M3SUM: A NOVEL UNSUPERVISED LANGUAGE-GUIDED VIDEO SUMMARIZATION

Hongru Wang, Baohang Zhou*, Zhengkun Zhang, Yiming Du, David Ho, Kam-Fai Wong*

The Chinese University of Hong Kong, Nankai University

ABSTRACT

Language-guided video summarization empowers users to use natural language queries to effortlessly summarize lengthy videos into concise and relevant summaries that cater specifically to their information needs, which is more friendly to access and digest. However, most of the previous works rely on tremendous (also expensive) annotated videos and complex designs to align different modals at the feature level. In this paper, we first explore the combination of off-the-shelf models for each modal to solve the complex multi-modal problem by proposing a novel unsupervised language-guided video summarization method: ***M**odular **M**ulti-**M**odal **Sum**marization* (M3Sum), which does not require any training data or parameter updates. Specifically, instead of training an alignment module at the feature level, we convert all modal information (*e.g.* audio and frames) into textual descriptions and design a parameter-free alignment mechanism to fuse text descriptions from different modals. Benefiting from the remarkable long-context understanding capability of large language models (LLMs), our approach demonstrates comparable performance to most unsupervised methods and even outperforms certain supervised methods.

***Index Terms*—** video summarization, ChatGPT

1. INTRODUCTION

The saying goes, *"If a picture is worth a thousand words, then a video is worth a million."* This emphasizes the immense value and richness of information that can be conveyed through videos. Given the vast amount of video content available, there is a pressing need for an efficient technique that can summarize or edit videos and extract essential and relevant information. Nowadays, lots of works focus on using natural language (*e.g.* user's query) to guide the video summarization [1] or edit the video [2], leading to more customized and accessible video artifact without need to learn complicated video editing tools like Premiere and Final Cut, as shown in Figure 1(a).

Most of the previous works reduce the task to a frame-wise score prediction problem with the user query as an additional input signal [1]. However, the existing learning paradigm suffers from two obvious limitations: 1) they require tremendous high-quality annotated videos which are expensive and labor-consuming to acquire [3, 4]; 2) There is a specific or complex network design to align different modals at the feature-level which requires time-consuming training and sometimes even additional data. Inspired by recent progress which simply uses transcript to edit the videos in both academics and industries[1], especially after the release of Whisper [5] API[2], we explore a novel data-agnostic paradigm to summarize the video solely based on textual descriptions by leveraging the exceptional capability of LLMs.

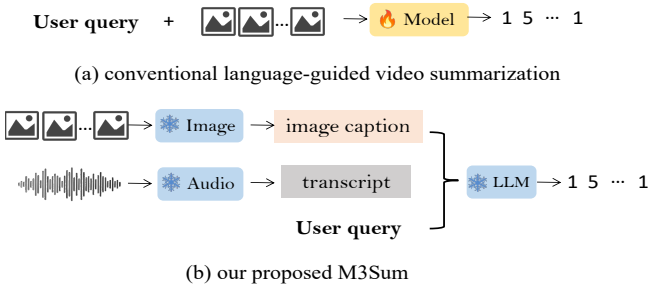

Fig. 1. (a) The conventional framework of language-guided video summarization, necessitates the abundant annotated data and time-consuming training; (b) our proposed M3Sum that bases on off-the-shelf models without any training.

Specifically, we present a novel unsupervised language-guided video summarization framework: ***M**odular **M**ulti-**M**odal **Sum**marization* (M3Sum) [3], as shown in Figure 2(b), to modularize each modal information in the video. Instead of training a linear layer or other complicated networks [6, 2] to align visual features and textual features, we focus on converting all modal information (*e.g.* audio, image, and text) into textual descriptions to unleash the potential of single text modal, supported by a carefully designed alignment mechanism without requiring any additional training. In this way, the video is decomposed into multiple modals first, and each module is responsible for converting a specific modal into textual information. These modules can be dynamically combined (*or* removed) as needed to adapt to different application scenarios and summary requirements. For example, in educational videos, the audio modality plays a crucial role,

*Equal contributions.

[1]https://app.pictory.ai/ and https://www.assemblyai.com

[2]https://openai.com/blog/introducing-chatgpt-and-whisper-apis

[3]The code will be published at https://github.com/ZovanZhou/M3Sum.

while in landscape videos, the visual frame becomes more significant. We summarize our contributions as follows: 1) To the best of our knowledge, this is the first work that explores unsupervised language-guided video summarization from a single-modal perspective; 2) We propose a novel unsupervised summarization framework: M3Sum that can be easily applied to any video, regardless of domain or topics, without the need for any training, thanks to the exceptional understanding and generalization capability of LLMs. 3) Our experimental results on two generic video summarization datasets reveal the robustness and effectiveness of M3Sum, which outperforms several unsupervised methods and some certain supervised methods.

2. RELATED WORK

2.1. Language-guided Video Editing

There are some earlier works that focus on query- or topic-focused video summarization [7], and language-guided video summarization is first introduced by [1], in which a multi-modal summarization model is proposed which takes two inputs, a video, and a natural language text, and synthesizes a summary video conditioned on the text. Following a similar definition, [2] extend to the language-guided video editing (LBVE) task and propose a multi-modal multi-level framework to learn the correspondence between video perception and language semantics. Furthermore, text-guided video completion (TVC) is also explored which requires the model to generate a video from partial frames guided by an instruction. However, most of these previous methods heavily rely on large amounts of training data and focused primarily on frame-level features of the video, resulting in computational complexity, difficulty in generalizing to unseen data.

2.2. Unsupervised Video Summarization

A further line of work investigates unsupervised video summarization [8], mainly adopting two kinds of approaches: Generative Adversarial Network (GAN) [9, 10, 11, 12] and Reinforcement Learning with complex hand-crafted rewards (RL) [13, 14, 15], but almost all of them require training to assess the importance of each frame. While large language models (LLMs) have demonstrated impressive capabilities in natural language processing tasks [16], their application to video summarization is relatively limited. Our goal is to explore the potential of LLMs to leverage their understanding of context for video summarization.

3. METHOD

3.1. Task Definition

Let $V = \{F_1, F_2, ..., F_N\}$ be a video sequence consisting of N frames, where F_i represents the i_{th} frame. Video summarization can be defined as the process of generating a summary $S = \{F_{s_1}, F_{s_2}, ..., F_{s_M}\}$ from the video sequence V, where $M << N$ is the desired length of the summary. The summary S consists of a subset of frames from V that captures the most salient and representative information. Following lots of previous works, we formulate this problem as a frame-wise score predication task, and the key shots will be picked to form the final summary using 0/1 knapsack algorithm after converting frame-level scores to shot-level scores [17].

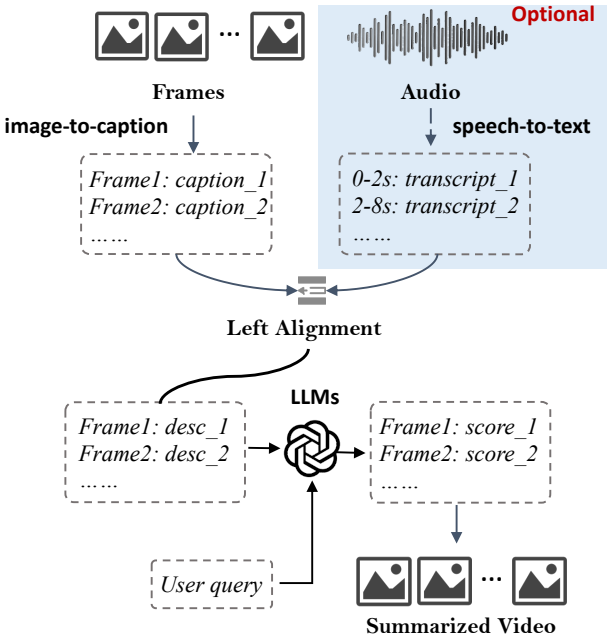

Fig. 2. The framework of our proposed method

3.2. Framework

Our framework can be divided into three steps: 1) converting all modal information into textual descriptions (*All to Text*); 2) the alignment of different modal information (*Alignment*); and 3) scoring the frames according to fusion results in the last step (*Score Frames*). The framework is shown in Figure 2.

Step 1: All to Text. There are two major modals in the video: frame and audio features. For the former one, we utilize the off-he-shelf `image-to-caption` model to generate a caption for each frame in the video. For the latter one, we utilize another off-the-shelf `speech-to-text` model to transfer audio into a transcript, including multiple sentence segments and corresponding timestamp labels (SRT format). Thus, we can convert different modals into textual descriptions.

Step 2: Alignment. The importance of different modal information varies across different types of videos, and even some videos may not have audio information. In this way, we consider frame features as a primary resource following lots of previous works [1, 2]. Additionally, we have developed an algorithm that determines whether or not to utilize audio in-

Algorithm 1 Alignments of Different Resources

Input: The threshold τ for selecting different modals,
Image caption set of frames $C = \{c_i\}_{i=1}^{N}$ where c_i is the caption sentence of the i-th frame,
Transcript set of frames $T = \{(t_i^1, t_i^2, w_i)\}_{i=1}^{L}$ where t_i^1 and t_i^2 are the start and end times, and w_i is the sentence of the i-th transcript.
Output: Description set of frames $D = \{d_i\}_{i=1}^{N}$.

for $i = 1, 2, \ldots, L$ **do**
  $s_1 = \frac{t_i^1}{2}$, $s_2 = \frac{t_i^2}{2}$ /* The frame is sampled every 2 seconds.*/
  for $j = s_1, s_1 + 1, \ldots, s_2$ **do**
    $\tilde{w}_j = w_i$ /* $\tilde{w}_j$ is the transcript of j-th frame. */
  end for
end for
$BS = BertScore(C, T)$.
/* The overlap is big and thus no need to fuse.*/
if $BS > \tau$ **then**
  for $i = 1, 2, \ldots, N$ **do**
    $d_i = c_i$
  end for
else
  for $i = 1, 2, \ldots, N$ **do**
    $d_i = c_i \cup \tilde{w}_i$
  end for
end if
return $D = \{d_i\}_{i=1}^{N}$

The user query is "[USER QUERY]". Please help me score the frames for selecting important frames from the video according to the given query. Here is the process you will follow:
1. Firstly, you will read the descriptions of frames from the same video across different ranges.
2. After that, you will review each frame's descriptions and establish a relationship between them based on the given user query.
3. The range for scoring the frames should be 1 to 5; a higher score represents greater significance for the video and the query. Using the frame descriptions, you must assign scores to each frame. It is suggested to allocate more low scores and fewer higher scores.
4. Lastly, you must output the frame scores in JSON format, excluding any descriptions of frames.
The presented descriptions of frames are as follows:
Frame1: "a blurry image of a cell phone in a dark room"
Frame2: "a white refrigerator sitting on top of a white wall"
.....

Table 1. The standard prompt of unsupervised M3Sum to score each frame according to the user query.

formation and aligns it with the frame when it is decided to be used, as shown in Algorithm 1.

Step 3: Score Frames. In the last step, we take advantage of the long-context understanding capability of LLMs, such as ChatGPT, to score each frame according to the user query and frame description. Specifically, we carefully design two types of prompting:

Standard prompting. We directly feed the frame description, user query, and instructions together to the LLM. The template can be found in Table 1.

Progressive CoT. We first prompt the LLM to generate a summary of all the frame descriptions in the video as shown in Table 2. This summary provides a concise text representation of the entire video. Next, we feed the summary along with frame descriptions, user queries, and instructions to LLM to score each frame, mimicking the way a human annotator would assess and assign scores to frames after watching the complete video. We also try directly appending the summary into each description of frames but the performance is not satisfactory due to the long context.

Please summarize the important information based on all descriptions of frames from the video. Making sure your answer is concise and accurate. The descriptions of frames are presented as follows:
Frame1: "a blurry image of a cell phone in a dark room"
Frame2: "a white refrigerator sitting on top of a white wall"
.....

Table 2. The prompt to summarize video information

4. EXPERIMENT

4.1. Implementation Details

Experimental Setup. We choose the state-of-the-art open-source large-v2 whisper model [5] as our `speech-to-text` model and we select ViT-GPT2[4] as our `image-to-caption` model due to its high number of downloads. We mainly adopt `ChatGPT` as our LLM with the temperature as 0.1 and top p as 0.1. We conduct our experiments on two generic video summarization datasets: TVSum [18] and SumMe [19], and we use the title of the video as the user query and F1 as evaluation metrics following [1, 15].

Baselines. We select several unsupervised models for comparison: SUM-FCN [20], DR-DSN [13], EDSN [14], UnpairedVSN [21], Online Motion-AE [22], AC-SUM-GAN [8], we select SUM-GAN$_{sup}$ [9], SUM-FCN, CLIP-it [1] for supervised setting for reference.

4.2. Main Results

Table 3 shows the results of TVSum and SumMe. We observe that our method outperforms 6 out of 7 unsupervised

[4]https://huggingface.co/models?other=image-captioning

Model	Training	TVSum	SumMe
Supervised Setting			
SUM-GAN$_{sup}$	✓	56.3	41.7
SUM-FCN	✓	56.8	47.5
CLIP-it	✓	66.3	54.2
Unsupervised Setting			
Online Motion-AE	✓	51.5	37.7
SUM-FCN	✓	52.7	41.5
DR-DSN	✓	57.6	41.4
EDSN	✓	57.3	42.6
UnpairedVSN	✓	55.6	47.5
CLIP-Image+bi-LSTM	✓	52.8	35.7
AC-SUM-GAN	✓	60.6	50.8
M3Sum (SP)	✗	56.9	43.6
M3Sum (PCoT)	✗	57.6	41.9

Table 3. Comparing F1 scores of our M3Sum with supervised and unsupervised baselines on the TVSum and SumME datasets. The results of baselines are copied from [1] and [15].

methods on TVSum except for AC-SUM-GAN and DRL-SLS, 5 out of 7 unsupervised methods on SumMe except for UnpairedVSN, AC-SUM-GAN, and DRL-SLS. Despite there still being a gap between the current supervised/unsupervised SOTA method, our method still outperforms some certain supervised methods such as SUM-FCN. In addition, we found that PCoT prefer video that contains informative captions and transcripts. The performance of SumMe is relatively lower since the videos in SumMe mostly do not contain informative transcripts, leading to poor summaries in PCoT. The strength of our method is that it does not require any prior knowledge, annotated data, or parameter updates, which makes it capable of processing any video and thus serves as a strong baseline for all videos, and a video annotator for some videos which are expensive or time-consuming to annotate.

Model	TVSum		SumMe	
	SP	PCoT	SP	PCoT
M3Sum	**56.9**	**57.6**	**43.6**	41.9
- caption	51.8	51.6	32.6	37.5
- transcript	56.7	57.2	43.2	**42.2**
- alignment	50.4	52.8	43.1	40.8

Table 4. The ablation study of M3Sum on the two datasets.

4.3. Ablation Study

We also conducted an ablation study to investigate the effects of different modules. Table 4 presents the results. We directly fuse the transcript with the caption without the filter using threshold (- alignment). The performance is dropped after removing different modules, revealing the effectiveness of each module, except for removing the transcript at the SumMe dataset due to similar reasons in the main experiment.

4.4. Analysis of Different Metrics

We also try different metrics to align different modal information. The result can be found in Table 5. We select the best metric BertScore [23] in our main experiment.

Metrics	TVSum		SumMe	
	SP	PCoT	SP	PCoT
PPL	57.1	56.8	42.5	41.2
BLEU	57.2	57.6	43.2	41.9
BertScore	**57.2**	**57.6**	**43.6**	**42.3**

Table 5. The comparison of different alignment metrics.

4.5. Analysis of Threshold of Different Metrics

Figure 3 shows the performance of F1 under different thresholds using BLEU and BertScore. It can be observed that different threshold has a serious effect on the final performance. In addition, different methods may prefer different thresholds. We set the threshold in the main experiment which can lead to the highest performance.

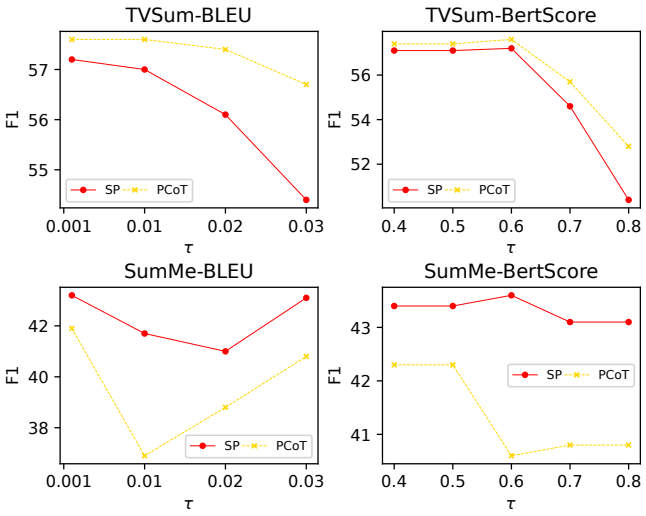

Fig. 3. The influence of different threshold values.

5. CONCLUSION

In this paper, we explore a novel paradigm: M3Sum that utilizes the exceptional ability of LLMs to summarize video information by converting all modals into textual descriptions. Our proposed method demonstrates promising and robust performance on two widely used video summarization datasets.

6. ACKNOWLEDGEMENTS

This paper was partially supported by ITF project no. PRP-054-21FX, HKSAR.

7. REFERENCES

[1] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell, "Clip-it! language-guided video summarization," *NeurIPS*, vol. 34, pp. 13988–14000, 2021.

[2] Tsu-Jui Fu, Xin Eric Wang, Scott T Grafton, Miguel P Eckstein, and William Yang Wang, "M3l: Language-based video editing via multi-modal multi-level transformers," in *CVPR*, 2022, pp. 10513–10522.

[3] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[4] Hao Jiang and Yadong Mu, "Joint video summarization and moment localization by cross-task sample transfer," in *CVPR*, 2022, pp. 16388–16398.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan, "Flamingo: a visual language model for few-shot learning," 2022.

[7] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao, "Video summarization via semantic attended networks," in *AAAI*, 2018, vol. 32.

[8] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras, "Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *TCVST*, vol. 31, no. 8, pp. 3278–3292, 2020.

[9] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *CVPR*, 2017, pp. 202–211.

[10] Bin Zhao, Xuelong Li, and Xiaoqiang Lu, "Property-constrained dual learning for video summarization," *TNNLS*, vol. 31, no. 10, pp. 3989–4000, 2019.

[11] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *MM*, 2019, pp. 2296–2304.

[12] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *MultiMedia Modeling*. Springer, 2020, pp. 492–504.

[13] Kaiyang Zhou, Yu Qiao, and Tao Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *AAAI*, 2018, vol. 32.

[14] N Gonuguntla, B Mandal, NB Puhan, et al., "Enhanced deep video summarization network," BMVC, 2019.

[15] Ye Yuan and Jiawan Zhang, "Unsupervised video summarization via deep reinforcement learning with shot-level semantics," *TCVST*, vol. 33, no. 1, pp. 445–456, 2022.

[16] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong, "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs," in *Findings of EMNLP 2023*, Singapore, Dec. 2023, pp. 12047–12064, Association for Computational Linguistics.

[17] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," in *ECCV*. Springer, 2016, pp. 766–782.

[18] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes, "Tvsum: Summarizing web videos using titles," in *CVPR*, 2015, pp. 5179–5187.

[19] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool, "Creating summaries from user videos," in *ECCV*. Springer, 2014, pp. 505–520.

[20] Mrigank Rochan, Linwei Ye, and Yang Wang, "Video summarization using fully convolutional sequence networks," in *ECCV*, 2018, pp. 347–363.

[21] Mrigank Rochan and Yang Wang, "Video summarization by learning from unpaired data," in *CVPR*, 2019, pp. 7902–7911.

[22] Yujia Zhang, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing, "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognition Letters*, vol. 130, pp. 376–385, 2020.

[23] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.