

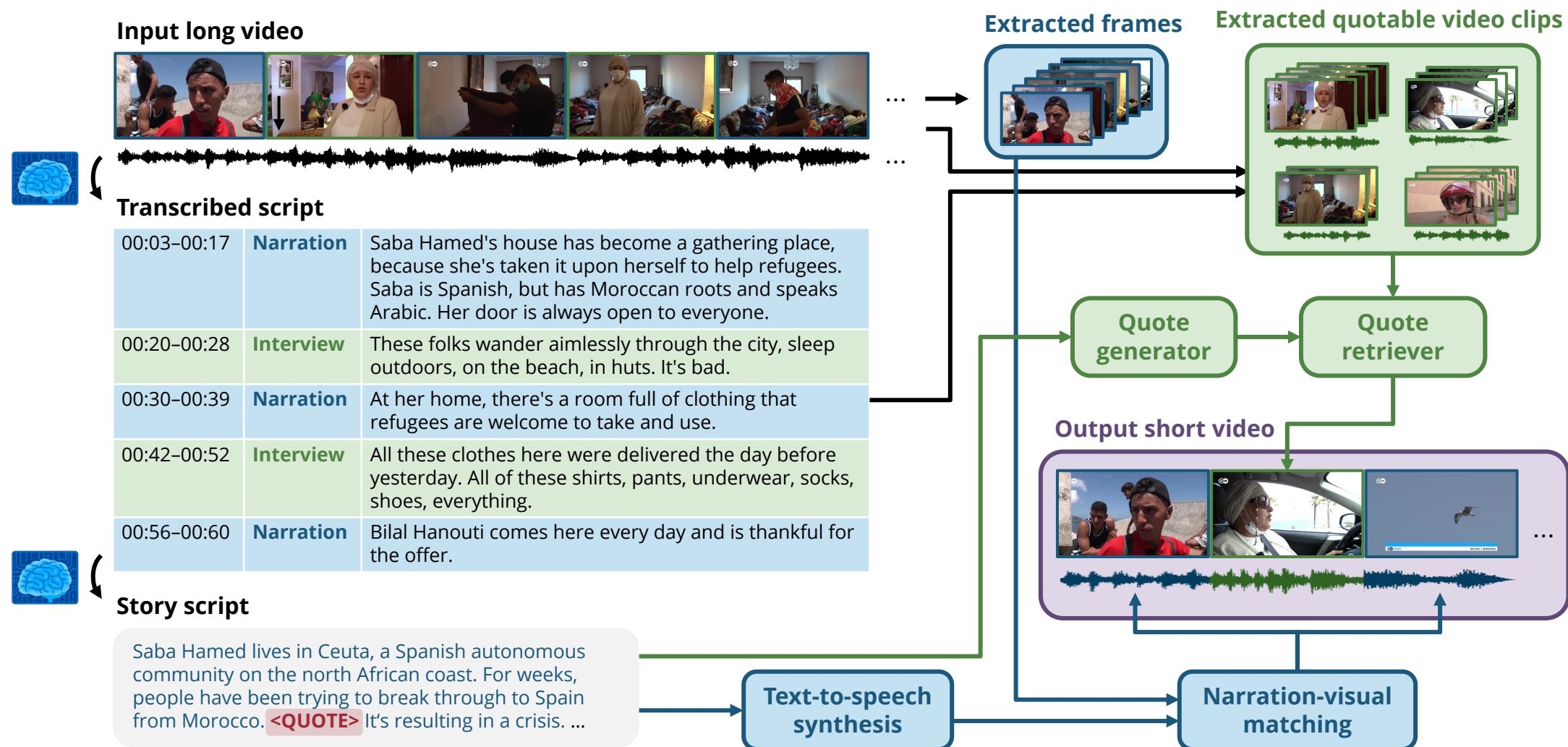
REGen: Multimodal Retrieval-Embedded Generation for Long-to-Short Video Editing

Wei-han Xu¹ Yimeng Ma¹ Jingyue Huang² Yang Li¹ Weyne Ma³
Taylor Berg-Kirkpatrick² Julian McAuley² Paul Pu Liang² **Hao-Wen Dong⁴**

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

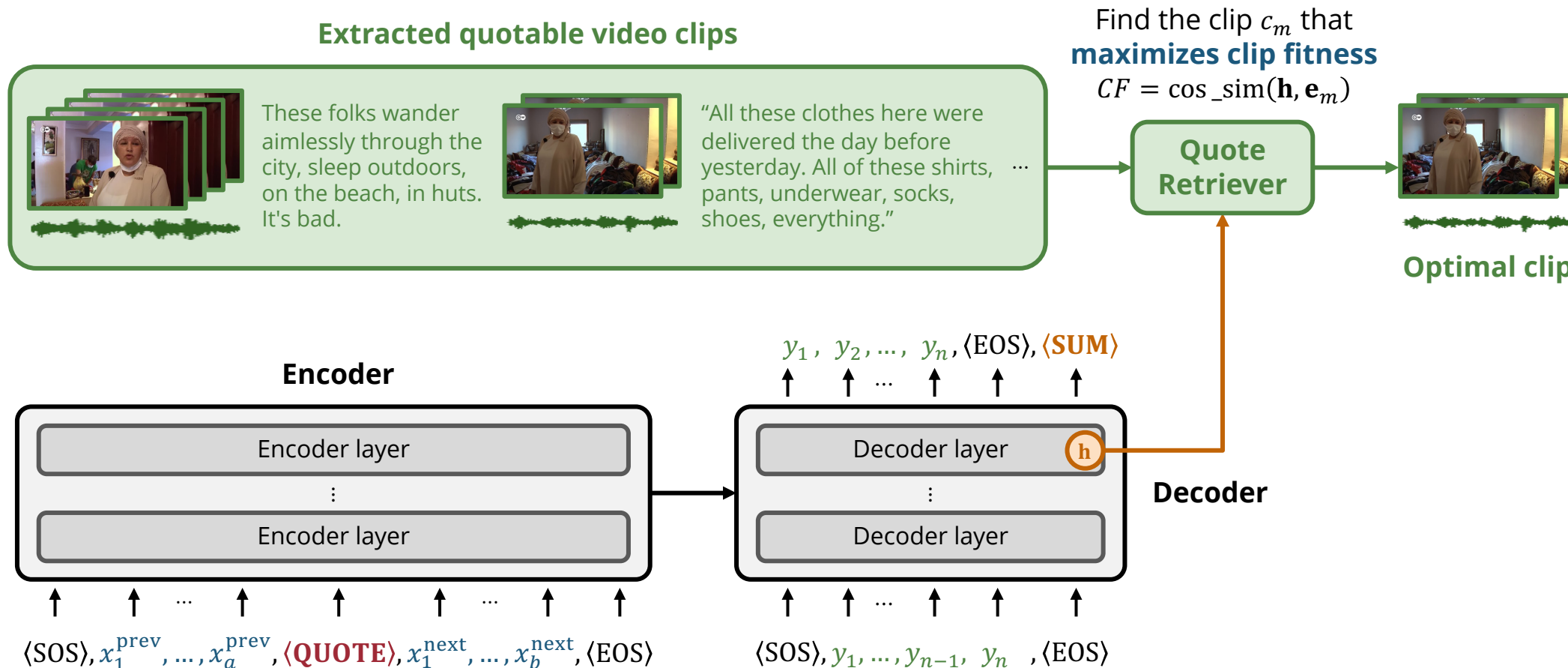
REGen-DQ
(direct quote)

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 \pm 0.00	0.28 \pm 0.48	7.22 \pm 5.54	3.08 \pm 0.25
GPT-4o infilling	Text only	2.78 \pm 0.48	13.89 \pm 1.27	22.50 \pm 1.44	2.48 \pm 0.31
QuoteRetriever-T	Text only	5.00	17.50	30.00	3.56 \pm 0.22
QuoteRetriever-TV	Text+Visual	5.00	15.00	23.33	3.49 \pm 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



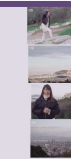
Narrator: Each park has its own artist and they're all based on themes, on human migration for example




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 Block 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? Yes, huge 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 resurrecting a modern-day 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 2002. I was taken to the headquarters of the state security in Cairo, down underground in their torture dungeons. I was Mindfucked. And that's where the worst ordeal began. Torture? They began electrocuting everyone.

Narrator: To better understand the relevance of these prophecies in 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 Manks, 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 (an 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. Ah, you know, a continuation of Western civilization is 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 book 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 in 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: 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 Jewish tradition, he has three things he's supposed to do. Number one, he's going to reconstruct the Jewish Kingdom, the Jewish state. Number two, he's going to bring peace with the neighbors. And number three, he's going to rebuild that temple. Retain 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 night. 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

Repetitiveness								
Model	Dur (sec)	Interview ratio (%)	F1 (%)	SCR (%)	REP (%)	VTGHLS	CLIPS-I	CLIPS-N
Random extraction	101	56 \pm 20	1.10	20.71	0.41	0.83	0.55	0.62
ETS	142	34 \pm 16	1.92	13.65	4.49	1.06	0.64	0.60
A2Summ [4]	73	42 \pm 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 \pm 42	1.56	16.55	20.75	1.01	0.58	0.42
GPT-4o-SP-DQ	619	61 \pm 17	2.07	12.38	18.33	1.02	0.62	0.62
REGen-DQ	95	37 \pm 26	1.45	19.13	10.35	1.05	0.48	0.57
REGen-IDQ-T	77	35 \pm 31	1.89	19.79	10.02	1.03	0.41	0.57
REGen-IDQ-TV	81	35 \pm 31	1.90	19.86	9.70	1.02	0.39	0.57
Ground truth	76	54 \pm 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