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] Massachusetts Institute of Technology
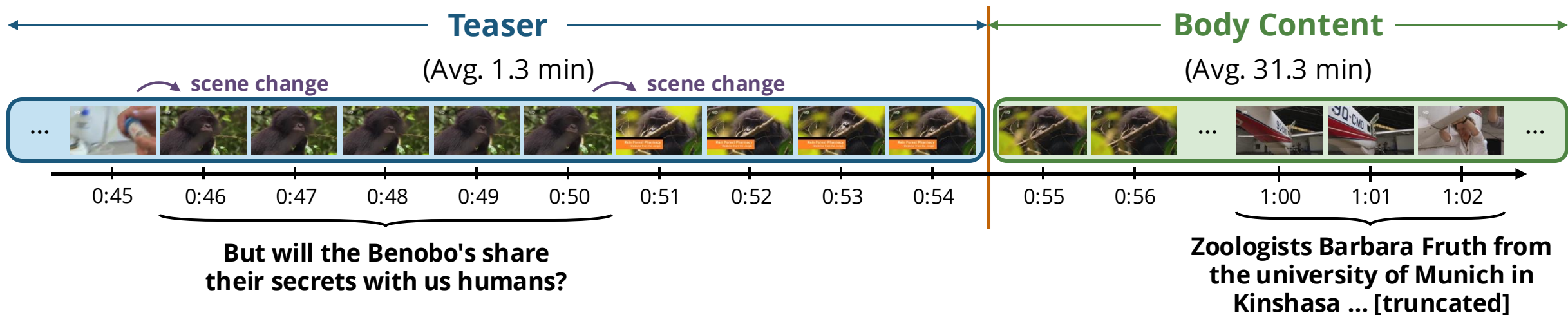[3] University of California San Diego   [4] University of Michigan

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

Teaser
(Avg. 1.3 min)

scene change   scene change

Body Content
(Avg. 31.3 min)

0:45   0:46   0:47   0:48   0:49   0:50   0:51   0:52   0:53   0:54   0:55   0:56   1:00   1:01   1:02

**But will the Benobo's share
their secrets with us humans?**

**Zoologists Barbara Fruth from
the university of Munich in
Kinshasa ... [truncated]**

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.
...

3

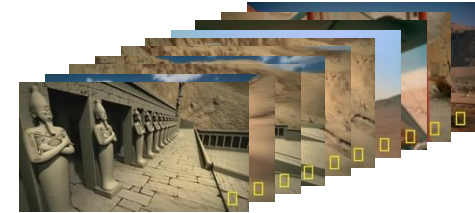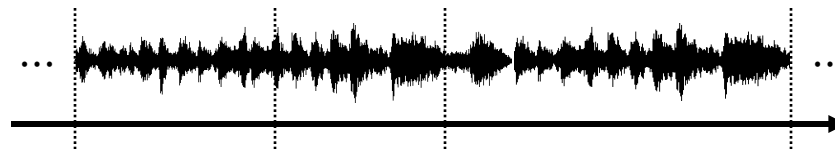# A Narration-Centered Approach to Teaser Generation

Body content
(Input)

Speech recognition
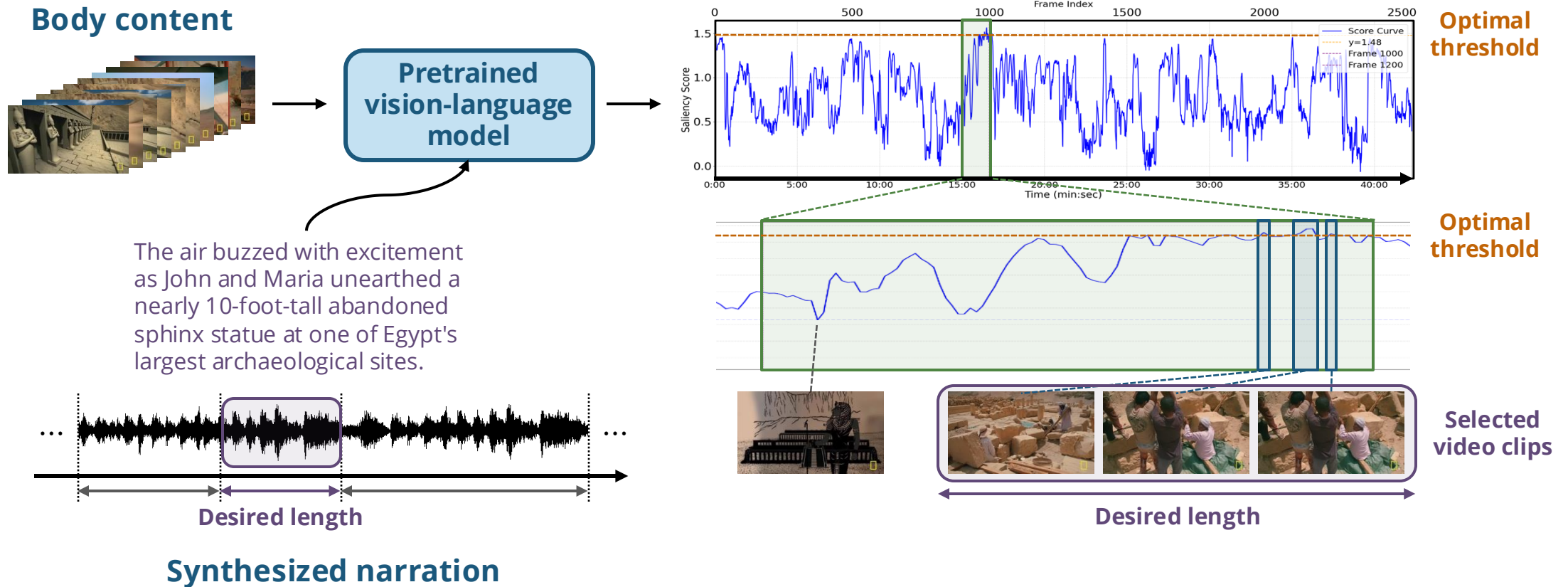
Narration generation

Text-to-speech synthesis

Narration-video matching

Teaser
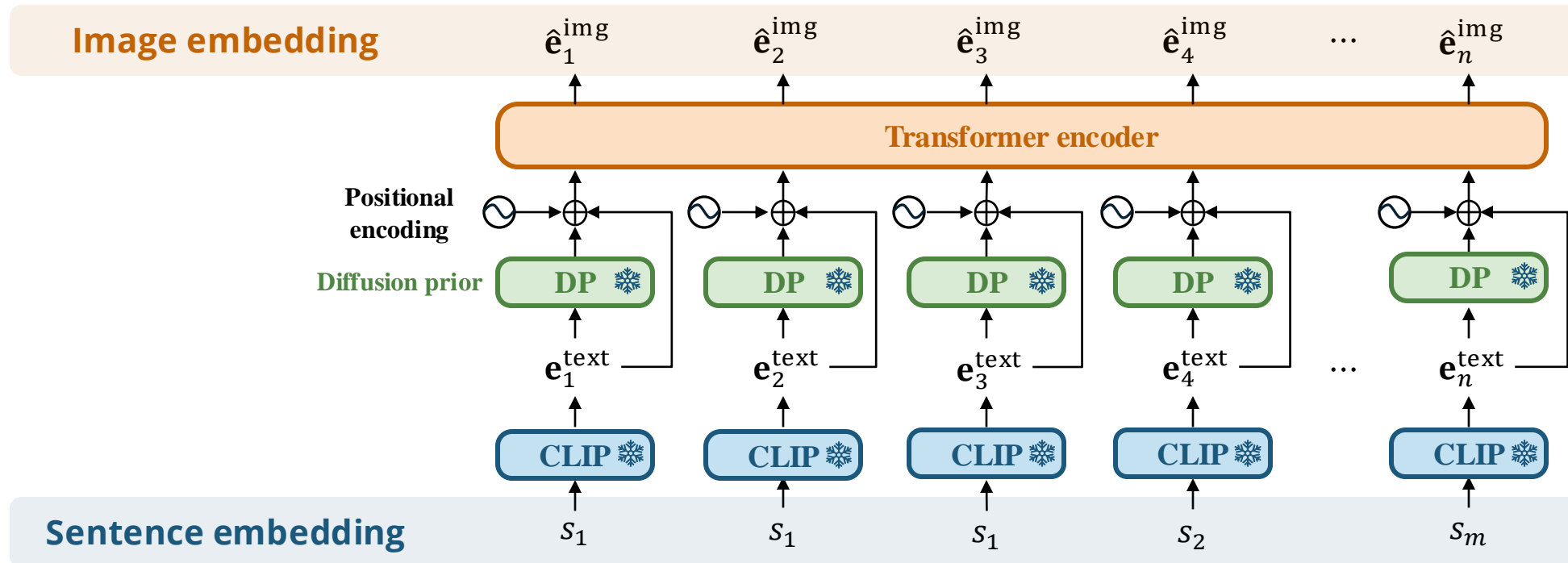**(Output)**

...   ...

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?*"

Body content

Pretrained vision-language model

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

Optimal threshold

Selected video clips

Desired length

Learning-based Approach using Deep Sequential Models

Learning-based Approach using Deep Sequential Models
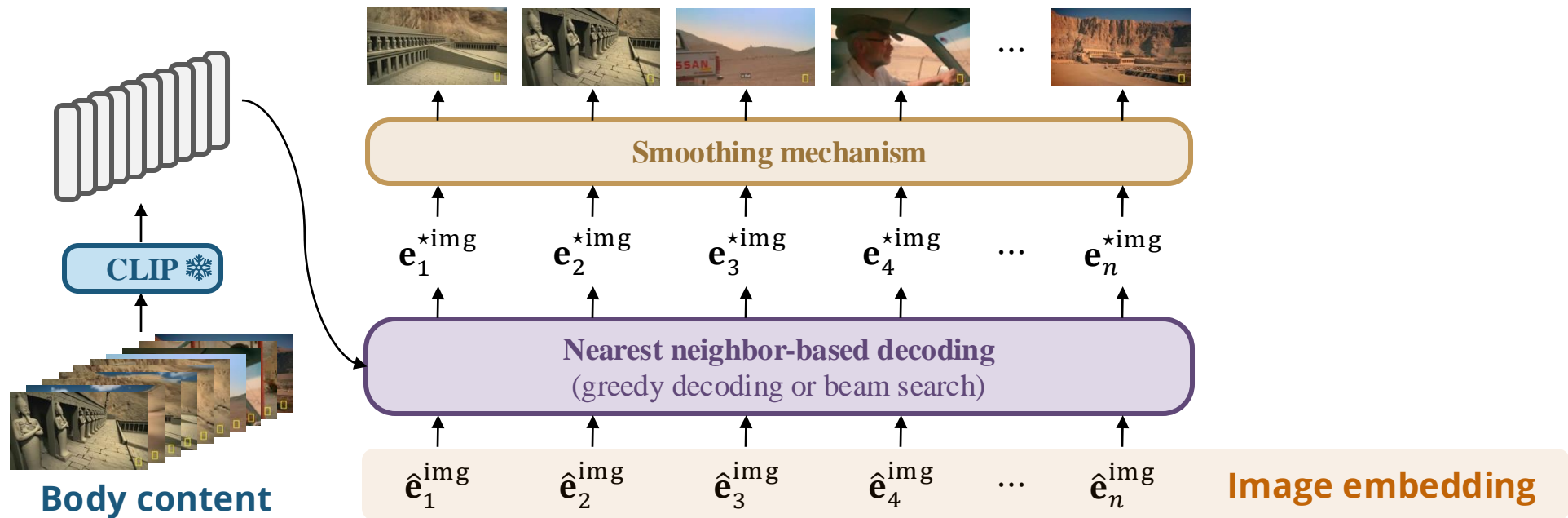
Demo

Interval-based

Learning-based

wx83.github.io/TeaserGen_Official/

Evaluation Metrics

- **Objective Evaluation**
  - Retrieval-based metrics
  - Repetitiveness
  - Scene change rate (SCR)
  - CLIPScore
  - VTGHLS

- **Subjective Evaluation**
  - Generated Narration
    - Organization
    - Informativeness
    - Engagingness
  - Generated Teaser
    - Coherence
    - Alignment
    - Engagingness
    - Realness

Objective Evaluation

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

Subjective Test

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

Our Contribution

- We **propose DocumentaryNet**, a **publicly-available dataset** consisting of 1,269 high-quality documentaries paired with their teasers.
- We propose **TeaserGen**, a narration-centered **teaser generation system** that can effectively compress >30-min documentaries into <3-min teasers
- We design **new evaluation metrics** to better evaluate the new task of teaser generation, where we conduct human evaluations **to validate these evaluation metrics**
- Our proposed method shows **generalizability across old movies and educational videos.**