Explicit Temporal-Semantic Modeling for Dense Video Captioning via Context-Aware Cross-Modal Interaction

Mingda Jia[1,2], Weiliang Meng[1,2*], Zenghuang Fu[1,2], Yiheng Li[1,2], Qi Zeng[1,2], Yifan Zhang[1,2], Ju Xin[3], Rongtao Xu[4], Jiguang Zhang[1,2*], Xiaopeng Zhang[1,2]

[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] The Navigation Guarantee Center of North China Sea
[4] Spatialtemporal AI
weiliang.meng@ia.ac.cn, jiguang.zhang@ia.ac.cn

Abstract

Dense video captioning jointly localizes and captions salient events in untrimmed videos. Recent methods primarily focus on leveraging additional prior knowledge and advanced multi-task architectures to achieve competitive performance. However, these pipelines rely on implicit modeling that uses frame-level or fragmented video features, failing to capture the temporal coherence across event sequences and comprehensive semantics within visual contexts. To address this, we propose an explicit temporal-semantic modeling framework called Context-Aware Cross-Modal Interaction (CACMI), which leverages both latent temporal characteristics within videos and linguistic semantics from text corpus. Specifically, our model consists of two core components: Cross-modal Frame Aggregation aggregates relevant frames to extract temporally coherent, event-aligned textual features through cross-modal retrieval; and Context-aware Feature Enhancement utilizes query-guided attention to integrate visual dynamics with pseudo-event semantics. Extensive experiments on the ActivityNet Captions and YouCook2 datasets demonstrate that CACMI achieves the state-of-the-art performance on dense video captioning task.

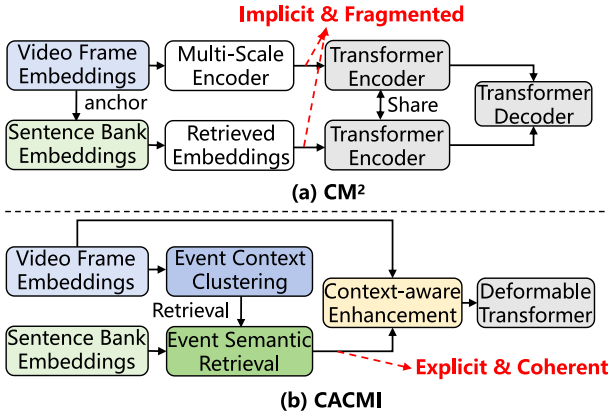

Figure 1: (a) CM^2 introduces a cross-modal memory-based model, the external sentence bank is specifically designed to select relevant implicit semantics. (b) Our CACMI harnesses explicit temporal-semantic information through context-aware cross-modal interaction to enhance the event localization and captioning performance.

Introduction

In recent years, video understanding has emerged as a rapidly growing focus within the fields of computer vision and multimodal analysis (Wang et al. 2023; Song et al. 2024; Nie et al. 2024; Chen et al. 2024; Li et al. 2024; Wang et al. 2024), with video captioning recognized as a foundational task. Traditional video captioning aims to generate a concise description that summarizes the main content of a video, and significant progress has been achieved in this area (Gao et al. 2017; Krishna et al. 2017; Pei et al. 2019; Seo et al. 2022). However, conventional methods often miss fine-grained details and struggle to handle multiple events or segments effectively. To overcome these limitations, dense video captioning (DVC) has been introduced, which seeks to produce descriptive annotations for all salient events in an untrimmed video, along with their precise temporal boundaries.

Dense video captioning methods typically begin by leveraging a pre-trained image encoder to extract visual features from input frames, followed by the detection of salient event boundaries within these features to achieve temporal localization and event representations (Zhou et al. 2018; Mun et al. 2019; Wang et al. 2021). With the rapid advancement of vision-language models, recent approaches have explored retrieval-augmented generation for video captioning, incorporating external semantic knowledge into the encoding-decoding pipeline to enhance understanding and generation capabilities. In dense video captioning, the representative work CM^2 (Kim et al. 2024) pioneered the integration of memory retrieval mechanism, effectively utilizing semantic cues from external sources to improve both event localization and caption generation.

Despite these advancements, recent memory-based methods depend on inherently implicit retrieval-augmented generation (RAG) frameworks (Chen et al. 2023; Ramos et al. 2023; Kim et al. 2024, 2025). These approaches employ manually designed windows for cross-modal retrieval at fragmented video segments, leading to two fundamental limitations: (i) Temporal modeling deficiency: Visual fea-

*Corresponding authors.

tures derived from fixed-size windows focus exclusively on localized segments, thus resulting in discontinuous semantic retrieval, (ii) Modality gap: Retrieved semantic features are fused with visual representations using simplistic operations (e.g., concatenation or basic attention mechanisms), which are inadequate for bridging the inherent divergence between visual and textual modalities, leading to inconsistencies that impair both localization accuracy and captioning quality.

To address these limitations, we propose that effective retrieval-augmented generation for dense video captioning involves exploiting the inherent temporal structure and rich semantic information embedded within video data. This is intuitively grounded in visual continuity: adjacent frames sharing similar visual and temporal contexts typically represent the same semantic event or action. Inspired by this observation, we introduce explicit temporal-semantic modeling based on pseudo events to enhance contextual coherence and yield retrieved text semantics with temporal characteristics. Furthermore, it is also essential to enhance visual representations using events rather than simply integrating frame-level or fragmented textual information.

As illustrated in Figure 1, we propose a novel framework called Context-aware Cross-Modal Interaction (CACMI), which leverages the explicit temporal-semantic structure in video data for the dense video captioning task. First, we design a Cross-modal Frame Aggregation (CFA) module, which consists of two components: Event Context Clustering employs a temporal constraint mechanism to integrate visual features that share contextual and temporal coherence, and Event Semantic Retrieval performs cross-modal matching between the clustered event features and sentence bank, extracting event-aligned textual information. Subsequently, a Context-aware Feature Enhancement (CFE) module is introduced to facilitate fine-grained integration of the visual features and retrieved textual features, enabling precise alignment of visual dynamics with linguistic semantics. Finally, we adopt a transformer encoder-decoder architecture with parallel multi-task heads to jointly perform event localization and caption generation. Our main contributions can be summarized as follows:

- We propose CACMI, a novel dense video captioning framework featuring explicit temporal-semantic modeling, which leverages Context-Aware Cross-Modal Interaction to fully exploit the rich temporal structure and semantic information inherent in videos.
- We introduce a Cross-modal Frame Aggregation module to extract temporally coherent, event-aligned semantic features through cross-modal retrieval, which is complemented by a Context-aware Feature Enhancement module that mitigates the visual-linguistic modality gap through query-guided feature refinement.
- Extensive experiments on the ActivityNet Captions and YouCook2 datasets demonstrate the effectiveness of our model in dense video captioning task, highlighting its superior event localization capability.

Related Work

Dense Video Captioning

Dense video captioning consists of two key subtasks: event localization and caption generation. Early methods typically followed a two-stage localization-description paradigm (Krishna et al. 2017; Wang et al. 2018, 2020), where salient temporal segments are first identified and then passed to language models for caption generation. However, this decoupled training strategy limits the mutual interaction between localization and captioning, hindering joint optimization. To overcome this, recent works adopt end-to-end frameworks that enable joint learning. PDVC (Wang et al. 2021) proposes a parallel decoding structure that shares intermediate representations across both subtasks. Vid2Seq (Yang et al. 2023) formulates dense video captioning as a sequence-to-sequence problem, leveraging transcribed speech from narrated videos as multimodal input and employing special text-time tokens to unify event detection and caption generation. CM^2 (Kim et al. 2024) introduces a memory bank mechanism that preserves end-to-end learning while incorporating external textual knowledge to enhance caption quality.

Despite these advancements, achieving precise event localization and comprehensive semantic understanding remains a significant challenge, particularly without extensive pretraining on large-scale video datasets. We propose a novel DVC approach that integrates the inherent temporal representations of videos with retrieved textual semantics to improve both localization accuracy and caption generation.

Retrieval-Augmented Captioning

Retrieval-augmented captioning enhances video understanding by integrating external textual knowledge, following the paradigm of Retrieval-Augmented Generation (Jing et al. 2023). In this setting, retrieved texts serve as supplementary context, enriching the original visual features and enabling deeper semantic comprehension. Recent DVC methods construct a text memory bank relevant to source videos, using retrieved sentences as multimodal input to boost both localization and captioning performance (Chen et al. 2023; Ramos et al. 2023; Kim et al. 2024, 2025).

However, a key limitation persists: current retrieval strategies rely on manually designed sliding windows for search matching, neglecting contextual relationships across video segments. This approach results in fragmented retrieval that fails to capture the inherent temporal structure and coherent semantics of video content. To overcome this, we propose a explicit temporal-semantic modeling framework for retrieval-augmented dense video captioning, which preserves the temporal structure of videos to ensure contextually consistent semantic retrieval.

Method

The goal of this study is to enhance event localization and caption generation from untrimmed videos by effectively leveraging the explicit temporal-semantic structure within video content. As illustrated in Figure 2, we propose a novel framework named CACMI (Context-Aware

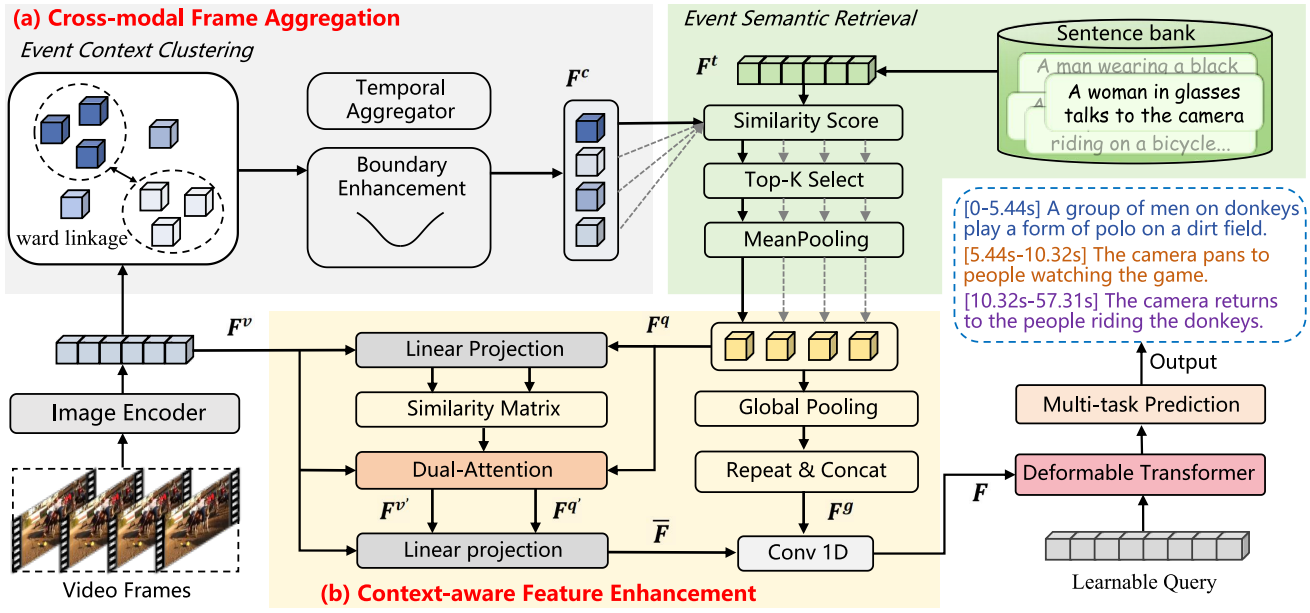

Figure 2: The overview of our CACMI framework. We employs a retrieval-augmented generation paradigm for DVC task. The pipeline begins with a pretrained CLIP image encoder extracting frame-level features. (a) Cross-modal Frame Aggregation (CFA). This module comprises two synergistic components: Event Context Clustering aggregates temporally and semantically consistent frame features to generate clustered event representations, and Event Semantic Retrieval matches relevant semantic information from a sentence bank via cosine similarity to produce retrieval-enhanced semantic features. (b) Context-aware Feature Enhancement (CFE). This module facilitates cross-modal interaction between retrieved textual features and visual representations, bridging the modality gap to generate enhanced frame features. Finally, a deformable transformer equipped with multi-task heads generates the joint outputs of event localization and captioning.

Cross-Modal Interaction), which incorporates external textual knowledge from sentence bank and enhances visual features with pseudo-event semantics. The model outputs a set of N tuples $(t_n^s, t_n^e, c_n)_{n=1}^N$, where N is the number of detected events in a video segment, t_n^s and t_n^e denote the start and end timestamps of the n-th event, and c_n is the corresponding textual description.

Cross-modal Frame Aggregation

Event Context Clustering. We utilize pre-trained CLIP ViT-L/14 (Dosovitskiy et al. 2020; Radford et al. 2021) to extract frame-level visual features $\mathbf{F}^v \in \mathbb{R}^{L\times d}$, where L and d are the number of clips and the feature dimension, respectively. Within a video, frames associated with the same event often share similar background and foreground, leading to high similarity in their encoded representations. To capture the temporal-semantic correlations across frames, we apply agglomerative clustering to $\mathbf{F}^v$ at the frame level. Agglomerative clustering does not assume a fixed cluster shape, making it suitable for discovering flexible and diverse patterns in feature space. We adopt Euclidean distance as the similarity metric, which efficiently captures smooth variations in frame-level features. For cluster merging, we employ Ward linkage to minimizes the increase in within-cluster variance during hierarchical merging. This combination encourages the formation of compact and semantically coherent clusters with high intra-cluster similarity.

To further enhance temporal coherence, we incorporate a temporal aggregation constraint. Specifically, after clustering, we ensure that any two frames within a cluster are no more than $\mathbf{t}_{\max}$ apart in time. Frames that exceed this temporal threshold are assigned to a new cluster. This constraint guarantees that the resulting clusters maintain both semantic similarity and temporal continui. Finally, the video is segmented into c clusters, each corresponding to a potential pseudo-event. The output is a set of cluster-level feature vectors $\mathbf{F}^c = \{C_i\}_{i=1}^c$, where each C_i represents the boundary-enhanced average of features within the i-th cluster. Specifically, we generate a bell-shaped weight distribution centered at the cluster's midpoint and then invert and normalize these weights, thereby assigning higher importance to features near the cluster boundaries.

Event Semantic Retrieval. To facilitate efficient cross-modal matching with an external textual corpus, we first preprocess the sentence bank using the CLIP text encoder, obtaining a set of textual feature embeddings $\mathbf{F}^t \in \mathbb{R}^{M\times d}$, where M is the total number of sentences in the corpus. We then compute a cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{c\times M}$ between each pseudo-event visual feature and all text features:

$$\mathbf{S} = \frac{\mathbf{F}^c \, \mathbf{F}^{t\top}}{||\mathbf{F}^c||\,||\mathbf{F}^t||} \tag{1}$$

Due to the large size of the sentence bank, we apply top-k retrieval to each row of the similarity matrix $\mathbf{S}$, selecting

the k most relevant textual features for each pseudo-event. These retrieved features are aggregated into a condensed representation $\mathbf{F}^s \in \mathbb{R}^{c\times k\times d}$:

$$\mathbf{F}^s = \text{Top-K}(\mathbf{S}) \tag{2}$$

To form a unified representation for each event, we perform average pooling over the top-k text features, resulting in the final retrieved semantic features $\mathbf{F}^q \in \mathbb{R}^{c\times d}$:

$$\mathbf{F}^q = \text{MeanPooling}(\mathbf{F}^s) \tag{3}$$

Context-aware Feature Enhancement

As shown in Figure 2, we introduce a fine-grained cross-modal fusion module to facilitate interactive refinement between retrieved textual features and visual representations. While CM^2 (Kim et al. 2024) employs a transformer encoder-decoder architecture with shared self-attention weights for feature enhancement, this parameter-sharing scheme is insufficient for bridging the inherent semantic gap between visual and linguistic modalities. Visual features often contain substantial noise that is misaligned with textual context, and naive fusion methods such as simple addition and concatenation fail to capture fine-grained and semantically aligned visual information.

To address this challenge, we adopt a query-guided multi-modal fusion module inspired by (Xiong, Zhong, and Socher 2016; Sun et al. 2024), which leverages textual queries to selectively suppress irrelevant visual elements and enhance semantically aligned regions, thereby enabling more accurate cross-modal alignment. The detailed structure of the module is shown in Figure 2. We begin by computing a similarity matrix between the frame-level visual features $\mathbf{F}^v$ and the event-level textual queries $\mathbf{F}^q$:

$$M = \frac{LP(\mathbf{F}^v)\, LP(\mathbf{F}^q)^\mathsf{T}}{\sqrt{d}}, \tag{4}$$

where $M \in \mathbb{R}^{L\times N}$ denotes the similarity matrix, and LP represents the linear projection layer. Using M, we compute the dual-attention features:

$$\mathbf{F}^{v'} = M_{col}\, \mathbf{F}^q, \tag{5}$$

$$\mathbf{F}^{q'} = M_{row}\, M_{col}^\mathsf{T}\, \mathbf{F}^v, \tag{6}$$

where M_{col} and M_{row} are the column-wise and row-wise softmax-normalized versions of M, respectively. Here, $\mathbf{F}^{v'}$ captures event-level visual context guided by text, while $\mathbf{F}^{q'}$ refines the query representation based on visual input.

Next, we concatenate the original frame features $\mathbf{F}^v$ with the cross-attended features $\mathbf{F}^{v'}$ and $\mathbf{F}^{q'}$, followed by a linear projection to obtain the refined visual features $\overline{\mathbf{F}}$:

$$\overline{\mathbf{F}} = LP([\,\mathbf{F}^v \,|\, \mathbf{F}^{v'} \,|\, \mathbf{F}^{q'}\,]), \tag{7}$$

where $[\cdot|\cdot]$ denotes feature concatenation. To incorporate global semantic guidance, we apply average pooling on $\mathbf{F}^q$ to obtain a global text vector $\mathbf{F}^g$, and replicate it across all frames to match the temporal dimension of $\mathbf{F}^v$. Finally, we fuse this global context with $\overline{\mathbf{F}}$ via channel-wise 1D convolution to produce the enhanced frame-level features:

$$\mathbf{F} = Conv_{1D}([\,\mathbf{F}^g \,|\, \overline{\mathbf{F}}\,]) \tag{8}$$

Multi-task Prediction

For event prediction, our CACMI framework incorporates a deformable transformer (Zhu et al. 2020) module followed by parallel multi-head predictors. The deformable transformer takes video features F and a set of learnable queries $\{q_i\}_{i=1}^N$ as input, which produces semantic and temporal representations of events with encoder-decoder framework. Given the extracted event features, three separate prediction heads are utilized for dense video captioning process.

Localization Head. The localization head consists of a multi-layer perceptron that predicts the temporal boundaries of events for each query. Specifically, it regresses the event center and temporal span, producing outputs in the form of tuples $(t_i^s, t_i^e, c_i)_{i=1}^N$, where t_i^s and t_i^e denote the predicted start and end times, and c_i is the confidence score representing foreground probability.

Captioning Head. The backbone of the captioning head is an LSTM augmented with deformable soft attention around the predicted reference points (Wang et al. 2021). At each decoding step t, the LSTM receives the context features $a_{i,t}$, the corresponding event query q_i, and the previous word $w_{i,t-1}$ to predict the next word $w_{i,t}$. This process continues until the full caption $S_i = w_{i,1}, ..., w_{i,S}$ is generated for the i-th event, where S denotes the caption length.

Event Counter. The event counter is designed to predict the number of events in an input video. To achieve this, essential information from the event query Q is first compressed using a max-pooling layer, followed by a fully-connected layer that outputs a fixed-size vector f. Each dimension of f corresponds to a possible event count. During inference, the predicted number of events is given by $N = argmax(f)$. Finally, we employ the Hungarian algorithm to match the predicted and ground-truth event tuples $(t_n^s, t_n^e, c_n)_{n=1}^N$, minimizing the matching loss defined as:

$$L_{\text{match}} = L_{\text{cls}} + \alpha L_{\text{loc}}, \tag{9}$$

where L_{cls} is the focal classification loss, and L_{loc} is the generalized IoU loss measuring the alignment between predicted and ground-truth temporal segments.

Loss Function

The overall training objective integrates four loss components: L_{cls}, L_{loc}, L_{count}, and L_{cap}. Specifically, L_{cls} is the loss between predicted event classification and ground-truth labels, L_{loc} is the generalized IoU loss for temporal boundary regression, L_{count} is the cross-entropy loss for event count prediction, and L_{cap} is the cross-entropy loss for word prediction across the generated captions. The final loss is a weighted sum of these components:

$$L = \alpha_{\text{cls}} L_{\text{cls}} + \alpha_{\text{loc}} L_{\text{loc}} + \alpha_{\text{count}} L_{\text{count}} + \alpha_{\text{cap}} L_{\text{cap}} \tag{10}$$

Experiments

Dataset

We evaluate our CACMI on two widely-used benchmark datasets for dense video captioning: ActivityNet Captions (Krishna et al. 2017) and YouCook2 (Zhou, Xu,

Models	PT	B4↑	M↑	C↑	S↑
UEDVC (ECCV'22)	✓	-	-	-	5.50
Vid2Seq (CVPR'23)	✓	-	8.50	30.10	5.80
OmniVID (CVPR'24)	✓	1.73	7.54	26.00	5.60
PDVC† (ICCV'21)	×	2.21	8.06	29.97	5.92
CM^{2†} (CVPR'24)	×	2.38	8.55	33.01	<u>6.18</u>
E^{2}DVC (CVPR'25)	×	<u>2.43</u>	<u>8.57</u>	<u>33.63</u>	6.13
CACMI (Ours)	×	**2.44**	**8.68**	**33.80**	**6.39**

Table 1: Performance of Event Captioning in ActivityNet Captions. B4, M, C, S denote BLEU4, METEOR, CIDEr and SODA_c, respectively. † indicates reproduced from official code. Bold means the highest score. Underline means 2nd score. PT denotes whether pretraining is conducted using additional video data.

Models	PT	B4↑	M↑	C↑	S↑
Vid2Seq (CVPR'23)	✓	-	9.30	47.10	7.90
PDVC† (ICCV'21)	×	1.40	5.56	29.69	4.92
CM^{2†} (CVPR'24)	×	1.63	6.08	31.66	5.34
E^{2}DVC (CVPR'25)	×	<u>1.68</u>	<u>6.11</u>	<u>34.26</u>	<u>5.39</u>
CACMI (Ours)	×	**1.70**	**6.21**	**34.83**	**5.57**

Table 2: Performance of Event Captioning in YouCook2. B4, M, C, S denote BLEU4, METEOR, CIDEr and SODA_c, respectively. † indicates reproduced from official code. Bold means the highest score. Underline means 2nd score. PT denotes pretraining from additional video datasets.

Models	PT	ActivityNet Captions			YouCook2		
		F1↑	*Rec.*↑	*Pre.*↑	*F1*↑	*Rec.*↑	*Pre.*↑
Vid2Seq	✓	53.29	52.70	53.90	27.84	**27.90**	27.80
PDVC†	×	54.78	53.27	56.38	26.81	22.89	32.37
CM^{2†}	×	55.21	53.71	56.81	28.43	24.76	33.38
E^{2}DVC	×	<u>56.42</u>	<u>55.14</u>	<u>57.77</u>	<u>28.87</u>	25.01	<u>34.13</u>
CACMI (Ours)	×	**57.10**	**55.89**	**58.05**	**29.34**	<u>25.54</u>	**34.63**

Table 3: Performance of Event Localization in ActivityNet Captions and YouCook2 datasets. † denotes results reproduced from official implementation. Bold means the highest score. Underline means 2nd score. PT denotes pretraining from additional video datasets. Rec. and Pre. denote average recall and average precision, respectively.

and Corso 2018). ActivityNet Captions comprises approximately 20,000 untrimmed YouTube videos, spanning over 700 hours and encompassing a broad range of event categories such as sports, cooking, and social activities. Each video is annotated with an average of 3.7 temporally localized captions, resulting in over $100,000$ sentence-level annotations with precise timestamps. Following the standard data split, we use $10,024$ videos for training, $4,926$ for validation, and $5,044$ for testing. YouCook2 focuses on instructional cooking videos and contains $2,000$ untrimmed YouTube videos, totaling 176 hours of content. Each video is densely annotated with an average of 7.7 captions, offering fine-grained temporal-textual alignments to support procedural video understanding. We follow the standard dataset split for training, validation, and testing purposes. It is worth noting that we only use videos that are still available on YouTube, which is 7% fewer than in the original dataset.

Implementation Details

We sample video frames at a rate of 1 frame per second (FPS) for both datasets. To standardize input length, we either subsample or pad the frame sequences to a fixed number of frames F, where $F = 100$ for ActivityNet Captions and $F = 200$ for YouCook2. The number of event queries used in the Deformable Transformer is set to 10 for ActivityNet Captions and 100 for YouCook2 to account for the varying density of event annotations. For the Event Context Clustering module, we set the number of clusters to 10 and 20 for ActivityNet Captions and YouCook2, respectively. During Event Semantic Retrieval, we apply a soft top-k selection with $k = 40$, allowing each pseudo-event center to retrieve the 40 most relevant text features from the memory bank. All remaining model hyperparameters are aligned with those used in CM^2 (Kim et al. 2024). All experiments are conducted using an NVIDIA RTX A6000 GPU.

Evaluation Metrics

The evaluation of our model is conducted from two complementary perspectives: (i) Event Captioning Performance: To assess the quality of generated captions, we utilize standard metrics including CIDEr (Vedantam, Zitnick, and Parikh 2015), which computes TF-IDF weighted n-gram consensus, BLEU4 (Papineni et al. 2002), which measures 1 to 4 gram precision, and METEOR (Banerjee and Lavie 2005), which incorporates synonym matching and word order alignment. These scores are averaged across multiple IoU thresholds $\{0.3, 0.5, 0.7, 0.9\}$ to ensure robustness. Additionally, we report SODA_c (Fujita et al. 2020), a metric designed to evaluate storytelling ability and the coherence of the overall caption sequence. (ii) Event Localization Performance: To measure localization accuracy, we report the average precision, average recall, and F1 scores, each calculated across IoU thresholds $\{0.3, 0.5, 0.7, 0.9\}$, providing a comprehensive view of temporal alignment performance.

Results

Dense Video Captioning Performance. In Table 1 and Table 2, we compare our CACMI framework with the state-of-the-art methods on the ActivityNet Captions and YouCook2 datasets. As shown in Table 1, our CACMI consistently outperforms the strong baseline CM^2 (Kim et al. 2024) across all four metrics CIDEr, METEOR, BLEU4, and SODA_c and achieves highly competitive results with the state-of-the-art E^{2}DVC (Wu et al. 2025). Remarkably, our method even surpasses some pretrained models that leverage large-scale external video data, highlighting the effectiveness of our temporal-semantic modeling. For Table 2, Vid2Seq (Yang et al. 2023) achieves higher scores than our method on YouCook2 dataset. The performance gap is attributed to the limited training videos coupled with highly diverse event semantics for YouCook2 dataset, where Vid2Seq benefits from broader domain coverage. Im-

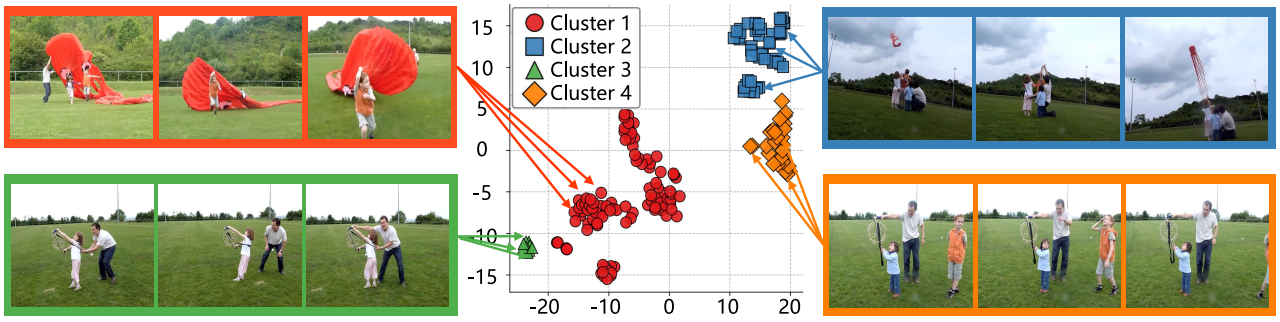

Figure 3: Visualization of event features. The t-SNE projection illustrates a two-dimensional embedding space, where grouped points within the same cluster indicate temporal correlation and semantic similarity. This demonstrates that the frame aggregation module effectively constructs discriminative event representations while preserving meaningful temporal information.

CFA	CFE	BLEU4	METEOR	CIDEr	SODA_c	F1
×	×	2.38	8.55	33.01	6.18	55.21
✓	×	2.37	8.63	33.62	6.26	56.07
×	✓	2.41	8.59	33.48	6.31	56.95
✓	✓	**2.44**	**8.68**	**33.80**	**6.39**	**57.10**

Table 4: Performance of different components. CFA and CFE denote cross-modal frame aggregation and context-aware feature enhancement, respectively.

N_{cluster}	BLEU4	METEOR	CIDEr	SODA$_c$	F1
3	2.32	8.53	32.84	6.12	54.91
5	2.35	8.58	33.15	6.21	54.87
7	2.36	8.63	33.24	6.28	56.02
10	**2.44**	**8.68**	**33.80**	**6.39**	**57.10**
15	2.28	8.49	32.98	6.19	55.15
γ	2.34	8.60	33.43	6.29	56.23

Table 5: Ablation study on the number of event clusters. We report the results on the ActivityNet Captions. The best performance is highlighted.

portantly, our CACMI significantly outperforms all non-pretrained baselines in SODA_c, a metric designed to evaluate the storytelling quality and coherence of multi-event captions. This result underscores our model's ability to effectively capture temporal and semantic dependencies within untrimmed video streams, leading to more contextually rich and consistent caption generation.

Event Localization Performance. We further evaluate the event localization capability of our CACMI. As shown in Table 3, our CACMI achieves the state-of-the-art performance on both benchmark datasets. On the ActivityNet Captions dataset, our CACMI attains an F1 score of 57.10, with a recall of 55.89 and precision of 58.05, indicating a strong balance between accurate detection and coverage of relevant events. On the more challenging YouCook2 dataset, our method achieves an F1 score of 29.34, with 25.54 recall and 34.63 precision, demonstrating robustness in complex,

top_k	BLEU4	METEOR	CIDEr	SODA_c	F1
10	2.23	8.49	32.20	6.32	55.95
20	2.31	8.60	32.26	6.21	55.50
40	**2.44**	**8.68**	**33.80**	**6.39**	**57.10**
60	2.27	8.50	32.25	6.25	56.39
80	2.25	8.53	32.57	6.27	56.15

Table 6: Ablation study on the top-k selection number for retrieval. We report the results on the ActivityNet Captions. The best performance is highlighted.

fine-grained scenarios. These results highlight the effectiveness of our temporal-semantic retrieval framework, which not only improves captioning quality by modeling temporally coherent semantics, but also strengthens localization accuracy by incorporating rich temporal cues into the event boundary prediction process.

Ablation Studies

Analysis of Different Components. Table 4 presents ablation results on the ActivityNet Captions dataset, evaluating the contributions of the key components in our framework: Cross-modal Frame Aggregation (CFA) and Context-aware Feature Enhancement (CFE). The CFA module captures temporally consistent event information and enhances visual representations by retrieving semantically relevant text features, while the CFE module aligns visual dynamics with textual semantics through a context-query attention mechanism. From the results, CFE alone substantially boosts both event localization and caption generation, particularly improving temporal boundary precision and enhancing narrative coherence. Although CFA independently contributes noticeable gains in both tasks, the combination of CFA and CFE yields the best overall performance, demonstrating strong synergistic effects. This integration enables CACMI to effectively retain fine-grained temporal structure, leading to superior dense video captioning performance.

Effect of the Number of Event Clusters. As shown in Table 5, we investigate the impact of the number of event clusters used during temporal clustering. The clustering pro-

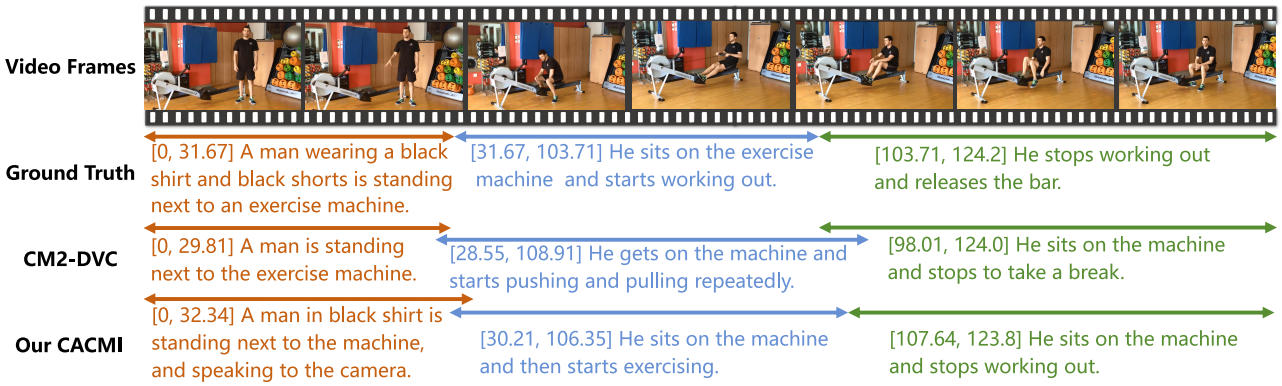

Figure 4: Visualizations of dense event captioning prediction on ActivityNet Captions. We present the results of the ground truth, the baseline CM^2 and our CACMI.

cess is guided by a hyperparameter γ, which adaptively determines the optimal number of clusters from 5 to 10. When the number of clusters is too small, the model captures only coarse event priors, limiting fine-grained temporal representation and resulting in marginal performance. Conversely, too many clusters lead to over-segmentation and noise from fragmented boundaries. Our results show that optimal performance is achieved with $N_{cluster}$ is 10, balancing temporal coherence and semantic granularity.

Effect of the Number of Retrieved Sentences. Table 6 shows the impact of varying the number of retrieved textual features on model performance. The number of clustered events is fixed at 10, then we evaluate the influence of different top-k values during the cross-modal retrieval process. When the top-k value is set too low, each cluster has access to only a limited number of supplemental sentences. This restricted retrieval fails to provide sufficient semantic diversity, limiting the model's ability to capture rich contextual cues. Conversely, an excessively high top-k value introduces redundant or less relevant sentences, which can dilute the informative content and obscure key semantic signals, ultimately hindering the model's performance. Empirically, the model achieves optimal performance when the top-k value is set to 40, striking an effective balance between contextual richness and semantic relevance.

Qualitative Comparison

Visualization of Event Clusters. Figure 3 shows the frame-level features aggregated by the Event Context Clustering module, providing qualitative evidence of the temporal coherence within the generated event representations. We first apply PCA to reduce the dimensionality of the high-dimensional visual features, followed by t-SNE projection into a two-dimensional embedding space for visualization. The results show that features grouped within the same cluster correspond to video segments exhibiting strong temporal continuity and semantic similarity, while features from different clusters represent visually distinct content.

Visualization of Predicted Results. Figure 4 illustrates a qualitative example of event predictions generated by our CACMI. Compared to the baseline CM^2, our CACMI produces more precise event boundaries, demonstrating significantly enhanced localization capabilities. Simultaneously, it captures richer semantic details, enabling more accurate and contextually relevant descriptions of video events. By leveraging contextual semantics, our CACMI achieves a deeper understanding of video content while maintaining robust event localization performance.

Conclusion

We propose Context-Aware Cross-Modal Interaction (CACMI), a novel framework for explicit temporal-semantic modeling in the dense video captioning task. Our CACMI effectively leverages the temporal dependencies within video content and the semantic knowledge embedded in text corpus through a unified cross-modal interaction strategy. The framework consists of two core components: Cross-modal Frame Aggregation, which enhances contextual understanding and semantic enrichment by grouping temporally coherent frames and retrieving relevant text features; and Context-aware Feature Enhancement, which bridges the semantic gap between modalities by aligning visual dynamics with textual semantics. Comprehensive experiments on the ActivityNet Captions and YouCook2 datasets demonstrate that our CACMI achieves the sota results, significantly improving event localization accuracy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. U21A20515, 62376271, U22B2034, 62171321, 62572059, and 62365014), the Beijing Natural Science Foundation (Grant Nos. L231013, L241056), the Shenzhen Science and Technology Program (Grant No. CJGJZD20240729141906008), the Open Project of the Key Laboratory of Computing Power Network and Information Security (Grant No. 2024PY021), and the Open Project Program of the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Grant No. VRLAB2025B03).

References

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.

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, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T. 2017. Video captioning with attention-based LSTM and semantic consistency. *IEEE Transactions on Multimedia*, 19(9): 2045–2055.

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, 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.

Kim, M.; Kim, H. B.; Moon, J.; Choi, J.; and Kim, S. T. 2025. HiCM^2: Hierarchical Compact Memory Modeling for Dense Video Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4293–4301.

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.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, 237–255. Springer.

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.

Nie, M.; Ding, D.; Wang, C.; Guo, Y.; Han, J.; Xu, H.; and Zhang, L. 2024. Slowfocus: Enhancing fine-grained temporal understanding in video llm. *Advances in Neural Information Processing Systems*, 37: 81808–81835.

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.

Pei, W.; Zhang, J.; Wang, X.; Ke, L.; Shen, X.; and Tai, Y.-W. 2019. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8347–8356.

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.

Ramos, R.; Martins, B.; Elliott, D.; and Kementchedjhieva, Y. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2840–2849.

Seo, P. H.; Nagrani, A.; Arnab, A.; and Schmid, C. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17959–17968.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4998–5007.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

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.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Wang, Z.; Shi, Y.; et al. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 396–416. Springer.

Wu, K.; Li, P.; Fu, J.; Li, Y.; Wu, Y.; Liu, Y.; Wang, J.; and Zhou, S. 2025. Event-Equalized Dense Video Captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8417–8427.

Xiong, C.; Zhong, V.; and Socher, R. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

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.

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.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.