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

NEURAL INFORMATION PROCESSING SYSTEMS

Weihan Xu[1] Yimeng Ma[1] Jingyue Huang[2] Yang Li[1] Wenye Ma[3]
Taylor Berg-Kirkpatrick[2] Julian McAuley[2] Paul Pu Liang[4] Hao-Wen Dong[5]

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

Overview

Generating shorts from long videos allows 1) audiences to digest information in a more engaging way and 2) content creators promote their long video contents.

Challenges

- Extractive methods stitch together video clips extracted from the input video, yet this may produce disjointed videos that do not together convey a coherent story.
- Abstractive approaches synthesize new narratives and even new scenes, but these methods cannot insert extracted video clips from the input video to support the generated narrative.

Contributions

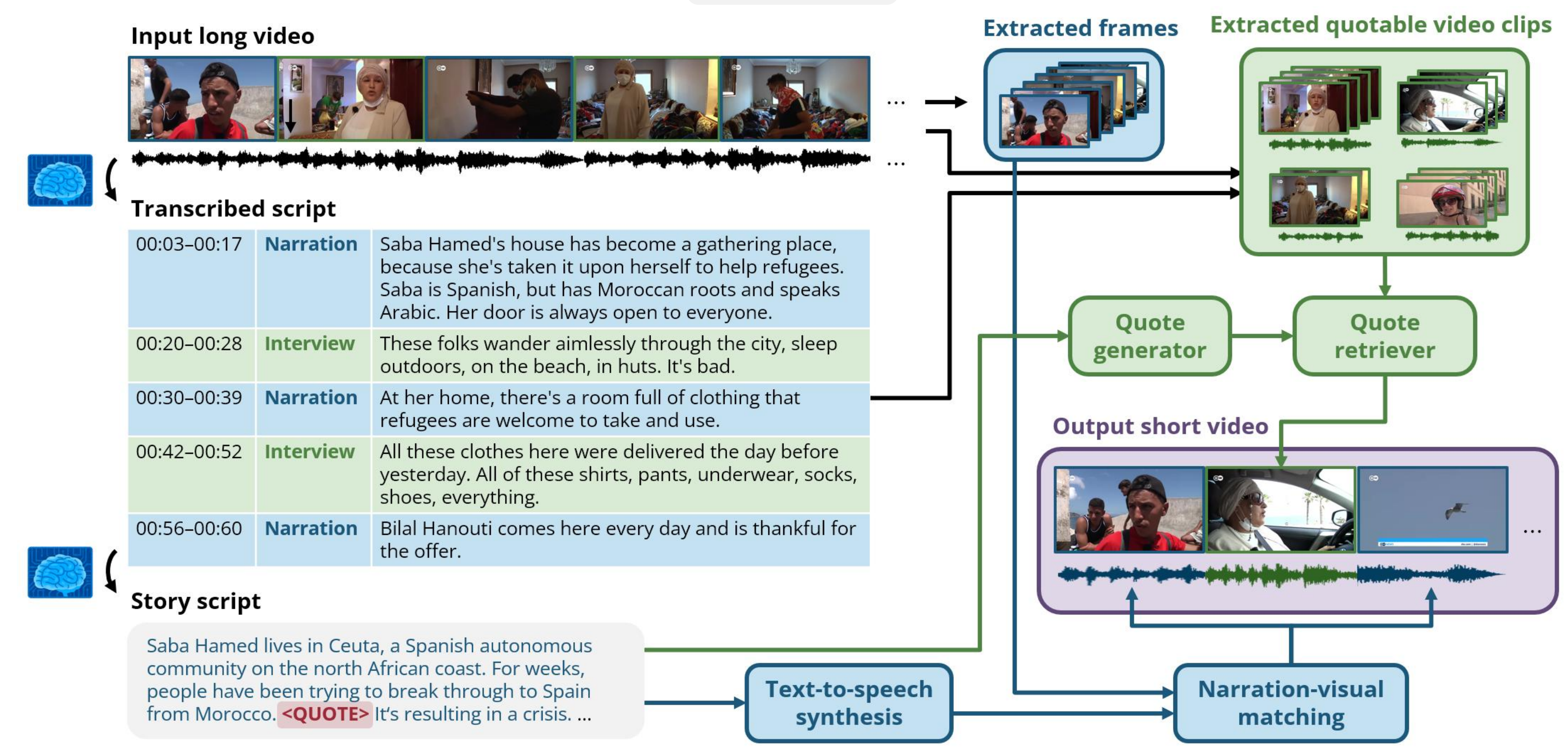
- We propose a new **retrieval-embedded generation (REG)** framework that allows an LLM to quote multimodal resources while maintaining a coherent narrative.
- We propose **REGen**, a novel long-to-short video editing model for generating shorts that feature a coherent narrative with **embedded video insertions** extracted from a long input video.

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

Method

REGen



Learning to Quote a Video

REGen-DQ (direct quote)

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

REGen-IDQ (indirect quote)

$$\ldots, x_i, \underbrace{\langle\mathrm{QUOTE}\rangle}_{\text{To be retrieved later!}}, x_{i+1}, \ldots$$

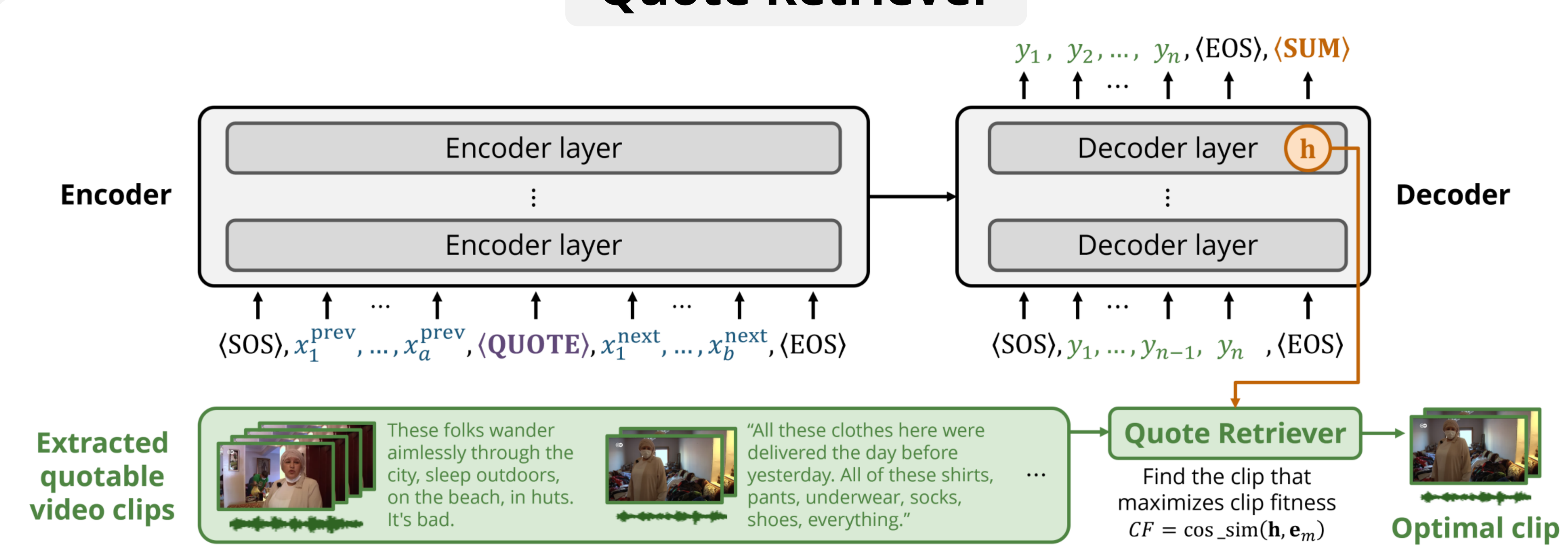
Clip Fitness

For a candidate clip c_m, the **clip fitness** is defined as $CF \coloneqq \cos\_\mathrm{sim}(\mathbf{h}, \mathbf{e}_m)$

REGen-IDQ-T (text only) $\mathbf{e}_m = \mathbf{e}_m^{\text{text}}$

REGen-IDQ-TV (text+video) $\mathbf{e}_m = \underbrace{f}_{\text{Learnable mapping}}\left(\mathrm{concat}(\mathbf{e}_m^{\text{text}}, \mathbf{e}_m^{\text{img}})\right)$

Quote Retriever



Results

Script Generation Methods

Model	Before fulfillment			After fulfillment				
	Tokens	QCR (%)	QDI	Tokens	R-1	R-2	R-L	G-Eval
Random extraction	-	98	11.71	235	0.27	0.04	0.12	0.56 ± 0.02
ETS	-	96	1.96	340	0.21	0.03	0.11	0.81 ± 0.01
A2Summ [4]	-	96	3.98	172	0.27	0.03	0.13	0.42 ± 0.01
TeaserGen [11]	-	-	-	304	0.21	0.03	0.11	0.85 ± 0.01
GPT-4o-DQ	292	98	4.02	402	0.22	0.05	0.12	0.77 ± 0.01
GPT-4o-SP-DQ	631	100	22.33	1372	0.13	0.03	0.07	0.75 ± 0.01
REGen-DQ	153	**76**	**2.31**	210	**0.28**	**0.05**	**0.13**	0.43 ± 0.02
REGen-IDQ-T	98	67	1.98	172	0.25	0.04	0.13	0.57 ± 0.02
REGen-IDQ-TV	98	67	1.98	179	0.25	0.04	0.13	$\mathbf{0.59 \pm 0.01}$
Ground truth	-	82	3.02	121	-	-	-	0.62 ± 0.03

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**	$\mathbf{3.56 \pm 0.22}$
QuoteRetriever-TV	Text+Visual	**5.00**	15.00	23.33	3.49 ± 0.26

Documentary Teaser Generation

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

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

Project page
Project page
wx83.github.io/REGen

Paper
arxiv.org/pdf/2505.18880
Paper