Implicit Location-Caption Alignment via Complementary Masking for Weakly-Supervised Dense Video Captioning

Shiping Ge[1]*, **Qiang Chen**[2], **Zhiwei Jiang**[1]†, **Yafeng Yin**[1], **Liu Qin**[2], **Ziyao Chen**[2], **Qing Gu**[1]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]Tencent WeChat, Guangzhou, China
shipingge@smail.nju.edu.cn, ethanqchen@tencent.com, {jzw, yafeng}@nju.edu.cn,
{stenliu, yateschen}@tencent.com, guq@nju.edu.cn

Abstract

Weakly-Supervised Dense Video Captioning (WSDVC) aims to localize and describe all events of interest in a video without requiring annotations of event boundaries. This setting poses a great challenge in accurately locating the temporal location of event, as the relevant supervision is unavailable. Existing methods rely on explicit alignment constraints between event locations and captions, which involve complex event proposal procedures during both training and inference. To tackle this problem, we propose a novel implicit location-caption alignment paradigm by complementary masking, which simplifies the complex event proposal and localization process while maintaining effectiveness. Specifically, our model comprises two components: a dual-mode video captioning module and a mask generation module. The dual-mode video captioning module captures global event information and generates descriptive captions, while the mask generation module generates differentiable positive and negative masks for localizing the events. These masks enable the implicit alignment of event locations and captions by ensuring that captions generated from positively and negatively masked videos are complementary, thereby forming a complete video description. In this way, even under weak supervision, the event location and event caption can be aligned implicitly. Extensive experiments on the public datasets demonstrate that our method outperforms existing weakly-supervised methods and achieves competitive results compared to fully-supervised methods.

Code — https://github.com/ShipingGe/ILCACM

Introduction

Dense Video Captioning (DVC) is a challenging task that aims to generate a series of temporally localized captions to describe the various events in a video (Krishna et al. 2017). It extends traditional video captioning by providing a more comprehensive understanding of the video content, making it particularly useful for applications such as video understanding, video summarization, and video search (Li et al. 2018; Wang et al. 2021; Yang et al. 2023). Recently, the Weakly-Supervised Dense Video Captioning (WSDVC) task, which only relies on video-level captions and does not require extensive temporal location annotations for training, has been proposed and appears to be more feasible for practical applications (Duan et al. 2018). However, the lack of supervision on event localization poses a great challenge in accurately locating the events in the video.

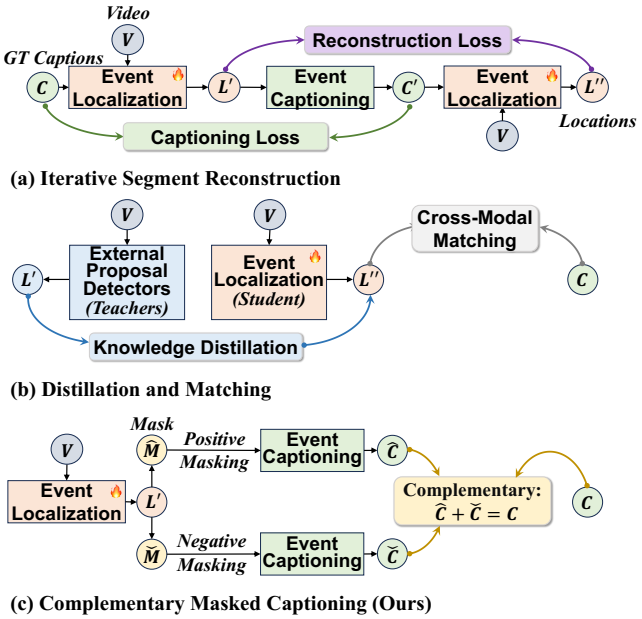


Figure 1: Comparison of our complementary masking paradigm with previous paradigms for event localization.

To address this problem, existing methods attempt to align event locations and captions using alignment constraints like reconstruction loss or cross-modal matching loss (Duan et al. 2018; Chen and Jiang 2021; Wu et al. 2021; Choi, Chen, and Yoon 2023). These approaches primarily fall into two categories: Iterative Segment Reconstruction (Figure 1(a)) and Distillation and Matching (Figure 1(b)). The first type involves a reconstruction cycle where event localization and captioning are interdependent, aiming to minimize reconstruction error (Duan et al. 2018; Chen and Jiang 2021; Choi, Chen, and Yoon 2023). The second type employs external proposal detectors to guide localization and uses met-

*Work done during an internship at Tencent WeChat.
†Corresponding author.

ric learning for cross-modal matching, maximizing semantic similarity between locations and captions (Wu et al. 2021). Despite their promising results, they still suffer from the cumbersome event proposal procedures during both training and inference. During the training phase, these methods rely on complex techniques for event localization, such as the use of pre-defined proposals (Duan et al. 2018; Chen and Jiang 2021; Choi, Chen, and Yoon 2023) or the external pretrained temporal event localization models (Wu et al. 2021). During the inference phase, many of them often require a large number of random proposals to be sampled (Duan et al. 2018; Choi, Chen, and Yoon 2023), which can be computationally expensive.

Unlike previous methods, in this paper, we propose a novel implicit location-caption alignment paradigm based on complementary masking, which simplifies the complex event proposal and localization process while maintaining effectiveness. Specifically, our approach involves an dual-mode video captioning module for event captioning and an extra mask generation module for event localization. We configure our video captioning module to operate in two captioning modes: full video captioning mode and masked video captioning mode. The first mode can provide global event information (e.g., event count) for the event localization in the second mode, eliminating the need for cumbersome event proposal procedures. Besides, as shown in Figure 1(c), with the mask generation module, we can first predict the temporal location of each event merely based on video and then construct the corresponding location mask. After applying a positive mask and its corresponding negative mask (i.e., inverse mask) to the video and performing masked video captioning, we constrain that the captions generated from these two types of masked videos should be complementary (i.e., the two parts of captions constitute the complete video caption). In this way, even under weak supervision, the event location and event caption can be aligned implicitly.

In summary, this paper makes the following contributions:

- We propose a novel implicit location-caption alignment paradigm based on complementary masking, which addresses the problem of unavailable supervision on event localization in the WSDVC task.
- We introduce a dual-mode dense video captioning model, which can simplify the process of event localization.
- Extensive experiments conducted on the public datasets demonstrate the effectiveness of our method and each of its components.

Related Work

Dense Video Captioning Dense Video Captioning is a challenging multi-task problem that involves event localization (Buch et al. 2017; Lin et al. 2018; Zeng et al. 2019; Zhao et al. 2024) and event captioning (Gao et al. 2017; Seo et al. 2022; Nie et al. 2022). A lot of existing methods follow the 'detect-then-describe' paradigm, which first localizes a set of event proposals and then generates captions for the event proposals (Krishna et al. 2017; Wang et al. 2018; Mun et al. 2019; Iashin and Rahtu 2020). Krishna et al. (2017) first introduce the dense video captioning problem and propose an event proposal module followed by a captioning module. Mun et al. (2019) propose to model temporal dependency across events explicitly and leverages visual and linguistic context from prior events for coherent storytelling. Iashin and Rahtu (2020) utilize audio and speech modalities and Transformer architecture to convert multi-modal input data into textual descriptions. Another line of work removes the explicit event proposing process and jointly performs event localization and captioning for each event (Wang et al. 2021; Zhu et al. 2022; Yang et al. 2023). Wang et al. (2021) formulate the dense caption generation as a set prediction task and feed the enhanced representations of event queries into the localization head and caption head in parallel. Zhu et al. (2022) propose to solve the dense video captioning task as a single sequence-to-sequence modeling task using a multimodal Transformer. Yang et al. (2023) introduce a single-stage dense event captioning model pretrained on narrated videos and generate event timestamps as special tokens.

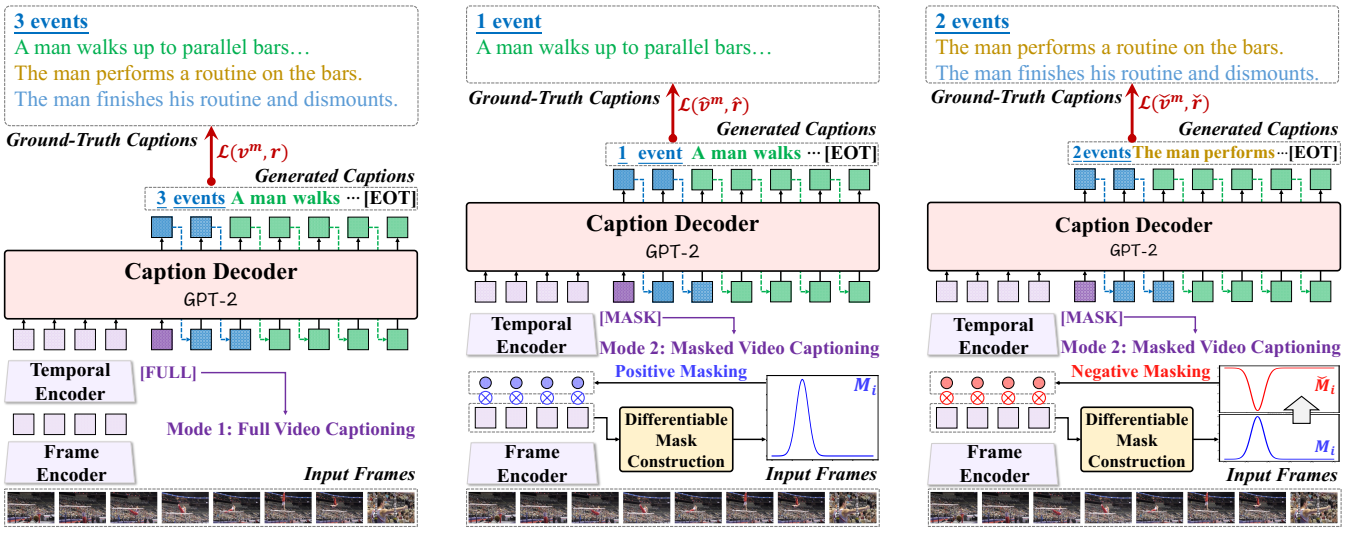
Weakly-Supervised Dense Video Captioning Recently, there has been an increased focus on the Weakly-Supervised Dense Video Captioning (WSDVC) setting, which is considered more challenging and practical than the conventional DVC setting (Duan et al. 2018; Wu et al. 2021; Chen and Jiang 2021; Choi, Chen, and Yoon 2023). Duan et al. (2018) first introduce the WSDVC problem and decompose it into the sentence localization and event captioning problems. Specifically, this paper present a cycle system based on the fixed-point iteration (Chidume 1987) to train the model. Several methods follow this cycle training system and enhance the performance by improving the sentence localization and event captioning modules. Chen and Jiang (2021) propose to use a concept learner as the basis of the sentence localizer, which can be utilized to construct an induced set of concept features to enhance video features and improve the event captioner. Choi, Chen, and Yoon (2023) further improve the performance by pretraining the event captioning model on an extra video description dataset MSR-VTT (Xu et al. 2016). Different from the above methods, Wu et al. (2021) adopt the knowledge distilled from relevant tasks to generate high-quality event proposals and build semantic matching between the proposals and sentences. However, these methods still suffer from the cumbersome event proposal procedures during both training and inference.

Proposed Method

Task Definition

In Dense Video Captioning (DVC), a video is represented as $\boldsymbol{v} = \{v_i\}_{i=1}^{N_v}$, where v_i denotes the i-th frame, and N_v is the total number of frames. The objective is to generate captions $\{S_i\}_{i=1}^{N_s}$ for temporally localized events within the video. Each captioning event S_i encompasses a tuple (t_i^s, t_i^e, C_i), detailing the start time, end time, and the associated caption.

Unlike the DVC task, the Weakly-Supervised Dense Video Captioning (WSDVC) requires the model to generate these temporally localized captioning events without relying on explicit annotations for the start and end times of



(a) Full Video Captioning (b) Positive Masked Captioning (c) Negative Masked Captioning

Figure 2: Illustration of our proposed framework, which consists of two main components: a Dense Video Captioning model for event captioning and a Complementary Mask Generation module for event localization.

each event, i.e., t_i^s and t_i^e for each S_i are unavailable during training. The model should leverage information from the video frames and the provided captions during training to infer the appropriate temporal location and generate accurate, contextually relevant captions during inference.

Overview

Our proposed method integrates two main components: a Dense Video Captioning (DVC) module for event captioning and a Complementary Mask Generation (CMG) module for event localization. The DVC module operates in two modes: full captioning mode captures global narratives, while masked captioning mode enhances localization through differentiable masks. The CMG module predicts temporal locations by utilizing positive and negative masks during the masked captioning mode. Positive masking focuses on specific event captions, and negative masking handles the remaining context, ensuring alignment under weak supervision. Together, these components enable our model to align captions with video locations effectively.

Full Video Captions Generation

Our proposed DVC module leverages a spatial-temporal video encoder and a pretrained language model to generate multiple content-continuous captions for a given video. As shown in Figure 2(a), the process can be divided into two steps: (1) spatial-temporal video encoding, which captures both spatial and temporal information from the video frames, and (2) prompt-based caption decoding, which finetunes the language model using video embeddings to generate contextually relevant captions.

Spatial-Temporal Video Encoding To fully capture visual information, we design a spatial-temporal encoder with a frame-level spatial encoder E^a and a video-level temporal encoder E^b. Given a video $v \in \mathbb{R}^{N_v \times 3 \times H \times W}$, where H and W represent the height and width of each video frame,the spatial encoder E^a extracts visual embeddings from each frame, resulting in frame-level embeddings $\boldsymbol{v}^a \in \mathbb{R}^{N_v \times d}$:

$$\boldsymbol{v}^a = \{E^a(v_i)\}_{i=1}^{N_v}, \quad (1)$$

where d is the dimension of the embeddings. We formalize E^a using the pretrained CLIP ViT-L/14 model (Radford et al. 2021) and keep the parameters of E^a frozen during training and testing. Next, the temporal encoder E^b processes these embeddings $\boldsymbol{v}^a$ to capture temporal information, generating contextualized video embeddings $\boldsymbol{v}^b$:

$$\boldsymbol{v}^b = E^b(\boldsymbol{v}^a + \theta_p) \in \mathbb{R}^{N_v \times d}, \quad (2)$$

where $\theta_p \in \mathbb{R}^{N_v \times d}$ are position embeddings and E^b is a randomly initialized Transformer encoder. The output $\boldsymbol{v}^b$ encodes frame characteristics and their temporal relationships, essential for generating accurate captions.

Prompt-Based Caption Decoding Inspired by advancements in multimodal language models (Li et al. 2023; Zhu et al. 2023; Liu et al. 2024), we adapt a pretrained GPT-2 (Radford et al. 2019) to serve as a prompt-based caption decoder. This model processes video embeddings to generate sequential captions for all events in a video. Given captions $\{C_i\}_{i=1}^{N_S}$, we concatenate a prompt P "`[FULL]` N_S `events:`" with all captions into a paragraph, tokenized and embedded into a sequence $\boldsymbol{r}$:

$$\boldsymbol{r} = \{r_1, \ldots, r_{N_r}\} = \text{tokenizer}(\{P, C_1, \ldots, C_{N_S}\}), \quad (3)$$

where `[FULL]` signals the model to generate all captions, "N_S `events:`" specifies the number of captions, and N_r is the number of all tokens. We use contextualized video embeddings $\boldsymbol{v}^b$ as prefix visual tokens, concatenated with caption token embeddings:

$$Z = \{\boldsymbol{v}^b; r_1, \ldots, r_{N_r}\}. \quad (4)$$

The objective is to minimize the negative log-likelihood of generating caption tokens given video embeddings and previous tokens:

$$\mathcal{L}(\boldsymbol{v}^b, \boldsymbol{r}) = -\frac{1}{N_r}\sum_{i=2}^{N_r} \log p(r_i \mid \boldsymbol{v}^b, r_{<i}; \theta_{E,G}), \quad (5)$$

where $\theta_{E,G}$ are parameters of the video encoder and caption decoder. By combining the video encoding and caption decoding steps, our full-video captioning model effectively generates multiple content-continuous captions for any given video, capturing the various events and their temporal relationships within the video.

Captioning-Guided Event Localization

We introduce a Captioning-Guided Event Localization method to ground the generated captions to the corresponding video segments. As shown in Figure 2(b)&(c), this method consists of two main steps: (1) Mask Prediction for Events, predicting the center and width of each caption to create a differentiable Gaussian mask representing its temporal location, and (2) Complementary Masked Captioning, applying these masks to video embeddings for training on positive and negative masked captioning tasks.

Mask Prediction for Event As shown in Figure 3, the mask prediction component consists of the event caption prediction and the differentiable mask construction processes. The goal is to generate a differentiable mask for each caption to represent its temporal location within the video and encourage diversity among the masks.

Event Location Prediction. We first use a randomly initialized learnable embedding as a event embedding $e_i \in \mathbb{R}^d$ for each caption proposal C_i. Then, for all event embeddings $\boldsymbol{e} \in \mathbb{R}^{N_S \times d}$ and the frame embeddings $\boldsymbol{v}^a \in \mathbb{R}^{N_v \times d}$, we use a Transformer decoder to generate the embeddings $\boldsymbol{h}$:

$$\boldsymbol{h} = \text{TransformerDecoder}(\boldsymbol{e}, \boldsymbol{v}^a + \theta_p) \in \mathbb{R}^{N_S \times d}. \quad (6)$$

Next, we utilize two linear layers to predict the center μ_i and width σ_i of the proposal i:

$$\mu_i = \text{Sigmoid}(\text{FC}_1(\boldsymbol{h}_i)), \quad \sigma_i = \text{Sigmoid}(\text{FC}_2(\boldsymbol{h}_i)). \quad (7)$$

Differentiable Mask Construction. Inspired by previous works (Zheng et al. 2022a,b; Kim et al. 2024), we use μ_i and σ_i to generate a Gaussian mask $M_i \in \mathbb{R}^{N_v}$ for each event, representing its temporal location:

$$M_i(t) = \exp\left(-\frac{(t/N_v - \mu_i)^2}{2(\sigma_i/\tau)^2}\right), \forall t \in \{1, \ldots, N_v\}, \quad (8)$$

where $M_i(t)$ is the mask value at frame v_t, and τ is a hyperparameter that controls the steepness of the Gaussian curve. Note that other functions generating soft masks can be used as alternatives to the Gaussian mask.

Diversity Loss. To ensure masks cover different video parts, we introduce a diversity loss based on cosine similarity, serving as a regularization term:

$$\mathcal{L}_{\text{div}} = \frac{1}{N_S(N_S-1)}\sum_{i=1}^{N_S}\sum_{j=1,j\neq i}^{N_S} \max(s(i,j) - \gamma, 0) \quad (9)$$

where $s(i,j) = \frac{M_i}{\|M_i\|} \cdot \frac{M_j}{\|M_j\|}^\top$ and γ is the hyperparameter that controls overlap between masks.

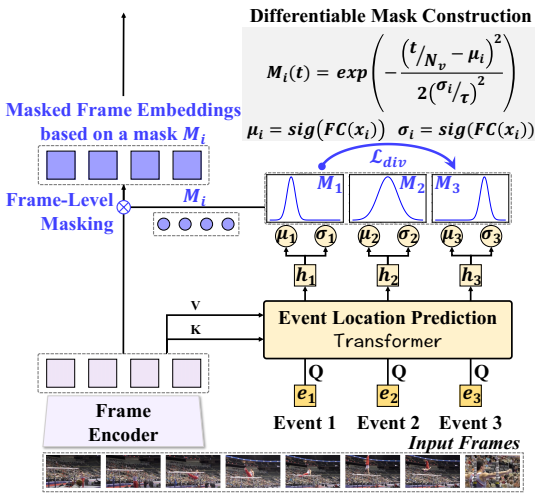


Figure 3: Detailed illustration of Event Location Prediction and Differentiable Mask Construction.

Complementary Masked Captioning Since the supervision for event localization is unavailable, we propose to guide event localization by complementary masked captioning, which mainly consists of two subtasks: positive masked captioning and negative masked captioning.

Positive Masked Captioning. In this task, we apply the Gaussian mask M_i to the video frame embeddings $\boldsymbol{v'}$ in the temporal dimension and input them into the video-level temporal encoder E^b:

$$\hat{\boldsymbol{v}}_i^b = E^b(M_i \cdot \boldsymbol{v}^a). \quad (10)$$

We then concatenate a prompt sentence $\hat{P}$ "`[MASK] 1 event:`" with caption C_i and tokenize it into a sequence $\hat{\boldsymbol{r}}_i$:

$$\hat{\boldsymbol{r}}_i = \{\hat{r}_1, \ldots, \hat{r}_{N_{\hat{r}}}\} = \text{tokenizer}(\{\hat{P}, C_i\}), \quad (11)$$

where `[MASK]` is a special prompt token that indicates the model to generate part of the captions.

Finally, the optimization objective for Positive Masked Captioning is defined as:

$$\mathcal{L}(\hat{\boldsymbol{v}}^b, \hat{\boldsymbol{r}}) = -\frac{1}{N}\sum_{n=1}^{N_S}\sum_{i=2}^{N_{\hat{r}_n}} \log p(\hat{r}_{n,i} \mid \hat{\boldsymbol{v}}_n^b, \hat{r}_{n,1}, \ldots, \hat{r}_{n,i-1}; \theta_{E,G,T}), \quad (12)$$

where N is the total number of the tokens and $T_{n,i}$ denotes the i-th token of the tokenized caption C_n and $\theta_{E,G,T}$ denotes the parameters of the video encoder, the caption decoder and the event localization model, respectively.

Negative Masked Captioning. This task further explores the alignment between captions and video frames by predicting the remaining captions $\{C_j\}_{j=1,j\neq i}^{N_S}$ using the inverse mask on video embeddings. Specifically, we compute the negative Gaussian mask $\check{M}_i$ and inverse masked video embedding $\check{\boldsymbol{v}}_i^b$ as:

$$\check{M}_i = 1 - M_i, \quad \check{\boldsymbol{v}}_i^b = E^b(\check{M}_i \cdot \boldsymbol{v}^a). \quad (13)$$

	Model	Features	SODA	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Fully-Supervised	DCE	C3D	–	5.69	12.43	–	10.81	4.57	1.90	0.71
	DVC	C3D	–	6.93	12.61	–	12.22	5.72	2.27	0.73
	PDVC	C3D	5.26	7.50	25.87	–	–	–	–	1.65
	Vid2Seq	CLIP	5.80	8.50	30.10	–	–	–	–	–
Weakly-Supervised	WSDEC	C3D	–	6.30	18.77	12.55	12.41	5.50	2.62	1.27
	ECG	C3D	–	7.06	14.25	–	11.85	5.64	2.71	1.33
	EC-SL	C3D	–	7.49	21.21	13.02	13.36	5.96	2.78	1.33
	PWS-DVC*	C3D	–	7.28	20.59	12.71	–	–	–	1.35
	Ours†	C3D	5.20	7.36	28.00	13.22	13.66	6.58	3.29	1.77
	Ours	C3D	5.29	7.71	30.17	13.91	14.37	7.05	3.58	1.96
	Ours†	CLIP	6.06	8.22	30.21	14.52	14.83	7.79	3.98	2.03
	Ours	CLIP	**6.08**	**8.48**	**33.42**	**14.77**	**15.36**	**8.12**	**4.17**	**2.26**

Table 1: Comparison with existing methods on the ActivityNet Caption dataset. The symbol † indicates the GPT-2 model used in our method is randomly initialized. The symbol $*$ indicates the results without training with extra video captioning datasets, ensuring a fair comparison with other methods. References for the compared methods are: (Krishna et al. 2017; Li et al. 2018; Wang et al. 2021; Yang et al. 2023; Duan et al. 2018; Wu et al. 2021; Chen and Jiang 2021; Choi, Chen, and Yoon 2023)

Next, we concatenate a prompt sentence $\check{P}$ "`[MASK]` N_S-1 `events:`" with captions $\{C_j\}_{j=1,j\neq i}^{N_S}$ and tokenize it into a sequence of tokens $\check{\boldsymbol{r}}_i$:

$$\check{\boldsymbol{r}}_i = \{\check{r}_1, \ldots, \check{r}_{N_{\check{k}}}\} = \text{tokenizer}(\{\check{P}, \{C_j\}_{j=1,j\neq i}^{N_S}\}). \tag{14}$$

Finally, the optimization objective for Negative Masked Captioning is defined as:

$$\mathcal{L}(\check{\boldsymbol{v}}^b, \check{\boldsymbol{r}}) = -\frac{1}{N}\sum_{n=1}^{N_S}\sum_{i=2}^{N_{\check{r}_n}} \log p(\check{r}_{n,i} \mid \check{\boldsymbol{v}}_n^b, \check{r}_{n,1}, \ldots, \check{r}_{n,i-1}; \theta_{E,G,T}). \tag{15}$$

By combining Positive and Negative Masked Captioning tasks, the Complementary Masked Captioning component enables the model to learn a better alignment between event captions and locations.

Model Training and Inference

Model Training Our model training consists of two stages: *captioning* and *localizing*. In the captioning stage, we train the DVC module by minimizing $\mathcal{L}(\boldsymbol{v}^b, \boldsymbol{r})$ to generate multiple captions. Next, in the localizing stage, we generate Gaussian masks based on the ground-truth captions in the training set and then compute the positive and negative captioning losses. We add these losses together to form the optimization objective, which is used to train the whole model:

$$\mathcal{L} = \mathcal{L}(\hat{\boldsymbol{v}}^b, \hat{\boldsymbol{r}}) + \mathcal{L}(\check{\boldsymbol{v}}^b, \check{\boldsymbol{r}}) + \mathcal{L}_{div}. \tag{16}$$

Model Inference The inference process consists of three stages: *captioning*, *localizing*, and *refining*. In the captioning stage, we use the video encoder and caption decoder to generate initial captions with the video embedding and "`[FULL]`" prompt. Unlike training, the model determines the number of captions based on video content. In the localizing stage, we predict timestamps and generate Gaussian masks for the coarse captions. Finally, in the refining stage, we perform positive captioning with masked video embeddings and "`[MASK] 1 events:`" to re-generate each caption, aiming to enhance caption quality.

Model	YouCook2			ViTT		
	SODA	METEOR	CIDEr	SODA	METEOR	CIDEr
WSDEC‡	2.11	1.47	8.43	4.13	1.95	10.31
PWS-DVC‡	3.14	2.48	9.81	6.11	2.36	12.53
Ours	**3.60**	**4.77**	**13.38**	**8.54**	**4.21**	**17.28**

Table 2: Comparison with existing methods on the YouCook2 and ViTT datasets. ‡ means we reimplement and rerun the baseline methods on the two new datasets.

Experiments

Experiment Setup

Datasets We evaluate our proposed method and baseline methods on the **ActivityNet Captions** dataset. The dataset connects videos to a series of temporally annotated sentence descriptions. Besides, we also conduct experiments on **ViTT** (Huang et al. 2020) and **YouCook2** (Zhou, Xu, and Corso 2018), which are two DVC datasets and have never been used for the evaluation of WSDVC methods.

Evaluation Metrics To make a fair comparison with previous methods, we use the evaluation tool provided by the 2018 ActivityNet Captions Challenge, which measures the capability to localize and describe events. To clarify, we calculate the METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), ROUGE-L (Lin 2004), and BLEU-N (Papineni et al. 2002) scores for the generated captions by comparing them to the reference captions. Moreover, we also adopt the recently proposed SODA metric (Fujita et al. 2020) to perform an overall evaluation of our proposed method.

Setting	SODA	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Full	**6.08**	**8.48**	**33.42**	**14.77**	**15.36**	**8.12**	**4.17**	**2.26**
− Temporal video encoder	5.88	8.33	32.64	14.40	14.98	7.88	4.05	2.18
− `[FULL]` or `[MASK]` prompt	5.96	8.43	33.04	14.53	15.08	7.95	4.08	2.16
− "N `events`" prompt	5.82	8.30	32.36	14.17	14.87	7.95	3.98	2.15
− Inference refinement	5.76	8.12	31.29	14.21	14.71	7.52	3.82	2.05
− Gaussian mask, + Hard binary mask	3.89	6.52	16.96	11.24	11.64	5.01	1.96	0.79
− Gaussian mask, + Sigmoid mask	5.98	7.95	27.79	14.03	14.68	7.41	3.58	1.85
− Gaussian mask, + Cauchy mask	6.02	8.32	32.22	14.50	15.15	7.98	4.07	2.21
− Positive masked captioning	4.08	7.10	20.22	12.43	12.48	5.71	2.31	1.14
− Negative masked captioning	5.92	8.13	30.29	14.37	15.00	7.71	3.88	1.91
− Diversity loss $\mathcal{L}_{div}$	5.73	8.22	30.93	14.48	15.14	7.79	3.84	2.01

Table 3: Ablation study of our proposed method. The symbol '−' means removing the component.

Setting	Backbone	Size (M)	SODA	CIDEr
PWS-DVC*†	vanilla Transformer	~ 59	–	20.59
Vid2Seq	T5-Base	264.48	5.80	30.10
Ours†(C3D)	Distilled-GPT2	62.10	5.20	28.00
Ours†	Distilled-GPT2	62.10	6.06	30.21
Ours	Distilled-GPT2	62.10	**6.08**	**33.42**
Ours†	GPT2-Base	104.62	5.74	26.92
Ours	GPT2-Base	104.62	6.00	32.77

Table 4: Comparison of the size and performance of different models. † indicates that the backbone is not pretrained.

Implementation Details We set the number of transformer blocks in the video-level temporal encoder and cross-modal localizer to 6 and 1, respectively. The number of attention heads, dimension of hidden states, and feed-forward layers are set to 12, 768, and 2, 048 in all transformer blocks, respectively. We utilize the Distilled-GPT2 model for the construction of our caption decoder model. For the training of the model, we adopt the AdamW (Loshchilov and Hutter 2017) optimizer with an initial learning rate of 1e-4 with a warmup rate of 0.1. We train the model for 10 epochs for the captioning stage and 10 epochs for the localizing stage on 8 Tesla V100 GPUs with a batch size of 8.

Comparison with Existing Methods

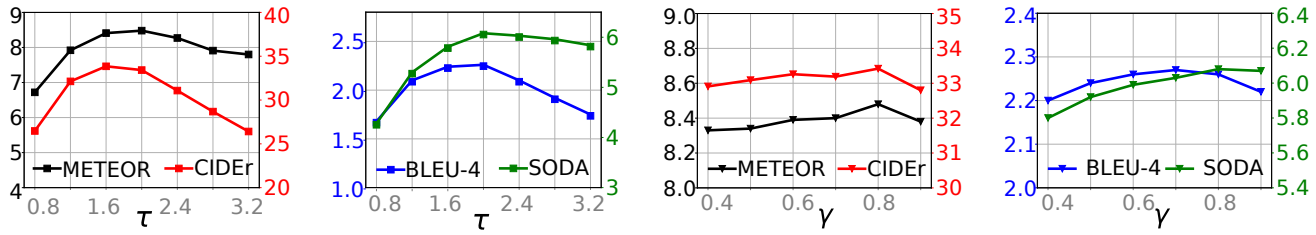
Table 1 presents a comparison of our proposed method with existing fully-supervised and weakly-supervised methods on the ActivityNet Caption dataset. As shown in Table 1, our method outperforms all existing weakly-supervised methods across all evaluation metrics. Specifically, our model trained with CLIP features achieves the highest METEOR, CIDEr, ROUGE-L, BLEU-N, and SODA scores, demonstrating its effectiveness in generating accurate and contextually relevant captions. Besides, our model trained with C3D features also shows strong performance, outperforming other weakly-supervised methods in most metrics. Notably, even the version of our method with a randomly initialized GPT-2 model (denoted by †) achieves competitive results, indicating the robustness of our approach. Moreover, as shown in Table 2, our method also outperform existing methods on the YouCook2 and ViTT datasets. In summary, our proposed method demonstrates excellent performance in the weakly-supervised dense video captioning task, outperforming existing weakly-supervised methods.

Ablation Study

In this section, we conduct an ablation study to investigate the contribution of each component in our proposed method. The results are shown in Table 3. More details are shown in Supplementary Material.

Effect of the model design As shown in Table 3, our full model achieves the best performance across all metrics, demonstrating the effectiveness of our proposed method. When we remove the temporal video encoder, the performance in all metrics decreases slightly. We can see that using CLIP features leads to better performance compared to using C3D features. This indicates that the choice of spatial frame encoder has a significant impact on the model's performance, and using a more powerful encoder like CLIP can further improve the quality of generated captions. Additionally, we observe a drop in performance when we remove the `[FULL]` or `[MASK]` prompt, which suggests that these prompts are important for guiding the model to generate appropriate captions. Similarly, removing the "N `events`" prompt leads to a decrease in performance, indicating its importance in informing the model about the number of events in the video. Moreover, removing the refining stage in the model inference process could also leads to a noticeable performance decline. Overall, the results indicates that each component plays crucial roles in capturing video information and improving the quality of generated captions.

Effect of the mask construction methods We also investigate the effect of different mask construction methods. The Gaussian mask used in our full model achieves the best performance. When we replace the Gaussian mask with a hard binary mask (where inside the predicted scope is 1, otherwise 0), the performance drops significantly, especially in terms of SODA, METEOR, CIDEr, and BLEU scores. This



(a) METEOR and CIDEr scores. (b) BLEU-4 and SODA scores. (c) METEOR and CIDEr scores. (d) BLEU-4 and SODA scores.

Figure 4: Impact of τ in the Gaussian mask construction (a-b) and impact of γ in the diversity loss (c-d).

result can be attributed to the fact that the hard binary mask is non-differentiable, which prevents the optimization of the event proposal generation process. Replacing the Gaussian mask with a Sigmoid mask (Duan et al. 2018) or a Cauchy mask (based on Cauchy Probability density function) also leads to a decrease in performance, although the decrease is less severe than that with the hard binary mask.

Effect of the loss functions Finally, we examine the effect of the captioning loss functions. Removing the positive masked captioning task results in a significant decrease in performance across all metrics. Removing the negative masked captioning task also leads to a decrease in performance across all metrics, although the decrease is less severe than that caused by removing the positive captioning task. Similarly, removing the diversity loss $\mathcal{L}_{div}$ leads to a decrease in performance. This indicates that the diversity loss is important for encouraging the model to generate diverse captions, which can cover different aspects of the video content. In summary, our ablation study demonstrates that all components of our proposed method, including the captioning loss functions, contribute to its strong performance in the weakly-supervised dense video captioning task.

Model Analysis

Impact of the model size Table 4 presents the comparison of the model size and performance of different models. Our method with a distilled GPT-2 (Radford et al. 2019) backbone, even without pretraining, achieves a SODA score of 6.06 and a CIDEr score of 30.21, outperforming the PWS-DVC method which uses a vanilla Transformer (Vaswani et al. 2017) backbone and has a similar model size. When equipped with the pretrained distilled GPT-2 backbone, our method achieves even better performance, with a SODA score of 6.08 and a CIDEr score of 33.42. This performance is competitive with the Vid2Seq method, which is trained under the fully-supervised settings and uses a larger T5-Base (Raffel et al. 2020) backbone. We also explore the effect of using a larger GPT-2 Base backbone in our method. The performance improves slightly compared to using the distilled GPT-2 backbone. However, the improvement is not proportional to the increase in model size, which may be due to overfitting caused by limited training data. In summary, our proposed method demonstrates strong performance on the WSDVC task, even with a relatively smaller model size.

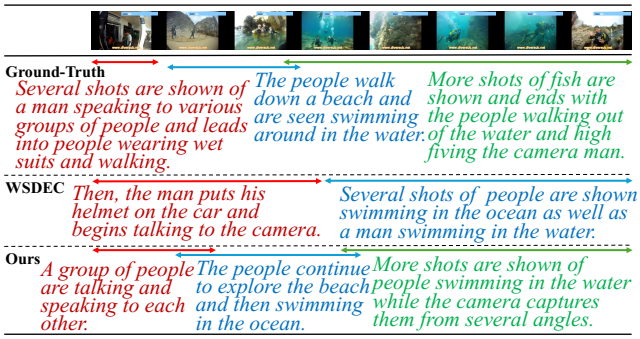


Figure 5: A Qualitative Example from Activity Caption.

Impact of the hyperparameters We examine the impact of values of τ on the performance of our method. The results are shown in Figure 4(a-b). We observe that all the scores increase as τ increases from 0.8 to 2.0, and then start to decrease when τ is larger than 2.0. Then, to investigate the impact of the γ value on our model's performance, we conduct experiments with different γ values ranging from 0.4 to 0.9. The results are shown in Figure 4(c-d). We observe that most scores generally increase as γ increases from 0.4 to 0.8, and then slightly decrease when γ reaches 0.9. The results show that both the value of τ and γ have the moderate impact on the performance of our model.

Case study Figure 5 compares the predictions of our model with the ground-truth annotations and WSDEC method on an example from the ActivityNet Caption dataset. As shown in the cases, our method accurately detects most of the scenes and activities in the video. Moreover, our method can roughly predict the number and the location of the events. In summary, the case study demonstrates the effectiveness of our proposed method for the WSDVC task.

Conclusion

In this paper, we propose a novel WSDVC method that effectively addresses the problem of unavailable supervision on event localization by implicitly aligning event location with event captions via complementary masking, which simplifies the complex event proposal and localization process while maintaining effectiveness. Extensive experiments on the public datasets validate the effectiveness of our method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants Nos. 61972192, 62172208, 61906085. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is supported by the Fundamental Research Funds for the Central Universities under Grant No. 14380001.

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.

Buch, S.; Escorcia, V.; Shen, C.; Ghanem, B.; and Carlos Niebles, J. 2017. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2911–2920.

Chen, S.; and Jiang, Y.-G. 2021. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8425–8435.

Chidume, C. 1987. Iterative approximation of fixed points of Lipschitzian strictly pseudocontractive mappings. *Proceedings of the American Mathematical Society*, 99(2): 283–288.

Choi, W.; Chen, J.; and Yoon, J. 2023. PWS-DVC: Enhancing Weakly Supervised Dense Video Captioning With Pretraining Approach. *IEEE Access*, 11: 128162–128174.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31.

Fujita, S.; Hirao, T.; Kamigaito, H.; Okumura, M.; and Nagata, M. 2020. SODA: Story oriented dense video captioning evaluation framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 517–531. Springer.

Gao, L.; 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.

Huang, G.; Pang, B.; Zhu, Z.; Rivera, C.; and Soricut, R. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

Iashin, V.; and Rahtu, E. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 958–959.

Kim, S.; Cho, J.; Yu, J.; Yoo, Y.; and Choi, J. Y. 2024. Gaussian Mixture Proposals with Pull-Push Learning Scheme to Capture Diverse Events for Weakly Supervised Temporal Video Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2795–2803.

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, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7492–7500.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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, L.; Qu, L.; Meng, D.; Zhang, M.; Tian, Q.; and Bimbo, A. D. 2022. Search-oriented micro-video captioning. In *Proceedings of the 30th ACM international conference on multimedia*, 3234–3243.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

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.

Wu, B.; Niu, G.; Yu, J.; Xiao, X.; Zhang, J.; and Wu, H. 2021. Weakly supervised dense video captioning via jointly usage of knowledge distillation and cross-modal matching. *arXiv preprint arXiv:2105.08252*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

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.

Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; and Gan, C. 2019. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7094–7103.

Zhao, Y.; Zhang, H.; Gao, Z.; Guan, W.; Wang, M.; and Chen, S. 2024. A Snippets Relation and Hard-Snippets Mask Network for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3517–3525.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15555–15564.

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.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, W.; Pang, B.; Thapliyal, A. V.; Wang, W. Y.; and Soricut, R. 2022. End-to-end dense video captioning as sequence generation. *arXiv preprint arXiv:2204.08121*.