
M3SUM: A Novel Unsupervised Language-guided Video Summarization

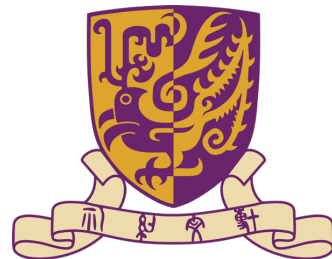
Hongru Wang*, Baohang Zhou*, Zhengkun Zhang,

Yiming Du, David Ho, Kam-Fai Wong

{hrwang, kfwong}@se.cuhk.edu.hk zhoubaohang@dbis.nankai.edu.cn

The Chinese University of Hong Kong, Nankai University

* Denotes equal contribution





- Introduction
- Related Work
- M3SUM
- Experiments
- Conclusion and Future Directions

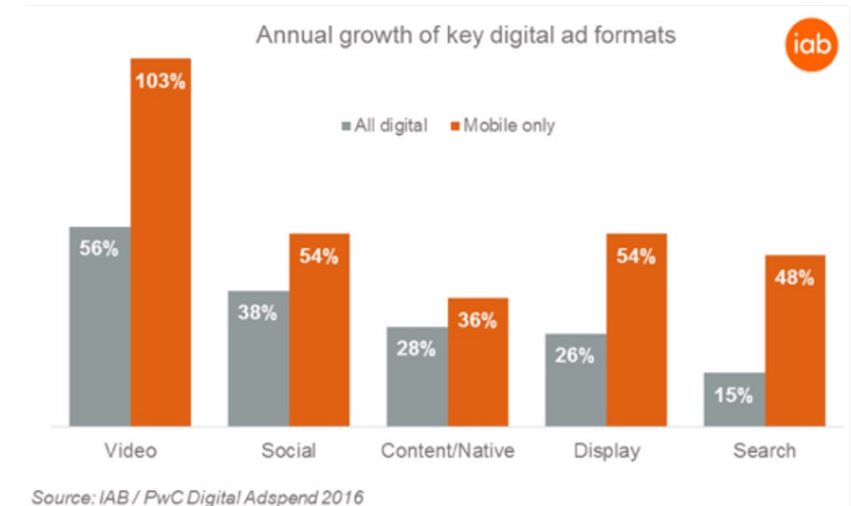


- Introduction
- Related Work
- M3SUM
- Experiments
- Conclusion and Future Directions

Introduction



- Video is arguably the primary source of information in today's digital era, with the significant growth annually. For example, 144,000 hours of video are uploaded to YouTube daily.
- Through video clips, people can acquire different new knowledge. Most users seek information, learn skills from the videos.



Challenges:

- Unable to quickly summarize a video clip by **skimming**
- Unable to **search and locate a specific piece of information** within a video clip using simple query.

How to summarize long video *in the way the user needs* ?



Language-guided Video Summarization





- Introduction
- **Related Work**
- M3SUM
- Experiments
- Conclusion and Future Directions

Language-guided Video Summarization/Editing

Clip-it

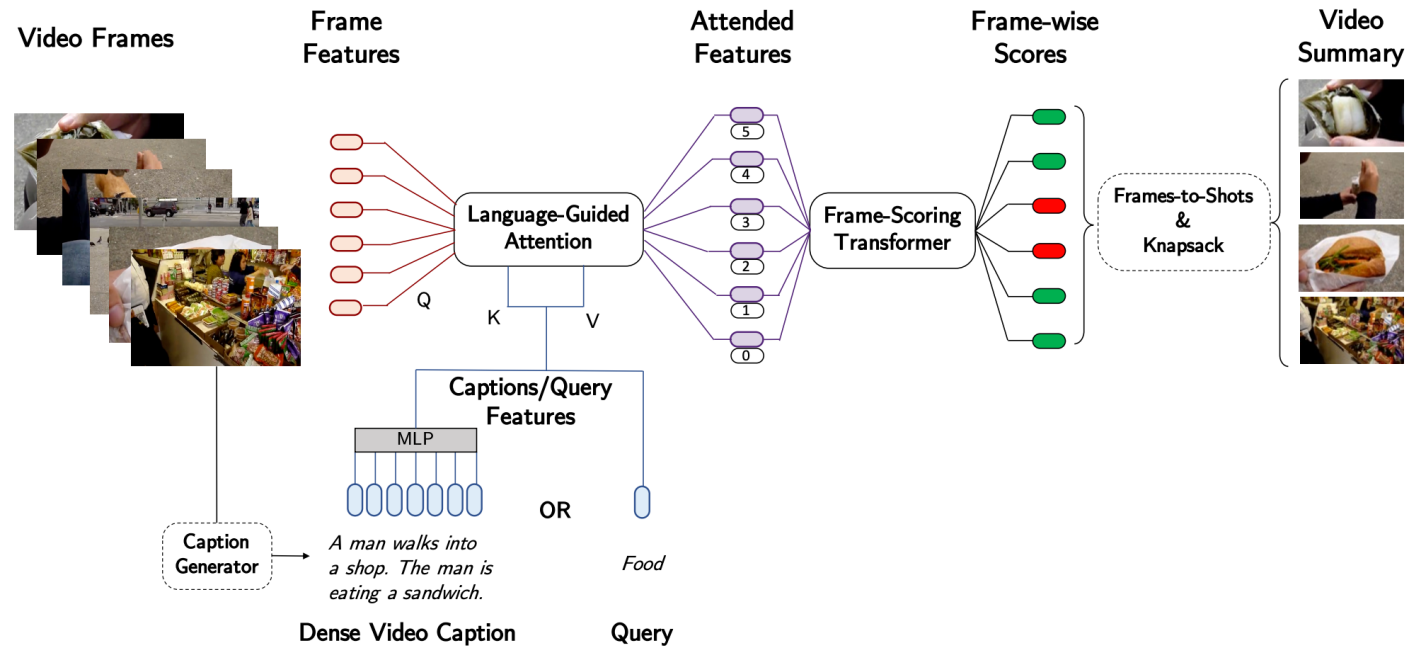


Figure 2: **Overview of CLIP-It.** Given an input video, CLIP-It generates a summary conditioned on either a user-defined natural language query or an automatically generated dense video caption. The Language-Guided Attention head fuses the image and language embeddings and the Frame-Scoring Transformer jointly attends to all frames to predict their relevance scores. During inference, the video summary is constructed by converting frame scores to shot scores and using Knapsack algorithm to select high scoring shots.

Language-guided Video Summarization/Editing

M³L

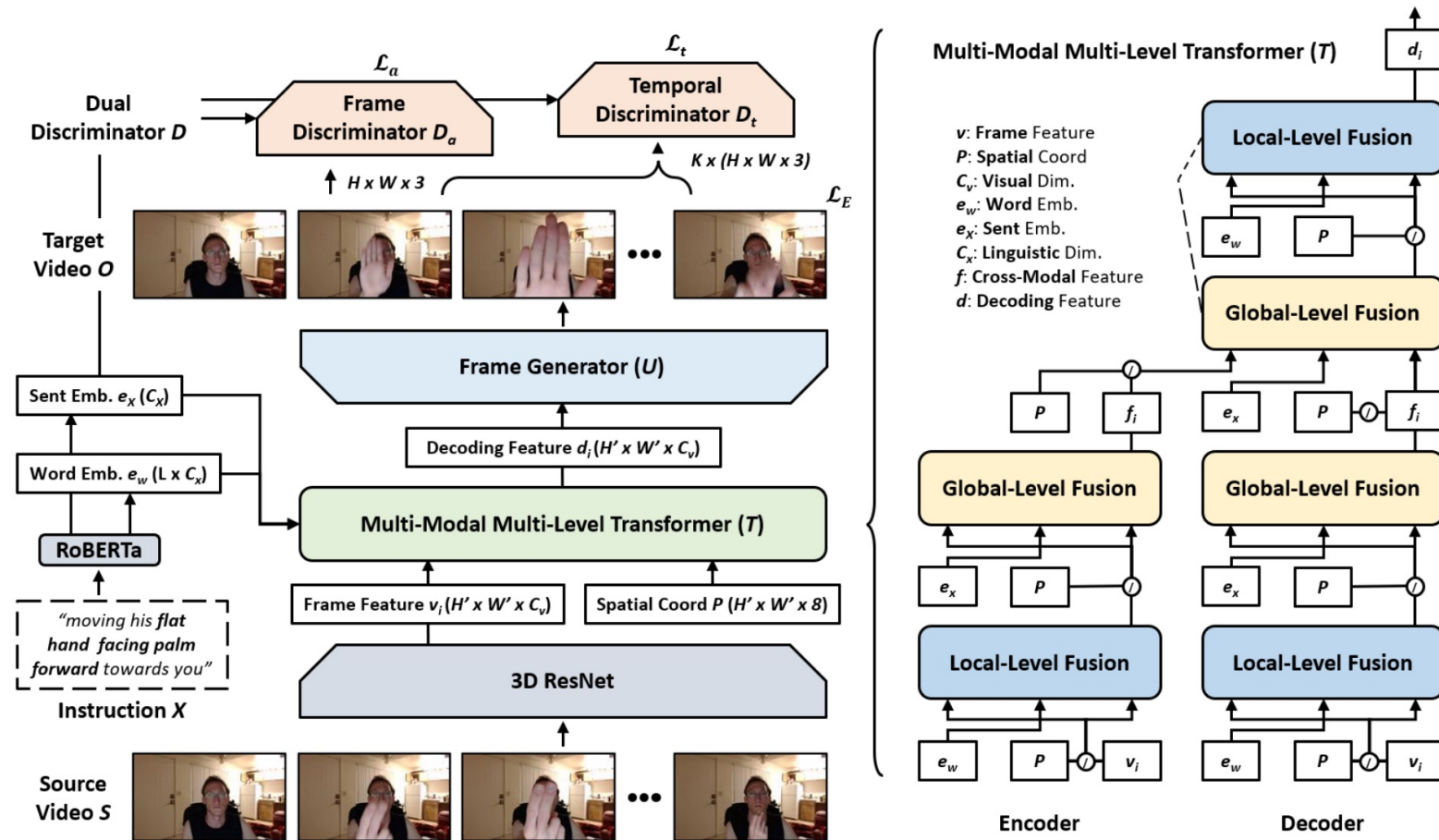


Figure 2. An overview architecture of our multi-modal multi-level transformer (M³L). M³L contains the multi-modal multi-level transformer T to encode the source video S and decode for the target video frame o by the multi-level fusion (MLF).



- Introduction
- Related Work
- **M3SUM**
- Experiments
- Conclusion and Future Directions

- **Unsupervised language guided video summarization**

- Without the need of training
- Without the need of labeled data

- **LLM-based summarization**

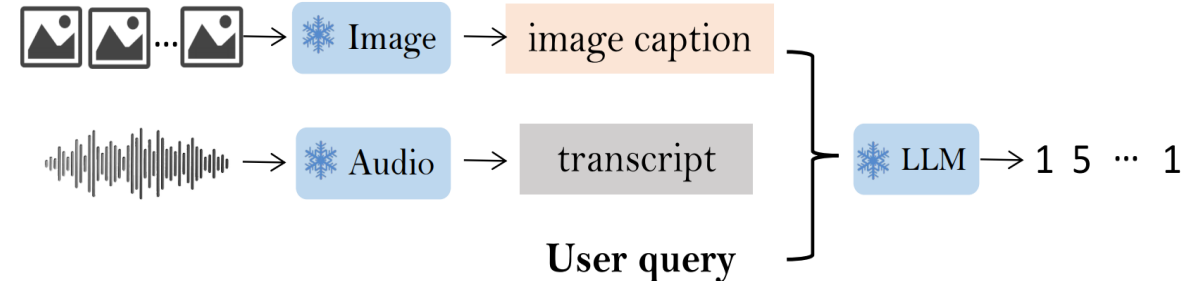
- High generalization capability
- High instruction-following capability

- **Lightweight deployment for efficient processing of video**

- Very useful for long videos such as education videos



(a) conventional language-guided video summarization



(b) our proposed M3Sum

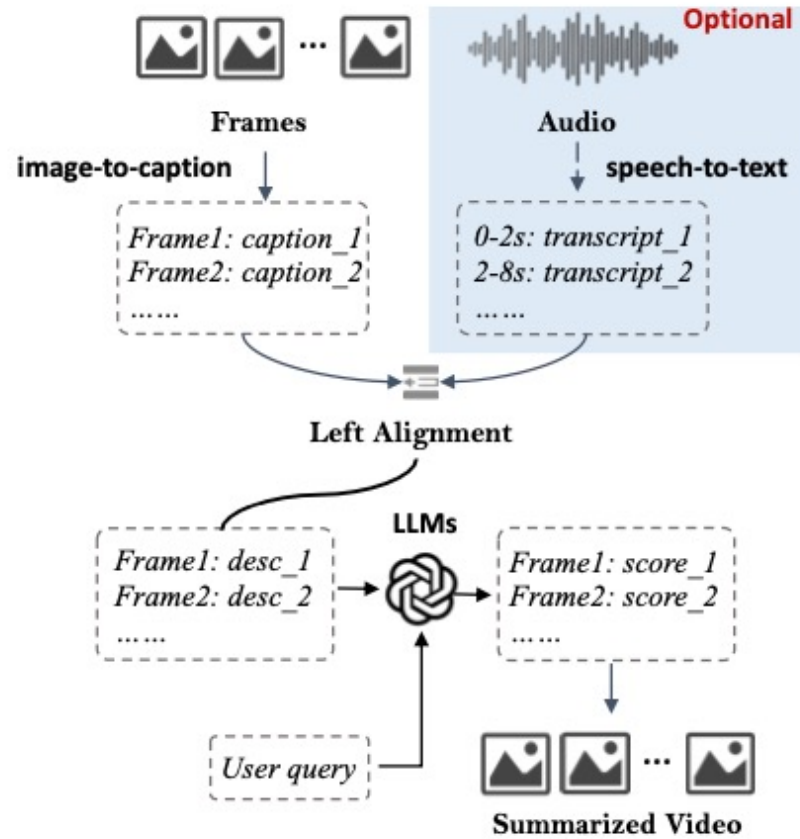


Fig. 2. The framework of our proposed method

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 $D = \{d_i\}_{i=1}^N$.

```

1: for  $i = 1, 2, \dots, L$  do
2:    $s_1 = \frac{t_i^1}{2}, s_2 = \frac{t_i^2}{2}$  /* The frame is sampled every 2
   seconds.*/
3:   for  $j = s_1, s_1 + 1, \dots, s_2$  do
4:      $\tilde{w}_j = w_i$  /*  $\tilde{w}_j$  is the transcript of  $j$ -th frame. */
5:   end for
6: end for
7:  $BS = BertScore(C, T)$ .
   /* The overlap is big and thus no need to fuse.*/
8: if  $BS > \tau$  then
9:   for  $i = 1, 2, \dots, N$  do
10:     $d_i = c_i$ 
11:   end for
12: else
13:   for  $i = 1, 2, \dots, N$  do
14:     $d_i = c_i \cup \tilde{w}_i$ 
15:   end for
16: end if
17: return  $D = \{d_i\}_{i=1}^N$ 

```



- Introduction
- Related Work
- M3SUM
- **Experiments**
- Conclusion and Future Directions

Experiments



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].

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

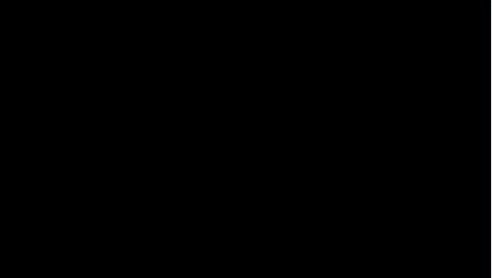

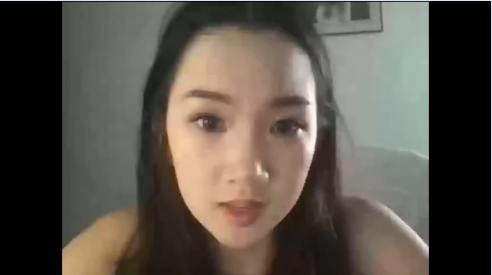

Table 4. The ablation study of M3Sum on the two datasets.

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.

Experiments



Original Video	User Query	Summarization
	Q1: Can you assist me in identifying the bees featured in the video?	
	Q3: What is the primary subject matter introduced in the video?	



- Introduction
- Related Work
- M3SUM
- Experiments
- Conclusion and Future Directions

Conclusion and Future Directions



- We first explore unsupervised language-guided video summarization from a single-modal perspective and achieve on-par performance with previous methods which require lots of training and annotated data.
- **[Multi-modal Question Answering and Other Tasks, Video Editing]** Our method can be extended to different downstream tasks and applications, similar methods include LLaVA, Video-ChatGPT, Video-LLaMA!
- **[Multi-modal Large Language Model]** We hope to train our own multi-modal LLM using the high-quality data generated by our proposed method. We also design some specific modules to capture more features in the video.