KYUNG HEE UNIVERSITY

HiCM^2 : Hierarchical Compact Memory Modeling for Dense Video Captioning

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)

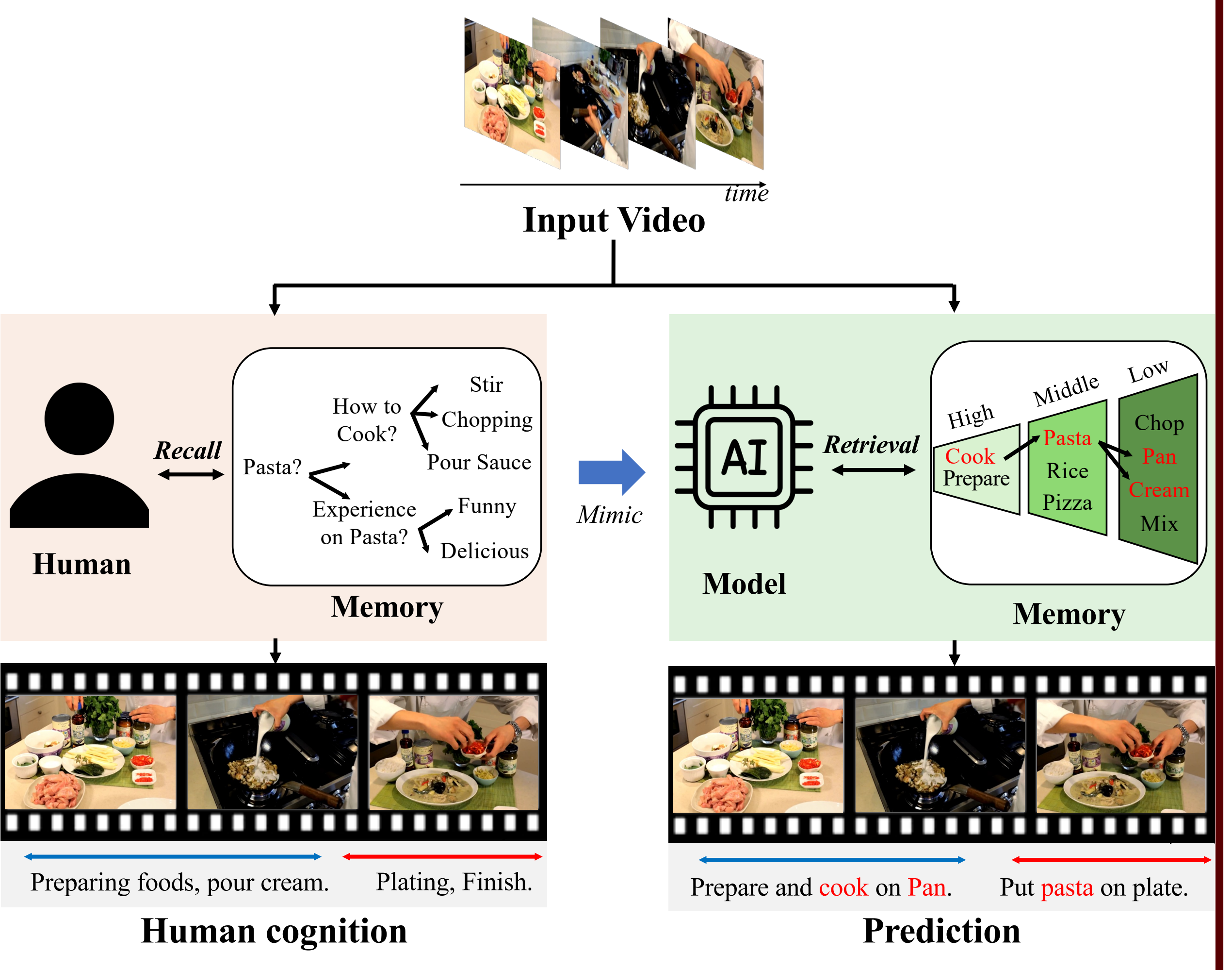
AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

At One Glance

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

" How human Recognizes, Remembers and Recalls in Memory Hierarchy. "

- Concept of our proposed method



Our contributions:

1. We introduce the first hierarchical memory structure for DVC, inspired by human cognition and enabling cross-modal retrieval with compact representations.

2. We propose a top-down hierarchical memory retrieval strategy, starting with abstract information and progressively accessing detailed levels.

3. Extensive experiments on YouCook2 and ViTT show that our approach achieves state-of-the-art performance, validating its effectiveness.

- Effect of Memory Retrieval in YouCook2

Memory Construction	Hierarachy	CIDEr	METEOR	SODA_c	F1
No Memory	✗	66.29	12.41	9.87	31.08
All Training Captions	✗	67.90	12.49	10.38	32.31
Clustering	✓	67.15	**12.97**	10.17	32.30
Clustering+LLM(Ours)	✓	**71.84**	12.80	**10.73**	**32.51**

✓ Our observations indicate that retrieving relevant content from memory benefits model performance, especially under a hierarchical, compact memory configuration.

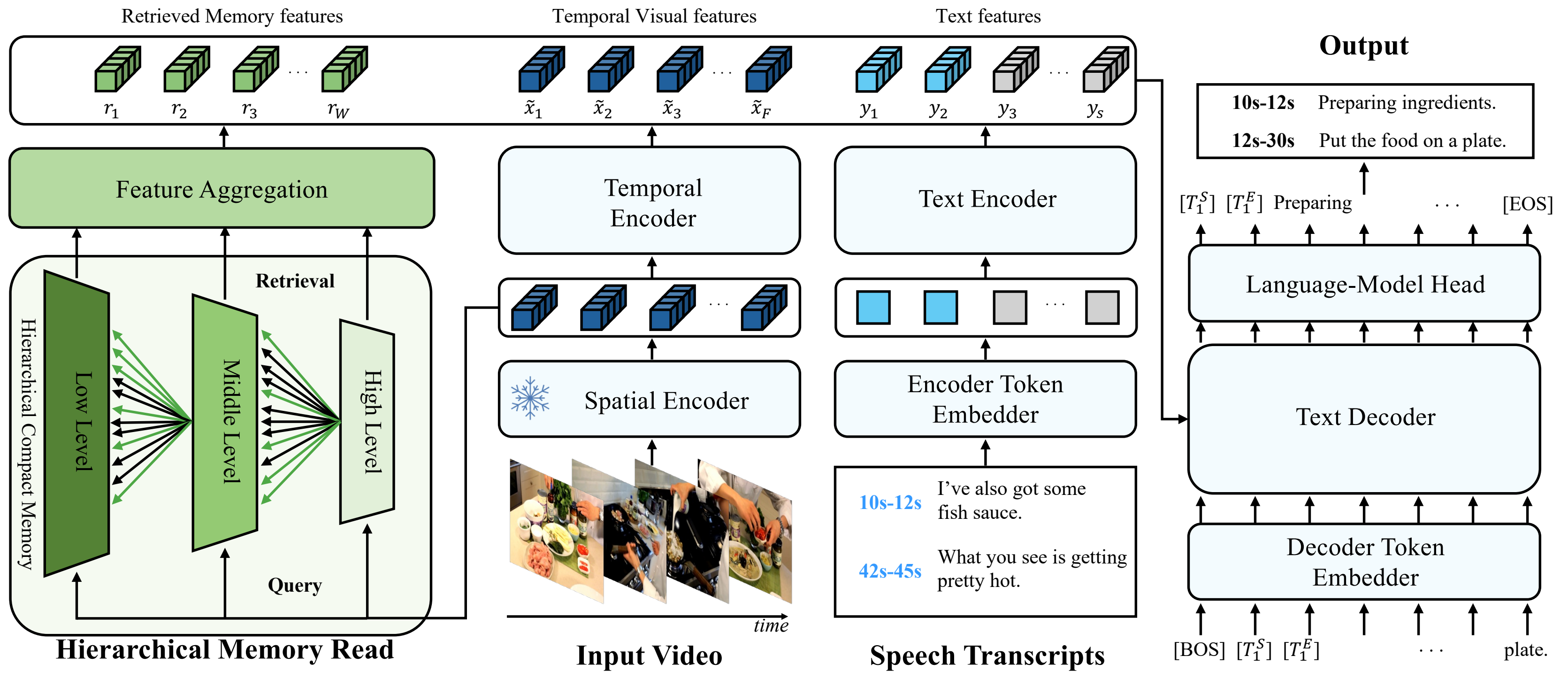
Paper   Github   Personal page

Hierarchical Compact Memory Modeling for Dense Video Captioning

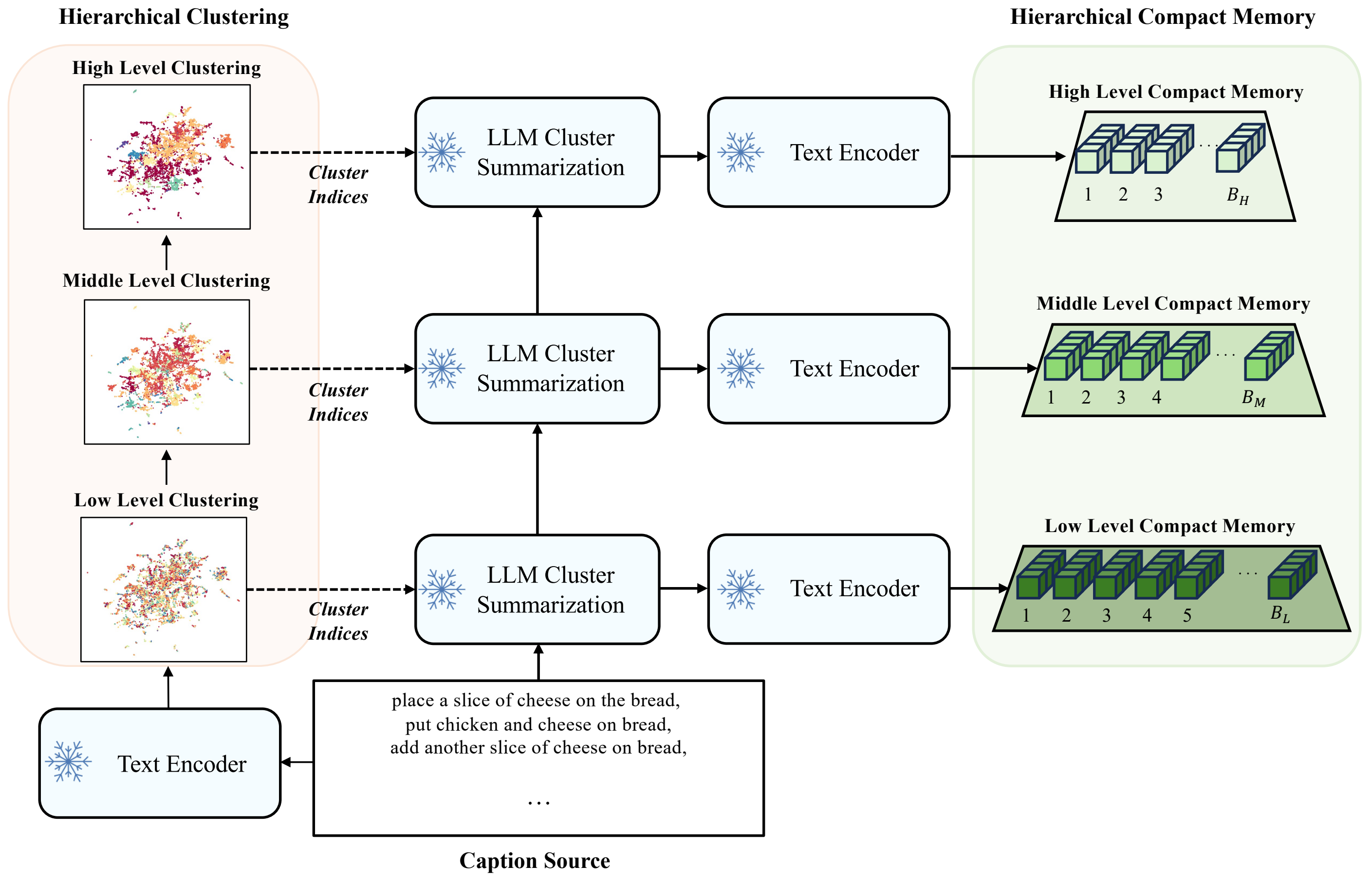
- We focus how to improve event localization and captioning from untrimmed video with prior knowledge memory bank.

- For this, we propose two sections :

1) How to model Hierarchical Compact Memory and 2) How to hierarchically retrieve?



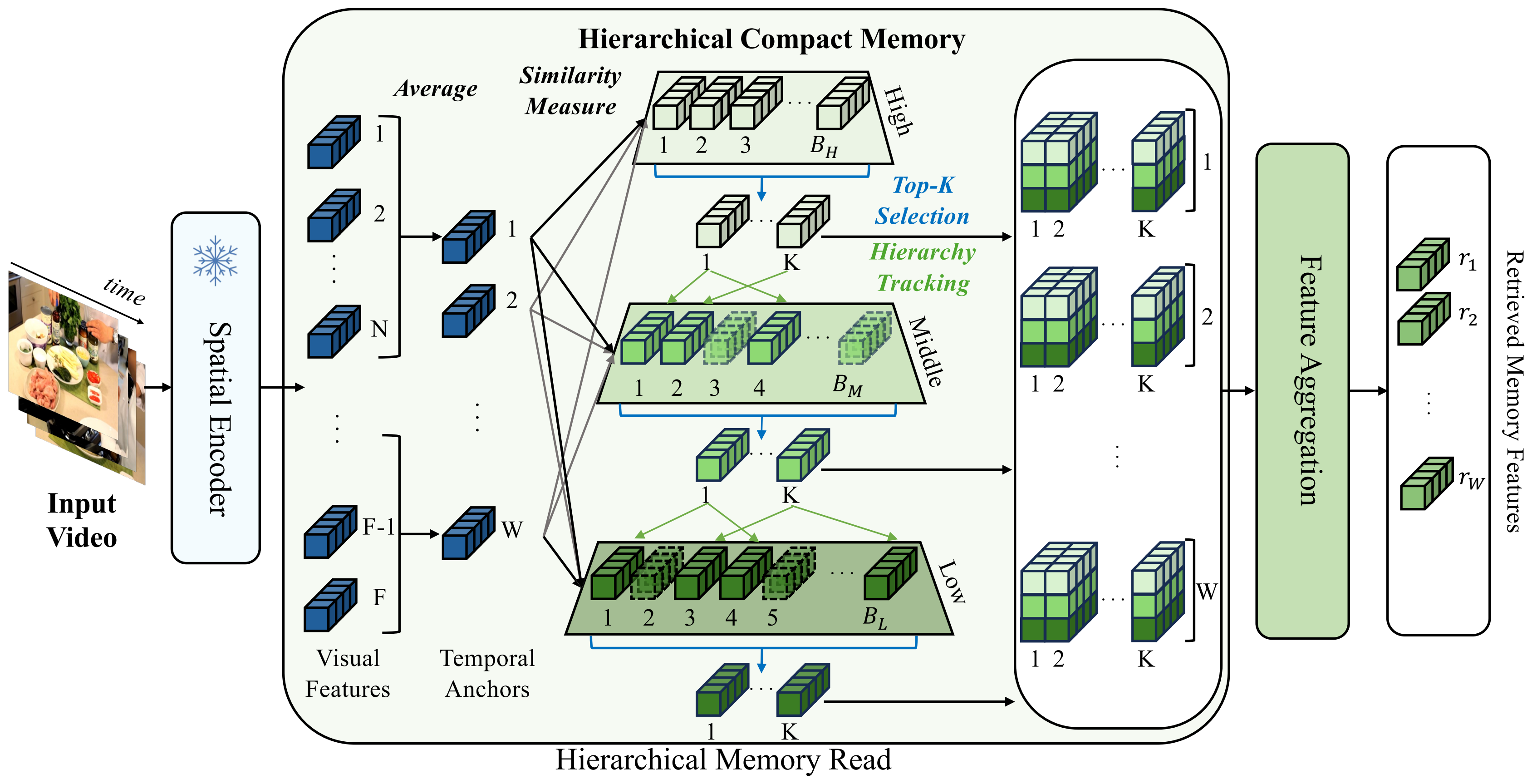
How to model Hierarchical Compact Memory?


Hierarchical Clustering
High Level Clustering
Middle Level Clustering
Low Level Clustering
Cluster Indices
Cluster Indices
Cluster Indices
LLM Cluster Summarization
LLM Cluster Summarization
LLM Cluster Summarization
Text Encoder
Text Encoder
Text Encoder
Text Encoder
Hierarchical Compact Memory
High Level Compact Memory
1 2 3 B_H
Middle Level Compact Memory
1 2 3 4 B_M
Low Level Compact Memory
1 2 3 4 5 B_L
place a slice of cheese on the bread,
put chicken and cheese on bread,
add another slice of cheese on bread,
...
Caption Source

- We iteratively cluster the caption source to create hierarchical results. In the first low level, we generate representative summaries for each cluster with clustered indices, and higher levels re-summarize these for compact memory components.

- This approach efficiently recalls abstract concepts and detailed episodes, minimizing redundancy while preserving semantic relevance.

How to hierarchically retrieve?



- We divide the video into W temporal anchors and retrieve for each anchor, then aggregate.

- We perform top-down search to efficiently retrieve, starting from abstract information and progressively retrieving connected detailed information. In here, we repeatedly select top-k and track hierarchy across all levels for efficient retrieval.

Experimental Results

- Event Captioning Performance on YouCook2 and ViTT

Method	PT	YouCook2(val)				ViTT(test)			
		CIDEr	METEOR	SODA_c	BLEU4	CIDEr	METEOR	SODA_c	BLEU4
PDVC	✗	29.69	5.56	4.92	1.40	-	-	-	-
CM^2	✗	31.66	6.08	5.34	1.63	-	-	-	-
Streaming V2S	✓	32.90	7.10	6.00	-	25.2	5.80	10.00	-
DIBS	✓	44.44	7.51	6.39	-	-	-	-	-
Vid2Seq†	✓	66.29	12.41	9.87	5.64	48.84	9.51	14.99	0.71
HiCM^2(Ours)	✓	**71.84**	**12.80**	**10.73**	**6.11**	**51.29**	**9.66**	**15.07**	**0.86**

- Event Localization Performance on YouCook2 and ViTT

Method	PT	YouCook2(val)			ViTT(test)		
		F1	Recall	Precision	F1	Recall	Precision
PDVC	✗	26.81	22.89	32.37	-	-	-
CM^2	✗	28.43	24.76	33.38	-	-	-
Streaming V2S	✓	24.10	-	-	35.40	-	-
DIBS	✓	31.43	26.24	**39.81**	-	-	-
Vid2Seq†	✓	31.08	30.38	31.81	**46.21**	**45.89**	46.53
HiCM^2(Ours)	✓	**32.51**	**32.51**	32.51	45.98	45.00	**47.00**

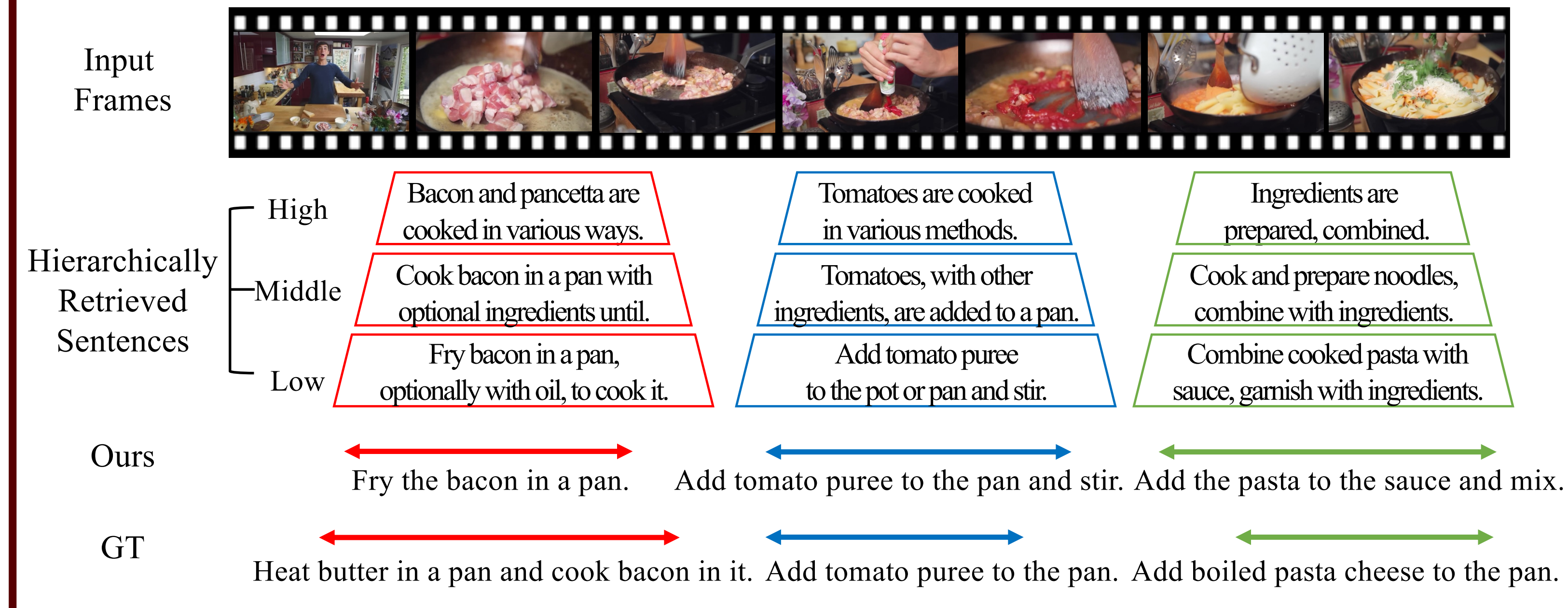
✓ Our model achieves State-of-the-art performance with Hierarchical Compact Memory.

- Ablation study for the use of hierarchical levels memory on YouCook2.

High	Middle	Low	CIDEr	METEOR	SODA_c	F1
			66.29	12.41	9.87	31.08
		✓	67.75	12.33	10.35	32.27
	✓		67.59	12.45	10.37	32.23
✓			67.21	12.21	10.28	31.98
	✓	✓	68.61	12.35	10.51	32.07
✓		✓	69.81	12.63	10.54	**33.00**
✓	✓		68.05	12.33	10.41	32.87
✓	✓	✓	**71.84**	**12.80**	**10.73**	32.51

✓ We showed that employing a comprehensive hierarchical memory yields complementary information and superior performance.

- Qualitative Results


Input Frames
Hierarchically Retrieved Sentences
High
Middle
Low
Bacon and pancetta are cooked in various ways.
Cook bacon in a pan with optional ingredients until.
Fry bacon in a pan, optionally with oil, to cook it.
Tomatoes are cooked in various methods.
Tomatoes, with other ingredients, are added to a pan.
Add tomato puree to the pot or pan and stir.
Ingredients are prepared, combined.
Cook and prepare noodles, combine with ingredients.
Combine cooked pasta with sauce, garnish with ingredients.
Ours
Fry the bacon in a pan.
Add tomato puree to the pan and stir.
Add the pasta to the sauce and mix.
GT
Heat butter in a pan and cook bacon in it.
Add tomato puree to the pan.
Add boiled pasta cheese to the pan.

✓ It can be observed that effectively retrieving meaningful and relevant memory references across multiple levels, from abstract concepts to detailed information, yields significant benefits.

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

✓ Our findings suggest that the synergy between pre-trained prior knowledge and retrieval-augmented knowledge could complement existing pretraining efforts, potentially contributing to further improvement in the field.