

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

September 12, 2025

Augmenting Human Creativity with AI

- **Novel Generative Models for New Domains**
 - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024, AIMG 2024), **text-to-symbolic music generation** (ISMIR LBD 2024, ISMIR 2025)
- **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), **a capella source separation** (ISMIR LBD 2025)
- **Multimodal Generative Models for Content Creation**
 - **Queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WSS 2023, WASPAA 2023), **text-to-music generation** (ISMIR LBD 2024, arXiv 2024), **video-to-music generation** (ISMIR 2025, ISMIR LBD 2025), **long-to-short video editing** (ICLR 2025, arXiv 2025)

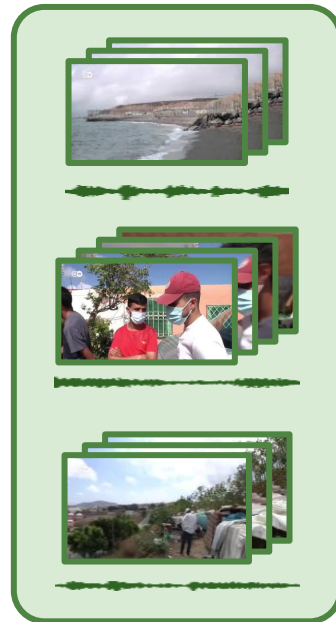
Video Editing



Interview footage
(main character)



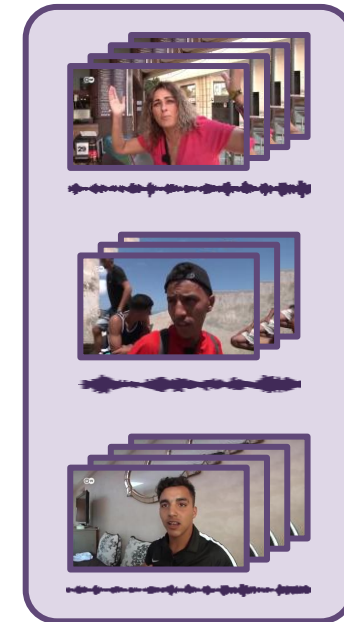
Background footage



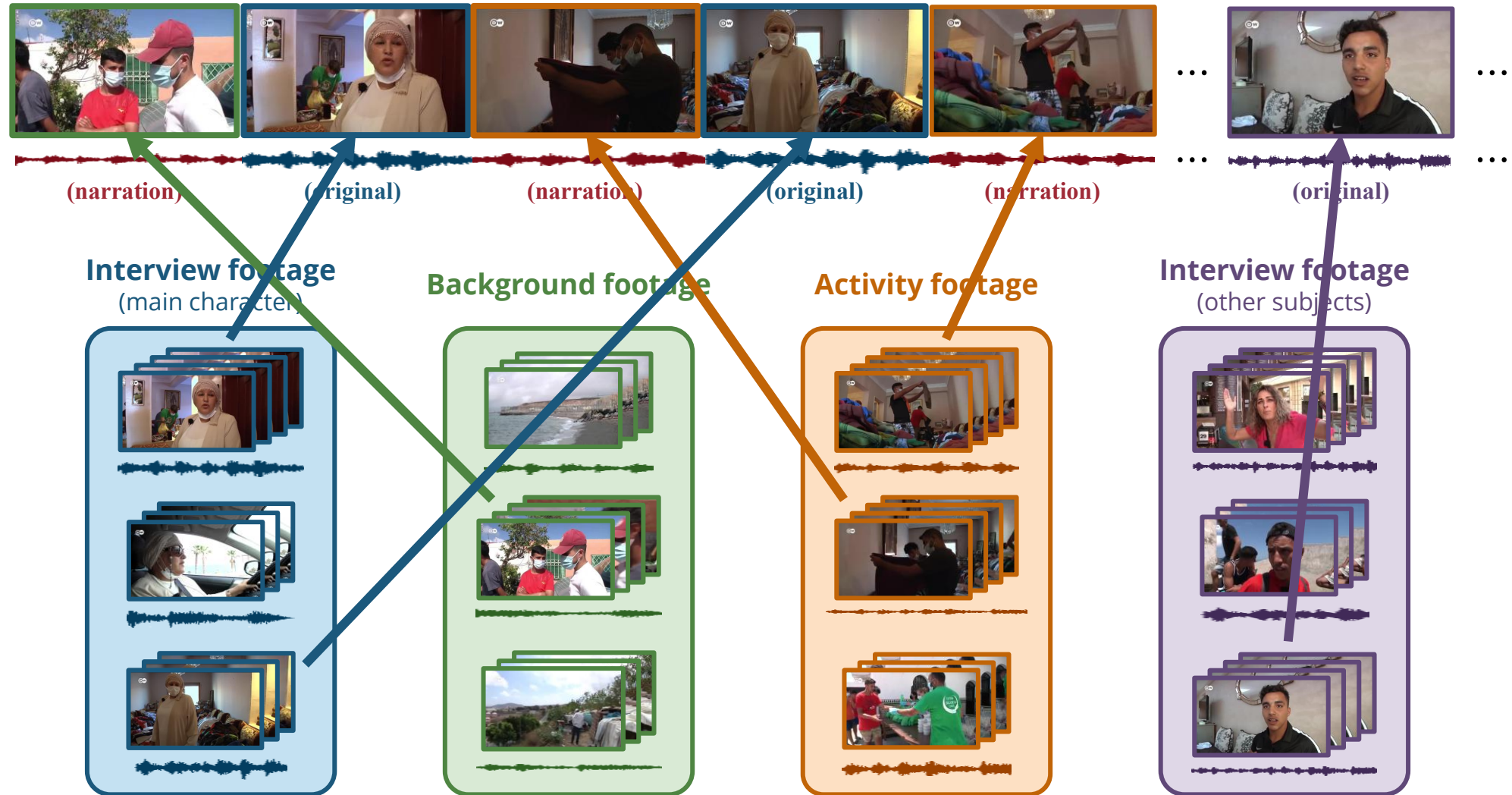
Activity footage



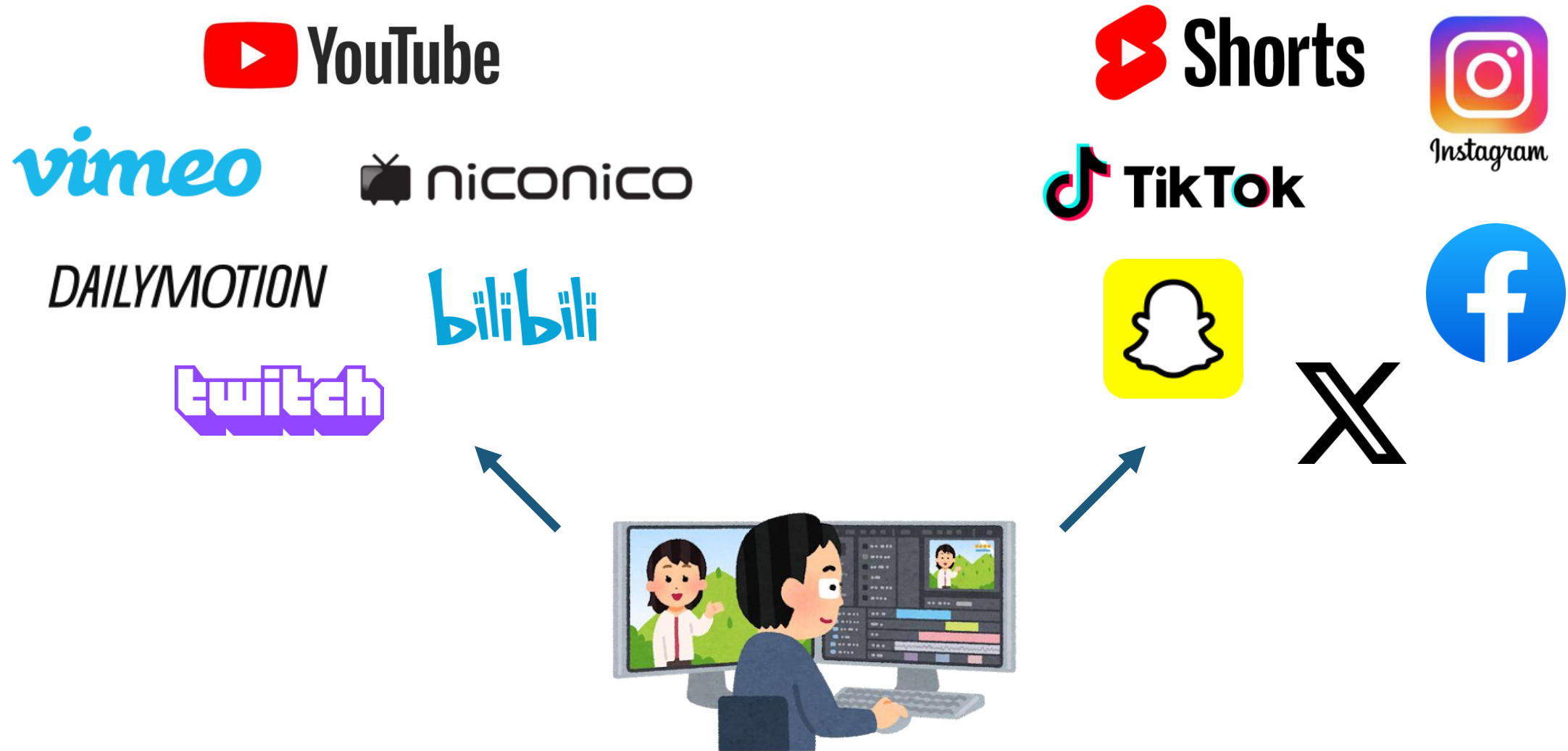
Interview footage
(other subjects)



Video Editing



Fast-growing Short Video Platforms



| Generating **Shorts** from Long Videos

- For content creators, help **promoting long video contents**
- For content consumers, help **digest information in a more engaging way**

TeaserGen: Generating Teasers for Long Documentaries

Weihan Xu¹ Paul Pu Liang² Haven Kim³
Julian McAuley³ Taylor Berg-Kirkpatrick³ **Hao-Wen Dong⁴**

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



Documentary Teaser Generation

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

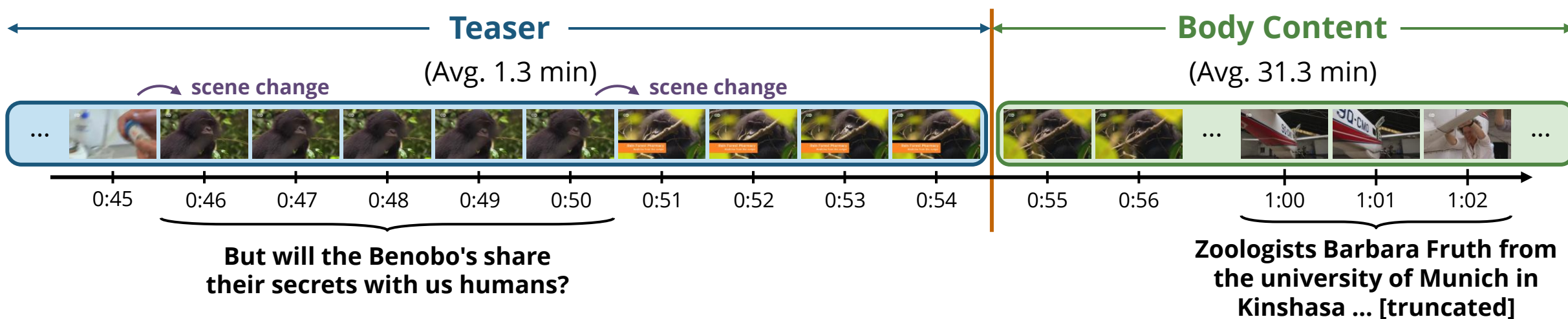


youtu.be/89xTTczbv0E

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**

DocumentaryNet: A New Dataset for Documentaries

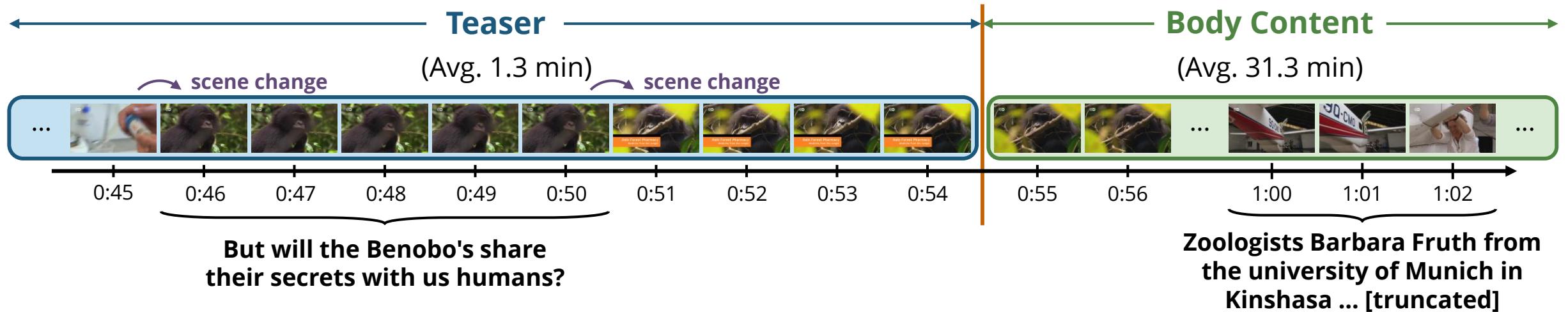


DocumentaryNet: A New Dataset for Documentaries

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



DocumentaryNet: A New Dataset for Documentaries



Title: Medicine from the jungle - Rainforest pharmacy | DW Documentary

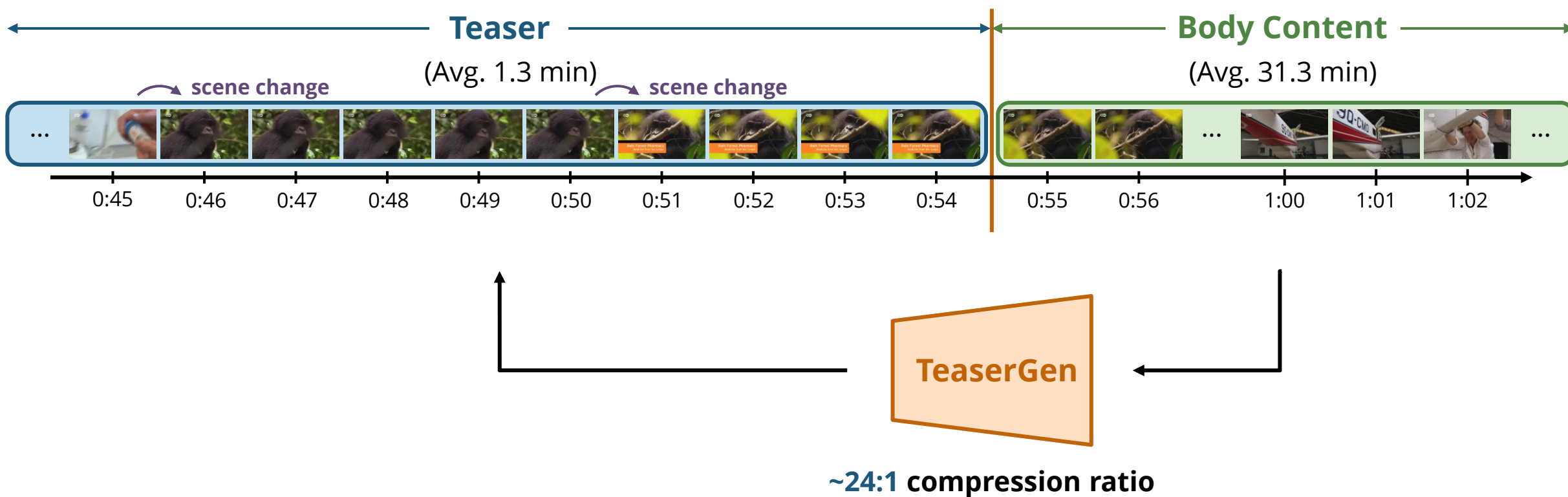
User Annotated Tags: Education, rainforest, medicine, psychology, apes, nature, Africa, monkey, Barbara Fruth, Democratic Republic of the Congo, natural medicine, bonobos, tropical rainforest, endangered species, animals, pharmacy , pharmaceutical, research, DW, Deutsche Welle, documentary

Narration with Timestamps: (in sec)

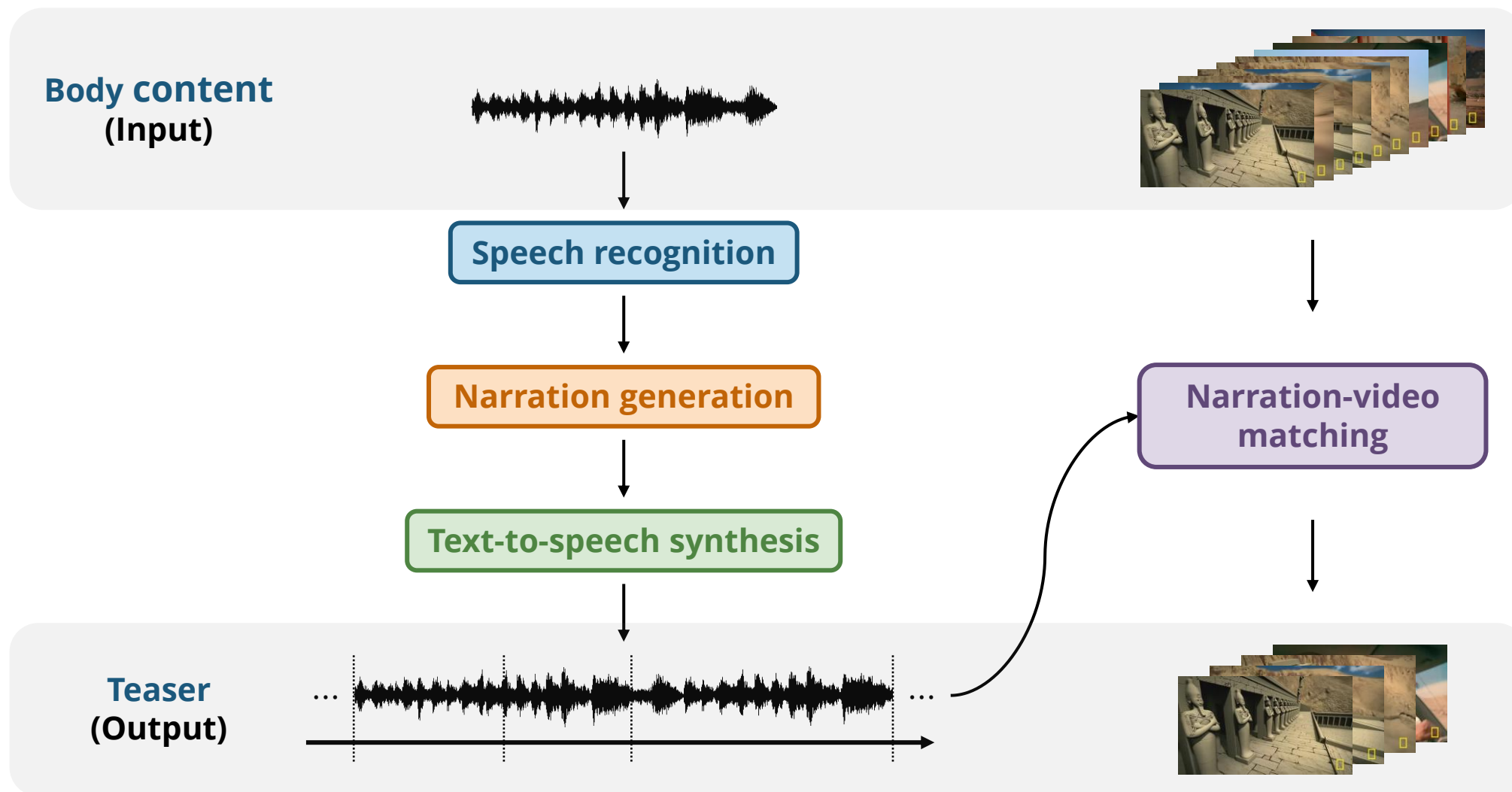
[2.6 – 15.92] Benobo Apes are among our closest living relatives, and the Benobo's here in Africa's Congo basin have a valuable treasure. They know which plants have healing properties. Maybe we can learn from them.

[17.89 – 26.59] It's always interesting when one specific individual within a group eats something different. That could turn out to be some kind of medicinal plant. ...

Generating Teasers from Long Documentaries



A Narration-Centered Approach to Teaser Generation



Leveraging LLMs for Narration Generation

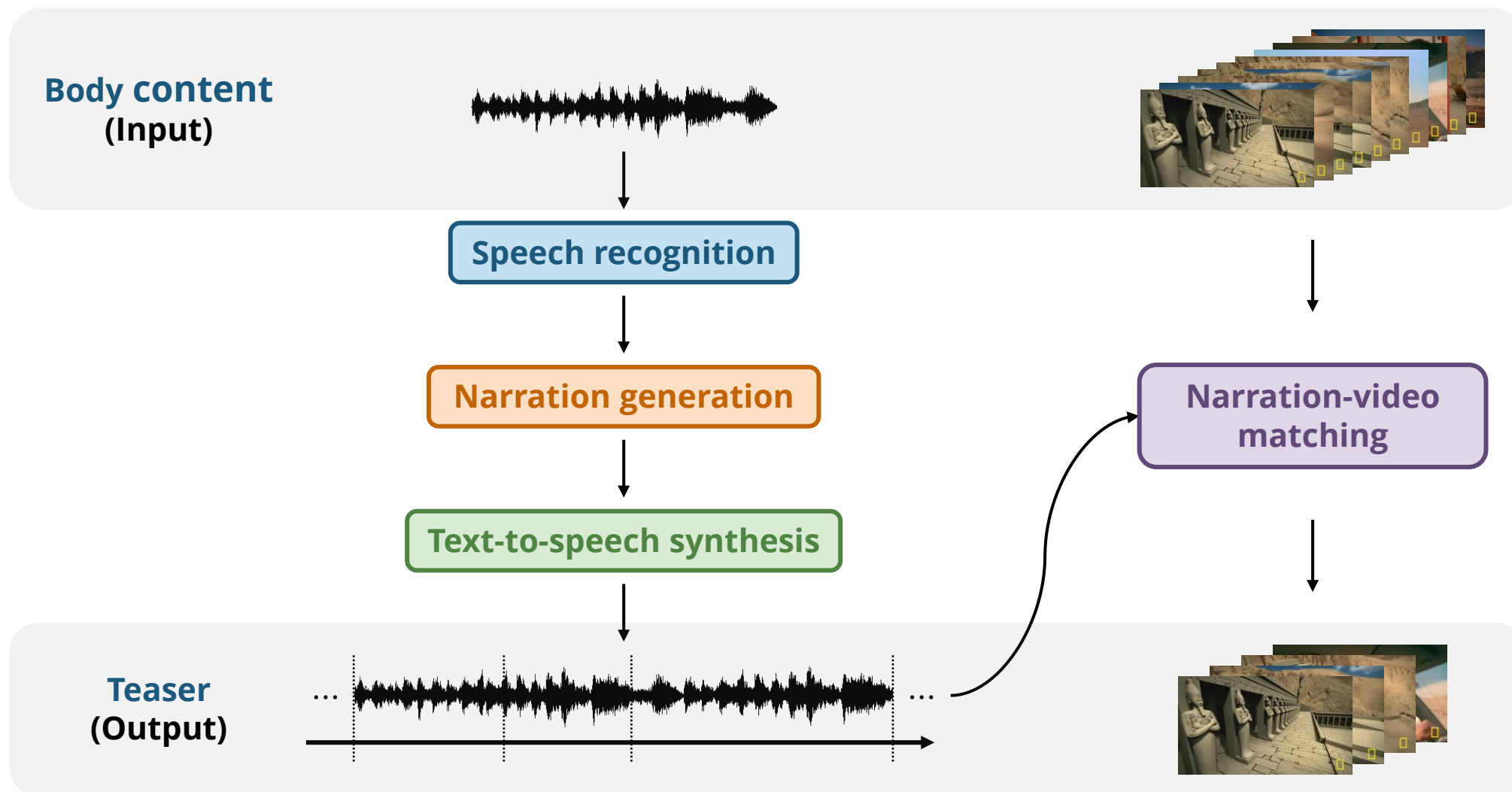
- Break the full narration into **10 segments** (avg. 3,900 words)
- Use GPT-4o to **summarize each segment**
- **Rewrite** the 10 summarized sentences **into a cohesive paragraph**
 - “Rewrite the paragraph into an engaging story opening in 10 sentences or less, keeping all names and avoiding being 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?”*

Leveraging LLMs for Narration Generation

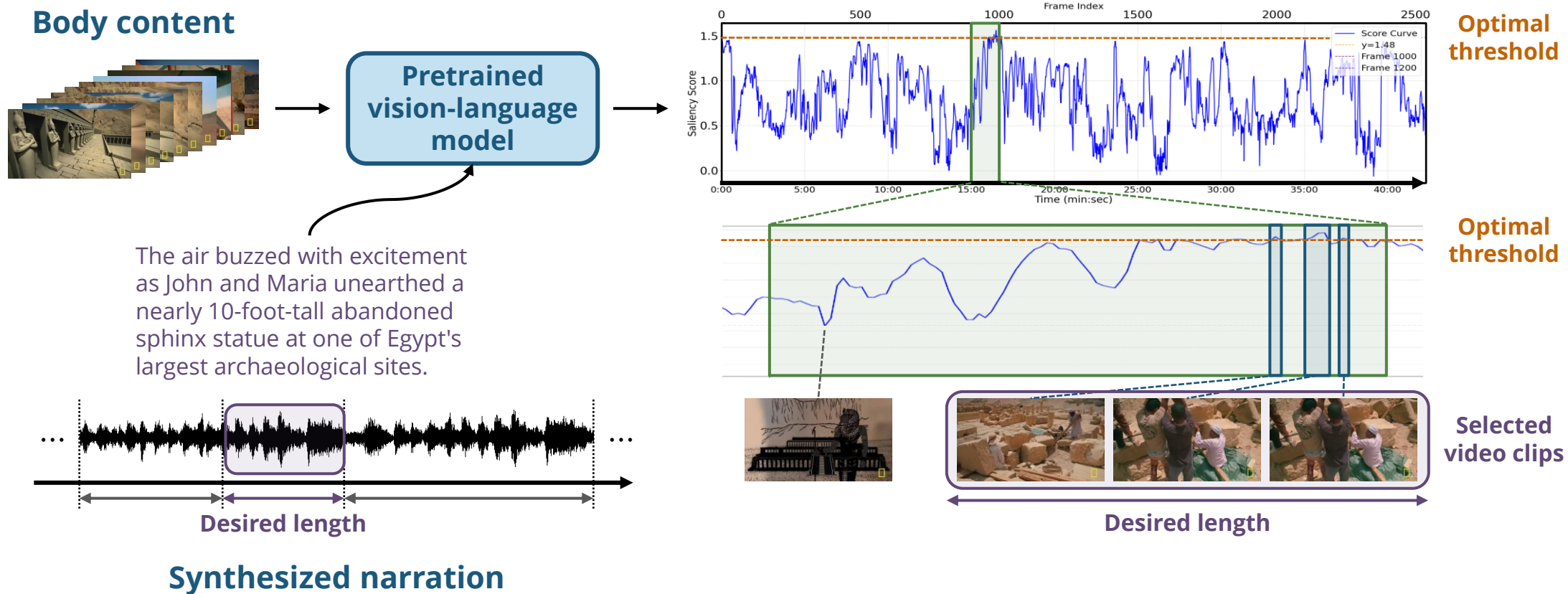
- Break the full narration into **10 segments** (avg. 3,900 words)
- Use GPT-4o to **summarize each segment**
- **Rewrite** the 10 summarized sentences **into a cohesive paragraph**
- **Propose an ending question**

Narration	Organization↑	Informativeness↑	Engagingness↑
Naive summarization	3.58 ± 0.57	3.72 ± 0.47	3.60 ± 0.56
Finely-tuned scripts	3.88 ± 0.44	3.82 ± 0.54	3.70 ± 0.46

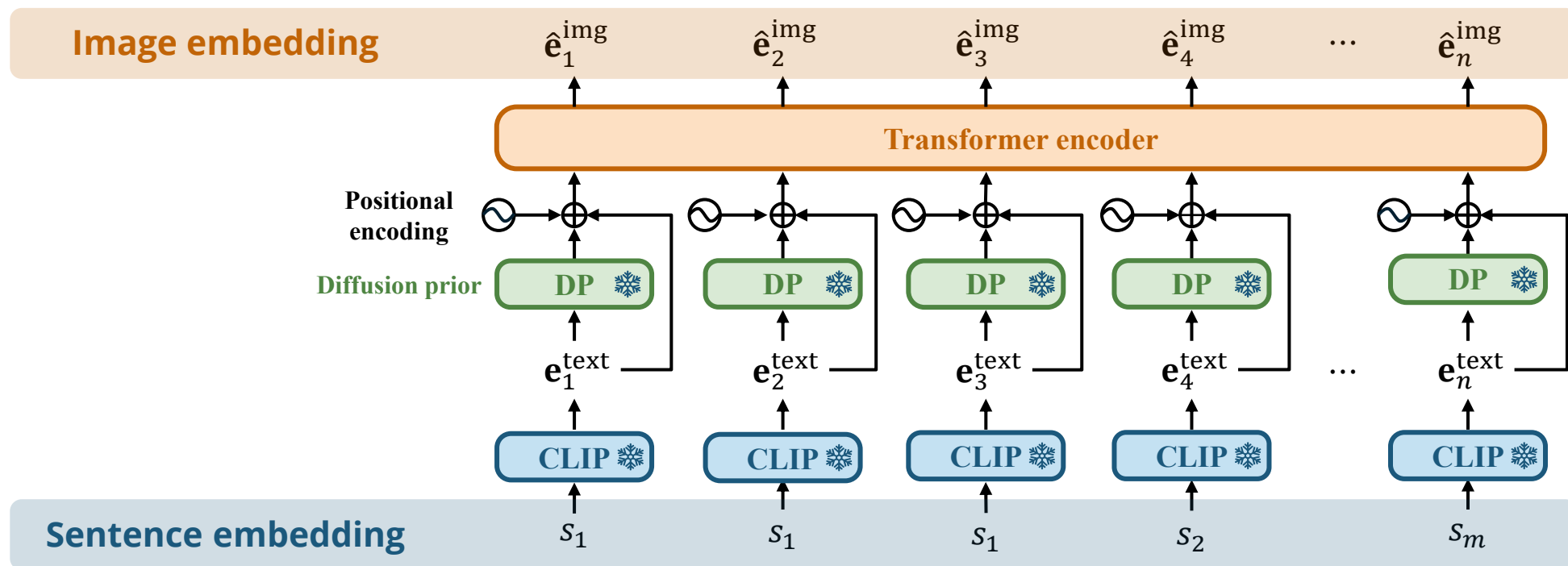
A Narration-Centered Approach to Teaser Generation



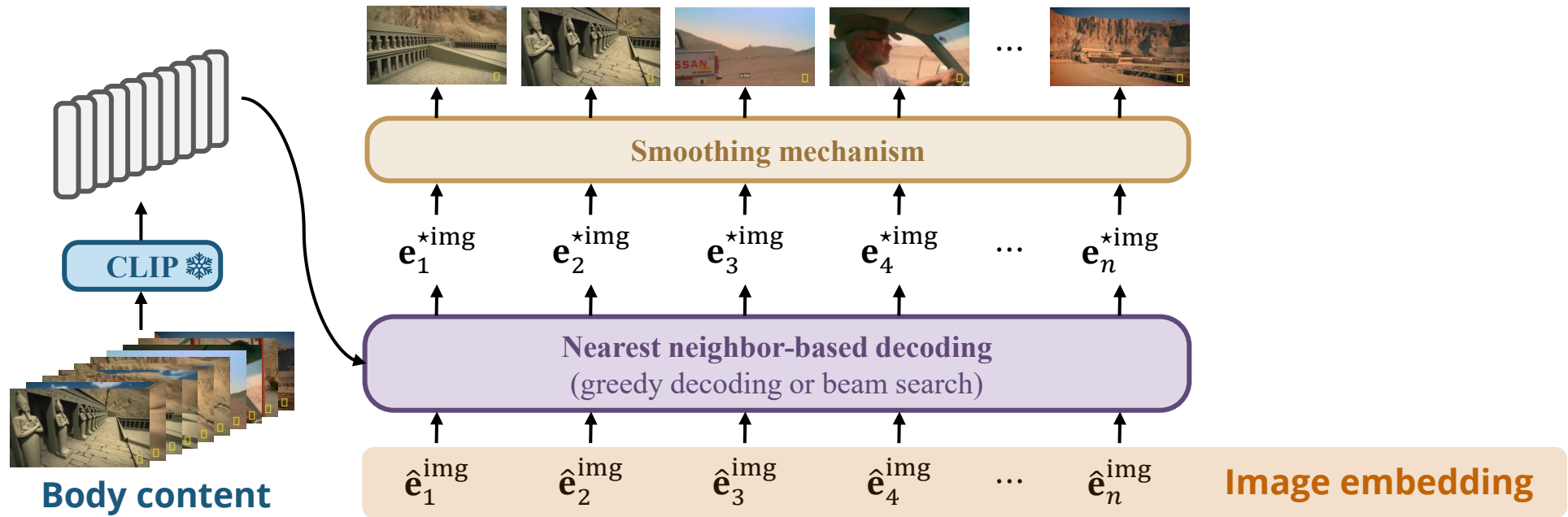
Interval-based Approach using Pretrained Models



Learning-based Approach using Deep Sequential Models
















Learning-based Approach using Deep Sequential Models













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 Results: TeaserGen-PT (Interval-based)

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



wx83.github.io/TeaserGen_Official/

Example Results: TeaserGen-LR (Learning-based)

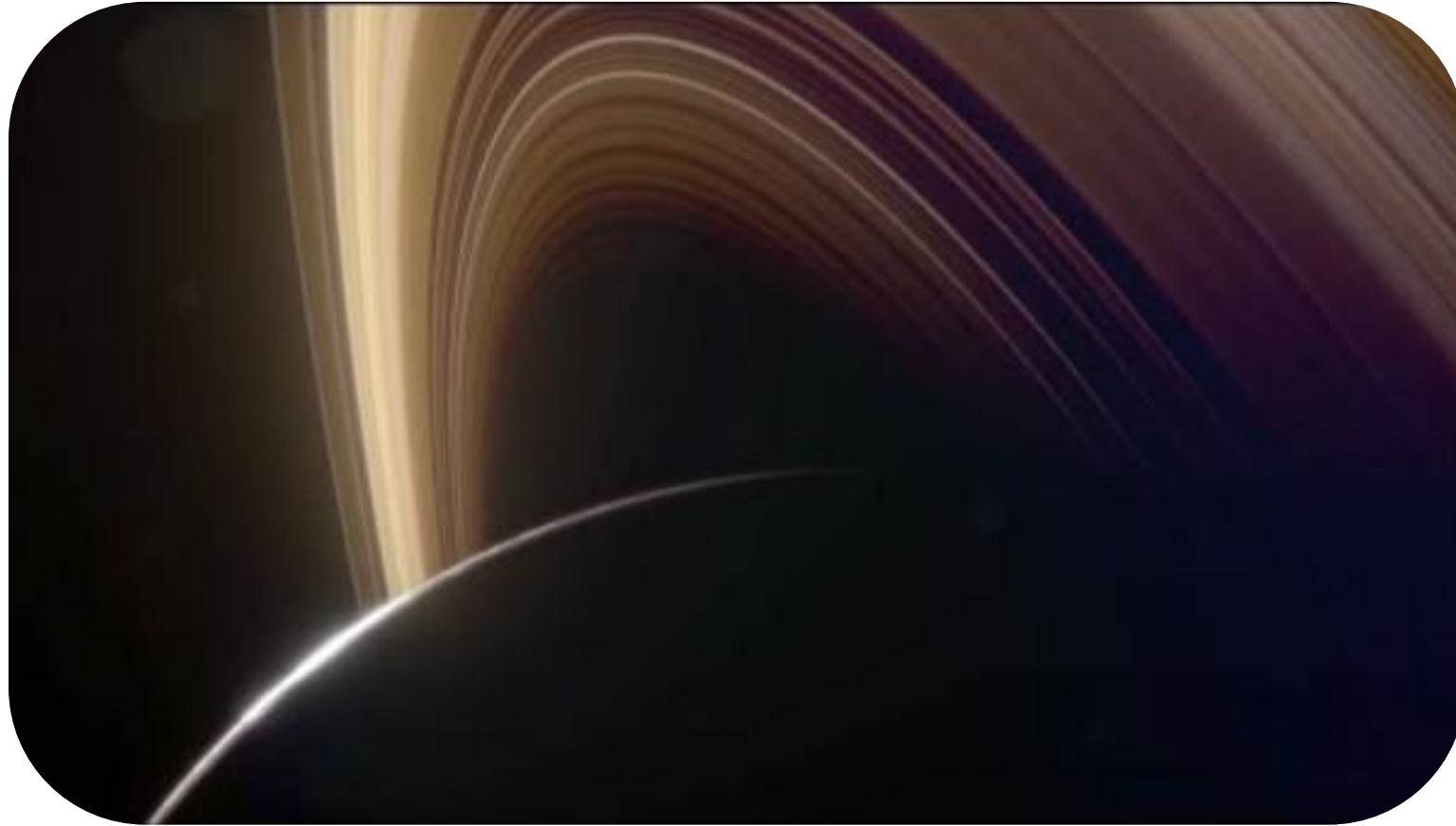
Title: "Hatshepsut: Mysteries of the Warrior Pharaoh Queen (Full Episode) | Lost Treasures of Egypt"



wx83.github.io/TeaserGen_Official/

Example: BBC Documentary on the Solar System

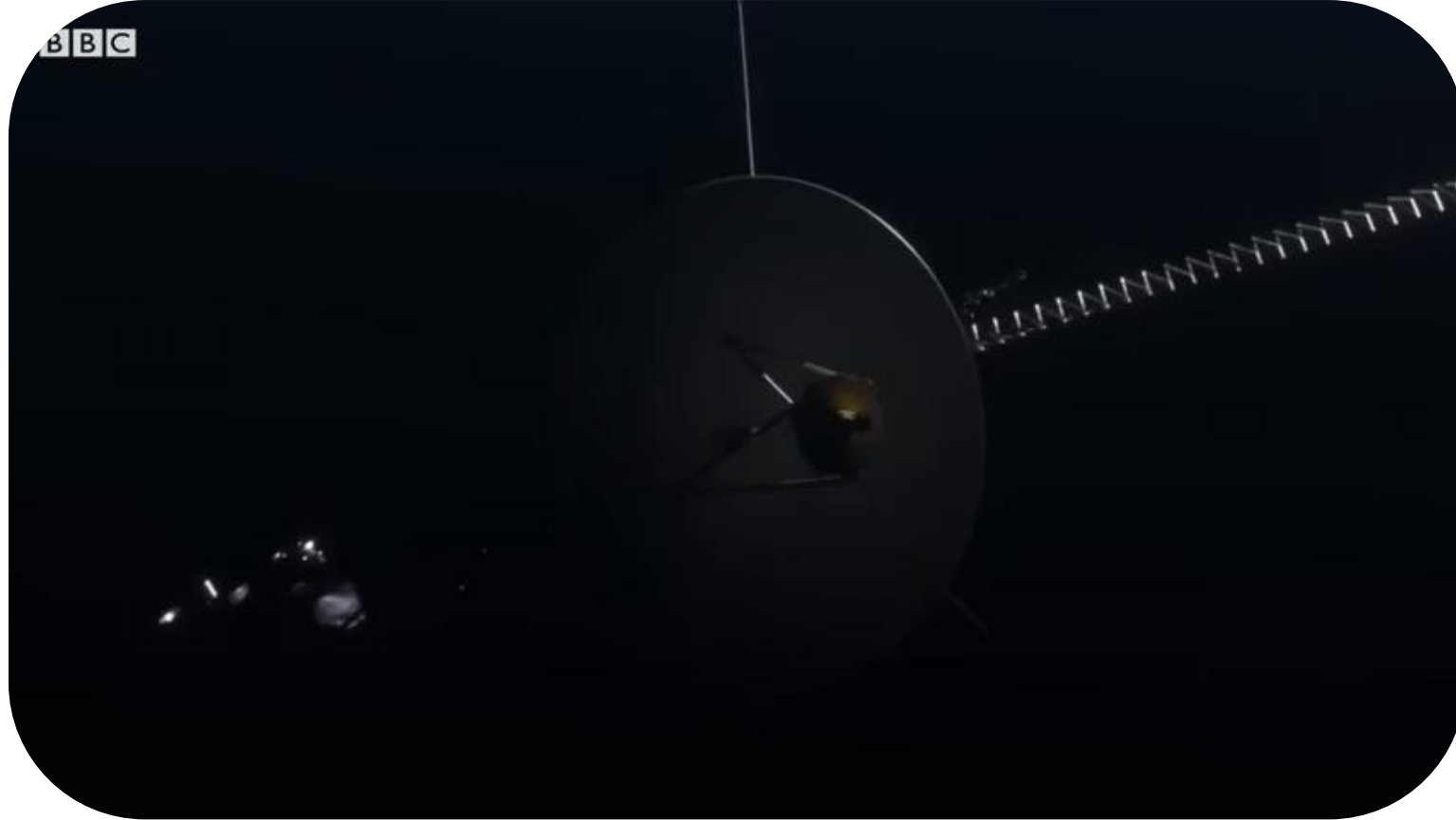
Title: "Eight Wonders Of Our Solar System | The Planets | BBC Earth Science"



youtu.be/wkQuOrsgVGy

Example Results: TeaserGen-PT (Interval-based)

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



wx83.github.io/TeaserGen_Official/

Example Results: TeaserGen-LR (Learning-based)

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



wx83.github.io/TeaserGen_Official/

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!

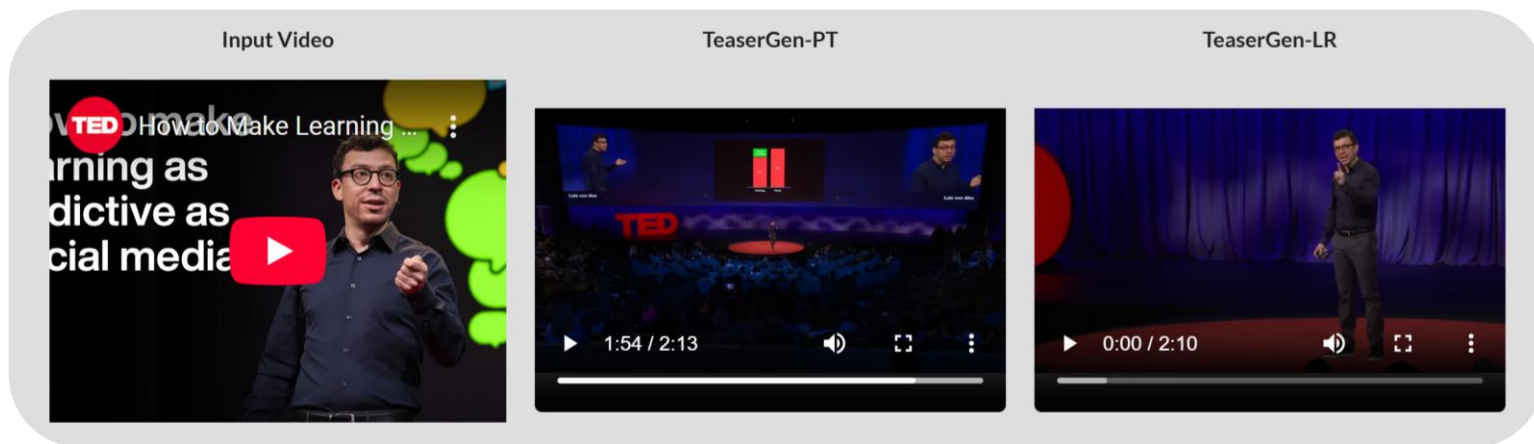
Subjective Evaluation

Model	Query	Decoding	Coherence \uparrow	Alignment \uparrow	Engagingness \uparrow	Realness \uparrow
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	3.14 ± 0.50	2.84 ± 0.57	2.81 ± 0.49	2.94 ± 0.50
TeaserGen-LR	Narration	Greedy	2.90 ± 0.45	2.88 ± 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)

Zero-shot Examples

TED Talks



Old movies



wx83.github.io/TeaserGen_Official/

Limitations

- Assumed that **narration plays a more significant role** than visuals
 - This assumption might not hold for movies and vlogs
- Cannot match **interview scenes** commonly seen in documentaries
 - Ongoing work to allow the model to “quote” an interview!
- 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



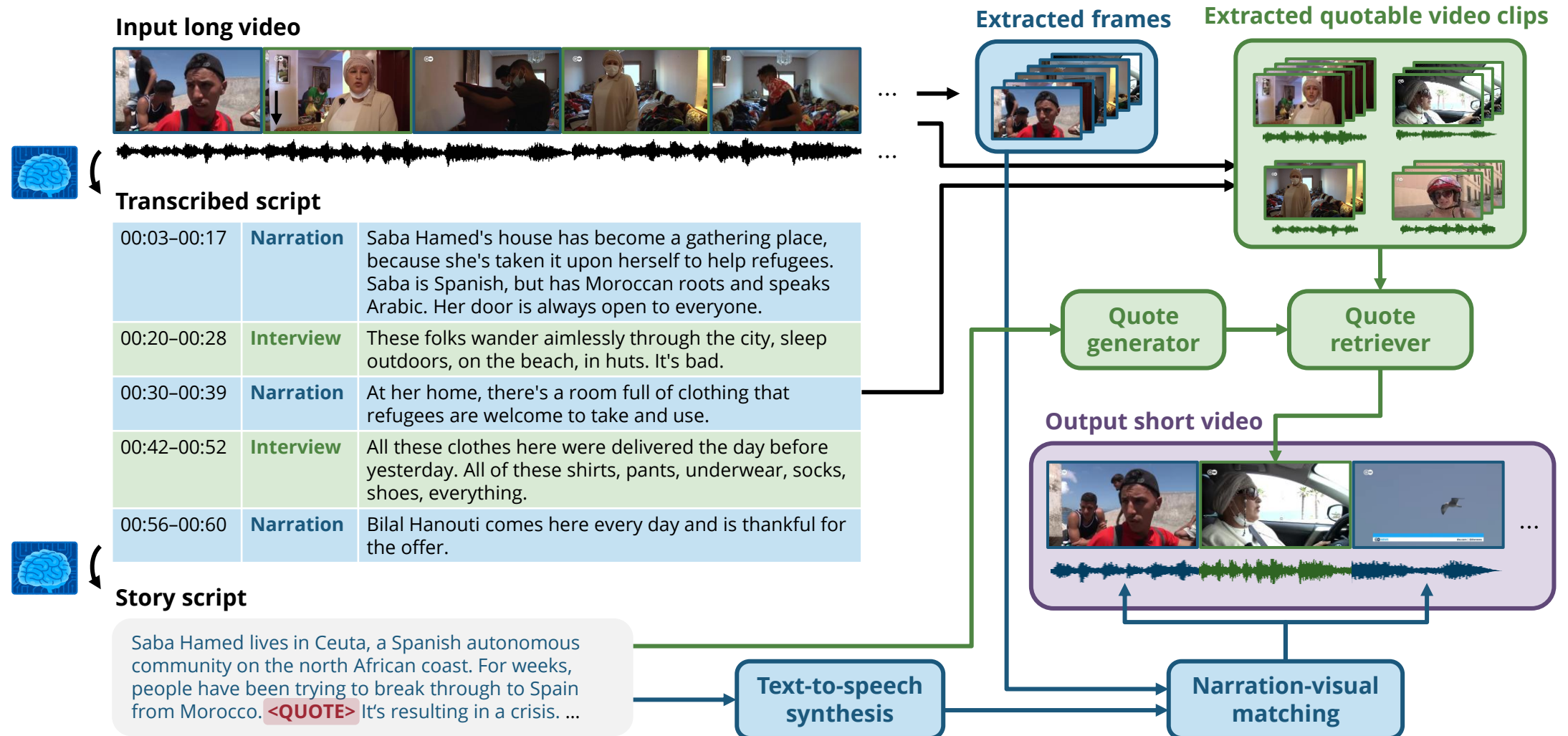
REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing

Wei-han Xu¹ Yimeng Ma¹ Jingyue Huang² Yang Li¹ Weyne Ma³
Taylor Berg-Kirkpatrick² Julian McAuley² Paul Pu Liang² **Hao-Wen Dong**⁴

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



Learning to *Quote* a Video



Learning to *Quote* a Video

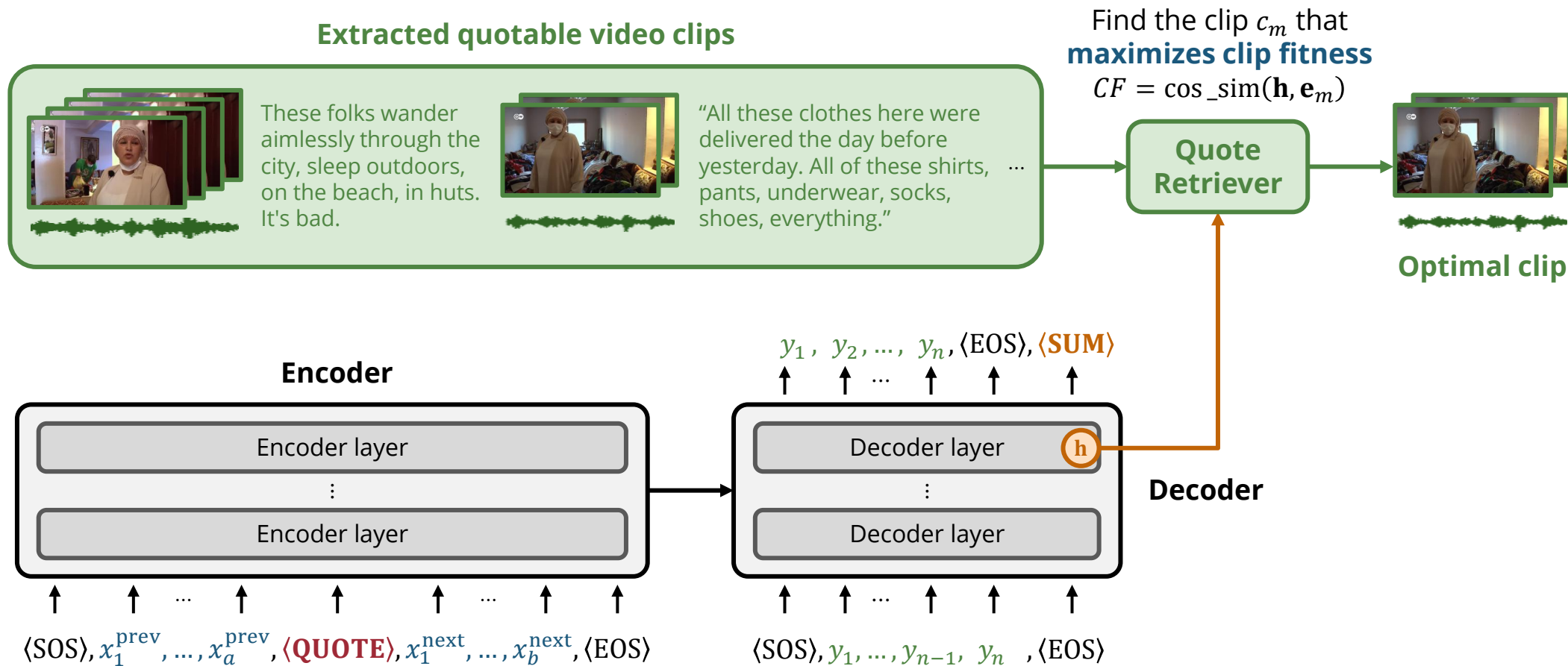
REGen-DQ
(direct quote)

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

REGen-IDQ
(indirect quote)

$\dots, x_i, \langle \text{QUOTE} \rangle, x_{i+1}, \dots$
↓
To be retrieved later!

Quote Retriever for REGen-IDQ



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)

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

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	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

Example: DW Documentary on a Modern Art Exhibition

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



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 Tremala, 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 Results: REGen-IDQ-TV (Indirect-quote, text+video)

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



wx83.github.io/REGen/

Example Results: REGen-IDQ-T (Indirect-quote, text-only)

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



wx83.github.io/REGen/

Example Results: REGen-DQ (Direct-quote)

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



wx83.github.io/REGen/

Example: National Geographic Documentary on Apocalypse

Title: "Apocalypse (Full Episode) | The Story of God with Morgan Freeman"



youtu.be/ATvKJ_HftNs

Example Results

REGen-IDQ-TV

Narrator: I'd like to know what people in all four faiths have to say about this image.

Narrator: So I've traveled to the Holy Land to compare and contrast the scriptural words with actual experience.

Quotable Interview Segments: Big question, isn't it? You have challenge. Yeah, maybe. I think the first part is maybe you need to recognize yourself, like where you come from."

Narrator: To learn about the origins of end-time beliefs, I'll visit with a scholar who's devoted his life to unlocking the Bible.

Quotable Interview Segments: Yes. It's the first Islamist organization that was responsible for popularizing the notion of constructing a modern-theocratic caliphate, as we now see that ISIS has laid claim to. But my former group, they were the first ones to popularize that term. I ended up in Egypt, where I continued to recruit people in this case. I was eventually arrested on the 1st of March in 2005. I was taken to the headquarters of the state security in Cairo, down underground in their torture dungeons. I was Misfakidat. And that's where the worst ordeal began. Torture? They began electrocuting everyone.

Narrator: To better understand the relevance of these prophecies to our day, I'll go to four different places of worship to hear from leaders who hear these words and words of violence directed at their communities.

Quotable Interview Segments: This is the Temple of the Moslems, and on the other side, the Temple of the Great Jagers. This would have been the very center of the ancient city of Tikal."

Narrator: This is the birthplace of three of the world's great faiths, all with end-time prophecies.

Narrator: What do Jews in Jerusalem say about the direction of the world and the coming of the messiah?

Quotable Interview Segments: Yeah. Now, there's a prophecy of the Prophet Muhammad that says that Constantinople will fall first, and then Rome will fall. So ISIS has interpreted this piece of scripture that because Constantinople has already fallen to Muslims, that the next big battle will be against the West, and the West would eventually fall. The idea would be that actually, in fact, that America today represents Rome. As you know, a continuation of Western civilization represented by the Roman Empire."

REGen-IDQ-T

Narrator: I'd like to know what people in all four faiths have to say about this image.

Narrator: So I've traveled to the Holy Land to compare and contrast the scriptural words with actual experience.

Quotable Interview Segments: Well, I've been looking at some fragments of the books of Revelation.

Narrator: To learn about the origins of end-time beliefs, I'll visit with a scholar who's devoted his life to unlocking the Bible.

Quotable Interview Segments: The end of days, the apocalypse. It's a prophetic book. It's got loads of symbolism. But it's also very much a political book and making a political claim about the cause of evil."

Narrator: To better understand the relevance of these prophecies to our day, I'll go to four different places of worship to hear from leaders who have these words and words of violence directed at their communities.

Quotable Interview Segments: So there are a lot of prophecies that most Muslims share in common with each other. The difference is what ISIS has done is it's manipulated those prophecies to serve its own political and ideological ends. So there's an example of this end of times battle that ISIS believes is going to take place in a small village called Dabiq in Syria. Now this village has absolutely no strategic value militarily whatsoever. Has hardly any economic, strategic value either. But ISIS has nevertheless committed resources to conquering this small village called Dabiq. They believe that the international community and the coalition must somehow be driven to come and meet them in Dabiq and engage in a final battle.

Narrator: This is the birthplace of three of the world's great faiths, all with end-time prophecies.

Narrator: What do Jews in Jerusalem say about the direction of the world and the coming of the messiah?

Quotable Interview Segments: According to scripture, he has three things that he's going to do. Number one, he's going to rebuild the Jewish Temple in Jerusalem. Number two, he's going to bring peace with his neighbors. And number three, he's going to rebuild that temple. Rebuild the temple.

REGen-DQ

Narrator: I'm going on a journey to find out why so many people are expecting the end of the world as they know it.

Quotable Interview Segments: The end of days, the apocalypse. It's a prophetic book. It's got loads of symbolism. But it's also very much a political book and making a political claim about the cause of evil."

Narrator: And I'm not the only one obsessed with the idea of the world coming to an end.

Narrator: I have yet to meet a 21st century person who yet has not imagined life without consciousness.

Quotable Interview Segments: It was really a dark time. It was really dark.

Quotable Interview Segments: All right. We just knew that God would get us out of there.

Quotable Interview Segments: They think of, you know, the end of days. What we have for the messiah is a man, a king of this earth, who's going to bring peace among the nations in this world.

Quotable Interview Segments: So there are a lot of prophecies that most Muslims share in common with each other. The difference is what ISIS has done is it's manipulated those prophecies to serve its own political and ideological ends. So there's an example of this end of times battle that ISIS believes is going to take place in a small village called Dabiq in Syria. Now this village has absolutely no strategic value militarily whatsoever. Has hardly any economic strategic value either. But ISIS has nevertheless committed resources to conquering this small village called Dabiq. They believe that the international community and the coalition must somehow be driven to come and meet them in Dabiq and engage in a final battle.

Quotable Interview Segments: To understand why I've come to New York

Example Results: REGen-IDQ-TV (Indirect-quote, text+video)

Title: "Apocalypse (Full Episode) | The Story of God with Morgan Freeman"



wx83.github.io/REGen/

Example Results: REGen-IDQ-T (Indirect-quote, text-only)

Title: "Apocalypse (Full Episode) | The Story of God with Morgan Freeman"



wx83.github.io/REGen/

Example Results: REGen-DQ (Direct-quote)

Title: "Apocalypse (Full Episode) | The Story of God with Morgan Freeman"



wx83.github.io/REGen/

Objective Evaluation

Model	Dur (sec)	Interview ratio (%)	F1 (%)	SCR (%)	REP (%)	Repetitiveness		
						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!

Subjective Evaluation

Model	Coherence \uparrow	Alignment \uparrow	Realness \uparrow	Interview effectiveness \uparrow
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	3.33 ± 0.29
REGen-IDQ-TV	3.29 ± 0.24	3.30 ± 0.26	3.05 ± 0.25	3.25 ± 0.30

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

Limitations

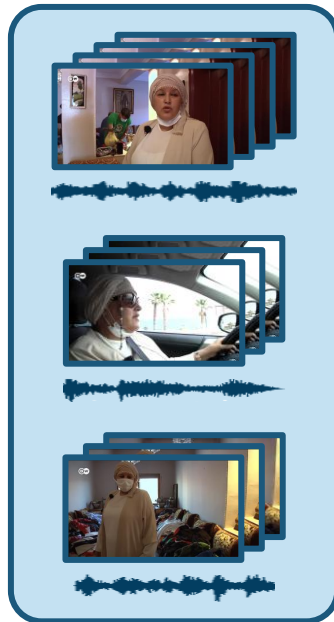
- Assumed that **narration plays a more significant role** than visuals
 - This assumption might not hold for movies and vlogs
- Risks of **misplacing a quote in a wrong context**
 - Grounding the script generation model with information about all quotable materials
 - May also be alleviated by context-aware 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

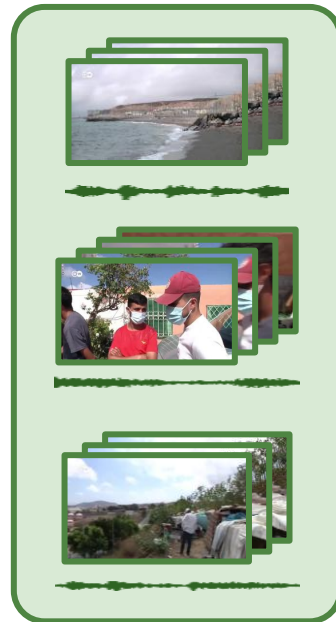
Video Editing



Interview footage
(main character)



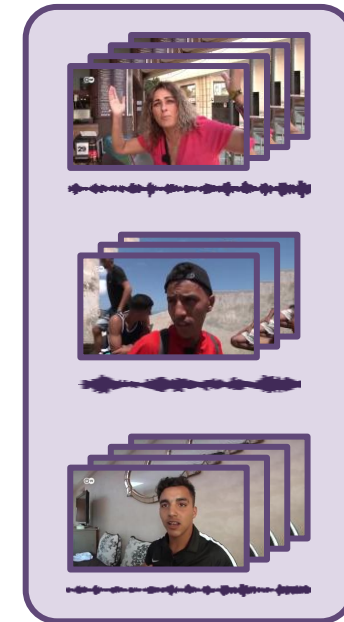
Background footage



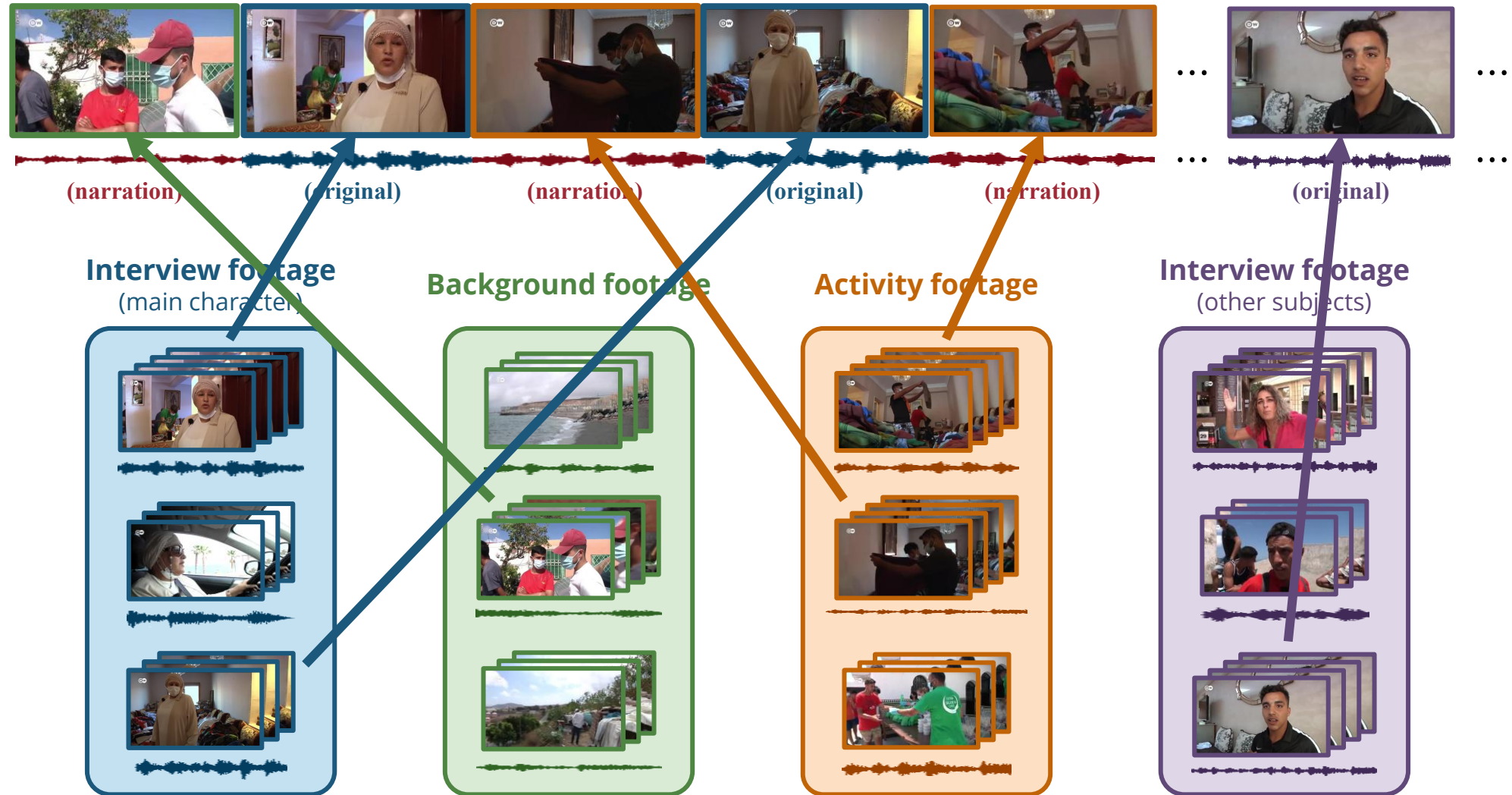
Activity footage



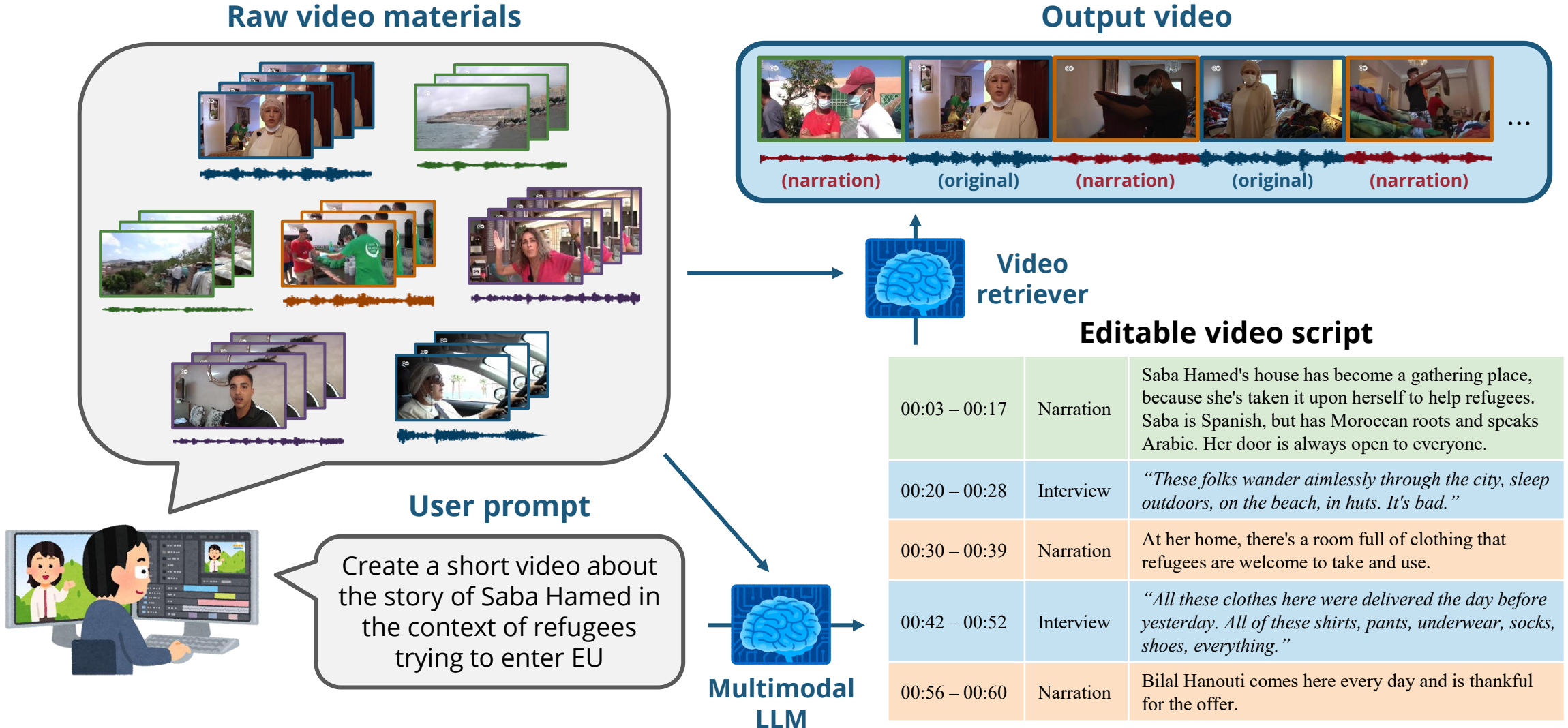
Interview footage
(other subjects)



Video Editing

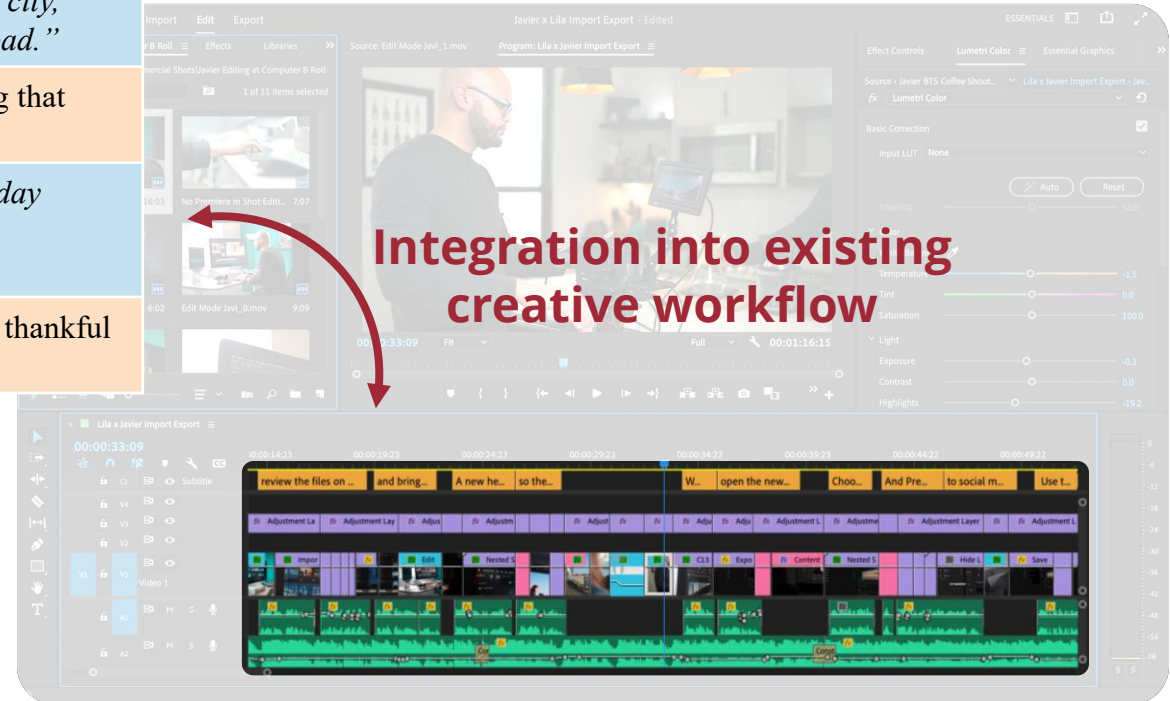
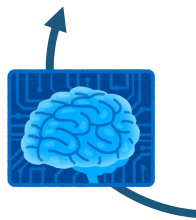


Multimodal RAG-based Video Editing



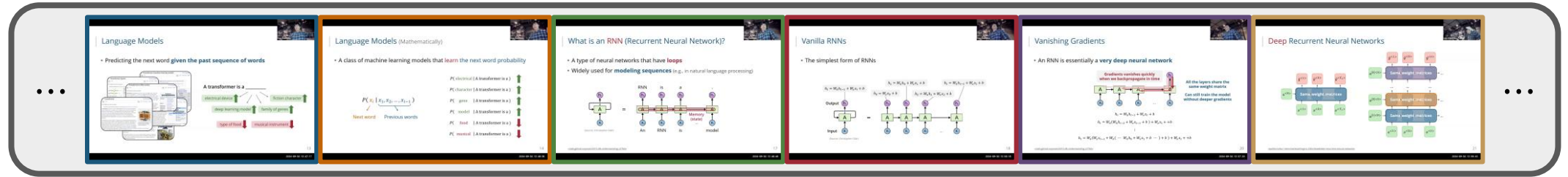
Future Work: Integration into Video Editing Software

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.



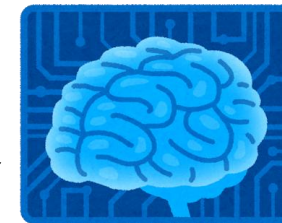
Future Work: LectureRecap

Lecture recording

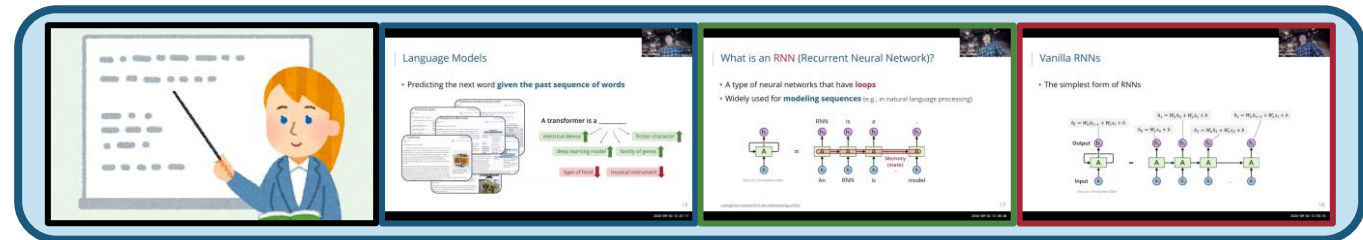


I would like to review the concept of recurrent neural networks. How does an RNN work?

Can you explain the math behind it?



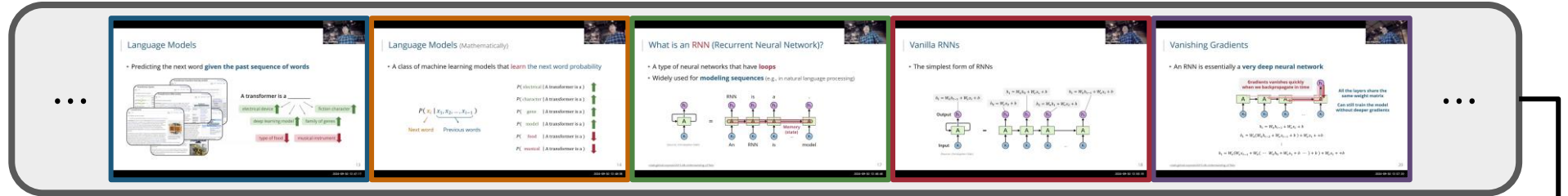
LectureRecap



Lecture recap

Future Work: LectureRecap

Lecture recording



User query

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

Script generation

Speech recognition

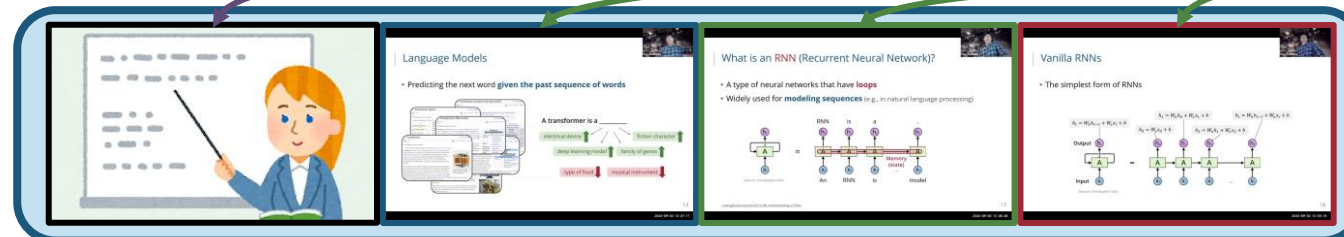
Video script

[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

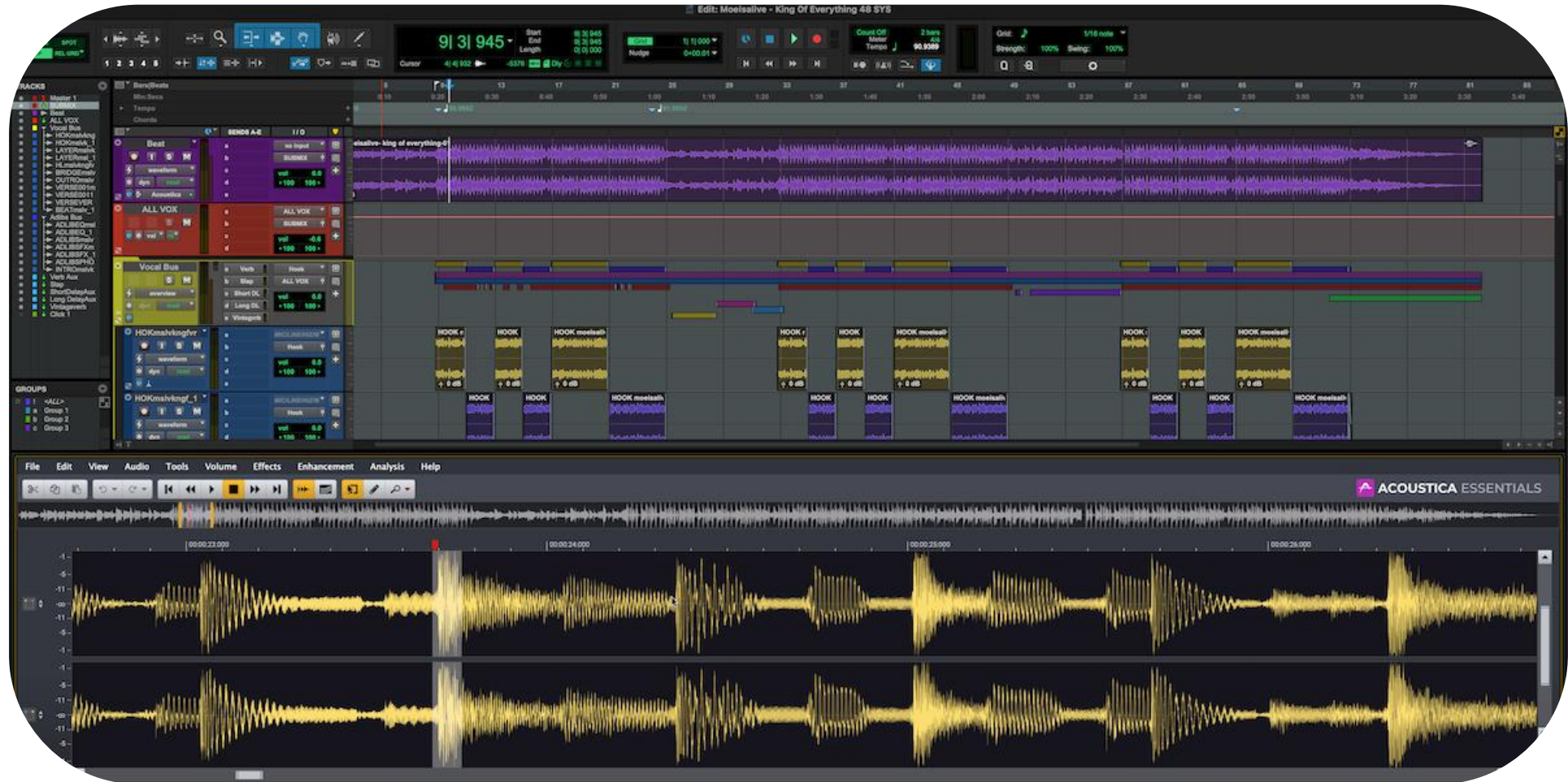


Retrieval-Augmented → Retrieval-Embedded Generation

- Can an LLM learn to quote and **embed the quote properly**?
- How to **quote materials in other modalities**?
 - Such as audio, image, videos, 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

How About Music?

Integrating GenAI into the Music Creative Workflow



(Source: Avid)

| Integrating GenAI into the Music Creative Workflow



(Source: Avid)

Augmenting Human Creativity with AI

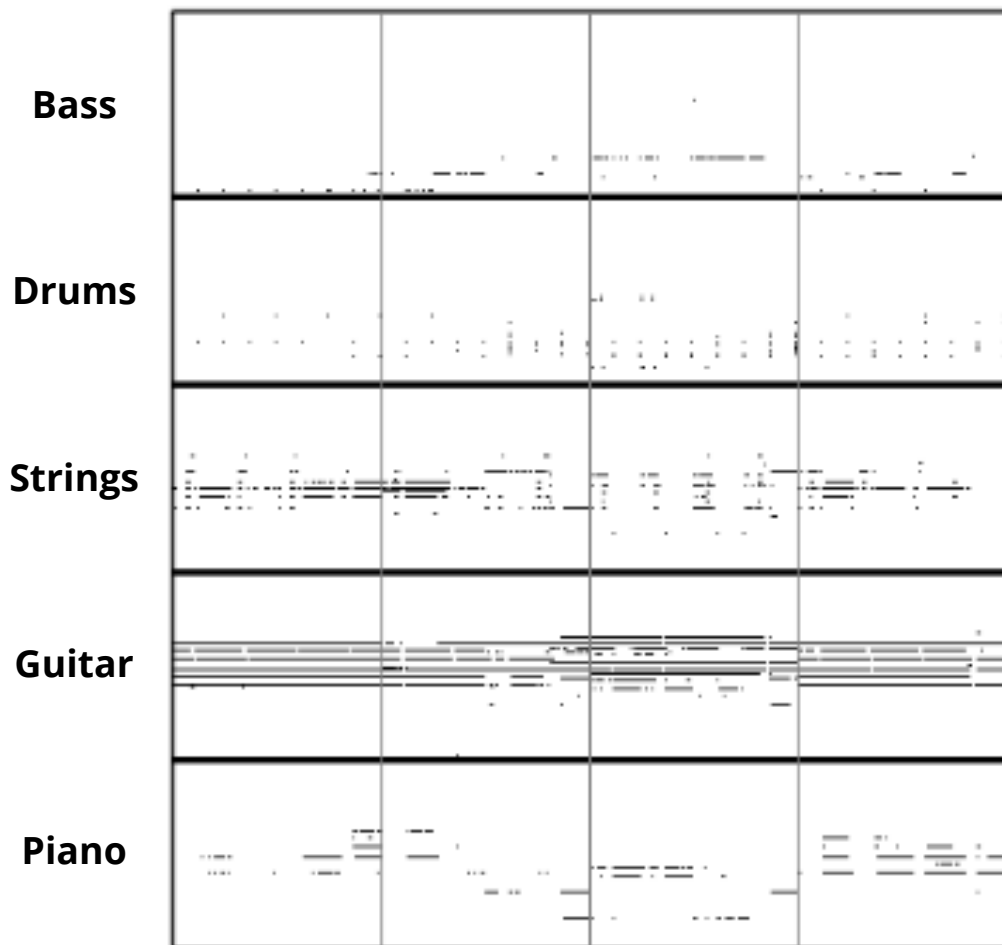
- **Novel Generative Models for New Domains**
 - **Multitrack music generation** (AAAI 2018, ISMIR 2018, ISMIR 2020, ICASSP 2023, ISMIR 2024, AIMG 2024), **text-to-symbolic music generation** (ISMIR LBD 2024, ISMIR 2025)
- **AI-assisted Tools for Content Creation**
 - **Violin performance synthesis** (ICASSP 2022, ICASSP 2025), **music instrumentation** (ISMIR 2021), **music arrangement** (AAAI 2018), **music harmonization** (JNMR 2020), **a capella source separation** (ISMIR LBD 2025)
- **Multimodal Generative Models for Content Creation**
 - **Queried sound separation** (ICLR 2023), **text-to-audio synthesis** (WSS 2023, WASPAA 2023), **text-to-music generation** (ISMIR LBD 2024, arXiv 2024), **video-to-music generation** (ISMIR 2025, ISMIR LBD 2025), **long-to-short video editing** (ICLR 2025, arXiv 2025)

| My Research on AI for Music

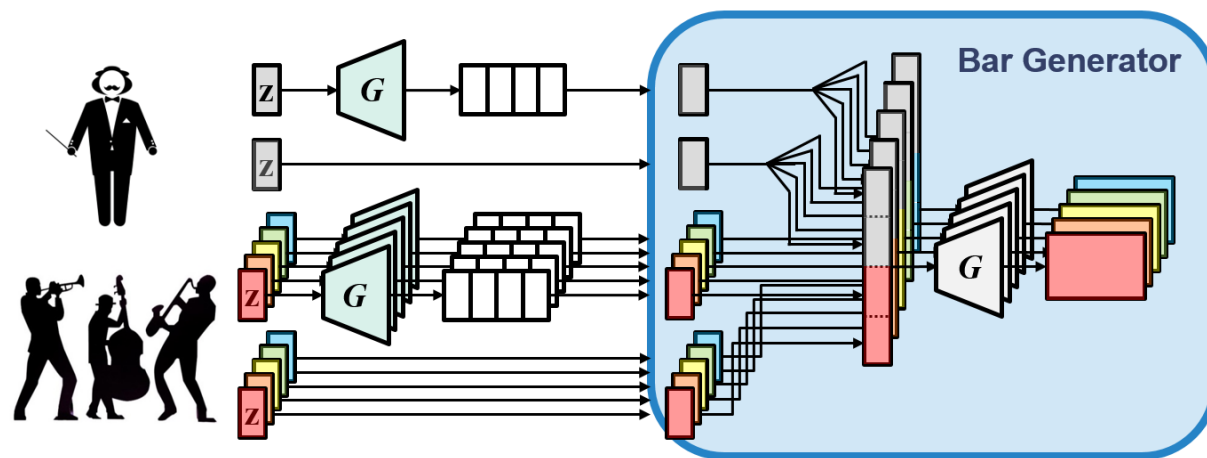
- **Multitrack music generation** (ISMIR 2024 , ICASSP 2023, ISMIR 2020 , ISMIR 2018, AAAI 2018, ISMIR LBD 2017)
- **Expressive violin performance synthesis** (ICASSP 2025, ICASSP 2022)
- **Text-to-music generation** (ISMIR 2025, AIMG 2024, ISMIR LBD 2024)
- **Video-to-music generation** (ISMIR 2025, ISMIR LBD 2025)
- **Music arrangement** (ISMIR 2021, JNMR 2020)
- **Music LLM** (ICASSP 2025, NLP4MusA 2024)
- **Choral music separation** (ISMIR LBD 2025, ISMIR 2022)
- **Optical music recognition** (ISMIR 2021)

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 screenshot shows the AWS DeepComposer 'Train a model' interface. The 'Generative algorithm' section has two options: 'MuseGAN' (selected) and 'U-Net'. The 'MuseGAN' option is highlighted with a red circle and a note that it is a GAN algorithm often used for complex music structures. The 'U-Net' option is also shown with a note that it is the distinguishing feature. Below the algorithm selection, there are two diagrams: one showing a sequence of layers with 512, 256, and 128 units, and another showing a U-Net architecture with an input noise layer. The background of the interface features a large 32-key, 2-octave keyboard and a video of a man playing the keyboard.

amazon.com/dp/B07YGZ4V5B/

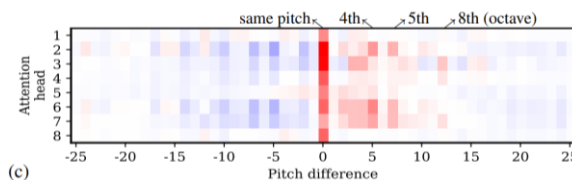
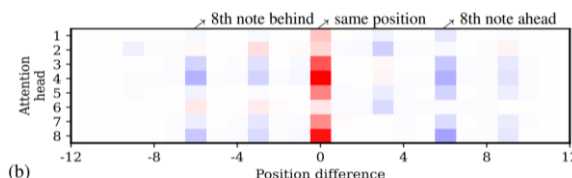
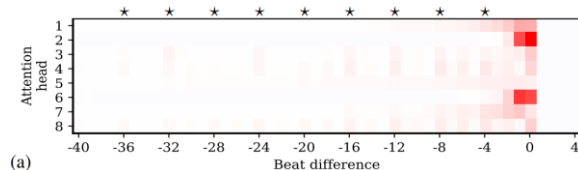
Julien Simon, "AWS DeepComposer – Now Generally Available With New Features," AWS News Blog, April 2, 2020.

Generating Multitrack Music with Transformers (ICASSP 2023)

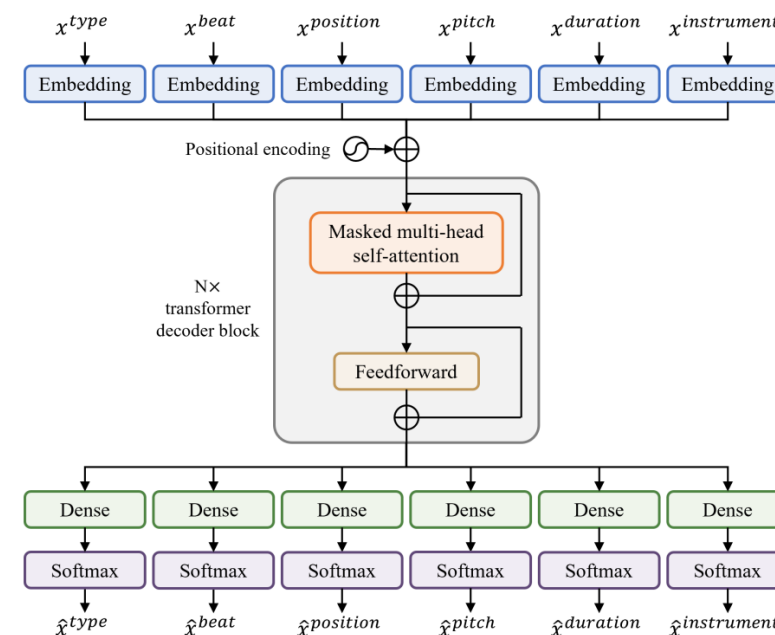
Multitrack Music Representation

(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

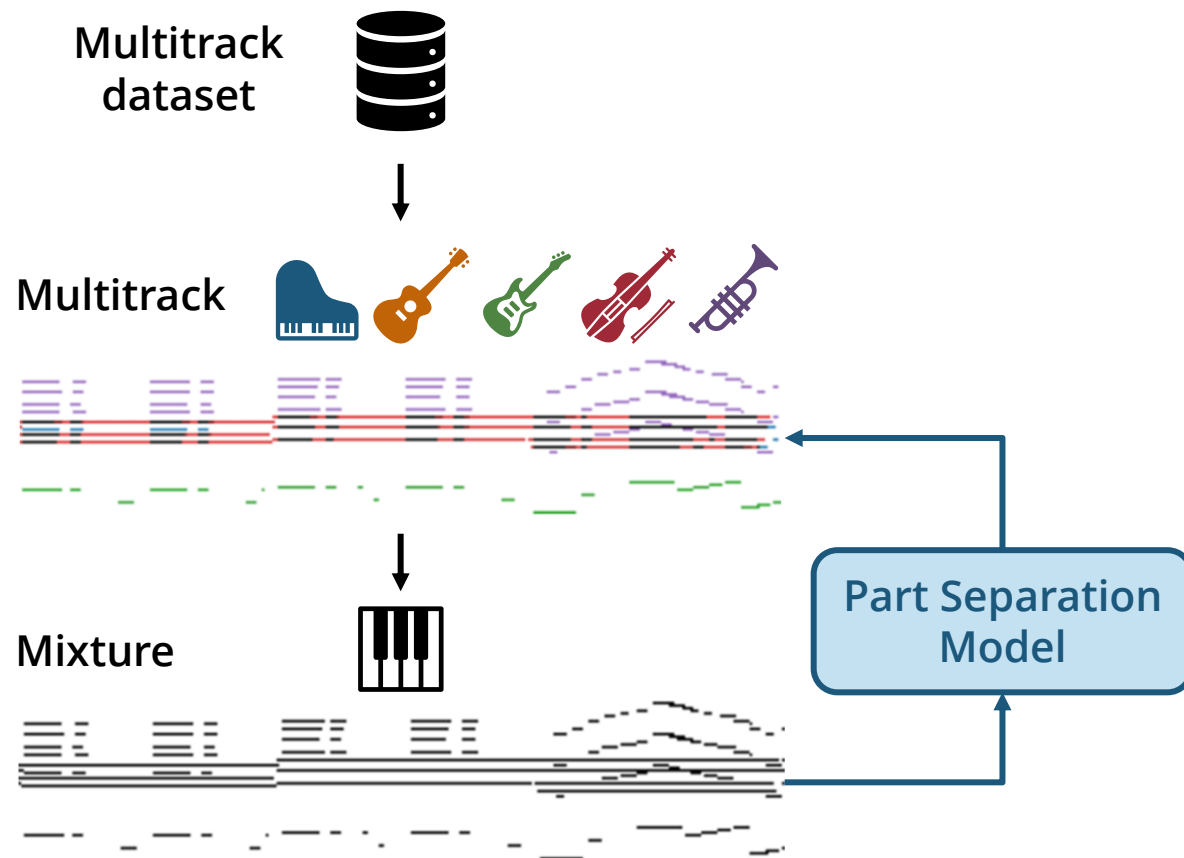
Automatic Instrumentation (ISMIR 2021)



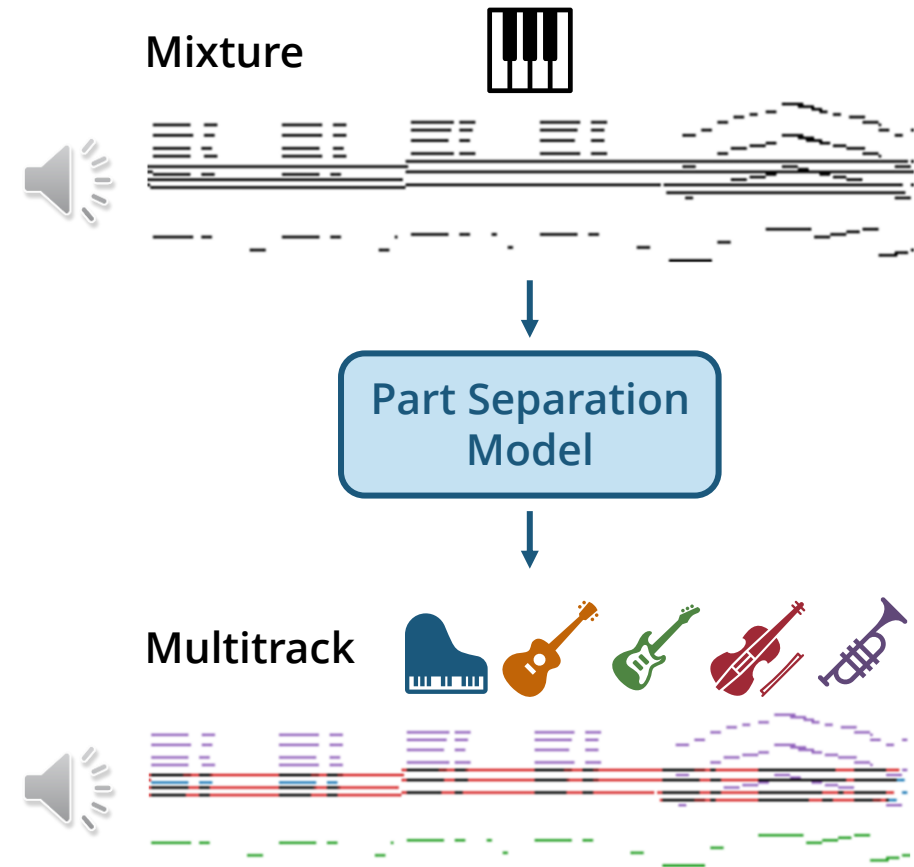
Stanford

UC San Diego

Training

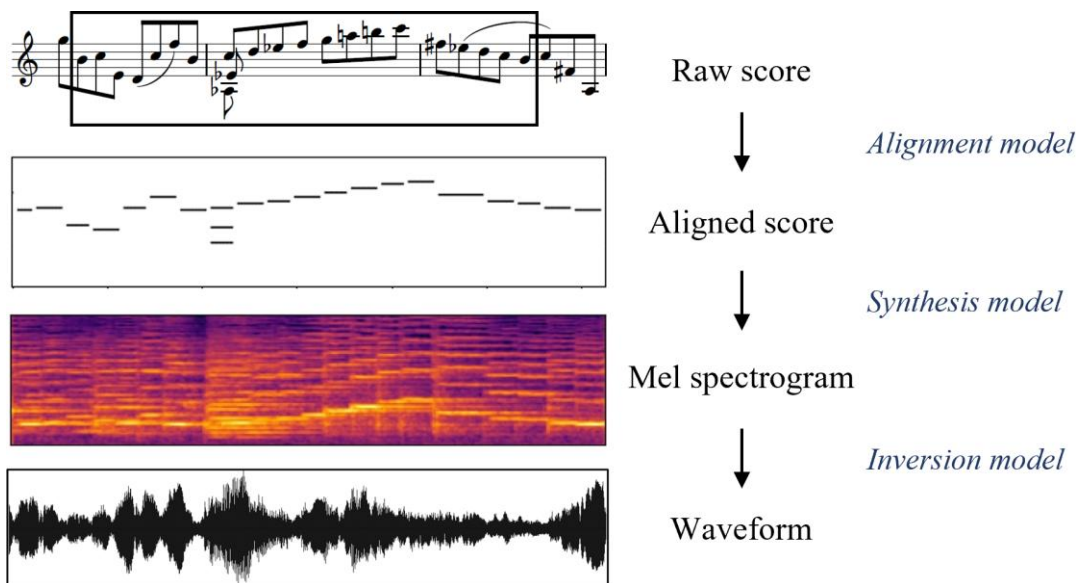


Inference

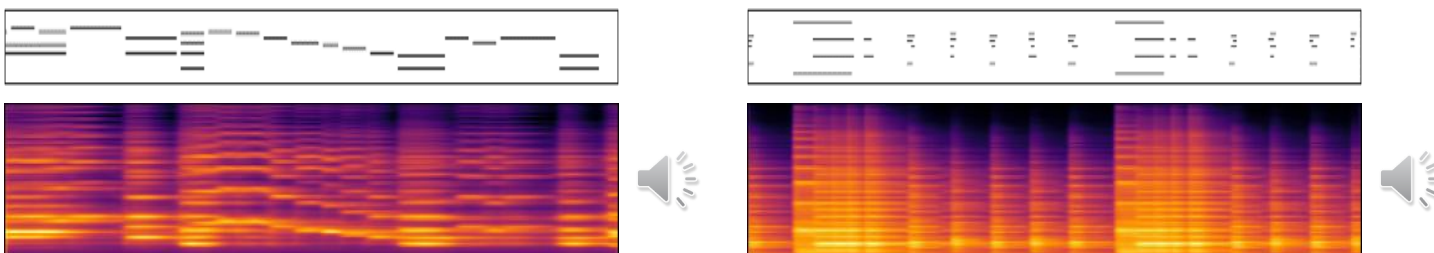


Synthesizing Expressive Violin Performance (ICASSP 2022)

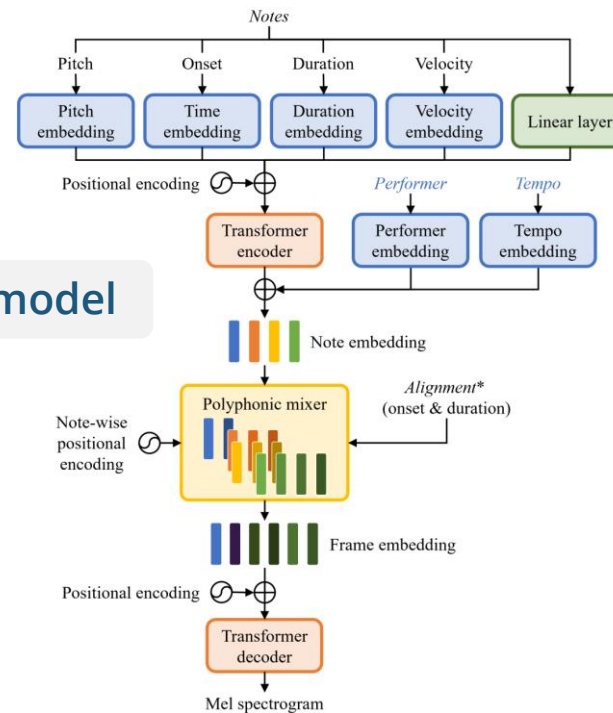
Performance synthesis



Example results



TTS-based model



Dolby

UC San Diego

Music & Technology Co-evolves



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

Art challenges Technology



Music

**Augmenting Human Creativity
with AI**



AI



Technology inspires the Art

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

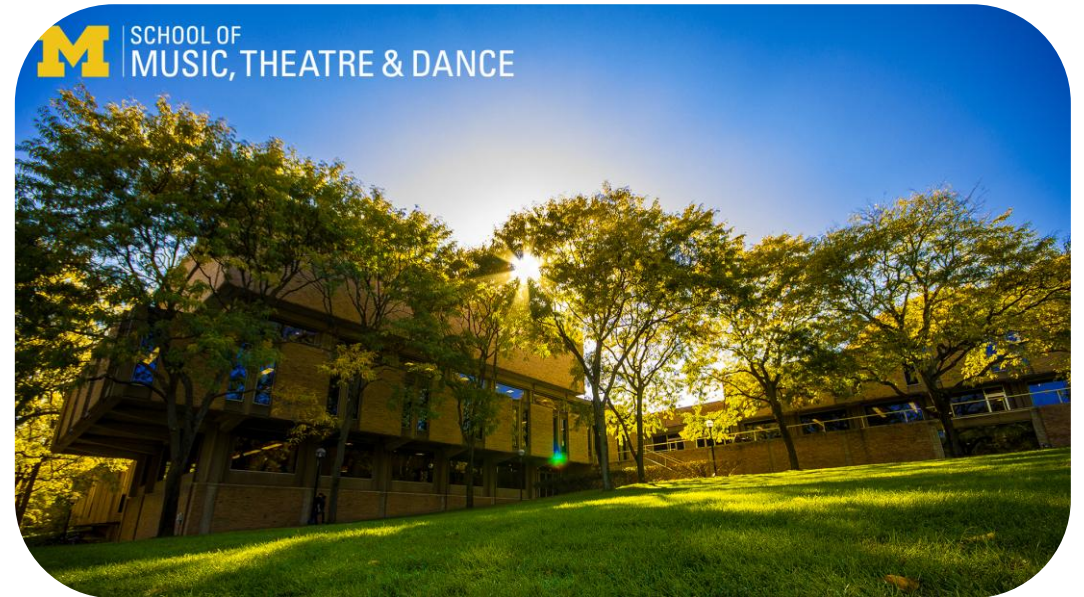
AI Music @ Michigan



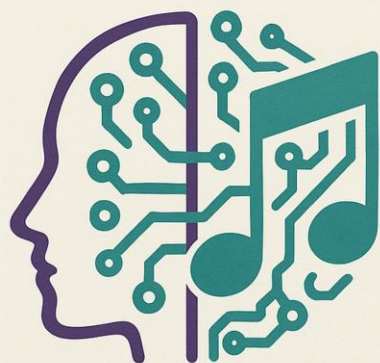
Hao-Wen Dong



Julie Zhu



AI4Music Workshop @ NeurIPS

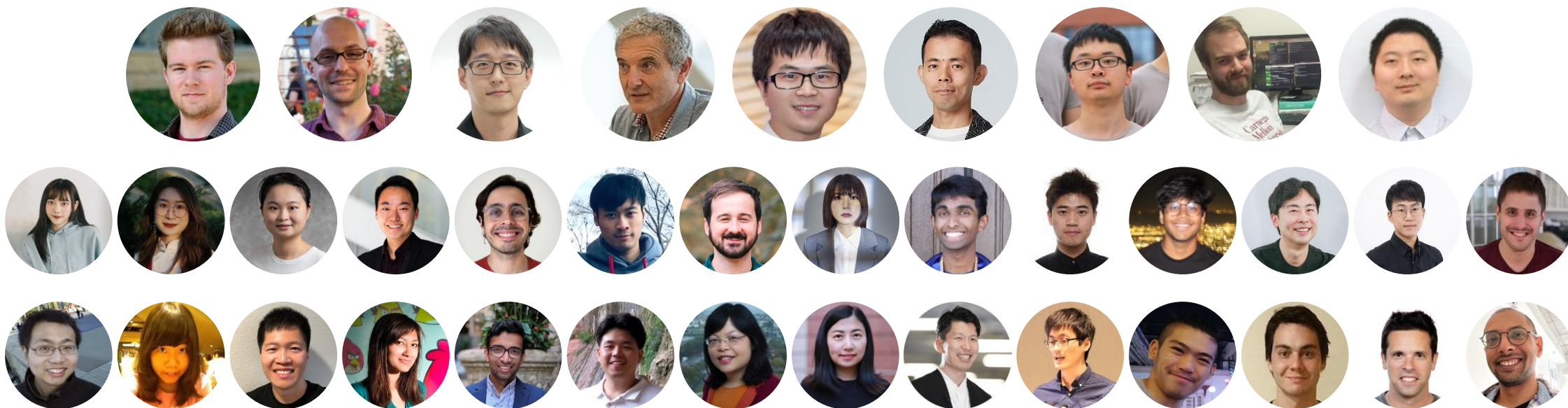


NEURIPS 2025 WORKSHOP ON
AI FOR MUSIC
CREATIVITY MEETS COMPUTATION

December 7 @ San Diego
aiformusicworkshop.github.io

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

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 / hwdong@umich.edu

M UNIVERSITY OF MICHIGAN