TeaserGen: Generating Teasers for Long Documentaries

(/pdf?id=G1n50BMqzm)

Weihan Xu (/profile?id=~Weihan_Xu1),
Paul Pu Liang (/profile?id=~Paul_Pu_Liang1),
Haven Kim (/profile?id=~Haven_Kim1),
Julian McAuley (/profile?id=~Julian_McAuley1),
Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1),
Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1)

Published: 22 Jan 2025, Last Modified: 22 Jan 2025 ICLR 2025 conditionalonethicsreview Everyone
Revisions (/revisions?id=G1n50BMqzm) BibTeX

Keywords: Teaser Generation, Multimodal Learning, Vision-Language Model

Abstract:
Teasers are an effective tool for promoting content in entertainment, commercial and educational fields. However, creating an effective teaser for long videos is challenging for it requires long-range multimodal modeling capability for the input videos, while necessitating maintaining audiovisual alignments, managing scene transitions and preserving factual accuracy for the output teasers. Due to the lack of a publicly-available dataset, progress along this research direction has been hindered. In this work, we present DocumentaryNet, a collection of 1,269 documentaries paired with their teasers, featuring multimodal data streams of video, speech, music, sound effects and narrations. With DocumentaryNet, we propose a new two-stage system for generating teasers from long documentaries. The proposed TeaserGen system first generates the teaser narration from the transcribed narration from the documentary using a pretrained large language model, and then selects the most relevant visual content to accompany the generated narration through language-vision models. For narration-video matching, we explore two approaches: a pretraining-based model using pretrained contrastive language-vision models and a deep sequential model that learns the mapping between the narrations and visuals. Our experimental results show that the pretraining-based approach is more effective at identifying relevant visual content than directly trained deep autoregressive models.

Primary Area: applications to computer vision, audio, language, and other modalities
Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.
Submission Guidelines: I certify that this submission complies with the submission instructions as described on https://iclr.cc/Conferences/2025/AuthorGuide (https://iclr.cc/Conferences/2025/AuthorGuide).
Reciprocal Reviewing: I understand the reciprocal reviewing requirement as described on https://iclr.cc/Conferences/2025/CallForPapers (https://iclr.cc/Conferences/2025/CallForPapers). If none of the authors are registered as a reviewer, it may result in a desk rejection at the discretion of the program chairs. To request an exception, please complete this form at https://forms.gle/Huojr6VjkFxiQsUp6 (https://forms.gle/Huojr6VjkFxiQsUp6).
Resubmission: No
Student Author: Yes
Anonymous Url: I certify that there is no URL (e.g., github page) that could be used to find authors' identity.
No Acknowledgement Section: I certify that there is no acknowledgement section in this submission for double blind review.
Large Language Models: Yes, at the sentence level (e.g., fixing grammar, re-wording sentences)
Submission Number: 1391

Paper Decision

Decision by Program Chairs 22 Jan 2025, 00:24 (modified: 22 Jan 2025, 11:07) Program Chairs, Authors
Revisions (/revisions?id=PaDIIsrrVb)

Decision: Accept (conditional on ethics review)

Meta Review of Submission1391 by Area Chair 5Fyo

Meta Review by Area Chair 5Fyo 22 Dec 2024, 23:40 (modified: 22 Jan 2025, 11:04)
Senior Area Chairs, Area Chairs, Authors, Program Chairs Revisions (/revisions?id=PE3qcHnfy4)

Metareview:
The paper introduces a novel method for generating teaser clips, accompanied by a new benchmark. It received mixed reviews: two negative and two positive. The concerns raised were mostly minor, focusing on engineering aspects and stemming from misunderstandings. Although the authors' rebuttal appears to have effectively addressed these issues, the reviewers have not actively engaged with the rebuttal. Given that the authors successfully resolved the concerns, the AC concludes that the merits of the proposed work outweigh its flaws.

Additional Comments On Reviewer Discussion:
One of the reviewers assigned a score of 3, expressing several concerns which were mostly addressed in the initial author rebuttal. Although the reviewer acknowledged that many issues were resolved, they continued to express concern over the manual selection of a threshold set at 0.64, questioning its justification. The authors clarified in a subsequent response that this threshold was not used during training or inference, but solely for comparison purposes to aid reader comprehension of the results. Unfortunately, there was no further response from the reviewer.

The AC believes that the authors' responses adequately addressed the concerns raised, and attributes the persistently low score to the reviewer's lack of further engagement.

Official Review of Submission1391 by Reviewer uVty

Official Review by Reviewer uVty 03 Nov 2024, 22:51 (modified: 12 Nov 2024, 11:03) Everyone
Revisions (/revisions?id=Ey3Q6gteW4)

Summary:
This paper introduces TeaserGen, a method for generating teasers for long videos. To address the lack of suitable datasets, it presents the DocumentaryNet dataset, which contains 1,269 documentaries paired with their teasers. The dataset includes streams for video, speech, music, sound effects, and narrations. The proposed method is a two-stage system:

first, it generates the teaser narration using a large language model (LLM), and then it uses vision-language models to select the most relevant visual content.
Soundness: 2: fair
Presentation: 3: good
Contribution: 3: good
Strengths:

- The newly introduced dataset could be valuable for the research community.
- The task is interesting and meaningful.
- The experiments are thorough, and the performance appears good.

Weaknesses:

- The system heavily relies on LLMs and vision-language models, which may lead to error accumulation. How can we evaluate whether the teaser narration generated by GPT is effective?
- Some video summarization and highlight detection methods could also be applied to generate teasers, but the paper lacks a comparison with these approaches.

Questions:
Please refer to the weaknesses.

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold
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.
Code Of Conduct: Yes

Official Review of Submission1391 by Reviewer 339o

Official Review by Reviewer 339o 03 Nov 2024, 15:18 (modified: 25 Nov 2024, 02:05) Everyone
Revisions (/revisions?id=6jkQmn7Fzg)

Summary:
The paper addresses the challenge of creating effective teasers for long videos, the authors introduce DocumentaryNet, a dataset comprising 1,269 documentaries paired with their teasers. The proposed system operates in two stages:

Teaser Narration Generation: A pretrained large language model (LLM), is prompted to create engaging, story-like narratives with a thought-provoking ending question from the transcription. Visual Content Selection: For relevant video segments to be selected to match the narration using two methods: a pretrained contrastive language-vision model (TeaserGen-PT) and a deep learning-based sequential model (TeaserGen-LR) that aligns video frames to narration. TeaserGen-LR frames the narration-video matching as a sequence-to-sequence learning task, where it learns a direct mapping between the sequence of sentences in the teaser narration and the video frames. It uses transformer-based architecture with a diffusion prior to embed narration and visual sequences into a shared embedding space, enabling more nuanced and context-aware matching.

The authors used both objective metrics (like F1 score, CLIPScore, and scene change rate) and subjective user surveys that evaluate in terms of consistency, informativeness, and engagingness, to assess the quality of the teasers. They found that TeaserGen-PT with threshold-based selection often provided better coherence and alignment between video and narration, while TeaserGen-LR benefited from enhanced narration-video correspondence.

Soundness: 2: fair
Presentation: 3: good
Contribution: 2: fair
Strengths:

1. The paper proposes frameworks to generate teasers from documentary using audiovisual alignments and scene-changes.
2. The paper demonstrates robust experiments and comparisons to baseline models.

3. The authors use thorough evaluation on their dataset using both objective metrics (like F1 score and scene change rate) and subjective evaluations (coherence, engagingness) to validate their results.
4. The paper has shown extensive ablation studies.

Weaknesses:

1. Limited dataset scale, both for training and testing. Though the dataset is domain specific, still the scale is limited with just 1.2k documentaries.
2. Reliance on pretrained LLM, for teaser narration generation without any check for hallucinations or error compounding due to this step.
3. The work is very domain specific, the framework's reliance on pretrained language-vision models for narration-video alignment, while effective for documentaries, may struggle with complex visual elements that don't directly correspond to narration, such as scenes with symbolic or artistic visuals or videos with limited narrative.
4. Comparison with other existing video summarization models is not shown in the paper.

Questions:

1. The paper claims generalizability of proposed models, however the work is only shown qualitatively through some examples. It would be interesting to get some quantitative results for the same.
2. Can we see the framework's results on how it adapts to other videos with complex visual elements that don't directly correspond to narration, such as scenes with symbolic or artistic visuals or videos with limited narrative.
3. Comparison with other existing video summarization models, on how the framework is against other models on this task.

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold
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.
Code Of Conduct: Yes

Official Review of Submission1391 by Reviewer VbpS

Official Review by Reviewer VbpS 30 Oct 2024, 22:57 (modified: 12 Nov 2024, 11:03) Everyone
Revisions (/revisions?id=ZRA3bg93yW)

Summary:

- The paper presents TeaserGen, a two-stage system for creating teasers from long documentaries, addressing challenges like audiovisual alignment, smooth transitions, and factual accuracy. To support this, the authors developed DocumentaryNet, a dataset of 1,269 documentaries with teasers, including multimodal elements like video, narration, and sound effects.
- TeaserGen first generates teaser narration from the documentary's transcript using a large language model, creating an engaging summary. It then pairs visuals with narration through either a pre-trained contrastive language-vision model or a deep sequential model to match visuals accurately.
- Results show that TeaserGen outperforms baseline models in maintaining coherence and alignment, offering a streamlined approach to automated teaser generation. DocumentaryNet and TeaserGen together provide valuable tools for advancing multimodal content modeling in documentary summarization.

Soundness: 3: good
Presentation: 3: good
Contribution: 3: good
Strengths:

- Tackles a unique problem in automated teaser generation for documentaries with TeaserGen, a creative, narration-centered two-stage approach that combines large language models with language-vision models for cohesive narration and visual alignment, showing effective and innovative use of existing technologies.

- Provides solid empirical support with comparisons to baseline models across objective (e.g., F1 score, CLIPScore) and subjective metrics, as well as the introduction of DocumentaryNet, a multimodal dataset with documentary-teaser pairs that enriches resources available for this research area.
- The work has significant potential impact by addressing a real-world gap in video summarization for documentary-style content, with applications in multimedia and educational fields, and establishes a foundation for further multimodal research, likely to stimulate new directions in long-form video modeling.

Weaknesses:

- The paper presents an innovative approach to video teaser generation using pretrained language-vision models, but several issues need to be addressed to enhance its clarity and robustness. In rows 210 and 211, there is a notation inconsistency where S is defined as a sequence of language tokens, yet later each S_i is referred to as a waveform (audio signal). This inconsistency creates confusion, and it's crucial for the notation to consistently represent either language tokens or audio waveforms throughout the paper to avoid misunderstandings.
- In Section 4.2.1, the method relies heavily on a single pretrained VTGHS model without sufficient ablation studies or comparisons with alternative architectures, which weakens the validation of the approach. The constraints imposed such as a minimum clip length of three seconds and a one-second overlap between clips, appear arbitrary and lack theoretical or empirical justification, raising questions about their effectiveness and impact on the results. Additionally, using a frame rate of only one frame per second (1 FPS), as mentioned in rows 453 to 455, is inadequate for videos that change dynamically. This low frame rate hinders the model's ability to effectively capture motion and make accurate predictions.
- In Section 4.2.2, the model extracts features at a low frame rate of 1 FPS, causing multiple frames to share identical sentence embeddings. This coarse temporal resolution fails to capture dynamic changes within the video, leading to overly repetitive scenes. The absence of fine-grained temporal annotations in the dataset prevents the model from effectively distinguishing and assigning unique embeddings to semantically similar frames occurring at different times. As a result, the approach struggles to maintain diversity and temporal coherence in the generated visual content.
- Regarding threshold selection, in row 320, the paper states, `We estimate a VTGHLS of 0.64 for the ground truth teasers` , but there is insufficient explanation about the criteria used to select this threshold value. The paper does not analyze how varying this threshold affects the results. Since there are ablation studies on changing the matching score function, as mentioned in Section 5.6, it is necessary to explore this threshold value in more detail to understand its impact on the overall pipeline.
- Table 4 may present an unfair comparison because the TeaserGen models utilize advanced decoding techniques like beam search, which can enhance performance, while the baseline models do not use these techniques. For a fair comparison that accurately reflects each model's true capabilities, all models should employ similar decoding methods.
- A significant limitation in the methodology is the heavy reliance on subjective metrics without incorporating standardized quantitative measures, as mentioned in rows 468 to 470. This over-reliance weakens the study's reproducibility and generalizability. While subjective listening tests provide valuable insights, the absence of automatic evaluation metrics commonly used in natural language processing and computer vision (such as ROUGE, BLEU, BERTScore, or perplexity for the generated narration tex) makes it difficult to objectively compare the results with other approaches or validate the findings across different contexts.
- The TeaserGen-LR model uses a limited architecture with only three transformer layers, which may not capture complex patterns as effectively as the diffusion prior's more extensive 12-block backbone. Relying solely on L2 distance as the loss function might not fully capture perceptual similarities, leading to less nuanced image generation. Additionally, training the model for only 15 epochs on a small test set of 49 documentaries raises concerns about underfitting and limits the generalizability of the results.
- The study does not discuss the potential computational overhead introduced by incorporating the diffusion prior, which could affect the practicality and scalability of the approach. While higher scene change rates may increase visual diversity, they might also compromise the narrative coherence of the generated teasers. Addressing these issues would enhance the study's robustness and applicability.

Questions:

- Could you clarify the notation used in rows 210 and 211? Specifically, is S intended to represent a sequence of language tokens or audio waveforms? Consistent notation throughout the paper would enhance understanding and prevent confusion.

- Have you considered evaluating alternative architectures or conducting ablation studies to assess the robustness of relying solely on the pretrained VTGHL model? Exploring different models could strengthen the validation of your approach.
- What is the rationale behind setting the minimum clip length to three seconds and the overlap between clips to one second? Providing theoretical or empirical justification for these specific constraints would help in understanding their impact on the results.
- Given that a frame rate of 1 FPS may be insufficient for capturing dynamic video content, have you experimented with higher frame rates? How does increasing the frame rate affect the model's ability to capture motion and improve prediction accuracy?
- How does your model address the issue of repetitive scenes arising from multiple frames sharing identical sentence embeddings? Have you explored methods to incorporate fine-grained temporal information or annotations to enhance diversity and temporal coherence?
- Could you elaborate on how the VTGHLS threshold of 0.64 was determined? Additionally, have you investigated how varying this threshold influences the results, perhaps through an ablation study?
- Given that the TeaserGen-LR model uses only three transformer layers, have you tested deeper architectures to see if they capture complex patterns more effectively? Would increasing the number of layers improve performance?
- Have you considered using standardized quantitative evaluation metrics like ROUGE, BLEU, BERTScore, or perplexity for assessing the generated narration text? Including these metrics could enhance the reproducibility and comparability of your study.
- Relying solely on L2 distance as the loss function might not fully capture perceptual similarities. Have you experimented with alternative loss functions, such as perceptual loss, SSIM loss and so on to potentially achieve more nuanced image generation?
- Considering that the models were trained for only 15 epochs on a relatively small test set of 49 documentaries, have you explored training for more epochs or using a larger dataset? How might this affect model performance and generalizability? To test the generalizability of your approach, have you considered evaluating the model on additional datasets beyond the 49 documentaries? How does the model perform on different genres or types of video content?
- Could you provide details on the computational resources required by the diffusion prior model? Understanding the computational overhead would help assess the practicality and scalability of your approach in real-world applications.
- Higher scene change rates may enhance visual diversity, have you evaluated their impact on the narrative coherence of the teasers? How do you balance diversity with maintaining a coherent and engaging storyline?
- You employ different CLIP models (CLIP-ViT-B/32 and CLIP-ViT-L/14) for different components of your system. Have you considered the potential inconsistencies this might introduce? Would using the same CLIP model throughout improve the alignment between textual and visual data?

-How does the frame extraction rate influence the model's performance in terms of capturing essential visual information? Have you analyzed the trade-offs between computational efficiency and the richness of visual features at different frame rates?

Flag For Ethics Review: No ethics review needed.
Rating: 3: reject, not good enough
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Official Review of Submission1391 by Reviewer Hqts

Official Review by Reviewer Hqts 30 Oct 2024, 04:45 (modified: 25 Nov 2024, 02:51) Everyone
Revisions (/revisions?id=kyc3QcZCqg)

Summary:

This paper presents TeaserGen, a two-stage system for generating promotional teasers for long documentaries. Leveraging a their proposed dataset, DocumentaryNet, the authors aim to generate teasers by first synthesizing teaser narrations from documentary transcripts using a large language model. They then use a language-vision model to select relevant visual content that aligns with the narration. In that process, to avoid repeat frames, they proposed some methods to alleviate. The study compares a pretraining-based model (TeaserGen-PT) and a deep sequential learning model (TeaserGen-LR) for narration-video alignment. Experimental results show some advantages of the pretraining-based approach over directly trained autoregressive models.

Soundness: 3: good

Presentation: 2: fair

Contribution: 2: fair

Strengths:

1)The DocumentaryNet dataset of 1,269 documentary-teaser pairs fills a gap in the community by providing a publicly available resource for multimodal summarization and teaser generation, and the system shows potential for applications in media, advertising, and education for automated content promotion.

2)The inclusion of various evaluation metrics, including scene change rate and repetitiveness, provides a more nuanced assessment of the teaser's quality.

Weaknesses:

1)The approach predominantly relies on pretrained language-vision models, limiting novelty. This reliance raises questions about the model's true capacity to generate creative outputs, as it functions more as an information retrieval system instead of a real generative system.

2)The proposed model does not consider the alignment of audio cues like music or sound effects with visual elements, which limits its ability to produce emotionally engaging teasers, I've checked in your demo, there also exists this issue, and none of above baseline methods consider this aspect, and if you could consider this, that would be a great contribution in this domain. In this aspect, you get all frames by your frame-matching system driven by the text narration you got from llm, so the audio should also be different correspond to each frame in this proposed hypothesis. And by the way, this multi-stage method, means multi-stage information loss and also your pretrained model from general domain also have information loss, so I don't think this method work well for the task that you work with in this paper.

3)The sentence-by-sentence matching approach does not effectively capture scene continuity, leading to potentially fragmented visual sequences that lack coherence even though you utilized the smoothing and regularisation.

4)The experiments are mainly based on one proprietary dataset (other methods also didn't train on your dataset) and lack extensive ablation studies for elements such as the diffusion prior model and threshold sensitivity, which are critical to understanding the model's flexibility and robustness.

Questions:

1)Line229-230: You should detail this part about how to find the threshold and does it sensitive and what is the pretrained model here. especially in the supplementary material.

2)Line267-268: Please illustate well about the diffusion prior and how it helps and also do you conduct the ablation study about this.

3)Line693-706: You collect these media from three main sources, but in the appendix A that you explain about the dataset, I didn't see any discussion about the copyright for these videos.

Flag For Ethics Review: Yes, Legal compliance (e.g., GDPR, copyright, terms of use)

Details Of Ethics Concerns:

In the dataset part where they specify in the Appendix A, they didn't discuss about all three parts sourcees of their collected medias, the copyright and availability for research should be discussed.

Rating: 5: marginally below the acceptance threshold

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.

Code Of Conduct: Yes