

# Augmenting Human Creativity with Generative AI

**Hao-Wen (Herman) Dong**

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

March 2, 2026

# 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/)

**Art challenges Technology**



**Creativity**

**Augmenting Human Creativity  
with AI**



**AI**



**Technology inspires the Art**

# 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

- **Novel Generative Models for New Domains**
  - **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 Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **Long-to-short video editing** (ICLR 2025, NeurIPS 2025), **text-queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WASPAA 2023)

# Video Editing



**Interview footage**  
(main character)



**Background footage**



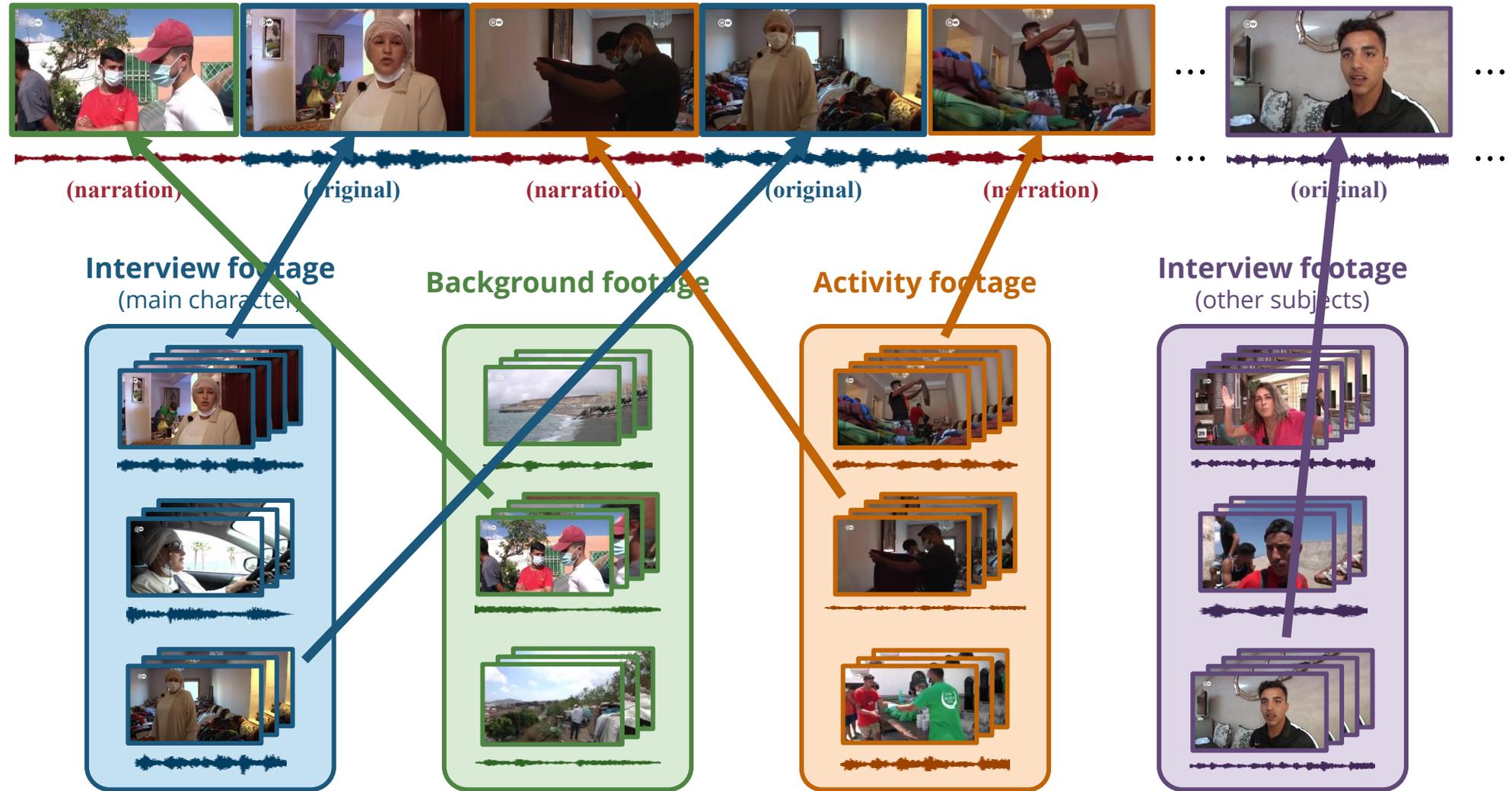
**Activity footage**



**Interview footage**  
(other subjects)



# 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



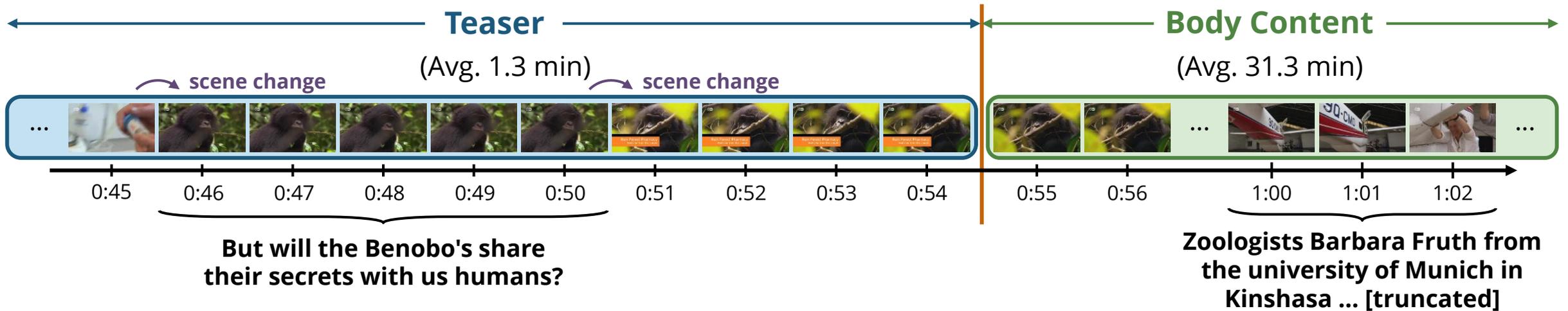
# TeaserGen: Generating Teasers for Long Documentaries

Weihan Xu<sup>1</sup> Paul Pu Liang<sup>2</sup> Haven Kim<sup>3</sup>  
Julian McAuley<sup>3</sup> Taylor Berg-Kirkpatrick<sup>3</sup> **Hao-Wen Dong<sup>4</sup>**

<sup>1</sup> Duke University <sup>2</sup> MIT <sup>3</sup> UC San Diego <sup>4</sup> University of Michigan



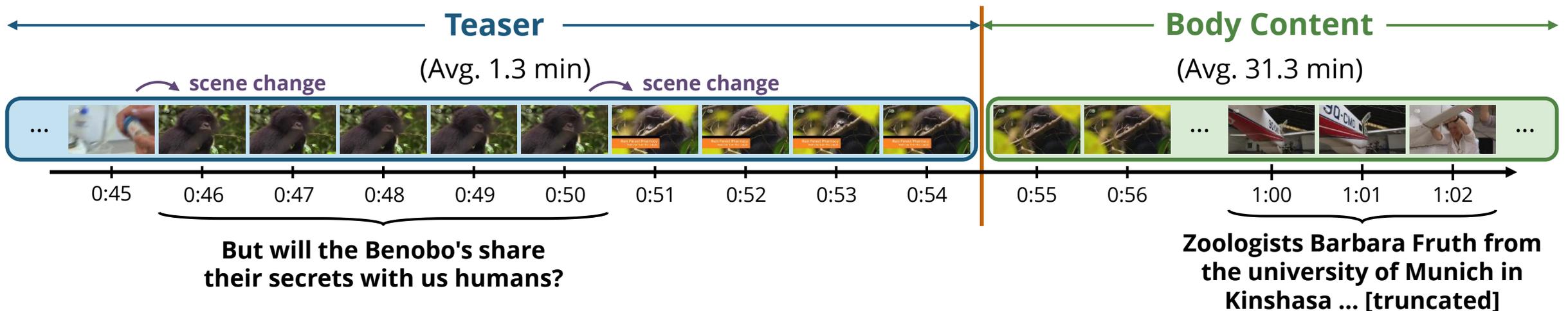
# DocumentaryNet: A New Documentary Dataset



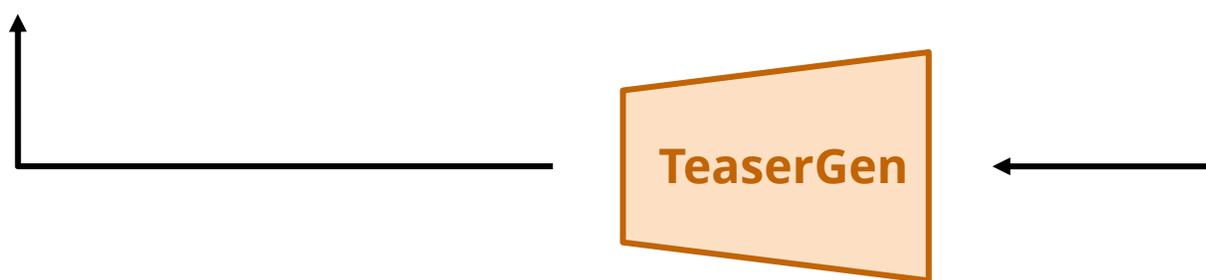
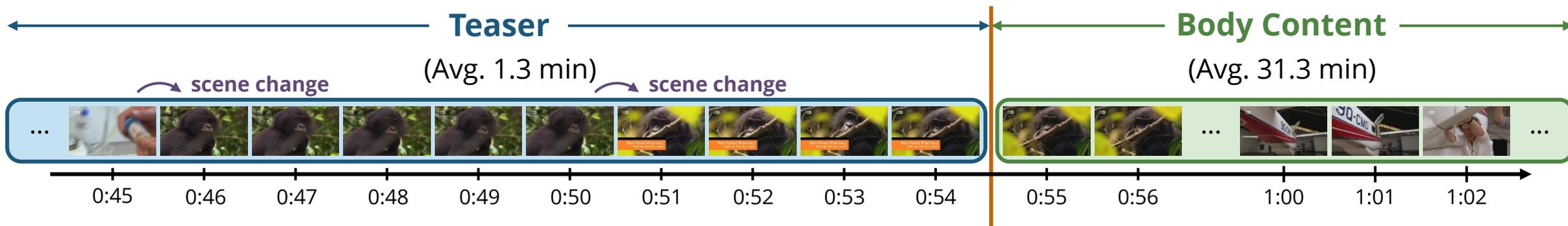
# DocumentaryNet: A New Documentary Dataset



- **1,269** high-quality documentaries paired with **teasers**
- **689 hours** in total
- Three reputable sources: **DW, PBS, National Geographic**



# Generating Teasers from Long Documentaries



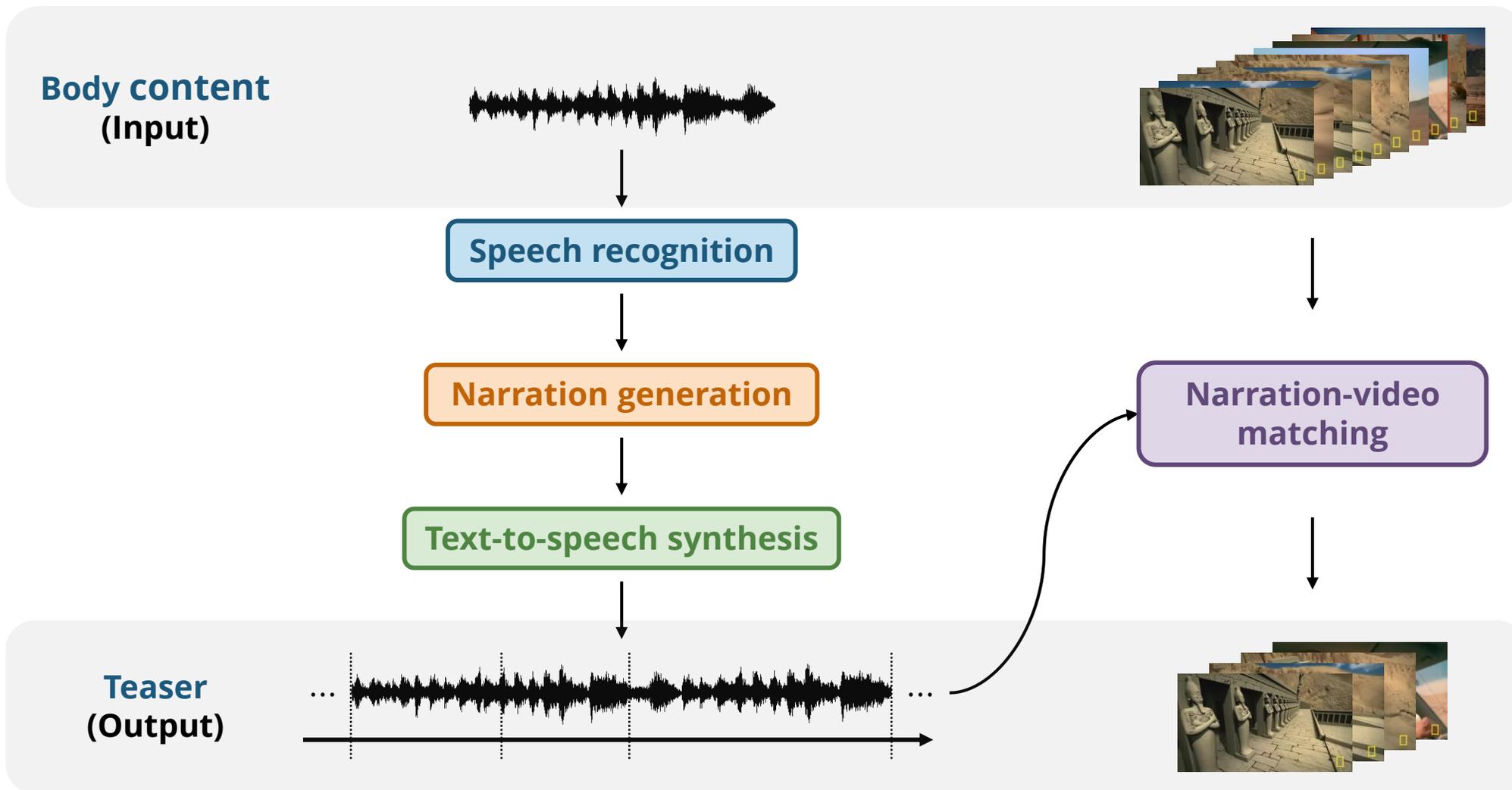
**~24:1 compression ratio**

# Documentary Teaser Generation



- Unlike **video highlight detection**, a teaser **needs a cohesive narrative**
- Unlike **video summarization**, a teaser **needs to be interesting and engaging**
- Unlike a **movie trailer**, a documentary teaser is more **narration-focused**
- A documentary teaser **needs to preserve the factual accuracy**

# Narration-Centered Long-to-Short Video Editing

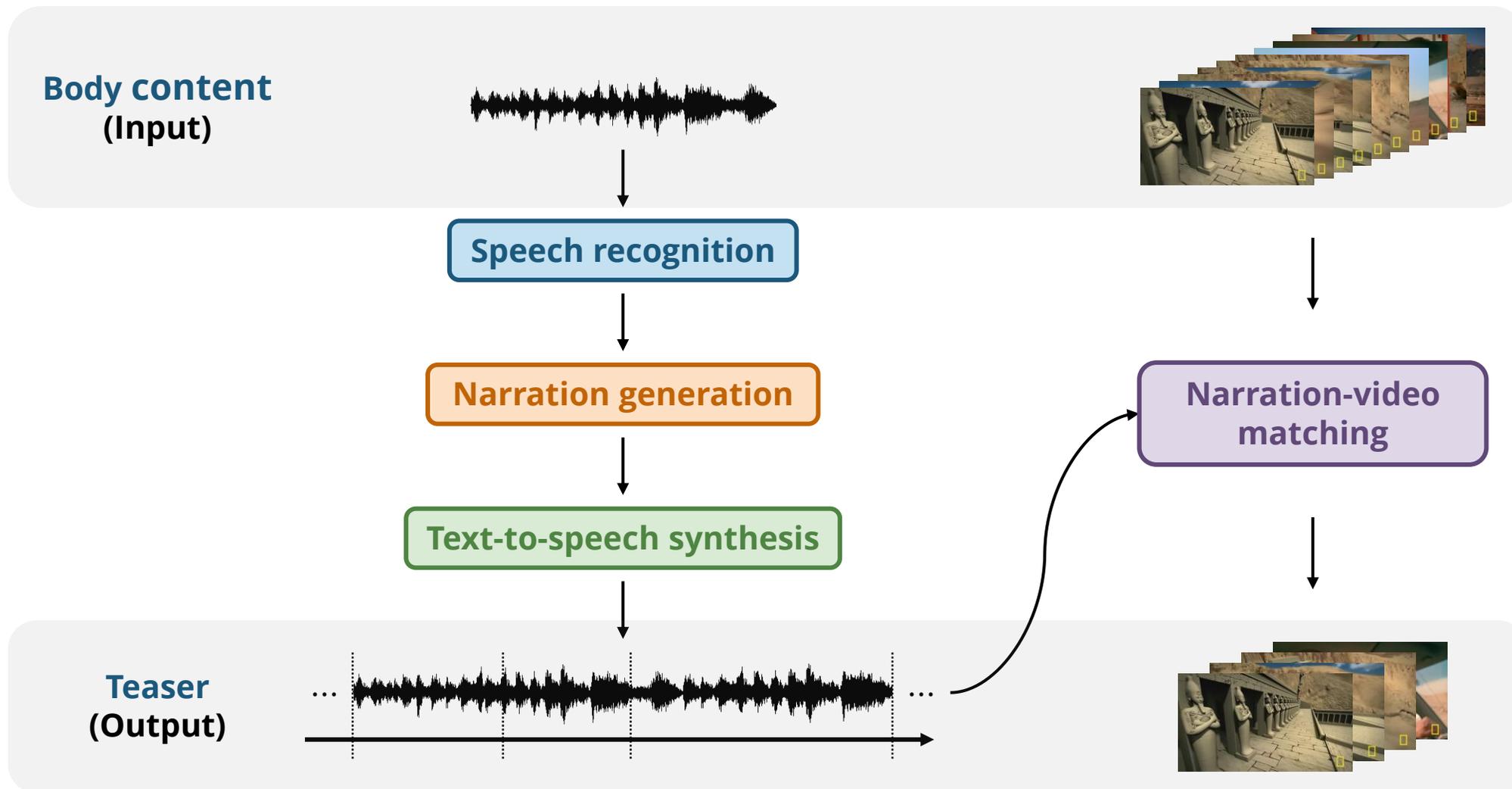


# Leveraging LLMs for Narration Generation



- Break the full narration into **10 segments** (avg. 3,900 words)
- Use GPT-4o to **summarize each segment**
- **Rewrite** the 10 summarized sentences **into a cohesive paragraph**
  - “Rewrite the paragraph into an engaging story opening in 10 sentences or less, keeping all names and avoiding being replaced by pronouns.”
- **Propose an ending question**
  - “Given the title and the provided summary, formulate one thought-provoking and concise question that relate directly to the summary.”
  - For example, *“But will the Benobo's share their secrets with us humans?”*

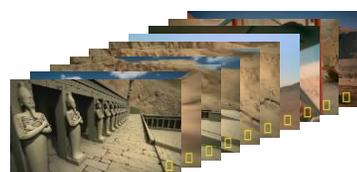
# Narration-Centered Long-to-Short Video Editing



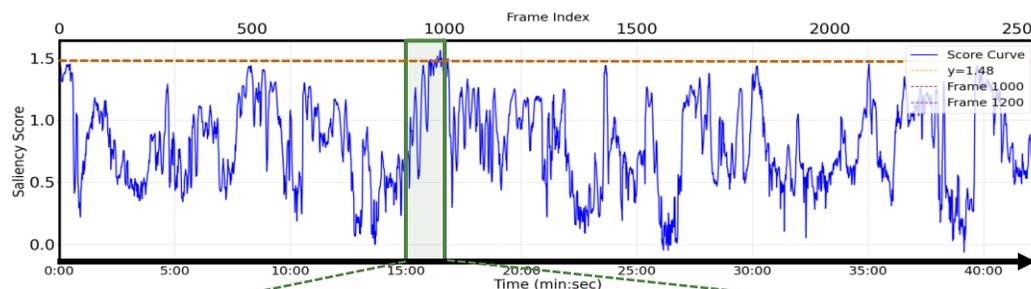
# Finding Accompanying Visuals for Narrations



Body content

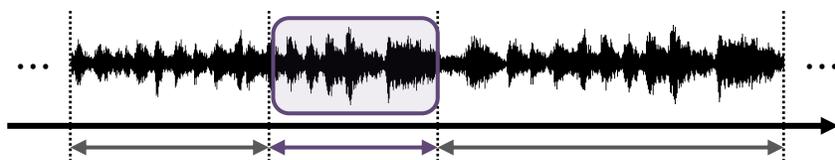


Pretrained vision-language model



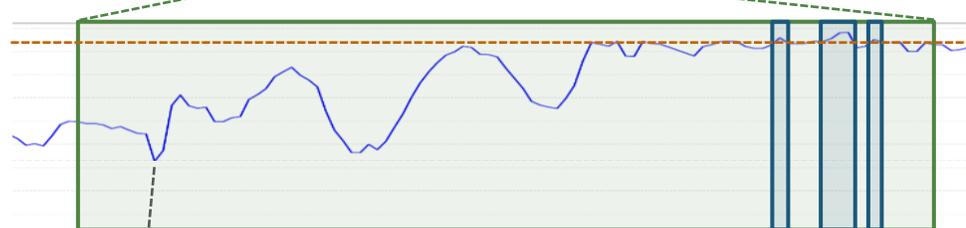
Optimal threshold

The air buzzed with excitement as John and Maria unearthed a nearly 10-foot-tall abandoned sphinx statue at one of Egypt's largest archaeological sites.



Desired length

Synthesized narration



Optimal threshold



Desired length

Selected video clips

# Example Results



## Ground truth

-  Egypt, the richest source of archaeological treasures on the planet
-  Beneath this desert landscape, why the secrets of this ancient civilization?
-  Wow! You can see why a Pharaoh's chosen place
-  for a full season of excavations
-  our cameras have unprecedented access, follow teams on the front line of archaeology.
-  I'm driving so fast because I'm excited!
-  It's an entrance, we can see an entrance.
-  I have just been told that they have found something
-  Revealing varied secrets.
-  Making discoveries, they could rewrite ancient history.
-  This time, new secrets about one of Egypt's greatest rulers the Pharaoh Queen, Hatshepsut
-  Doctor Ranski discovers very treasures that her magnificent temple had reached was to be remembered for millions of years.
-  For my beloved daughter, not son, and John and Maria, honor a rare and intriguing statue.

## TeaserGen-PT

-  Under the scorching Egyptian sun, Dr. Zbigniew Szybranski led a team of archaeologists in Luxor, meticulously excavating the unique temple of the revolutionary female pharaoh Hetshepsood.
-  The air buzzed with excitement as John and Maria unearthed a nearly 10-foot-tall abandoned sphinx statue at one of Egypt's largest archaeological sites.
-  Meanwhile, in Aswan, Martina Bartanova's team stumbled upon ancient human remains that could halt their exploration of an unopened tomb.
-  The discovery of a child's remains and a miniature sphinx added layers of mystery to their quest.
-  At Karnak Temple, John and Maria delved into inscriptions revealing Hetshepsood's portrayal as a male pharaoh, while Yale professors used digital technology to study ancient texts.
-  In Dra'abu El Naga, another team meticulously organized fragmented human remains, uncovering stories of women in ancient Egyptian society.
-  American archaeologist Susanne Onstein explored the monumental building campaigns of Hetshepsood, whose colossal obelisks stood as testaments to her reign.
-  Amidst pottery fragments and evidence of temples built by Hetshepsood and her stepson Tutmose III, the archaeologists faced personal reflections and challenges.
-  The winds of history whispered through the sands, as each discovery brought them closer to unraveling the enigmatic legacy of Pharaoh Queen Hetshepsood.
-  How do the discoveries and restoration efforts at Hatshepsut's temple and other archaeological sites in Egypt contribute to our understanding of her reign and the broader role of women in ancient Egyptian society?

## TeaserGen-LR

-  Under the scorching Egyptian sun, Dr. Zbigniew Szybranski led a team of archaeologists in Luxor, meticulously excavating the unique temple of the revolutionary female pharaoh Hetshepsood.
-  The air buzzed with excitement as John and Maria unearthed a nearly 10-foot-tall abandoned sphinx statue at one of Egypt's largest archaeological sites.
-  Meanwhile, in Aswan, Martina Bartanova's team stumbled upon ancient human remains that could halt their exploration of an unopened tomb.
-  The discovery of a child's remains and a miniature sphinx added layers of mystery to their quest.
-  At Karnak Temple, John and Maria delved into inscriptions revealing Hetshepsood's portrayal as a male pharaoh, while Yale professors used digital technology to study ancient texts.
-  In Dra'abu El Naga, another team meticulously organized fragmented human remains, uncovering stories of women in ancient Egyptian society.
-  American archaeologist Susanne Onstein explored the monumental building campaigns of Hetshepsood, whose colossal obelisks stood as testaments to her reign.
-  Amidst pottery fragments and evidence of temples built by Hetshepsood and her stepson Tutmose III, the archaeologists faced personal reflections and challenges.
-  The winds of history whispered through the sands, as each discovery brought them closer to unraveling the enigmatic legacy of Pharaoh Queen Hetshepsood.
-  How do the discoveries and restoration efforts at Hatshepsut's temple and other archaeological sites in Egypt contribute to our understanding of her reign and the broader role of women in ancient Egyptian society?

# Objective Evaluation



Model	Query	Decoding	DP	Repetitiveness			Text-visual correspondence	
				F1 (%)↑	REP (%)	SCR (%)	CLIPScore	VTGHLS
<b>Baseline models</b>								
Random	Random	-	-	1.67	4.05	7.81	0.56	0.75
CLIP-NN	Narration	Greedy	×	0.11	92.73	8.29	0.69	0.79
UniVTG (2023b)	Title	Rank	-	1.82	0	89.68	0.58	1.01
CLIP-it (2021b)	Narration	Rank	×	1.24	0	99.39	0.56	0.61
<b>Pretraining-based models</b>								
TeaserGen-PT	Title	Thresholding	-	1.85	0	13.16	0.56	1.02
TeaserGen-PT	Narration	Thresholding	-	1.07	21.38	22.58	0.58	1.45
TeaserGen-PT-CLIP	Narration	Threshold	×	1.31	27.23	24.10	0.58	0.74
<b>Learning-based models</b>								
TeaserGen-LR	Narration	Greedy	×	1.56	31.97	27.18	0.58	0.74
TeaserGen-LR	Narration	Greedy	✓	1.38	26.83	35.48	0.62	0.78
TeaserGen-LR	Narration	Beam search	×	<b>1.88</b>	24.16	41.97	0.58	0.74
TeaserGen-LR	Narration	Beam Search	✓	<b>1.88</b>	19.39	46.56	0.63	0.77
Ground truth	-	-	-	100	>7.86	27.6	0.58	0.64

Scene change rate

**Check out our paper for more results!**

# Subjective Evaluation



Model	Query	Decoding	Coherence $\uparrow$	Alignment $\uparrow$	Engagingness $\uparrow$	Realness $\uparrow$
UniVTG (2023b)	Title	Rank	$2.61 \pm 0.50$	$2.62 \pm 0.47$	$2.67 \pm 0.57$	$2.66 \pm 0.54$
CLIP-it (2021b)	Narration	Rank	$2.61 \pm 0.46$	$2.67 \pm 0.44$	$2.57 \pm 0.46$	$2.51 \pm 0.46$
TeaserGen-PT	Title	Threshold	<b><math>3.14 \pm 0.50</math></b>	$2.84 \pm 0.57$	<b><math>2.81 \pm 0.49</math></b>	<b><math>2.94 \pm 0.50</math></b>
TeaserGen-LR	Narration	Greedy	$2.90 \pm 0.45$	<b><math>2.88 \pm 0.48</math></b>	$2.71 \pm 0.42$	$2.71 \pm 0.44$
TeaserGen-LR	Narration	Beam search	$2.84 \pm 0.46$	$2.69 \pm 0.51$	$2.71 \pm 0.42$	$2.64 \pm 0.41$

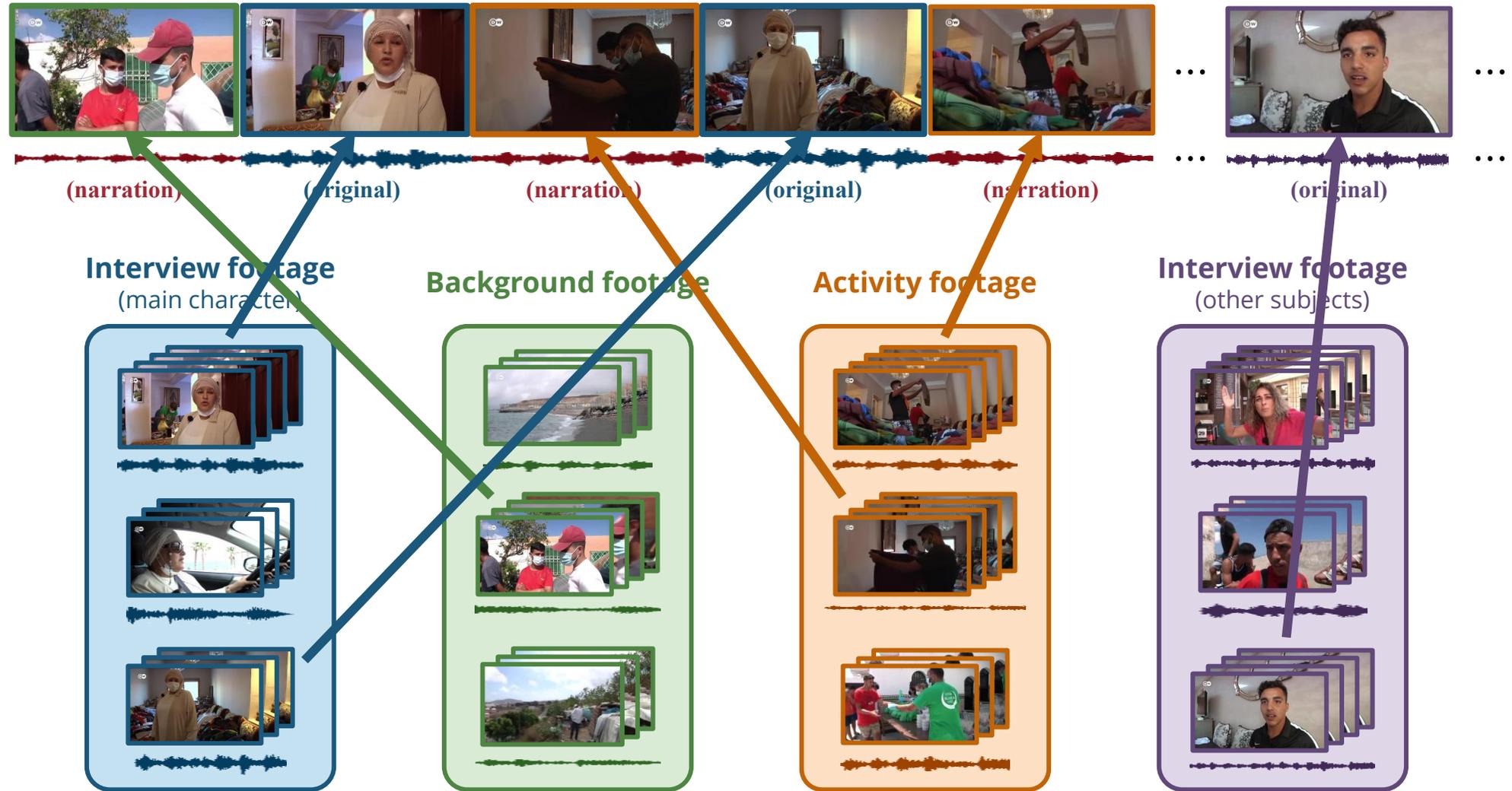
**TeaserGen-PT (interval-based) is more effective at identifying relevant visual content than TeaserGen-LR (learning-based)**

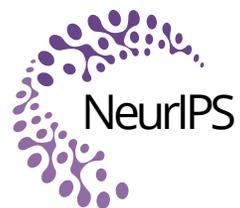
# Limitations



- Assumed that **narration plays a more significant role** than visuals
  - This assumption might not hold for movies and vlogs
- Teaser generation is a **one-to-many** mapping, i.e., a generative process
  - The model still falls short in terms of artistic quality and creativeness
- Cannot match **interview scenes** commonly seen in documentaries
  - Can a model learn to “quote” an interview?

# Video Editing





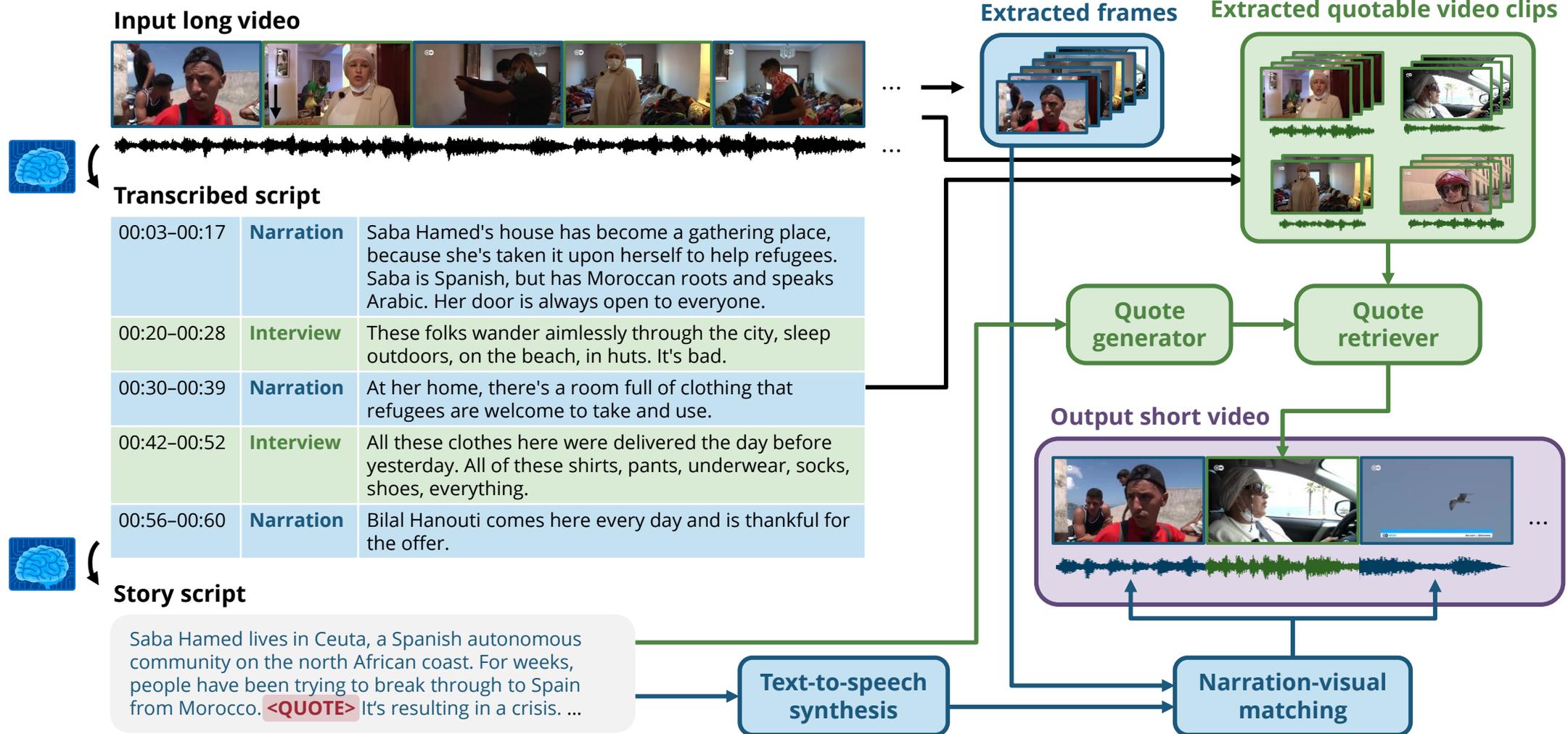
# REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing

Weihan Xu<sup>1</sup> Yimeng Ma<sup>1</sup> Jingyue Huang<sup>2</sup> Yang Li<sup>1</sup> Weyne Ma<sup>3</sup>  
Taylor Berg-Kirkpatrick<sup>2</sup> Julian McAuley<sup>2</sup> Paul Pu Liang<sup>2</sup> **Hao-Wen Dong**<sup>4</sup>

<sup>1</sup> Duke University <sup>2</sup> UC San Diego <sup>3</sup> MBZUAI <sup>4</sup> MIT <sup>5</sup> University of Michigan



# Learning to *Quote* a Video



# Learning to *Quote* a Video



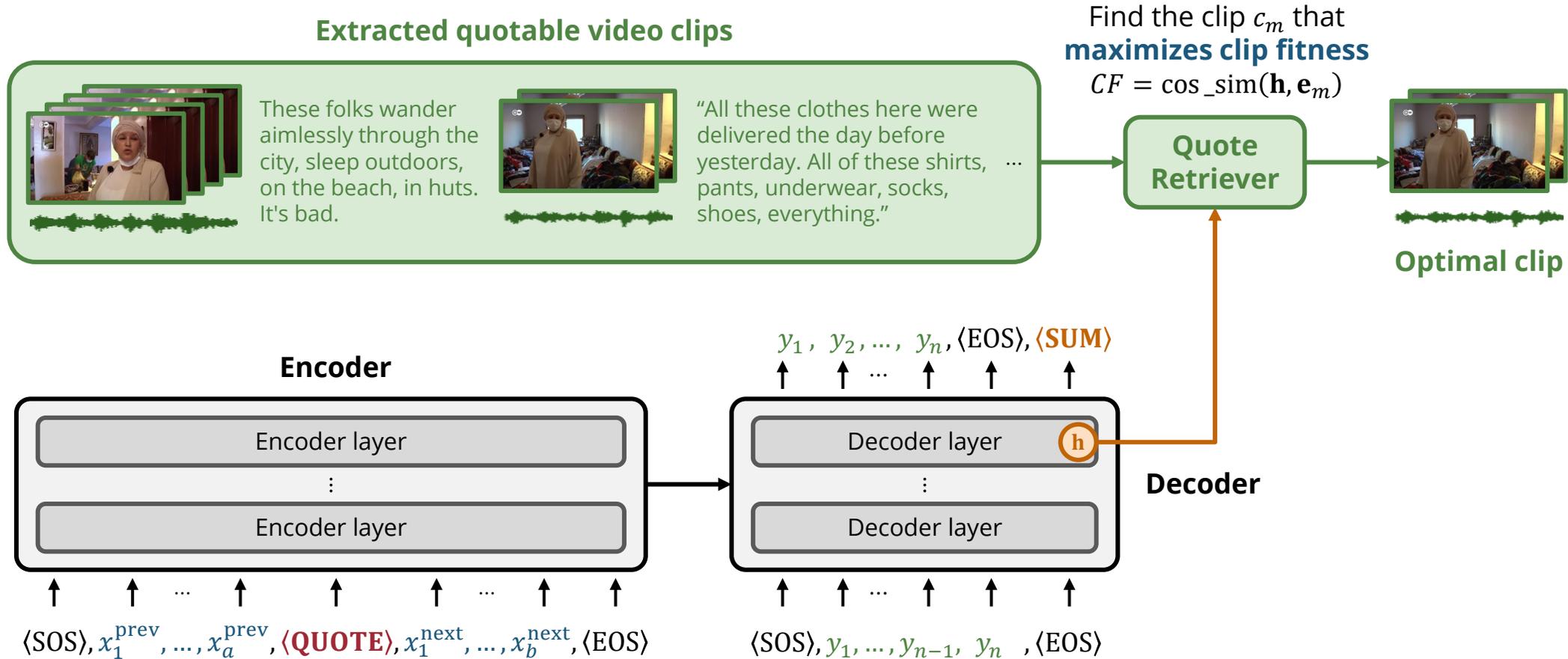
**REGen-DQ**  
(direct quote)

Quote  
↑  
...,  $x_i$ , **<SOQ>**,  $y_1, \dots, y_n$ , **<EOQ>**,  $x_{i+1}$ , ...

**REGen-IDQ**  
(indirect quote)

...,  $x_i$ , **<QUOTE>**,  $x_{i+1}$ , ...  
↓  
**To be retrieved later!**

# Retrieving a Video Quote



# Measuring Clip Fitness



For a candidate clip  $c_m$ , the **clip fitness** is defined as

$$CF := \cos\_sim(\mathbf{h}, \mathbf{e}_m)$$

**REGen-IDQ-T**  
(text only)

$$\mathbf{e}_m = \mathbf{e}_m^{\text{text}}$$

**REGen-IDQ-TV**  
(text+video)

$$\mathbf{e}_m = f\left(\text{concat}\left(\mathbf{e}_m^{\text{text}}, \mathbf{e}_m^{\text{img}}\right)\right)$$

Learnable mapping

# Comparing Quote Retrieval Methods



Retriever	Similarity measure	Recall@1 (%)	Recall@5 (%)	Recall@10 (%)	Insertion effectiveness
Random	-	0.00 ± 0.00	0.28 ± 0.48	7.22 ± 5.54	3.08 ± 0.25
GPT-4o infilling	Text only	2.78 ± 0.48	13.89 ± 1.27	22.50 ± 1.44	2.48 ± 0.31
QuoteRetriever-T	Text only	<b>5.00</b>	<b>17.50</b>	<b>30.00</b>	<b>3.56 ± 0.22</b>
QuoteRetriever-TV	Text+Visual	<b>5.00</b>	15.00	23.33	3.49 ± 0.26

Retrieving with only text is better than retrieving with both text and video

# Example: Modern Art Exhibition



**Title:** "documenta 14 - learning from Athens | DW Documentary"



[youtu.be/agij\\_lxGjCI](https://youtu.be/agij_lxGjCI)

# Example Results



## REGen-IDQ-TV

**Narrator:** The crisis has given me a lot

**Narrator:** I've never before seen rents like they are right now.

**Narrator:** Lacks of money has opened up these opportunities for people like me to rent apartments in the city for such cheap

**Quotable Video Clip:** Athens is very free, very free. The down-economical level and the big freedom that we enjoy here in Athens gives you the point that you can do whatever you like

**Quotable Video Clip:** The Greeks have been through a lot of crisis and a lot of problems, and we are not like North Europeans, where we expect the state to come and take care of us. This has never been the case in Greece, because on a government level, it has never been very successful, let's say. So Greeks are kind of used to doing things by themselves. And so perhaps the crisis is one of those cases where the Greeks are called to find their own way to do stuff

**Quotable Video Clip:** We have an economic crisis. Many people have lost their jobs and there's no such thing as unemployment benefit. Everyday people I talk to in the taxi tell me they don't know how to go on.

## REGen-IDQ-T

**Narrator:** The crisis has given me a lot

**Narrator:** I've never before seen rents like they are right now.

**Narrator:** Lacks of money has opened up these opportunities for people like me to rent apartments in the city for such cheap

**Quotable Video Clip:** Athens is very free, very free. The down-economical level and the big freedom that we enjoy here in Athens gives you the point that you can do whatever you like

**Quotable Video Clip:** It's perhaps too early to see the changes in the city because of this current refugee crisis. The fact is that nothing is being built now in Athens. So the refugees that are coming now, they haven't had the chance yet to establish themselves

**Quotable Video Clip:** I was born in Athens, so my kind of sentiment is connected to this city. You know, my experiences, my childhood, my teenagehood is cultivated from the city, from the way the architecture, the everyday is created. It's not a decision to be here. It's an emotional responsibility."

## REGen-DQ

**Narrator:** It's dazzling, early morning light in Athens and a myriad of colors

**Narrator:** For a weekend, little parks have been created throughout the city for documenta 14

**Narrator:** Each park has its own artist and they're all based on themes, on human migration for example

**Narrator:** These banners will get flying off the start,

**Quotable Video Clip:** This title, Learning from Athens, describes a situation, a situation of people, even the Greek people, that we are learning on how one of these capitals of Europe now has been once the cradle of civilization and now is also this kind of place that has accumulated all the, so many miseries The Koumenda cannot change the economical crisis. It can give hope to people, mainly to the artistic scene. It's an exhibition that can steer up things, but not really change situations. So, I'm looking very forward to these 100 days of the Koumenda that will be for us a kind of an escape, a break

**Narrator:** It's art, it's documents, the art exhibition that takes place every other year

**Narrator:** This year it's right in the middle and in Greece for the first time since it was first held in 1972,

**Narrator:** The city Goths and documenta 14 head honcho Caroline Bock has given Athens Treemala, mid-generation, tremala, and research-age

**Narrator:** But even before documenta 14 has arrived, Athens has been fit for documenta, and this white and grey city could actually benefit from it

**Narrator:** The city needs the larger framework of a significant event,

# Example: Athens History



**Title:** “documenta 14 - learning from Athens | DW Documentary”



[wx83.github.io/REGen/](https://wx83.github.io/REGen/)

# Objective Evaluation



## Repetitiveness

Model	Dur (sec)	Interview ratio (%)	F1 (%)	SCR (%)	REP (%)	VTGHLS	CLIPS-I	CLIPS-N
Random extraction	101	56 ± 20	1.10	20.71	0.41	0.83	0.55	0.62
ETS	142	34 ± 16	1.92	13.65	4.49	1.06	0.64	0.60
A2Summ [4]	73	42 ± 25	1.70	14.20	1.73	0.89	0.56	0.63
TeaserGen [11]	155	-	1.64	<b>22.61</b>	21.38	0.80	-	0.67
GPT-4o-DQ	151	42 ± 42	1.56	16.55	20.75	1.01	0.58	0.42
GPT-4o-SP-DQ	619	61 ± 17	<b>2.07</b>	12.38	18.33	1.02	0.62	0.62
REGen-DQ	95	37 ± 26	1.45	19.13	10.35	1.05	0.48	0.57
REGen-IDQ-T	77	35 ± 31	1.89	19.79	10.02	1.03	<b>0.41</b>	<b>0.57</b>
REGen-IDQ-TV	81	35 ± 31	1.90	19.86	<b>9.70</b>	1.02	0.39	0.57
Ground truth	76	54 ± 37	69.00*	27.60	> 7.86	<0.98	0.43	0.57

Scene change rate

Text-visual correspondence

**Check out our paper for more results!**

# Subjective Evaluation



Model	Coherence $\uparrow$	Alignment $\uparrow$	Realness $\uparrow$	Interview effectiveness $\uparrow$
A2Summ [4]	$2.72 \pm 0.24$	$2.87 \pm 0.26$	$2.67 \pm 0.23$	$3.07 \pm 0.24$
TeaserGen [11]	$3.22 \pm 0.23$	$2.92 \pm 0.24$	$2.86 \pm 0.23$	-
GPT-4o-SP-DQ	$3.08 \pm 0.24$	$3.23 \pm 0.25$	$2.81 \pm 0.25$	$3.32 \pm 0.25$
REGen-DQ	$2.97 \pm 0.27$	$3.03 \pm 0.27$	$2.75 \pm 0.30$	<b><math>3.33 \pm 0.29</math></b>
REGen-IDQ-TV	<b><math>3.29 \pm 0.24</math></b>	<b><math>3.30 \pm 0.26</math></b>	<b><math>3.05 \pm 0.25</math></b>	$3.25 \pm 0.30$

**REGen-IDQ-TV (indirect quote-based) outperforms REGen-DQ in most criteria**

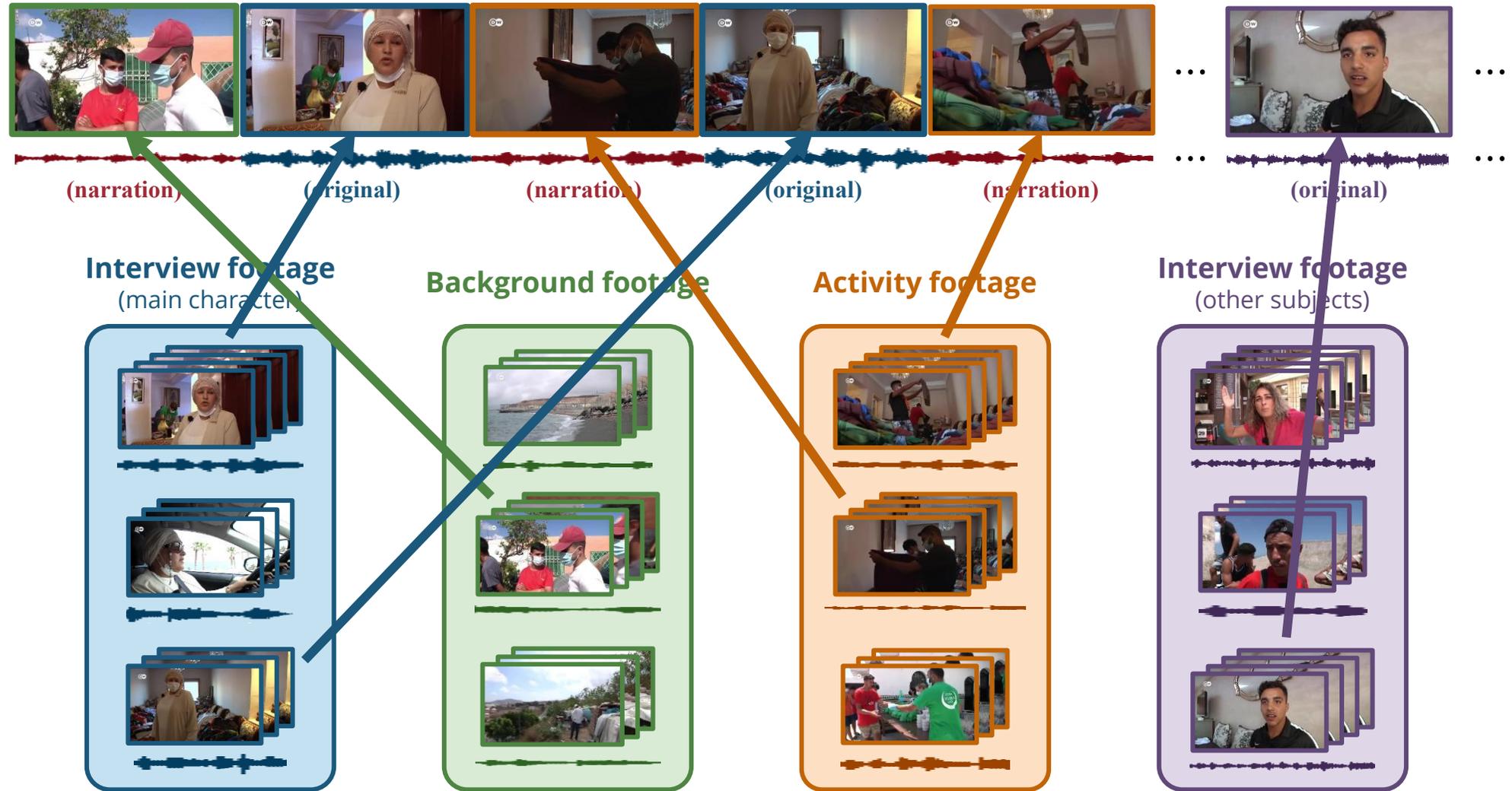
# Limitations



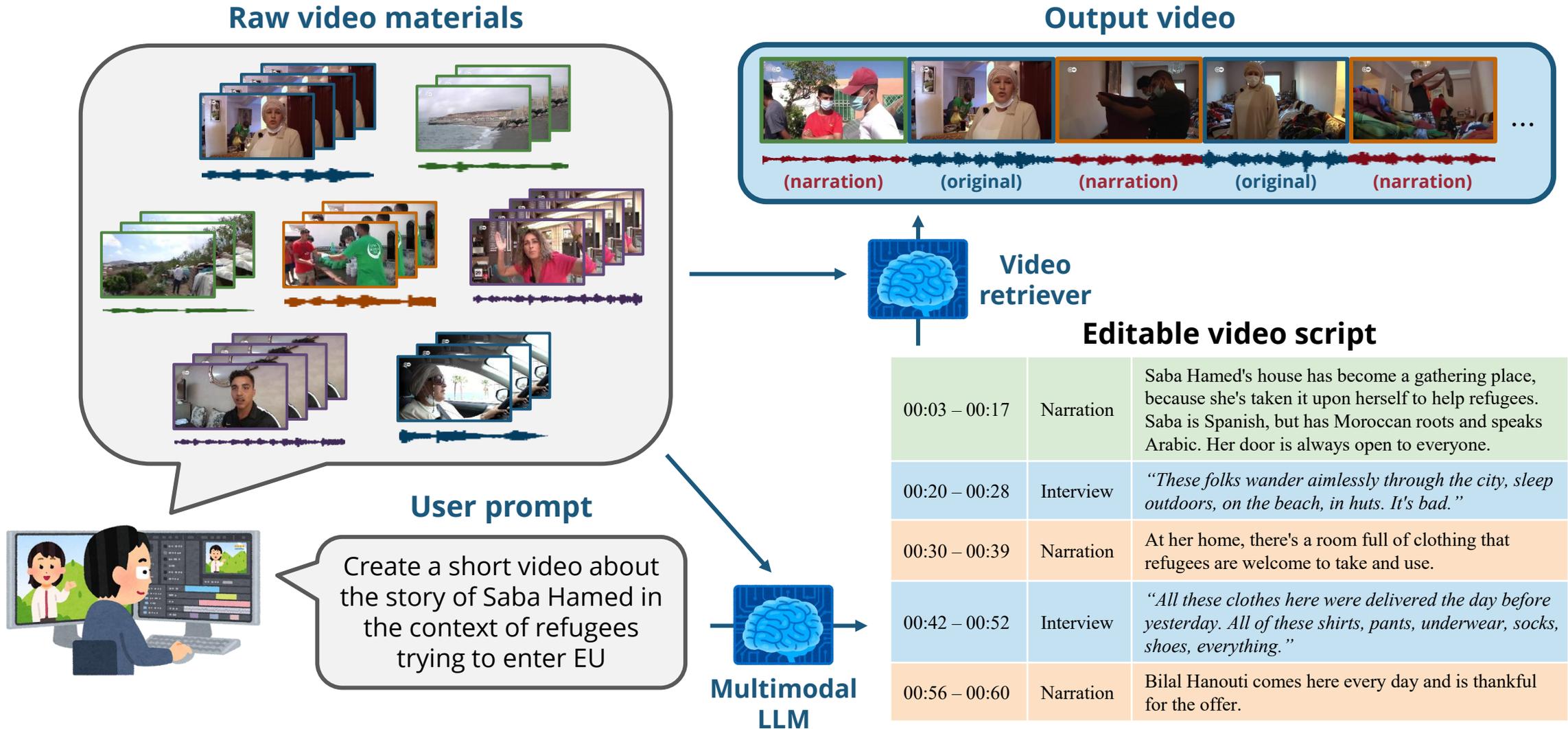
- Assumed that **narration plays a more significant role** than visuals
  - This assumption might not hold for movies and vlogs
- Risks of **misplacing a quote in a wrong context**
  - Grounding the script generation model with information about all quotable materials
  - May also be alleviated by context-aware video embeddings
- Reliance on successful **scene segmentation** of the input video
  - Speaker diarization might not do the trick for lecture recordings

# Towards AI-assisted Video Editing

# Video Editing



# Ongoing Work: Multimodal RAG-based Video Editing



⋮

# Creating a Dataset for Video Editing



Raw video materials

# Ongoing Work: Multimodal RAG-based Video Editing



Input video ("Mammal Origins | Full Documentary | NOVA | PBS")



[youtu.be/23BGbVBxXdQ](https://youtu.be/23BGbVBxXdQ)

Example output 1 (Query: "mammal")

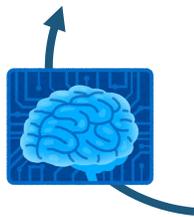
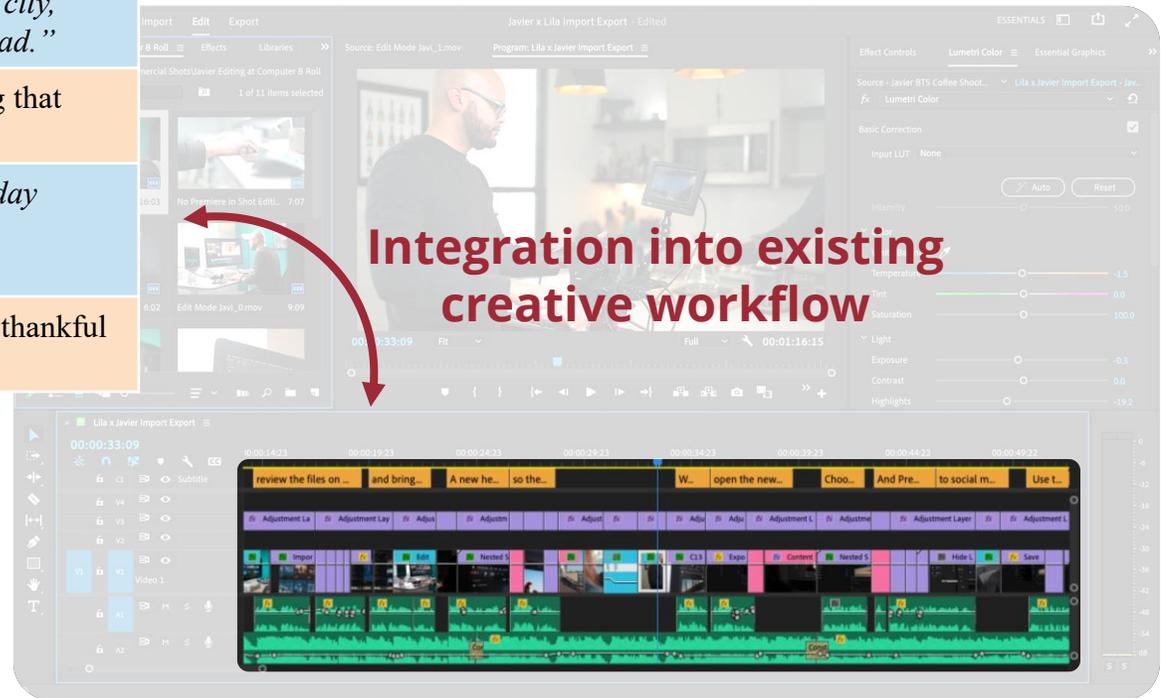


Example output 2 (Query: "therapsids")



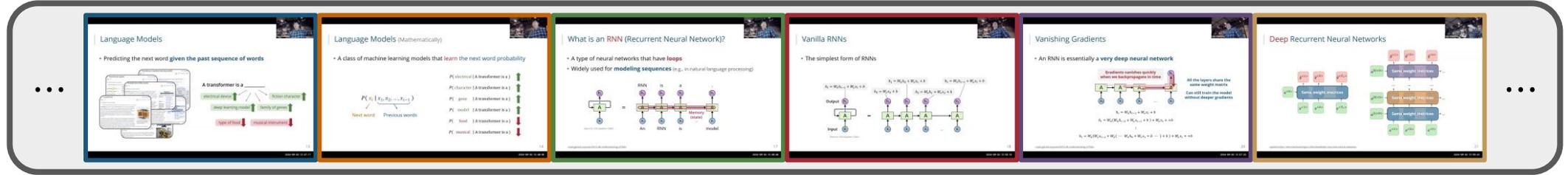
# Future Work: Integration into Video Editing Software

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.



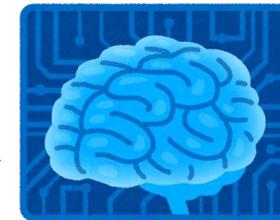
# Future Work: LectureRecap

## Lecture recording

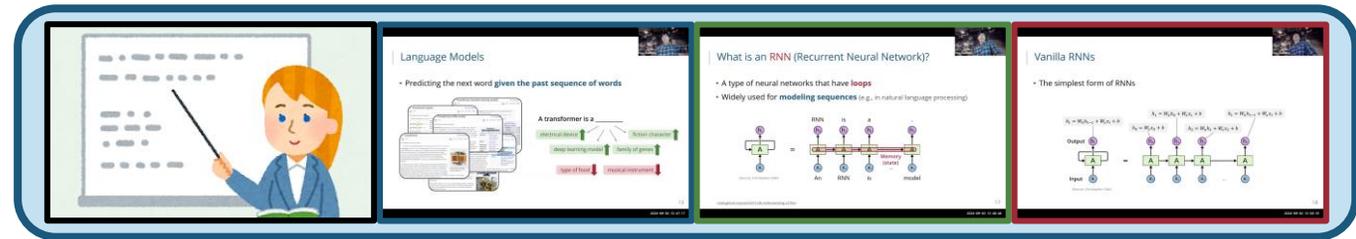


I would like to **review** the concept of **recurrent neural networks**. How does an RNN work?

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



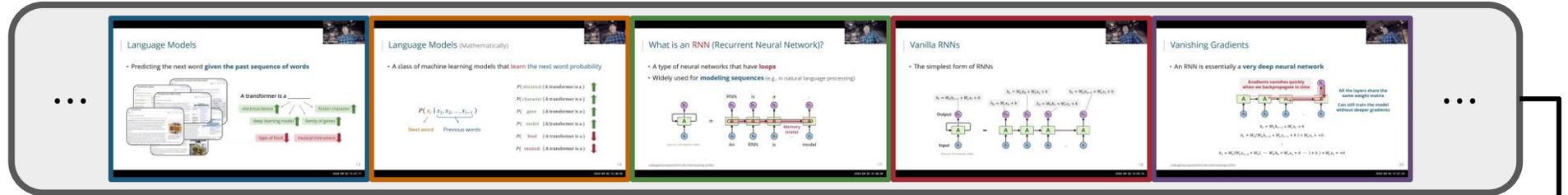
LectureRecap



Lecture recap

# Future Work: LectureRecap

Lecture recording



User query

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

Script generation

Speech recognition

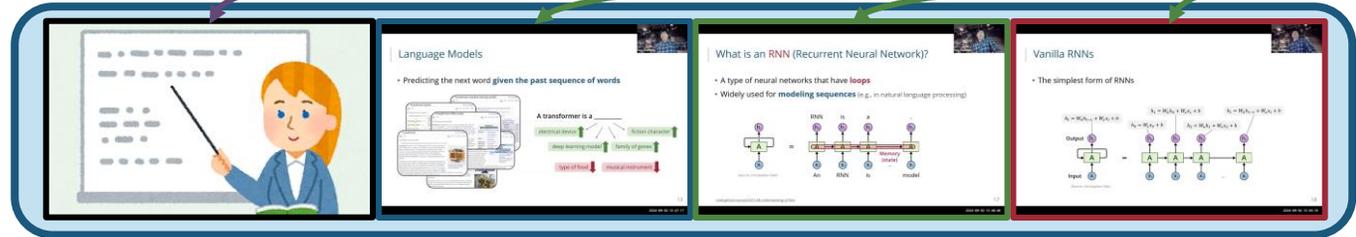
Video script

[**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



# Retrieval-Augmented → Retrieval-Embedded Generation

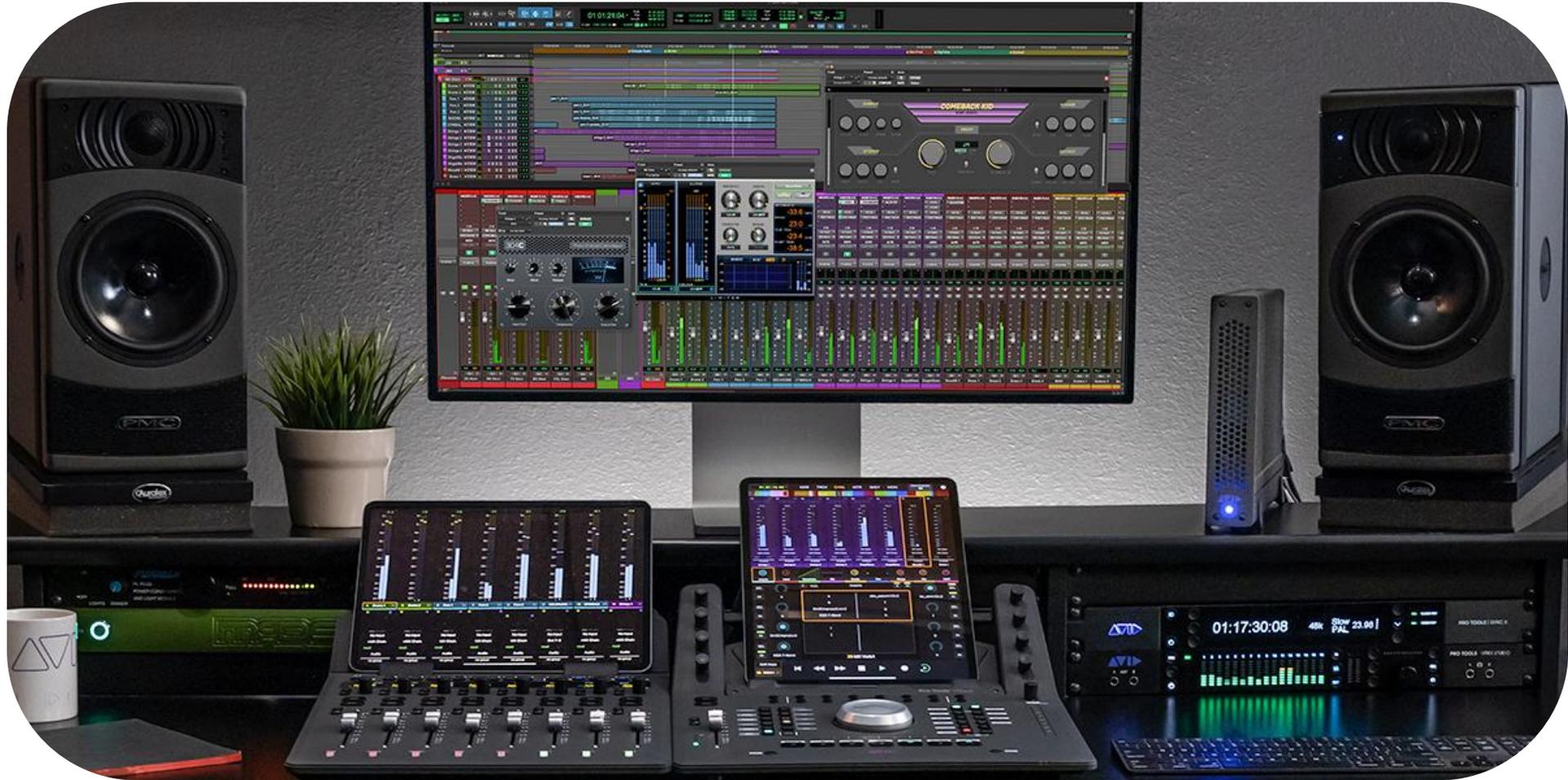
- Can an LLM **learn to quote** and **embed the quote properly**?
- How to **quote materials in other modalities**?
  - Audio, image, videos, sensor data, etc.
- What do we need?
  - A **retriever that can identify candidate quotable materials**
  - A **multimodal LLM that can understand multimodal data**

# Future Work: Integrating GenAI into Music Production



(Source: Avid)

# Future Work: Integrating GenAI into Music Production



(Source: Avid)



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

Haven Kim<sup>1</sup> Zachary Novack<sup>1</sup> Weihan Xu<sup>2</sup> Julian McAuley<sup>1</sup> **Hao-Wen Dong<sup>3</sup>**

<sup>1</sup> UC San Diego   <sup>2</sup> Duke University   <sup>3</sup> University of Michigan



UC San Diego

Duke  
UNIVERSITY

**M** UNIVERSITY OF MICHIGAN

# Accessibility of Music-Video Datasets

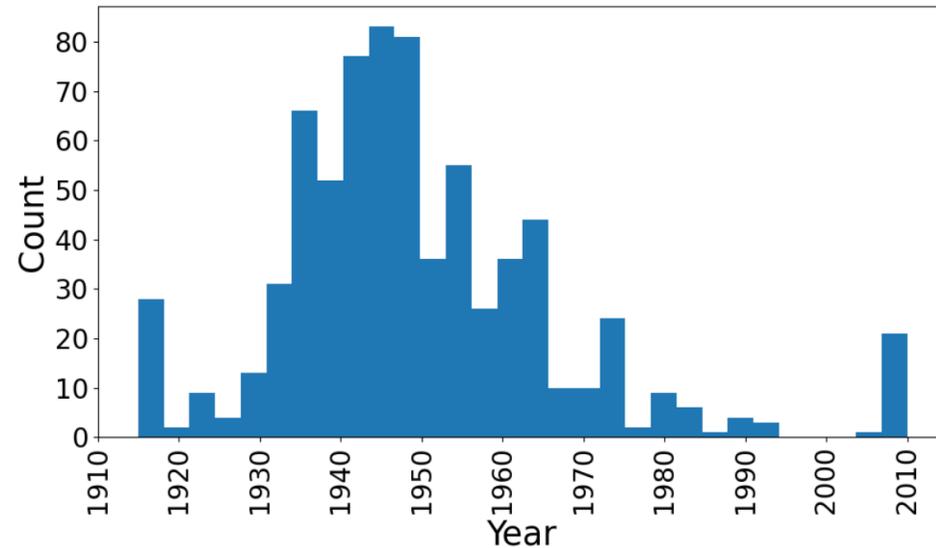


Dataset	Audio	MIDI	Self-Hosted	Mood	Video Content	Length (Hours)
HIMV-200K [48]	✓	✗	✗	✗	Music Video, User-Generated Video	-
URMP [49]	✓	✓	✗	✗	Music Performance	33.5
TikTok [50]	✓	✗	✗	✗	Dance Video	1.5
AIST++ [51]	✓	✗	✓	✗	3D Dance Motion	5.2
SymMV [52]	✓	✓	✗	✗	Music Video	76.5
MuVi-Sync [53]	✓	✓	✗	✗	Music Video	-
BGM909 [54]	✓	✓	✗	✗	Music Video	-
NES-VMDB [55]	✗	✓	✗	✗	Gameplay Video	474.0
OSSL (Ours)	✓	✗	✓	✓	Films	36.5

# 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)

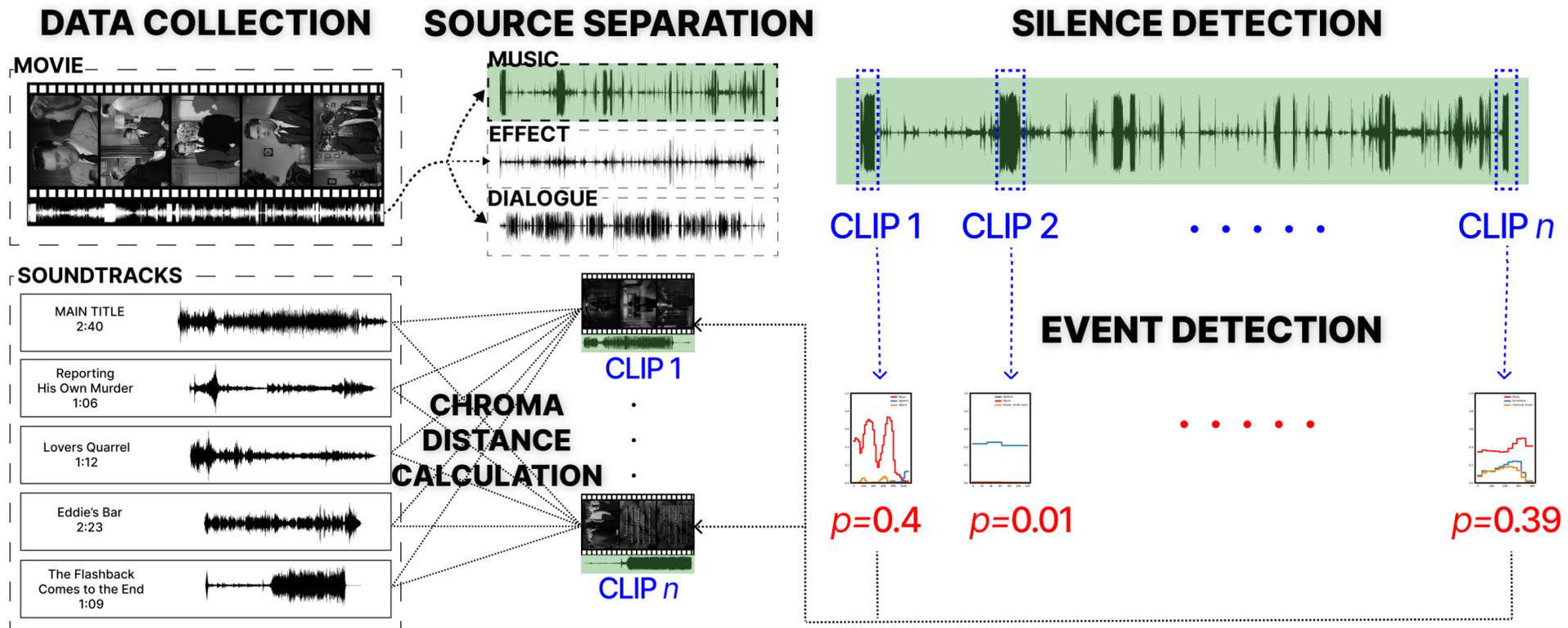


# Open Screen Soundtrack Library (OSSL)

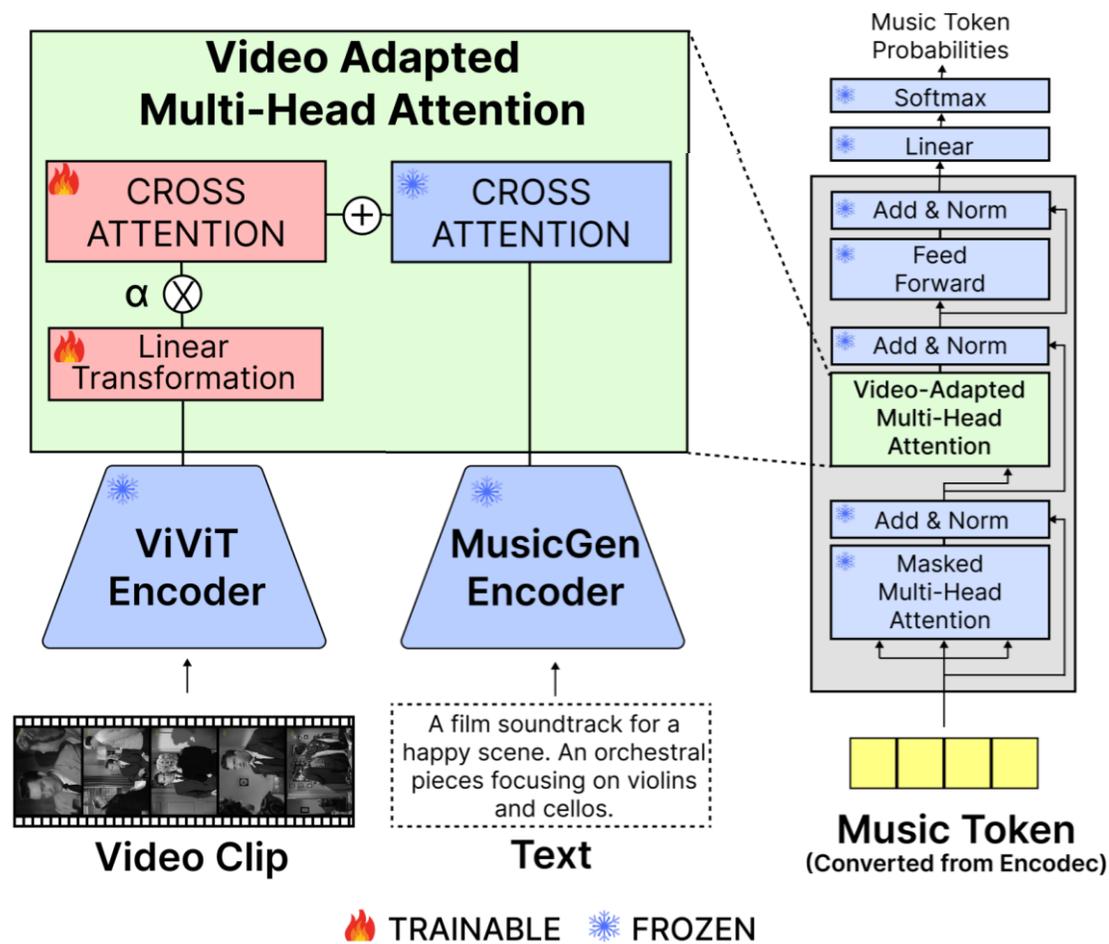


[havenpersona.github.io/oss1-v1](https://havenpersona.github.io/oss1-v1)

# 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)

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

(ISMIR 2025)

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

# Objective Evaluation



OSSL Fine- tuned	Video Adapter Inte- grated	Compared against commercial soundtracks				Compared against the ground truth						
		Objective				Diversity		Recall		Precision		
		Distributional Fidelity		Paired Fidelity		Similarity		KL		Recall		
		FAD ↓	Precision ↑	Similarity ↑	KL ↓	Recall ↑	Precision ↑	Similarity ↑	KL ↓	Recall ↑	Precision ↑	
		pub	com	pub	com	pub	com	pub	com	pub	com	
M-BASE	✗	✗	60.91	76.79	21.00	12.00	43.61	34.39	1.06	1.88	8.04	8.46
M-TEXT	✓	✗	61.15	77.79	24.00	17.00	<b>45.31</b>	33.72	1.04	1.90	7.28	13.78
M-MULTI	✓	✓	<b>59.51</b>	<b>73.26</b>	<b>25.00</b>	<b>21.00</b>	<b>45.31</b>	<b>36.25</b>	<b>1.00</b>	<b>1.81</b>	<b>9.96</b>	8.72

The proposed video adapter improves both **distributional** and **paired fidelity**.

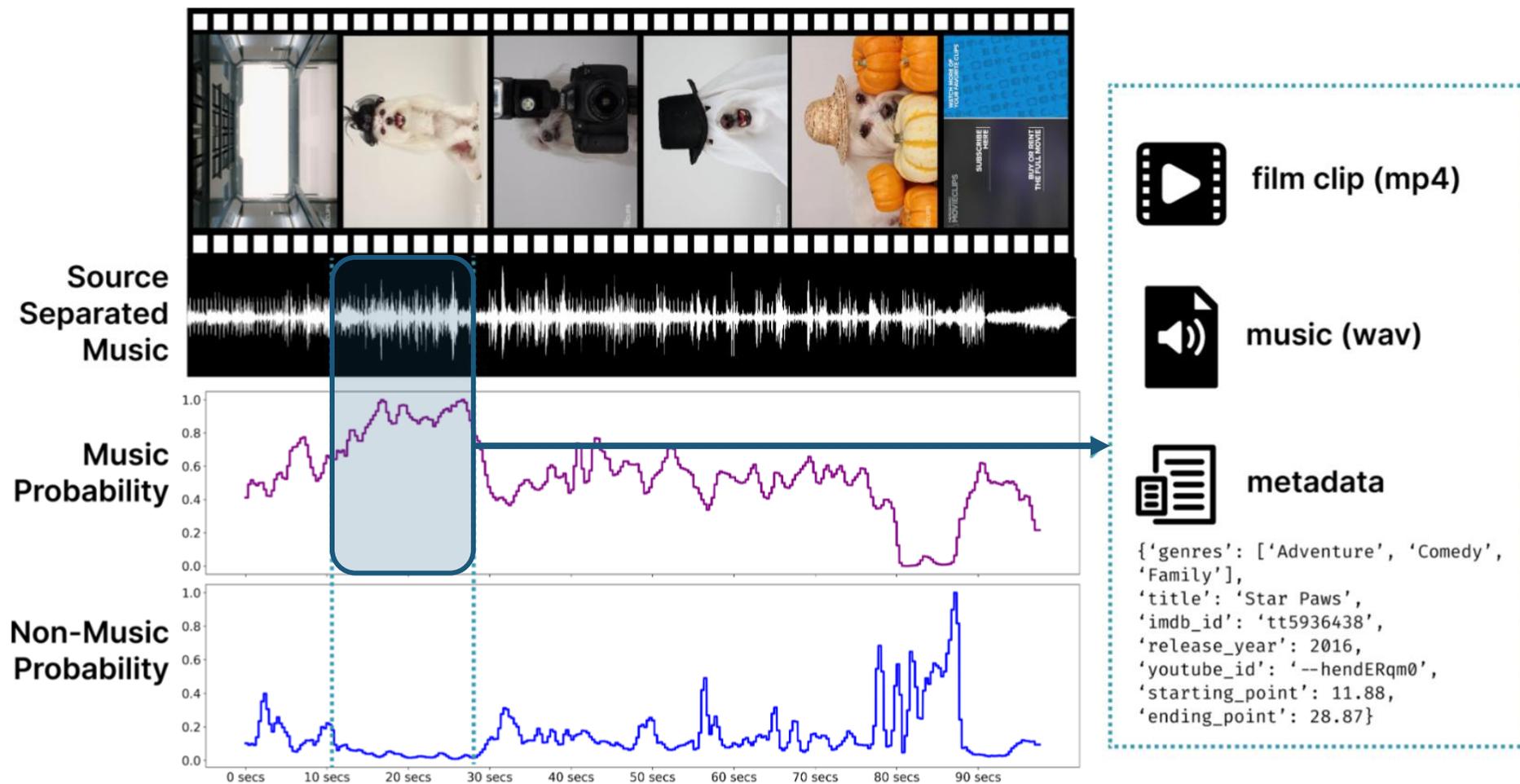
# Subjective Evaluation



	OSSL Fine- tuned	Video Adapter Inte- grated	Subjective		
			Human Ratings		
			Mood ↑	Genre ↑	Quality ↑
			avg ± CI	avg ± CI	avg ± CI
M-BASE	✗	✗	5.13 ± 0.94	5.60 ± 1.12	6.20 ± 0.84
M-TEXT	✓	✗	5.20 ± 0.99	6.03 ± 0.98	6.00 ± 1.02
M-MULTI	✓	✓	6.20 ± 1.05	6.70 ± 1.06	7.07 ± 0.93

The proposed video adapter significantly improves the **generated music quality**.

# Extending OSSL to OSSL v2



# Extending OSSL to OSSL v2

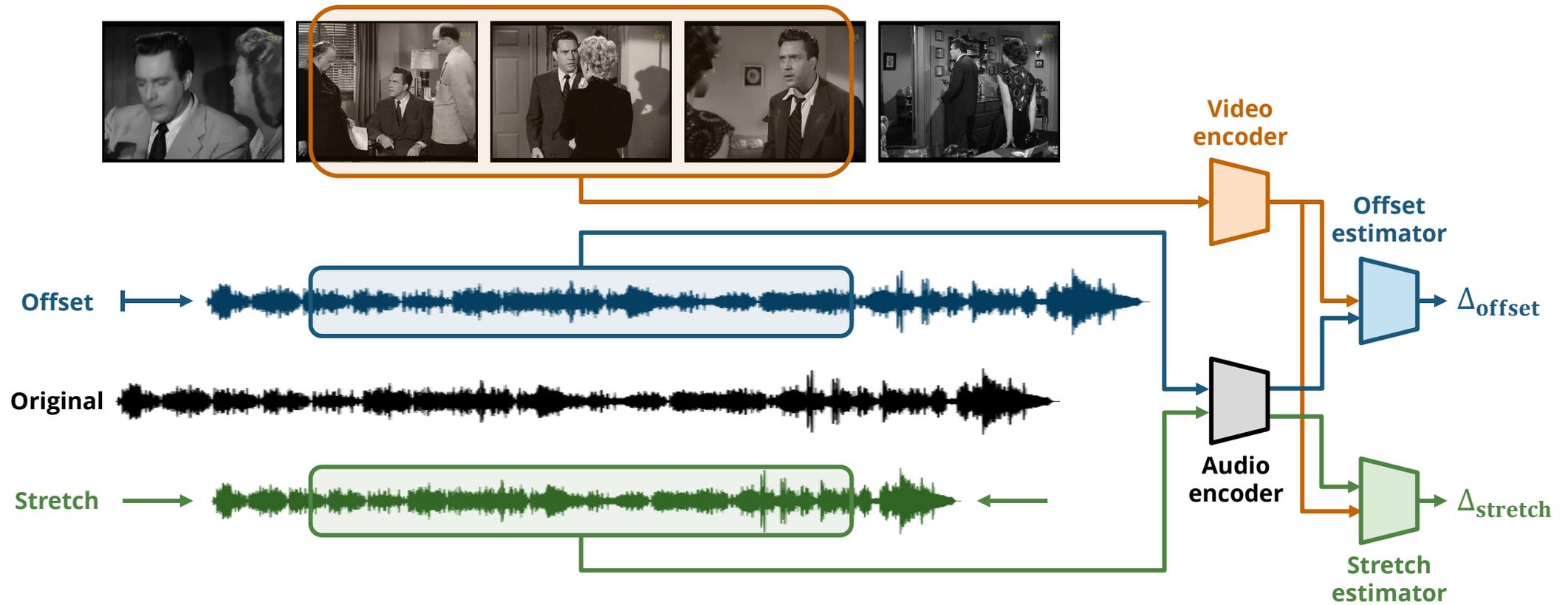


	Directly downloadable	YouTube- based	
	Public Domain	Commercial [2]	Total
Number of Clips	35,705	40,703	76,408
Number of Unique Films	1,886	2,633	4,519
Average Length (seconds)	28.77	23.65	26.04
Total Length(hours)	285.31	267.39	552.70

**7x larger than OSSL**

Enables training **foundational models for video-to-music generation from scratch!**

# Future Work: Measuring Video-Music Alignment

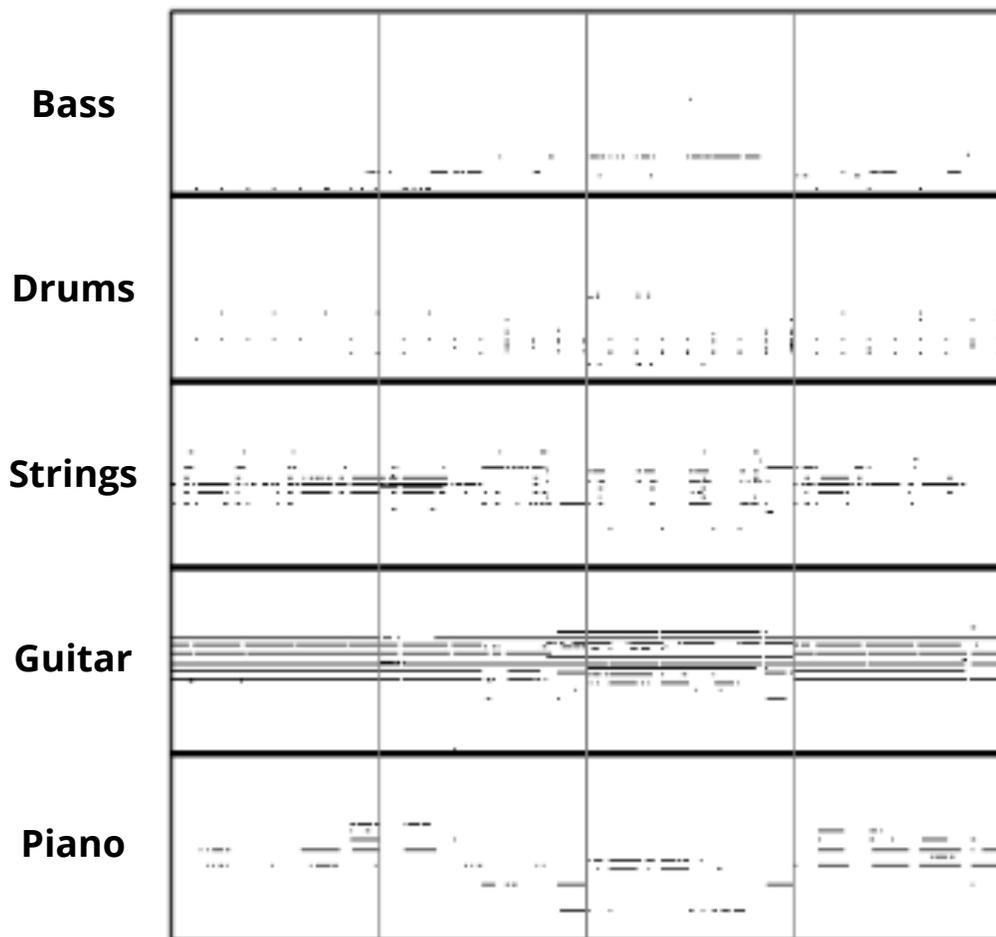


# Augmenting Human Creativity with AI

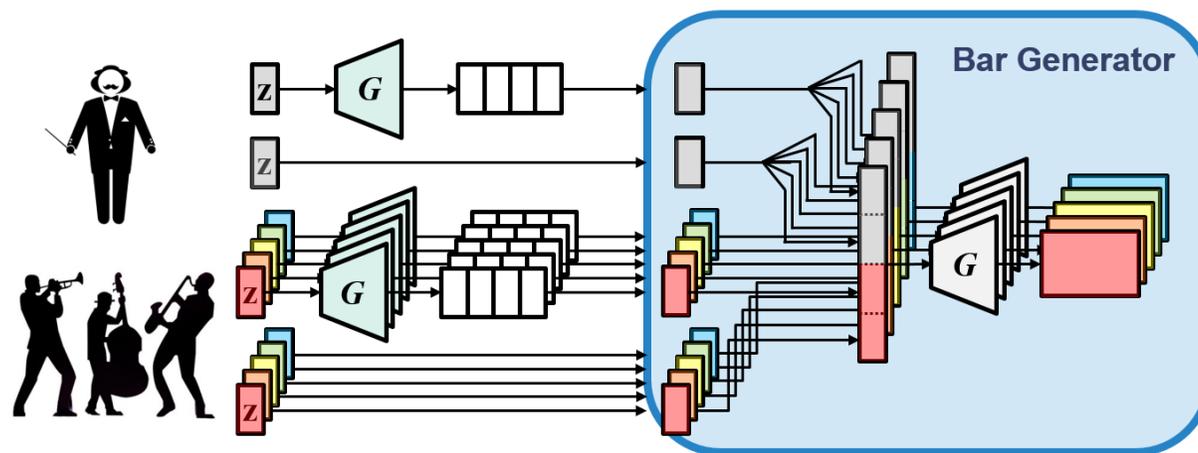
- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024), **text-to-music generation** (ISMIR 2025), **video-to-music generation** (ISMIR 2025), **symbolic music processing tools** (ISMIR LBD 2019, ISMIR 2020)
- **AI-assisted Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **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



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)

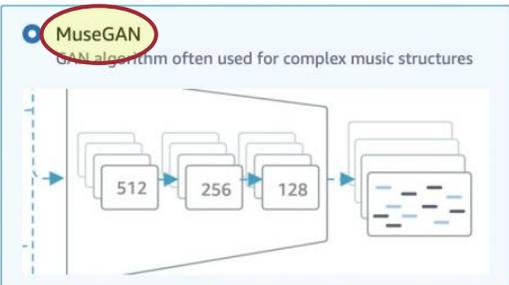
AWS DeepComposer > Models > Train a model

## Train a model

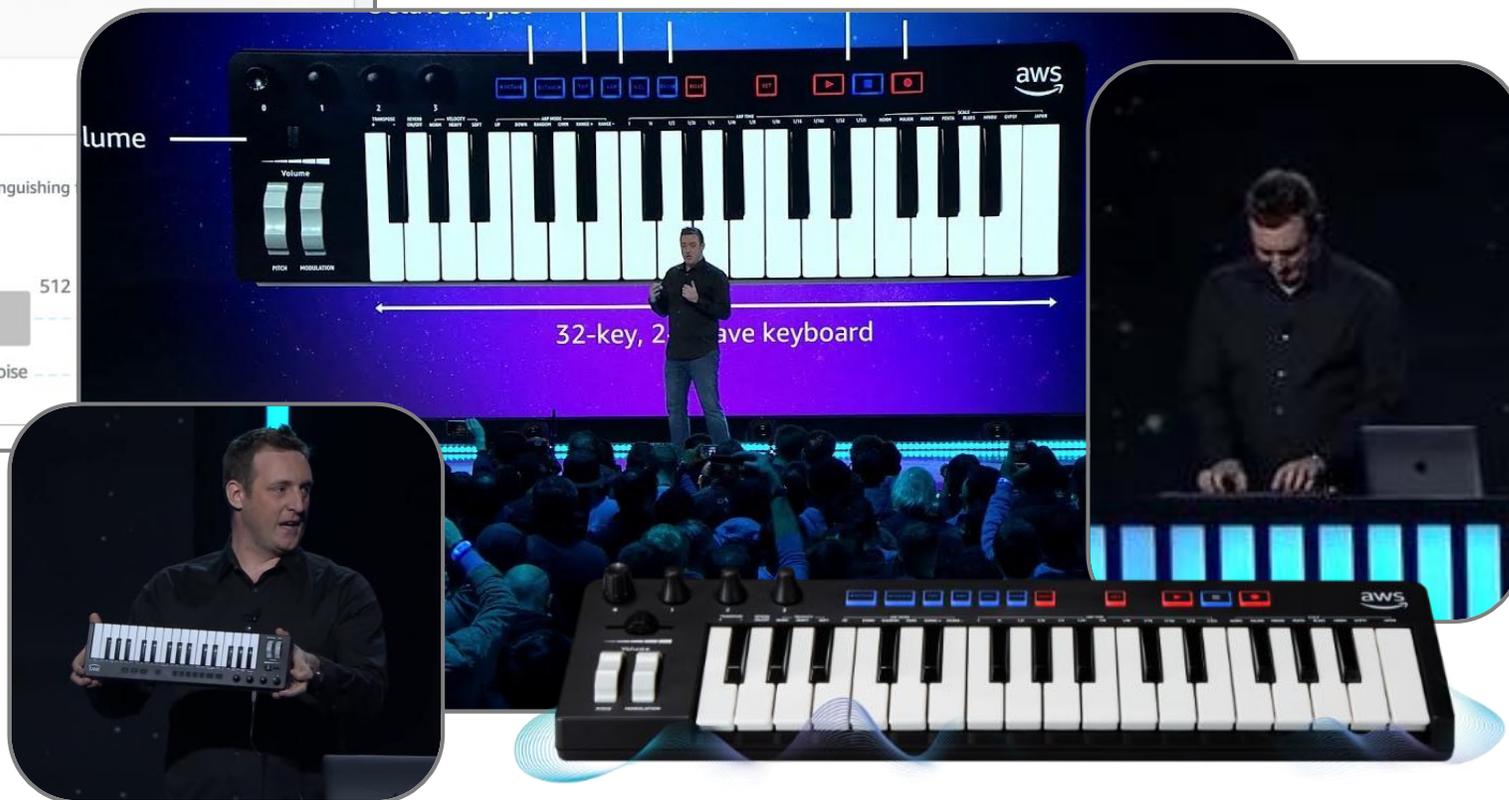
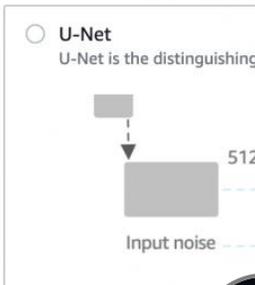
**Generative algorithm** [Info](#)

Choose a generative algorithm to train a model

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



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



[amazon.com/dp/B07YGZ4V5B/](https://amazon.com/dp/B07YGZ4V5B/)

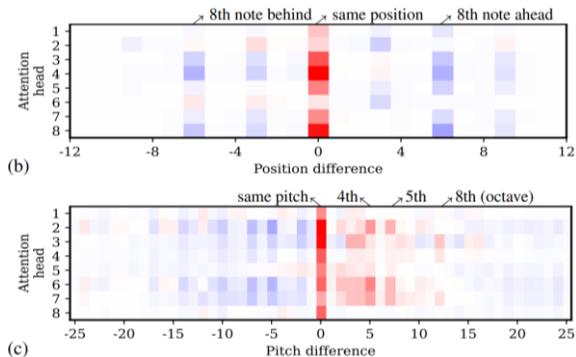
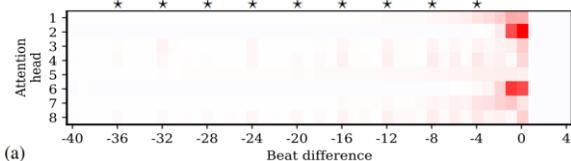
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," *AWS News Blog*, April 2, 2020.

# Generating Multitrack Music with Transformers (ICASSP 2023)

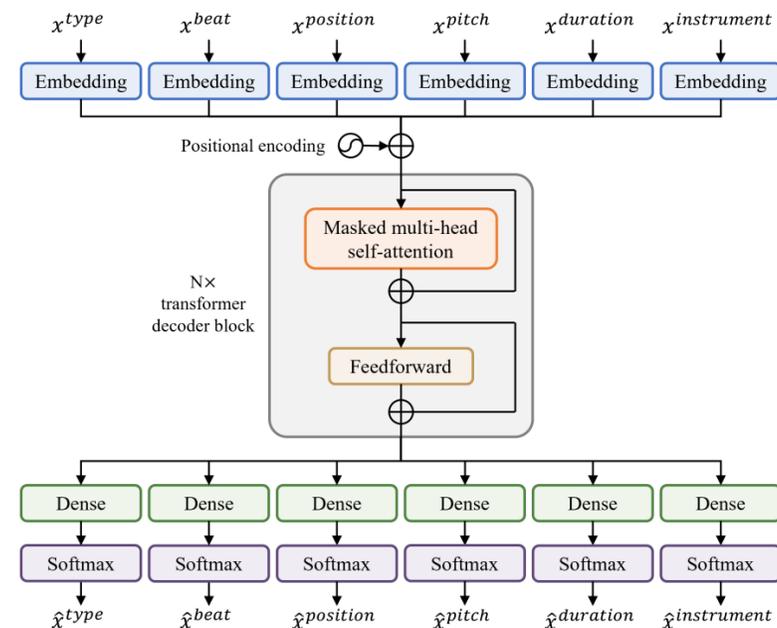
## Multitrack Music Representation

(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	

## Musical Self-attention



## Multitrack Music Transformer



UC San Diego

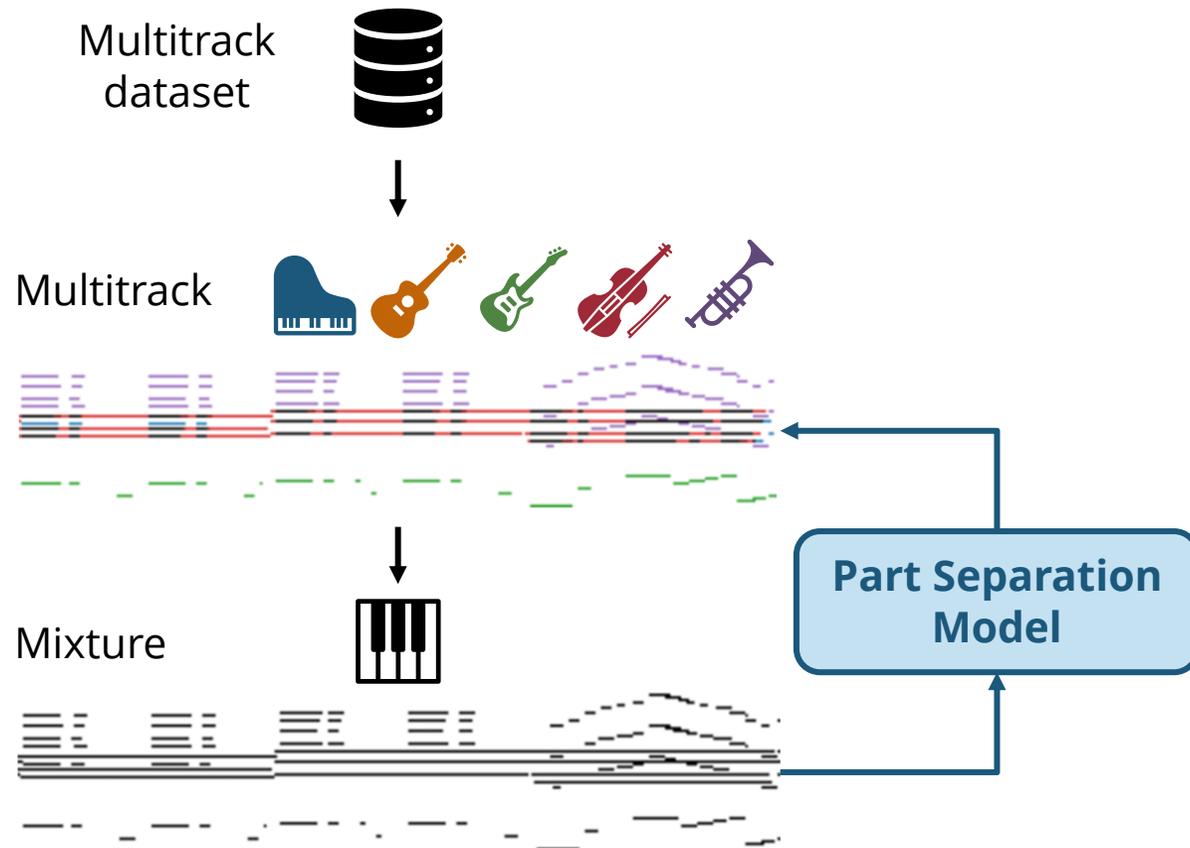
# Automatic Instrumentation (ISMIR 2021)



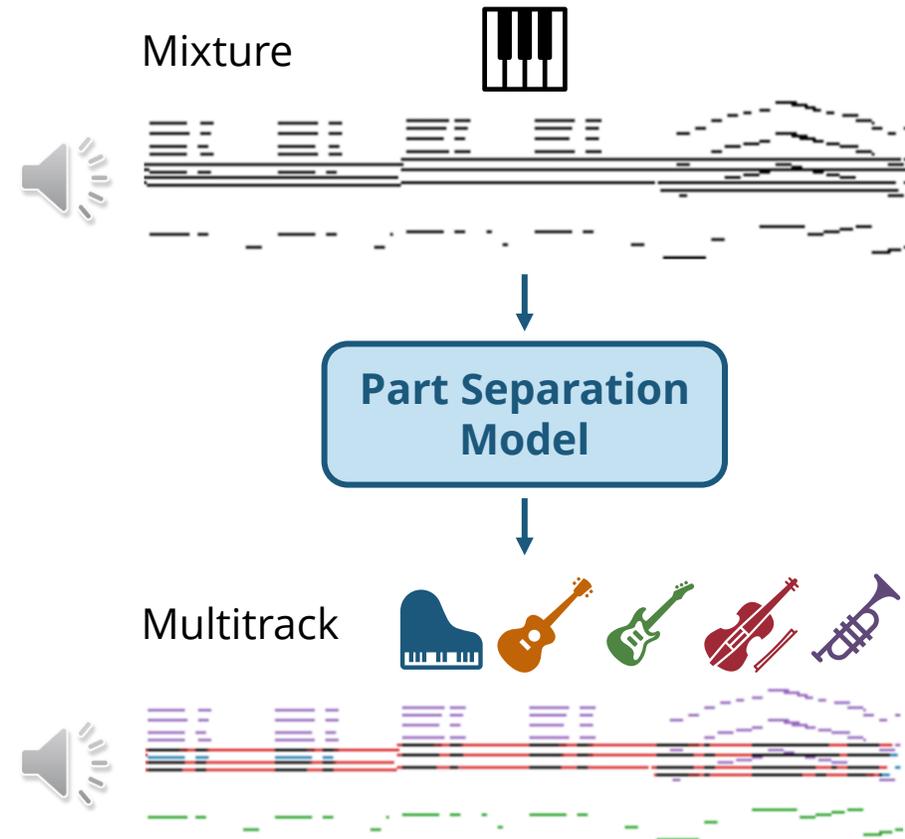
Stanford

UC San Diego

## Training

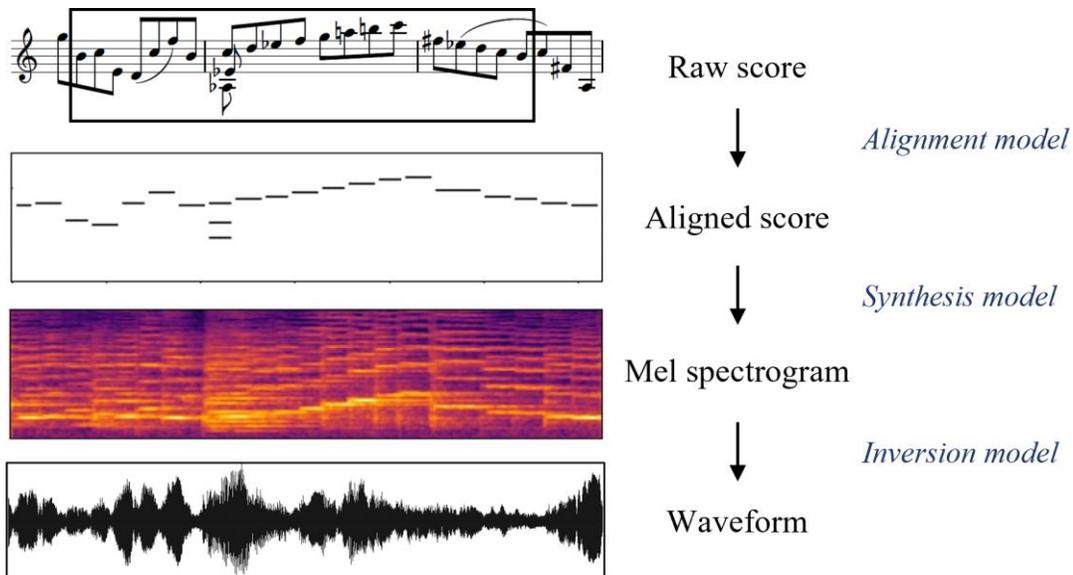


## Inference

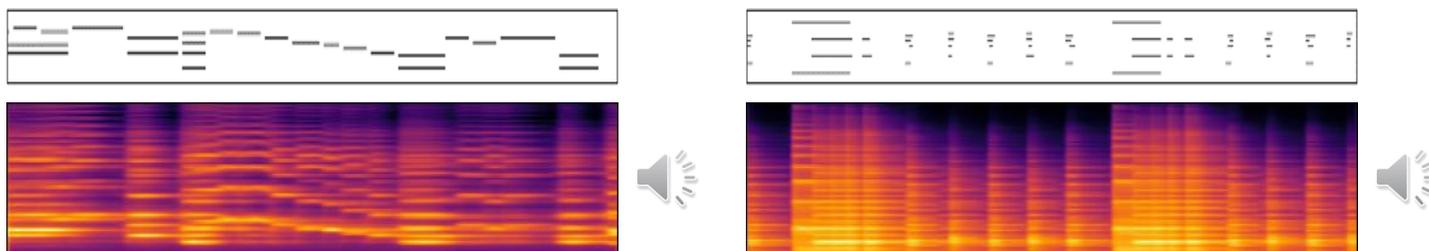


# Synthesizing Expressive Violin Performance (ICASSP 2022)

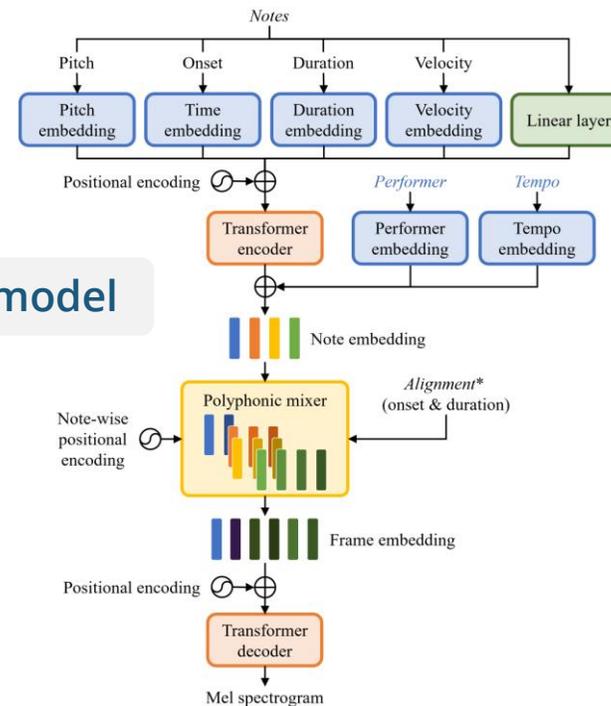
## Performance synthesis



## Example results

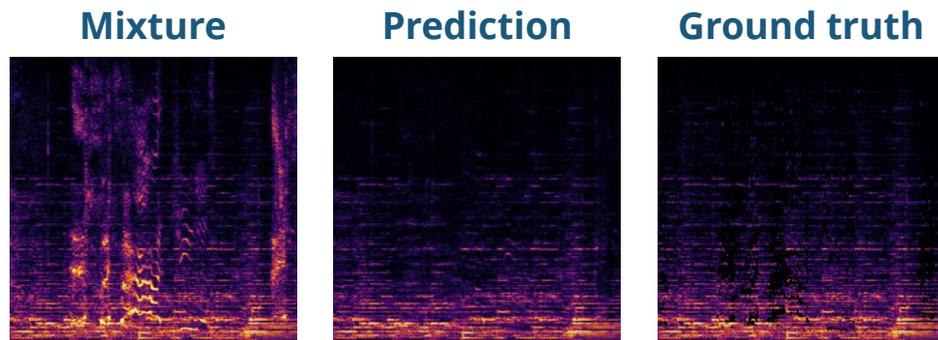


## TTS-based model



# Text-queried Sound Separation (ICLR 2023)

Query: "playing harpsichord"

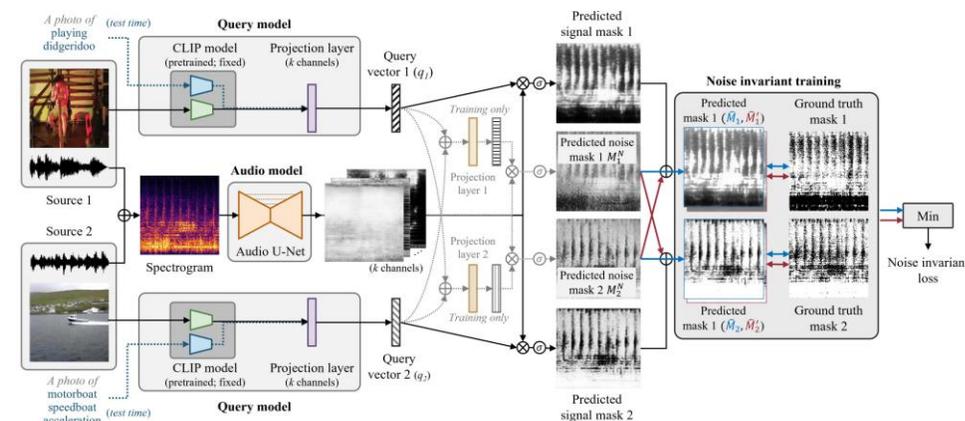
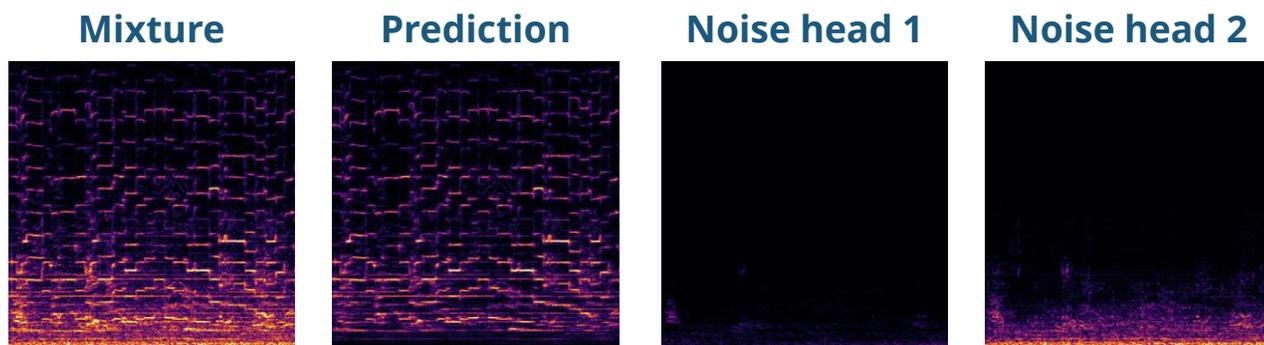


SONY

UC San Diego

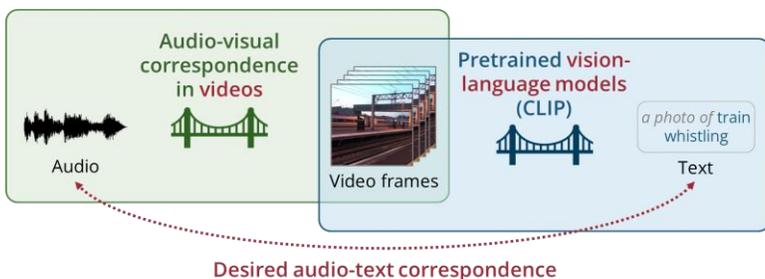
## Text-queried sound separation model

Query: "playing bagpipe"

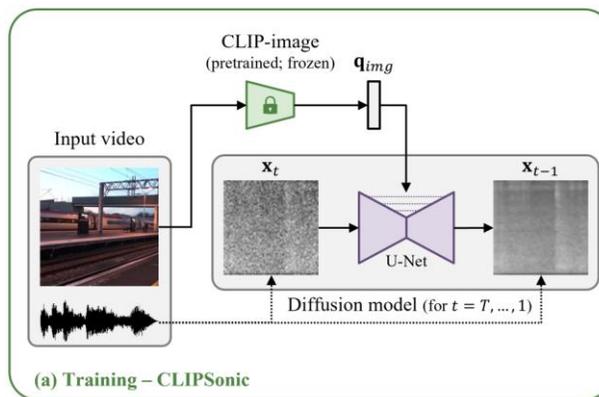


# Text-to-Audio Synthesis (WASPAA 2023)

## Learning Sounds from Noisy Videos



## Training

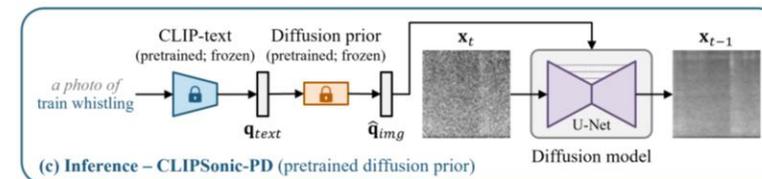


## Image-to-sound results

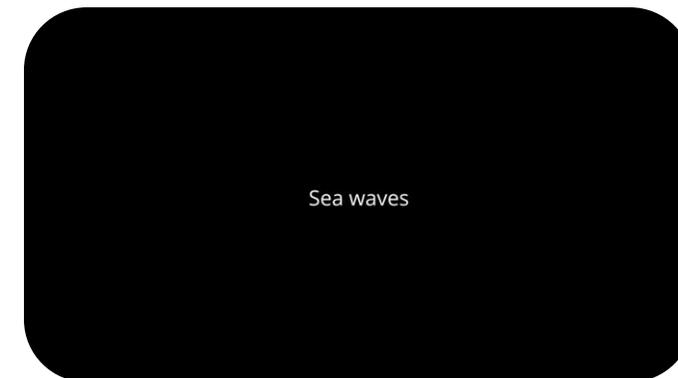


Dolby UC San Diego

## Inference



## Text-to-sound results



**Art challenges Technology**



**Creativity**

**Augmenting Human Creativity  
with AI**



**AI**



**Technology inspires the Art**

# Augmenting Human Creativity with AI

- **Novel Generative Models for New Domains**
  - **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 Tools for Content Creation**
  - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020)
- **Multimodal Generative Models for Content Creation**
  - **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



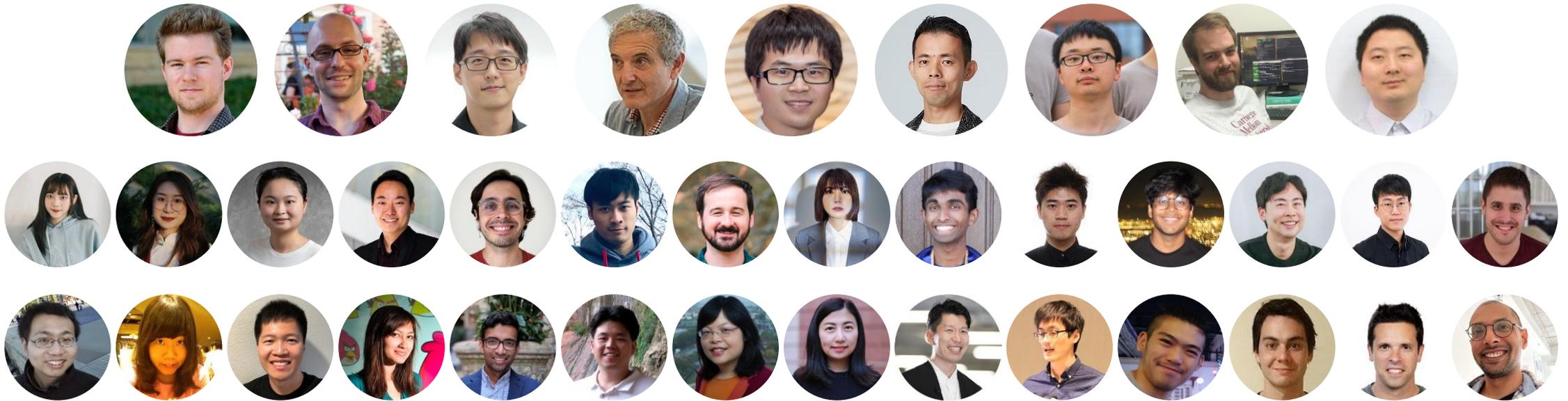
Universitaetsmedizin, [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/), via Wikimedia Commons  
[uploadvr.com/iron-man-vr-breaks-free-from-cords-load-screens-on-quest-2/](https://uploadvr.com/iron-man-vr-breaks-free-from-cords-load-screens-on-quest-2/)  
[descript.com/blog/article/what-is-the-best-audio-interface-for-recording-a-podcast](https://descript.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**

# Augmenting Human Creativity with Generative AI

Nothing would have been possible without all my fantastic collaborators!



UC San Diego

中央研究院  
ACADEMIA SINICA

Dolby

SONY

amazon

nvidia



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

**M** UNIVERSITY OF MICHIGAN