

Overview

Teasers are an effective tool for promoting content in entertainment, commercial and educational fields.

Challenges

- **Input:** Require long-range multimodal modeling
- **Output:** Maintain audiovisual alignments; manage scene changes; preserve factual accuracy

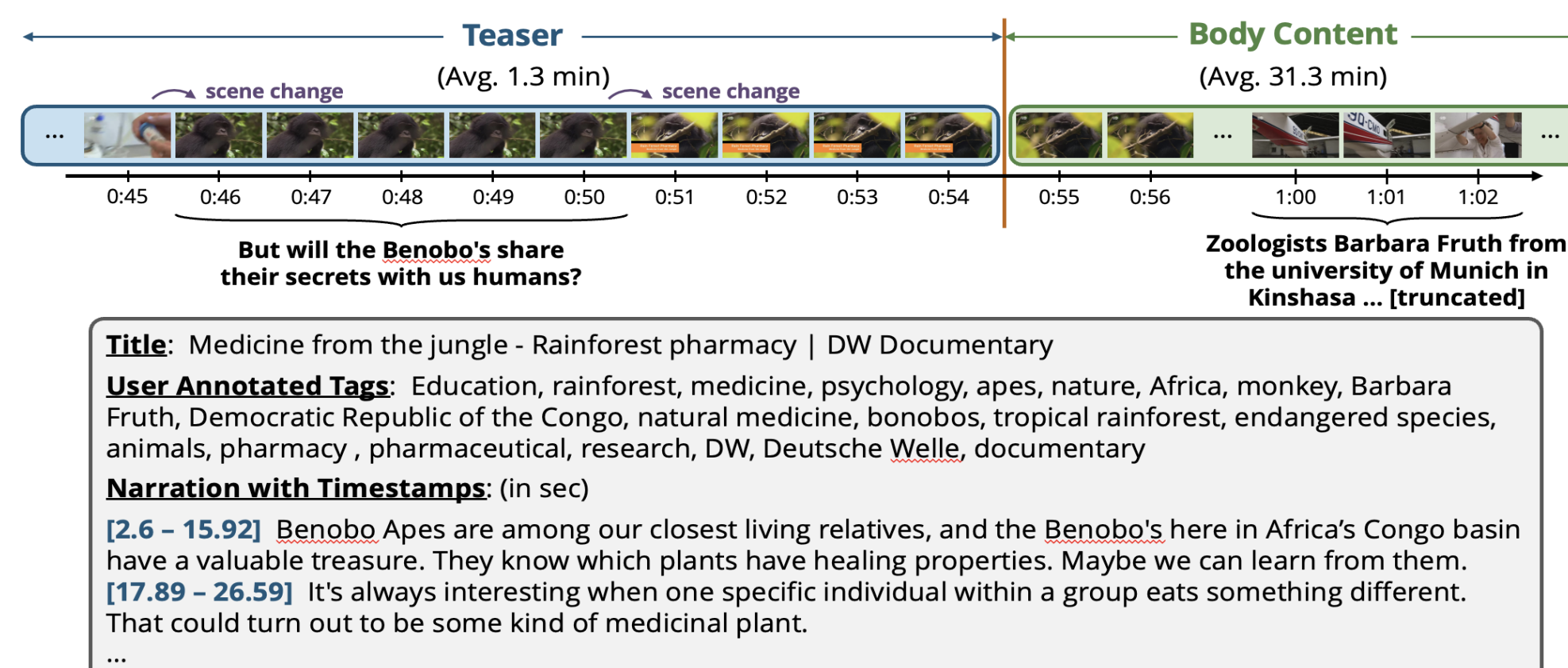
Contributions

- 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 **compress >30-min documentaries into <3-min teasers**.

DocumentaryNet

- **1,269** high-quality documentaries (600+ hours)
- **Sources:** DW Documentary, Public Broadcasting Service (PBS), and National Geographic
- **Annotations:** Metadata, audio tracks (separated into music, sound effect, and dialogue), and dialogue transcription with timestamps

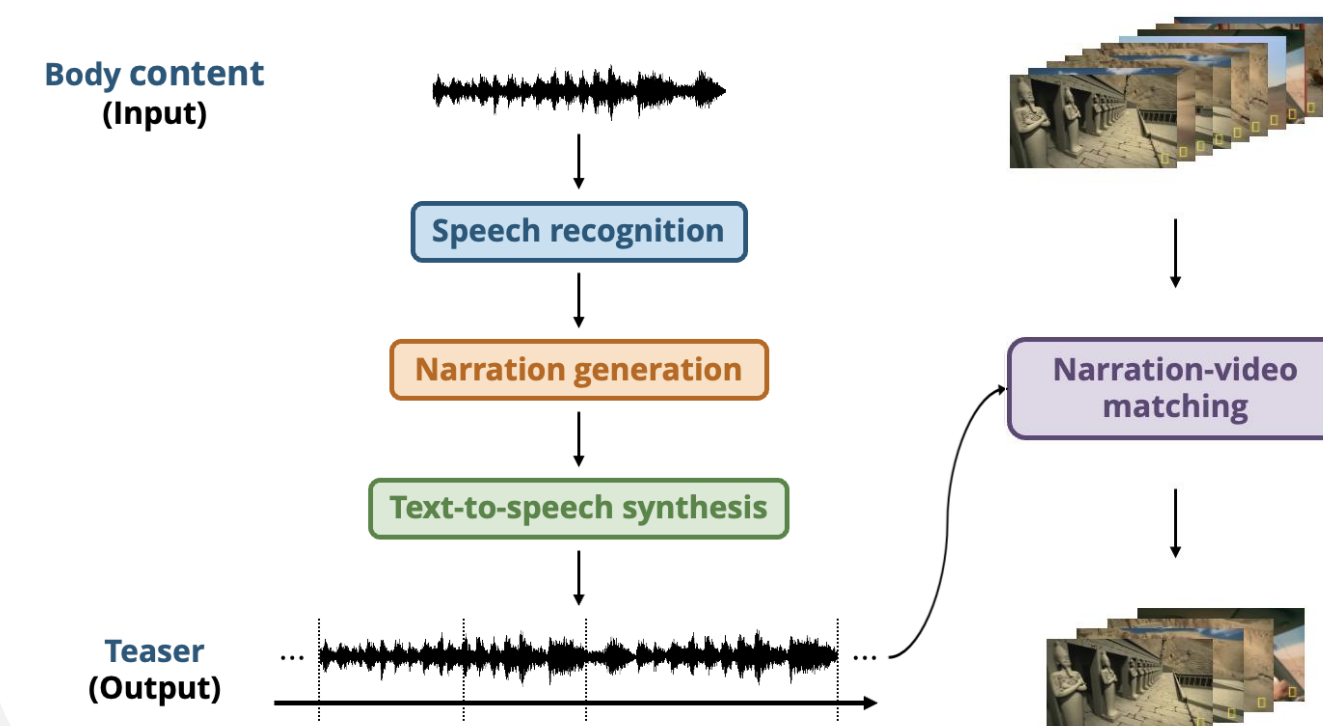
An example of DocumentaryNet



TeaserGen

Two-stage system

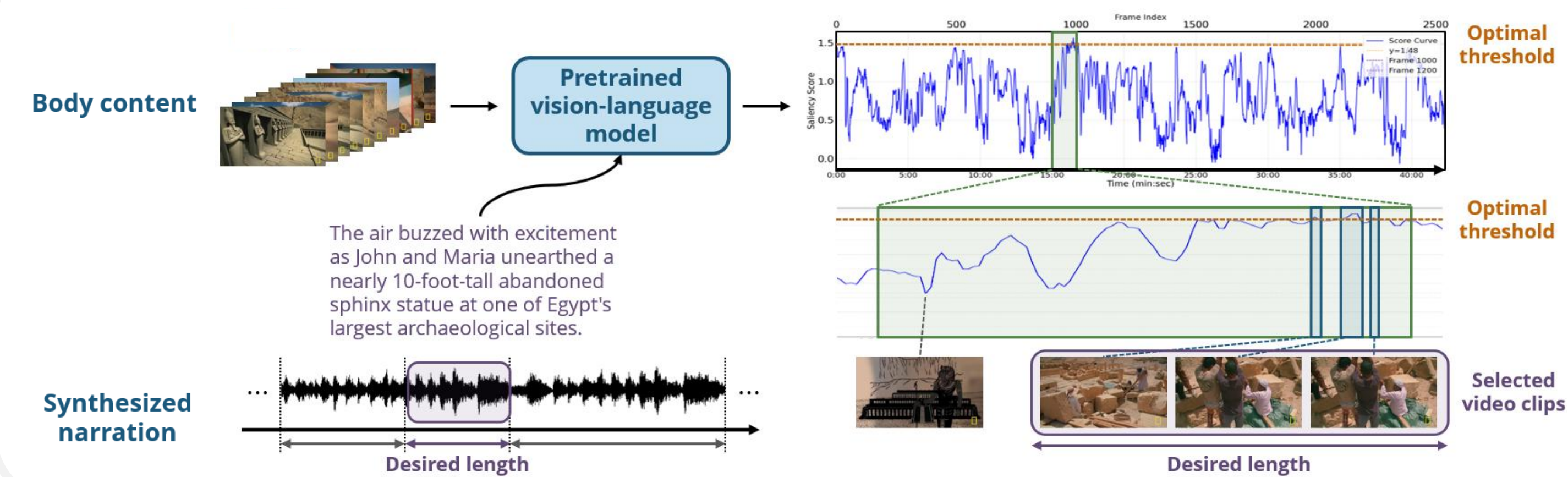
- Generate the teaser narration from the transcribed narration
- Select relevant visual content to accompany the teaser narration



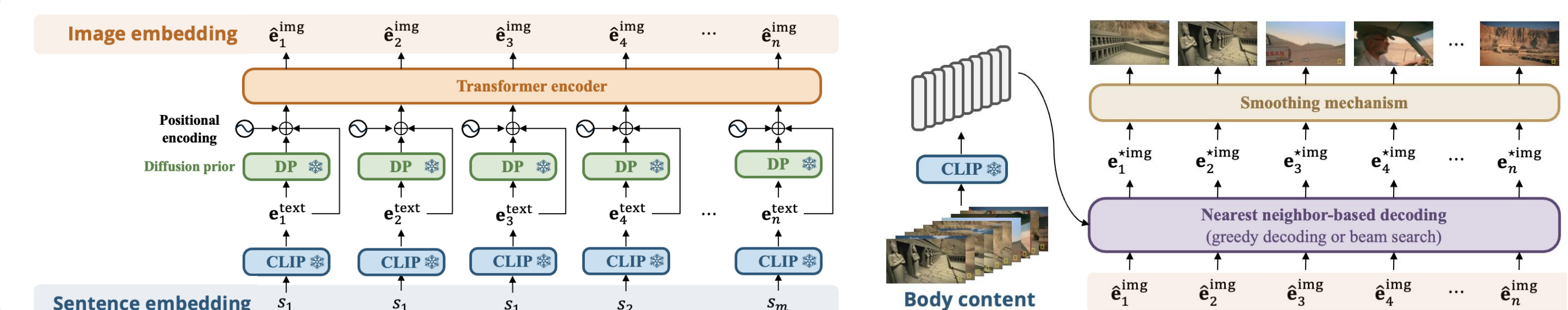
Narration generation

- Transcribe the documentary's narration
- Split it into 10 segments (3900 words each on average)
- Summarize each segment in one sentence using GPT-4o
- Rewrite the 10 summaries into a story-like teaser narration with an ending question using GPT-4o
- Synthesizes the generated teaser script into audio using a text-to-speech model

TeaserGen-PT: Interval-based approach



TeaserGen-LR: Learning-based approach



Results

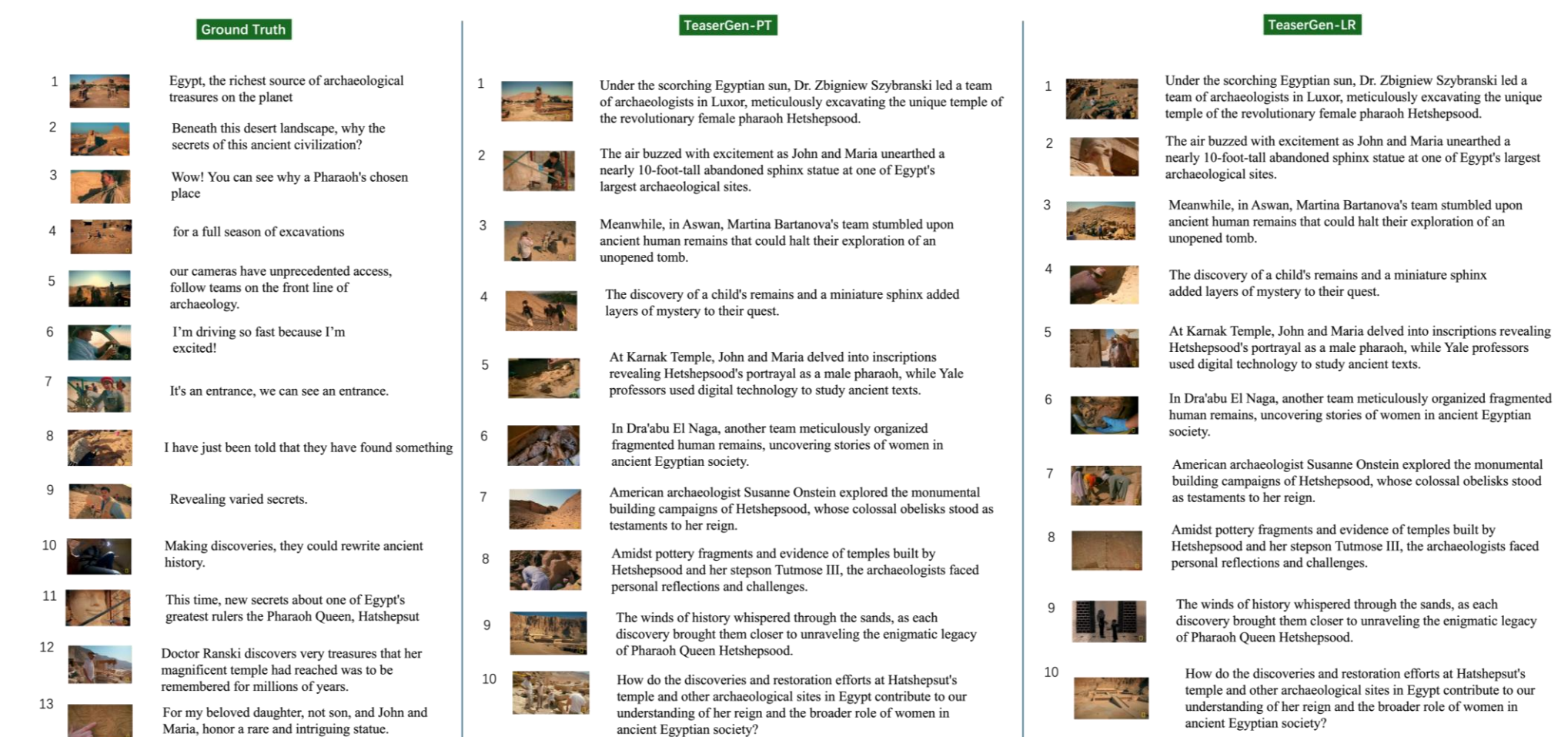
Objective evaluation results

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 evaluation results

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	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

Qualitative examples



Project page
wx83.github.io/TeaserGen_Official

Paper
arxiv.org/pdf/2410.05586

