Towards AI-assisted Video Editing: Generating Shorts from Long Videos

Hao-Wen (Herman) Dong

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

February 5, 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.

Art challenges Technology

Creativity

Augmenting Human Creativity with AI

AI

Technology inspires the Art

Generative AI for Music, Audio & Video Creation

(Source: The Denver Post)

(Source: UploadVR)

(Source: Descript)

(Source: Daily Bruin)

(Source: Wikimedia Commons)

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

Augmenting Human Creativity with AI

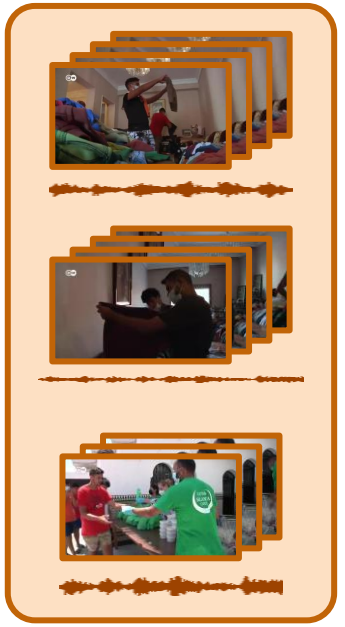
- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024), **text-to-music generation** (ISMIR 2025), **video-to-music generation** (ISMIR 2025), **symbolic music processing tools** (ISMIR LBD 2019, ISMIR 2020)
- **AI-assisted 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)

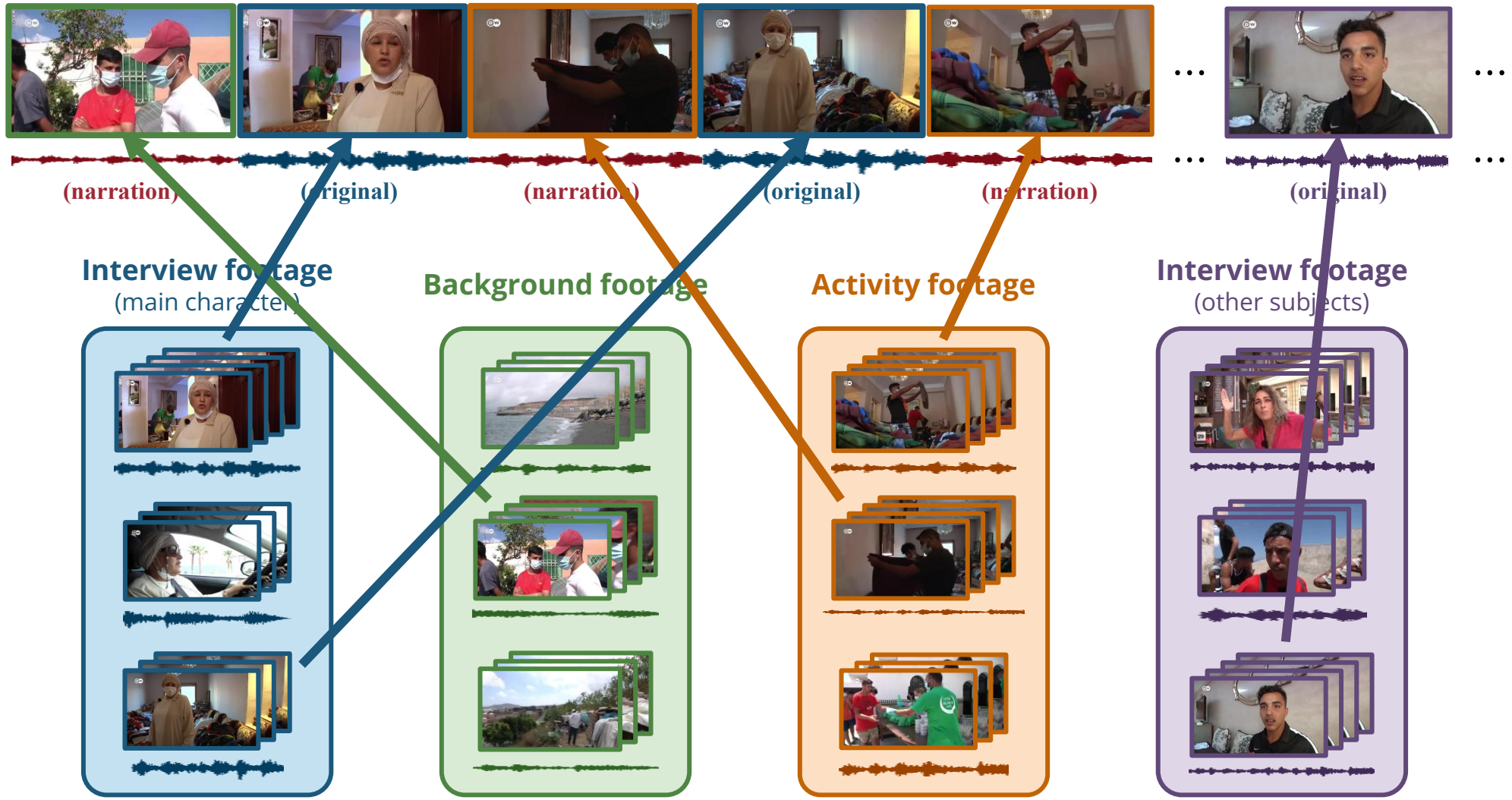
Background footage

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

ICLR

TeaserGen: Generating Teasers for Long Documentaries

Weihan Xu[1] Paul Pu Liang[2] Haven Kim[3]
Julian McAuley[3] Taylor Berg-Kirkpatrick[3] **Hao-Wen Dong**[4]

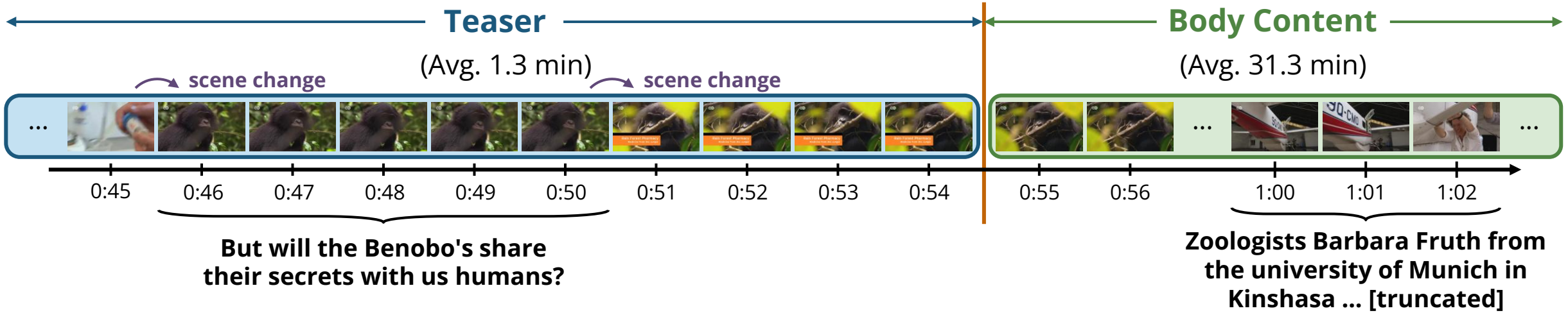
[1] Duke University [2] MIT [3] UC San Diego [4] University of Michigan





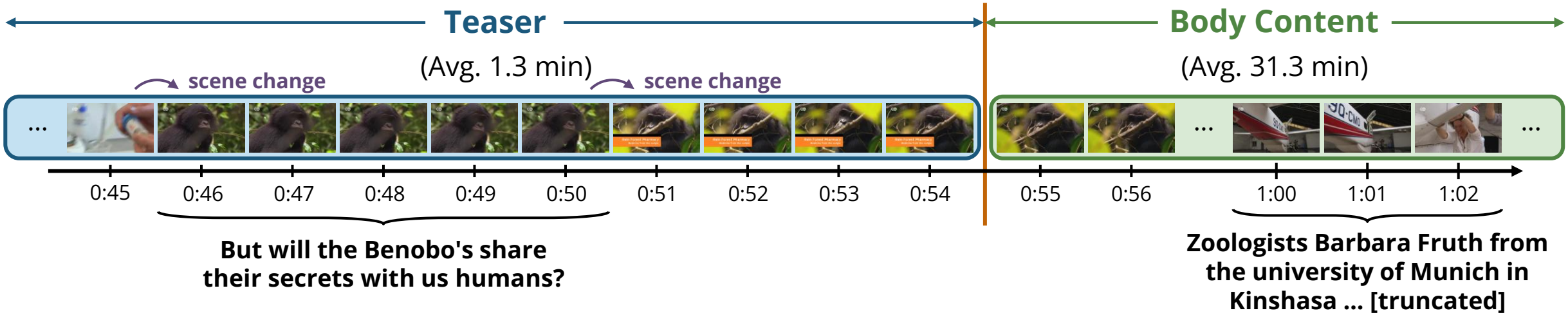

Duke UNIVERSITY Massachusetts Institute of Technology UC San Diego UNIVERSITY OF MICHIGAN

DocumentaryNet: A New Documentary Dataset



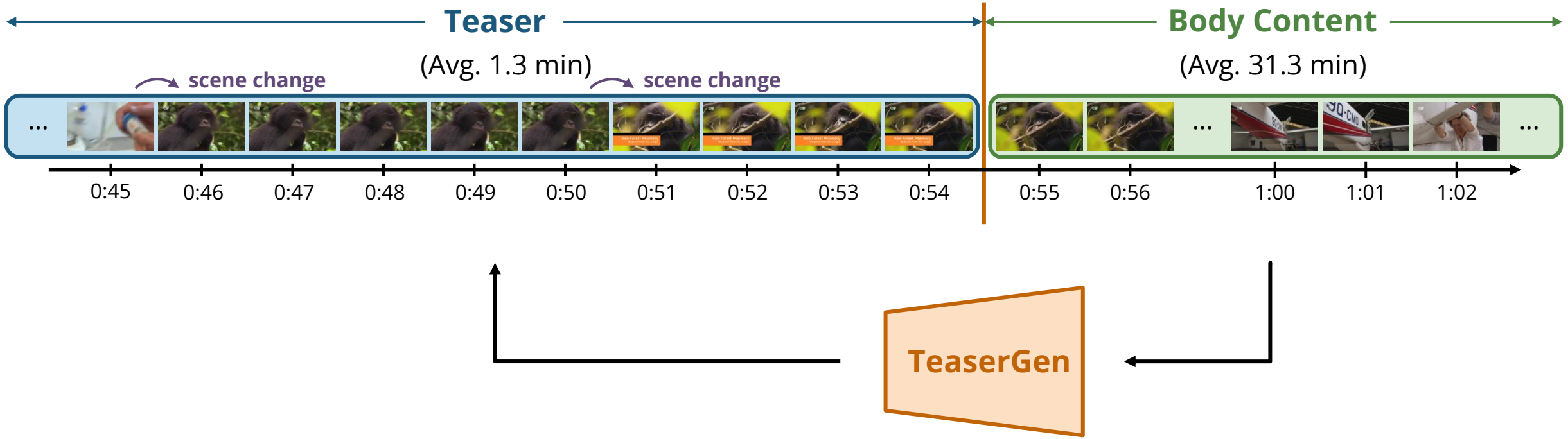
DocumentaryNet: A New Documentary Dataset


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



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

Generating Teasers from Long Documentaries



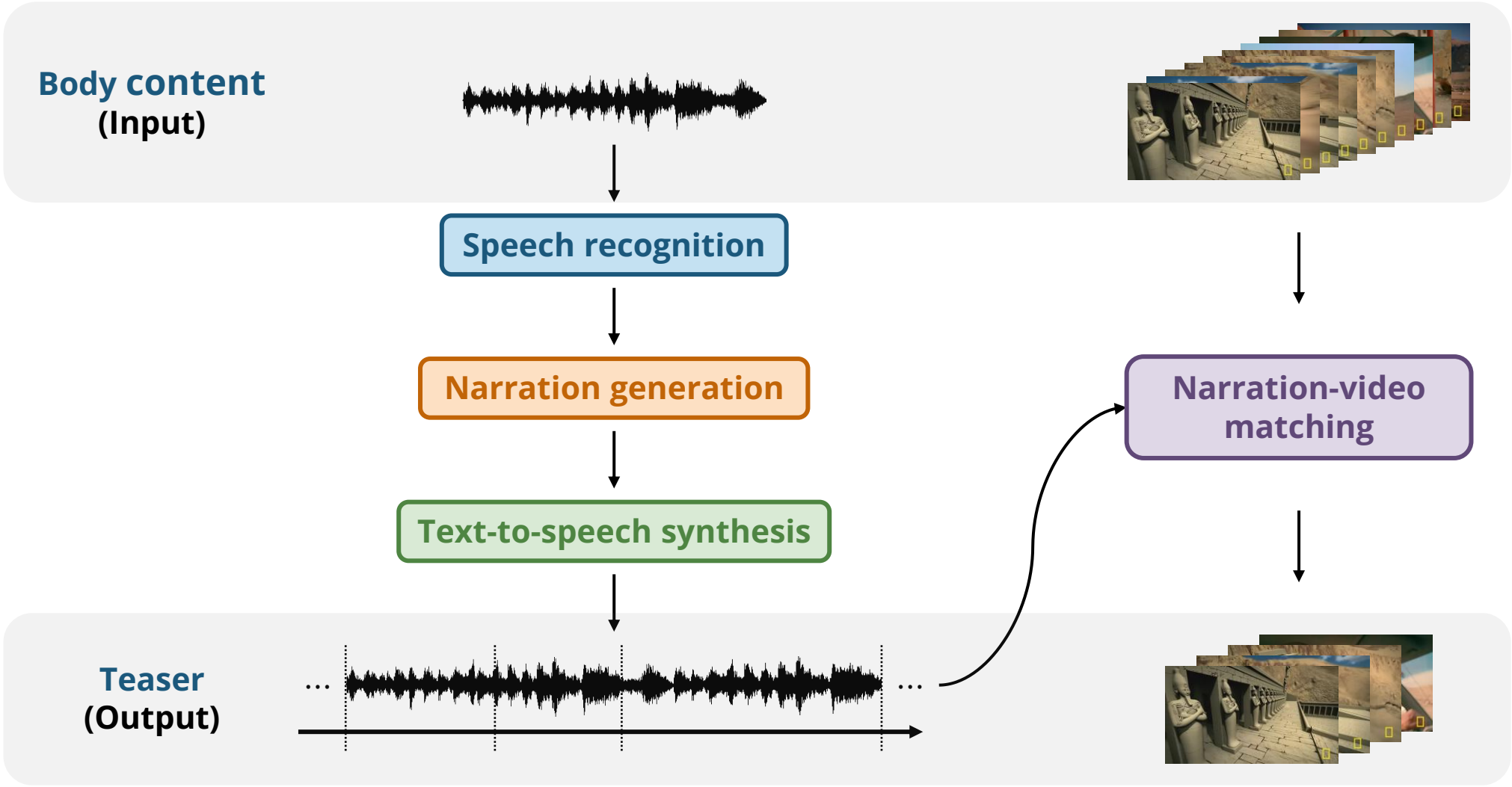
~24:1 compression ratio

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

Documentary Teaser Generation

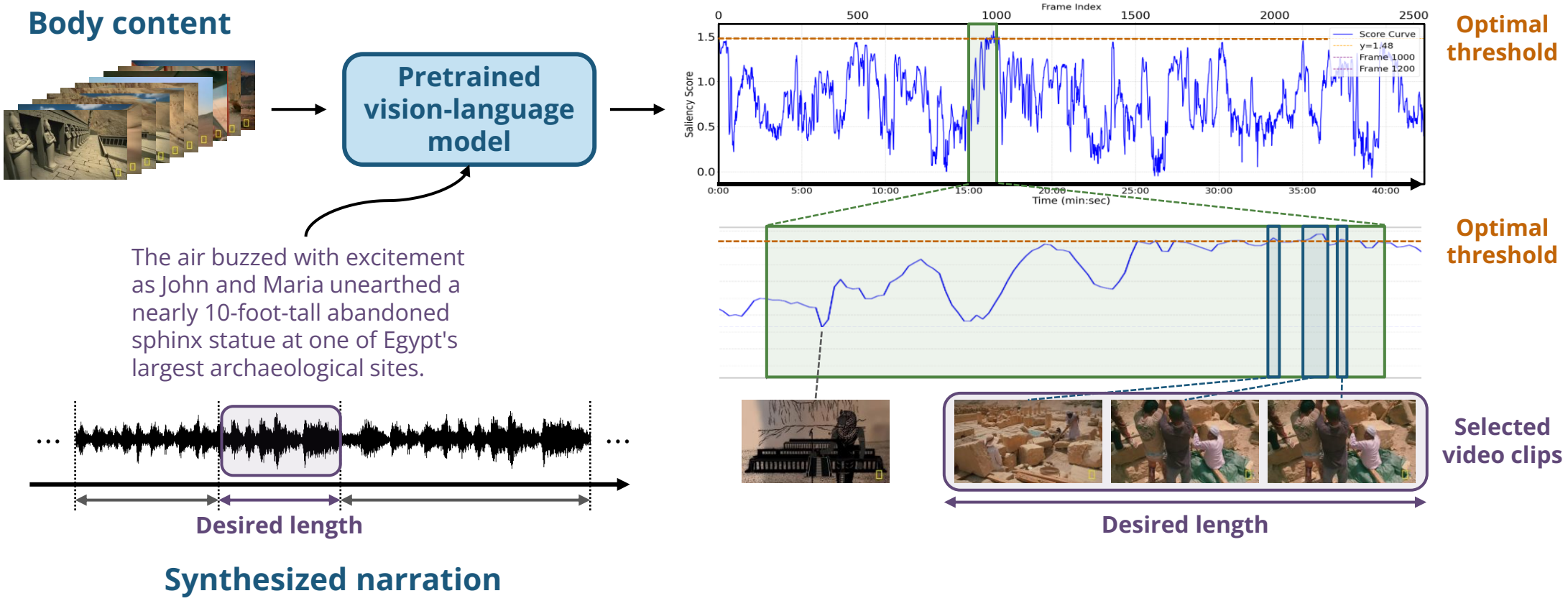

- Unlike **video highlight detection**, a teaser **needs a cohesive narrative**
- Unlike **video summarization**, a teaser **needs to be interesting and engaging**
- Unlike a **movie trailer**, a documentary teaser is more **narration-focused**
- A documentary teaser **needs to preserve the factual accuracy**

Narration-Centered Long-to-Short Video Editing

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

Finding Accompanying Visuals for Narrations

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

Example Results


Ground truth

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

TeaserGen-PT

1		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.
2		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.
3		Meanwhile, in Aswan, Martina Bartanova's team stumbled upon ancient human remains that could halt their exploration of an unopened tomb.
4		The discovery of a child's remains and a miniature sphinx added layers of mystery to their quest.
5		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.
6		In Dra'abu El Naga, another team meticulously organized fragmented human remains, uncovering stories of women in ancient Egyptian society.
7		American archaeologist Susanne Onstein explored the monumental building campaigns of Hetshepsood, whose colossal obelisks stood as testaments to her reign.
8		Amidst pottery fragments and evidence of temples built by Hetshepsood and her stepson Tutmose III, the archaeologists faced personal reflections and challenges.
9		The winds of history whispered through the sands, as each discovery brought them closer to unraveling the enigmatic legacy of Pharaoh Queen Hetshepsood.
10		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

1		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.
2		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.
3		Meanwhile, in Aswan, Martina Bartanova's team stumbled upon ancient human remains that could halt their exploration of an unopened tomb.
4		The discovery of a child's remains and a miniature sphinx added layers of mystery to their quest.
5		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.
6		In Dra'abu El Naga, another team meticulously organized fragmented human remains, uncovering stories of women in ancient Egyptian society.
7		American archaeologist Susanne Onstein explored the monumental building campaigns of Hetshepsood, whose colossal obelisks stood as testaments to her reign.
8		Amidst pottery fragments and evidence of temples built by Hetshepsood and her stepson Tutmose III, the archaeologists faced personal reflections and challenges.
9		The winds of history whispered through the sands, as each discovery brought them closer to unraveling the enigmatic legacy of Pharaoh Queen Hetshepsood.
10		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?

Example: Egyptian History


Title: “Hatshepsut: Mysteries of the Warrior Pharaoh Queen (Full Episode) | Lost Treasures of Egypt”

wx83.github.io/TeaserGen_Official/

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

Example: Solar System


Title: “Eight Wonders Of Our Solar System | The Planets | BBC Earth Science”

wx83.github.io/TeaserGen_Official/

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

Objective Evaluation


				Repetitiveness			Text-visual correspondence	
Model	Query	Decoding	DP	F1 (%)↑	REP (%)	SCR (%)	CLIPScore	VTGHLS
Baseline models								
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
Pretraining-based models								
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
Learning-based models								
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	×	**1.88**	24.16	41.97	0.58	0.74
TeaserGen-LR	Narration	Beam Search	✓	**1.88**	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!

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

Subjective Evaluation


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

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

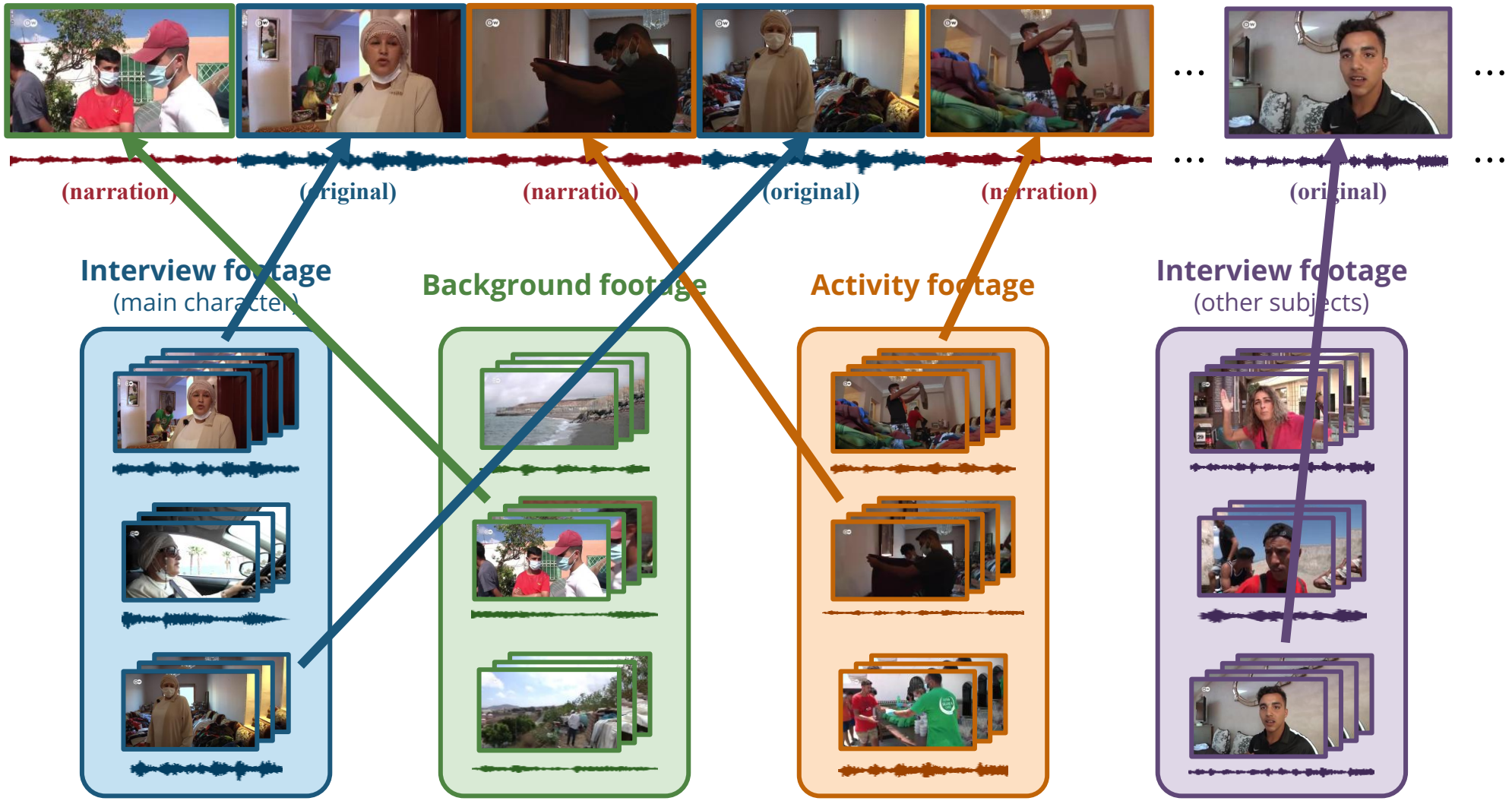
Weihan Xu, Paul Pu Liang, Haven Kim, Julian McAuley, Taylor Berg-Kirkpatrick, and Hao-Wen Dong, "TeaserGen: Generating Teasers for Long Documentaries," *ICLR*, 2025.

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?

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

Video Editing



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

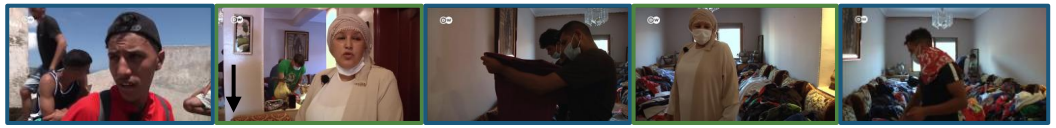
Weihan Xu[1] Yimeng Ma[1] Jingyue Huang[2] Yang Li[1] Weyne Ma[3]
Taylor Berg-Kirkpatrick[2] Julian McAuley[2] Paul Pu Liang[2] **Hao-Wen Dong**[4]

[1] Duke University [2] UC San Diego [3] MBZUAI [4] MIT [5] University of Michigan

Duke UNIVERSITY UC San Diego MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE Massachusetts Institute of Technology UNIVERSITY OF MICHIGAN

Learning to *Quote* a Video

Input long video



Transcribed script

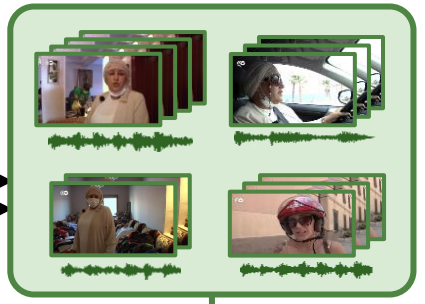
00:03–00:17	Narration	Saba Hamed's house has become a gathering place, because she's taken it upon herself to help refugees. Saba is Spanish, but has Moroccan roots and speaks Arabic. Her door is always open to everyone.
00:20–00:28	Interview	These folks wander aimlessly through the city, sleep outdoors, on the beach, in huts. It's bad.
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	All these clothes here were delivered the day before yesterday. All of these shirts, pants, underwear, socks, shoes, everything.
00:56–00:60	Narration	Bilal Hanouti comes here every day and is thankful for the offer.

Story script

Saba Hamed lives in Ceuta, a Spanish autonomous community on the north African coast. For weeks, people have been trying to break through to Spain from Morocco. **<QUOTE>** It's resulting in a crisis. ...

Extracted frames

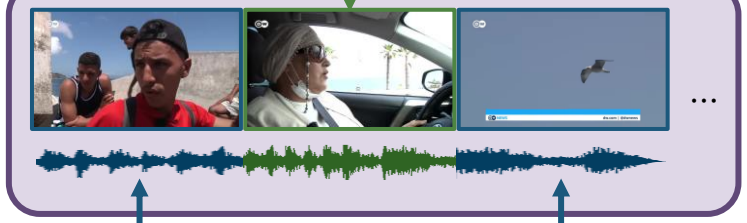
Extracted quotable video clips



Quote generator

Quote retriever

Output short video



Text-to-speech synthesis

Narration-visual matching

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

Learning to *Quote* a Video

REGen-DQ
(direct quote)

$$\ldots, x_i, \langle \text{SOQ} \rangle, y_1, \ldots, y_n, \langle \text{EOQ} \rangle, x_{i+1}, \ldots$$

Quote

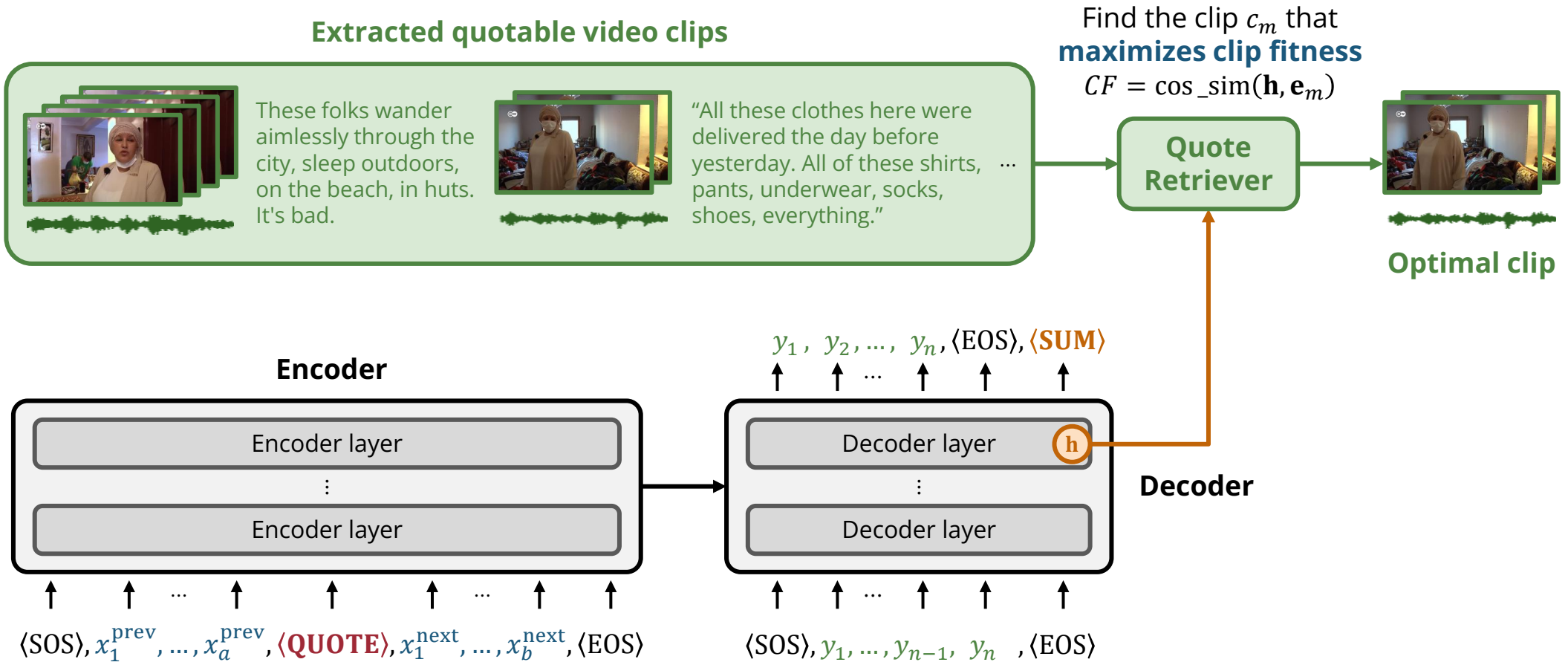
REGen-IDQ
(indirect quote)

$$\ldots, x_i, \langle \text{QUOTE} \rangle, x_{i+1}, \ldots$$

To be retrieved later!

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

Retrieving a Video Quote



Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

Measuring Clip Fitness

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

$$CF := \cos\_\text{sim}(\mathbf{h}, \mathbf{e}_m)$$

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

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

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	**5.00**	**17.50**	**30.00**	**3.56 ± 0.22**
QuoteRetriever-TV	Text+Visual	**5.00**	15.00	23.33	3.49 ± 0.26

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

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

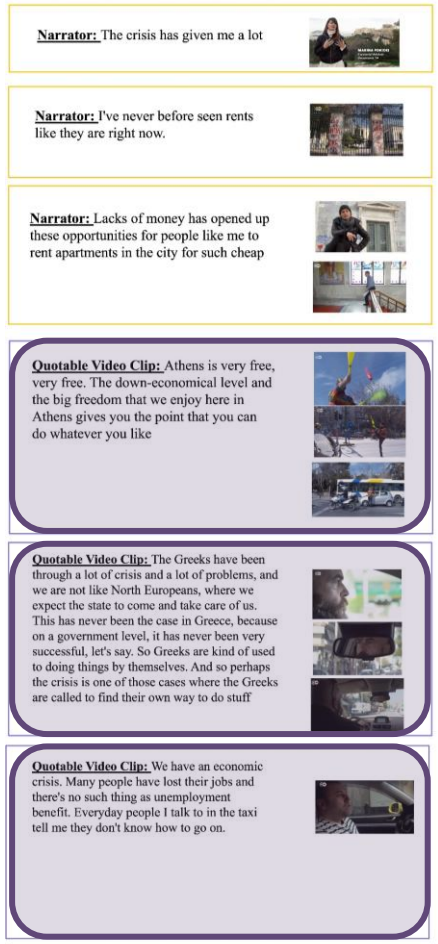
Example: Modern Art Exhibition

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

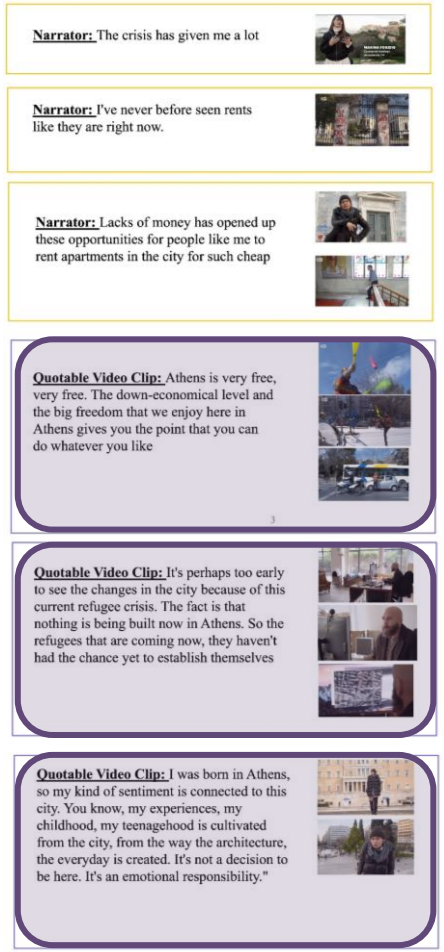
youtu.be/agij_IxGjCI

Example Results

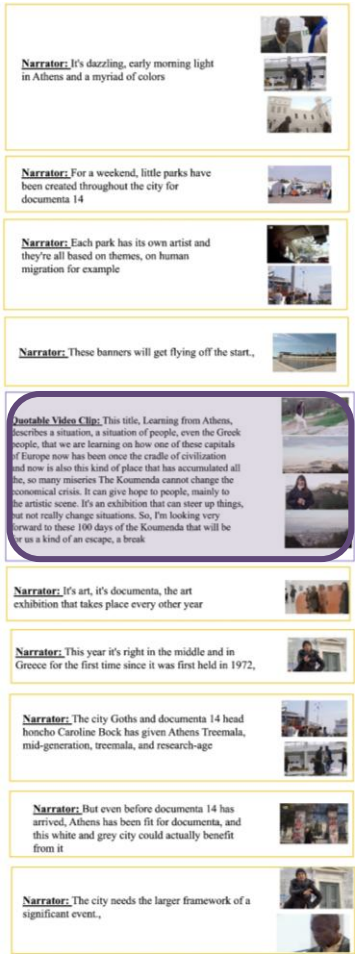
REGen-IDQ-TV



REGen-IDQ-T



REGen-DQ



Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

Example: Athens History

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

wx83.github.io/REGen/

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, “REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing,” *arXiv preprint arXiv:2505.18880*, 2025.

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	**22.61**	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	**2.07**	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	**0.41**	**0.57**
REGen-IDQ-TV	81	35 ± 31	1.90	19.86	**9.70**	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!

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

Subjective Evaluation

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

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

Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

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

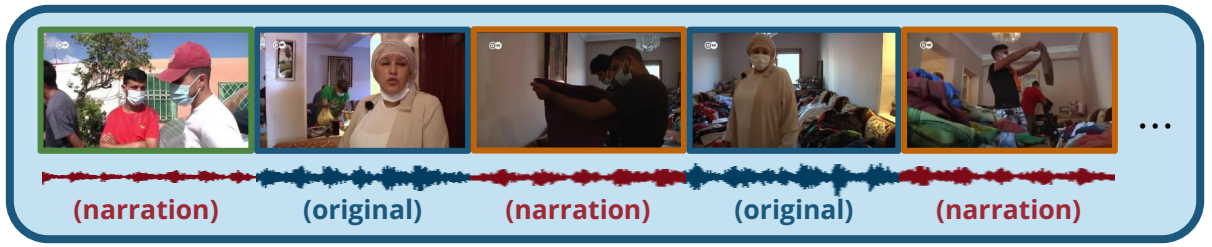
Weihan Xu, Yimeng Ma, Jingyue Huang, Yang Li, Wenye Ma, Taylor Berg-Kirkpatrick, Julian McAuley, Paul Pu Liang, and Hao-Wen Dong, "REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing," *arXiv preprint arXiv:2505.18880*, 2025.

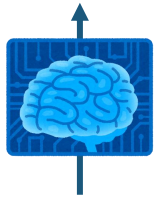
Towards AI-assisted Video Editing

Future Work: Multimodal RAG-based Video Editing

Raw video materials

Output video





Video retriever

Editable video script

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	*"These folks wander aimlessly through the city, sleep outdoors, on the beach, in huts. It's bad."*
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	*"All these clothes here were delivered the day before yesterday. All of these shirts, pants, underwear, socks, shoes, everything."*
00:56 – 00:60	Narration	Bilal Hanouti comes here every day and is thankful for the offer.

⋮

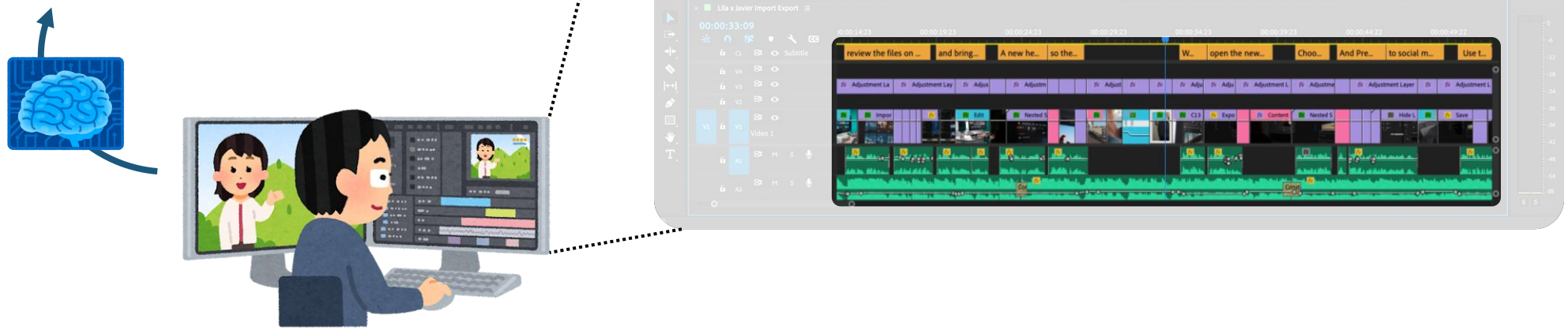
User prompt

Create a short video about the story of Saba Hamed in the context of refugees trying to enter EU

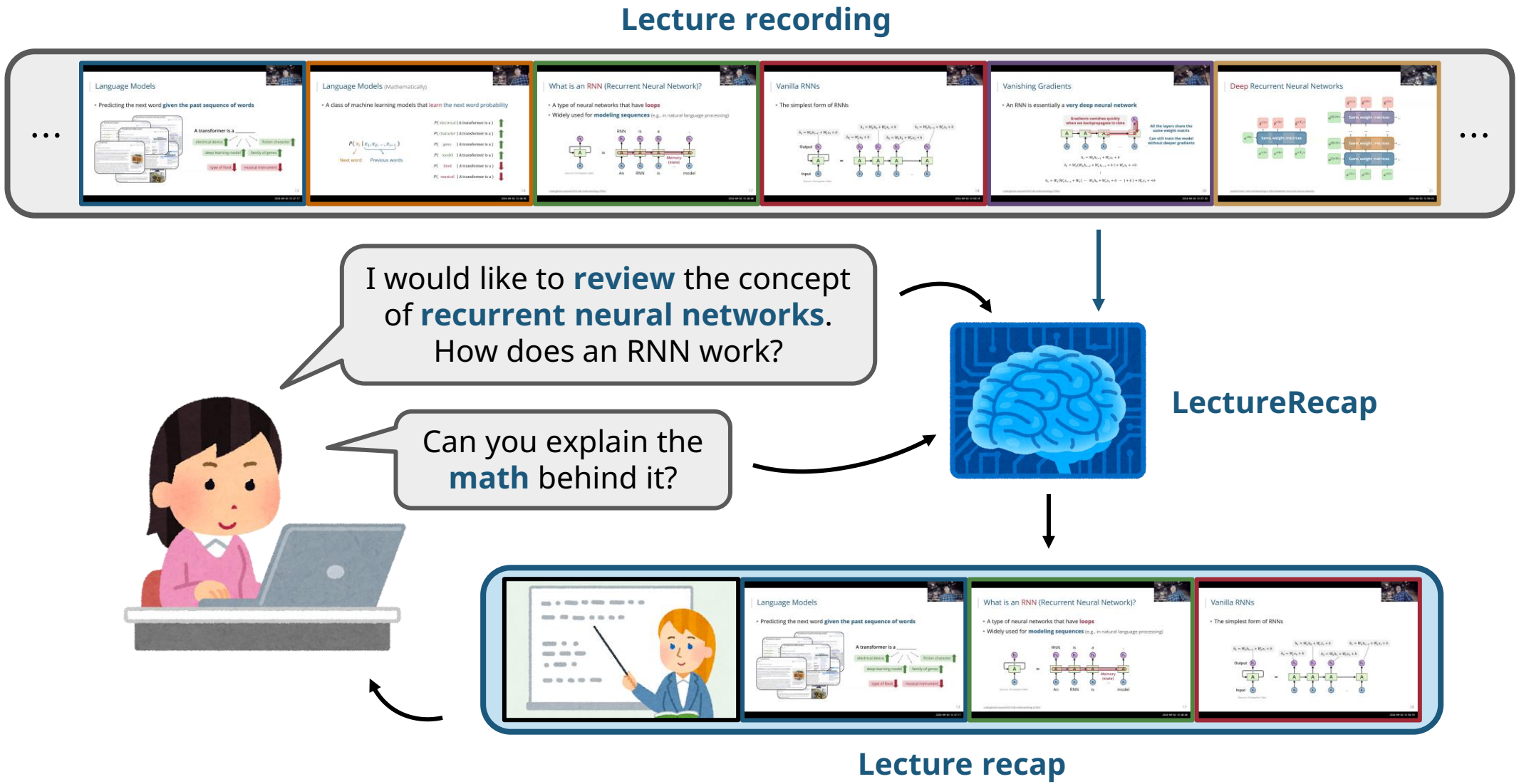
Multimodal LLM

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	*"These folks wander aimlessly through the city, sleep outdoors, on the beach, in huts. It's bad."*
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	*"All these clothes here were delivered the day before yesterday. All of these shirts, pants, underwear, socks, shoes, everything."*
00:56 – 00:60	Narration	Bilal Hanouti comes here every day and is thankful for the offer.

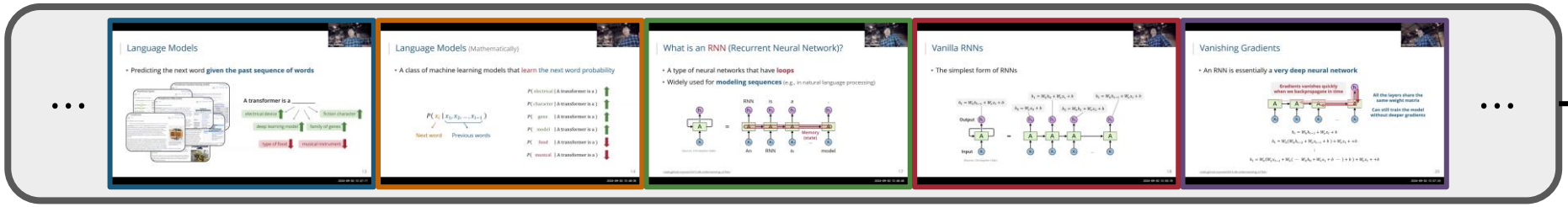


Future Work: LectureRecap



Future Work: LectureRecap

Lecture recording



User query

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

Speech recognition → **Script generation**

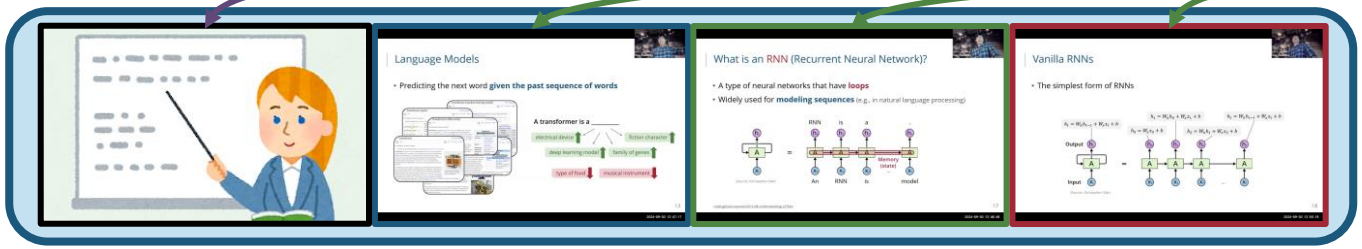
Video script

[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

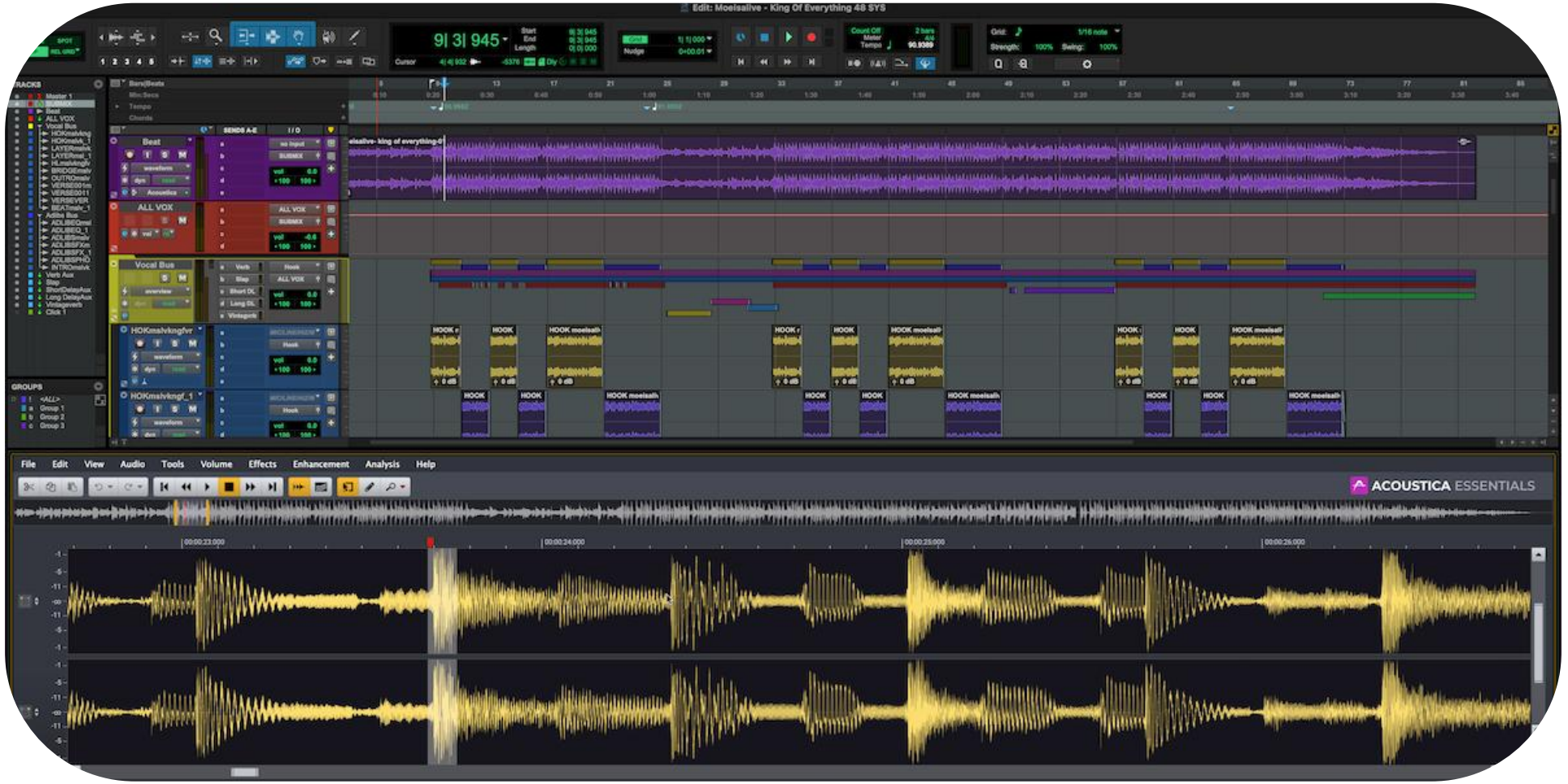
Lecture recap



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)

avid.com/pro-tools/whats-new

Future Work: Integrating GenAI into Music Production

(Source: Avid)

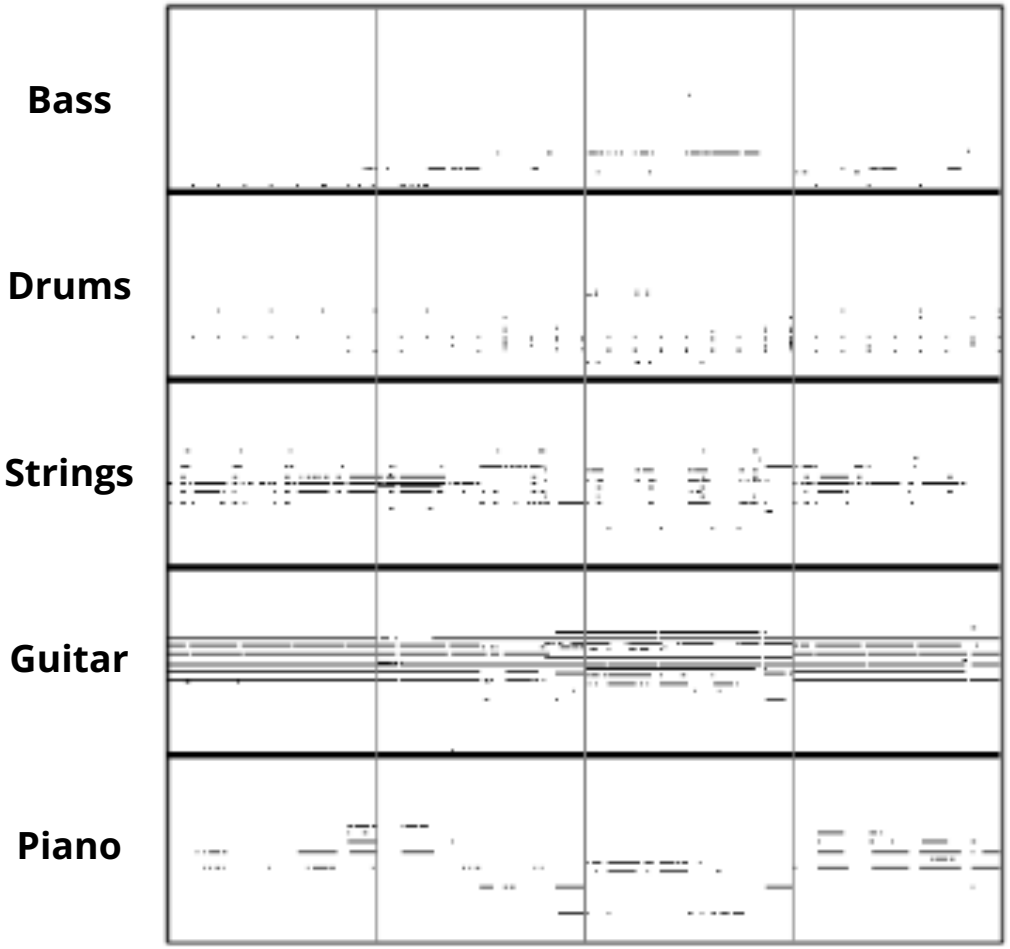
avid.com/pro-tools

Augmenting Human Creativity with AI

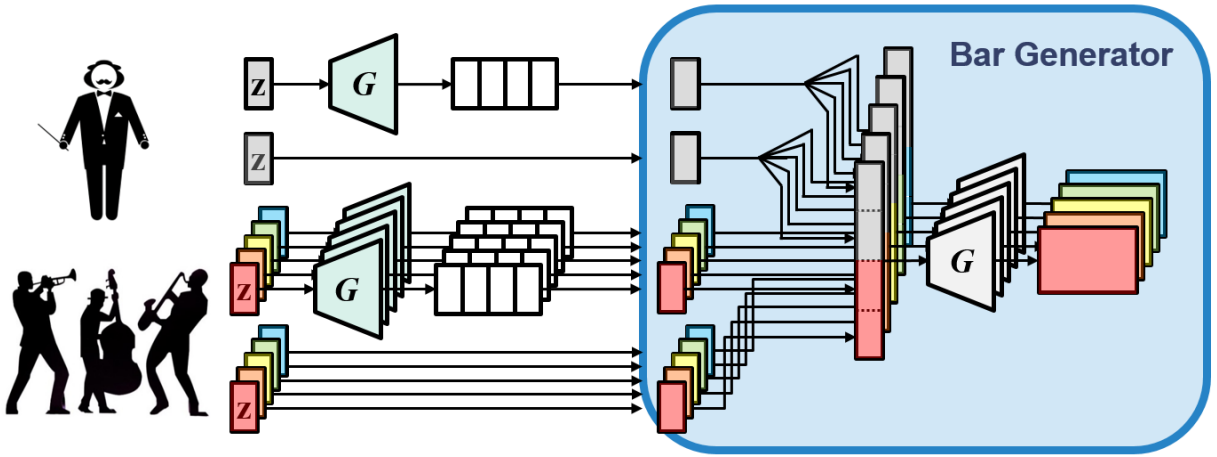
- **Novel Generative Models for New Domains**
  - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024), **text-to-music generation** (ISMIR 2025), **video-to-music generation** (ISMIR 2025), **symbolic music processing tools** (ISMIR LBD 2019, ISMIR 2020)
- **AI-assisted 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



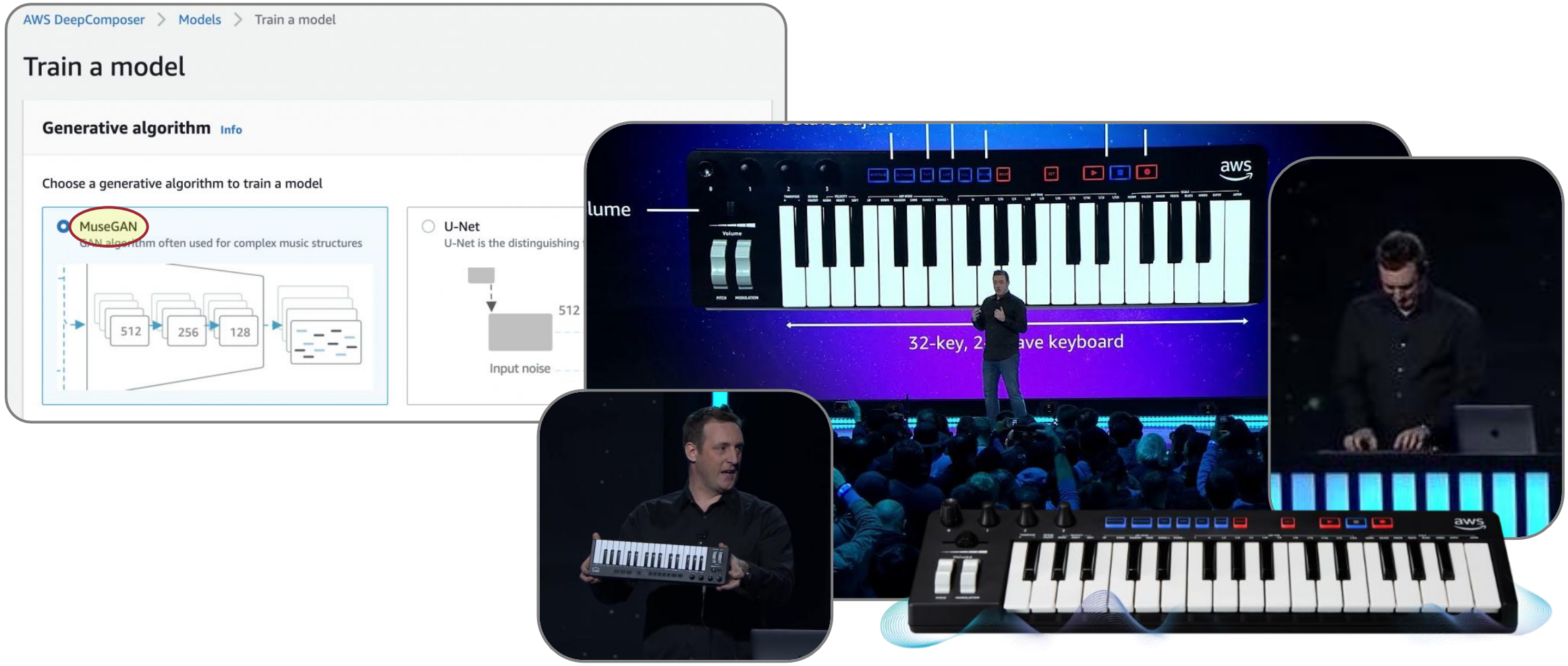
MuseGAN Generator



中央研究院
ACADEMIA SINICA

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

MuseGAN Features in AWS DeepComposer (2020)



amazon.com/dp/B07YGZ4V5B/

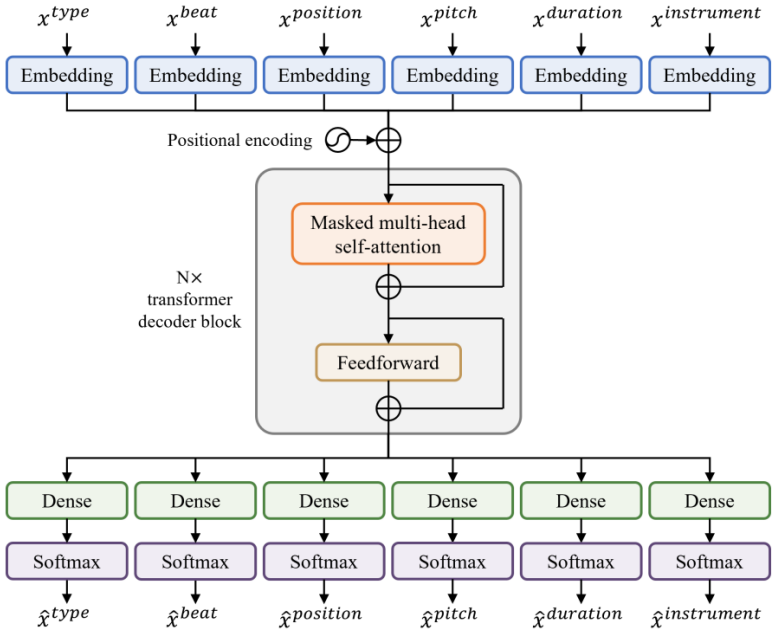
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," *AWS News Blog*, April 2, 2020.

Generating Multitrack Music with Transformers (ICASSP 2023)

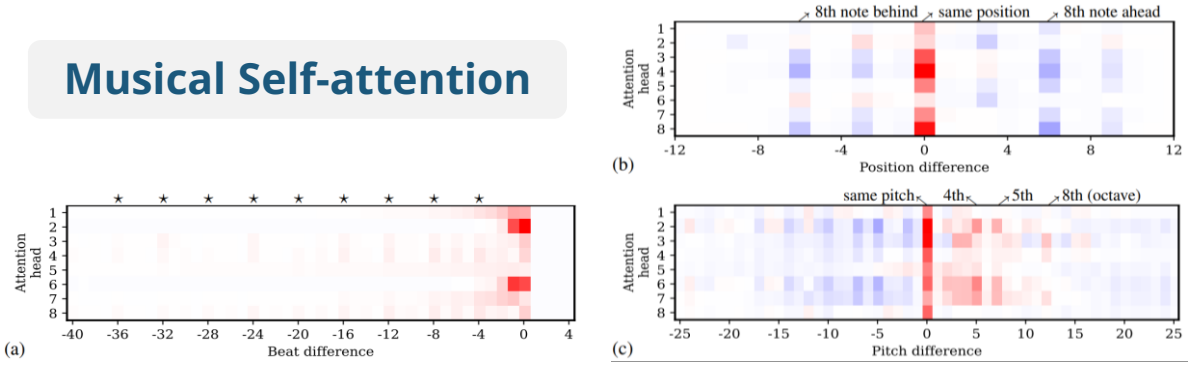
Multitrack Music Representation

```
(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

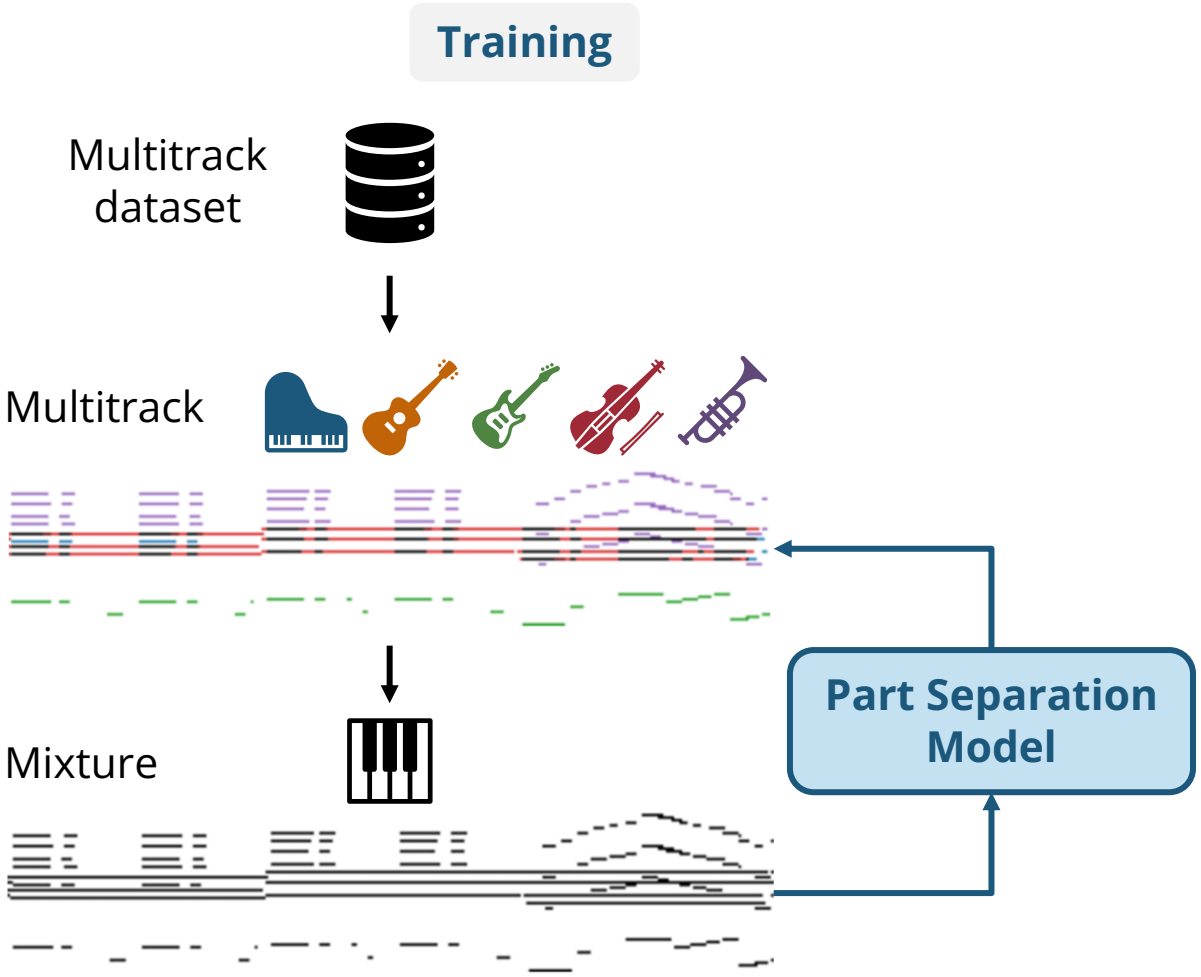


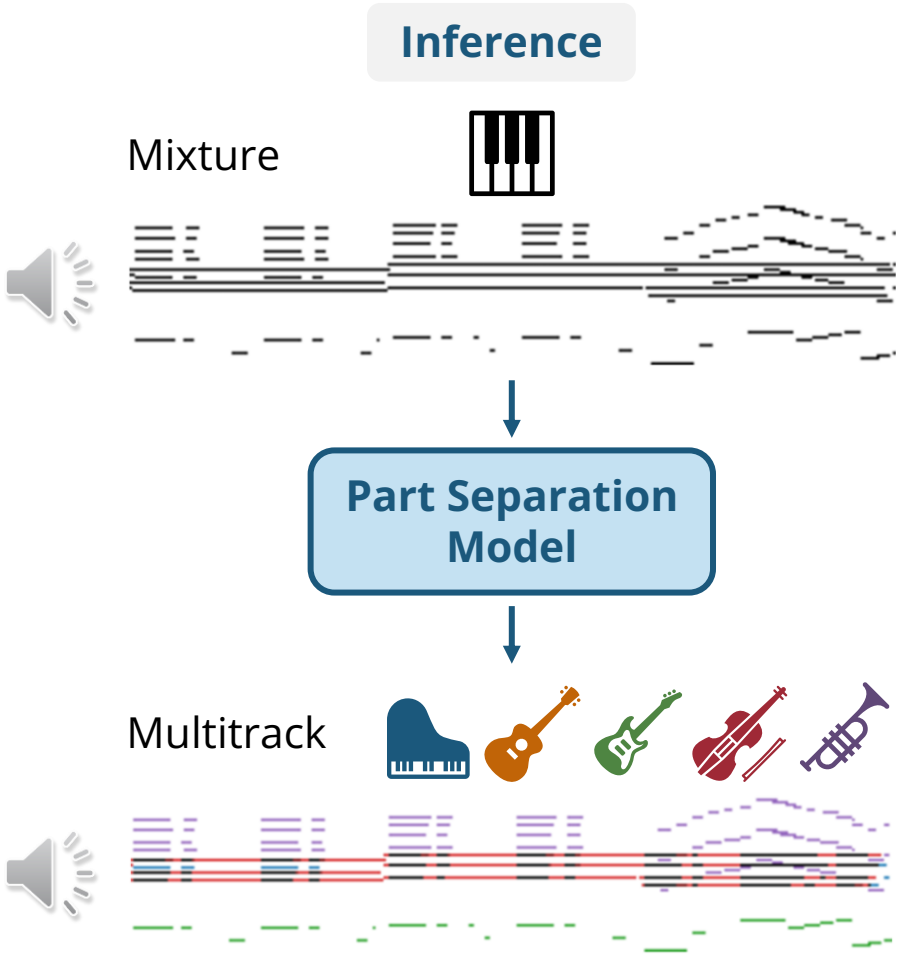
UC San Diego

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

Automatic Instrumentation (**ISMIR 2021**)

Stanford UC San Diego

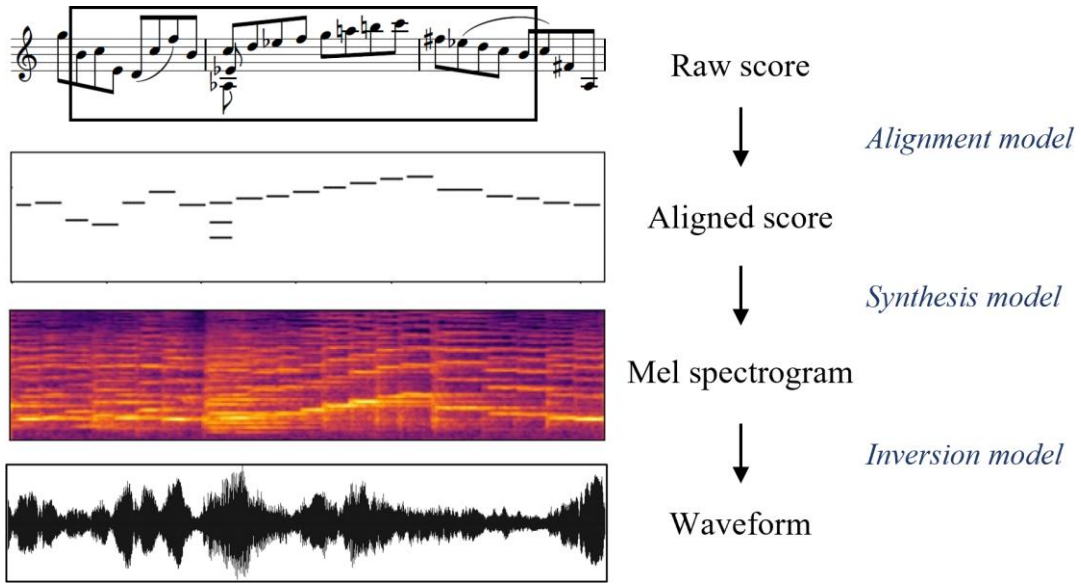




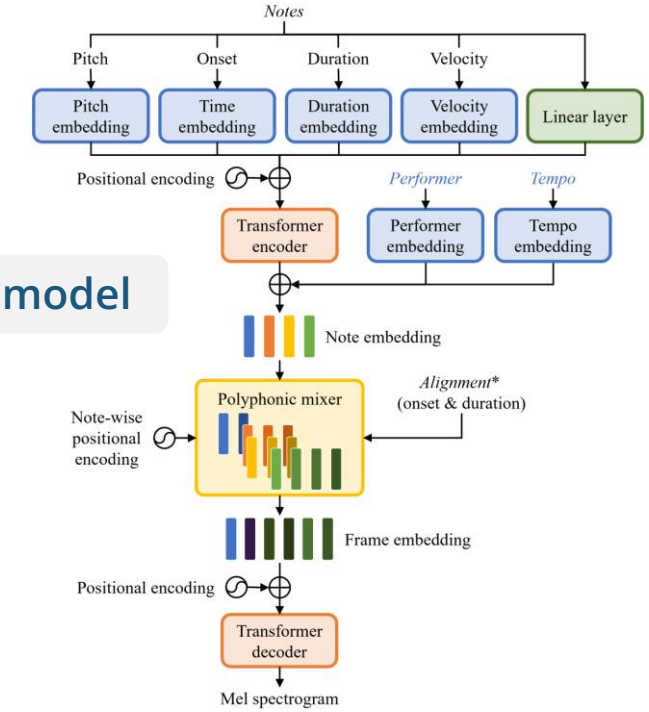
Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *ISMIR*, 2021.

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

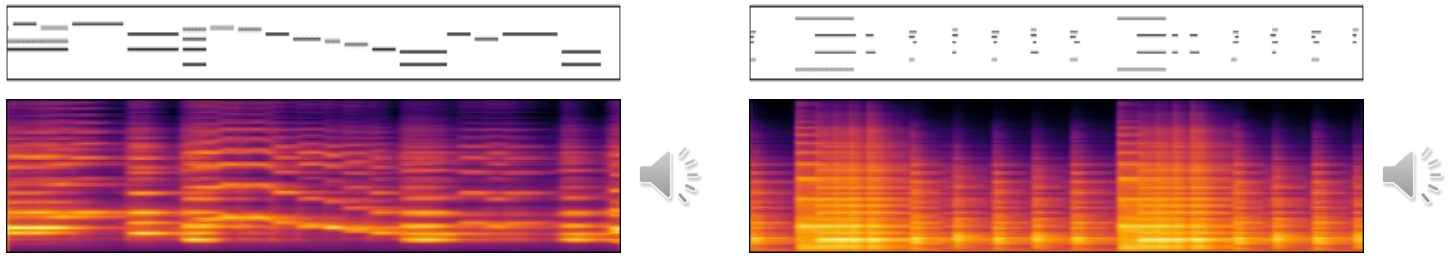
Performance synthesis



TTS-based model



Example results



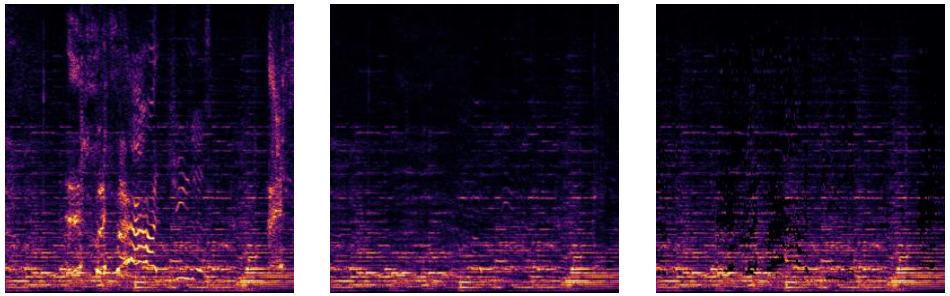
Dolby UC San Diego

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

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

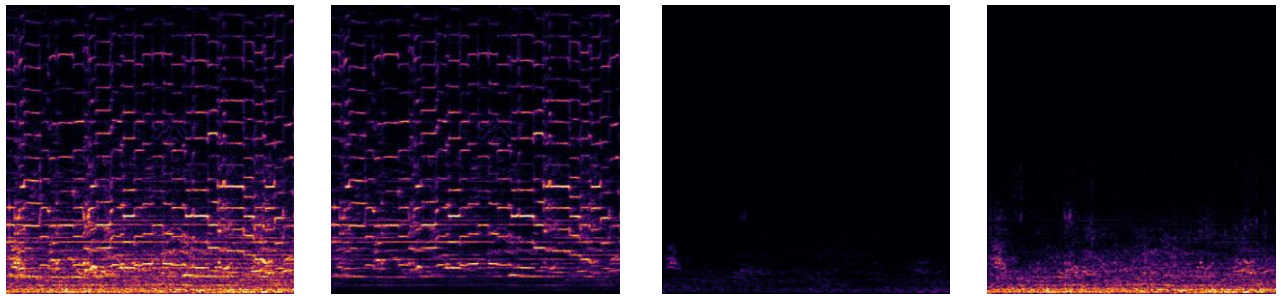
Query: "*playing harpsichord*"

Mixture **Prediction** **Ground truth**

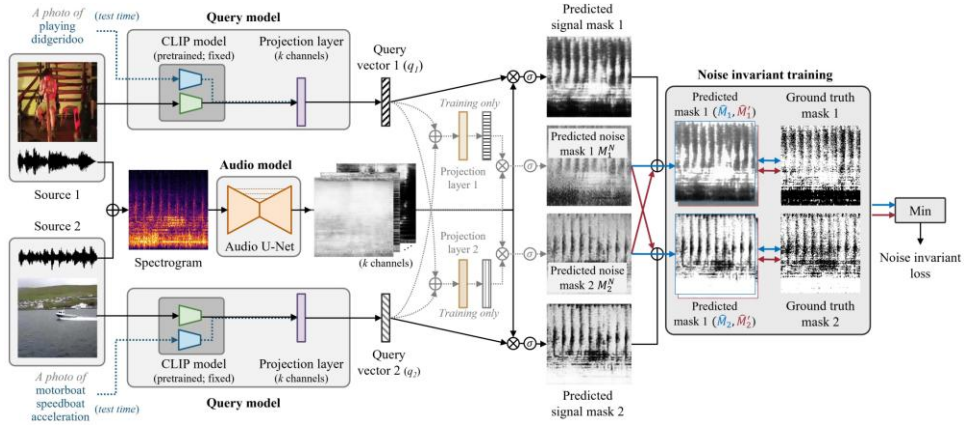


Query: "*playing bagpipe*"

Mixture **Prediction** **Noise head 1** **Noise head 2**

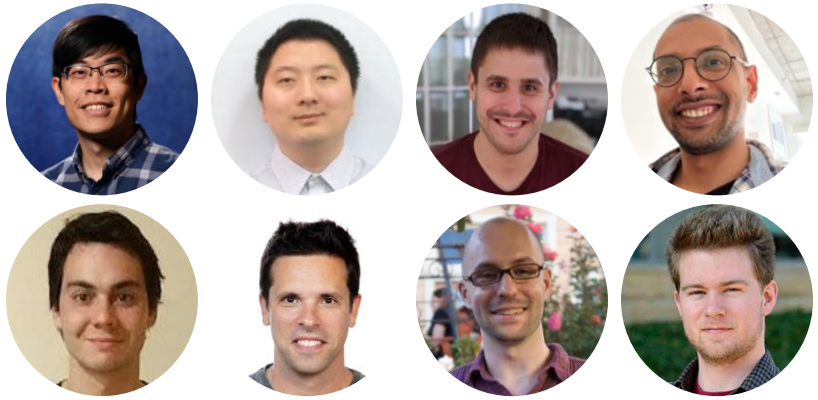


Text-queried sound separation model



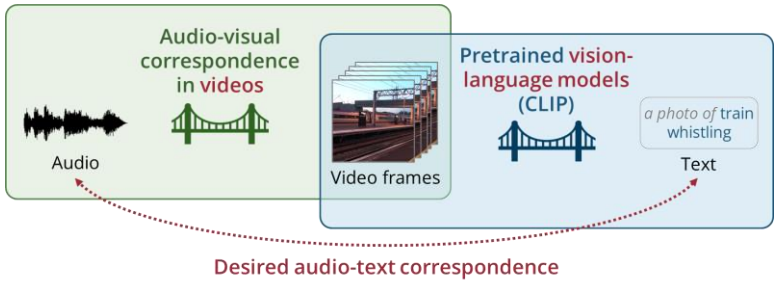
Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick, "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *ICLR*, 2023.

Text-to-Audio Synthesis (**WASPAA 2023**)

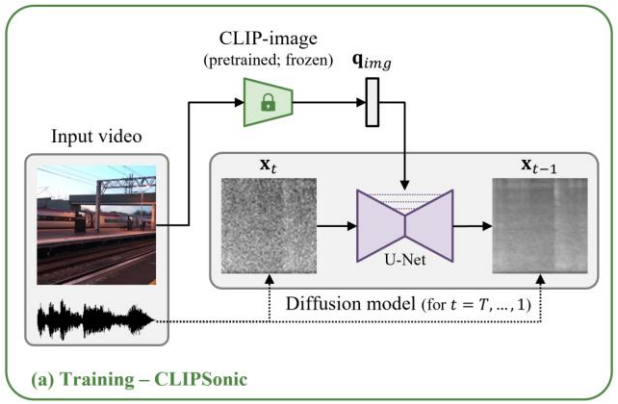


Dolby UC San Diego

Learning Sounds from Noisy Videos



Training



Inference

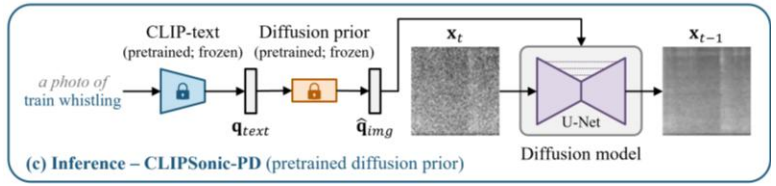


Image-to-sound results

Text-to-sound results

Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley, "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.

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

(Source: The Denver Post)

(Source: UploadVR)

(Source: Descript)

(Source: Daily Bruin)

(Source: Wikimedia Commons)

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

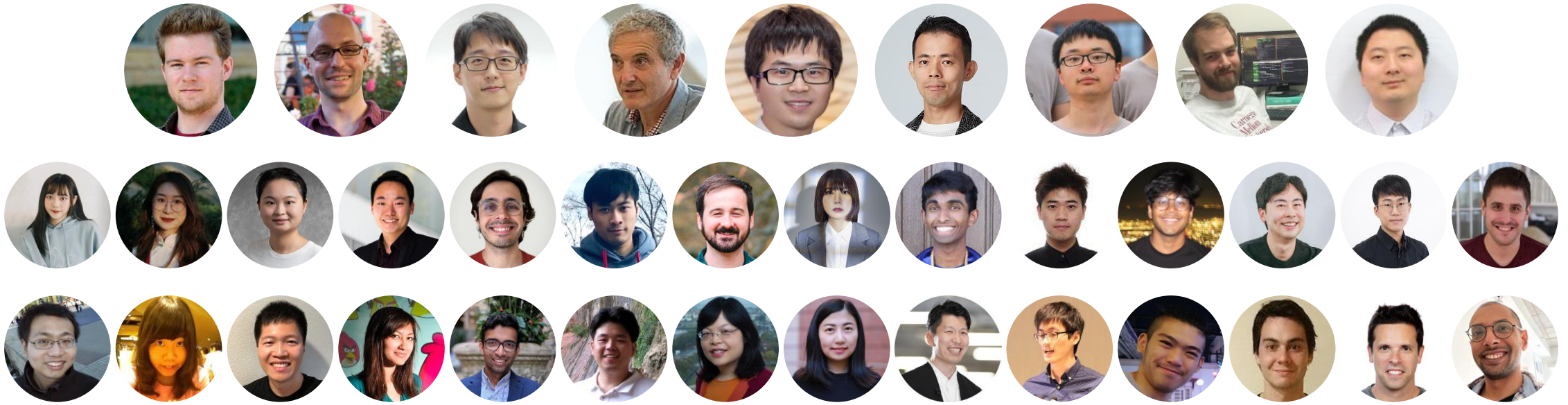
Augmenting Human Creativity with AI

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

Performing Arts Technology (PAT)

Augmenting Human Creativity with AI

Nothing would have been possible without all my fantastic collaborators!



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

hermandong.com / hwdong@umich.edu