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

(/pdf?id=DxiP59Z81m)

Weihan Xu (/profile?id=~Weihan_Xu1), Yimeng Ma (/profile?id=~Yimeng_Ma1), Jingyue Huang (/profile?id=~Jingyue_Huang1), Yang Li (/profile?id=~Yang_Li124), Wenye Ma (/profile?id=~Wenye_Ma1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1), Julian McAuley (/profile?id=~Julian_McAuley1), Paul Pu Liang (/profile?id=~Paul_Pu_Liang1), Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1)

08 May 2025 (modified: 28 May 2025) NeurIPS 2025 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=DxiP59Z81m)

Keywords: Multimodal Large Language Model; Retrieval Augmented Generation; Video Editing

Abstract:

Short videos are an effective tool for promoting contents and improving knowledge accessibility. While existing extractive video summarization methods struggle to produce a coherent narrative, existing abstractive methods cannot `quote' from the input videos, i.e., inserting short video clips in their outputs. In this work, we explore novel video editing models for generating shorts that feature a coherent narrative with embedded video insertions extracted from a long input video. We propose a novel retrieval-embedded generation framework that allows a large language model to quote multimodal resources while maintaining a coherent narrative. Our proposed REGen system first generates the output story script with quote placeholders using a finetuned large language model, and then uses a novel retrieval model to replace the quote placeholders by selecting a video clip that best supports the narrative from a pool of candidate quotable video clips. We examine the proposed method on the task of documentary teaser generation, where short interview insertions are commonly used to support the narrative of a documentary. Our objective evaluations show that the proposed method can effectively insert short video clips while maintaining a coherent narrative. In a subjective survey, we show that our proposed method outperforms existing abstractive and extractive approaches in terms of coherence, alignment, and realism in teaser generation.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Reviewer Nomination: hwdong@umich.edu
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Applications (e.g., vision, language, speech and audio, Creative AI)
LLM Usage: Editing (e.g., grammar, spelling, word choice)
Declaration: I confirm that the above information is accurate.
Submission Number: 10908

Decision ✕ Official Review ✕ ✕ ⌄ Filter by author... ⌄ Search keywords...

Sort: Newest First

Everyone | Program Chairs | Submission10908... | Submission10908... | Submission10908... *5 / 18 replies shown*

Submission10908... Submission10908... Submission10908... Submission10908...

Submission10908... ✖

Add: Withdrawal

Paper Decision

Decision by Program Chairs 17 Sept 2025, 08:44 (modified: 18 Sept 2025, 09:41) Program Chairs, Authors Revisions (/revisions?id=GVpibMjzbD)

Decision: Accept (poster)

Comment:

All four reviews [yMCw,QQGE,tTNr,sppt] leaned towards acceptance and ultimately had the same borderline accept rating.

The reviewers appreciated several aspects of the work:

- The motivation was considered clear [tTNr]
- The idea was considered interesting [sppt]
- The task was considered novel [yMCw] and important [yMCw,sppt]
- The approach was considered a significant advance over existing video summarisation addressing a key weakness [yMCw]; similarly, the framework was considered effective [QQGE] and to have potential application in e.g. education [QQGE]
- The combination of extractive and abstractive summarisation was appreciated [QQGE]
- The proposed REGen system was considered technically sound and well engineered [yMCw]
- The experiments and ablation studies were considered extensive [tTNr]
- The experiment results were considered solid [sppt] and convincing [tTNr]
- The visualisation examples were appreciated [sppt]
- The paper was considered very well written and easy to follow by one reviewer [yMCw]

However, several weaknesses were pointed out, and authors responded to them in the rebuttal stage:

- The dependence of system performance on the quality of the initial speaker diarization was criticised [yMCw] and analysis of sensitivity to such upstream errors was desired [yMCw]; authors provided an experiment involving noise injection.
- Limitation primarily to documentaries was criticised [yMCw], as in appendix results on lectures a purely abstractive method was better; authors provided brief speculation that preprocessing of the lectures may be a cause of the difference.
- The complexity of the multi-stage pipeline was criticised [yMCw], and discussion of the cost and scalability was desired [yMCw,tTNr]; authors provided a brief argument about the pipeline and some details and discussion of the time consumption of the steps.
- Rationality of some evaluation metrics was criticised [sppt]; authors provided some discussion.
- Further analysis of low recall in quote retrieval was desired [sppt]; authors provided some discussion.
- Clarification on the role of the visual modality was desired [yMCw]; authors provided brief discussion.
- The claim of narrative coherence was considered not well supported by the sequential pipeline versus joint optimisation and the overlap ratio results [QQGE]; authors argued the overlap ration did not evaluate narrative coherence, and argued their multitask loss involves joint optimisation.
- Clarification of ensuring temporal coherence was desired with a concern about potential logical inconsistencies [QQGE]; authors replied they "intend to alleviate the issue" in the future via contextual embeddings.
- Analysis of the poor performance of LLM (GPT-4o) based quote filling-in was desired [yMCw]; authors analyse GPT-4o yielded generalised insertions.
- The use of fixed 10 chunks was criticised [QQGE]; authors replied they follow TeaserGen but did not seem to justify the amount of chunks.
- It was criticised the method could still place quotes in incorrect contexts [tTNr]
- Reliance on accurate video segmentation was criticised [tTNr] and its suitability to domains like lecture recordings was questions [tTNr]

- It was questioned whether the method could work by detecting moments instead of relying on diarization; authors provide some discussion on how in lecture videos they can detect slide transitions via OCR.
- Discussion on avoiding the limitations was desired [tTNr]; reviewers provided brief discussion of e.g. using OCR-based slide transition detection
- Discussion of some 'innovative method designs' were desired for the methods section [sppt]; authors stated they would expand some descriptions.
- One reviewer considered the methodological description dense and lacking emphasis on novelty [QQGE]; authors restated their main goal.

Ultimately after the rebuttals reviewers [yMCw,QQGE] considered their questions answered, [tTNr] considered most of the concerns addressed, and [sppt] considered their main issues to have been solved.

Overall, despite some concerns reviewers appreciated the work overall, and although some concerns could be further explored, it seems if the authors incorporate the material from their rebuttals the paper could be in a sufficiently good state to be presented at NeurIPS.

Official Review of Submission10908 by Reviewer yMCw

Official Review by Reviewer yMCw 02 Jul 2025, 16:32 (modified: 18 Sept 2025, 12:04)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer yMCw

Revisions (/revisions?id=tOcbG9tDzE)

Summary:
This paper introduces REGen, a novel framework for generating short promotional videos (e.g., teasers) from long-form videos like documentaries. The core problem it addresses is the inability of existing methods to create a coherent abstractive narrative while also embedding ("quoting") verbatim clips from the original source to ground the story in facts.

REGen tackles this with a two-stage, hybrid approach with Script Generation followed by Quotation Retrieval.

The final output is a short video that combines a synthesized narrative with directly extracted clips. The authors demonstrate the effectiveness of this system on the DocumentaryNet dataset, showing through comprehensive objective and subjective evaluations that REGen outperforms purely extractive and abstractive baselines in generating teasers that are coherent, factually grounded, and realistic.

Strengths And Weaknesses:
Strengths

1. The paper introduces a novel and highly relevant task: enabling generative models to "quote" multimodal content. This hybrid abstractive-extractive approach is a significant conceptual advance over existing video summarization techniques. It directly addresses a key weakness of purely abstractive methods (lack of factual grounding) and extractive methods (lack of narrative coherence), proposing an elegant and practical solution. The problem formulation is a strong contribution in itself
2. The REGen system is technically sound and well-engineered. The two-stage pipeline is a logical and effective way to decompose this complex task. The design of the Quote Retriever, which is jointly trained to perform masked infilling and retrieval, is particularly clever. The paper is supported by a thorough set of experiments with strong baselines.
3. The paper is exceptionally well-written and easy to follow. The motivation is clear, the methodology is explained in detail, and the figures (especially the system overview) are highly effective at illustrating the proposed framework.

Weakness

1. The system's performance is dependent on the quality of the initial speaker diarization used to identify quotable interview segments. The authors are transparent about the accuracy of this step, but errors here could negatively impact the entire pipeline by creating a flawed pool of candidate quotes. A brief analysis of the system's sensitivity to these upstream errors would be beneficial.

2. The method is developed and primarily evaluated on documentaries, a genre where the distinction between a narrator and an interviewee is often clear. The appendix includes a generalizability study on news and lecture videos, which honestly reports that users preferred a purely abstractive method for lectures. This suggests the "quoting" paradigm, as currently framed, may be best suited for specific content types.
3. The end-to-end system is a multi-stage pipeline involving several distinct models (ASR, diarization, LLM for scripts, retriever model). While this modularity is effective, it introduces complexity. Future work could explore more integrated, end-to-end models, although this would be a significant research challenge.

Quality: 3: good
Clarity: 4: excellent
Significance: 3: good
Originality: 3: good
Questions:
I have a few questions on the empirical analysis below:

1. The *QuoteRetriever-TV* model fuses text and visual features. However, the objective results in Table 3 show that it does not significantly outperform the text-only *QuoteRetriever-T* on recall metrics. Could you elaborate on the role of the visual modality? Is it more important for qualitative aspects not captured by recall, or does its benefit depend on the specific type of content being retrieved?
2. The results show that prompting a powerful LLM like GPT-4o to fill in the quote content and then using that text for retrieval performs poorly. This is an interesting finding. Why do you think this approach fails? Does the LLM generate text that is too generic or semantically divergent from the actual interview clips, leading to poor retrieval?
3. The proposed pipeline involves multiple models, including calls to GPT-4o for some baselines and processing steps. Could you comment on the computational cost and scalability of the REGen system? Is it feasible for processing very large video archives, or is it better suited for single-document editing?

Limitations:
I have covered potential empirical limitations in questions and weaknesses section above.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
NA

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
My concerns are addressed and I shall maintain the score.

Official Review of Submission10908 by Reviewer QQGE

Official Review by Reviewer QQGE 29 Jun 2025, 00:27 (modified: 18 Sept 2025, 12:04)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer QQGE

Revisions (/revisions?id=WYrZyQRDXN)

Summary:
This work presents a new multimodal retrieval-embedded generation framework, named REGen, for editing long videos into shorts. It combines abstractive and extractive methods via a two-stage pipeline: 1) generating a story script with quote placeholders using fine-tuned LLaMA, and 2) retrieving supporting video clips from the input to fill these placeholders. Experiments on documentary teaser generation show REGen effectively inserts video quotes while maintaining narrative coherence, outperforming baselines.

Strengths And Weaknesses:

Strengths:

1. Effective framework for creating documentary teasers and has promising potential for applications in education and other domains.
2. The combination of extractive and abstractive summarization methods is inspiring.

Weaknesses:

1. The description of the methodology is overly dense and lacks emphasis on the novel innovations.
2. The choice of using fixed 10 chunks is not empirically justified or supported by ablation studies, leaving the rationale unclear.
3. Although the paper emphasizes maintaining narrative coherence while integrating factual quotes, the architecture relies on a sequential pipeline rather than joint optimization. This limitation is reflected in Table 6, where the overlap ratio results fail to fully support the claim of narrative coherence.

Quality: 2: fair
Clarity: 3: good
Significance: 2: fair
Originality: 3: good
Questions:
Please refer to the weaknesses for detailed concerns.

Additional Questions:

1. For videos without clear speaker segmentation (e.g., lectures), could REGen be adapted to detect "quotable moments" based on visual/audio cues (e.g., slide transitions, emphasis in speech) rather than relying solely on diarization?
2. How does the system ensure temporal coherence when inserting retrieved clips? For instance, if a selected interview clip references an event that has not yet been introduced in the narrative, could this lead to logical inconsistencies in the final output?

Limitations:
Yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
No or very minor formatting issues

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The authors' response has successfully addressed my concerns. I maintain my initial rating with higher confidence.

Official Review of Submission10908 by Reviewer tTNr

Official Review by Reviewer tTNr 28 Jun 2025, 02:39 (modified: 24 Jul 2025, 09:12)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer tTNr
Revisions (/revisions?id=xxwTTqQedz)

Summary:
This paper proposes a novel approach for generating short videos that combine a coherent narrative with embedded video excerpts from longer videos. While extractive summarization methods often fail to produce cohesive stories, existing abstractive methods cannot directly incorporate video clips from the source material. To address this, the

authors introduce a retrieval-embedded generation framework (REGen). REGen first uses a fine-tuned large language model to generate a narrative script containing placeholders for video quotes, and then employs a retrieval model to select suitable video clips that replace these placeholders, ensuring alignment with the narrative. The method is evaluated on the task of documentary teaser generation, where short interview clips are commonly used to enhance storytelling. Results show that REGen effectively integrates video excerpts while maintaining narrative coherence, and outperforms existing extractive and abstractive approaches in terms of coherence, alignment, and realism, as demonstrated by both objective metrics and subjective user studies.

Strengths And Weaknesses:
Stengths:

1 The motivation of the paper is clear, that existing abstractive methods cannot directly incorporate video clips from the source material.

2 The authors conducted extensive experiments, including thorough ablation studies and comparisons with baselines, which convincingly demonstrate the effectiveness of the proposed REGen framework.

Weakness:

**1 ** Although video insertions help improve factual grounding, the method can still mistakenly place quotes in incorrect contexts. Incorporating all quotable materials into the initial script generation could reduce this issue, but this is technically challenging for LLaMA-based models due to their limited context window.

**2 ** The approach relies on accurate video segmentation, such as speaker diarization, which may not be applicable in other domains like lecture recordings.

Quality: 3: good
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:
1 Is the computation required by the proposed method expensive?

2 Have you ever had some ideas to improve the limitations you list?

Limitations:
yes

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
No

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes

Official Review of Submission10908 by Reviewer sppt

Official Review by Reviewer sppt 18 Jun 2025, 10:48 (modified: 18 Sept 2025, 12:04)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer sppt
Revisions (/revisions?id=xb8CO3WsxA)

Summary:
This paper proposed REGen, a multimodal retrieval-embedded generation method for long-to-short video editing, which utilizes a large language model for multimodal resources quoting while maintaining a coherent narrative. The proposed method is evaluated on the task of documentary teaser generation and outperforms existing abstractive and

extractive approaches in many evaluation metrics on teaser generation.

Strengths And Weaknesses:

Strengths:

1. The overall idea is interesting, and the task is important for real applications.
2. The experiment result is solid, both objective and subjective evaluations are provided on quoting short interview clips within a coherent narrative.
3. Convincing visualization examples are provided to prove the effectiveness of the proposed method.

Weaknesses:

1. There is some controversy over the rationality of some evaluation metrics, for example, QCR and QDI may be close to ground truth due to the methods used that generate scripts with a similar quote distribution to the ground truth scripts. But it does not affirmatively mean that the closer QCR and QDI are to the ground truth, the better the method would be for generating a similar quote distribution to the ground truth. On the contrary, the paper should use the average absolute difference (of number of quotes inserted per documentary, or proportion of test videos in which at least one quotation is correctly inserted) between the method and the ground truth, to demonstrate the performance difference for generating scripts with a similar quote distribution to the ground truth scripts.
2. Recall of quote retrieval methods seems relatively low, but this should be considered as a core component in the framework of this article and deserves further analysis and improvement.
3. The complete framework of Figure 1 lacks some innovative method designs, which the author should incorporate with the method section of the article.

Quality: 2: fair
Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:

See weaknesses. I believe that weakness 1 is especially crucial to address for theoretically supporting the evaluated performance. I would increase the rating if the reply is convincing.

Limitations:

The authors adequately addressed the limitations and potential negative societal impact of their work.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:

No major formatting concerns in this paper.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:

The additional experiment solving weakness 1 is encouraged to have a deeper analysis in the final paper, such as adding a comparison on the newly provided Quote Distribution Plot between baselines and the proposed method.

Meanwhile, the main issues from the weaknesses of my concern have been solved. I will raise my initial rating to 4.