Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval

Minkuk Kim[1], Hyeon Bae Kim[1], Jinyoung Moon[2], Jinwoo Choi[1], Seong Tae Kim[1]

[1]Kyung Hee Univerity, [2]Electronics and Telecommunications Research Institute (ETRI)

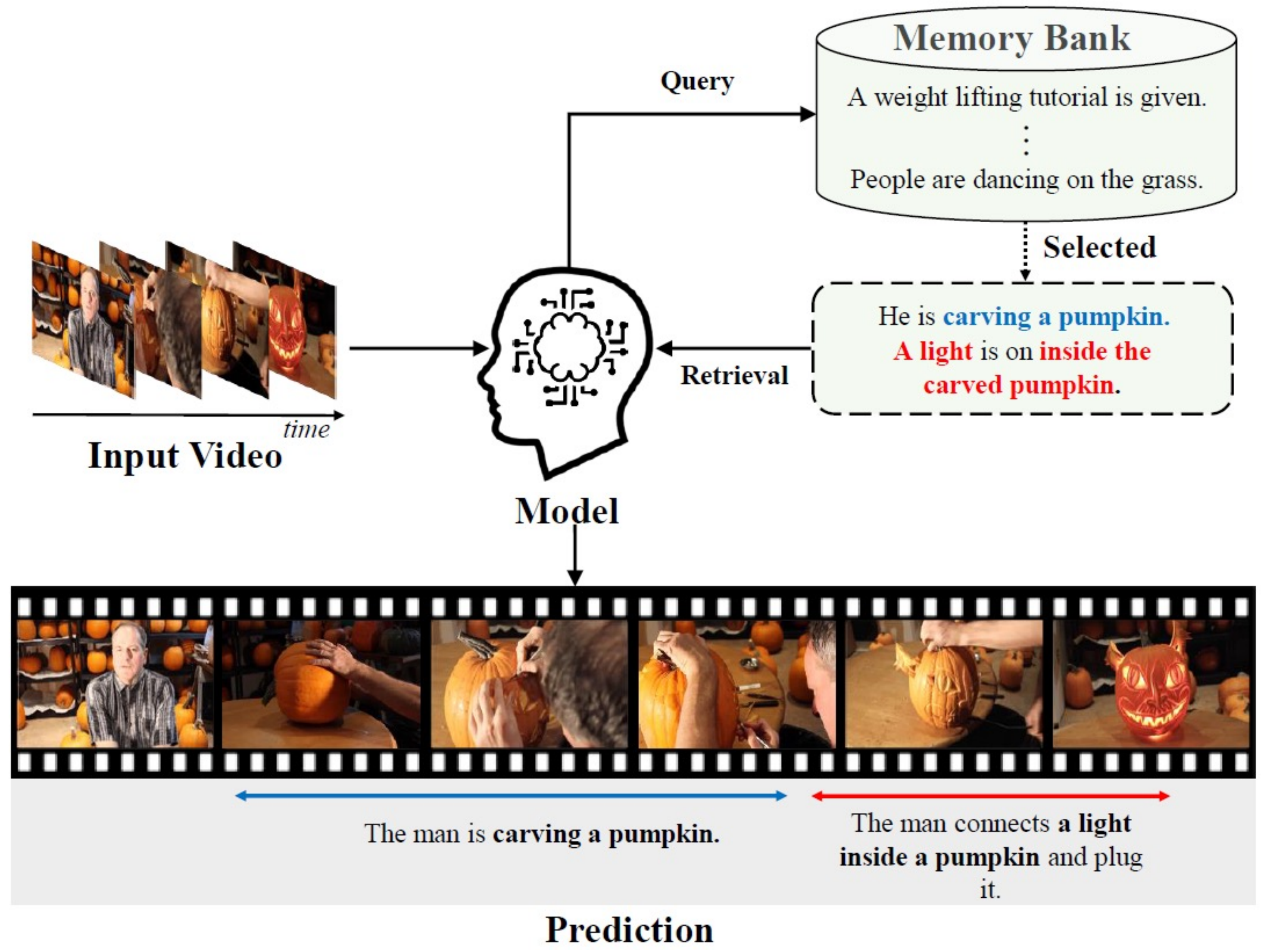
CVPR SEATTLE, WA JUNE 17-21, 2024

At One Glance

We address challenging Video Localization and Description tasks by proposing a novel framework based on

" How Humans Recognize,
Remember and Recall. "

- Concept of our proposed method


Memory Bank
A weight lifting tutorial is given.
⋮
People are dancing on the grass.
Query
Selected
He is carving a pumpkin.
A light is on inside the carved pumpkin.
Retrieval
time
Input Video
Model
The man is carving a pumpkin.
The man connects a light inside a pumpkin and plug it.
Prediction

Our Contributions:

1. Inspired by the human cognitive process, we introduce a new dense video captioning method with cross-modal retrieval from external memory.

2. We propose a versatile encoder-decoder structure that can learn cross-modal correlation and inter-task interactions.

- Effect of Memory Retrieval in YouCook2

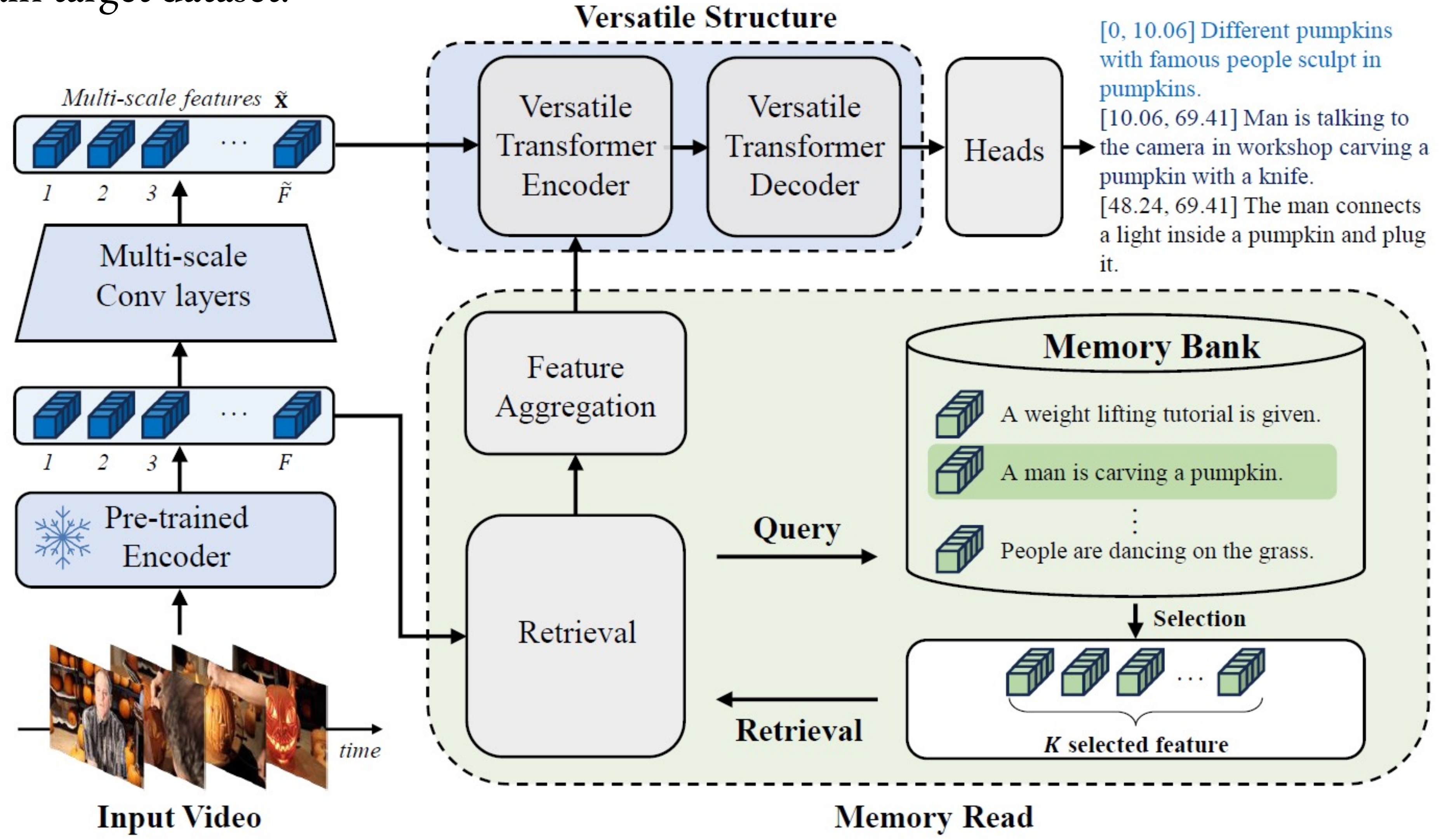
Retrieval Type	YouCook2			
	CIDEr	METEOR	BLEU4	SODA_c
No Retrieval	23.67	5.30	1.17	4.77
Proposed Retrieval (Ours)	31.66	6.08	1.63	5.34
Oracle w/o GT proposal	53.55	9.18	3.49	6.81
Oracle w/ GT proposal	183.95	23.53	13.05	25.51

Dense Video Captioning with *Cross-Modal Memory* Retrieval(CM^2)

- We focus how to improve event localization and captioning from untrimmed video with prior knowledge memory bank.
- For this, we propose two sections

1) Cross-Modal Retrieval for semantic clues and 2) Leverage them with the Versatile Structure

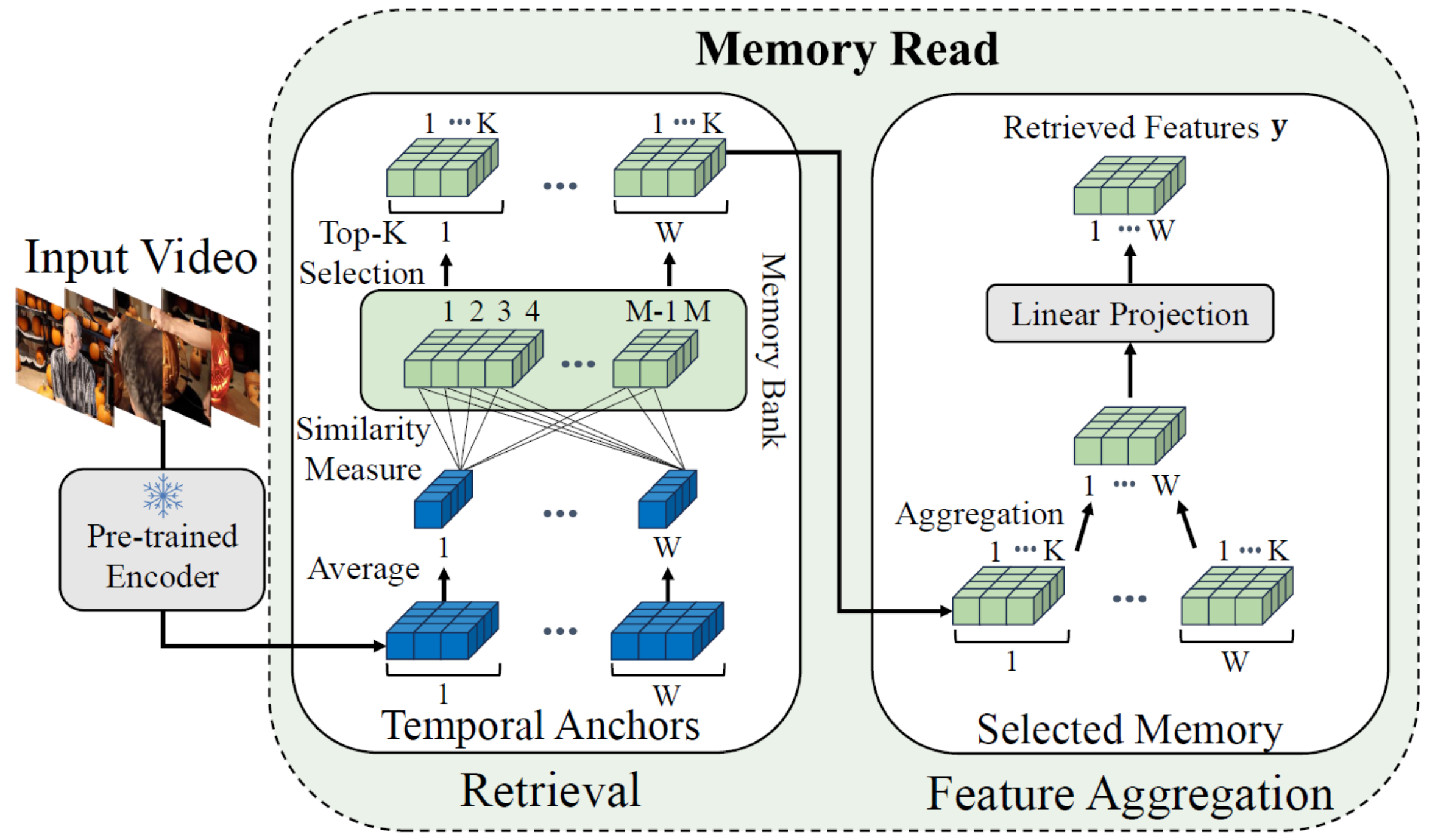
- We construct an external memory by encoding sentence features from the training data of in-domain target dataset.



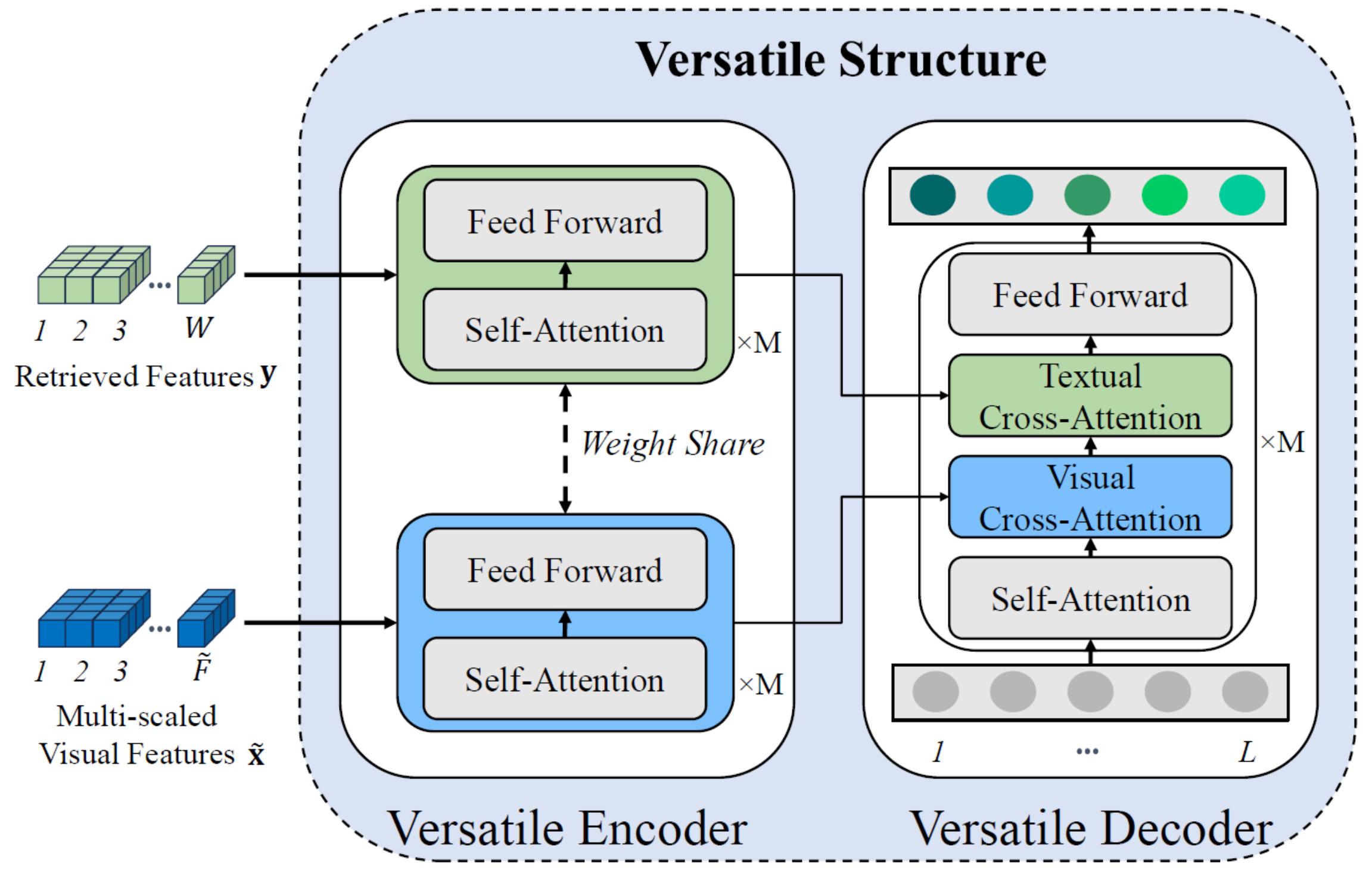
[0, 10.06] Different pumpkins with famous people sculpt in pumpkins.
[10.06, 69.41] Man is talking to the camera in workshop carving a pumpkin with a knife.
[48.24, 69.41] The man connects a light inside a pumpkin and plug it.

(a) Overall Architecture.

- We divide the video into W temporal anchors and retrieve for each anchor, then aggregate.
- To leverages the retrieved semantic information for both localization and captioning tasks, we design a versatile encoder-decoder architecture and a modal-level cross-attention method.



(b) Memory Read Module.



(c) Versatile Encoder-Decoder Module.

Experimental Results

- Event Captioning Performance on ActivityNet Captions and YouCook2

Method	Backbone	ActivityNet Captions				YouCook2			
		PT	CIDEr	METEOR	SODA_c	PT	CIDEr	METEOR	SODA_c
Vid2Seq [48]	CLIP	15M	30.10	8.50	5.80	1M	**47.10**	**9.30**	**7.90**
MT [54]	TSN	✗	6.10	3.20	-	✗	9.30	5.00	-
ECHR [45]	C3D	✗	14.70	7.20	3.20	✗	-	3.82	-
PDVC† [46]	CLIP	✗	29.97	8.06	5.92	✗	29.69	5.56	4.92
Ours	CLIP	✗	**33.01**	**8.55**	**6.18**	✗	31.66	6.08	5.34

- Event Localization Performance on ActivityNet Captions and YouCook2

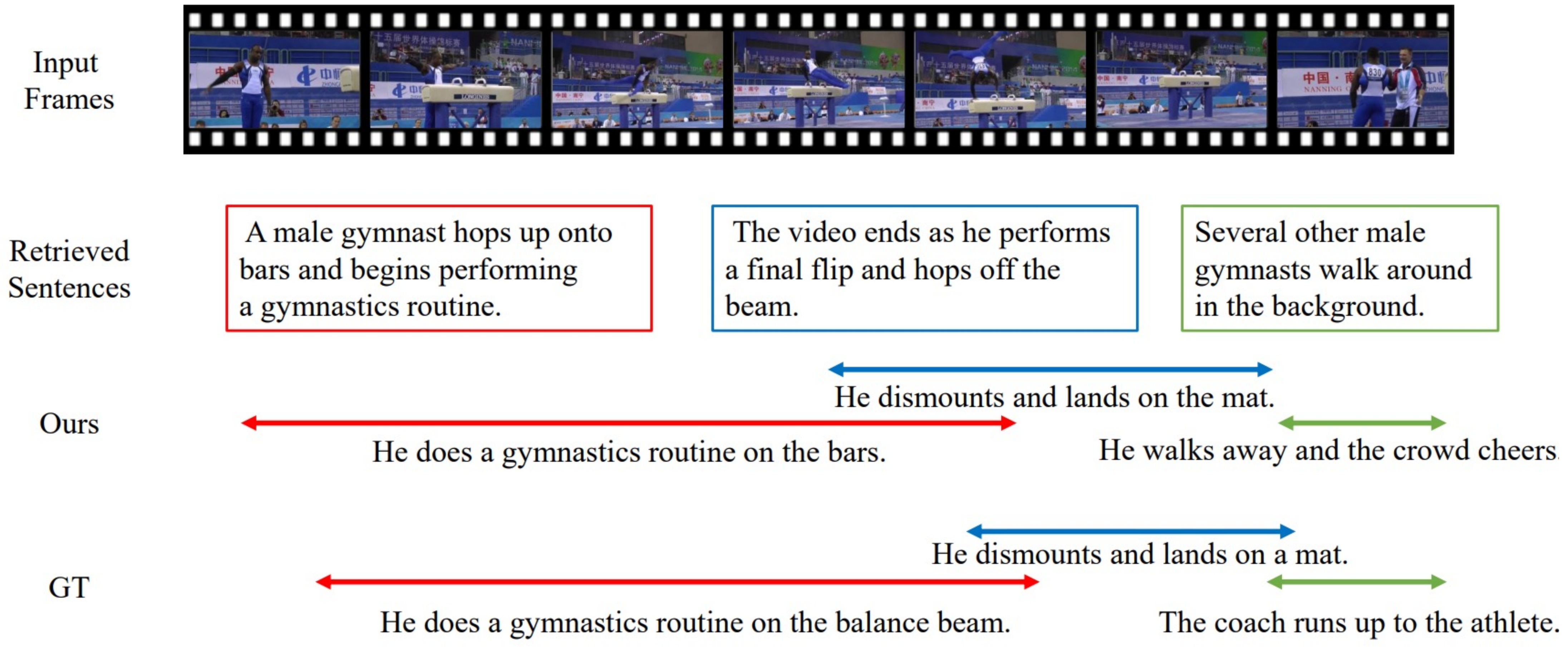
Method	ActivityNet Captions				YouCook2			
	PT	F1	Recall	Precision	PT	F1	Recall	Precision
Vid2Seq [48]	15M	53.29	52.70	53.90	1M	27.84	**27.90**	27.80
PDVC† [46]	✗	54.78	53.27	56.38	✗	26.81	22.89	32.37
Ours	✗	**55.21**	**53.71**	**56.81**	✗	**28.43**	24.76	**33.38**

Code

Paper

✓ Our model achieves comparable performance without pretraining on large video datasets.

- Qualitative Results



A male gymnast hops up onto bars and begins performing a gymnastics routine.

The video ends as he performs a final flip and hops off the beam.

Several other male gymnasts walk around in the background.

Ours: He dismounts and lands on the mat. He does a gymnastics routine on the bars. He walks away and the crowd cheers

GT: He dismounts and lands on a mat. He does a gymnastics routine on the balance beam. The coach runs up to the athlete.

✓ It can be observed that memory retrieval effectively references meaningful and helpful sentences from memory for each event.

✓ As a result, our model generates relatively accurate event boundaries and captions.