

Exploring Temporal Event Cues for Dense Video Captioning in Cyclic Co-Learning

Zhuyang Xie¹, Yan Yang^{1,2*}, Yankai Yu¹, Jie Wang¹, Yongquan Jiang^{1,2}, Xiao Wu^{1,2}

¹School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China

²Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, China
{zyxie, yyk}@my.swjtu.edu.cn, {yyang, jackwang, yqjiang, wuxiaohk}@swjtu.edu.cn

Abstract

Dense video captioning aims to detect and describe all events in untrimmed videos. This paper presents a dense video captioning network called Multi-Concept Cyclic Learning (MCCL), which aims to: (1) detect multiple concepts at the frame level and leverage these concepts to provide temporal event cues; and (2) establish cyclic co-learning between the generator and the localizer within the captioning network to promote semantic perception and event localization. Specifically, weakly supervised concept detection is performed for each frame, and the detected concept embeddings are integrated into the video features to provide event cues. Additionally, video-level concept contrastive learning is introduced to produce more discriminative concept embeddings. In the captioning network, a cyclic co-learning strategy is proposed, where the generator guides the localizer for event localization through semantic matching, while the localizer enhances the generator's event semantic perception through location matching, making semantic perception and event localization mutually beneficial. MCCL achieves state-of-the-art performance on the ActivityNet Captions and YouCook2 datasets. Extensive experiments demonstrate its effectiveness and interpretability.

Introduction

In recent years, with the rapid growth of video data, automatically generating high-quality video captions has emerged as a significant research focus in the fields of computer vision and natural language processing. Dense video captioning (Krishna et al. 2017; Shen et al. 2017; Mun et al. 2019; Suin and Rajagopalan 2020; Deng et al. 2021), a critical task within this domain, requires systems not only to accurately identify multiple events but also to generate natural language descriptions for each event. The primary challenge is to simultaneously handle video content understanding, temporal event localization, and event captioning.

Despite recent advancements in dense video captioning, existing methods still exhibit limitations when handling complex video scenarios. Early methods typically adopt unimodal information (e.g., visual features) as input (Krishna et al. 2017; Tu et al. 2021; Zhou et al. 2018; Tu et al. 2023)

*Corresponding author.

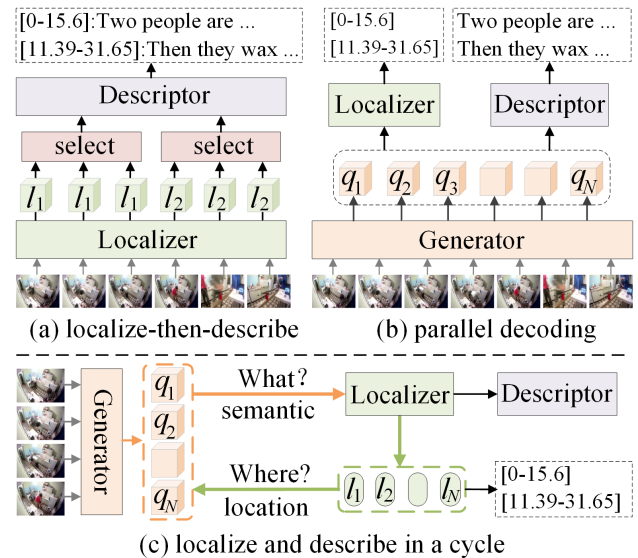


Figure 1: (a) The two-stage approach predicts multiple event candidates and eliminates redundancy using non-maximum suppression. (b) Learnable event queries are embedded in the generator (visual transformer), with two prediction heads for localization and captioning. (c) MCCL uses cyclic co-learning, where the generator, localizer, and descriptor enhance performance collaboratively by leveraging the mutual benefits of captioning and localization.

for image semantic understanding. Additionally, current approaches face challenges in feature representation and cross-modal alignment, which can lead to inconsistencies between the generated captions and video content. Inspired by recent progress in vision and language learning (Radford et al. 2021; Luo et al. 2022; Xie et al. 2023a; Deng et al. 2023), we enhance video semantic comprehension by retrieving text information that is highly relevant to the video content, thereby improving both event localization and caption quality.

Event localization and caption generation are commonly integrated into a unified framework. Mainstream methods can be divided into two categories: two-stage method (Krishna et al. 2017; Duan et al. 2018; Chen and Jiang

2021; Iashin and Rahtu 2020b) and parallel decoding method (Wang et al. 2021; Kim et al. 2024). As illustrated in Fig.1(a), the two-stage method initially employs a localizer to produce a large set of event proposals, which are then filtered using non-maximum suppression (NMS) to select the most accurate ones. Subsequently, a descriptor produces captions based on the visual context of these proposals. This approach heavily depends on anchor design, resulting in substantial computational overhead and limiting end-to-end training capabilities. More recently, a parallel end-to-end method PDVC (Wang et al. 2021), shown in Fig.1(b), decouples dense video captioning into two parallel tasks that operate on intermediate features (event queries). However, we argue that there is a lack of an effective interaction mechanism between event semantic perception and localization. Fig.1(c) illustrates our cyclic co-learning mechanism, in which the generator perceives potential events (What ?) through semantic matching and guides the localizer in event localization. The localizer then feeds back the location matching to the generator (Where ?), enabling it to better perceive events. This interactive mechanism makes semantic perception and event localization mutually beneficial.

In this paper, we propose a dense video captioning network called Multi-Concept Cyclic Learning (MCCL). The network first enhances video features through video-text retrieval. It then detects concepts at the frame level to provide temporal event cues that enhance event localization and captioning. Finally, cyclic co-learning is established between the generator and localizer, enabling mutual benefits in semantic perception and event localization. The key contributions are as follows:

- A cyclic mechanism is built between the generator and the localizer to promote co-learning of semantic perception and event localization.
- To explore temporal event cues, multiple concepts are detected at the frame level in a weakly supervised manner, which improves both event localization and caption quality.
- Experiments conducted on ActivityNet Captions and YouCook2 demonstrate that the proposed method achieves state-of-the-art performance.

Related Work

Dense Video Captioning. Dense video captioning involves two main tasks: event localization and captioning. Early methods typically adopt a two-stage framework (Krishna et al. 2017), where event localization and captioning are performed separately. Some approaches aim to improve event representation to produce more informative captions (Wang et al. 2018, 2020; Ryu et al. 2021). To achieve more robust and context-aware representations, recent studies incorporate multimodal inputs (Rahman, Xu, and Sigal 2019; Chang et al. 2022; Aafaq et al. 2022). However, these two-stage methods often fail to optimize event localization and captioning together, leading to a lack of interaction between the two tasks. To address this, PDVC (Wang et al. 2021) redefines dense video captioning as a set prediction problem, simultaneously optimizing both tasks on shared

intermediate features. Some studies enhance the captioning network by using cycle consistency (Kim et al. 2021; Yue et al. 2024) to match generated captions with image features. In this study, we emphasize the guidance and feedback between semantic perception and event localization. Motivated by (Tian, Hu, and Xu 2021), we introduce a cyclic co-learning mechanism that strengthens the interaction between localization and captioning, enabling more accurate event localization and high-quality video captions.

Retrieval-Enhanced Captioning. In recent years, retrieval-based methods have emerged as a promising approach to enhance video captioning by improving the informativeness of generated captions (Zhang et al. 2021; Yang, Cao, and Zou 2023; Chen et al. 2023; Jing et al. 2023; Kim et al. 2024). By aligning video content with semantically relevant textual information, these methods enhance contextual understanding, addressing challenges such as insufficient context and ambiguous descriptions. This alignment enables the generation of more accurate and informative captions. In this paper, we retrieve text associated with video frames to provide additional semantic information. The retrieved sentences are then integrated with video features, facilitating more comprehensive video understanding.

Concept Detection for Video Captioning. Concept detection plays a pivotal role in enhancing video captioning by providing a deeper understanding of video content. Recent studies have demonstrated its potential to improve both the quality and informativeness of captions (Gao et al. 2020; Yang, Cao, and Zou 2023; Wu et al. 2023; Lu et al. 2024). These approaches leverage video-level concepts to generate more detailed and contextually enriched captions. In this work, we apply frame-level concept detection in a weakly supervised manner, enabling the model to identify concepts such as objects, actions, and scenes within individual frames. By incorporating these concepts, our model captures fine-grained details and temporal relationships among video elements, leading to more accurate and contextually relevant captions.

Methodology

As shown in Fig.2, the proposed MCCL consists of three components: video-to-text retrieval, multiple concept detection, and cyclic co-learning. The details are described in the following subsections.

Video-to-Text Retrieval

To collect semantic information as prior knowledge, a sentence corpus is constructed from in-domain training sets. By using a pretrained CLIP text encoder, sentence features are extracted and stored in the corpus. The corpus is denoted as $U = \{u_j\}_{j=1}^M$, where $u_j \in \mathbb{R}^{1 \times d}$ represents the j -th sentence feature, and M is the total number of sentences.

Given an input video V with T frames, a pretrained CLIP image encoder extracts frame-level visual features $F^v = \{f_t^v\}_{t=1}^T$, where $f_t^v \in \mathbb{R}^{1 \times d}$. To reduce the computational cost during video-to-text retrieval, these features are evenly

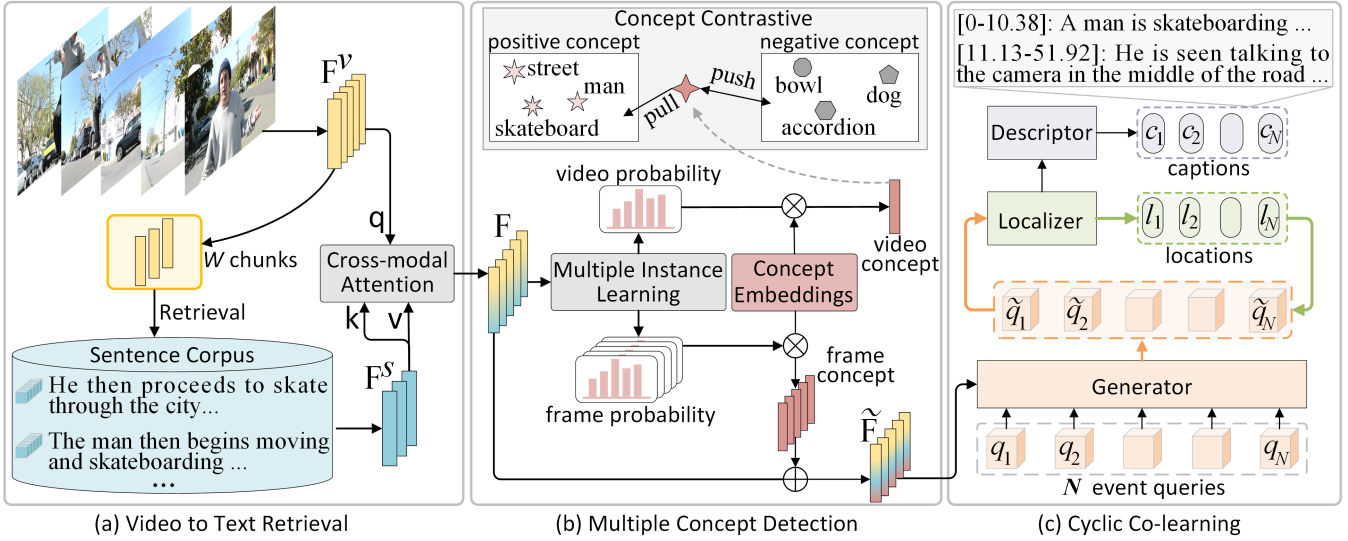


Figure 2: Overview of the framework. (a) A pretrained image encoder extracts video features and performs cross-modal retrieval to obtain sentence features. (b) Video-level and frame-level concepts are detected via multiple instance learning. (c) The features are fed into the generator to update event queries. The localizer predicts locations for each query and selects the optimal ones, while the descriptor produces captions based on these optimal queries. The generator and localizer co-learn in a cycle.

divided into W temporal chunks, and the features within each chunk are averaged to produce chunk-level visual features $\{s_i\}_{i=1}^W$, where $s_i \in \mathbb{R}^{1 \times d}$. These chunk features serve as input queries for text retrieval:

$$\text{sim}(s_i, u_j) = \frac{s_i \cdot u_j}{\|s_i\| \cdot \|u_j\|}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ indicates the cosine similarity. For each chunk s_i , the top N_K sentences are retrieved from corpus U with respect to their similarity scores. These N_K sentence features are subsequently mean pooled to obtain the semantic feature $f_i^s \in \mathbb{R}^{1 \times d}$. All chunk-level semantic features are collectively denoted as $F^s = \{f_i^s\}_{i=1}^W$.

To effectively integrate semantic information with visual features, cross-modal attention is applied for dynamic fusion:

$$F = \text{softmax} \left(\frac{F^v (F^s)^T}{\sqrt{d}} \right) F^s, \quad (2)$$

where $F \in \mathbb{R}^{T \times d}$ represents the aggregated video features.

Multiple Concept Detection

Concept detection can provide valuable guidance for localization (Chen and Jiang 2021) and caption generation (Yang, Cao, and Zou 2023). Frame-level concept detection is introduced to explore temporal event cues. Specifically, the top N_C most frequent words (nouns, adjectives, and verbs) from training captions are selected as concepts $E = \{e_1, e_2, \dots, e_{N_C}\}$. Then, for each video, a video-level concept label $Y^C \in \{0, 1\}^{N_C}$ is constructed, indicating the presence of each concept in the ground-truth captions. Specifically, Y^C is a multi-hot label where $Y_i^C = 1$ if concept e_i is present and $Y_i^C = 0$ otherwise.

Multiple Instance Learning Multiple concept detection takes the video features $F = \{f_t\}_{t=1}^T$ as input and predicts both video-level and frame-level concepts. For frame-level concept detection, a shared fully connected layer followed by a sigmoid function predicts the probability:

$$p_t = \sigma(FC(f_t)), \quad (3)$$

where $p_t \in \mathbb{R}^{1 \times N_C}$ represents the concept probability at the frame t , $FC(\cdot)$ is the fully connected layer, and $\sigma(\cdot)$ is the sigmoid function. Then, to predict video-level concepts, temporal attention aggregates all frame probabilities:

$$P^v = \sum_{t=1}^T \alpha_t p_t, \quad \alpha = \text{softmax}(FW^{tp}), \quad (4)$$

where $P^v \in \mathbb{R}^{1 \times N_C}$ is the video concept probability, $W^{tp} \in \mathbb{R}^{d \times 1}$ is learnable parameter, $\alpha \in \mathbb{R}^T$ represents the temporal attention weights, and α_t is the weight at frame t .

Since only the video-level label Y^C is available, frame-level concept detection is performed in a weakly supervised manner using multiple instance learning:

$$\mathcal{L}_{mil} = - \sum_{i=1}^{N_C} [Y_i^C \log P_i^v + (1 - Y_i^C) \log(1 - P_i^v)], \quad (5)$$

where P_i^v represents the probability of the i -th concept.

Considering the high correlations between concepts and video events, the concept probabilities are weighted with learnable concept embeddings $W^C \in \mathbb{R}^{N_C \times d}$ to derive frame-level and video-level concepts:

$$f_t^c = p_t W^C, \quad f^{vc} = P^v W^C, \quad (6)$$

where $f^{vc} \in \mathbb{R}^{1 \times d}$ represents the video concept, and $f_t^c \in \mathbb{R}^{1 \times d}$ is the frame concept at time t . Then, each frame concept is combined with its corresponding frame feature:

$$\tilde{f}_t = f_t^c + f_t, \quad (7)$$

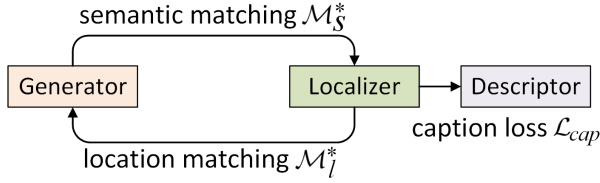


Figure 3: Cyclic co-learning.

where $\tilde{f}_t \in \mathbb{R}^{1 \times d}$ is the enhanced frame feature. The enhanced video features $\tilde{F} = \{\tilde{f}_t\}_{t=1}^T$ are used for event localization and caption generation.

Concept Contrastive Given that concept detection is weakly supervised, contrastive learning is introduced to obtain discriminative concept embeddings, enabling the model to accurately distinguish various concepts in videos. For each video, positive and negative concepts are generated according to the ground-truth concept label Y^C . A positive concept is defined as one that has at least one intersection with the ground-truth label, while a negative concept has none. During training, N_S pairs of positive and negative concept labels are sampled: Y^{C^+} and Y^{C^-} . By combining the sampled labels and the concept embeddings W^C as defined in Eq. (6), we derive N_S positive and negative concepts: $\{f_n^{vc^+}\}_{n=1}^{N_S}$ and $\{f_n^{vc^-}\}_{n=1}^{N_S}$. The triplet loss in the concept embedding space is defined as:

$$\mathcal{L}_{tri} = \frac{1}{N_S} \sum_{n=1}^{N_S} \left[\text{sim}(f_n^{vc}, f_n^{vc^-}) - \text{sim}(f_n^{vc}, f_n^{vc^+}) + \delta \right]_+, \quad (8)$$

where $[\cdot]_+ = \max(0, \cdot)$, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and δ denotes the margin. This loss encourages the model to learn concept embeddings such that the distance between the video concept and the positive concept is minimized, while the distance to the negative concept is maximized.

Cyclic Co-learning

The captioning network, based on PDVC (Wang et al. 2021), consists of three components: a deformable transformer (generator), a localization head (localizer), and a captioning head (descriptor). The generator takes video features \tilde{F} and N learnable event queries $\{q_i\}_{i=1}^N$ as input, and outputs updated queries $\{\tilde{q}_i\}_{i=1}^N$ containing event semantics and temporal information. For each query \tilde{q}_i , the localizer predicts the event location l_i (start time and end time), and the descriptor produces the caption $C_i = \{c_{i,1}, \dots, c_{i,L}\}$, where L is the sentence length. As illustrated in Fig.3, the generator and localizer co-learn in a cycle: the generator perceives events and guides the localizer via semantic matching, while the localizer provides feedback through location matching to enhance the generator’s semantic perception. Finally, the descriptor produces captions based on the queries selected by the localizer.

Given N^* ground truth events $\{l_j^*, C_j^*\}_{j=1}^{N^*}$, where l_j^* is the j -th event location and C_j^* is the corresponding caption. To match the predicted events with ground truth in a

global scheme, the localizer matches the predicted locations $\{l_i\}_{i=1}^N$ with the ground truth locations $\{l_j^*\}_{j=1}^{N^*}$ (location matching) using the Hungarian algorithm (Kuhn 1955):

$$\mathcal{M}_l^* = \arg \min_{\mathcal{M}} \sum_{(i,j) \in \mathcal{M}} \text{cost}(l_i, l_j^*), \quad (9)$$

where \mathcal{M} represents all possible matchings, and $\text{cost}(\cdot, \cdot)$ is the gIOU (Rezatofighi et al. 2019) cost. The optimal location matching $\mathcal{M}_l^* = \{(i, j) | i = \pi(j), j \in \{1, \dots, N^*\}\}$, where $\pi(j)$ returns the index of the predicted location that matches the ground truth location l_j^* . Based on the matching set \mathcal{M}_l^* , the localization-guided loss \mathcal{L}_{lg} is defined:

$$\mathcal{L}_{lg} = \frac{1}{|\mathcal{M}_l^*|} \sum_{(i,j) \in \mathcal{M}_l^*} (\lambda_1 \mathcal{L}_{giou}(l_i, l_j^*) + \lambda_2 \mathcal{L}_{cap}(C_i, C_j^*)), \quad (10)$$

where λ_1 and λ_2 are hyperparameters, \mathcal{L}_{giou} is the gIOU localization loss for localizer, and \mathcal{L}_{cap} denotes the cross-entropy caption loss for descriptor. Obviously, the localizer selects queries for the descriptor based on location matching and improves the descriptor’s captioning ability. However, without semantic guidance, the localizer may not effectively distinguish between semantically distinct nearby events. Thus, semantic matching is introduced in the generator:

$$\mathcal{M}_s^* = \arg \max_{\mathcal{M}} \sum_{(i,j) \in \mathcal{M}} \text{sim}(\tilde{q}_i, z_j), \quad (11)$$

where z_j is the semantic representation of C_j^* from the CLIP text encoder, and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity. To enhance the generator’s semantic perception, the cosine distance between the queries selected by the localizer and the ground-truth semantics is minimized:

$$\mathcal{L}_{sem} = \frac{1}{|\mathcal{M}_l^*|} \sum_{(i,j) \in \mathcal{M}_l^*} (1 - \text{sim}(\tilde{q}_i, z_j)). \quad (12)$$

Using the optimal semantic matching \mathcal{M}_s^* , the semantic-guided loss \mathcal{L}_{sg} is defined as:

$$\begin{aligned} \mathcal{L}_{sg} &= \frac{1}{|\mathcal{M}_s^*|} \sum_{(i,j) \in \mathcal{M}_s^*} (\lambda_1 \mathcal{L}_{giou}(l_i, l_j^*) + \lambda_2 \mathcal{L}_{cap}(C_i, C_j^*)) \\ &+ \lambda_3 \mathcal{L}_{sem}, \end{aligned} \quad (13)$$

Semantic matching encourages the localizer and descriptor to perceive events from a semantic perspective. Nevertheless, it is sensitive to temporally distinct but semantically similar events (similar events recurring temporally). To address this, semantic and location matching are combined to derive the cyclic loss:

$$\begin{aligned} \mathcal{L}_{cyc} &= \frac{1}{|\mathcal{M}_s^*|} \sum_{(i,j) \in \mathcal{M}_s^*} \lambda_1 \mathcal{L}_{giou}(l_i, l_j^*) \\ &+ \frac{1}{|\mathcal{M}_l^*|} \sum_{(i,j) \in \mathcal{M}_l^*} \lambda_2 \mathcal{L}_{cap}(C_i, C_j^*) \\ &+ \lambda_3 \mathcal{L}_{sem}. \end{aligned} \quad (14)$$

In the cyclic loss \mathcal{L}_{cyc} , the generator improves the localizer’s accuracy through semantic matching (the first term). The localizer, in turn, provides feedback to refine the generator’s event semantic perception via location matching (the third term). This strategy enables mutual benefits between semantic perception and event localization.

Models	Ground-truth proposals				Learned proposals			
	B@4	M	C	S	B@4	M	C	S
<i>visual feature C3D</i>								
HRNN	1.59	8.81	24.17	-	0.70	5.68	12.35	-
DCE	1.60	8.88	25.12	-	0.71	5.69	12.43	-
DVC	1.62	10.33	25.24	-	0.73	6.93	12.61	-
GPaS	1.53	11.04	28.20	-	0.93	7.44	13.00	-
E2E-MT	2.71	11.16	47.71	-	1.15	4.98	9.25	4.02
PDVC [†]	2.74	10.34	48.55	8.94	1.65	7.50	25.87	5.26
<i>visual C3D + audio VGGish</i>								
MDVC	1.98	11.07	45.39	-	1.01	7.46	7.38	-
BMT	1.99	10.90	42.74	-	1.88	8.44	11.35	-
GS-MS-FTN	1.91	10.93	-	-	1.64	8.69	-	-
PDVC [†]	2.86	10.27	48.64	9.70	1.96	8.08	28.59	5.42
<i>visual feature CLIP</i>								
Vid2Seq	-	-	-	-	-	8.50	30.10	5.80
DIBS	-	-	-	-	-	8.93	31.89	5.85
PDVC [†]	3.29	11.53	57.17	10.16	1.75	7.63	30.22	6.16
CM ^{2†}	3.55	12.06	58.97	10.02	2.29	8.51	32.89	6.21
MCCL	3.89	12.52	63.01	10.35	2.68	9.05	34.92	6.16

Table 1: Caption results on ActivityNet Captions. B@4, M, C, S denote BLEU4, METEOR, CIDEr and SODA.c, respectively. † indicates reproduced from official code.

Training

The overall loss is obtained:

$$\mathcal{L} = \mathcal{L}_{cyc} + \lambda_4 \mathcal{L}_{tri} + \lambda_5 \mathcal{L}_{mil}, \quad (15)$$

where λ_4 , and λ_5 are hyperparameters.

Experiment

Experimental Setup

Dataset Experiments are conducted on two popular benchmark datasets, ActivityNet Captions (Krishna et al. 2017) and YouCook2 (Zhou, Xu, and Corso 2018), to demonstrate the effectiveness of the proposed method. ActivityNet Captions contains 20k untrimmed videos of human activities, averaging 120 seconds per video with 3.65 temporally localized sentences, totaling 100k sentences. We use the official split: 10,009/4,925/5,044 videos for training, validation, and testing. YouCook2 includes 2k untrimmed cooking videos, each averaging 320 seconds and annotated with 7.7 sentences. We follow the official split: 1,333/457/210 videos for training, validation, and testing.

Evaluation Metrics The method is evaluated from two aspects: 1) For dense captioning performance, evaluation tools from the ActivityNet Challenge 2018 are used, measuring CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), BLEU4 (Papineni et al. 2002), and METEOR (Banerjee and Lavie 2005) scores. These metrics assess the average precision of the matched pairs between generated captions and ground truth at IOU thresholds of {0.3, 0.5, 0.7, 0.9}. Additionally, SODA.c (Fujita et al. 2020) is used to evaluate storytelling ability. 2) For localization performance, the harmonic mean of average precision and average recall at IOU

Models	Ground-truth proposals				Learned proposals			
	B@4	M	C	S	B@4	M	C	S
<i>Pretrain</i>								
Vid2Seq	-	-	-	-	-	9.30	47.10	7.90
DIBS	-	-	-	-	-	7.51	44.44	6.39
<i>without Pretrain</i>								
PDVC [†]	3.79	14.05	86.39	16.38	1.62	5.72	30.82	5.03
CM ^{2†}	3.56	14.09	89.43	15.10	1.64	5.56	30.15	5.21
MCCL	4.69	14.69	96.95	16.85	2.04	6.53	36.09	5.21

Table 2: Caption results on YouCook2 dataset. † denotes the results reproduced from official code.

	Event Captions			Event Localization		
	B@4	M	C	Recall	Precision	F1
PDVC [†]	1.75	7.63	30.22	53.69	55.16	54.41
CM ^{2†}	2.29	8.51	32.89	53.73	56.14	54.91
MCCL(\mathcal{L}_{sg})	1.36	6.75	23.43	46.50	61.04	52.79
MCCL(\mathcal{L}_{lg})	2.23	8.56	32.15	52.15	56.61	54.28
MCCL(\mathcal{L}_{cyc})	2.68	9.05	34.92	53.19	57.36	55.23

Table 3: Localization performance on ActivityNet Captions. † denotes the results reproduced from official code.

thresholds of {0.3, 0.5, 0.7, 0.9}, along with the F1 score, is used.

Implementation Details The experiments are developed with Python 3.9 and PyTorch 1.12 and performed on a single RTX 4090 GPU. For both datasets, each video is uniformly sampled or interpolated to T frames, with T set to 100 for ActivityNet Captions and 200 for YouCook2. CLIP ViT-L/14’s image encoder extracts 768-dimensional frame features, while its text encoder extracts 768-dimensional sentence features from the corpus. For both datasets, the number of chunks is set to $W = 20$, the number of retrieved sentences per chunk N_K is set to 10. The number of concepts is set to $N_C = 500$ for ActivityNet Captions and $N_C = 600$ for YouCook2. The number of positive and negative samples N_S is set to 10, with the margin δ is set to 0.5. The number of event queries is set to $N = 10$ for ActivityNet Captions and $N = 100$ for YouCook2. The hyperparameters λ_1 , λ_2 , λ_3 , λ_4 , and λ_5 are set to 4, 1, 0.5, 1, and 1, respectively.

Experimental Results

We compare the performance of our method with that of state-of-the-art methods based on three types of approaches: (1) Methods using C3D features: HRNN (Venugopalan et al. 2015), DCE (Krishna et al. 2017), DVC (Li et al. 2018), GPaS (Zhang et al. 2020), E2E-MT (Zhou et al. 2018), and PDVC (Wang et al. 2021). (2) Methods using multimodal features: MDVC (Iashin and Rahtu 2020b), BMT (Iashin and Rahtu 2020a), GS-MS-FTN (Xie et al. 2023b), and PDVC (Wang et al. 2021). (3) Methods using CLIP features: PDVC (Wang et al. 2021), Vid2Seq (Yang et al. 2023), DIBS (Wu et al. 2024), and CM² (Kim et al. 2024).

N_K	B@4	M	C	SODA_c
5	2.41	8.84	32.03	5.97
10	2.68	9.05	34.92	6.16
15	2.61	8.93	35.21	6.14
20	2.53	8.69	34.68	6.24
30	2.56	8.87	34.90	6.34
50	2.52	8.71	34.32	6.28

(a) Number of retrieved sentences N_K .

N_C	B@4	M	C	SODA_c
300	1.78	8.09	27.86	5.96
400	2.41	8.65	33.02	6.12
500	2.68	9.05	34.92	6.16
600	2.75	9.02	34.41	6.01
700	2.62	8.81	32.53	6.06
800	2.15	8.54	32.75	6.12

(b) Number of concepts N_C .

N_S	B@4	M	C	SODA_c
0	2.01	8.25	30.11	5.96
5	2.42	8.75	34.69	6.20
10	2.68	9.05	34.92	6.16
15	2.65	8.89	34.83	6.12
20	2.66	8.87	34.41	6.14
30	2.42	8.60	33.73	6.18

(c) Number of positive and negative samples N_S .

Table 4: Hyperparameter studies on the ActivityNet Captions dataset.

V2T	MCD	CyC	B@4	M	C	SODA_c
✗	✗	✗	1.93	7.68	27.74	5.61
✓	✗	✗	2.26	8.54	31.32	6.01
✗	✓	✗	2.01	8.03	29.33	5.91
✗	✗	✓	2.28	8.39	30.98	5.96
✓	✓	✗	2.32	8.55	32.97	6.24
✓	✓	✓	2.68	9.05	34.92	6.16

Table 5: Performance of different components. V2T, MCD, and CyC denote video-to-text retrieval, multiple concept detection, and cyclic co-learning, respectively.

Dense Captioning Performance As shown in Tab.1, the performance of MCCL on ActivityNet Captions is reported. Models utilizing CLIP features as input outperform other methods, indicating that CLIP features provide superior semantic understanding. With ground-truth proposals, MCCL surpasses state-of-the-art models across all metrics, achieving a significant 4.04% improvement in CIDEr. With learned proposals, despite a lower SODA_c compared to CM², MCCL achieves the best results in BLEU4, METEOR, and CIDEr, showing a 2.03% improvement in CIDEr.

Tab.2 presents the performance of MCCL on YouCook2 dataset. Notably, Vid2seq (Yang et al. 2023) achieves the best performance by using an additional 15M videos for pre-training, and DIBS (Wu et al. 2024) shows sub-optimal results with an extra 56k videos for pretraining. In contrast, our method, similar to PDVC (Wang et al. 2021) and CM², does not involve any extra pretraining. With ground-truth proposals, MCCL outperforms competitors across all metrics, even achieving a notable 7.52% improvement in CIDEr. With learned proposals, MCCL achieves the best results in BLEU4, METEOR, and CIDEr, demonstrating a 5.94% improvement in CIDEr compared to CM².

Event Localization Performance Tab.3 compares MCCL with methods that use CLIP features as input on ActivityNet Captions datasets. We evaluate the impact of the cyclic loss (\mathcal{L}_{cyc}) by comparing it with the semantic-guided loss (\mathcal{L}_{sg}) and the localization-guided loss (\mathcal{L}_{lg}). \mathcal{L}_{sg} achieves high precision but lower recall due to its reliance on semantic matching, making it challenging to handle semantically similar but temporally distinct events in complex videos. \mathcal{L}_{lg} , on the other hand, focuses on location information to

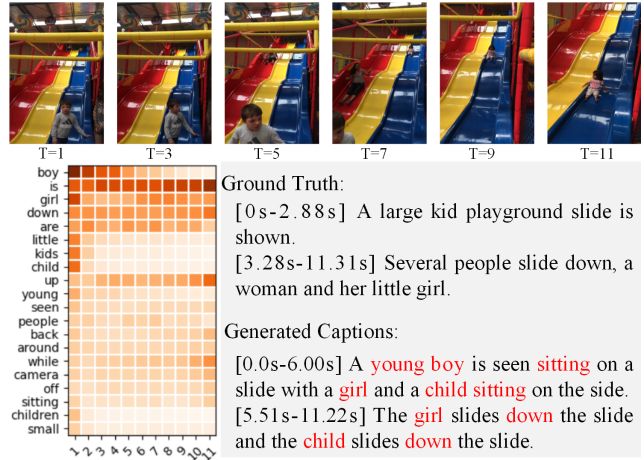


Figure 4: Concept guidance for video captioning.

	B@4	M	C	SODA_c
MCCL(w/o concept)	2.57	8.76	32.46	5.71
MCCL(w/ concept)	2.68	9.05	34.92	6.16

Table 6: Effect of concept features on ActivityNet Captions.

identify temporal boundaries, significantly improving recall. This highlights the importance of location information in understanding complex videos. Finally, \mathcal{L}_{cyc} enhances the interaction and optimization between the generator and localizer, helping the localizer identify temporal boundaries under semantic guidance and the generator perceive event semantics based on localizer’s feedback. This mutual reinforcement of semantics and localization results in the best F1 score.

Ablation Studies

Ablation of Hyperparameters As shown in Tab.4a, a small number of sentences is sufficient to achieve good performance. However, retrieving too many sentences results in higher SODA_c, but lower METEOR and CIDEr scores. We hypothesize that a larger number of sentences may enrich semantics, which benefits SODA_c, but also introduces more irrelevant information, which affects caption quality.

In Tab.4b, the impact of the number of concepts on cap-

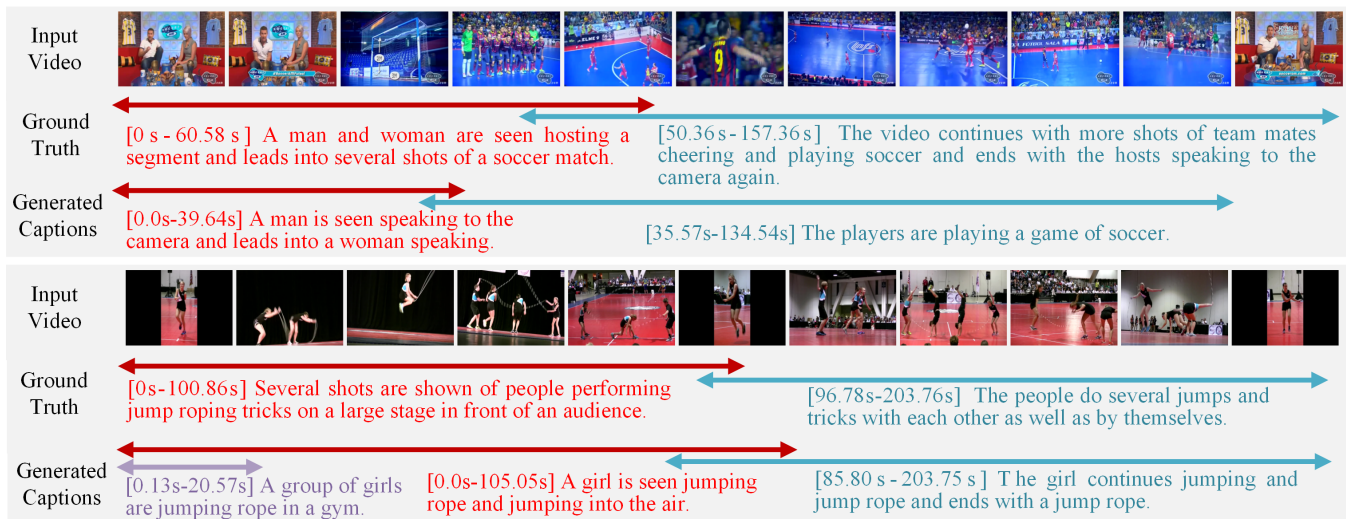


Figure 5: Qualitative captioning results. Two examples from ActivityNet Captions. The ground truth and generated captions for each video are shown separately.

tioning performance is illustrated. It can be observed that an appropriate number of concepts improves the model’s performance, whereas too many concepts may somewhat degrade caption quality.

Tab.4c shows the impact of concept contrastive learning. When the number of samples is set to $N_S = 0$, it indicates that no contrastive learning is applied. The results show that concept contrastive learning significantly enhances performance by developing discriminative concepts, which help in identifying visual targets and provide cues for caption generation.

Ablation Study of Different Components Tab.5 compares the functionality of different components. Through ablation studies on individual modules, it can be seen that integrating video-text retrieval (V2T) significantly enhances video captioning performance compared to the baseline without any components. This demonstrates the effectiveness of leveraging external knowledge to improve video understanding. While concept detection (MCD) shows limited standalone impact, its combination with V2T significantly improves all metrics, highlighting the crucial role of conceptual information in identifying visual targets and providing valuable event cues for caption generation. Cyclic co-learning (CyC) further enhances semantic perception and event localization, yielding notable gains in CIDEr and METEOR. Finally, integrating all modules achieves the best overall performance, despite a slight decrease in the SODA_c metric.

Qualitative Results

Effect of Concept Guidance To see how concepts guide video captions, Fig.4 presents an example. The 11-second video uses the top 20 video-level concepts as reference, with frame-level probabilities visualized. MCCL identifies concepts such as ‘boy’, ‘girl’, and ‘child’, as well as actions like ‘sitting’ and ‘down’, providing valuable temporal cues for

understanding the video. Compared to the ground truth captions, our concept detection offers more fine-grained explanations. Although there is a misidentification of ‘woman’ as ‘girl’ and a lack of recognition of their relationship, we believe that discerning object relationships from visuals alone is challenging. Additional information, such as audio, could enhance video understanding. In addition, the comparison of experiments with and without incorporating concept features into video features is provided in Tab.6. It is clear that MCCL with concept features shows improvements in all metrics.

Captioning Results Fig.5 shows two prediction examples from MCCL. It can be observed that our method effectively captures the temporal boundaries of events, leading to more accurate event localization. Additionally, the cyclic loss effectively handles complex videos, enhancing both event detection and caption quality.

Conclusion

In this paper, we propose a Multi-Concept Cyclic Learning (MCCL) network for dense video captioning, capable of performing frame-level multi-concept detection and enhancing video features to capture temporal event cues. Additionally, we introduce a cyclic co-learning strategy where the generator and localizer mutually reinforce each other. Specifically, the generator uses semantic information to guide the localizer in event localization, while the localizer provides feedback through location matching to refine the generator’s semantic perception. This mutual enhancement improves both semantic perception and event localization.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.61976247).

References

- Aafaq, N.; Mian, A.; Akhtar, N.; Liu, W.; and Shah, M. 2022. Dense video captioning with early linguistic information fusion. *IEEE Transactions on Multimedia*, 25: 2309–2322.
- Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.
- Chang, Z.; Zhao, D.; Chen, H.; Li, J.; and Liu, P. 2022. Event-centric multi-modal fusion method for dense video captioning. *Neural Networks*, 146: 120–129.
- Chen, J.; Pan, Y.; Li, Y.; Yao, T.; Chao, H.; and Mei, T. 2023. Retrieval augmented convolutional encoder-decoder networks for video captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1s): 1–24.
- Chen, S.; and Jiang, Y.-G. 2021. 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*, 8425–8435.
- Deng, C.; Chen, Q.; Qin, P.; Chen, D.; and Wu, Q. 2023. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15648–15658.
- Deng, C.; Chen, S.; Chen, D.; He, Y.; and Wu, Q. 2021. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 234–243.
- Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31.
- Fujita, S.; Hirao, T.; Kamigaito, H.; Okumura, M.; and Nagata, M. 2020. 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*, 517–531. Springer.
- Gao, L.; Wang, X.; Song, J.; and Liu, Y. 2020. Fused GRU with semantic-temporal attention for video captioning. *Neurocomputing*, 395: 222–228.
- Iashin, V.; and Rahtu, E. 2020a. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. In *The 31st British Machine Vision Virtual Conference*.
- Iashin, V.; and Rahtu, E. 2020b. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 958–959.
- Jing, S.; Zhang, H.; Zeng, P.; Gao, L.; Song, J.; and Shen, H. T. 2023. Memory-based augmentation network for video captioning. *IEEE Transactions on Multimedia*.
- Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Viewpoint-Agnostic Change Captioning with Cycle Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2075–2084.
- Kim, M.; Kim, H. B.; Moon, J.; Choi, J.; and Kim, S. T. 2024. Do You Remember? Dense Video Captioning with Cross-Modal Memory Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13894–13904.
- Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.
- Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.
- Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7492–7500.
- Lu, Y.; Zhang, Z.; Yuan, C.; Li, P.; Wang, Y.; Li, B.; and Hu, W. 2024. Set Prediction Guided by Semantic Concepts for Diverse Video Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3909–3917.
- Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.
- Mun, J.; Yang, L.; Ren, Z.; Xu, N.; and Han, B. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6588–6597.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rahman, T.; Xu, B.; and Sigal, L. 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8908–8917.
- Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.
- Ryu, H.; Kang, S.; Kang, H.; and Yoo, C. D. 2021. Semantic grouping network for video captioning. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2514–2522.
- Shen, Z.; Li, J.; Su, Z.; Li, M.; Chen, Y.; Jiang, Y.-G.; and Xue, X. 2017. Weakly supervised dense video captioning.

- In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1916–1924.
- Suin, M.; and Rajagopalan, A. 2020. An efficient framework for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12039–12046.
- Tian, Y.; Hu, D.; and Xu, C. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2745–2754.
- Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.
- Tu, Y.; Li, L.; Yan, C.; Gao, S.; and Yu, Z. 2021. R³Net: Relation-embedded Representation Reconstruction Network for Change Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9319–9329.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 4534–4542.
- Wang, J.; Jiang, W.; Ma, L.; Liu, W.; and Xu, Y. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7190–7198.
- Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.
- Wang, T.; Zheng, H.; Yu, M.; Tian, Q.; and Hu, H. 2020. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1890–1900.
- Wu, B.; Liu, B.; Huang, P.; Bao, J.; Xi, P.; and Yu, J. 2023. Concept parser with multimodal graph learning for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4484–4495.
- Wu, H.; Liu, H.; Qiao, Y.; and Sun, X. 2024. DIBS: Enhancing Dense Video Captioning with Unlabeled Videos via Pseudo Boundary Enrichment and Online Refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18699–18708.
- Xie, C.-W.; Sun, S.; Xiong, X.; Zheng, Y.; Zhao, D.; and Zhou, J. 2023a. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19265–19274.
- Xie, Y.; Niu, J.; Zhang, Y.; and Ren, F. 2023b. Global-shared Text Representation based Multi-Stage Fusion Transformer Network for Multi-modal Dense Video Captioning. *IEEE Transactions on Multimedia*.
- Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10714–10726.
- Yang, B.; Cao, M.; and Zou, Y. 2023. Concept-aware video captioning: Describing videos with effective prior information. *IEEE Transactions on Image Processing*.
- Yue, S.; Tu, Y.; Li, L.; Gao, S.; and Yu, Z. 2024. Multi-grained Representation Aggregating Transformer with Gating Cycle for Change Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*.
- Zhang, Z.; Qi, Z.; Yuan, C.; Shan, Y.; Li, B.; Deng, Y.; and Hu, W. 2021. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9837–9846.
- Zhang, Z.; Xu, D.; Ouyang, W.; and Zhou, L. 2020. Dense video captioning using graph-based sentence summarization. *IEEE Transactions on Multimedia*, 23: 1799–1810.
- Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8739–8748.