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

Supplementary Material

Table 8. **Ablation study to verify the effect of cross-attention design in the versatile decoder.** VC→TC denotes the structure where textual cross-attention follows after visual cross-attention (Figure 2 (c)). TC→VC denotes the structure that is implemented by making visual cross-attention follow after textual cross-attention. Parallel cross-attention is implemented by merging each cross-attention output. The performance is measured in YouCook2.

Cross-Attention Design	CIDEr	METEOR	SODA_c	F1
Sequential (VC→TC)	31.66	**6.08**	**5.34**	**28.43**
Sequential (TC→VC)	**33.33**	5.77	5.14	27.03
Parallel cross-attention	30.63	5.55	5.13	26.96

Table 9. **Analysis of different aggregation methods of memory read module in YouCook2 dataset.**

Aggregation Type	CIDEr	METEOR	BLEU4	SODA_c	F1
Attention	32.21	5.84	1.90	5.19	27.05
Average Pooling	31.66	6.08	1.63	5.34	28.40

A. Additional Analysis of CM^2

Cross-attention Design. We explore the effect of the order of constructing cross-attention modules in the versatile decoder on the model's localization and caption generation abilities. VC→TC is the method used in this study where visual features are incorporated first with visual cross-attention and textual cross-attention follows after visual cross-attention. TC→VC is designed by making temporal cross-attention first, then visual cross-attention follows. In Table 8, it is observed that VC→TC shows better performance compared to TC→VC on METEOR, SODA_c, and F1. We also conducted an experiment of the Parallel cross-attention. Parallel cross-attention is implemented by conducting cross-attention separately for VC and TC, and then merging them. Compared to parallel cross-attention, sequential cross-attention demonstrates better performance in both localization and captioning tasks overall.

Effect of Selected Features Aggregation Methods. In Table 9, we explore methods for aggregating retrieved text features as segment-level semantic information. We compare two different aggregation methods. The attention method shows fairly comparable results in event captioning but exhibited relatively lower performance in event localization. On average pooling method, as outlined in Section 3.1, showed consistently comparable performance in both event captioning and event localization.

Memory Bank Size. We further analyze our model with respect to the memory bank size. To investigate the effect of memory size on our model, we randomly sample text features to construct external memory. For example, the memory size 10% is implemented by sampling 10% of training data captions to construct the external memory, and the inference is conducted with the reduced memory. To reduce the effect of randomness, we report average scores calculated from 50 repetitions with different memory sampling, excluding cases where the memory size ratio is 0% or 100%. We investigate the effect of memory size on a log scale. As shown in Table 10, as the amount of information in the memory increases, CIDEr and METEOR improve. With a small memory size, our method could achieve higher performance compared with the model with no memory (no memory is implemented by replacing text features in the external memory with zero features).

Table 10. **Analysis of our model with respect to the memory bank size in YouCook2 dataset.** The performance is measured by changing the memory bank size. The memory ratio means that we randomly sample the sentence features from the training data to construct text features in the external memory. The average scores are calculated from 50 repetitions with different memory sampling

Memory Ratio	CIDEr	METEOR	BLEU4	SODA_c
0%	27.91	5.66	1.24	4.92
0.01%	29.59	5.74	1.53	5.09
0.1%	30.70	5.75	1.67	5.33
1%	31.17	5.79	1.69	5.39
10%	31.34	5.92	1.67	5.37
100%	31.66	6.08	1.63	5.34

Table 11. **Effect of memory bank construction.** YC2, ANet, Epic, MSR denote YouCook2, ActivityNet captions, Epic Kitchens, and MSR-VTT datasets. For memory bank, only training set is used.

Test	Memory Bank	CIDEr	METEOR	SODA_c	F1
YC2	YC2	31.66	6.08	5.34	28.43
	YC2+ANet	31.66	6.01	5.38	28.53
	Epic	31.58	5.91	5.36	28.51
	YC2+Epic	32.21	6.08	5.39	28.50
ANet	ANet	33.01	8.55	6.18	55.21
	ANet+YC2	33.13	8.57	6.19	55.18
	MSR	33.09	8.55	6.12	54.94
	ANet+MSR	33.34	8.60	6.19	55.18

Scalability of Memory Bank. We show additional results in Table 11 below when the memory is built by combining training sets of YouCook2 and ActivityNet Captions as a unified external knowledge. The unified memory still shows comparable performance. Moreover, we further present experimental results of constructing the memory with another

Table 12. **Performance of Paragraph Captioning in ActivityNet.** Bold means the highest score. Underline means 2nd score. # PT denotes the number of videos used for pre-training. † denotes results reproduced from official implementation in our environment.

Method	Backbone	#PT	ActivityNet (val-ae) CIDEr	METEOR
Vid2Seq[48]	CLIP	15M	**28.00**	**17.00**
PDVC[46]	TSN	-	20.50	15.80
PDVC†[46]	CLIP	-	23.74	16.03
Ours	CLIP	-	<u>25.31</u>	<u>16.47</u>

dataset. When the memory is built with the dataset from the same domain (e.g., YC2+Epic for YC2, ANet+MSR for ANet), we observe performance improvement without additional training. Note that our method builds the memory with CLIP text features from caption datasets, which increases the scalability of the method.

B. Performance of Paragraph Captioning

We further present the results of our model in terms of paragraph captioning. Note that any additional training is not conducted for paragraph captioning. We just measure the performance of our model by collecting generated captions in order and calculating the performance for the query video at a paragraph level. Table 12 shows the results of the models at a paragraph level. As shown in the table, Vid2seq [48], which utilizes an additional 15 million videos for pre-training, achieves the best performance. Our method shows comparable performance without pre-train with extra videos. In our future work, we plan to enhance paragraph generation by incorporating optimized sentence retrieval and training schemes specifically tailored for paragraph generation.

C. Qualitative Results

In Figure 4 and Figure 5, we show additional qualitative examples of our approach. As shown in the figures, memory retrieval could provide relevant semantics for analyzing input query video. As a result, our approach could yield precise event boundaries and captions. The semantic information retrieved from memory assists in semantic predictions during the caption generation process as shown Section 4.

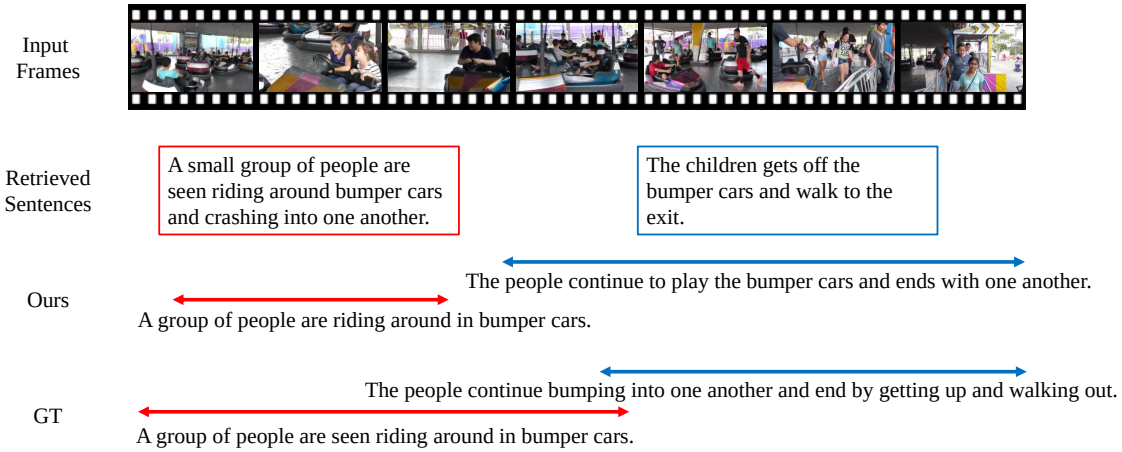

Figure 4. **Example of predictions from our method on ActivityNet Captions dataset.** We show a comparison with the ground truth. Retrieved sentences are example results from retrieval that have the highest similarity to the corresponding segments of input frames. Each retrieved sentence is utilized in our model's predictions for the segments with the corresponding color.

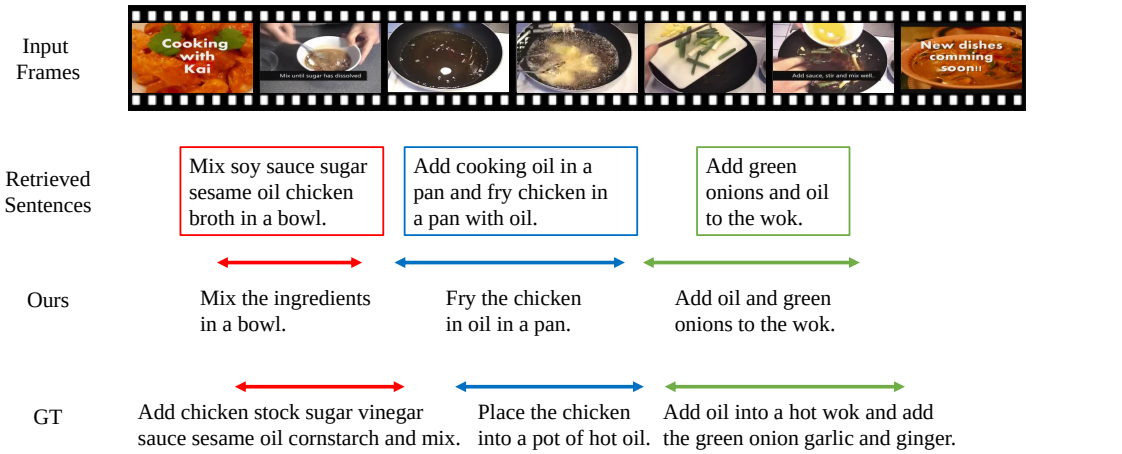

Figure 5. **Example of predictions from our method on YouCook2 dataset.** We show a comparison with the ground truth. Retrieved sentences are example results from retrieval that have the highest semantic similarity to the corresponding segments of input frames. Each retrieved sentence is utilized in our model's predictions for the segments with the corresponding color.