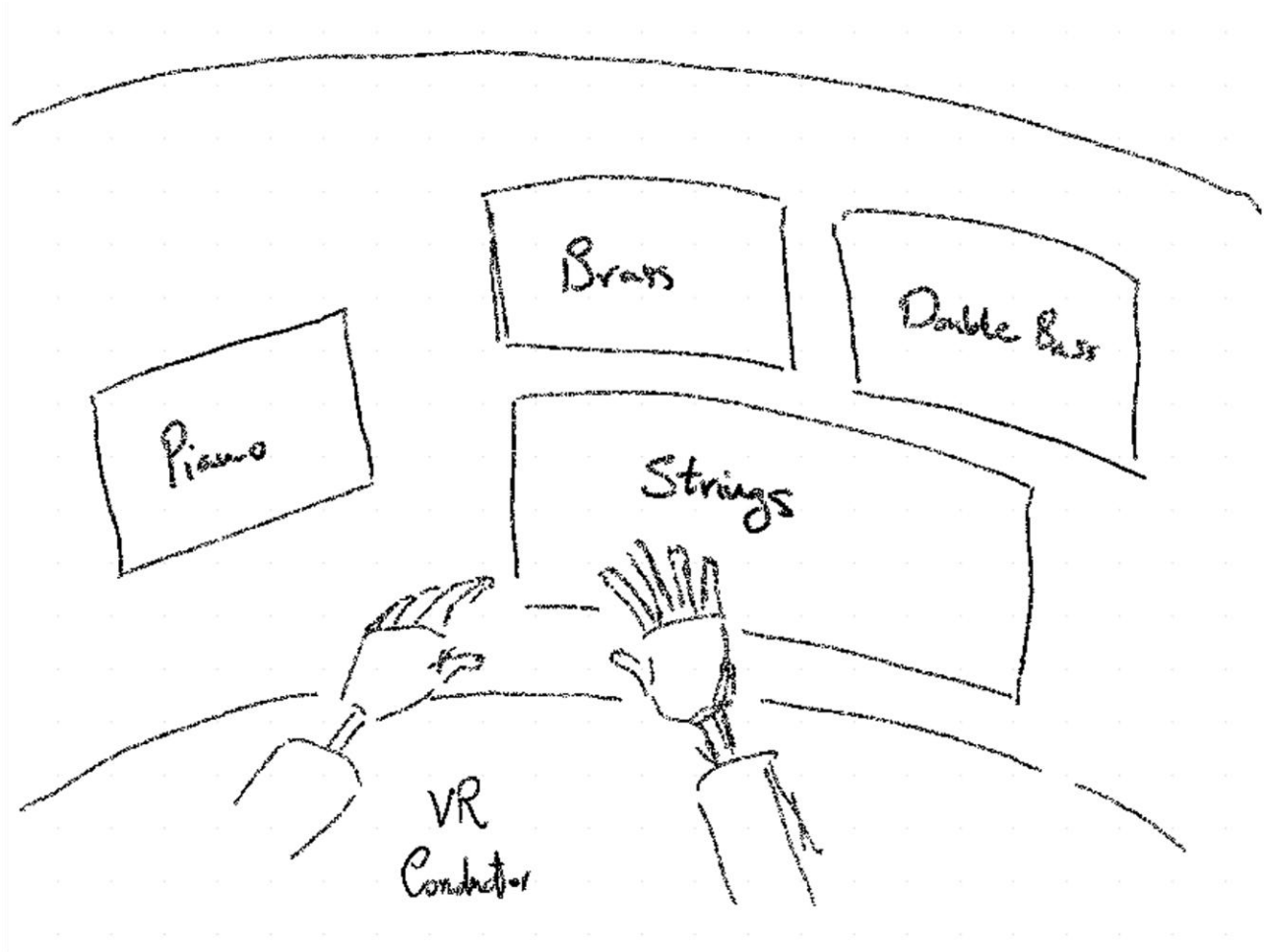


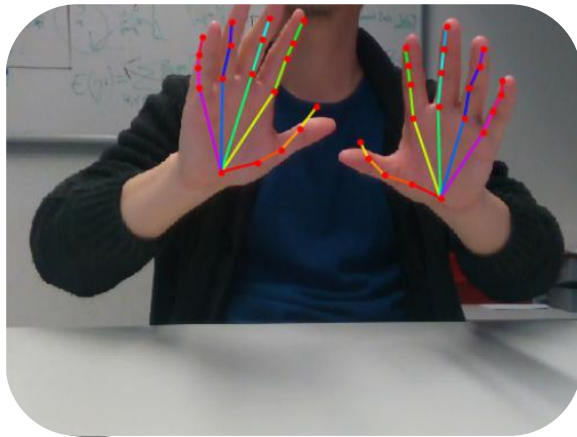
Illustration by Erfun Ackley



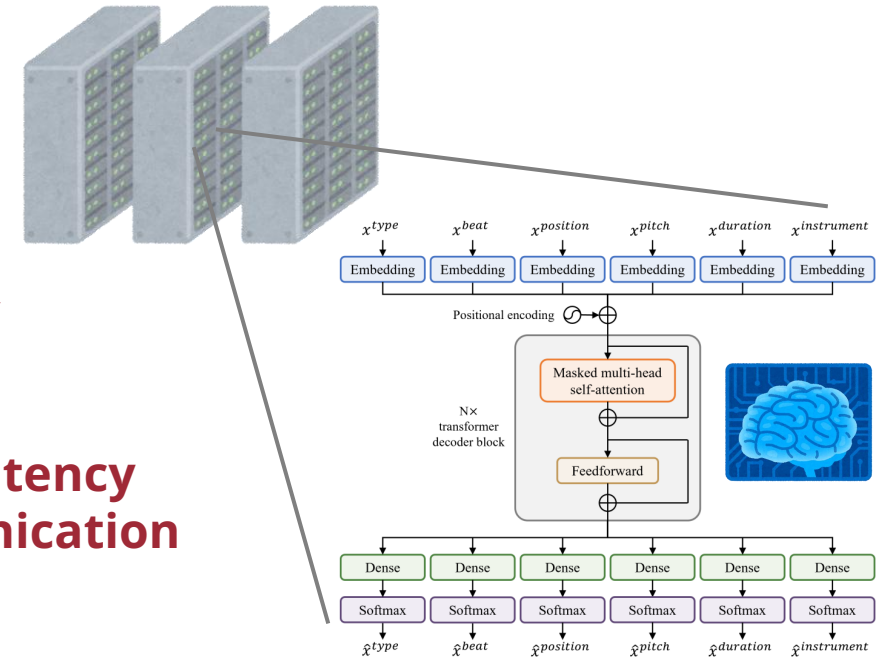
# SuperConductor



(Source: Wang et al., 2020)



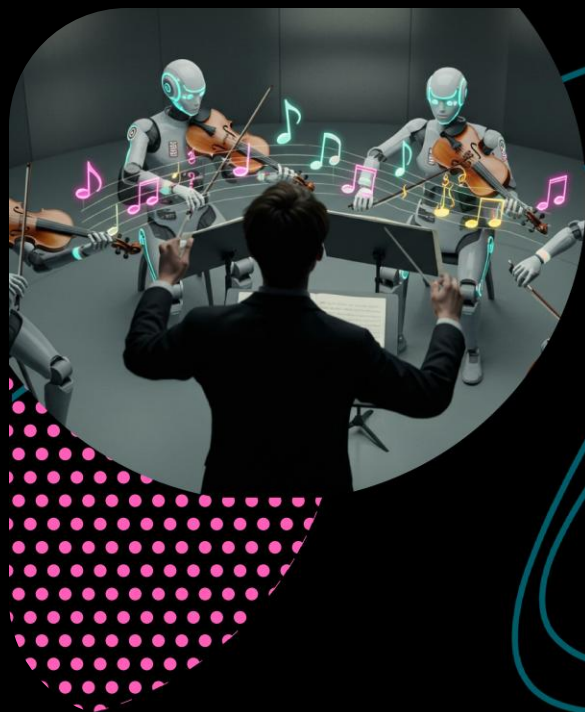
**Low latency  
communication**



(Source: Dong et al., 2023)

Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt, "RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video," *SIGGRAPH Asia*, 2020.

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *ICASSP*, 2023.



# **SUPERCONDUCTOR:** **EXPLORING PLAYFUL HUMAN-AI** **MUSIC CO-CREATIVITY**

UARTS FACULTY ENGINEERING/ARTS STUDENT TEAMS (FEAST)  
SCHOOL OF MUSIC, THEATRE & DANCE

**HAO-WEN DONG**  
Assistant Professor  
School of Music, Theatre & Dance

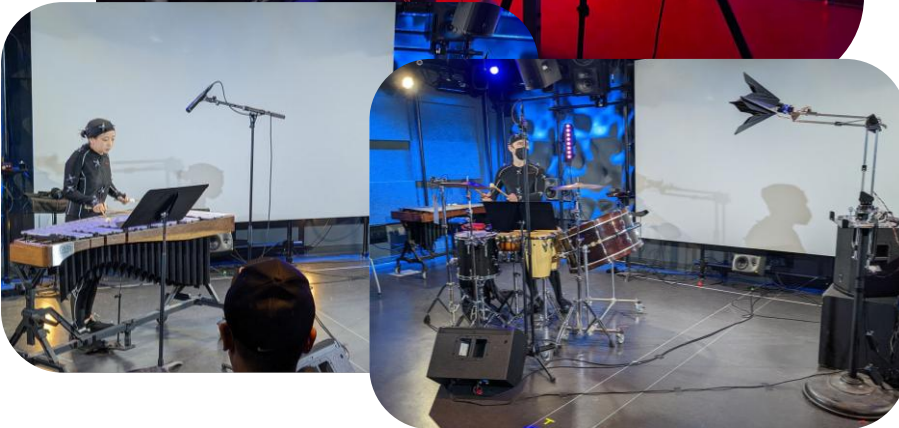


**Winter-Fall 2026**

**UARTS FEAST**

Join an interdisciplinary research team  
Apply by October 12

# Performing Arts Technology (PAT)



# Performing Arts Technology (PAT)

## Intro

PAT 100: Music in Technology  
PAT 200: Intro to Electronic Music

## PAT Theory/Studies

PAT 150: Experiential Music Theory  
PAT 205: Intermedia AI Music Practice  
PAT 305: Video Game Music  
PAT 315: Diversity in Music Technology  
PAT 316: NOISE

## PAT Practice

PAT 202: Computer Music  
**PAT 204: Creative Coding**  
PAT 220: Songwriting Workshop  
PAT 280: Sound Reinforcement  
PAT 412: Digital Music Ensemble  
PAT 413: Electronic Chamber Music

## Electives

PAT 421: Advanced Psychoacoustics  
PAT 422: Technical Ear Training & Critical Listening  
PAT 424: Dialog of the Senses  
PAT 431/432: Contemporary Practice in Studio Production I/II  
PAT 441: Sound for Film and Games  
PAT 443: Immersive Media  
PAT 451/452: Interactive Media Design I/II  
PAT 454: Digital Fabrication for Acoustics  
PAT 461: Performance Systems  
PAT 462: Digital Sound Synthesis  
**PAT 463: Music & AI**  
**PAT 464: Generative AI for Music & Audio Creation**  
PAT 472: Business of Music





# Generative AI for Music & Audio Creation

**PAT 464/564 (Winter 2026)**

## **Generative AI for Music and Audio Creation**

Learn about all the latest music and  
audio generation models

PAT 464/564 (Winter 2026)  
Instructor: Hao-Wen Dong



**PERFORMING ARTS TECHNOLOGY**  
UNIVERSITY OF MICHIGAN



**Art challenges Technology**



**Music**

**Augmenting Human Creativity  
with AI**



**AI**



**Technology inspires the Art**

# Augmenting Human Creativity with AI

- **Generative Models for Music Creation**

- **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024), **text-to-music generation** (ISMIR 2025), **video-to-music generation** (ISMIR 2025), **symbolic music processing tools** (ISMIR LBD 2019, ISMIR 2020)

- **AI-assisted Music Creation Tools**

- **Expressive violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020), **a cappella source separation** (ISMIR LBD 2025)

- **Multimodal Generative Models for Content Creation**

- **Long-to-short video editing** (ICLR 2025, NeurIPS 2025), **text-queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WASPAA 2023)

# Generative AI for Music, Audio & Video Creation



Universitaetsmedizin, CC BY-SA 4.0, via Wikimedia Commons  
[uploadvr.com/iron-man-vr-breaks-free-from-cords-load-screens-on-quest-2/](https://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](https://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/](https://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](https://dailybruin.com/2023/08/04/theater-review-the-musical-les-misrables-offers-stellar-displays-and-impassioned-vocals)

# Augmenting Human Creativity with AI

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



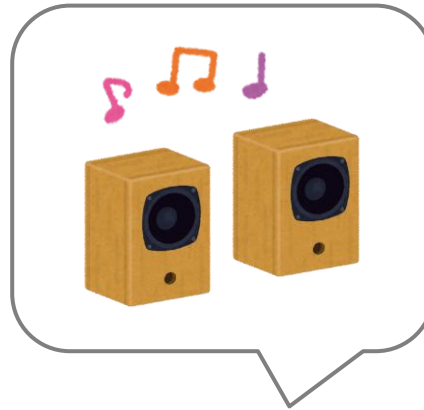
**Works of art make rules;  
rules do not make works of art.**

– Claude Debussy

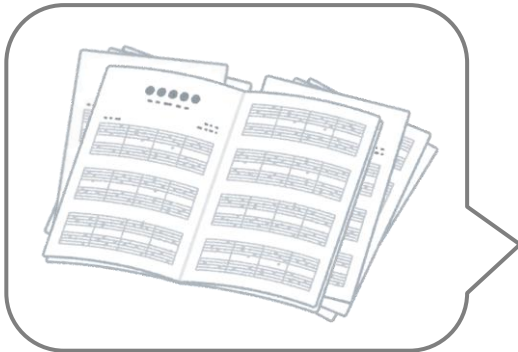


# Human-inspired Machine Learning for Music & Audio

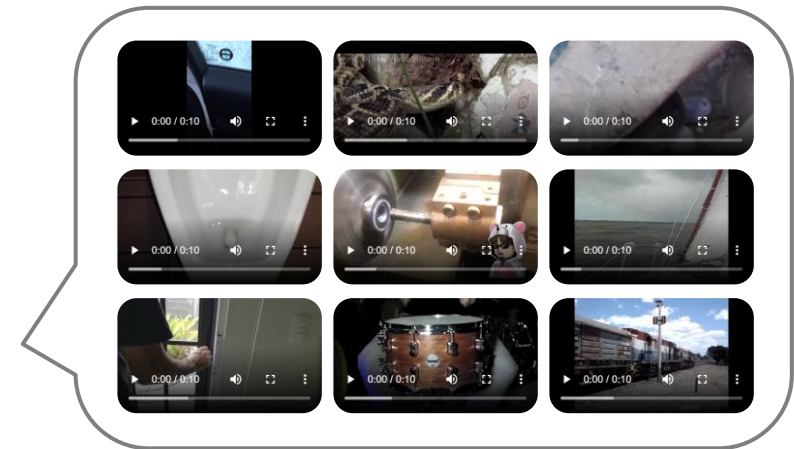
Learning from listening



Learning from reading



Learning from watching



# Misusable Music Tools (Nao Tokui, 2024)

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.**

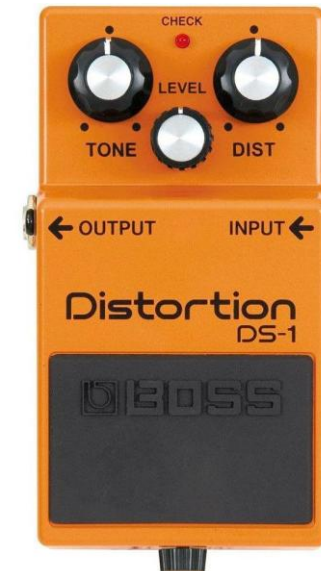
– Nao Tokui, 2024



(Source: Flintmi via [Wikimedia Commons](#))

# Overfitting vs Distortion

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



Without **deviation from the norm**,  
progress is not possible.

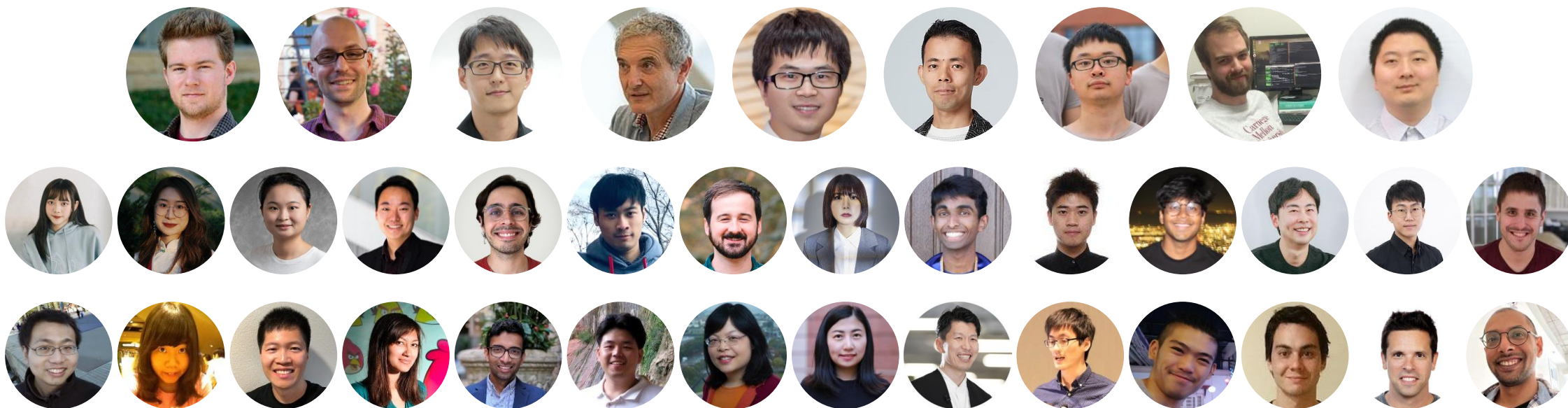
– Frank Zappa





# Augmenting Human Creativity with AI

**Nothing would have been possible without all my fantastic collaborators!**



UC San Diego



SONY



[hermandong.com](http://hermandong.com) / [hwdong@umich.edu](mailto:hwdong@umich.edu)

