Task-Specific Information Decomposition for End-to-End Dense Video Captioning

Zhiyue Liu[1,2]*, Xinru Zhang[1], Jinyuan Liu[1]
[1]School of Computer, Electronics and Information, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
liuzhy@gxu.edu.cn, {2313301061,2213394017}@st.gxu.edu.cn

Abstract

Dense video captioning aims to localize events within input videos and generate concise descriptive texts for each event. Advanced end-to-end methods require both tasks to share the same intermediate features that serve as event queries, thereby enabling the mutual promotion of two tasks. However, relying on shared queries limits the model's ability to extract task-specific information, as event semantic perception and localization demand distinct perspectives on video understanding. To address this, we propose a decomposed dense video captioning framework that derives localization and captioning queries from event queries, enabling task-specific representations while maintaining inter-task collaboration. Considering the roles of different queries, we design a contrastive semantic optimization strategy that guides localization queries to focus on event-level visual features and captioning queries to align with textual semantics. Besides, only localization information is considered in existing methods for label assignment, failing to ensure the relevance of the selected queries to descriptions. We jointly consider localization and captioning losses to achieve a semantically balanced assignment process. Extensive experiments on the YouCook2 and ActivityNet Captions datasets demonstrate that our framework achieves state-of-the-art performance.

1 Introduction

Videos have been deeply integrated into various aspects of contemporary society, emerging as one of the primary means of sharing information. Dense video captioning (DVC) (Krishna et al., 2017; Shen et al., 2017; Xu et al., 2019; Suin and Rajagopalan, 2020; Huang et al., 2020) is a fundamental task in video understanding that requires the detection of multiple events within a video, and the generation of temporally localized and semantically consistent textual descriptions. Unlike traditional video captioning (Rohrbach et al., 2013; Gao et al., 2017; Chen et al., 2017; Wang et al., 2018a; Pei et al., 2019; Qi et al., 2020; Lin et al., 2022; Seo et al., 2022), which summarizes an entire video with a single caption, DVC demands precise event localization and fine-grained captioning, making it essential for applications such as video summarization, assistive technology, and multimedia retrieval. This task remains especially challenging owing to the complex temporal structure and diverse event compositions in dense videos.

Traditional DVC methods primarily adopt a two-stage architecture, which first localizes events and then generates captions separately (Krishna et al., 2017; Li et al., 2018; Wang et al., 2018b; Iashin and Rahtu, 2020b). While these methods allow for independent optimization of event localization and captioning, they often suffer from the limited interaction between the two tasks, leading to sub-optimal performance and efficiency. Recent advancements have shifted towards end-to-end architectures (Wang et al., 2021a; Yang et al., 2023; Kim et al., 2024), which jointly optimize event localization and caption generation within a single framework. By leveraging the shared event queries and parallel decoding heads, these methods improve task synergy between event localization and caption generation. Although those methods enhance the interaction between tasks, the shared queries across tasks fail to meet the differing information requirements of each task, which hinders the performance improvement of models. To elaborate:

- **Event Localization** operates as a temporal regression task and focuses on precise temporal segmentation, which requires the model to predict event boundaries accurately.
- **Event Captioning** functions as a sequence generation task and demands a deeper under-

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16524–16536

standing of event semantics and linguistic fluency to produce meaningful descriptions.

Using shared content for both tasks often leads to task interference (Li et al., 2018; Kanakis et al., 2020; Chen et al., 2023; Yan et al., 2024), since the model struggles to balance fine-grained temporal localization with contextually relevant caption generation. Additionally, previous end-to-end methods guide label assignment using a localization loss, binding the ground truth localization and description labels to queries by minimizing the localization loss. Such assignment overlooks the possibility that the assigned queries may lack sufficient semantic capacity to generate descriptions, leading to semantic imbalance in label assignment.

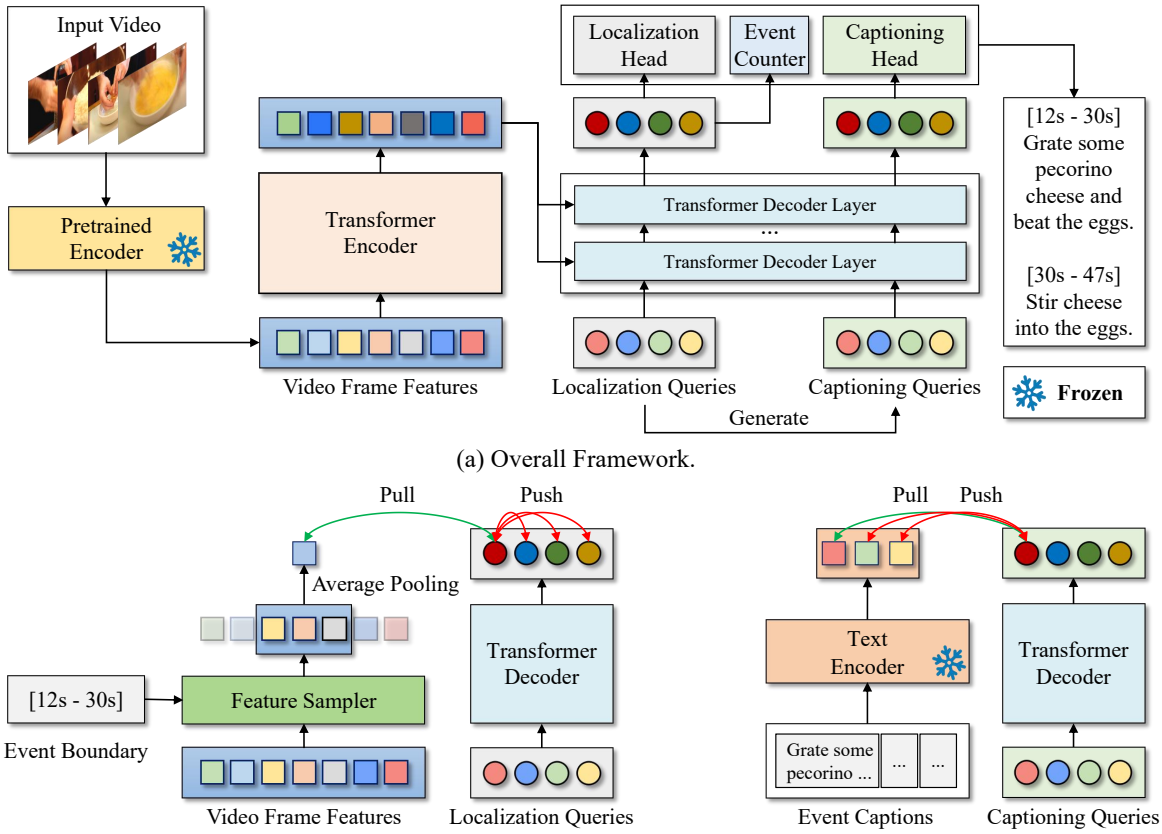
Toward the above issues, we propose decomposed dense video captioning (DDVC), a novel framework that decomposes event queries to enable task-specific semantic representations. Instead of relying on a shared representation for both tasks, DDVC constructs separate localization and captioning queries from a shared source, allowing for task-specific feature extraction while maintaining synergy between these two tasks. Besides, a joint supervision label assignment method is used to allocate ground truth for queries. Our key contributions are summarized as follows:

- **Task-Specific Query Decomposition:** We introduce localization queries for event localization and captioning queries for text generation. This decomposition enables the extraction of task-specific features, effectively mitigating the conflict between precise temporal localization and rich semantic caption generation.
- **Contrastive Semantic Optimization for Multimodal Alignment:** A cross-modal contrastive learning strategy is designed to boost the model for differentiating events. It makes localization queries focus on event-level visual features, and captioning queries align with textual semantics, resulting in improved segmentation and caption quality.
- **Label Assignment with Joint Supervision:** In contrast to previous works, we simultaneously consider both localization and description losses when assigning ground truth, evaluating the comprehensive potential of query decoding into event boundaries and descriptions to achieve reasonable label assignments.
- **State-of-the-Art Performance:** We conduct extensive experiments on several benchmark datasets, which demonstrates that our method achieves state-of-the-art performance across multiple metrics. Unlike recent methods that depend on external retrieval mechanisms to enhance performance, DDVC achieves competitive results without such dependencies.

2 Related Work

Dense video captioning is inherently a multi-task learning problem that involves both event localization and description. Early methods adopted a two-stage paradigm, where event detection and caption generation were performed separately (Krishna et al., 2017; Wang et al., 2018b). To enhance event representation within this framework, subsequent works introduced hierarchical event modeling (Wang et al., 2021b) and multimodal fusion techniques (Iashin and Rahtu, 2020a,b), leading to more accurate and informative captions. However, these methods lacked explicit interaction between localization and description tasks, often leading to discrepancies between detected events and their textual descriptions. To overcome this limitation, pretrained methods such as Vid2seq (Yang et al., 2023) and DIBS (Wu et al., 2024) unified both tasks within a text generation framework, producing event timestamps and captions simultaneously through a single decoding process. Their models have also inspired dense captioning in online settings (Zhou et al., 2024). Despite achieving strong descriptive performance, they demonstrate limited localization accuracy and impose computationally demanding training requirements.

An alternative direction focuses on end-to-end architectures that jointly optimize localization and caption generation (Zhou et al., 2018b; Mun et al., 2019; Wang et al., 2021a; Zhu et al., 2022; Kim et al., 2024; Xie et al., 2025). PDVC (Wang et al., 2021a) formulates dense video captioning as a set prediction problem and employs shared event queries for both tasks. Recent advances, CM2 (Kim et al., 2024) and MCCL (Xie et al., 2025), employed retrieval-augmented frameworks with external memory banks to improve performance, but they still struggle to effectively handle the issue of inconsistent information requirements between different tasks (Chen et al., 2023; Yan et al., 2024), limiting the further development of end-to-end approaches. Furthermore, existing label assignment



(a) Overall Framework.

(b) Semantic Alignment for Localization.

(c) Semantic Alignment for Captioning.

Figure 1: Overview of DDVC. (a) shows the structure of DDVC, where localization queries and generated captioning queries interact with video features to capture different semantic information. Then, event boundaries and descriptive text are predicted through task-specific heads. (b) shows the semantic alignment for localization queries, where the feature sampler extracts video features within the ground-truth temporal boundaries corresponding to the localization queries. The queries are then aligned with these video features. (c) shows the semantic alignment for captioning queries, where captioning queries are aligned with the features of corresponding ground-truth event captions.

strategies prioritize localization loss minimization when matching queries to ground truth, neglecting semantic richness critical for description quality. In comparison, the proposed framework decomposes event queries into localization and captioning queries, enabling task-specific representations. Our contrastive semantic optimization further shapes the constructed queries to help the model extract accurate localization and description features. We also consider both localization and description objectives for label assignment, which guides ground truth allocation to queries while ensuring semantically balanced optimization.

3 Methodology

Dense videos contain multiple sequential or overlapping events, and DVC needs to localize these events while producing coherent textual descriptions. Given a video V, the objective is to generate a set of event proposals $\{\hat{t}_i^s, \hat{t}_i^e, \hat{c}_i\}_{i=1}^{\hat{N}}$, where $\hat{t}_i^s$, $\hat{t}_i^e$, and $\hat{c}_i$ denote the predicted starting time, ending time, and caption of the i-th event, respectively. $\hat{N}$ denotes the total number of events, which is also predicted by the model.

3.1 Overall Framework

Our goal is to decompose shared event queries, typically adopted by end-to-end DVC for video feature extraction, into localization and captioning queries, allowing task-specific information extraction. As exhibited in Figure 1(a), DDVC preserves the standard end-to-end pipeline, including a pretrained visual encoder, a deformable transformer (Zhu et al., 2021), a localization head, a captioning head, and an event counter, but modifies only the query construction. For an input video V, frame-level features $\{f_l\}_{l=1}^{T}$ are extracted by the visual encoder, where T is the video temporal length. These features are subsequently enriched via the transformer encoder as $f_l^E = \text{Trans}_e(f_l)$ for temporal

context modeling. Unlike existing methods (Kim et al., 2024; Xie et al., 2025) that utilize shared event queries $\{q_j^o\}_{j=1}^K$ for localization and captioning, DDVC would decompose shared queries into localization queries $\{q_j^{\text{loc}}\}_{j=1}^K$ and captioning queries $\{q_j^{\text{cap}}\}_{j=1}^K$. In the transformer decoder, these queries attend to $\{f_l\}_{l=1}^T$ to yield the task-specific features $\tilde{q}_j^{\text{loc}} = \text{Trans}_d(q_j^{\text{loc}}, \{f_l^E\}_{l=1}^T)$ and $\tilde{q}_j^{\text{cap}} = \text{Trans}_d(q_j^{\text{cap}}, \{f_l^E\}_{l=1}^T)$. Each $\tilde{q}_j^{\text{loc}}$ is fed into the localization head to predict the temporal boundaries $\hat{t}_j^s, \hat{t}_j^e = \text{Head}_{\text{loc}}(\tilde{q}_j^{\text{loc}})$, while the event number $\hat{N} = \text{Head}_{\text{count}}(\{\tilde{q}_j^{\text{loc}}\}_{j=1}^K)$ is predicted by the event counter based on the set of all localization features. The captioning head generates $\hat{c}_j = \text{Head}_{\text{cap}}(\tilde{q}_j^{\text{cap}})$ using $\tilde{q}_j^{\text{cap}}$.

3.2 Query Decomposition

Rather than employing a single set of learnable queries $\{q_j^o\}_{j=1}^K$ for the feature extraction of both tasks, our framework adopts a simple-yet-effective decomposition strategy that yields task-specific features while preserving inter-task synergy. Specifically, we initialize localization queries $\{q_j^{\text{loc}}\}_{j=1}^K$ directly using the learnable queries $\{q_j^o\}_{j=1}^K$, and derive the captioning queries from $\{q_j^{\text{loc}}\}_{j=1}^K$ via a lightweight transformation. This process could be formulated as follows:

$$\begin{aligned} q_j^{\text{loc}} &= q_j^o, \\ q_j^{\text{cap}} &= \text{Generator}(q_j^{\text{loc}}), \end{aligned} \tag{1}$$

where the Generator consists of three linear layers followed by a ReLU activation function. Although each task is equipped with its own queries, an appropriate optimization is still necessary to guide those queries to focus on task-specific information.

3.3 Joint Supervised Label Assignment

Before model optimization, ground-truth annotations should be assigned to predicted events. Existing end-to-end methods consider only event localization for label assignment, formulating the matching between predicted localization and ground-truth ones using the focal loss (Lin et al., 2017) and IoU loss (Rezatofighi et al., 2019), since this paradigm is directly derived from the object detection task, which does not involve textual semantics. Ignoring text matching hinders accurate label assignment for DVC, thus we propose a joint supervision strategy for label assignment that incorporates both localization and caption semantics. Specifically, we integrate both aspects into the cost computation to enable more accurate alignment between predictions and ground-truth events.

In our framework, each query is responsible for predicting a single event. We assign ground-truth labels to queries via a one-to-one matching scheme that minimizes a global cost function. Given a set of ground-truth event labels $\{t_i^s, t_i^e, c_i\}_{i=1}^N$, each label e_i corresponds to a ground-truth event annotated by its start time, end time, and caption. For each query pair $q_j \in \{q_j^{\text{loc}}, q_j^{\text{cap}}\}_{j=1}^K$ obtained from our decomposition, we compute a matching cost with respect to each e_i. The objective is to establish the optimal assignment by finding the minimum-cost matching between queries and ground-truth events, mathematically formulated as follows:

$$\begin{aligned} \min_{x_{ij}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{K} \text{cost}(e_i, q_j)\, x_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{K} x_{ij} = 1, \quad \forall\, i \in (1, \ldots, N), \\ & \sum_{i=1}^{N} x_{ij} \le 1, \quad \forall\, j \in (1, \ldots, K), \\ & x_{ij} \in \{0, 1\}, \quad \forall\, i \in (1, \ldots, N),\ j \in (1, \ldots, K), \end{aligned} \tag{2}$$

where the first constraint ensures that each event label is assigned to exactly one query pair, while the second constraint ensures that each query pair is assigned at most one event label. The $\text{cost}(e_i, q_j)$ includes the focal loss, localization loss, and captioning loss predicted by q_j, defined as follows:

$$\text{cost} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is the focal loss between the predicted classification score and the label, $\mathcal{L}_{\text{giou}}$ denotes the generalized IoU loss between predicted temporal segments and ground-truth segments, and $\mathcal{L}_{\text{cap}}$ is the captioning loss computed between the predicted and ground-truth captions. The above optimization problem could be solved by the Hungarian algorithm (Kuhn, 1955), resulting in the matching set $\mathcal{M}$ that assigns labels to predictions as follows:

$$\mathcal{M} = \{(i, j) | x_{ij} = 1\}. \tag{4}$$

Considering the matching of both localization and captioning in label assignment promotes effective collaboration between these two tasks.

3.4 Contrastive Semantic Alignment

To enhance the ability of localization and captioning queries to capture their respective task-specific

semantics, we introduce a contrastive learning strategy. In detail, the localization queries are pulled closer to the average visual features of their corresponding video segments and pushed away from those of other events. Meanwhile, the captioning queries are aligned with the textual representations of their matched ground-truth descriptions and contrasted against those of unrelated events.

Semantic Alignment for Localization. We employ the event duration provided in the annotations as a key cue to guide the model in extracting visual semantics specific to each event, leveraging event-independent visual features to enhance localization quality, as illustrated in Figure 1(b). We utilize a feature sampler to extract frame-level video features within the ground-truth temporal boundaries corresponding to the matched localization queries. Assuming that the ground truth e_i matches $q_{i^*}^{\text{loc}}$, where i^* denotes the index of the optimally matched query for e_i and $(i, i^*) \in \mathcal{M}$, we take the mean video feature $\overline{f_i} = \frac{1}{e-s+1}\sum_{j=t_i^s}^{t_i^e} f_j$ within the time range $[t_i^s, t_i^e]$ as the positive example and all other queries as negative examples. Then, we pull positive examples closer while pushing negative examples apart with the cosine distance to measure similarity, which is defined as follows:

$$\mathcal{L}_{\text{contrast}}^{\text{loc}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(q_{i^*}^{\text{loc}}, \overline{f_i})/\tau)}{\exp(\text{sim}(q_{i^*}^{\text{loc}}, \overline{f_i})/\tau) + \sum_{j\neq i}\exp(\text{sim}(q_i^{\text{loc}}, q_j^{\text{loc}})/\tau)}, \tag{5}$$

where N is the number of events, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity, and τ is the temperature parameter. This contrastive loss draws localization queries closer to the visual features of video frames corresponding to their timestamps, which enables localization queries to decode event-related visual patterns more accurately, reduces background interference, and enhances localization precision. Meanwhile, it pushes apart the distributions of different localization queries, helping the model distinguish adjacent or similar events more effectively.

Semantic Alignment for Captioning. Rather than using label captions solely to guide text decoding, we enhance the model's ability to extract descriptive semantics by pulling in the distribution of captioning queries and corresponding event captions, which is illustrated in Figure 1(c). We employ a pretrained text encoder to extract the textual features f^c of the caption c for contrastive learning. The query $q_{i^*}^{\text{cap}}$ and its corresponding description feature f_i^c form a positive example, while $q_{i^*}^{\text{cap}}$ and all other description features form negative examples. This contrastive loss is computed as follows:

$$\mathcal{L}_{\text{contrast}}^{\text{cap}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(q_{i^*}^{\text{cap}}, f_i^c)/\tau)}{\sum_{j\neq i}\exp(\text{sim}(q_{i^*}^{\text{cap}}, f_j^c)/\tau)}. \tag{6}$$

The above loss aligns captioning queries with the textual semantic space, ensuring that the generated descriptions remain highly relevant to the target text. By distancing caption queries from the text representations of other events, the model improves its ability to distinguish event semantics, thereby reducing confusion between event descriptions.

The two contrastive losses $\lambda_{\text{contrast}}^{\text{loc}}$ and $\lambda_{\text{contrast}}^{\text{cap}}$ explicitly constrain localization queries to focus on video temporal dynamics and captioning queries to linguistic information.

3.5 Prediction

Training. Our loss function comprises six terms, and the overall function is defined as follows:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}} \\ & + \lambda_{\text{count}}\mathcal{L}_{\text{count}} + \lambda_{\text{contrast}}^{\text{loc}}\mathcal{L}_{\text{contrast}}^{\text{loc}} + \lambda_{\text{contrast}}^{\text{cap}}\mathcal{L}_{\text{contrast}}^{\text{cap}},\end{aligned} \tag{7}$$

where the $\mathcal{L}_{\text{count}}$ represents the cross-entropy between the predicted count number distribution and the ground truth number of events. In the above loss, the first four terms follow the conventional design of end-to-end DVC, whereas the final two constitute our proposed contrastive learning.

Inference. During inference, DDVC processes the input video features $\{f_l\}_{l=1}^{T}$ to generate a set of candidates $\{\hat{t}_i^s, \hat{t}_i^e, \hat{c}_i\}_{i=1}^{K}$ along with a predicted event count $\hat{N}$. Each candidate is assigned a ranking score by combining its localization confidence with the cumulative probability of its generated caption. Then, the top $\hat{N}$ candidates with the highest scores are selected as the final predictions.

4 Experiments

4.1 Experimental Settings

Datasets. Our experiments are conducted on two widely used benchmark datasets, ActivityNet Captions (Krishna et al., 2017) and YouCook2 (Zhou et al., 2018a). ActivityNet Captions consists of approximately 20,000 untrimmed videos depicting diverse human activities, with each video averaging 120 seconds and annotated with about 3.65 temporally localized sentences. We leverage the official

Method	Event Captioning				Event Localization		
	CIDEr(↑)	**METEOR**(↑)	**BLEU-4**(↑)	**SODA_c**(↑)	**Recall**(↑)	**Precision**(↑)	**F1**(↑)
Pretrain							
Vid2seq	47.10	9.30	-	7.90	27.90	27.80	27.84
DIBS	44.44	7.51	-	6.39	26.24	39.18	31.43
without Pretrain							
PDVC	29.69	5.56	1.40	4.92	22.89	32.37	26.81
CM2	31.66	6.08	1.63	<u>5.34</u>	<u>24.76</u>	<u>33.38</u>	<u>28.43</u>
MCCL	<u>36.09</u>	<u>6.53</u>	**2.04**	5.21	-	-	-
DDVC (Ours)	**38.75**	**6.92**	<u>1.92</u>	**6.68**	**30.81**	**37.25**	**33.73**

Table 1: Performance of event captioning and event localization on YouCook2. ↑ means higher is better. The best result is in bold, and the second best result is underlined.

Method	Event Captioning				Event Localization		
	CIDEr(↑)	**METEOR**(↑)	**BLEU-4**(↑)	**SODA_c**(↑)	**Recall**(↑)	**Precision**(↑)	**F1**(↑)
Pretrain							
Vid2seq	30.10	8.50	-	5.80	52.70	53.90	53.29
DIBS	31.89	8.93	-	5.85	53.14	58.31	55.61
without Pretrain							
PDVC	29.97	8.06	2.21	5.92	53.27	56.38	54.78
CM2	33.01	8.55	2.38	<u>6.18</u>	<u>53.71</u>	56.81	55.21
MCCL	<u>34.92</u>	**9.05**	**2.68**	6.16	53.19	<u>57.36</u>	<u>55.23</u>
DDVC (Ours)	**35.48**	<u>8.62</u>	<u>2.44</u>	**6.55**	**54.77**	**57.54**	**56.12**

Table 2: Performance of event captioning and event localization on ActivityNet Captions.

split of 10,009 videos for training, 4,925 for validation, and 5,044 for testing. YouCook2 contains around 2,000 untrimmed cooking procedure videos, each with an average duration of 320 seconds and approximately 7.7 annotated sentences per video. We follow the standard split with 1,333 videos for training, 457 for validation, and 210 for testing. Note that our experiments utilize about 7% fewer videos than the original dataset due to relying on accessible YouTube content (Kim et al., 2024).

Evaluation Metrics. We evaluate our method from two aspects, including dense video captioning and event localization, as in previous works (Kim et al., 2024; Xie et al., 2025). For captioning, we employ the official ActivityNet Challenge evaluation tools to calculate the CIDEr (Vedantam et al., 2015), BLEU-4 (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) scores, which assess the quality of matched pairs between generated captions and ground truth at IoU thresholds of 0.3, 0.5, 0.7, and 0.9. In addition, we measure storytelling ability using the SODA_c (Fujita et al., 2020) metric. For event localization, we calculate average precision, average recall, and the F1 score at the same IoU thresholds.

Implementation Details. In our experimental setup, the video frames from both datasets are extracted at 1 frame per second. We utilize the pretrained CLIP ViT-L/14 (Radford et al., 2021) visual encoder to extract 768-dimensional frame features, while the corresponding CLIP text encoder generates 768-dimensional textual features. For event detection, the number of event queries K is set to 10 for ActivityNet Captions and 50 for YouCook2. The framework structure employs a deformable transformer with two encoder layers and two decoder layers, integrating multi-scale deformable attention across four feature levels. The localization head is implemented as a multilayer perceptron with three linear layers. The captioning head comprised two decoding modules, with each containing a self-attention layer, a cross-attention layer, and a feed-forward network. The event counter is designed as a single linear layer. The balancing hyperparameters $\lambda_{\text{cls}}, \lambda_{\text{giou}}, \lambda_{\text{cap}}, \lambda_{\text{count}}, \lambda^{\text{loc}}_{\text{contrast}}$, and $\lambda^{\text{cap}}_{\text{contrast}}$ are set to 2, 4, 2, 0.5, 1, and 1, respectively. Adam (Kingma and Ba, 2015) is used as the optimizer with a learning rate of 5×10^{-5} and a weight

QD	JSLA	CSA	Event Captioning				Event Localization		
			CIDEr(↑)	**METEOR**(↑)	**BLEU-4**(↑)	**SODA_c**(↑)	**Recall**(↑)	**Precision**(↑)	**F1**(↑)
✘	✘	✘	31.44	6.24	1.49	5.48	27.98	36.54	31.69
✔	✘	✘	33.54	6.48	1.61	5.74	28.61	**38.33**	32.76
✘	✔	✘	34.74	6.36	1.62	6.16	30.10	36.65	33.06
✔	✔	✘	36.78	6.55	1.63	6.42	30.08	37.81	33.50
✔	✘	✔	37.88	6.57	1.83	6.50	29.83	37.45	33.21
✔	✔	✔	**38.75**	**6.92**	**1.92**	**6.68**	**30.81**	37.25	**33.73**

Table 3: The ablation result of the different components on YouCook2.

Method	Event Captioning				Event Localization		
	CIDEr(↑)	**METEOR**(↑)	**BLEU-4**(↑)	**SODA_c**(↑)	**Recall**(↑)	**Precision**(↑)	**F1**(↑)
w/o CSA	36.78	6.55	1.63	6.42	30.08	**37.81**	33.50
CSA (Loc)	34.65	6.25	1.67	6.51	**32.53**	36.81	**34.54**
CSA (Cap)	38.50	6.73	**2.04**	6.67	29.81	35.87	32.56
DDVC	**38.75**	**6.92**	1.92	**6.68**	30.81	37.25	33.73

Table 4: The ablation result of the different query semantic alignment on YouCook2.

decay of 1×10^{-4}. The temperature in our proposed contrastive loss is set to 0.1. The code is available at https://github.com/siplysagari/DDVC.

Baselines. We compare the performance of our DDVC with state-of-the-art works, including end-to-end methods with or without pretraining. The pretrained methods contain Vid2Seq (Yang et al., 2023) and DIBS (Wu et al., 2024). The methods without pretraining include PDVC (Wang et al., 2021a), CM2 (Kim et al., 2024), and MCCL (Xie et al., 2025), where CM2 and MCCL build external memory banks for retrieval augmentation.

4.2 Experimental Results

Event Captioning Performance. Table 1 shows the experimental results of our proposed DDVC on the YouCook2 dataset. Compared to the end-to-end method PDVC, our DDVC achieves a significant improvement, with increases of 9.06 in CIDEr and 1.76 in SODA_c. Since our method and PDVC both retain a standard end-to-end pipeline, the proposed query decomposition and label assignment are particularly beneficial for dense captioning. While DDVC shows a slight disadvantage in BLEU-4 compared to retrieval-augmented methods that leverage external corpora, it demonstrates clear advantages in other metrics. Specifically, it outperforms MCCL by 2.66 in CIDEr and 0.39 in METEOR, and surpasses CM2 by 1.34 in SODA_c. Our method avoids the burden of additional retrieval and leverages decomposed queries to fully extract task-specific information from videos.

Table 2 reports the model performance of DDVC on the ActivityNet Captions dataset. Our method significantly outperforms PDVC. Compared to the retrieval-augmented methods, DDVC surpasses CM2 in all description metrics. Against MCCL, DDVC achieves gains of 0.56 in CIDEr and 0.39 in SODA_c but lags behind in METEOR and BLEU-4. We attribute this disparity to the benefits of retrieval augmentation, as METEOR and BLEU-4 emphasize surface-level textual matching, which can be effectively leveraged from retrieved corpora.

Both Vid2seq and DIBS require substantial computation for pretraining and finetuning. Their pretraining datasets contain plenty of cooking videos, with fewer human activities, which enhances captioning on YouCook2 but offers limited improvement on ActivityNet Captions. This aligns with the results in Tables 1 and 2. In contrast, our method shows competitive results with a lightweight architecture and an efficient training process.

Event Localization Performance. As shown in Table 1 and Table 2, our method leads in all metrics, demonstrating superior event localization performance across both the YouCook2 and ActivityNet Captions datasets compared to state-of-the-art baselines. Note that the YouCook2 dataset features a denser distribution of events, making precise localization more challenging than in ActivityNet Captions. On YouCook2, DDVC significantly outperforms Vid2Seq and PDVC and surpasses the most advanced model, CM2, by 6.05 in recall, 3.87

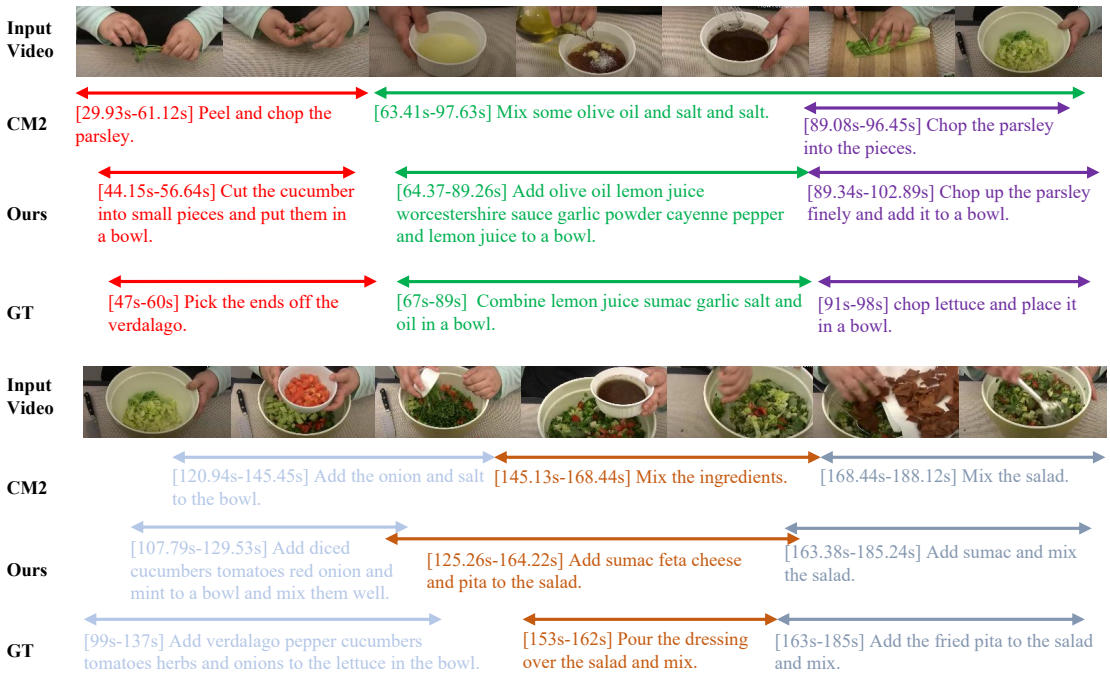


Figure 2: Visualization of the prediction results from our method and CM2 on the YouCook2 dataset, along with the corresponding ground truth (GT) for comparison.

in precision, and 5.3 in F1 score. Our method also achieves superior performance on ActivityNet Captions which contains fewer events. For Vid2seq and DIBS, since the pretraining datasets do not involve event localization, they do not demonstrate strong capabilities in locating events. Our framework produces localization queries to assist the model in exploring required visual information, resulting in more accurate localization results.

4.3 Ablation Studies

Effect of Different Components. We conduct ablation experiments on the YouCook2 dataset to evaluate the effectiveness of the components in DDVC. A baseline model using task-shared queries is constructed, and then we attach the three proposed components to this baseline: 1) query decomposition (QD), 2) joint supervision label assignment (JSLA), and 3) contrastive semantic optimization (CSA). As shown in Table 3, the experimental results demonstrate that each component contributes to the improvement in performance. Incorporating query decomposition into the baseline model enables query decoupling, allowing the model to flexibly capture task-specific semantics and thereby enhancing performance. The joint supervision label assignment method applies a global optimal task allocation strategy, mitigating suboptimal solutions in description tasks under localization guidance. The result shows that, without using our label assignment, the method with query decomposition achieves the highest precision score of 38.33 in event localization, but the description performance lags behind. The addition of contrastive semantic optimization on top of query decomposition further enhances the model's ability to capture task-specific semantics, boosting results in both localization and description. DDVC combines these components and achieves the best overall performance. The above results validate the effectiveness of each component and their compatibility.

Effect of Contrastive Semantic Alignment. To further investigate the impact of contrastive semantic augmentation (CSA) on DDVC, we conduct an additional experiment on YouCook2, with results presented in Table 4. The CSA (Loc) variant refers to applying semantic augmentation solely to the localization query. Compared to the method without semantic augmentation (w/o CSA), CSA (Loc) significantly improves event localization performance but may have reduced performance in the descriptive task. Conversely, in the CSA (Cap) setting, where only the caption query is augmented, the opposite effect is observed. Given that both localization and captioning share the same input but pursue different objectives, they are always competing, since DVC is a multi-task learning problem. Although our contrastive semantic alignment en-

hances the model's ability to extract task-specific information, exclusively optimizing semantic features for one task may hinder the learning of the other. DDVC simultaneously employs two semantic alignment methods to enhance the overall capability of dense video captioning.

4.4 Qualitative Evaluation

We visualize the prediction results from our method and CM2 on the YouCook2 dataset, as shown in Figure 2. The observations indicate that both methods accurately predict the number of events in this case. Regarding event localization, CM2 exhibits boundary definition errors in Event 2 (marked in green), whereas our predictions demonstrate coherent boundaries with low redundancy that better align with the ground truth distribution. For event captioning, CM2 tends to generate overly concise texts (e.g., Events 5 and 6, marked in brown and gray, respectively) that inadequately summarize event content, while our method produces more semantically rich descriptions resembling human annotations. DDVC leverages query decomposition and contrastive semantic alignment to empower the model with task-aware information extraction. Our joint supervised label assignment also promotes collaboration between localization and captioning.

5 Conclusion

This paper presents a novel decomposed framework for dense video captioning that addresses the limitations stemming from shared query representations in existing end-to-end paradigms. By decoupling event queries into task-specific localization and captioning queries while preserving inter-task collaboration, our method enhances both event boundary detection and description generation. Contrastive semantic optimization further refines query representations by aligning localization with visual cues and captioning with textual semantics, while our joint label assignment provides semantically balanced supervision. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance. Our work establishes a new standard for dense video understanding through its dual-path framework that respects task-specific requirements while maintaining inter-task synergy.

Limitations

Our DDVC is built on an end-to-end model architecture, thereby inheriting the limitations of this architecture. It restricts the number of queries and requires the complete video content as input, which limits the number of detectable events and renders the model unsuitable for streaming video scenarios. Regarding semantic disentanglement, our method only decomposes the query representations without addressing the structural disentanglement of the model architecture. Sharing model parameters to accommodate both types of query features may lead to suboptimal performance due to potential interference between tasks. In future work, we plan to explore more profound disentanglement methods to separate semantics tailored for different tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62406081), and the Guangxi Natural Science Foundation (No. 2025GXNSFBA069232).

References

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Runhang Chen, Xiao-Yuan Jing, Fei Wu, Wei Zheng, and Yaru Hao. 2023. Task-specific parameter decoupling for class incremental learning. *Information Sciences*, 651:119731.

Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. 2017. Video captioning with guidance of multimodal latent topics. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1838–1846, New York, NY, USA. Association for Computing Machinery.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision – ECCV 2020*, pages 517–531, Cham. Springer International Publishing.

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

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th*

International Joint Conference on Natural Language Processing, pages 470–490, Suzhou, China. Association for Computational Linguistics.

Vladimir Iashin and Esa Rahtu. 2020a. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*.

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

Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. 2020. Reparameterizing convolutions for incremental multi-task learning without task interference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 689–707. Springer.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 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 (CVPR)*, pages 13894–13904.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, pages 1–13.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7492–7500.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6588–6597.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. 2020. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666.

Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 433–440.

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

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

Maitreya Suin and A. N. Rajagopalan. 2020. An efficient framework for dense video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12039–12046.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018a. Reconstruction network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7622–7631.

Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018b. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7198.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021a. End-to-end dense video captioning with parallel decoding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6827–6837.

Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. 2021b. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900.

Hao Wu, Huabin Liu, Yu Qiao, and Xiao Sun. 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 (CVPR)*, pages 18699–18708.

Zhuyang Xie, Yan Yang, Yankai Yu, Jie Wang, Yongquan Jiang, and Xiao Wu. 2025. Exploring temporal event cues for dense video captioning in cyclic co-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8771–8779.

Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. 2019. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 25–26.

Ruiran Yan, Rui Fan, and Defu Lian. 2024. Multi-task recommendation with task information decoupling. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2786–2795, New York, NY, USA. Association for Computing Machinery.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 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 (CVPR)*, pages 10714–10726.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):7590–7598.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8748.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252.

Wanrong Zhu, Bo Pang, Ashish V. Thapliyal, William Yang Wang, and Radu Soricut. 2022. End-to-end dense video captioning as sequence generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5651–5665, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.

Method	Event Captioning				Event Localization		
	CIDEr(↑)	**METEOR**(↑)	**BLEU-4**(↑)	**SODA_c**(↑)	**Recall**(↑)	**Precision**(↑)	**F1**(↑)
YouCook2							
CM2	31.66	6.08	1.63	5.34	24.76	33.38	28.43
MCCL	36.09	6.53	2.04	5.21	-	-	-
DDVC	38.75	6.92	1.92	**6.68**	**30.81**	**37.25**	**33.73**
+RA (CM2)	**39.50**	**6.98**	**2.05**	6.52	30.38	37.08	33.40
ActivityNet Captions							
CM2	33.01	8.55	2.38	6.18	53.71	56.81	55.21
MCCL	34.92	**9.05**	2.68	6.16	53.19	57.36	55.23
DDVC	35.48	8.62	2.44	**6.55**	**54.77**	**57.54**	**56.12**
+RA (CM2)	**37.14**	8.83	**2.75**	6.43	53.77	57.02	55.34

Table 5: The results of employing retrieval augmentation on DDVC.

A Computational Overhead

The experiments are conducted using one NVIDIA GeForce RTX 3090 GPU. Our proposed DDVC contains approximately 73 million parameters. On the YouCook2 dataset, each training epoch takes 8.75 minutes, and with 15 epochs, the total training cost amounts to approximately 2.19 GPU hours. For the ActivityNet Captions dataset, each epoch requires 42.35 minutes of training time. Following the same 15-epoch training protocol, the total computational cost reaches 10.59 GPU hours.

B Retrieval Augmentation on DDVC

To investigate the impact of retrieval-augmented design, we apply the retrieval strategy of CM2 to our DDVC framework. Specifically, we fuse the retrieved textual features with video and query features, denoted as "+RA (CM2)", with the results in Table 5. The results demonstrate that retrieval augmentation improves text-only metrics such as CIDEr, METEOR, and BLEU-4 on both YouCook2 and ActivityNet Captions, indicating its effectiveness in enhancing surface-level textual alignment. However, we observe that retrieval augmentation does not lead to performance gains on localization-related metrics such as SODA_c, Recall, Precision, and F1. Note that SODA_c evaluates both localization and captioning, and the degradation in event boundary prediction would offset the gains in description quality. This suggests that the external texts retrieved with visual content may contain useful semantics for captioning but not for localization, potentially interfering with boundary judgment. Existing retrieval-augmented methods are not fully compatible with our query decomposition. It is worth exploring retrieval mechanisms that incorporate more fine-grained feature manipulation (e.g., decomposed feature fusion and modality-specific alignment) to better leverage external knowledge while maintaining localization precision.

C Additional Qualitative Examples

We provide additional qualitative examples from the YouCook2 and ActivityNet Captions datasets to demonstrate the effectiveness of our method. As shown in Figure 3 and Figure 4, DDVC produces more accurate temporal boundaries and semantically coherent captions compared to the baseline CM2. In YouCook2, where videos often contain many short-duration events, our method exhibits better temporal alignment and produces more semantically accurate descriptions. In ActivityNet Captions, where its videos feature fewer but longer events with rich semantics, our method captures fine-grained actions and transition cues that are frequently overlooked by CM2. These results highlight the superiority of our model in handling the videos with diverse event densities and complexities in real-world scenarios.

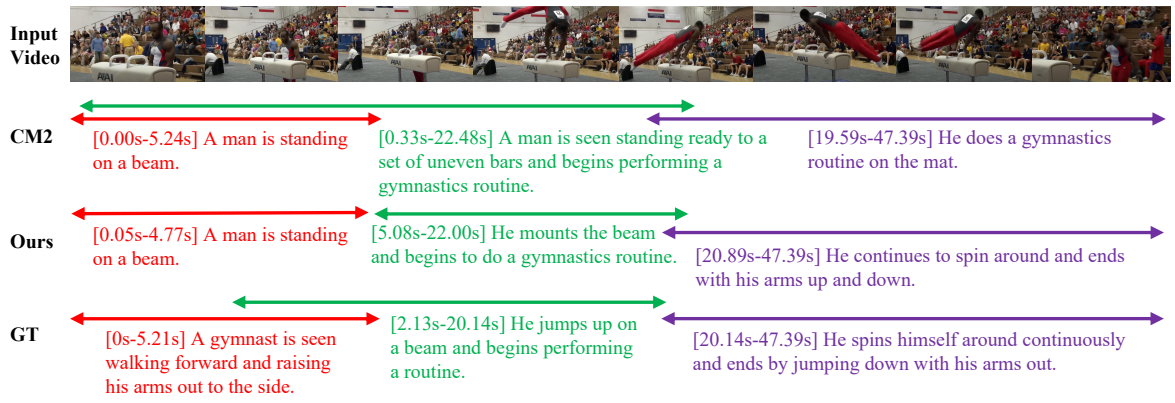


Figure 3: Visualization of the prediction results from our method and CM2 on ActivityNet Captions.

Figure 4: Visualization of the prediction results from our method and CM2 on YouCook2.