Towards Diverse Paragraph Captioning for Untrimmed Videos

Yuqing Song[1], Shizhe Chen[1], Qin Jin

Renmin University of China

{syuqing, qjin}@ruc.edu.cn, cshizhe@gmail.com

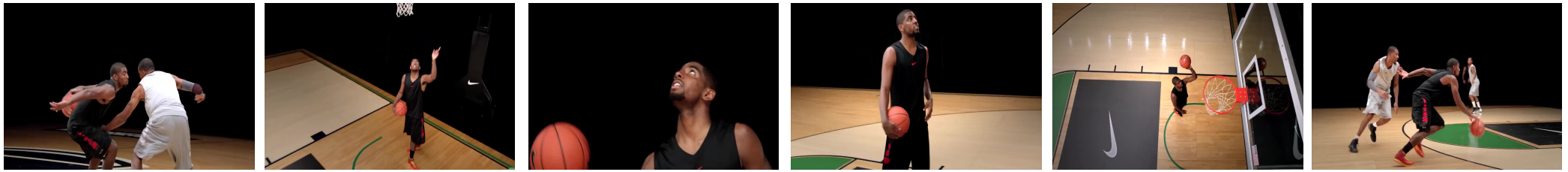
[1] Equal Contribution

Introduction

- Dominant Video Captioning VS. Video Paragraph Captioning
  - Dominant Video Captioning
    - Describe a 10-20s short video in one sentence, which usually contains a single event activity

A basketball player is doing a slam dunk on a basketball hoop.

  - Video Paragraph Captioning
    - Describe a long untrimmed video with a logical paragraph, which contains multiple event activities



A small group of men are seen running around a basketball court playing a game of basketball.

One player moves all around the net holding the ball and demonstrates how to properly shoot a hoop.

He bounces the ball around a bit and more shots of the people playing are shown.

Introduction

- Video Paragraph Captioning
  - Two-stage framework
    - Short events detection
    - Events captioning
  - Drawback of the dominant two-stage framework
    - It is hard to correctly recognize and localize different event activities.
    - Poor detection accuracy greatly hinders the final captioning performance.

How about eschewing the event detection stage and directly generating the paragraph?

Introduction

- Challenges without the event detection stage
  - ➢ Inefficient attention computation
    - The whole video is input to the model
    - Each generation step needs to compute attention weights on all the candidate frames
  - ➢ Redundant event description
    - With limited training examples, it is hard to learn effective attention mechanism with so much attention candidates.
    - The model tends to get stuck in a few salient frames and describes redundant events.
  - ➢ Boring language generation
    - For long text generation, the model prefers to repeat high-frequency words and phrases.

Method

- Challenges without the event detection stage
  - Inefficient attention computation → keyframe-aware video encoder
  - Redundant event description → dynamic memory enhanced decoder
  - Boring language generation → diversity-driven training strategy

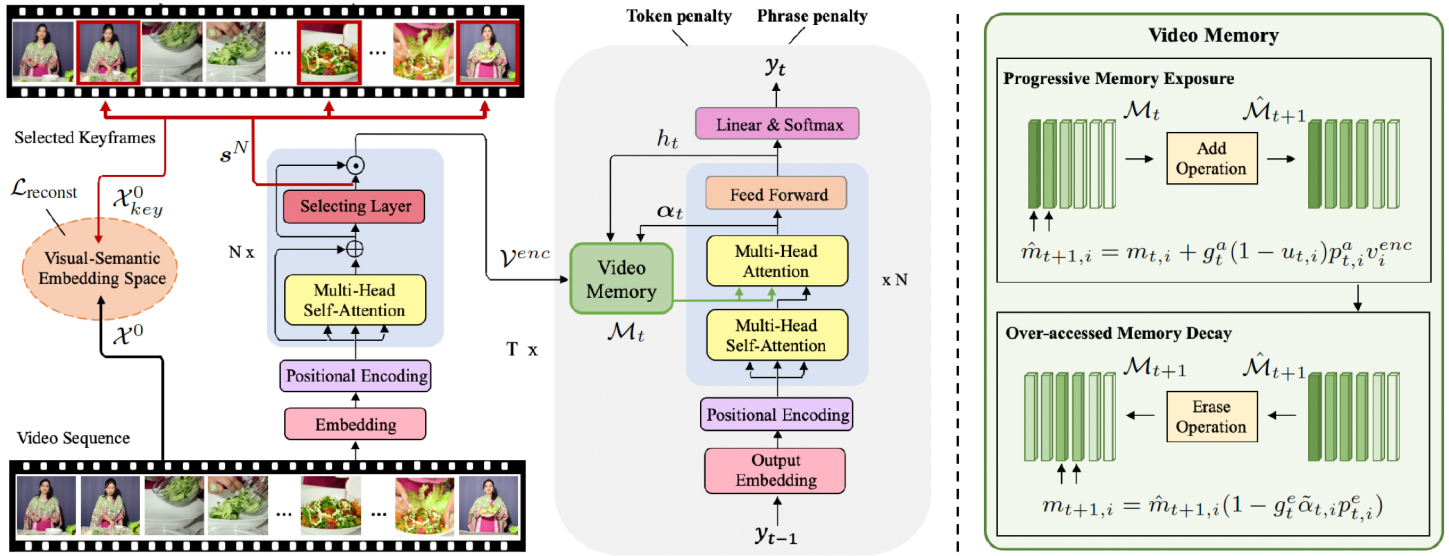


Figure 1. *Left*: The framework of our proposed video paragraph captioning model. *Right*: Details of the proposed dynamic video memories with two updating mechanisms for description coherence and diversity respectively. $\oplus$ denotes addition and $\odot$ denotes hadamard product.

Method

- Challenges without the event detection stage
  - Inefficient attention computation → keyframe-aware video encoder
  - Redundant event description → dynamic memory enhanced decoder
  - Boring language generation → diversity-driven training strategy

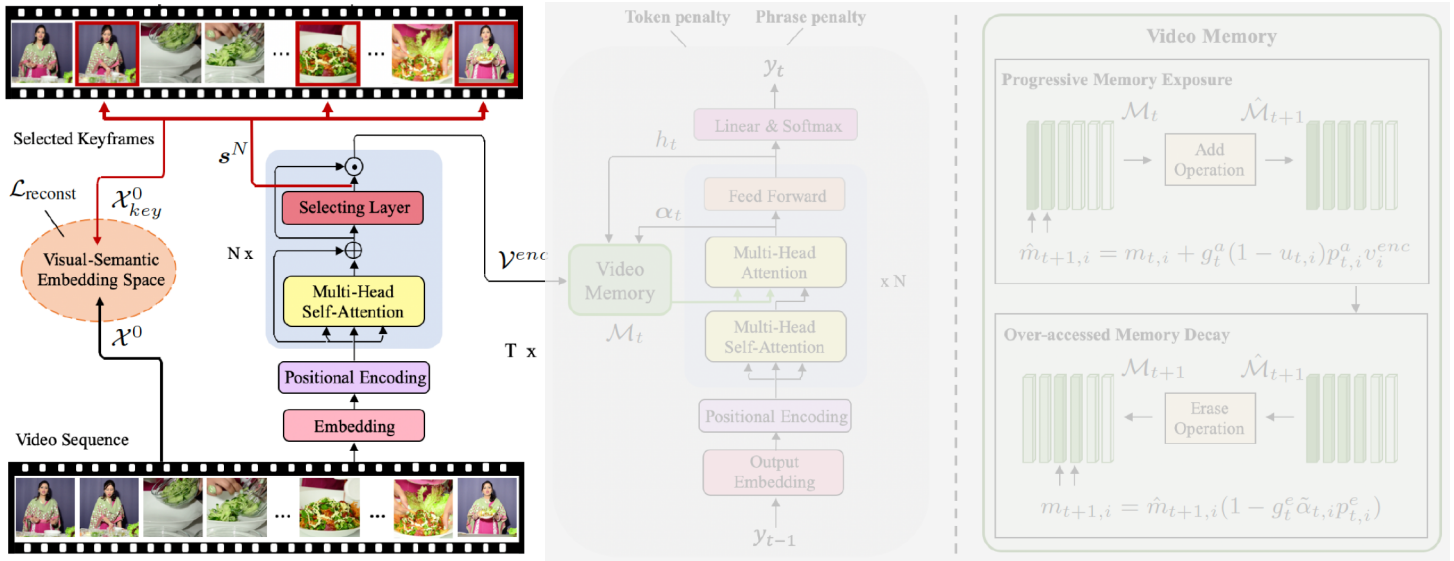


Figure 1. *Left*: The framework of our proposed video paragraph captioning model. *Right*: Details of the proposed dynamic video memories with two updating mechanisms for description coherence and diversity respectively. ⊕ denotes addition and ⊙ denotes hadamard product.

Keyframe-aware video encoder

- Simultaneously encoding the video content and selecting key frames
- Reconstruct the original video with keyframes at the semantic level

$$\widehat{x^i} = x^{i-1} + \text{MultiHead}\left(x^{i-1}, x^{i-1}, x^{i-1}\right)$$

$$s^i = \sigma(\text{FFN}(\widehat{x^i})), \quad x^i = \widehat{x^i} \cdot s^i$$

Method

- Challenges without the event detection stage
  - Inefficient attention computation → keyframe-aware video encoder
  - Redundant event description → dynamic memory enhanced decoder
  - Boring language generation → diversity-driven training strategy

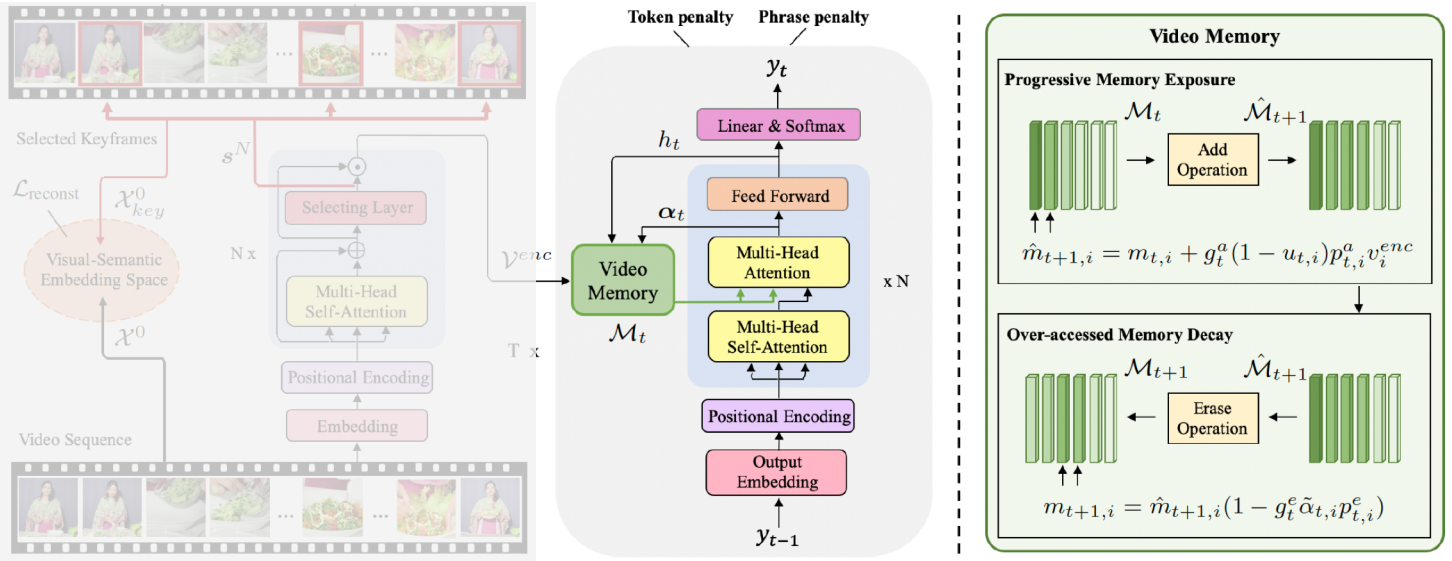


Figure 1. *Left*: The framework of our proposed video paragraph captioning model. *Right*: Details of the proposed dynamic video memories with two updating mechanisms for description coherence and diversity respectively. ⊕ denotes addition and ⊙ denotes hadamard product.

Dynamic Memory Enhanced Decoder

- Progressive memory exposure (coherence)

  gradually **add** new video frames to the memory according to the semantic relevance to the current describing content.

- Over-accessed memory decay (diversity)

  remove highly attended video frames from the memory to force the model to explore new events.

$$\hat{m}_{t+1,i} = m_{t,i} + g_t^a(1-u_{t,i})p_{t,i}^a v_i^{enc}$$
$$m_{t+1,i} = \hat{m}_{t+1,i}(1 - g_t^e \tilde{\partial}_{t,i} p_{t,i}^e)$$

Method

- Challenges without the event detection stage
  - Inefficient attention computation → keyframe-aware video encoder
  - Redundant event description → dynamic memory enhanced decoder
  - Boring language generation → diversity-driven training strategy

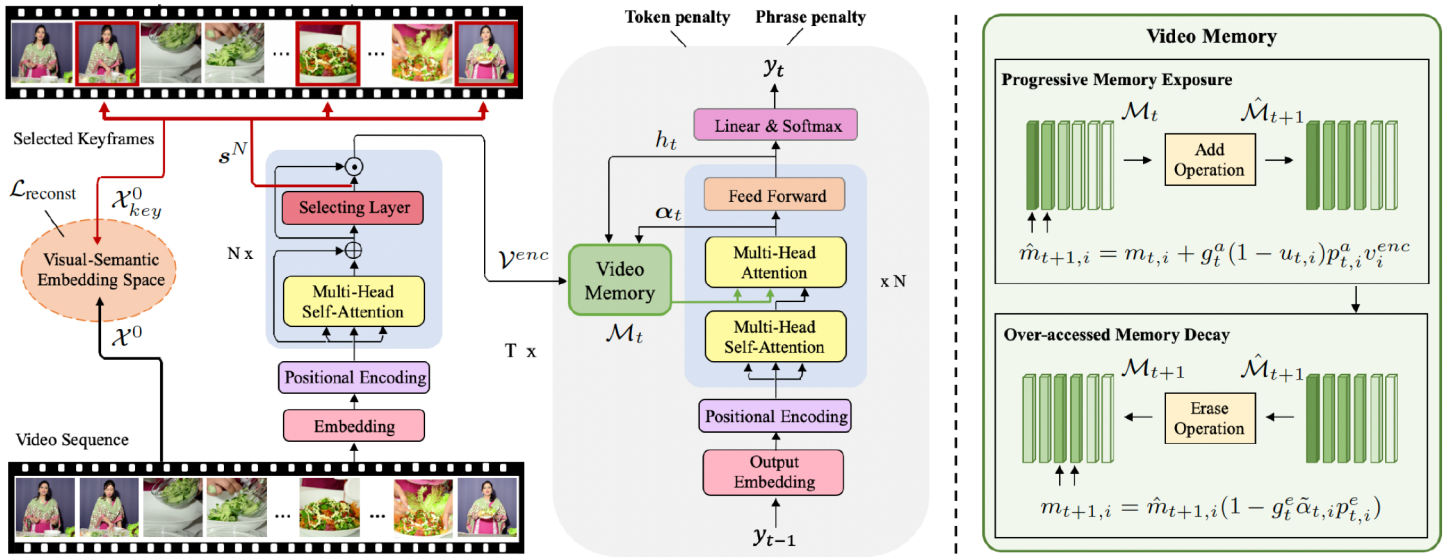


Figure 1. *Left*: The framework of our proposed video paragraph captioning model. *Right*: Details of the proposed dynamic video memories with two updating mechanisms for description coherence and diversity respectively. ⊕ denotes addition and ⊙ denotes hadamard product.

Diversity-driven training strategy

- High frequency word penalty

$$\mathcal{L}_{\text{mle}} = -\frac{1}{T}\sum_{t=1}^{T}(\log p(y_t^*|y_{<t}^*, v) + \sum_{c\in C^t}\log(1-p(c|y_{<t}^*, v)))$$

- High frequency phrase penalty

$$\mathcal{L}_{\text{rl}} = -\frac{1}{T}\sum_{t=1}^{T}(r_{rlv}(y^s) + \beta r_{div}(y_t^s))\log p(y_t^s|y_{<t}^s, v)$$

$$r_{div}(y_t^s) = \frac{1}{k}\sum_{ph\in H_n(y_t^s)}\frac{1}{\text{freq}(ph)}$$

Results

- Datasets
  - ActivityNet Captions dataset: 10,009 train, 2,460 validation, 2,457 test, 5,044 hidden test
  - Charades Captions dataset: 6,963 train, 500 validation, 1,760 test

- Metrics
  - Accuracy: BLEU@4, METEOR, CIDEr
  - Diversity: Div@1, Div@2, Rep@4

Results

- Results on ActivityNet Captions dataset

Table 1. Comparison with state-of-the-art approaches for video paragraph generation on ActivityNet Captions *ae-test* split. "Train" and "Infer" indicate if the video segment annotations are needed at training and inference time.

#	Methods	Segment Annotation		Accuracy			Diversity		
		Train	Infer	BLEU@4	METEOR	CIDEr	Div@1↑	Div@2↑	Rep@4↓
1	MFT [39]	✓	✓	10.33	15.09	19.56	-	-	15.88
2	VTransformer‡ [50]	✓	✓	10.38	16.33	21.05	61.45	77.36	7.42
3	AdvInf‡ [21]	✓	✓	10.89	**17.41**	20.40	60.59	78.29	5.09
4	MART‡ [13]	✓	✓	10.54	17.12	24.14	61.41	77.43	5.32
5	MFT [39]	✓	×	8.45	14.75	14.15	-	-	17.59
6	Vanilla	×	×	11.53	15.91	24.11	64.92	82.34	3.17
7	**Ours**	×	×	**12.20**	16.10	**27.36**	**68.33**	**84.26**	**2.63**
8	Human	-	-	-	-	-	68.60	85.40	0.83

- Our model outperforms the SOTA methods which even use ground-truth event temporal segments on both the accuracy and diversity metrics.
- Our one-stage framework shows superiority compared with the previous two-stage framework.
- We achieve competitive diversity results with the human level.

Results

- Results on ActivityNet Captions dataset

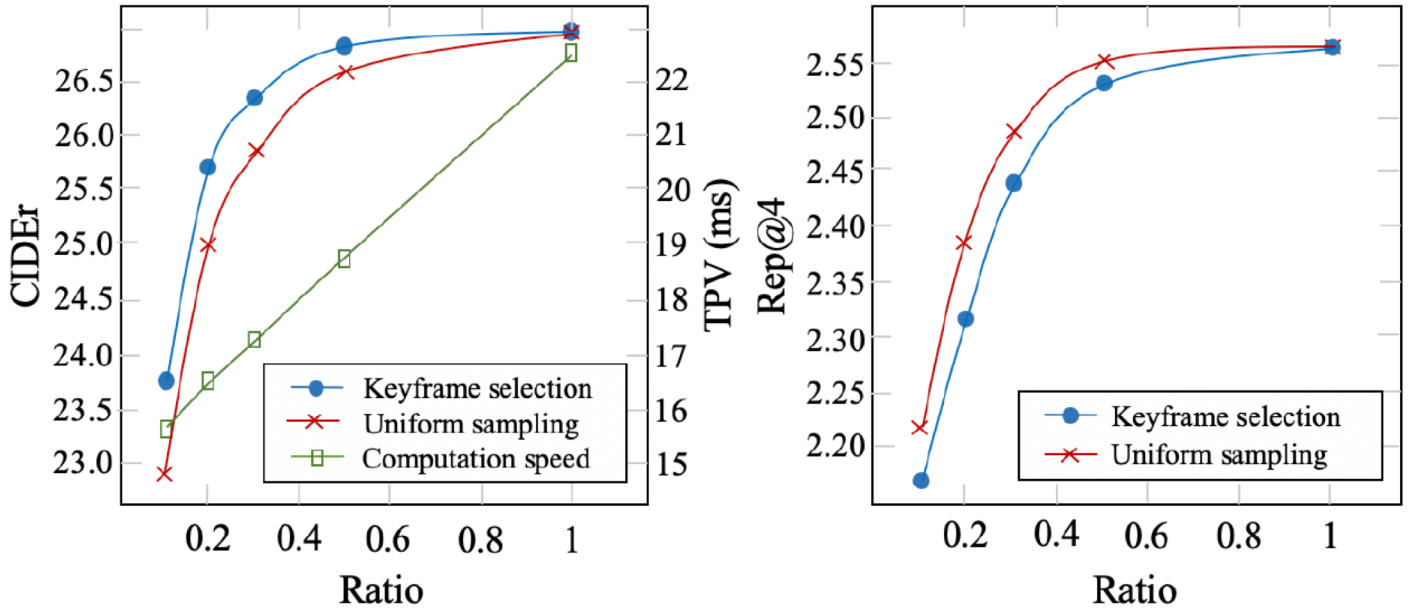
Table 3. Ablation study on ActivityNet *ae-test* set to demonstrate the effectiveness of different components. (pme: progressive memory exposure, omd: over-accessed memory decay, token: token penalty objective, r_{rlv}: relevance reward, r_{div}: phrase penalty objective.)

#	Decoder		MLE	RL		Accuracy			Diversity		
	pme	omd	token	r_{rlv}	r_{div}	BLEU@4	METEOR	CIDEr	Div@1↑	Div@2↑	Rep@4↓
1						11.53	15.91	24.11	64.92	82.34	3.17
2	✓					11.95	15.94	25.52	66.79	82.81	3.39
3		✓				11.91	16.01	24.47	66.18	82.95	2.87
4	✓	✓				11.61	15.72	25.65	67.90	83.37	2.80
5	✓	✓	✓			11.74	15.64	26.55	**68.42**	83.95	2.75
6	✓	✓	✓	✓		12.10	15.85	27.06	67.81	83.45	2.97
7	✓	✓	✓	✓	✓	**12.20**	**16.10**	**27.36**	68.33	**84.26**	**2.63**

- The dynamic memory and diversity-driven training strategy shows their effectiveness in improving the event diversity and language diversity respectively.

Results

- Results on ActivityNet Captions dataset



- With the key frames selection, our model saves half of the decoding time while maintains the captioning performance.

Results

- Results on Charades Captions dataset

Table 2. Captioning results on Charades Captions dataset.

Methods	Accuracy			Diversity		
	B@4	M	C	D@1	D@2	R@4
HRL [36]	18.80	19.50	23.20	-	-	-
Vanilla	19.19	19.80	25.30	72.90	86.13	1.23
Ours	**20.34**	**20.05**	**27.54**	**76.18**	**87.31**	**0.92**
Human	-	-	-	79.90	90.81	0.10

- Our model also achieves the state-of-the-art captioning results on the Charades Captions dataset on both the accuracy metrics and diversity metrics.

Visualization

- Generated Paragraph Visualization

VTransformer (GT events): He starts cooking in the kitchen. A chef is standing at a counter in a kitchen. A man is standing in a kitchen.

AdvInf (GT events): A man is standing in front of a counter while speaking to the camera and leads into him taking a pan and presenting it to the camera. The camera pans around the food and ends by presenting it to the camera. The man takes a sip of the food and begins to stir the pan.

MART (GT events): A man is seen standing behind a table speaking to the camera and begins mixing ingredients into a pan. The man continues to mix ingredients and ends by presenting it to the camera. He continues speaking to the camera and showing off his finished in the end.

Vanilla (no event detection): A man is seen standing behind a counter speaking to the camera and leads into him holding up a food. The man then mixes ingredients into a bowl and finally putting food into a pan.

Ours (no event detection): A woman is seen standing behind a counter and putting various ingredients into a pan. She mixes the ingredients together and ends by spreading it onto a plate.

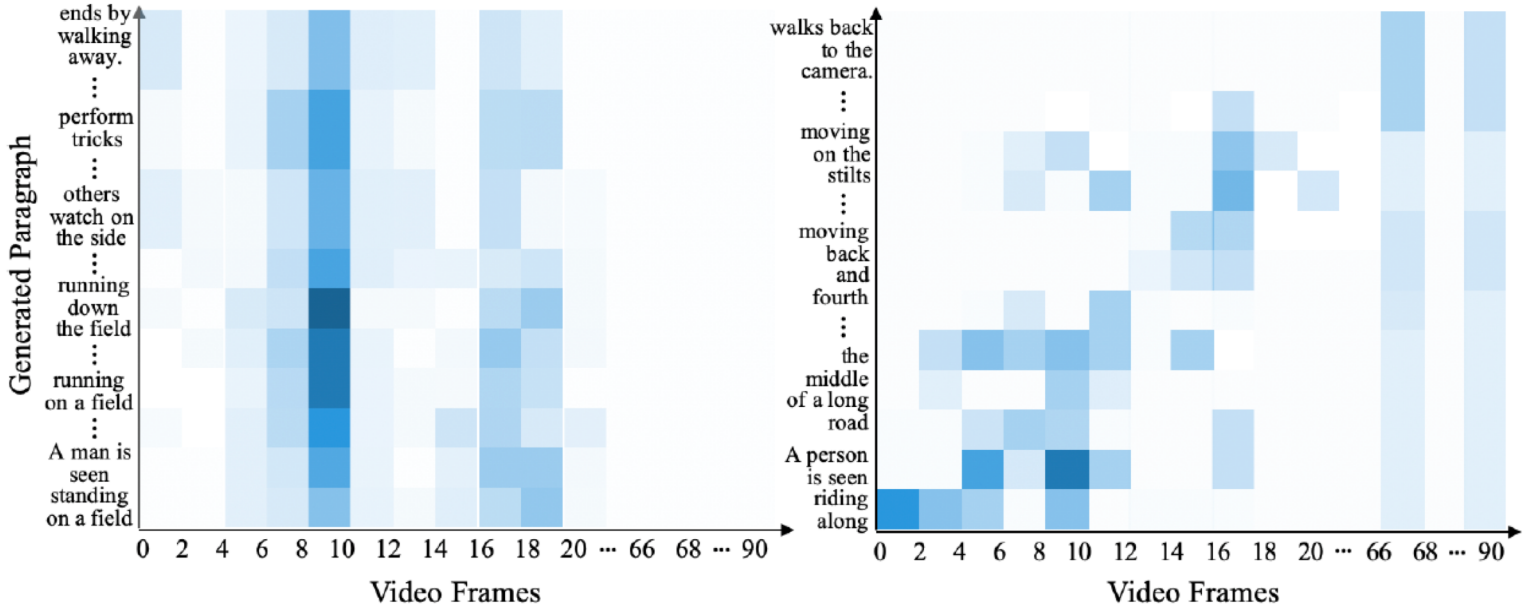
Ground-Truth: A woman is seen cooking items onto a stove with various ingredients laid out. The camera pans around kitchen and shows the woman cooking more ingredients. She continues mixing it around in the pan.

Figure 4. Qualitative examples of the generated paragraphs by our model and other state-of-the-arts methods. The words in red represent high-frequency tokens and phrases which are generated regardless of video content.

- Our model can generate more diverse and coherent paragraph than other methods even without any event temporal annotations.

Visualization

- Attention Map Visualization



- With the dynamic video memory, our model focuses on diverse periods of the video rather than getting stuck in several salient frames.

Summary

- We are the first to eschew the event detection stage to generate paragraphs for untrimmed videos.
- We propose an attention mechanism with dynamic video memories and diversity-driven training objectives for coherent and diverse paragraph generation.
- To improve the decoding efficiency, we further propose a keyframe-aware video encoder to simultaneously encode the video and select keyframes.
- We achieve the state-of-the-art results on two benchmark datasets even without using any event boundary annotations.
- We will release the code and data at https://github.com/syuqings/video-paragraph