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 University, Republic of Korea
[2]Electronics and Telecommunications Research Institute (ETRI), Republic of Korea

Abstract

There has been significant attention to the research on dense video captioning, which aims to automatically localize and caption all events within untrimmed video. Several studies introduce methods by designing dense video captioning as a multitasking problem of event localization and event captioning to consider inter-task relations. However, addressing both tasks using only visual input is challenging due to the lack of semantic content. In this study, we address this by proposing a novel framework inspired by the cognitive information processing of humans. Our model utilizes external memory to incorporate prior knowledge. The memory retrieval method is proposed with cross-modal video-to-text matching. To effectively incorporate retrieved text features, the versatile encoder and the decoder with visual and textual cross-attention modules are designed. Comparative experiments have been conducted to show the effectiveness of the proposed method on ActivityNet Captions and YouCook2 datasets. Experimental results show promising performance of our model without extensive pretraining from a large video dataset. Our code is available at `https://github.com/ailab-kyunghee/CM2_DVC`.

1. Introduction

With the increasing demand for video understanding and multimodal analysis, the field of video captioning is growing rapidly. The task of conventional video captioning involves generating precise descriptions for trimmed video segments and several studies show successful results [7, 12, 21, 22, 24, 26, 27, 32, 36, 40–42]. However, it faces considerable challenges when applied to dense video captioning. Dense video captioning aims to localize important event segments (i.e., to find event boundaries) from untrimmed videos and describe the event segment (i.e., what happens in the event) with natural language. For achieving high-performance dense video captioning, it is important to properly model inter-task interactions between event localization and caption generation.

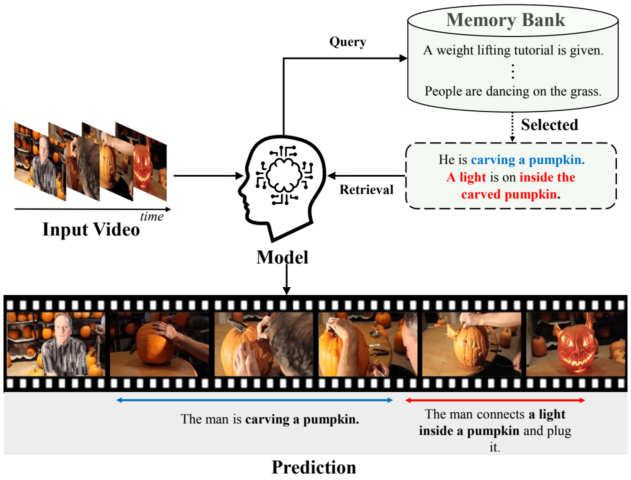

Memory Bank
A weight lifting tutorial is given.
⋮
People are dancing on the grass.
Query
Selected
He is carving a pumpkin.
A light is on inside the carved pumpkin.
Retrieval
Input Video
time
Model
The man is carving a pumpkin.
The man connects a light inside a pumpkin and plug it.
Prediction

Figure 1. Conceptual figure of the proposed cross-modal memory-based dense video captioning (CM^2). Our method can search for relevant clues from an external memory bank to provide precise descriptions and localization for untrimmed video.

Recent studies in vision and language learning have shown impressive results in cross-modal correlation tasks [19, 28, 34]. However, connecting natural language and video is still challenging due to the difficulties in modeling spatiotemporal information [13]. Video-and-language learning requires complex model architectures, specialized training protocols, and large computational costs [8]. Even dense video captioning requires connecting untrimmed videos and natural language to localize events and describe them [17, 46, 48].

This study is motivated by the observation of how humans recognize and describe scenes. Humans are capable of identifying important events and describing them by recalling relevant memories based on cues they have observed. In cognitive information processing, this processing is called cued recall [1, 33]. By recalling relevant memories, humans can describe the scenes with human-understandable natural language.

[*] Corresponding authors.

To verify the feasibility of our idea, we have conducted a preliminary experiment. To measure the usefulness of text clues from external memory, we search the relevant information by using the ground truth caption of the query video, which is the ideal case where we can achieve in the external memory. In the real-world condition, we could not use text query and video features will be used as a query. As shown in Table 1, the performance of the dense video captioning could be significantly improved (CIDEr of 183.95 is achievable with Oracle retrieval on the ground truth event segment in YouCook2 dataset [53]).

Following this insight, we devise a new dense video captioning framework, named Cross-Modal Memory-based dense video captioning (CM^2). Our model can recall relevant events from external memory to improve the generation quality of captions in dense video captioning as shown in Figure 1. To mimic the human's process, an external memory is designed based on prior knowledge which is extracted from training data. Then, the proposed model extracts potential event candidates from given untrimmed videos and retrieves relevant information from the external memory to provide the model with diverse and semantic information. By incorporating the retrieved memory into visual features, our method further introduces a versatile encoder and decoder structure. The encoded features are aggregated by using visual cross-attention and textual cross-attention in a versatile transformer decoder, which helps the model learn inter-task interactions from visual and text clues. Our main contributions can be summarized as:

- Inspired by the human cognitive process, we introduce a new dense video captioning method with cross-modal retrieval from external memory. To the best of our knowledge, this is the first study that uses cross-modal retrieval from external memory for dense video captioning. By retrieving relevant text clues from the memory, the proposed model could elaborately localize and describe important events in a more fluent and natural way.
- To effectively leverage multi-modal features, we propose a versatile encoder-decoder structure with a visual cross-attention and a textual cross-attention. Our model could effectively learn cross-modal correlation and model inter-task interactions for improving dense video captioning.
- Comprehensive experiments have been conducted on ActivityNet Captions [17] and YouCook2 [53] datasets to verify the effectiveness of memory retrieval in dense video captioning. Our model also achieves comparable performance without pretraining on large video datasets.

2. Related Work

2.1. Desne Video Captioning

Dense video captioning is a multi-task problem that combines two sub-tasks: Event localization and event captioning. Krishna *et al.* [17] introduced a dense video captioning model by first generating proposals and then using an attention-based LSTM to generate captions, following the ”localize-then-describe” strategy. Subsequent studies [14, 15, 43, 45, 49] aimed to produce more precise and informative captions within this strategy. However, two-stage approaches have major limitations, as they do not jointly train event localization and event captioning, resulting in less attention to inter-task interactions.

To address the aforementioned limitations, recent studies propose joint training of two sub-tasks [4, 6, 9, 20, 23, 29, 37, 38, 43, 46, 48, 54]. Deng *et al.* [9] initially generate a paragraph for a given video and then utilize it for grounding. Wang *et al.* [46] define dense video captioning as a parallel set prediction task and propose an end-to-end method for event localization and event captioning, using only visual input to solve the two sub-tasks. Yang *et al.* [48] make use of transcribed speech for multi-modal inputs, predicting both time tokens and caption tokens as a single sequence. For the pretraining of the model, an additional YT-Temporal-1B dataset which contains 18 million narrated videos collected from YouTube is used.

However, training high-quality dense video captioning models without pretraining from a large number of videos still remains very challenging. Our study presents a novel approach to exploit prior knowledge to enhance the quality of dense video captioning.

2.2. Retrieval-Augmented Generation

The retrieval-augmented approach is often used in language generation tasks. Lewis *et al.*[18] propose retrieval-augmented generation, which combines pre-trained parametric and external non-parametric memory to effectively leverage pre-trained model knowledge. Some works [30, 31, 35, 47, 52] in image captioning also employ this external datastore approach. Similar to ours, Sarto *et al.*[35] and Ramos *et al.*[30] propose an approach to train a retrieval-augmented image captioning model by processing encoded retrieved captions through cross-attention. Recent studies also show retrieval augmented generation in the context of video captioning [5, 16, 51]. They propose to improve video captioning by incorporating external knowledge, such as video-related training corpus [16] and memory-augmented encoder-decoder structure [51]. They reference the retrieved text obtained from memory in the word prediction distribution of the captioning decoder.

In this study, our model references retrieved information throughout all layers of the decoder with cross-attention. While they only concentrate on enhancing word prediction for video captioning, our method adopts a structure that utilizes the retrieved text as semantic information, benefiting both event localization and event captioning. Note that, retrieval-augmented generation has been largely unexplored in dense video captioning. Previous studies that use the retrieval-augmented generation approach in the image and short video captioning only utilize retrieved textual information for improving caption quality. In this study, we present a new structure to exploit the retrieval of text clues for generating dense captions and localizing events from untrimmed videos.

3. Method

Our goal is to improve event-level localization and event captioning from untrimmed video by exploiting prior knowledge. For this, we introduce a new framework (CM^2) which is designed with cross-modal memory retrieval. CM^2 could search relevant information by segment-level video features and retrieve text features from external memory in a video-to-text cross-modal manner (Section 3.1). Furthermore, to ensure that the model efficiently 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 (Section 3.2). As illustrated in Figure 2, our model takes input video frames and extracts video frame features $\mathbf{x} = \{x_i\}_{i=1}^{F}$ and retrieved text features $\mathbf{y} = \{y_j\}_{j=1}^{W}$ where F and W denote the number of frames in the given video and the number of retrieved text features, respectively. For the given input video, the model generates segment and caption pairs $\{(t_n^s, t_n^e, S_n)\}_{n=1}^{N}$ where N denotes the number of events detected by our method and t_n^s and t_n^e denote the start and the end timestamp of n-th event. S_n denotes the generated captions for n-th event segment. Details of dense event prediction will be introduced in Section 3.3.

3.1. Memory Retrieval

3.1.1 Memory Construction

To store high-quality semantic information as prior knowledge in the memory, we first construct an explicit external memory bank by encoding sentence-level features. The sentences are collected from the training data of the in-domain target dataset [17, 53] by taking into account the semantic distributions appropriate to the query videos. For example, the captions in AcitivityNet Caption training set are used for constructing external memory in experiments on AcitivityNet Captions in this study. For segment-level video-to-text retrieval, we define a memory unit in a sentence level that corresponds to the event clip instead of whole paragraphs from an untrimmed video in the dataset. For segment-level video-to-text embedding, we adopt pretrained CLIP Vit-L/14 [10, 28] which shows promising alignment ability by mapping image and text to the shared feature space. For storing semantic information of captions at a sentence level, we tokenize the captions of the event segment using the CLIP tokenizer, ensuring padding to match the maximum token number of ground truth captions. Subsequently, all tokenized caption sentences are encoded by a CLIP text encoder, and the resulting sentence-level embeddings are stored in the external memory bank.

3.1.2 Segment-level Retrieval

Untrimmed videos could consist of multiple events, each containing distinct semantic information. As both sub-tasks of dense video captioning operate at the event level, it is crucial to design an appropriate retrieval method that considers segment-level semantic information. We propose a novel cross-modal memory-based dense video captioning (CM^2), designed to take into account the semantic information of the segment that can potentially include events. By utilizing image-to-text retrieval strategies with CLIP [10, 28] and temporal anchors, our method ensures the incorporation of semantic details from dense events. The proposed approach involves two key steps: segment-level retrieval and feature aggregation as shown in Figure 2 (b).

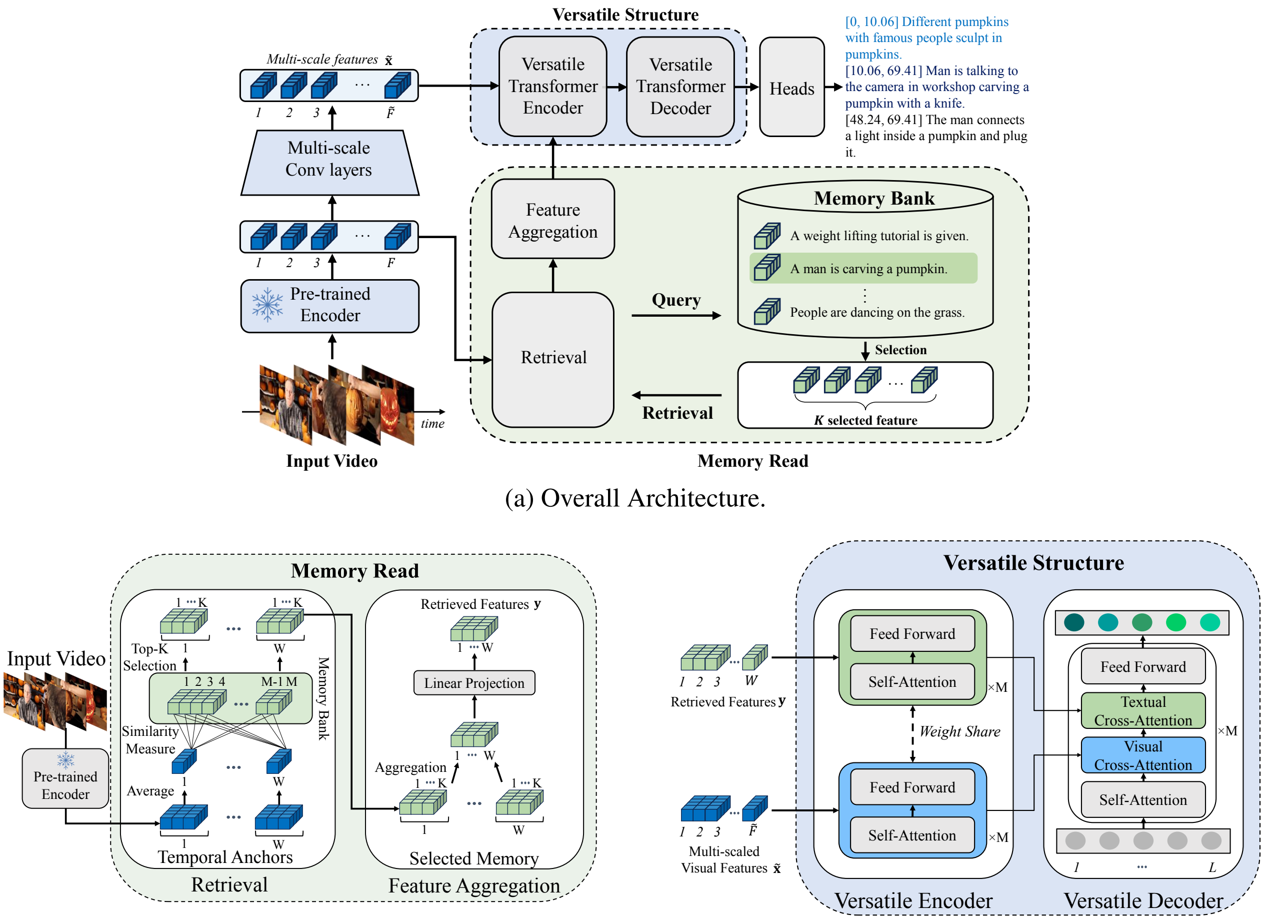
In segment-level retrieval, to acquire semantic information related to events within the input video, we divide the input video into W temporal anchors. For frame-level visual feature extraction, we adopt CLIP ViT-L/14. To obtain the representative information contained in each anchor, we compress the temporal dimension at each anchor through averaging, yielding segment-level visual features. Then, for each anchor, the segment-level visual feature is used as a query for retrieving relevant information from the external memory. For finding relevant information, the similarity between the segment-level visual feature and CLIP text features in the memory is calculated (In this study, cosine similarity between two feature vectors is used as a similarity metric). Based on the similarity scores, K sentence features are retrieved for each anchor, which results in a selected memory feature set for j-th anchor as $\mathbf{m}^j = \{m_1^j, ..., m_K^j\}$.

Next, we perform feature aggregation to summarize useful information from K retrieved sentence features $\mathbf{m}^j$ associated with each anchor in the selected memory. The average pooling over the K sequences is conducted in each anchor. Finally, we obtain the retrieved text features $\{y_j\}_{j=1}^{W}$.

3.2. Versatile Encoder-Decoder

In this section, we describe how we build a structure to incorporate visual features and retrieved text features for event localization and event captioning. We generate event query features with well-incorporated temporal information using an encoder-decoder structure based on the deformable transformer [56], as in [46]. However, our approach differs from [46] as our model incorporates not only visual features but also retrieved text features for making positive effects in both captioning and localization. To achieve this, we propose a versatile encoder-decoder structure that effectively uses retrieved text features and visual features.

Feature Encoding. First, we sample the frame-level features extracted by the pre-trained CLIP ViT-L/14 with 1 FPS to a fixed frame number as $\mathbf{x} = \{x_i\}_{i=1}^{F}$ for batch processing. Then, we added L temporal convolutional layers for the multi-scaling processing of video frame features. The multi-scale convolutional layers output multi-scale visual features as $\tilde{\mathbf{x}} = \{\tilde{x}_i\}_{i=1}^{\tilde{F}}$.



(a) Overall Architecture.

(b) Memory Read Module.

(c) Versatile Encoder-Decoder Module.

Figure 2. **Overview of CM^2.** We approach the dense video captioning task in a memory-retrieval-augmented caption generation manner. We show the overall architecture in (a). We conduct video-to-text cross-modal retrieval using input video features obtained through a pre-trained encoder. As illustrated in (b), we generate segment-level W temporal anchors from the input video features. Then we measure similarities between the anchors and the text features stored in a memory to obtain W retrieved features through aggregation. As illustrated in (c), we encode the multi-scale video features $\tilde{\mathbf{x}}$ and retrieved features using a versatile transformer encoder. Each encoded feature vector undergoes the corresponding cross-attention layers to obtain refined event queries. Finally, we obtain the set of start time, end time, and caption by passing the event queries through a head.

Versatile encoder. CM^2 enhances the interplay between visual and text modalities while preserving their original information, achieved through the use of versatile weight-shared encoders. These weight-shared encoders, illustrated in Figure 2 (c), are employed to process each modality feature. The versatile encoder is designed with M blocks where each block consists of feedforward and self-attention layers. By employing weight-shared encoders, the visual and text modality features undergo training in a shared embedding space, fostering potential cross-modality connections. Furthermore, since each modality process is processed separately by the weight-shared encoder, it could effectively retain distinctive modality-specific information. The visual encoder takes a sequence of multi-scale frame features $\tilde{\mathbf{x}} = \{\tilde{x}_i\}_{i=1}^{\tilde{F}}$ as input and generates encoded visual features as output. Simultaneously, the same versatile encoder processes a set of retrieved text features $\mathbf{y} = \{y_j\}_{j=1}^{W}$, producing W encoded text features.

Versatile decoder. Through the versatile decoder, we design learnable embeddings, event queries $\mathbf{q} = \{q_l\}_{l=1}^{L}$, to include temporally and semantically rich information. When video and text modalities are given, a single cross-attention is insufficient to generate the necessary representations for the two sub-tasks. Therefore, CM^2 separates the visual cross-attention layer from the textual cross-attention layer, as described in Figure 2 (c). We aim for each modality to handle tasks related to temporal and semantic information processing separately. In visual cross-attention, considering the cross-attention between encoded visual features and event queries enhances the temporal information

Table 1. **Effect of memory retrieval in ActivityNet Captions and Youcook2.** No retrieval refers to a case where the model is forwarded without any retrieval. Oracle methods are implemented to measure the upper bound which could be achieved by the retrieval with an ideal query to the memory bank. The captions retrieved from the memory by ground truth captions of query video are directly used as the output of the model.

Retrieval Type	ActivityNet				YouCook2			
	CIDEr	METEOR	BLEU4	SODA_c	CIDEr	METEOR	BLEU4	SODA_c
No Retrieval	31.24	8.03	2.15	6.01	23.67	5.30	1.17	4.77
Proposed Retrieval (Ours)	33.01	8.55	2.38	6.18	31.66	6.08	1.63	5.34
Oracle w/o GT proposal	40.24	9.43	2.88	6.96	53.55	9.18	3.49	6.81
Oracle w/ GT proposal	84.47	15.69	5.86	12.41	183.95	23.53	13.05	25.51

of event queries. In textual cross-attention, considering the cross-attention between encoded text features and event queries enriches the semantic information of event queries. The output of the versatile decoder produces event queries $\tilde{\mathbf{q}} = \{q_l\}_{l=1}^{L}$ with both temporal and semantic information.

3.3. Dense Event Prediction

Parallel Heads. CM^2 employs a parallel decoding structure with sub-task heads for a given event query $\tilde{q}_l$. Our approach includes three parallel heads: a localization head, a captioning head, and an event counter.

Localization Head. The localization head is implemented by a multi-layer perceptron to predict the box prediction, including the center and length of the ground-truth segment, for a given event query. Additionally, it conducts binary classification to predict the foreground confidence of each event query. Finally, the localization head outputs a set of tuples $(t_l^s, t_l^e, c_l)_{l=1}^{L}$, where each tuple represents the start time t_l^e, end time t_l^e, and localization confidence c_l of l-th event segment, respectively.

Captioning Head. For the captioning head, we employ the deformable soft attention LSTM which uses the soft attention around the reference points, enhancing word generation performance. [46]. For the input of the captioning head, we utilize attention feature $a_{l,s}$, event query $\tilde{q}_l$, and the previous word $w_{l,s-1}$ to predict the next word. As the sentence progresses, the captioning head generates the entire sentence $\mathbf{S}_l = w_{l,1}, ..., w_{l,S}$, where S represents the length of the sentence.

Event Counter. The event counter predicts the appropriate number of events in the video. To achieve this, it compresses essential information from the event query $\tilde{q}_l$ through a max-pooling layer and a fully-connected layer. It predicts a vector r_{len} representing a specific number of events. During inference, the predicted event count is selected by $N = argmax(r_{len})$. Finally, the N predicted sets $\{(t_n^s, t_n^e, S_n)\}_{n=1}^{N}$ are determined by the Hungarian algorithm [3], using a matching cost $C = L_{cls} + \alpha L_{loc}$ with generalized IOU loss and focal loss. The focal loss L_{cls} is computed between the predicted classification score and the ground-truth label. The generalized IOU loss L_{loc} measures the predicted segment against the ground-truth segment.

Training and Inference. During training, we train CM^2 using four losses: L_{loc}, L_{cls}, L_{count}, and L_{cap}. L_{count} represents the cross-entropy between the predicted count number distribution and the ground truth. L_{cap} is the cross-entropy between the predicted word probability and the ground truth. The total loss is defined as follows:

$$L_T = L_{cls} + \lambda_{loc} L_{loc} + \lambda_{count} L_{count} + \lambda_{cap} L_{cap} \quad (1)$$

During inference, given visual input x and retrieved text input y, our model predicts N sets of predictions $\{(t_n^s, t_n^e, S_n)\}_{n=1}^{N}$. For both training and inference, we conducted retrieval using the same external memory bank.

4. Experiments

To verify the effectiveness of our method, comparative experiments have been conducted. First, Section 4.1 introduces the experimental setting used in this study. Section 4.2 shows the effectiveness of memory retrieval in dense video captioning. Section 4.3 shows the comparison with state-of-the-art methods. Section 4.4 shows ablation studies for our model to validate the effectiveness of each component. Qualitative results of our method and discussion are followed.

4.1. Experimental Settings

Dataset. We employed two dense video captioning benchmark datasets, namely ActivityNet Captions [17] and YouCook2 [53], for training and evaluation. ActivityNet Captions consists of 20k untrimmed videos of diverse human activities. On average, each video spans 120s and is annotated with 3.7 temporally localized sentences. For training, validation, and testing, we follow the standard split of videos. YouCook2 consists of 2k untrimmed cooking procedure videos, with an average duration of 320s per video and 7.7 temporally localized sentences per annotation. We followed the standard split for training, validation, and testing videos. Notably, we use approximately 7% fewer videos than the original count, as we use those accessible on YouTube.

Evaluation Metrics. We evaluated our method for two subtasks in dense video captioning. By using ActivityNet Challenge official evaluation tool [44], we evaluated generated captions using the metrics CIDEr [39], BLEU4 [25], and

Table 2. **Performance of Event Captioning in ActivityNet Captions.** Bold means the highest score. Underline means 2nd score. # PT denotes the number of videos used for pretraining. † denotes results reproduced from official implementation in our environment.

Method	Backbone	# PT	CIDEr	METEOR	BLEU4	SODA_c
Vid2Seq [48]	CLIP	15M	30.10	8.50	-	5.80
MT [54]	TSN	-	6.10	3.20	0.30	-
ECHR [45]	C3D	-	14.70	7.20	1.82	3.20
UEDVC [50]	C3D	-	-	-	-	5.5
PDVC† [46]	CLIP	-	29.97	8.06	2.21	5.92
Ours	CLIP	-	**33.01**	**8.55**	**2.38**	**6.18**

Table 3. **Performance of Event Captioning in YouCook2.** Bold means the highest score. Underline means 2nd score. # PT denotes the number of videos used for pretraining. † denotes results reproduced from official implementation in our environment.

Method	Backbone	# PT	CIDEr	METEOR	BLEU4	SODA_c
Vid2Seq [48]	CLIP	1M	**47.10**	**9.30**	-	**7.90**
MT [54]	TSN	-	9.30	5.00	1.15	-
ECHR [45]	C3D	-	-	3.82	-	-
E2ESG [55]	C3D	-	25.00	3.50	-	-
PDVC† [46]	CLIP	-	29.69	5.56	1.40	4.92
Ours	CLIP	-	31.66	6.08	**1.63**	5.34

METEOR [2], which calculate matched pairs between generated captions and ground truth across IOU thresholds of 0.3, 0.5, 0.7, 0.9. Additionally, for measuring storytelling ability, we employed SODA_c [11]. For event localization, we measured average precision, average recall, and F1 score, which represents the harmonic mean of precision and recall. These scores are averaged over IOU thresholds of 0.3, 0.5, 0.7, 0.9.

Implementation Details. For both datasets, we extract video frames at a rate of 1 frame per second and then subsample or pad the sequence of frames to achieve a total of F frames, where we set $F = 100$ in ActivityNet Captions and $F = 200$ in YouCook2. We employ a two-layer deformable transformer with multiscale deformable attention spanning four levels. The number of event queries is set to 10 for ActivityNet Captions and 100 for YouCook2, respectively. In this study, the balancing hyperparameters of α, λ_{loc}, λ_{count}, and λ_{cap} are set to 2, 2, 1, and 1, respectively. The number of anchors is empirically set to 10 for ActivityNet Captions and 50 for YouCook2. In retrieval, We set the anchor number of 50, with k set to 80 for each anchor. Therefore, we utilize 4000 retrieved text features. During training, the ground truth of the corresponding input video was excluded from the memory bank.

4.2. Effect of Memory Retrieval

To assess the effectiveness of memory retrieval, comparative experiments have been conducted. Four different memory retrieval approaches are implemented as shown in Table 1. No retrieval refers to a method where the model is forwarded without any retrieval. Oracle retrieval is implemented to measure the upper bound of memory retrieval. The captions are retrieved from the external memory based on the similarity with ground truth captions of query video. The retrieved captions are directly used as an output of the model without model forwarding. For Oracle without GT proposal, matched the retrieved text to the event segments predicted by our model. For Oracle with GT proposal match retrieved text to ground truth event segments for measuring the performance.

As shown in the table, the proposed retrieval method achieves higher scores compared with the model without retrieval. This is mainly due to the reason that retrieved text features could provide semantically useful features to the model, which helps the model exploit visual and text relations. When we evaluate the performance with Oracle, even when we used the same event segments as our model, by using the retrieved text without the model forward, we observed a large enhancement in captioning performance. Moreover, when we match retrieved text to ground truth event segments, the performance is significantly improved.

These results show huge potential for retrieval-based dense video captioning. In this study, we use clip visual features from Vit-L/14 and average the features from each anchor are aggregated by averaging them. In other words, video-to-text matching is implemented by projecting video features to image-to-text feature space. Some important video features might be lost during this process. In the future, according to the advances in video modeling and video-to-text matching, our method which uses retrieval from external memory for dense video captioning could be further improved.

4.3. Comparison with State-of-the-art-Methods

In Table 2 and Table 3, we compare our method with state-of-the-art dense video captioning approaches [45, 46, 48, 50, 54, 55] on both YouCook2 and ActivityNet Captions datasets. As shown in Table 2, our method achieves the best scores over four metrics of CIDEr, METEOR, BLEU4, and SODA_c. Even the method could achieve higher scores compared with [48] which leverages an additional 15 million videos for pretraining. In YouCook2 dataset, Vid2seq [48] which uses extra 1 million videos for pretraining achieves the best performance. Our method achieves comparable performance on YouCook2 without using extra videos. By using prior knowledge from external memory, our method could improve the quality of caption generation.

We also compare the localization ability of our method. Table 4 shows the comparison of our model with other models that use CLIP features as a visual feature in YouCook2 and ActivityNet Captions datasets. As shown in the table, our method achieves the best scores in ActivityNet Captions in both precision and recall. Also, in YouCook2 dataset, our method achieves the best precision and second recall scores,

Table 4. **Performance of Event Localization in ActivityNet Captions and YouCook2.** Bold means the highest score. Underline means 2nd score. PT denotes pretraining from the additional video datasets. † denotes results reproduced from official implementation in our environment. All methods used the CLIP as the backbone.

Method	PT	ActivityNet Captions			YouCook2		
		F1	Recall	Precision	F1	Recall	Precision
Vid2Seq [48]	✓	53.29	52.70	53.90	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**

Table 5. **Ablation study to verify the effect of structure component for incorporating retrieved features.** WS denotes a weight sharing for the versatile encoder. SE denotes the case where we encode textual and visual features, separately. Without SE is implemented by concatenating textual and visual features and passing through a single encoder. TCA denotes the use of textual cross-attention where the model uses additional cross-attention for encoded text features. Without TCA is implemented by concatenating visual and text features and passing through a single cross-attention. The performance is measured in YouCook2.

WS	SE	TCA	CIDEr	METEOR	BLEU4	SODA_c	F1
✗	✗	✗	29.49	5.65	1.34	5.26	27.66
✗	✗	✓	28.40	5.56	1.37	4.74	25.14
✗	✓	✗	30.86	5.61	1.40	5.14	27.06
✗	✓	✓	30.94	5.71	**1.65**	5.07	27.11
✓	✓	✓	**31.66**	**6.08**	1.63	**5.34**	**28.43**

which results in the best F1 score. Our memory retrieval approach not only improves caption generation but also helps the model to localize event boundaries by providing semantic cues that can be exploited during inter-task interactions.

4.4. Ablation Studies

Our method aims to leverage the retrieved segment-level text features as semantic information for improving dense video captioning. In this section, we present ablation studies for the component that is designed to incorporate retrieved features from the memory. We design a versatile encoder structure where the encoder processes retrieved text features and visual features. In other words, one encoder is shared between two modalities, and the model is trained to process both modalities. Table 5 shows the ablation study results. It is observed that the use of a weight-shared versatile encoder structure could improve the model performance. The cases where cross-modal information (i.e., visual and textual features) is processed by the separate encoder (with SE) achieve higher performance compared with the model without the separate encoder in which the two features are concatenated before entering the transformer encoder. Also, weight sharing for the encoder is better than having two separate encoders for each modality. These results indicate that it is important to encode visual and textual features separately by preserving own information. However, the model could learn the interconnection between textual and visual features by training the encoder in a versatile manner.

Table 6. **Effect of anchor number for retrieval in YouCook2.** # Anchor denotes the number of anchors. The performance is measured by changing the number of anchors.

# Anchor	CIDEr	METEOR	BLEU4	SODA_c	F1
1	27.97	5.54	1.39	5.14	28.10
10	31.36	5.75	1.63	5.17	27.33
30	28.41	6.02	1.43	5.08	26.87
50	31.66	**6.08**	1.63	**5.34**	**28.43**
70	**32.73**	5.83	**1.66**	5.28	27.55
90	29.88	5.57	1.43	**5.34**	27.63

Furthermore, we also compare the presence of textual cross-attention. We compare cases where our model is designed with separate textual and visual cross-attention with the cases where the model is implemented by using combined cross-attention where the textual and visual features are concatenated before being put into the decoder. As shown in Table 5, the performance is increased with separate textual cross-attention. This is mainly due to the reason that the decoder could incorporate visual and textual features by specialized cross-attention explicitly.

4.5. Discussion

4.5.1 Qualitative Examples

Figure 3 shows predicted examples of our approach. It can be observed that memory retrieval effectively references meaningful and helpful sentences from memory, obtained through segment-level video-text retrieval for the given video. As a result, our method generates relatively accurate event boundaries and captions. The semantic information obtained from memory through retrieval assists in semantic predictions during caption generation. More examples are provided in Supplementary Material.

4.5.2 Effect of Anchor Number for Retrieval

We explore the effect of the number of temporal anchors generated during memory retrieval. The number of temporal anchors is related to the basic unit for giving a query to the memory bank and it also attributes to the number of retrieved features. Table 6 shows the performance by changing the anchor number in YouCook2 dataset. When the anchor number is set to 1, the untrimmed video information is averaged to a single visual feature for querying to the memory bank. This approach could not exploit fine-grained details for retrieving the semantic text cues. As we increase the number of anchors, the fine-grained details can be captured for querying to the memory bank, which improves the performance of dense video captioning model. However, when an excessive number of features are retrieved, noisy features contribute to the degradation of performance. It is

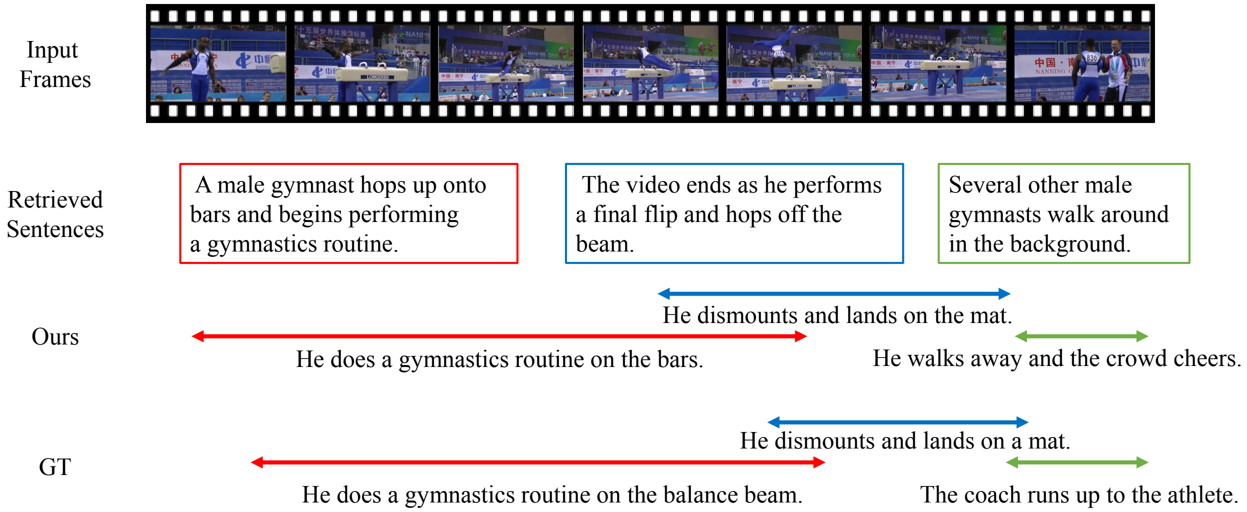

Input Frames
Retrieved Sentences
A male gymnast hops up onto bars and begins performing a gymnastics routine.
The video ends as he performs a final flip and hops off the beam.
Several other male gymnasts walk around in the background.
Ours
He dismounts and lands on the mat.
He does a gymnastics routine 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.

Figure 3. **Example of dense video captioning predictions with ours on ActivityNet Captions Validation set.** 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.

Table 7. **Effect of the number of selected features in YouCook2.** #SF denotes the number of retrieved features from the memory bank. The performance is measured by changing the number of retrieved text features per anchor.

#SF	CIDEr	METEOR	BLEU4	SODA_c	F1
1	19.76	4.36	0.65	4.79	26.79
20	30.22	5.64	1.62	5.20	27.33
40	31.25	5.73	**1.79**	5.29	28.10
60	31.24	5.77	1.63	5.26	**28.58**
80	31.66	**6.08**	1.63	**5.34**	28.43
100	**32.07**	5.81	1.58	5.32	27.86

observed that the anchor number of 50 consistently yields outstanding performance in both event localization and caption generation in YouCook2 dataset.

4.5.3 Effect of Number of Retrieved Features for Each Anchor

We also investigate the effect of the number of retrieved features per anchor on the performance. When we set the number of retrieved features per temporal anchor to 1, it means we only consider the text from memory that has the highest similarity to the visual feature of the temporal anchor. Table 7 shows the results according to the number of retrieved features per anchor. As we increase the number of retrieved features per anchor, the memory read could provide stable and robust semantic information to the model. When the number of retrieved features per anchor is set to 80, our method consistently achieves good performance in both sub-tasks in YouCook2 dataset. However, with a too large number, the noisy features could be retrieved because we retrieved the features with the similarity in descending order, which could degrade the performance.

5. Conclusion

In this study, we introduced a novel approach to dense video captioning inspired by the human cognitive process of scene understanding. Leveraging cross-modal retrieval from external memory, CM^2 demonstrated a significant improvement in both event localization and caption generation. Through comprehensive experiments on ActivityNet Captions and YouCook2 datasets, we validated the effectiveness of our memory retrieval approach. Notably, CM^2 achieved competitive results without the need for pretraining on a large number of video data, highlighting its efficiency. We believe that our work opens avenues for future study in dense video captioning and encourages the exploration of memory-augmented models for improving video understanding and captioning.

Acknowledgements

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant 2020-0-00004 (Development of Provisional Intelligence Based on Long-term Visual Memory Network), Grant 2022-0-00078, Grant IITP-2024-RS-2023-00258649, Grant 2021-0-02068, Grant RS-2022-00155911, by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant 2021R1G1A1094990, and by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD).

References

[1] Ken Allan and MD Rugg. An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia*, 35(4):387–397, 1997. 1

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[4] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*, 2020. 2

[5] Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. Retrieval augmented convolutional encoder-decoder networks for video captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1s):1–24, 2023. 2

[6] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021. 2

[7] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. Video captioning with guidance of multimodal latent topics. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1838–1846, 2017. 1

[8] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750, 2023. 1

[9] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 234–243, 2021. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 517–531. Springer, 2020. 6

[12] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 1

[13] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Qiyuan Hu, and Devi Parikh. Mugen: A playground for video-audio-text multimodal understanding and generation. In *European Conference on Computer Vision*, pages 431–449. Springer, 2022. 1

[14] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020. 2

[15] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 2

[16] Shuaiqi Jing, Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Memory-based augmentation network for video captioning. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 2

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 2, 3, 5

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[20] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7492–7500, 2018. 2

[21] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 1

[22] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation, 2020. 1

[23] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597, 2019. 2

[24] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016. 1

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[26] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 1

[27] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[29] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019. 2

[30] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*, 2023. 2

[31] Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, and Bruno Martins. Retrieval augmentation for deep neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[32] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 433–440, 2013. 1

[33] Michael D Rugg, Paul C Fletcher, Kevin Allan, Chris D Frith, RSJ Frackowiak, and Raymond J Dolan. Neural correlates of memory retrieval during recognition memory and cued recall. *Neuroimage*, 8(3):262–273, 1998. 1

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[35] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, pages 1–7, 2022. 2

[36] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1

[37] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1924, 2017. 2

[38] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6382–6391, 2019. 2

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[40] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 1

[41] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.

[42] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 1

[43] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7190–7198, 2018. 2

[44] Teng Wang, Huicheng Zheng, and Mingjing Yu. Dense-captioning events in videos: Sysu submission to activitynet challenge 2020. *arXiv preprint arXiv:2006.11693*, 2020. 5

[45] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900, 2020. 2, 6

[46] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 1, 2, 3, 5, 6, 7

[47] Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wangrong Cheng, and Jinwen Tian. A unified generation-retrieval framework for image captioning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2313–2316, 2019. 2

[48] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1, 2, 6, 7

[49] Dali Yang and Chun Yuan. Hierarchical context encoding for events captioning in videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1288–1292. IEEE, 2018. 2

[50] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *European Conference on Computer Vision*, pages 363–379. Springer, 2022. 6

[51] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9837–9846, 2021. 2

[52] Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, and Jiaxuan Zhang. Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, 10 (18):6235, 2020. 2

[53] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2, 3, 5

[54] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 2, 6

[55] Wanrong Zhu, Bo Pang, Ashish V Thapliyal, William Yang Wang, and Radu Soricut. End-to-end dense video captioning as sequence generation. *International Conference on Computational Linguistics (COLING)*, 2022. 6

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3